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

Lee K. Jones*, member I.E.E.E.
Department of Mathematical Sciences University of Massachusetts Lowell

Optimal local estimation is formulated in the minimax sense for inverse problems and nonlinear regression.This theory provides best mean squared finite sample error bounds for some popular statistical learning algorithms and also for several optimal improvements of other existing learning algorithms such as smoothing splines and kernel regularization. The bounds and improved 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 parameters. Results are given for optimal local learning of approximately linear functions with side information (context) using real algebraic geometry. In particular finite sample error bounds are given for ridge regression and for a local version of lasso regression.The new regression methods require only quadratic programming with linear or quadratic inequality constraints for implementation. Greedy additive expansions are then combined with local minimax learning via a change in metric. An optimal strategy is presented for fusing the local minimax estimators of a class of experts- providing optimal finite sample prediction error bounds from (random) forests. Local minimax learning is extended to kernel machines. Best local prediction error bounds for finite samples are given for Tikhonov regularization. The geometry of reproducing kernel Hilbert space is used to derive improved estimators with finite sample mean squared error bounds for class membership probability in two class pattern classification problems. A purely local, cross validation free algorithm is proposed which uses Fisher information with these bounds to determine best local kernel shape in vector machine learning. Finally a locally quadratic solution to the finite Fourier moments problem is presented. After reading the first three sections the reader may proceed directly to any of the subsequent applications sections.

* research partially supported by Massachusetts Highway Department and the Federal Highway Administration

CONTENTS

I. Introduction and Summary

As increased computing power has lead to more sophisticated artificial neural network and machine learning algorithms for function estimation and approximation in higher dimensional spaces so has the challenge become more difficult of proving theorems concerning the predictive accuracy at a query point of these procedures applied to a finite training sample (the "local estimation problem" ; see [2], [23], [44] .). At the heart of the problem is the curse of dimensionality for local estimation which is exhibited in the following special case of estimating a scalar function f(x) at the point x_0 : suppose the predictor distribution (distribution of x's at which noisy values of f(x) are given) is uniform on the ball of radius one with center x_0 in d-dimensional space R^d . Assume we are interested in estimating the expected response $f(x_0)$. It might be reasonable to use only sample vectors inside a ball of radius $r < 1$, assuming euclidean distance as a measure of closeness. But, since the probability of a sample vector x lying in the smaller ball is r^d , it is necessary that sample sizes are exponential in d to get enough close vectors for accurate estimation (Indeed the sample size would have to be at least r^{-d} to have on average at least one vector x lie in the smaller ball.). Clearly the curse persists for many other general predictor distributions and distances.

One way we may hope to avoid the curse is to assume that f(x) has approximately a very simple model close(in terms of an appropriate distance measure) to x_0 like linear or quadratic (if a very complicated model is necessary then insufficient data close to x_0 may prevent accuracy), but still the accuracy could be low with f(x) linear close to x_0 (or even globally linear) as is the case with simple linear prediction at a query point relatively far away from the data cluster. Hence any other information, like known bounds on f(x) in a neighborhood of the query (as is certainly the case if f is class 2 probability given x (Pr(2|x)) in a two class decision problem), must be incorporated. We refer to such information as context and call the associated (constrained) model a contextual model. Our results will treat only cases of band limited range (i.e. the contextual information consists of bounds on f(x) or the change in f(x) in various

neighborhoods of x_0) although results for other types of contextual information are anticipated. For many engineering applications there are physical or model assumptions that allow one to get bounds on f (or the change in f) in some region even when f is globally linear. In many other cases such bounds may be estimated initially from the data.

Note in the 2-class problem that a logistic model $\exp\{a*x + b\} / (1 + \exp\{a*x + b\})$, although having the bound of one automatically, could well be too complicated for accuracy since slopes of locally fitting functions can be arbitrarily large. Assuming a simple contextual condition on f(x) in addition to a locally approximating (linear or nonlinear) model, we will define a notion of local minimaxity at the query point x_0 (with respect to a given data set) which gives a best upper bound on mean squared error of any affine (in the training response values) estimator of $f(x_0)$. Our local minimax approach will be shown in simple examples to be superior to three popular methods of local estimation.(They will be reviewed in Sec.II. See [2] for a survey.) The approach also does not rely on crossvalidation (global leave-one-out averaging) to find weighting bandwidths as two of the popular methods do. We will present a large naturally occurring class of examples with globally linear f (x) for which the boundedness assumption in a neighborhood of x_0 leads to a large reduction in mean squared error (by a factor of O(1/d)) when the method is compared to ordinary least squares and near neighbor methods. Since versions of our method are scale invariant a similar reduction will often occur when they are compared to regularization and shrinkage methods. (See [42] for a survey.)

Implementing the boundedness (or other contextual) assumptions requires the solution of an optimization and classical real algebraic geometry problem. This associated geometry problem is solved here for the local approximately linear model. In this case the estimator coefficients are determined by solving a minimization problem (using quadratic programming (QP) for regression) where the number of variables is the number of predictors in the training set.

For learning nonlinear functions at x_0 dimensionality reduction may be necessary to get enough “close” points (so that a local approximation is accurate for a large enough sample). To

this end we examine a second way to avoid the curse by assuming that f(x) has an approximation by ridge functions which holds globally (or just in a weak neighborhood of x_0)

$$(\#) \qquad f(x) \sim \sum_{n=1}^{N} c_n g_n(a_n^t x)$$

with weights c_n which are l_1 bounded (w.l.o.g. $\sum |c_n| \leq 1$), $a_n \varepsilon A$, $g_n \varepsilon G$, where A is a given class of ridge vectors in R^d and G is a given class of uniformly bounded real valued functions on R . (e.g. G could be the class of translations of a fixed bounded neural activation function. Existence of approximations of the form (#),which are uniformly accurate to any desired degree, was proven in [10]. A constructive proof of this followed in[21]. Efficient algorithms for and theorems on the average global accuracy of the expansion (#) have been obtained in [16], [24], [38],[20],[3],[4],[29],[30],[28],[12],[22],[13],[31],[11],[14], [8]. An algorithm for obtaining a best local expansion of the form (#) in a weak neighborhood of x_0 is given in [23]. For an extensive survey see [35].) Hence we are assuming that f is (locally) nearly a convex combination of functions of projections onto R .

More generally the a's could be dx m matrices and the g's could be real functions on R^m (where m may vary with n). Some algorithms for obtaining (#) are given in [15] where such g_n's are pieceswise constant in m(n) splitting variables used to iteratively construct a regression or classification tree at stage n. Theorems relating the mean global accuracy of averaging ($c_n = 1/N$) the g_n's corresponding to randomly generated trees (a random forest) are presented in [6]. Also of interest are algorithms for constructing the expansions (#) where the $g_n(a_n^t x)$ are replaced by functions of quadratic forms in x , $g_n(x^t A_n x)$, like Gaussian kernels or other radial basis functions used in support vector machines. See [36] and [43]. We refer to the set of all such functions to be used in the expansions as a dictionary.

Assuming the existence of uniform approximations with any given degree of accuracy of the form (#) for f(x) and assuming a machine algorithm finds a parsimonious

estimate of f of the form (#), we will show how this algorithm, which finds estimates of the above ridge vectors(matrices) a_n using part of the training data, leads to a distance measure which reduces the dimensionality in an approximate sense so that our proposed local contextual methods may be accurate for the remaining data.

We develop a method of aggregating the local minimax estimators of a group of experts which yields a solution to the optimal estimation of the regression function from a random forest. An example in bioinformatics is given using these techniques.

We then apply our methods to optimally estimating a nonlinear function at a point using a sum of kernels. This leads to an improvement of the Tikhonov regularization procedure when the function is class 2 probability given x. An algorithm based on Fisher information is proposed to determine optimal local kernel shape.

We conclude with remarks on extensions to general loss functions and the real algebraic geometry of quadratic approximation. We present a solution to the linear inverse problem in one dimension of optimal reconstruction of a locally approximately quadratic function from noisy integral transform data (the finite Fourier moments problem). After reading the first three sections the reader may proceed directly to any of the subsequent applications sections.

II. Local Minimax Function Estimation with Contextual Model Assumptions

A. Statement of Problem and Review of Popular Solution Methods

The most general setting is that of linear inverse problems with (possibly) indirect measurements in which we are given real values Y_j for $j = 1,2,.....k$ where

$$Y_j = \int \theta_j(t)\, f(t)\, dt_1\, dt_2 \ldots dt_d + N_j .$$

with $\theta_j(t)$ known weight functions, f (x) an unknown real valued function on R^d, the value of which is to be estimated at x_0, and mean zero noise having covariance **N** with known positive definite upper bound σ (i.e. σ - **N** is positive semi-definite).

For ease of exposition and since the majority of our results are for regression and

classification, we stay in this subsection with the following regression case for now with $\theta_j(t)$ being the δ function at x_j ; that is we observe

$$Y_j = f(x_j) + N_j \qquad j = 1,2,.....k$$

where $\{ x_j \}$ are k predetermined design points in R^d or $x_j = X_j$ with $\{ X_j \}$ n i.i.d. random predictor vectors in R^d drawn from some unknown continuous probability distribution P on R^d , f (x) an unknown real valued function the value of which is to be estimated at x_0 , and $\{ N_j \}$ k independent real valued noises which have mean zero and variance bounded by a known constant σ_j^2. Our results will actually be proven for the somewhat more general dependent case where σ - **N** is positive semi-definite for some known positive definite covariance σ . But we keep the independence assumption in this subsection unless we specify otherwise (as in the 2 general forms of ridge regression and later in our theorems). (Subscripts of the letter x will denote enumeration of predictor vectors or queries. The notation $(-)_i$ will be used in a few instances to denote the ith component of -.)

Let the distance between x and y in R^d be given by a nonnegative function D(x , y) (which we may assume to be Euclidean, || ||, in this section ; other distances are introduced in Section V.) By translation we may take x_0 to be the zero vector. We seek optimal affine estimators of f(0) of the form

$$F = F(\mathbf{w}) = w^* + w^t \mathbf{Y} = w^* + \sum_1^k w_j Y_j \qquad (\mathbf{w} = (w^*, w), w = (w_1, ...w_k)^t)$$

where the w*, w_j's are functions of the design vectors (predictors) but not of the Y's. Linearity will be in terms of the Y_j. This assumption is made to obtain optimality results which are noise distribution free as long as we assume mean 0 noises with known covariance bounds. For w* = 0 such estimators are also called local smoothers. (A rigorous minimax theory of local

estimation such as that which we are developing needs to be presented first for the affine case before being extended to the nonlinear case. Furthermore such estimators might produce fewer artifacts common to many current nonlinear methods in inverse problems. And it must be noted that nonlinear estimators of probabilities in classification problems often do not generalize well since they depend too heavily on misclassified subsamples.)The estimators will be nonlinear in $\{x_j\}$ and $D(0,x_j)$. The optimality will be in terms of minimax criteria applied to $L (f(0), F)$ where $L (x ,y)$ is a general nonnegative loss function which we assume to be continuous in (x,y) .

One important learning theory example with known bounded noise variance is the two class classification problem in which case Y_j is 0 for class one and 1 tor class 2 when x_j is observed, $f(x) = Pr \{class\ 2|\ x\ observed\}$ and $Var\ N_j = f(x_j)(1-f(x_j))$, so that we may take $\sigma_j^2 = .25$ (or even the smaller bound $p(1-p)$ if the probability $f(x_j)$ is known to lie outside the interval $(p,1-p)$). It is important in many applications to estimate this function together with a confidence bound. Other cases where a σ_j^2 can be determined occur in regression when the Y's are known to lie in a bounded interval , in inverse problems where σ_j is the maximum magnitude of the instrument noise in the indirect measurements Y_j and in many other modeling situations.

Optimality is in terms of the general criterion which we may now write as $L(f (0),F)$. We first give three popular methods for which the solutions for the first two are linear (in the Y's but not in the x's)with squared error loss and for which the third is based on both linearity and the squared error loss assumption.(See [2], [17], [18].) Our method is then formulated for general loss functions. We present the solutions using our method for $L (x,y) =(x - y)^2$ in the remaining sections. As we will see, since the loss with our method is only applied at $(f (0),F)$ where F is a superposition, the squared error is highly appropriate as F is bell shaped for even modest sample sizes under a bounded noise assumption while many other local and global methods first apply the loss at each predictor-response pair and then sum thus making robustness a key issue.The problem for more general loss functions is discussed in section VIII.

---

1. Distance Weighted Learning (or Kernel Method)

Estimate f(0) by $\arg\min_{y} \sum_{j=1}^{k} L(Y_j, y)\, K_h(D(0,x_j))$ where

$K_h(\ )$ is a one dimensional kernel function of bandwidth h centered at the origin. ($K_h(u) = h^{-1} K(u/h)$ for a fixed univariate K (x) .) For L(x,y) equal the squared error loss function the estimator is linear and the method is the popular standard kernel regression.

---

2. Locally Weighted Residuals (LWR)

Let $f(x ; a)$ be a parametric model for $f(x)$. (i.e. the vector parameter a varies in some set with nonempty interior $A \subset R^q$ and for each a the domain of f (x; a) contains $\{x_j\}$.)

A generalization of 1. is to estimate f(0) by $f(0 ; a)$ where a minimizes the weighted residuals:

$$a = \arg\min_{a} \sum_{j=1}^{k} L(Y_j, f(x_j ; a))\, K_h(D(0,x_j))$$

Constraints(which are nonlinear inequalities in all examples given) may be imposed on a corresponding to information about f which is then assumed only for the models f (x; a). (Since the models are usually only approximate this will not impose all knowledge about f while our method will impose all such knowledge.) These situations will be described as alternatives to each solution using our method. In section II-C we give several simple examples where such constraints are imposed and for which our method is superior to all residual weightings. In both methods characterizing the constraints is a problem in real algebraic geometry.

A popular choice for $f(x ; a)$ is the linear model $a_0 + (a_1, a_2, ... a_{q-1}) \cdot x$ where q = d + 1 and,for squared error loss, the method is the usual locally weighted residual linear regression. The bias and variance of the estimator may be estimated using Taylor's theorem. In many applications with data of 1. and 2. both the bandwidth parameter h and the distance D

have been adaptively chosen. (See [2], [9]) This may introduce inaccuracy due to overtuning. Often h has been chosen by cross validation - performing the local estimation at each training point (leaving it out) for fixed h and obtaining a global estimate of estimation error as a function of h; then choosing the minimizing h for this function for the local analysis. The objection to this practice is clear: if one is only interested in f at 0 one should not let global information have too much influence. Our method does not need a bandwith parameter. Also for the general setting of inverse problems (where L (Y_j , f (x_j ;a)) is replaced above by L (Y_j, $\mathcal{S}_{\theta_j}$ (t) f(t;a) $dt_1..dt_d$)) it is unclear how to assign the weighting while our approach takes advantage of the redundancy and geometric configuration of the integral transform data and still produces the optimal affine estimator.

---

3. Best Linear Unbiased Estimator Assuming Exact Local Model (Local Kriging [17],[18])

Suppose only "near neighbor" predictors x_j , lying in V_o = {x: D(0, x) ≤ r_0},

are to be used in the analysis where the user specifies r_0 as a threshold of closeness; so we assume the sample has been appropriately redefined. Let us limit ourselves to squared error loss, linear estimators F of f (0) (i.e. w* = 0), and to target functions f belonging to a class of general linear models i.e. parametric models as defined in 2. which are linear in a basis of bounded functions $\{h_i\}$ on V_o.

f (x) = f (x ; a) = $a_o h_o(x) + a_1 h_1(x)$ +..... + $a_q h_q(x)$ for some a where $h_o(x) \equiv 1$

and $h_i(0) = 0$ for i > 0 (hence a_o = f (0)). Assume further that Var (N_j) = σ_0^2 . Now let us consider estimators F which are(conditionally) unbiased for the target value f(0) (conditioned on the predictors $X_j = x_j$): E ((F - a_o) | $x_1, x_2, ...x_k$) = 0 for any parameter vector a in $\mathcal{A}$. Then, using the linearity of expectation, it is straightforward to see that this restriction on the estimators is equivalent to

(1) $\Sigma\ w_j\ h_i(x_j) = 0$ for i =1,2 ....q and $\Sigma\ w_j = 1$

Then the weights for the unbaised linear estimator minimizing the expected loss

$$E\left(\left(F - a_o \right)^2 \mid x_1, x_2, ...x_k \right)$$

are easily seen to be the $\{ w_j \}$ that minimize Σw_j^2 subject to the constraints (1). The associated estimator is the local Kriging estimator. It can be extended to include correlated noises [18] for applications to one or two dimensional settings. We also include the case of correlated noise in our main theorems.

It should also be noted that for many simple local models Kriging produces only the ordinary least squares formulas applied to the close predictors. (The same occurs with our method below when the local model is exact and no context is applied. Otherwise different solutions arise. Local Kriging is unaltered by context because of the the unbiasedness restriction: Note that in local Kriging any boundedness constraints on f have no effect on the estimator, when $A = R^q$ and the h_i are bounded in V_o , since the set of parameters a describing the possible f's would still have a nonempty interior.) For instance for a local linear model it is well known that the ordinary linear prediction formula from least squares for the expected response at 0 results. (One possible proof is to apply our Theorem II and then let M become infinite.)

Let $\mathbf{X}$ be the k x (d+1) dimensional design matrix wth 1's in the first column and the components of the k vectors in the rest of the columns. Assuming $\mathbf{X}$ has full rank we have the ordinary linear regression estimator a :

$a = \mathbf{H}\,\mathbf{Y}$ with $\mathbf{Y} = (Y_1, Y_2, ... Y_k)^t$ and $\mathbf{H} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$ (see [27], sec.2.2.1)

and a_o is the estimate of f(0) .

The above ordinary linear prediction formula for f(0) is often modified using ridge regression in high dimensions or when the matrix $\mathbf{X}^t\mathbf{X}$ is (nearly) singular. (This is required in the ordinary formula if there is no data in some direction but this is not an issue with a convex

programming approach such as ours. As we shall see,with the exception of Theorem II, our formulas will be analytically different from ridge regression and the scale invariant versions can potentially outperform it.) The ridge regression technique corresponds to adding fake data to the problem by placing a diagonal matrix Γ at the bottom of the design matrix **X** (call it **X***) and d+1 0's at the end of the response vector **Y** (call it **Y***) and minimizing $(\mathbf{X}^*a - \mathbf{Y}^*)^t(\mathbf{X}^*a - \mathbf{Y}^*)$. (See [2] section 8.1.) The estimator of f(0) is a^*_o where $a^* = (\mathbf{X}^t \mathbf{X} + \Gamma^2)^{-1} \mathbf{X}^t \mathbf{Y}$.

In the general dependent noise case of cov ({ N_j}) = σ we would apply weighted least squares and minimize $(\mathbf{X}^*a - \mathbf{Y}^*)^t \sigma^{*-1}(\mathbf{X}^*a - \mathbf{Y}^*)$ where σ^* extends σ by a diagonal matrix of d+1 one's down the diagonal.The regularized estimator would then be

$$a^* = (\mathbf{X}^t \sigma^{-1}\mathbf{X} + \Gamma^2)^{-1} \mathbf{X}^t \sigma^{-1}\mathbf{Y}$$

If Γ is a multiple of the identity matrix **I** then we call this a <u>standard regularization</u>. Standard regularization corresponds to minimizing $(\mathbf{X}a - \mathbf{Y})^t \sigma^{-1}(\mathbf{X}a - \mathbf{Y}) + \lambda a^t a$. Another form of ridge regression is obtained by minimizing $(\mathbf{X}a - \mathbf{Y})^t\sigma^{-1}(\mathbf{X}a - \mathbf{Y}) + \lambda (a^t a - a_o^2)$. We call it <u>penalized regression</u> or <u>gradient regularization</u>. It is given by

$$(\mathbf{X}^t \sigma^{-1}\mathbf{X} + \lambda \mathbf{I}_o)^{-1} \mathbf{X}^t \sigma^{-1} \mathbf{Y}$$ where $\mathbf{I}_o$ equals the identity matrix **I** but with 0 in the upper left corner.

Ridge regression requires some understanding of the proper scale of the various data components. The choice of Γ and the result would otherwise be highly heuristic and without sound justification. One of our results (Theorem II) gives an an explicit formula which is in the form of a gradient regularization, yielding a minimax justification for ridge regression and providing the value of λ. We now describe our approach.

B. Minimax Function Approximation and Estimation for Approximate Local Models with Context and Finite Sample Error Bounds

Let us allow both parametric and nonparametric local models $\mathcal{M} = \{f(x ; a)\}$ where the parameter vector a varies in $\mathcal{A}$ ($\mathcal{A}$ is a subset of R^q in the parametric case or $\mathcal{A}$ is an abtract

infinite dimensional set in the nonparametric case.).Assume the local smoothness condition that the true f(x) can be uniformly approximated with accuracy $\varepsilon(x)$, in some region V containing 0 and the supports of the θ_j (for regression the predictors $\{x_j\}$), by some member of the family of models:

(2) $| f (x) - f (x; a) | \leq \varepsilon(x)$ for some a and for all x in V for given $\varepsilon(x)$ with $\varepsilon(0)=0$.

In addition the true f is often known or assumed to satisfy additional regularity and/or structural conditions $\mathcal{C}$ in V , like belonging to a certain function class and taking values in a bounded interval (we also denote by $\mathcal{C}$ the class of functions satisfying the conditions). With these assumptions on f the optimal (minimax) strategy for choosing the weight vector w for F is

(3) $\mathbf{w} = \arg \min_{\mathbf{w}} \max_{f \in \mathcal{C}\ ;\ (2) \text{ holds for } f} E \{ L (f(0) , F(\mathbf{w})) \mid \theta_1(t), \theta_2(t), ...\theta_k(t)\}$

This is the more general inverse problem case. For regression the objective function is denoted by $E \{ L (f(0) , F(\mathbf{w})) \mid x_1, x_2, ...x_k \}$ reflecting the fact that the general distributions $\theta_j (t)$ on predictor space are just δ functions at x_j.

(Here the noise is mean zero with covariance **N**. In all our maximizations we will optimize over the values of **N** as well as f which will essentially force the heteroscedastic case, $\mathbf{N} = \sigma$, as Nature's strategy. Of course Nature may never be able to employ this strategy if we are dealing with a classification problem but still our analysis provides estimators and upper bounds on their performances.)

We call the minimax value the local complexity, $\mathcal{LC}$, based on noisy data from the weightings $\theta_1, \theta_2, ...\theta_k$:

$$\mathcal{LC} = \min_{\mathbf{w}} \max_{f \in \mathcal{C}\ ;\ (2) \text{ holds for } f} E \{ L (f(0) , F(\mathbf{w})) \mid \theta_1, \theta_2, ...\theta_k \}$$

It gives a least upper bound on expected loss based on information about the nature (conds. $\mathcal{C}$) of the target function and how it can be approximated (the family f (x; a) and ε(x)).

Some aspects of the proposal (3) have been previously suggested. Gauss had argued with Laplace that a gametheoretic approach, as (3) indeed is, incorporating side information about the function being estimated, was preferable to maximum likelihood although at that time not computationally feasible ([1]). Also for regression with no context assumptions, uncorrelated noise with known σ_i^2 , distinct predictors and a linear $\mathcal{M}$ the proposal (3) is part of the global parameter estimation (in $\mathcal{M}$) convex program proposed by Sacks and Ylvisaker [40]. Our solution algorithms are more general than theirs, are provably convergent and apply to context cases and inverse problems also. All but one of our results involve context (non trivial $\mathcal{C}$). It is the context assumptions or side information which require solutions to classical real algebraic geometry problems and lead to potential increases in accuracy for high dimensional problems.

Of particular interest are the cases derived from Taylor's theorem when f (x; a) is a polynomial of degree s in the d coordinate variables. If f is sufficiently smooth, with a uniformly convergent power series about 0 in V, then the error of approximation at x using the terms up to degree s in $x_1, x_2, ...x_d$ is given by the error of the s'th order Maclaurin expansion for G (t) = f (t x / ||x||) at t = ||x||. But this error is clearly bounded by $\varepsilon(x) = c ||x||^{s+1}/ (s+1)!$ where c would be a user assumed bound on the maximum absolute value of all the s+1 'st directional derivatives of f at points in V along rays through 0. We may take this ε(x) as the modulus of approximation. In the ensuing analysis c could serve as a robustness parameter, indicating the richness of the possible local behavior of f through the relationship (2); or c could be determined via physical or other modelling assumptions. But, as we shall see in the approximately linear case, c can often be quite large while the mean squared error bound is not much larger than that for c = 0.

Unfortunately (3) appears to be computationally difficult to solve exactly when $\mathcal{C}$ requires

a boundedness condition. We can not always describe the class of f's in the maximization problem with a simple set of constraints. We do solve (3) explicitly in the important context-free case in Theorems I and in the exact ($\varepsilon(x) = 0$) local model contextual cases in Theorems II - V. Otherwise we obtain an upper bound on the minimax value.

C.Contrast with Global and Locally Weighted Residual Approximation and Estimation: simple examples for which the proposed minimax method outperforms all local least squares residual weightings

We first review the most common global approximation and estimation counterpart of (3): (see [37]) The global estimator of f , f*, satisfies f*(x) = f(x;a*) where

$$a^* = \arg \min_{a \varepsilon A} \left\{ 1/k \sum \sigma_j^{-2} L (Y_j , f(x_j; a)) + \gamma P(a) \right\}$$

A local version is obtained by weighting the terms of the sum by $K_h (D(0,x_j))$ as in Section A-2. Here { f(x;a) } is either a parametric or nonparametric global family and a penalty term $\gamma P(a)$ is added to the objective function where γ is a regularization parameter which acts like a Lagrange multiplier. (In the homoscedastic case or approximation-interpolation problem the σ_j^{-2} are omitted.) This counterpart arises in regularized minimum empirical loss solutions for a where P(a) is a complexity or lack of smoothness measure of the function f(x;a). It also arises in Bayesian maximum posterior likelihood solutions where the prior on a has the form exp { - ζ P(a) + τ } and the independent measurements Y_j have densities of the form exp { - $\sigma_j^{-2} L (y_j ,$ $f(x_j ; a)) + \beta_j$ } . In the linear inverse problem case (where $L (Y_j , f (x_j ;a))$ is replaced by $L (Y_j , \boldsymbol{S} \theta_j (t) f(t;a) dt_1 ....dt_d))$, if P(a) is an abstract distance from f(x;a) to a universal function f_o (e.g. cross entropy) and L(x,y) =0 only if x = y, then as the σ_j approach zero we get the minimum distance solution subject to data consistency [25] (e.g. minimum cross entropy or maximum entropy when f_o is a constant function).

For various losses one can bound the expectation of the above empirical loss in

brackets (with expectation over the design points as well as responses) as a sum of two terms, one varying directly as an appropriate distance between the true f in $\mathcal{C}$ and its best approximation f(x;a**) and the second as an expected distance between f(x;a*) and f(x;a**). One could then vary the set of models {f(x;a)} over a class $\mathcal{M}$ and minimize the bound. Usually the first term is bounded by a multiple of the Vapnik-Chernovenkis dimension of {f(x;a)}.

Next it is important to know if there is something to be gained by the proposed minimax method. Without context, with a general linear model and with $\varepsilon(x) = 0$ the method reduces to local Kriging (Thm.I). A example of the benefits of the method without context but with nonzero $\varepsilon(x)$ is given in the next section. Here we give two simple contextual examples of estimating f(0) for a truly linear function f on V = [-1, 1] whose slope is assumed to have an absolute value at most one: In both cases the predictor data are $x_1 = 1/2$ and $x_2 = 1$. Assume independent mean 0 noises with variance bounds $\sigma_j^2 = 1/4$. Thm. II will yield optimal weights for the minimax estimator F: $w_1 = 1$, $w_2 = w^* = 0$. For the first example the function is f(x) = x and the noises are $\pm$ 1/2 with equal probability.The four equally probable regressor- response training sets are {(1/2, 1/2 $\pm$ 1/2), (1, 1 $\pm$ 1/2)}. For the local minimax method the mean squared error of estimating f(0) is 1/2 while it is easily checked that the mean squared error of estimating f(0) with the ordinary least squares linear fit, subject to the constraint that its slope is at most one in absolute value, is 11/16. It is also easily checked that the error is at least 11/16 for any so constrained weighted residual least squares linear fit (minimizing α((residual)^2 at 1/2) + (1-α) ((residual)^2 at 1) with $0<\alpha<1$ subject to above constraint). For the second example again take f(x) = x ($0 \leq x \leq 1$) but let the noise be -x with probability 1-x or else 1-x with probability x. The two equally probable training sets are {(1/2, 1/2 $\pm$ 1/2), (1, 1)}.The proposed method has error 1/2 while the ordinary least squares constrained (as above) linear fit has error 17/32. Any least squares weighted residual constrained linear fit, favoring the data at 1/2 ($\alpha > 1/2$), has an error greater than 17/32 while all other least squares weighted residual constrained linear fits have

error greater than 1/2. See Figure 1 as an aid in checking these assertions. Finally note in each example that only one of the possible data sets has nonzero residuals. Hence our assertions are valid when we allow the weighting to depend on the data set.

III. Context Free Solution for Regression and Linear Inverse Problems; the Redundancy Function, Robustness and Advantages over Smoothing Splines

We now present the solution to (3) for the simple (but important) context free case involving local general linear models but with nonzero $\varepsilon(x)$. Our first result extends local Kriging to include approximate general linear models. It is also a generalization, which includes linear inverse problems, of a result in [40;eq. (3.4)]. It has a routine proof that involves an equation that is used in less trivial ways in later theorems. We define a term in our bound as the redundancy function $R(w) = \{ \int | \Sigma w_j \theta_j(t) | \varepsilon(t)\, dt_1 \ldots dt_d \}^2$. From its form one sees that it is minimized by oppositely weighting combinations of $\theta_j(t)$'s which are equal in regions of large $\varepsilon(t)$ (exploiting redundancy in the model weight functions). Further define $H_{ji} = \int \theta_j(t)\, h_i(t)\, dt_1 \ldots dt_d$.

Theorem I Let the general linear parametric family be given by

$f(x ; a) = a_o h_o(x) + a_1 h_1(x) + \ldots + a_q h_q(x)$ where a lies in $A = R^{q+1}$, $h_o(x) \equiv 1$, h_i bounded and $h_i(0) = 0$ for $i > 0$. Consider (3) for the general linear inverse problem: Assume that in V, a region containing 0 and the supports of bounded functions $\theta_j(t)$ (or the predictors $\{x_j\}$ in the regression case where $\theta_j(t)$ are δ functions at x_j), the true f(x) is within $\varepsilon(x)$ of some family member $f(x ; a)$. Assume mean zero noise having covariance **N** with known upper bound σ. Use squared error loss. Make no context assumption. Then the solution to (3) is

$$w = \arg\min_w [\, w^t \sigma w + R(w) \,] \quad , w^* = 0 \, , w = (w_1, \ldots w_k)^t$$

subject to (1G) $\Sigma_j w_j H_{ji} = 0$ for $i = 1, 2 \ldots q$ and $\Sigma w_j H_{j0} = 1$.

For regression with distinct predictors this has the form

$$w = \arg\min_w [\, w^t \sigma w + (\Sigma |w_j| \varepsilon(x_j))^2 \,] \quad , w^* = 0$$

subject to (1) ; which was $\Sigma_j w_j h_i(x_j) = 0$ for $i = 1,2 \ldots q$ and $\Sigma w_j = 1$.

(for nondistinct predictors, $|w_j|$ in the objective function above must be replaced by | sum of w_i over all i for which the predictor x_i equals x_j | and Σ is over the equivalence classes of equal predictors.) In all cases, for optimal **w** the local complexity $\mathcal{LC} = w^t \sigma w + R(w)$ is an upper bound on the mean squared error of the estimator F (**w**) and this bound is the best possible. Furthermore if **v** is any suboptimal solution satisfying the constraints with $v^* = 0$, then an upper bound on the mean squared error of F (**v**) is given by $v^t \sigma v + R(v)$.

In addition the solution for nontrivial $\varepsilon(x)$ is fundamentally different from LWR of Section II- A.

For regression, when the local model is exact ($\varepsilon(x) \equiv 0$), this is just local Kriging for the homoscedastic i.i.d. noise case. The result says that local Kriging may be extended to inexact models by requiring unbiasedness for exact submodels and minimaxing mean squared error.

---

proof of Theorem I : Write $f(x) = f(x ; a) + \zeta(x)$ where $|\zeta(x)| \leq \varepsilon(x)$ in V. Then $Y_j = a_o H_{jo} + a_1 H_{j1} + \ldots + a_q H_{jq} + \int \theta_j(t)\, \zeta(t)\, dt_1\, dt_2 \ldots dt_d + N_j$.

(Use a_o for f(0) interchangeably here. They are equal since $\varepsilon(0) = 0$. Sums in i or j will start at i = 1, j = 1.) Then by routine calculation

$$(4)\quad E\left((F(\mathbf{w}) - f(0))^2 \mid \theta_1, \theta_2, \ldots \theta_k\right) = w^t \mathbf{N} w + \left\{ \Sigma a_i \left(\Sigma w_j H_{ji} \right) + \left(\Sigma w_j H_{j0} - 1 \right) a_o + w^* + \int \Sigma w_j \theta_j(t)\, \zeta(t)\, dt_1 \ldots dt_d \right\}^2$$

Now if the i'th constraint in (1G) fails to hold the maximum of the above is infinite. (Pick f's of the form $a_i h_i(x)$ with larger and larger a_i .) Hence w must satisfy (1G). For w satisfying (1G) the maximum of the expectation is achieved by taking $\mathbf{N} = \sigma$ and taking

$f(x) = \zeta(x) = \mathrm{sgn}(w^*)\,\mathrm{sgn}\left(\Sigma w_j \theta_j(x)\right)\, \varepsilon(x)$. ($\mathrm{sgn}(z) = +1$ for $z \geq 0$, -1 for $z < 0$)

For the regression case with distinct predictors the proof is the same except we take f as any function bounded in absolute value by $\varepsilon(x)$ in B which satisfies

$f(x_j) = \mathrm{sgn}(w^*)\,\mathrm{sgn}(w_j)\,\varepsilon(x_j)$. $(\mathrm{sgn}(w^*)\,\mathrm{sgn}(\Sigma w_i)\,\varepsilon(x_j)$ for nondistinct predictors with Σ over all i with $x_i = x_j)$

Now, in minimizing these maxima for w satisfying (1*) (or (1)), we clearly must have w* = 0 and so the form of the argmin has been established.

Finally we compare the solution in the distinct predictors regression case to LWR assuming a linear f(x;a), squared error loss for LWR and nontrivial analytic $\varepsilon(x)$ and analytic $K_h(||x||)$ with fixed h. The solution to LWR is that of a linear system with coefficients analytic in the design matrix **X** while the system obtained by setting to 0 the w gradient of the objective function, $\mathcal{L} = w^t \sigma w + (\Sigma |w_j| \varepsilon(x_j))^2$, has jumps. Hence the estimate of f(0) by LWR is meromorphic in **X** while the $F(\mathbf{w})$ of the Theorem is not. **QED**.

To obtain solutions replace the |x| function by $\eta \ln(2\cosh(x/\eta))$ (with η suitably small). Then any differientiable optimization technique could be used to find **w** with accuracy easily quantified in terms of η. Approximate solutions **v** are thus generated and valid bounds are obtained by evaluating the objective at **v** using the true |x| function. This same trick (called smoothing) can be used in all context cases to be presented.

Solutions in the regression case for **w** are easily obtained by quadratic programming (QP): write $w_j = w_j^+ - w_j^-$, change $|w_j|$ to $w_j^+ + w_j^-$ in the objective function and include further constraints that w_j^+ and w_j^- are nonegative.

Here are some applications of the above obtained with civil engineers who are calibrating pavement profiling devices (vans with on-board lasers to estimate road profile and signal processing computers to translate road profile into ride quality) based on ride roughness measurements on a set of control paved sites. (The software was developed using smoothing by Brad Jones; quadratic programming solutions in this paper were suggested by Dave Einstein.)

Consider Graph 1 of the local solution at increments of .05. The 6 data points (denoted by solid triangles) are characterized by x coordinates which are the roughness levels (in units of

100 IRI (International Roughness Index)) as outputed by the device ICCS495R at each of 6 sites. These are not the true roughness levels since the device measures only certain frequency components (indeed very accurately) of the road profile while ignoring others. The y coordinates are unbiased estimates of the actual roughness levels for the corresponding sites as determined by measurement of the site road profiles with manual (time consuming) procedures followed by ride computer simulations. It is assumed that the manual procedures yield independent measurement errors with a true standard deviation of .03 (or less). Now, if the device now measures a roughness x_o on a newly paved roadway, a predicted true roughness is desired for the new pavement together with a root mean square error(RMSE) (or upper bound thereof) for the prediction. This RMSE (or bound thereof) should be mathematically guaranteed based on reasonable mathematical assumptions and not just an estimate as roughness quality measurements which are used to determine levels of highway contractor compensation are subject to legal challenge. This RMSE is local as it need only be valid when the device outputs x_o.

Ordinary least squares (assuming the unknown function is actually linear) with fixed controls is linear in the y values and (hence) furnishes (a bound) on RMSE which depends only on the control values $\{x_i\}$ and is probabilistically meaningful before the y measurements are taken into account and hence is valid for confidence analysis of all the measurements in the common frequentist's sense. However if the regression curve (called profiling correlation curve by pavement researchers) is nonlinear many(e.g.adaptive spline) curve estimation (and even robust linear) methods are (locally) nonlinear in the y values and RMSE(or bounds thereof) can only be estimated (and often only globally). Even the linear method of smoothing splines which yields a global estimate f(x) minimizing

$$\sum (Y_j - f(x_j))^2 + \gamma \int (f''(x))^2 \qquad \text{with } \gamma \text{ prespecified,}$$

requires Bayesian assumptions to produce a confidence statement. (recall Section II-C) Local learning is (globally nonlinear but) locally linear in the Y's. It is finite sample locally optimal(among

all other locally linear curve estimation methods, varying with each x_o) and furnishes guaranteed RMSE at x_o in the frequentist's sense for regression with fixed controls x_j. Consider the results: For the local bounds we assumed $\varepsilon(x)= x^2$. Hence the RMSE bounds hold when the true curve is twice differentiable with second derivative bounded in absolute value by c=2.Thus a high degree of non-linearity could be present . The upper curve in Graph 2 represents the worst case local error of ordinary least squares under the $\varepsilon(x) = x^2$ assumption (obtained by evaluating the achievable bound for the suboptimal weights v_j corresponding to the ordinary least squares solution).The lower curve represents the RMSE(bound) if the device were truly linear ($\varepsilon(x) = .5cx^2 = 0$). Amazingly the middle curve (obtained after convex optimization by the quasi-Newton method with initial weights v_j and using smoothing parameter $\eta = .02$), which is an upper bound for the worst case for the local theory method under the $\varepsilon(x) = x^2$ assumption, is very close to the lower (ideal) curve inside or close to the control interval (.8 - 1.7).

The method is optimally "downweighting" the predictors which are "far" from the query. This high degree of robustness is somewhat surprising and demonstrates the power of numerical optimization. Of course this is only a low dimensional problem where there naturally might exist points close enough to a given query for accurate estimation in the frequentist sense.The rest of this paper treats the local problem with side information on f(x) that in essence reduces dimensionality so that the methods can be proven to yield similar advantages in high dimensions.

As for training points far from the query Graph 3 demonstrates the advantages of using boundedness information (in only one dimension). Here the IRI data was provided at three sites. It was assumed that the true IRI of a query is between .6 and 1.2 and Theorem V of the next section was applied (expressing max and min in terms of | | and smoothing by changing |x| as above) . For each curve the first label is σ and the second is c (contextual linear method with $\varepsilon(x) = .5cx^2$) or slr (standard least squares regression). For the high noise case the contextual RMSE is substantially better than the standard linear predictor RMSE reasonably close to the

data cluster. This indicates potential accuracy in classification problems where σ varies from .2 to.5.

IV. Estimation for Linear and Approximately Linear Functions

A. Some Results with Rotational Invariance to the Design Set ; relationships to ridge regression, an application to stock price prediction

Now we consider the setting of an approximate (exact when $\varepsilon(x) = 0$) linear target function in V but with side information that the function takes values in a given bounded interval or has a certain bound on its oscillation in V. We will later describe a large class of naturally occurring examples where the side information leads to large reduction in mean squared error for the contextual estimator compared to that with standard local linear prediction. For these examples there is the same high degree of robustness in the presence of considerable nonlinearity for the contextual estimator in high dimensions as for the context free road quality estimator in one dimension of the last section. Our first result gives a robust extension of penalized local linear regression when nonlinearity is present as well as a theory for the correct choice of penalty parameter and new error bounds, optimality properties and interpretations for ridge regression by a gradient regularization.

Theorem II Let the linear parametric family be given by $f (x ; a) = (a_1, a_2,.....a_d) \cdot x + a_o$ for x in R^d. Let V be a ball of radius r centered at 0 which contains the predictors $\{ x_j \}$. $\mathbf{X}$ is the k x d+1 design matrix. Assume that $f(x)$ is within $\varepsilon(x)$ of some family member $f (x ; a)$ in V where $\varepsilon(x) = \kappa||x||^2$ ($\kappa = c/2$ where c is a bound on the magnitude of all directional second derivatives; see section II-B). Assume mean zero noise having covariance $\mathbf{N}$ with known upper bound σ. Use squared error loss. $\mathcal{C}$ is the condition that $| f(x) - f(z) | \leq 2 M$ for x and z in V. Recall $w = (w_1, w_2, ... w_k)^t$.

The following is an upper bound on mean squared error of $F(\mathbf{w})$ in the general ($\kappa > 0$) case for given $\mathbf{w}$ with $C(w) = \sum w_j - 1 = 0$ and $w^* = 0$. It may be minimized in w to give

near minimax optimality:

$$L_{\mathcal{C}}(\mathbf{w}) = w^t \sigma w + B(w)$$

$$\text{with } B(w) = \{ (M/r + \kappa r) A + \kappa \Sigma |w_j| \, ||x_j||^2 \}^2 \quad \text{if } \kappa \le M r^{-2}$$

$$\{ 2(\kappa M)^{1/2} A + \kappa \Sigma |w_j| \, ||x_j||^2 \}^2 \quad \text{if } \kappa > M r^{-2}$$

where $A = A(w) = ||\Sigma\, w_j x_j||$ and $w_k = 1 - \Sigma_{j<k} w_j$.

If f is linear in V ($\kappa = 0$), the optimal strategy (3) for squared error loss takes the form

$$w = \arg\min_w \left[w^t \sigma w + (MA/r)^2 \right] , \qquad w^* = 0 , \; C(w) = 0.$$

The solution w^t is the first row of $\mathbf{X}^t ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)^{-1}$ normalized to have sum one:

$$w^t = \nu\, (1, 1,...1) ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)^{-1}$$

$$\text{where } 1/\nu = (1, 1,...1) ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)^{-1} (1, 1,...1)^t .$$

Also the minmax value $\mathcal{L}_{\mathcal{C}}$ of (3) is given by

$$\mathcal{L}_{\mathcal{C}} = \left(\Sigma\Sigma ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)^{-1}_{ij} \right)^{-1} - (M/r)^2$$

and F(**w**) may be expressed as the constant term in a gradient regularization where the penalty (or ridge) parameter , $\lambda = (r/M)^2$, has a rigorous minimax justification:

$$F(\mathbf{w}) = (1, 0,...0) (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I}_o)^{-1} \mathbf{X}^t \sigma^{-1} \mathbf{Y}$$

where $\mathbf{I}_o$ equals the identity matrix $\mathbf{I}$ but with 0 in the upper left corner.

---

proof of Thm. II: With notation from the proof of Theorem I we rewrite (4) using the linear family

$$E ((F(\mathbf{w})-(f(0))^2 \mid x_1,... x_k) = w^t \mathbf{N} w + \{ \Sigma a_i (\Sigma w_j x_j)_i + w^* + C(w) a_o + \Sigma w_j \zeta (x_j) \}^2 .$$

For each fixed w*, w we want first an upper bound on the maximum over a of the absolute value of the quantity inside the brackets subject to (2) holding for a and some f satisfying $\mathcal{C}$. In this case there is no restriction on a_o since it has no effect on the change in f. Hence the

minimax value is infinite unless C(w) = 0 which is the condition imposed on w_k. We only need to maximize the magnitude of $\Sigma\, a_i\, (\Sigma\, w_j\, x_j)_i$ over this set of a's and, since this set is clearly symmetric from the following characterization, pick its sign to be the same as that of w*. Then add $\pm\ \kappa\Sigma\, |w_j|\ \ ||x_j||^2$ (note the interval generated by these two numbers contains $\Sigma w_j\, \zeta\, (x_j)$) to it inside the brackets to obtain the upper bound. It is then clear that w* = 0 makes this upper bound smallest for given w. The condition on a for which (2) holds for some f satisfying $\mathcal{C}$ is

$$| (a_1,a_2, ...,a_d) \cdot x | \ \leq\ \ M\ +\ \kappa||x||^2 \qquad \text{whenever } ||x|| \leq r\ .$$

(reason: if condition is satisfied then pick f = max{ min{a·x, M}, -M}; if violated at y then any f satisfying (2) takes values at +y, -y which differ by more than 2M)

The characterization of all such a's is a classical real algebraic geometry problem which is equivalent to $||\ (a_1,a_2, ...,a_d)\ ||\ \ \leq\ \ ||x||^{-1} M\ +\ \kappa||x||$ for $||x|| \leq r$. The solution is

$$||\ (a_1,a_2, ...,a_d)\ ||\ \ \leq\ \ \begin{cases} 2(\kappa M)^{1/2} & \text{if } \kappa r^2 > M \\ r^{-1} M\ +\ \kappa r & \text{else} \end{cases}$$

The maximization is achieved by multiplying the latter bound by A and adding (inside the brackets) $\kappa\Sigma\, |w_j|\ \ ||x_j||^2$ to obtain the mean squared error upper bound. The exact linear case best bound (solution to (3)) follows by setting $\kappa = 0$.

For the exact case the algebraic expressions for w_j follow by setting equal to 0 the gradient of the augmented Lagrange objective function $w^t\, \sigma\, w + (\ MA/r\ )^2\ +\ \lambda\, C(w)$. First rewrite this as $w^t\, \sigma\, w\ +\ (M/r)^2\ w^t\, \mathbf{x}\mathbf{x}^t\, w\ \ -\ (M/r)^2 (C(w)+1)^2\ +\ \lambda\, C(w)$ which we can do since $||\Sigma\ w_j x_j\ ||^{\,2} = \ w^t\, \mathbf{x}\mathbf{x}^t\, w\ -\ (\ \Sigma\ w_j\ )^2$. Now take the gradient and we obtain $\sigma\ w\ +\ (M/r)^2\, \mathbf{x}\mathbf{x}^t\, w\ =\ \boldsymbol{n}\, \mathbf{1}$ with $\mathbf{1}$ a column vector of k ones and $\boldsymbol{n}$ the normalizing constant which makes $w^t\, \mathbf{1} = 1$. Solving for w we get $w^t\ =\ \boldsymbol{n}\, \mathbf{1}^t\ \ (\ (M/r)^2\, \mathbf{x}\mathbf{x}^t\ +\ \sigma\ \ )^{-1}\ =$ $\boldsymbol{n}\ (1,\ 0,...0)\ \mathbf{x}^t\ (\ (M/r)^2\, \mathbf{x}\mathbf{x}^t\ +\ \sigma\ \ )^{-1}$. The expression for $\mathcal{L}\mathcal{C}$ follows by a simple plug-in.

We now show that this form is actually a gradient regularization. (The proof is with help from Alex Kheifets and Dan Klain.) In particular we show that the normalized first row of $\mathbf{X}^t ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)^{-1}$ is the first row of $\boldsymbol{G} = (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I}_o)^{-1} \mathbf{X}^t \sigma^{-1}$:

First consider the matrix equation $(\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} (r/M)^2 \mathbf{X}^t \sigma^{-1} ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)$ $= (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I}) \mathbf{X}^t = \mathbf{X}^t$, yielding the "exchange" identity

$$(\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} \mathbf{X}^t \sigma^{-1} = (M/r)^2 \mathbf{X}^t ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)^{-1} .$$

Now let $\mathbf{J}_o = \mathbf{I} - \mathbf{I}_o$. Write $\boldsymbol{G} = (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I} - (r/M)^2 \mathbf{J}_o)^{-1} \mathbf{X}^t \sigma^{-1}$

$= (\mathbf{I} - (r/M)^2 (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} \mathbf{J}_o)^{-1} (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} \mathbf{X}^t \sigma^{-1}$. Note

$(1, 0,...0) (\mathbf{I} - (r/M)^2 (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} \mathbf{J}_o) = (1 - (r/M)^2 (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1}{}_{11})$

multiplied by $(1, 0,...0)$ or $(1, 0,...0) (\mathbf{I} - (r/M)^2 (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} \mathbf{J}_o)^{-1}$ equals

$(1 - (r/M)^2 (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1}{}_{11})^{-1} (1, 0,...0)$. Applying this to the latter expression for $\boldsymbol{G}$ we get first row of $\boldsymbol{G} = (1, 0,...0)\boldsymbol{G} = (1 - (r/M)^2 (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1}{}_{11})^{-1}$ multiplied by the first row of $(\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1} \mathbf{X}^t \sigma^{-1}$, which by the "exchange" identity is

$(M/r)^2 (1 - (r/M)^2 (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I})^{-1}{}_{11})^{-1}$ times the first row of $\mathbf{X}^t ((M/r)^2 \mathbf{X}\mathbf{X}^t + \sigma)^{-1}$.

Finally, since $\mathbf{I}_o (1, 0,...0)^t = 0$, the first row sum of $\boldsymbol{G}$ is $(1, 0,...0)\boldsymbol{G} (1, 1,...1)^t =$ $(1, 0,...0)\boldsymbol{G}\mathbf{X} (1, 0,...0)^t = (1, 0,...0)(\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I}_o)^{-1} (\mathbf{X}^t \sigma^{-1} \mathbf{X} + (r/M)^2 \mathbf{I}_o) (1, 0,...0)^t$

$= (1, 0,...0) \mathbf{I} (1, 0,...0)^t = 1$. Hence the first row of $\boldsymbol{G}$ is already properly normalized. **QED**

QP may be used to minimize $L_{\mathcal{C}}(\mathbf{w})$ by writing $w_j = w_j^+ - w_j^-$, changing $|w_j|$ to $w_j^+ + w_j^-$ in the objective, including further constraints that w_j^+ and w_j^- are nonegative and treating A as a variable with the constraints $A \geq 0$ and $A^2 \geq \| \Sigma\, w_j x_j \|^2$.

For the exact case we could find the LWR least squares fit of the data subject to the quadratic inequality constraint $a^t a - a_o^2 \leq (M/r)^2$ (which we have called "constrained" regression) that ensures the context assumption for (just) the linear family in V and compare a_o (contextual LWR estimator) to F(**w**) from Theorem II. The estimators are not the same (a_o is a nonlinear estimator). In fact we gave simple examples in section II-C where a_o is inferior to F(**w**) no matter how we nontrivially weight the residuals . But penalized regression (gradient regularization) is a linear method and has the same form as F(**w**) for estimating at the query provided the penalty (ridge) parameter $\lambda = (r/M)^2$. Note that this penalty parameter does not in general equal the Lagrange multiplier value μ for the appropriate constrained regression (where $a^t a - a_o^2 \leq (M/r)^2$) even when the constraint is active (consider the first example of section II-C when the data are (.5, 0), (1, 1.5)); $\lambda = r = M = 1$ but $\mu = .25$) .

We give a simple application to stock price prediction of ridge regression with the minimax interpretation presented here. Suppose our predictor data consists of vectors x_j each of whose components are features like price/earnings ratio, % increase in sales over previous year, etc. at a past time T for the j'th stock in a set of k securities. The observed response y_j is the log of the ratio of the price of j at $T + \Delta T$ to that at time T. Now we wish to estimate f(x), the expected log price ratio for a time ΔT from now for a given stock of interest with current predictor vector x. We assume f(x) is (nearly) linear in x. Suppose the compounded change (trend) of each stock with predictor vector x is $\exp(\mu_x \Delta T) = (m\beta_x)^{\Delta T}$ where m represents the compounded change/yr. of the market as a whole and β_x is the compounded change/yr. relative to the market for the particular stock. Financial theory argues that for very small time increments δT the relative change in stock price $\delta S / S$ is normal with mean $\mu_x \delta T$ and variance $\sigma_x^2 \delta T$ and then proves [19,p.275] that the log ratio is normal with mean $f(x) = \mu_x \Delta T - \sigma_x^2 \Delta T / 2$ and variance

$\sigma_x^2 \Delta T$ for arbitrary time increments ΔT. In general we can apply our theory below with only an upper bound on the covariance of the log ratio for the family of stocks (to appear elsewhere). For the moment just assume the prices of the family of stocks considered are approximately stochastically independent (for$\Delta T \geq 3$ m.) and have approximately equal volatilities $\sigma_x = \sigma$ which can be estimated. It is reasonable to assume bounds on β_x over time intervals ΔT = 1/4 - 1 year. Suppose we assume $1/\beta_O \leq \beta_x \leq \beta_O$. Then $|f(x) - f(y)| \leq 2\Delta T \ln(\beta_O)$ and we apply the ridge regression above with $M = \Delta T \ln(\beta_O)$. An analysis has been carried out with 6 dimensional feature vectors (to appear elsewhere, computational support provided by Boniface Nganga). We show a typical comparison of m.s.e. in predicting log ratio for the stock PDX using 83 similar stocks . The components of the predictor vectors were standardized and the predictor vector of PDX was subtracted from all predictor vectors. r = 2.886, σ^2 = .102. Table 0 shows the ratio of m.s.e. bounds (efficiency) using Theorem II (exact linear case) for the usual estimator (M infinite) and the ridge estimator for 3 bounds β_O on β_x and 3 values of ΔT.

Table 0 Efficiency

ΔT	1.00	.50	.25
β_O			
1.82	1.09	1.17	1.31
1.29	1.41	1.70	2.10
1.20	1.70	2.09	2.55

Next we give a solution to (3) in the exact linear case and an upper bound on mean squared error for the general case when the function f is known to take values in a given interval. Here the solution always differs from penalized or constrained least squares.

Theorem III Use the same assumptions and the same notation as in Theorem II except that $\mathcal{C}$ is the condition: f (V) is contained in [v,y]. Let $C = C(w) = \sum w_j - 1$ and $A = A(w) = ||\sum w_j x_j ||$. Then, if the true f is linear in V ($\kappa = 0$ so $\varepsilon(x) = 0$), the optimal strategy (3) for

squared error loss takes the form

$$(5)\quad w = \arg\min_{w} \left[\; w^t \sigma w + \left(\max \{ \, |(y-v)C/2| \, , (y-v)A/2r \, \} \right)^2 \right]$$

with $w^* = -(v+y)C/2$. Furthermore the quantity inside the square brackets is an upper bound on mean squared error for any suboptimal w.

The following is an upper bound on mean squared error in the general ($\kappa > 0$) case for given **w** with $w^* = -(v+y)C/2$. It may be minimized in w to provide near minimax optimality.

$$L_{\mathcal{e}}(\mathbf{w}) = w^t \sigma w + B(w)$$

with $B(w) =$

$$(6)\quad \begin{cases} \left(((y-v)/2r + \kappa r)A + \kappa \Sigma |w_j| \; ||x_j||^2 \right)^2 & \text{if } \kappa r^2 < (y-v)/2 \\ \left((2\kappa(y-v))^{1/2} A + \kappa \Sigma |w_j| \; ||x_j||^2 \right)^2 & \text{if } \kappa r^2 \geq (y-v)/2 \\ & \text{when } |C| \leq A \max\{ 1/r \, , ((y-v)/2\kappa)^{-1/2} \} \\ \left((y-v)|C|/2 + A^2 \kappa/|C| + \kappa \Sigma |w_j| \; ||x_j||^2 \right)^2 & \text{otherwise} \end{cases}$$

---

proof of Theorem III: Using the notation from the proof of Theorem I we rewrite (4) using the basis of linear functions as

$$E((F(\mathbf{w})-(f(0))^2 \mid x_1,\ldots x_k) = w^t \mathbf{N} w + \left\{ \Sigma a_i (\Sigma w_j x_j)_i + w^* + C a_o + \Sigma w_j \zeta(x_j) \right\}^2 .$$

For each fixed **w** we want first an upper bound on the maximum over a of the absolute value of the quantity inside the brackets subject to (2) holding for a and some f satisfying C . This we achieve by maximizing the magnitude of $\Sigma a_i (\Sigma w_j x_j)_i + C a_o + w^*$ subject to (2) (for some a and f satisfying C) and adding $\kappa \Sigma |w_j| \; ||x_j||^2$ to the maximizing magnitude inside the brackets. So we need to characterize a such that (2) holds for the given a and some f satisfying $\mathcal{e}$. (This is the associated classical real algebraic geometry problem.) Then we need to maximize the magnitude of $\Sigma a_i (\Sigma w_j x_j)_i + C a_o + w^*$ over this set of a's. The condition on a for which (2) holds for some f satisfying $\mathcal{e}$ is

$| (a_1,a_2, ...,a_d) \cdot x | \leq \min \{ a_o - v + \kappa||x||^2 , y - a_o + \kappa||x||^2 \}$ whenever $||x|| \leq r$.

This is equivalent to $|| (a_1,a_2, ...,a_d) || \leq ||x||^{-1} \min \{ a_o - v , y - a_o \} + \kappa||x||$

for $||x|| \leq r$. Now minimizing the right hand side of this latter inequality over $0 \leq ||x|| \leq r$ we get

$$|| (a_1,a_2, ...,a_d) || \leq U(a_o) = \begin{cases} 2(\kappa(a_o - v))^{1/2} & v \leq a_o \leq \min \{ (v + y)/2 , v + \kappa r^2 \} \\ 2(\kappa(y - a_o))^{1/2} & \max \{ (v + y)/2 , y - \kappa r^2 \} \leq a_o \leq y \\ r^{-1} (a_o - v) + \kappa r & v + \kappa r^2 < a_o \leq (v + y)/2 \\ r^{-1} (y - a_o) + \kappa r & (v + y)/2 \leq a_o < y - \kappa r^2 \end{cases} .$$

(This is the solution of the associated real algebraic geometry problem.)

Now to maximize the magnitude of $\sum a_i (\sum w_j x_j)_i + C a_o + w^*$ over this set of a's for given a_o we pick $(a_1, ...,a_d)$ with length $U(a_o)$ and pointing in the direction of + or - $\sum w_j x_j$ (+ if $w^* + C a_o$ is positive and - otherwise).This is equivalent to taking the maximum of the two quantities: $(A U(a_o) + w^* + C a_o)$, $(A U(a_o) - w^* - C a_o)$.

We now optimize over a_o. Note that $U(a_o)$ (hence also $A U(a_o)$) is symmetric about $a_o = (v + y)/2$ and has a decreasing continuous slope in (v , y) except at $a_o = (v + y)/2$. Let us first consider maximizing $A U(a_o) + w^* + C a_o$:

1. Suppose C is nonegative but $\leq A/r$. Then if $\kappa r^2 < (y - v)/2$ the max occurs at $(y + v)/2$ since the right hand slope of $A U(a_o)$ is $-A/r$ at this point while if $\kappa r^2 \geq (y - v)/2$ the max still occurs at $(y + v)/2$ since the right hand slope of $A U(a_o)$ at this point is $-A((y - v)/2\kappa)^{-1/2}$ which is $\leq -A/r$.
2. Suppose C is nonegative but $> A/r$. If $C \leq A((y - v)/2\kappa)^{-1/2}$ then the max again occurs at $(y + v)/2$. Otherwise the max occurs at a_o which solves $C = A((y - a_o)/\kappa)^{-1/2}$. (i.e. when C is just minus the slope of $U(a_o)$)

Summarizing 1. and 2. and performing some easy but lengthy calculations when $C \geq 0$

$$\max \{A\,U(a_o) + w^* + C\,a_o\} = w^* + (y + v)C/2 + ((y-v)/2r + \kappa r)A \quad \text{if } \kappa r^2 < (y - v)/2$$

$$w^* + (y + v)C/2 + (2\kappa(y - v))^{1/2}A \quad \text{if } \kappa r^2 \geq (y - v)/2$$

$$\text{as long as } C \leq \max\{A/r\,,\, A((y - v)/2\kappa)^{-1/2}\}$$

$$= w^* + yC + A^2\kappa/C$$

$$\text{as long as } C > \max\{A/r\,,\, A((y - v)/2\kappa)^{-1/2}\}$$

Suppose now C is negative. By symmetry the max is the value at (y + v)/2 given by the first two formulas above as long as $C \geq \min\{-A/r\,,\, -A((y - v)/2\kappa)^{-1/2}\}$.Otherwise we solve $-C = A((a_o - v)/\kappa)^{-1/2}$ and obtain $w^* + vC - A^2\kappa/C$ as the max value.

Now consider maximizing $A\,U(a_o) - w^* - C\,a_o$. This is achieved simply by changing C to -C and w* to -w* in the analysis for maximizing $A\,U(a_o) + w^* + C\,a_o$. After summarizing over all possibilities the following is an expression for the maximizing magnitude.

$$|w^* + (y + v)C/2| + ((y-v)/2r + \kappa r)A \quad \text{if } \kappa r^2 < (y - v)/2$$

$$|w^* + (y + v)C/2| + (2\kappa(y - v))^{1/2}A \quad \text{if } \kappa r^2 \geq (y - v)/2$$

$$|C| \leq \max\{A/r\,,\, A((y - v)/2\kappa)^{-1/2}\}$$

$$\max\{w^* + yC + A^2\kappa/C\,,\, -w^* - vC + A^2\kappa/C\} \quad C > \max\{A/r\,,\, A((y - v)/2\kappa)^{-1/2}\}$$

$$\max\{w^* + vC - A^2\kappa/C,\; -w^* - yC - A^2\kappa/C\} \quad C < -\max\{A/r\,,\, A((y - v)/2\kappa)^{-1/2}\}$$

Since w* is only present in these terms of the upper bound the optimal w* can be found at this point in the derivation before minimizing over w. Indeed for fixed w all four of the above formulas are minimized at $w^* = -(y + v)C/2$. The last two formulas can now be simplified to: $(y - v)|C|/2 + A^2\kappa/|C|$. The bound (6) now follows easily and the solution in the exact case (5) follows by taking κ equal 0. **QED**.

To use quadratic programming first introduce the same variables and constraints for the objective function of Theorem III as for Theorem II. Then choose the smaller of two minima of the

objective function - one with the additional constraint $A^2 \leq \max\{ 1/r^2 , ((y - v)/2\kappa)^{-1} \} C^2$ and one with $A^2 \geq \max\{ 1/r^2 , ((y - v)/2\kappa)^{-1} \} C^2$.

As V is a ball about 0 and $\varepsilon(x) = \kappa||x||^2$ the bounds of the preceding two theorems are invariant to rotations.

B.Scale Invariant Versions which may Outperform Shrinkage and Regularization; differences from lasso regression

We now give a version of our minimax results where V is a rectangular region containing 0 and the predictors $\{ x_j \}$. If V is dilation by a fixed factor of the smallest rectangular region containing 0 and the predictors, then the estimators in Theorems IV and V below (in the exact linear case and in general for certain $\varepsilon(x)$) will be invariant to the scale of the design data and hence may outperform regularization and shrinkage methods (which weight components equally) in situations where the physical meaning of scale is not understood while still maintaining the same high efficiency (as we shall see in Sec. C) with respect to the standard and near neighbor methods. (They are however no longer rotation invariant. The standard Gauss method is both rotation and scale invariant.) We use a general $\varepsilon(x)$ in a rectangular region V which ,for the approximate linear case, converts the associated real algebraic geometry problems into the types one encounters in linear programming. Because the following bounds are shown to be valid for a large class of $\varepsilon(x)$'s they are not as sharp as the previous rotation invariant ones where $\varepsilon(x) = \kappa ||x||^2$.

Theorem IV Let the parametric family be $f (x ; a) = a_o + (a_1, a_2,.....a_d) \cdot x$. Let V = $[\mathbf{m}, \mathbf{m}+\mathbf{r}] = [m_1, m_1 + r_1] \times [m_2, m_2 + r_2] \times \cdots\cdots\cdots\cdots \times [m_d, m_d + r_d]$ be a rectangular region containing 0 and the predictors. Assume f(x) is within $\varepsilon(x)$ of some family member $f (x ; a)$. Use squared error loss. Let e be the maximum value of $\varepsilon(x)$ for x in V. $\mathcal{C}$ is the condition that f(x) satisfies $| f(x) - f(z) | \leq 2M$ in V.

The following is an upper bound on mean squared error of $F(\mathbf{w})$ in the general

($\varepsilon(x) > 0$) case for given **w** with $\sum w_j = 1$ and $w^* = 0$. It may be minimized over such w to provide an approximately minimax-optimal solution.

$$L_{\mathcal{C}}(\mathbf{w}) = w^t \Sigma w + B(w)$$

$$\text{with } B(w) = \left(2(M+e) \max_i \left\{ \left| \left(1/r_i\right)\left(\sum w_j x_j\right)_i \right| \right\} + \sum |w_j| \varepsilon(x_j) \right)^2$$

If f is linear in V ($\varepsilon(x) = 0$), the optimal strategy (3) for squared error loss takes the form

$$w = \arg\min_w \left[w^t \Sigma w + 4M^2 \left(\max_i \left\{ \left| \left(1/r_i\right)\left(\sum w_j x_j\right)_i \right| \right\} \right)^2 \right], \quad w^* = 0 .$$

where $w_k = 1 - \sum_{j<k} w_j$.

---

proof of Theorem IV: We write (in a different parametric form where $h_i(0)$ is nonzero) $f(x) = a_o + (a_1, a_2, \ldots a_d) \cdot (x - \mathbf{m}) + \zeta(x)$ with $|\zeta(x)| \le \varepsilon(x)$ in V. Note $f(0) = a_o - \sum a_i m_i$. Consider

$$E\left((F(\mathbf{w}) - (f(0))^2 \mid x_1, \ldots\right) = w^t \mathbf{N} w + \left\{ \sum a_i \left(\left(\sum w_j x_j\right)_i - C m_i \right) + w^* + C a_o + \sum w_j \zeta(x_j) \right\}^2 .$$

Here $C = C(w) = \sum w_j - 1$. Since there is no restriction on a_o the maximum over a's satisfying (2) with f satisfying $\mathcal{C}$ will be infinite unless C =0 which yields the condition on w_k. If f satisfies $\mathcal{C}$ and (2) then for some a

$$\left| (a_1, a_2, \ldots a_d) \cdot (x - y) \right| \le \left| a_o + (a_1, a_2, \ldots a_d) \cdot (x - \mathbf{m}) - f(x) \right| +$$

$$\left| f(y) - a_o - (a_1, a_2, \ldots a_d) \cdot (y - \mathbf{m}) \right| + \left| f(x) - f(y) \right| \le 2e + 2M$$

for all x and y in V. (Unlike the case in the proof of Theorem II not all such a's arise when we considers f's that satisfy $\mathcal{C}$ and (2). This set of a's is a bit bigger than necessary and hence the result may not be as "tight" as that in Theorem II.) The preceding set of inequalities in x and y has the real algebraic geometry solution - $\sum |a_i| r_i \le 2(M + e)$. Now the maximum error for given **w** is obtained by maximizing $\left|\sum a_i r_i \left(\left(\sum w_j x_j\right)_i / r_i \right) + w^* \right|$ and adding $\sum |w_j| \varepsilon(x_j)$

to it inside the above brackets. The solution is clearly to put all of the "weight" $|a_i|\, r_i = 2(M + e)$ on the component i maximized in the theorem statement and pick the sign of a_i appropriately according to the sign of w*. Clearly this maximized quantity is smallest for fixed w when w* = 0. The upper bound follows and the optimal bound in the exact case is proven by setting e = 0.**QED**

The estimator of this theorem (and the next), found by minimizing the upper bound on MSE, is invariant to a change in scale of the predictors if $\varepsilon(x) = h\,(\,(x)_1/\,r_1,\,(x)_2/\,r_2,\ldots\ldots(x)_d/\,r_d\,)$ and the region V is the smallest rectangular region containing 0 and the predictors or a contraction or dilation by a fixed factor thereof.

Under the latter condition with $\varepsilon(x) = 0$ consider generalizations of penalized regression

$\min ((\,\mathbf{X}a - \mathbf{Y})^t \sigma^{-1} (\,\mathbf{X}a - \mathbf{Y}\,) + \lambda\,(\,r_1|a_1|^q + \ldots\ldots\, r_d\,|a_d|^q\,)$ or constrained linear regression

$\min\ (\,\mathbf{X}a - \mathbf{Y})^t \sigma^{-1} (\,\mathbf{X}a - \mathbf{Y}\,)$ such that $r_1|a_1|^q + \ldots\ldots\, r_d\,|a_d|^q \le s$. For q=1: s may be interpreted as the maximum oscillation of the linear target function in V; both are scale invariant to the predictors, are nonlinear in the y's and are commonly referred to as lasso regression. Hence our linear method is fundamentally different from lasso regression. Since we have shown that ridge regression by gradient regularization is locally minimax (expressing λ in terms of M),we believe that the minimax method (exact case) of Theorem IV is the "correct'' scale invariant analog (not the lasso) of ridge regression by gradient regularization. In section II-C a simple one dimensional example was given where the minimax method outperformed constrained lasso.

For QP introduce the w_j^+ and w_j^- and their constraints in the objective function of Theorem IV exactly as for Theorem II. Then replace the max term by variable A and add the 2d constraints $A \ge \pm\,(\,1/r_i\,)(\,\Sigma\ w_j x_j\,)_i$.

Theorem V Assume the conditions of Theorem IV except that $\mathcal{C}$ is the condition that f (x) takes values in [v,y] for all x in V . Let C = C(w) = $\Sigma\ w_j$ - 1. The following is an upper bound on mean squared error of F(**w**) in the general ($\varepsilon(x) > 0$) case for given **w** with w* = -(v+y)C/2.

It may be minimized in w to give approximate minimax optimality.

$$L_{\mathcal{C}}(\mathbf{w}) = w^t \sigma\, w + B(w) \quad \text{with}$$

$$B(w) = \left(\max\left\{|((y-v+2e)/2)C - (y-v+2e)A_{min}|,\ |((y-v+2e)/2)C - (y-v+2e)A_{max}|\right\} + \Sigma\, |w_j|\, \varepsilon(x_j)\right)^2$$

where $A_{min} = \min_i \{ (1/r_i)((\Sigma\, w_j x_j)_i - Cm_i) \}$ or 0 whichever is smaller ,

$$A_{max} = \max_i \{ (1/r_i)((\Sigma\, w_j x_j)_i - Cm_i) \}$$ or 0 whichever is greater.

If f is linear in V , the optimal strategy (3) for squared error loss takes the form $w^* = -(v+y)C/2$,

$$w = \underset{w}{\operatorname{argmin}} \left[w^t \sigma\, w + \left(\max\left\{ |((y-v)/2)C - (y-v)A_{min}|\ ,\ |((y-v)/2)C - (y-v)A_{max}| \right\}\right)^2 \right]$$

---

proof of Theorem V: Again, as in the last proof, we write (in a different parametric form where $h_i(0)$ is nonzero) $f(x) = a_o + (a_1, a_2, ...a_d) \cdot (x - \mathbf{m}) + \zeta(x)$ with $|\zeta(x)| \le \varepsilon(x)$ in V. Note now that $f(0) = a_o - \Sigma a_i m_i$. If f satisfies $\mathcal{C}$ and (2) with some a then that a satisfies

$$v - e \le a_o + (a_1, a_2, ...a_d) \cdot (x - \mathbf{m}) \le y + e \quad \text{for all x in B.}$$

(Unlike the case in the proof of Theorem III not all such a's arise when considering f that satisfy $\mathcal{C}$ and (2). This set of a's is a bit bigger than necessary and hence the result may not be as "tight" as that in Theorem III but it holds for a more general class of $\varepsilon(x)$'s.) Our algebraic geometric argument is now to write the preceding inequalities in x in the equivalent form

$$\Sigma\, r_i \max\{a_i, 0\} \le y + e - a_o \quad \text{and} \quad \Sigma\, r_i \min\{a_i, 0\} \ge v - a_o - e\ .$$

Now by some simple algebra we may rewrite (4) in terms of the parametrization as

$$E((F(\mathbf{w})-(f(0))^2 \mid x_1,..) = w^t \mathbf{N}\, w + \left\{\Sigma\, a_i ((\Sigma\, w_j x_j)_i - Cm_i) + Ca_o + w^* + \Sigma w_j \zeta(x_j)\right\}^2.$$

The bracket term will be bounded by maximizing $|\Sigma\, a_i ((\Sigma\, w_j x_j)_i - Cm_i) + Ca_o + w^*|$ and adding $\Sigma\, |w_j|\, \varepsilon(x_j)$ to it inside the brackets. Now write the former as $|\Sigma\, a_i r_i A_i + Ca_o + w^*|$

where $A_i = (1/r_i) ((\Sigma w_j x_j)_i - Cm_i)$. (The idea of the construction is to attach all of the"weight" $a_i r_i$, which is either $y + e - a_o$ or $v - e - a_o$, to at most two i's.) For fixed a_o (lying in [v - e , y + e]) the maximum is achieved by one of the following choices for $(a_1,a_2, ...,a_d)$ where at most two components are nonzero:

1. $a_i = (y + e - a_o)/ r_i$ for one i when $A_{max} > 0$ and where $i = \arg\max A_i$

   $a_i = (v - e - a_o)/ r_i$ for one i when $A_{min} < 0$ and where $i = \arg\min A_i$

   Otherwise $a_i = 0$. (If A_{max} or A_{min} is 0 then one or no a_i's are nonzero.)

   Then the value is $| Ca_o + w^* + (y + e - a_o) A_{max} + (v - e - a_o) A_{min} |$.

2. $a_i = (v - e - a_o)/ r_i$ for one i when $A_{max} > 0$ and where $i = \arg\max A_i$

   $a_i = (y + e - a_o)/ r_i$ for one i when $A_{min} < 0$ and where $i = \arg\min A_i$

   Otherwise $a_i = 0$. (If A_{max} or A_{min} is 0 then one or no a_i's are nonzero.

   Then the value is $| Ca_o + w^* + (v - e - a_o) A_{max} + (y + e - a_o) A_{min} |$.

Now we maximize each of the two values over a_o and take the max of the max's. Since what is inside | | is linear in a_o we take the max over the following 4 quantities obtained by taking $a_o = y + e$ in 1. and 2. and then $a_o = v - e$ in 1. and 2.:

$| w^* + C(y + e) - (y - v + 2e) A_{min} |$, $| w^* + C(y + e) - (y - v + 2e) A_{max} |$,

$| w^* + C(v - e) + (y - v + 2e) A_{max} |$, $| w^* + C(v - e) + (y - v + 2e) A_{min} |$

Since w* will only appear in the max of the preceding 4 terms we can determine it at this point: Indeed for any fixed w the max of the first and fourth such terms is minimized when w* is half way between the zeros of these two terms, i.e. at $w^* = -(v+y)C/2$, as is also the max of the

second and third terms.Using this w* we can reduce the max of the four terms to the max of

$$|((y-v+2e)/2)C - (y-v+2e)A_{min}| \quad , \quad |((y-v+2e)/2)C - (y-v+2e)A_{max}|$$

Adding $\Sigma |w_j| \varepsilon(x_j)$ yields the expression for $L_{\mathcal{C}}(\mathbf{w})$. For the exact case we let e = 0. **QED**.

For QP write the objective as $(w^+ - w^-)^t \sigma (w^+ - w^-) + (A + \Sigma (w_j^+ + w_j^-)\varepsilon(x_j))^2$

and minimize subject to $w_j^+ \geq 0$, $w_j^- \geq 0$, $A \geq - ((y-v+2e)/2)C$ and $A \geq ((y-v+2e)/2)C$,

$A \geq - (y-v+2e) \{C/2 - (1/r_i)((\Sigma (w_j^+ - w_j^-) x_j)_i - Cm_i) \}$ and

$A \geq (y-v+2e) \{C/2 - (1/r_i)((\Sigma (w_j^+ - w_j^-) x_j)_i - Cm_i) \}$ for i= 1,2,..d. Here C = C (w)

$= \Sigma (w_j^+ - w_j^-) - 1$.

C.The Predominance of Examples with High Relative Accuracy of the Contextual Linear Estimators Compared to Standard Least Squares or (in cases where scale is poorly understood) Regularization Methods

We will now show that in high dimensions there are many locally linear situations for which the minimax estimators of Theorems II, IV (exact case) are much more efficient than both standard least squares applied to a set of k "close" points (which is also the local Kriging estimator with the exact local linear model) and the near neighbor average estimator (which is the case of $w_j = 1/k$). Here relative efficiency is given by the ratio of mean squared errors without the unbiasedness restriction.(The same advantages can be shown for Theorems II, III, IV and V in approximately locally linear cases if $\varepsilon(x)$ is sufficiently small.) We start with a simple artificial example for the close predictors whose construction leads to a large class of examples:

Assume i.i.d. gaussian noise of unit variance and an exact linear model in R^d with oscillation bounded by 2 for x in the unit ball V.The predictor data falls into d groups, the first d-1 of which are described as follows - group i contains two vectors whose components are 0 except for the i'th for which the values are 1and $1 - d^{-1/2}$ respectively. The d'th group consists of 2s vectors whose first d-1 components are 0 and half of whose d'th components are 1 and the

other half of whose d'th components are $1 - d^{-1/2}$. Note that for large d and large s the least squares predictor at 0 for the simple regression problem with the data of group i (i<d)has a (in all cases) variance ~ 2d while the corresponding estimator S for the d'th group has a (in all cases) variance ~2d/s .(See [27] p. 11, equ. 3.8). So one might say that subproblems i (prediction from data in group i) are "extremely hard" for i<d and subproblem d is only "moderately hard" if s~d/2.

Now let us consider the mean squared errors of the 3 competing estimators. For the near neighbor average we plug $w'_j = 1/(2d+2s-2)$ and our predictor data into the expression $w'^t \sigma w' + (MA/r)^2$ of Theorem II (with r = 1, M = 2) to get a mean squared error ~ $(18d -18 +16s^2)/(2d+2s-2)^2$, which is ~4/9 for s ~ d/2 . (From the proof of Theorem II this bound is achieved.)

For the standard least squares formula , using the facts that the weights satisfy the conclusions of Theorem I with $\varepsilon(x) = 0$ (recall the discussion in section II-A-3) and that the support subspaces of the d subproblems are orthogonal, one sees first that the weights associated with the design points in each subproblem satisfy the constraints (1) for that subproblem except possibly the normalization constraint. Now if the design matrices of the subproblems all have full rank (which holds for our example and in general with probability one when 2d+2s-2 such points are chosen at random wrt. an absolutely continuous measure) we may alter the weights by arbitrarily little so that the sum of the weights in each group is nonzero, the total sum remains unity and the other null constraints continue to hold for each group. We may now write the overall variance of the slightly altered estimator as

$$\sum_j \sum_{i \text{ in group } j} w_i^2 = \sum_j \alpha_j^2 \sum_{i \text{ in group } j} w'^2_i \quad \text{where} \quad \sum_j \alpha_j = 1$$

and the w'_i in group j satisfy all the constraints (1) for that group. Now we may alter the w'_i to make the inner sums as small as possible subject to the constraints. Now we minimize over the α_j and the result is still arbitrarily close to the original minimum overall variance. Applying this procedure to our particular subproblems we see that the overall variance is ~ $2d \sum_1^{d-1} \alpha_j^2$ +

$2d(1 - \sum_{1}^{d-1} \alpha_j)^2/s$ which is greater than or equal (by the Cauchy-Schwartz inequality)

$(2d / (d-1))(\sum_{1}^{d-1} \alpha_j)^2 + 2d(1 - \sum_{1}^{d-1} \alpha_j)^2/s$ which is always greater than one if $s < d$.

Finally we give an upper bound on the mean squared error of the estimator of Theorem II by using weights $w'_j = 1/(2d -1)$ for the 2d-2 vectors in the first d-1 groups and weights given by $1/(2d -1)$ times the weight of the vector in the standard linear predictor S of $f(0)$ for the d'th group. Denoting this estimator by E we may write

$$E - f(0) = ((2d-2) /(2d-1)) (W - f(0)) + (1/(2d-1)) (S - f(0))$$

where W is is the estimator corresponding to weights $1/(2d -2)$ for the 2d-2 vectors in the first d-1 groups. By plugging into the expression $w^t \sigma w + (MA/r)^2$ of Theorem II (with r = 1, M = 2) we see that the MSE bound of W is ~ $9/(2d - 2)$. From the unbiasedness of S and its independence from W we get a bound on mean squared error for E ~ $9/(2d - 2) + 2d/s(2d - 1)^2$. So for appropriate choices of s the minimax estimator is O(d) more efficient than either local least squares or the near neighbor average. We note that if, in Theorem IV with V the unit cube, we use the expression $w^t \sigma w + 4M^2 (\max \{ | (1/r_i)(\Sigma w_j x_j)_i | \})^2$ to bound the mean squared error of W we would get ~ $1/ (2d - 2) + 16/ (d - 1)^2$ so that the bound on MSE of E would be ~ $1/ (2d - 2) + 16/ (d - 1)^2 + 2d / s (2d - 1)^2$. Hence the same advantages would apply to the scale invariant local minimax estimator of Theorem IV.

Now the existence of many more such examples is clear : if the predictor data can be partitioned into(nearly)orthogonal groups with most groups providing a very difficult prediction problem and a few furnishing a moderately difficult challenge, then we may "weight" the subproblems appropriately to show that the rotation invariant minimax estimator may perform much better than the other two popular methods.

Shrinkage and regularization(which includes principal components, partial least squares and

continuum regression; see [41], [42].) are used to reduce dimensionality and hence increase accuracy when the relative magnitudes of the data coordinates are understood in physical or other scientific terms. But when the relative scales of the data components are collectively poorly understood as in many learning situations shrinkage and regularization become totally ad hoc. But our scale invariant versions will still have the same predominant improvement in accuracy over standard least squares when we combine regression problems as above but this time with disjoint supports instead of orthogonal supports(distinct components for each problem).

Finally note that if the moderately difficult subproblems described above have significant nonlinearities while the predominant difficult subproblems had only small nonlinearities then the local minimax estimators would again have similar significant advantages so that one expects significant robustness in the minimax solutions to high dimensional problems with context.

V. Using Boosting and Greedy Additive Expansions to estimate $\varepsilon(x)$ and obtain local minimax estimators

We divide (jacknife) regression data into two groups - to the first we apply a machine algorithm which gives us a global estimate of the form (#) of the function to be learned (or

$$(\#) \qquad f(x) \sim \sum_{1}^{N} c_n \, g_n(a_n^t x) \qquad (\#)$$

perhaps an estimate of the form (#) with different weights c_n in different regions so as to emphasize approximation accuracy in a weak neighborhood of 0 as in [22]); now from the data in the second group we obtain a local estimate and accuracy bound which uses information learned from the first group by the machine. We derive a distance measure and modulus of accuracy which forces near linearity of the target function at close points and for which there may exist sufficiently many such close points.This should hold if the expansion is parsimonious, most of the $g_n(a_n^t x)$ are nearly linear, or (#) has much near redundancy in the projection directions. This distance measure appears to be robust in that it changes little when the model estimate varies. This can be made more precise in the following development:

We will only treat the case where the g's in the expansion are functions of a one dimensional projection. The more general case could be similarily carried out as in the case of tree boosts but because partitioning reduces effective sample size we believe the fusion methods of the next section are more appropriate for tree learners, especially for Breiman's random forests.Suppose f can be expressed (locally) exactly in the form (#) where the a_n are unit vectors. Consider k_n , called the n'th coefficient of nonlinearity, as the smallest constant such that the following local univariate approximation bound holds for $c_n\, g_n(u)$ -

$$| c_n\, g_n(u) - bu - s | \leq k_n u^2 \text{ for some b,s and } -u_o \leq u \leq u_o .$$

Such a k_n always exists if $g_n''(u)$ is continuous which we assume. If the predictor data has been sphered so that the sample covariance is the identity matrix, u_o might be chosen as 2.0 ,for instance, so that the above inequality would hold for a high percentage of the projected data $u_j = a_n^t x_j$. The smaller k_n the more linear $c_n g_n(u)$ is and the larger the components of "close" points may be in the a_n direction.

Lemma I : Let $D(0, x) = \{ \sum k_n (a_n^t x)^2 \}^{1/2}$. Then f satisfies (2) for the affine family in $\mathcal{V} = \{ x : |a_n^t x| \leq u_o \text{ for n = 1, 2, ...N.} \}$ where the modulus of accuracy $\varepsilon(x) = D^2(0, x)$.

---

proof of Lemma I : Determine b_n, s_n such that $| c_n g_n(u) - b_n u - s_n | \leq k_n u^2$ for $u_o \leq u \leq u_o$. Then $| f(x) - (s_1 + s_2 + s_3 + ... + b_1 a_1^t x + b_2 a_2^t x + b_3 a_3^t x + ...) | \leq$

$| c_1 g_1(a_1^t x) - b_1 a_1^t x - s_1 | + | c_2 g_2(a_2^t x) - b_2 a_2^t x - s_2 | + ... \leq$

$k_1 (a_2^t x)^2 + k_2 (a_2^t x)^2 + ... = D^2(0, x)$. **QED**.

Hence for regression predictor data in $\mathcal{V}$ the local estimators and accuracy bounds from Theorem I (with the affine family and V = $\mathcal{V}$) or Theorems IV-V (with $\mathcal{V} \subseteq$ V) may be used with $\varepsilon(x) = D^2(0, x)$ if we can estimate D(0, x) and $\mathcal{V}$. We propose the estimators

$D^*(0, x) = \{ \sum k_n^*(a^{*t}_n x)^2 \}^{1/2}$, $\mathcal{V}^* = \{ x : |a^{*t}_n x| \leq u_o \text{ for } n = 1, 2, ...N. \}$,

where k_n^* and a^*_n come from the machine learned global (weak neighborhood) estimate of f based on the first data group. One would expect D* to be quite close to D but this relationship requires further study. One only needs D* to be within a moderate factor of D.

We have shown that greedy additive expansions produce an $\varepsilon(x)$ for local estimation. If there are only a few ridge directions in the expansion for which the corresponding ridge function is significantly nonlinear then $\varepsilon(x)$ is small for the (many) predictors that are nearly orthogonal to these directions and hence local estimation will not suffer (as much as say for radially nonlinear functions) from the curse of dimensionality- i.e. we can expect a significant number of predictor data with small $\varepsilon(x)$. When $\varepsilon(x)$ is small for the predictors we get nearly the same efficiency(in the squared error sense) as with linear estimation as the theorems II-V indicate.

VI. Fusion of Local Estimators ; Improved Estimation for Classification and Regression Forests

A.Combining the local estimators of a class of (possibly corrupted) experts, Overcoming the Curve of Dimensionality

In this chapter (and only here) we assume exclusively the random predictor-response regression model, with $x_j = X_j$ and (X_j , Y_j) i.i.d., and want to learn $f(0) = E(Y \mid X=0)$ by combining various conditional expectation estimates of Y from the data. Because many or all of these estimates are conditioned only upon information about projections of X we may only be able to learn a weighted sum of $E(Y| X \varepsilon \mathcal{a}_i)$ for some maximally “informative” collection $\{ \mathcal{a}_i \}$. We try to make this clearer with the description and random forest example that follow.

Suppose we have m experts, each with a model chosen independenty of the training data, who attempt to (approximately) solve (3) as follows: Expert i chooses a neighborhood $\mathcal{u}_i$ in R^d of the query 0 and considers only the training predictors x_j belonging to $\mathcal{v}_i$.$\mathcal{v}_i$ has the form of a cylinder in R^d generated by a neighborhood U_i of 0 in a d_i dimensional subspace A_i of R^d , i.e. $\mathcal{u}_i$ consists of the set of all points whose orthogonal projection onto A_i lies in U_i. The expert then considers local minimax estimation of $f^i(0)$, where $f^i(x) := E$ (Y | the projection of X onto A_i is x),

using the predictor data whose projectioins onto A_i lie in U_i . He then applies one of the optimal bounds in Theorems I - V using the affine family of approximands with his own $\varepsilon^i(x)$. Let $\mathcal{N}_i$ be the set of indices of the predictors in $\mathcal{U}_i$. Denote by $\{x^i_j\}$ the projections of predictors in $\mathcal{U}_i$ onto A_i (which lie in U_i). We write $Y_j = f^i(0) + a^i \cdot x^i_j + \zeta^i_j + N^i_j$ where j varies over the indices in $\mathcal{N}_i$. The error in the affine representation of f^i at predictor x^i_j , ζ^i_j, is assumed to be bounded in absolute value by $\varepsilon^i(x^i_j)$ (which would be $\kappa^i||x^i_j||^2$ if he is using Theorem II or III); ε^i could be chosen either by appropriate modeling or large enough to include low dimensional nonlinear submodels since the results would then be robust to the size of the nonlinearity as in the road roughness examples.

Now let's assume that the target function $f^i(x)$ satisfies appropriate conditions in a set V_i (containing $\{x^i_j\}$) of one of the theorems I-V and expert i uses an estimator

$F_i = w^{*i} + \sum_j w^i_j Y_j$ ($w^i_j = 0$ if j is not in $\mathcal{N}_i$), where the w^{*i}, w^i_j satisfy the appropriate constraints guaranteeing the error bound $L^i_{\mathcal{C}}(\mathbf{w}^i) = w^{i\,t}\boldsymbol{\sigma}^i w^i + B_i(w^i)$ of Theorem I (where $B_i(w^i) = R(w^i)$), orTheorems II- V. Each $\boldsymbol{\sigma}^i$ is diagonal and is an upperbound in the semidedinite order for the diagonal covariance $\mathbf{N}^i$. Denote by σ^i_j the bound on the standard deviation of N^i_j. Hence we are fixing the constraints of expert i according to the corresponding theorem but we do not further require that he minimize his own error bound leaving open the possibility of jointly optimizing a bound obtained by combining (fusing) the experts.

Before comparing and combining the experts' accuracies we need to compare their degree of conditioning with respect to the query. An expert who takes $A_i = R^d$ is estimating f(0) while another who chooses $A_i = R^2$ is (possibly more accurately) estimating Y conditioned on a d-2 dimensional event. We assume an information measure $I(\mathcal{A})$ is defined on subsets $\mathcal{A}$ containing the query. The appropriate $\mathcal{A}$ for expert i is $\mathcal{A}_i = \{ x:$ the projection of x onto A_i is 0 $\}$. Indeed he

is estimating $f^i(0) = E(Y \mid X \varepsilon \mathcal{A}_i)$. We use $I(\mathcal{A}_i) = d_i + 1 = \text{codim}(\mathcal{A}_i) + 1 = \dim(A_i) + 1$ but others may also be justified. By convention, for $d_i = 0$ ($A_i = \{0\}$), the expert estimates E(Y). So in this case there is one unit of conditioning information.

Since we are combining experts an extension of the concept of the conditioned event $\mathcal{A}$ is a probability measure $\mathcal{P}$ on the class of such events. Now define conditional expectation $\mathcal{E}(Y \mid \mathcal{P}) = \int_{\mathcal{S}} E(Y \mid X \varepsilon \mathcal{A})\, d\mathcal{P}(\mathcal{A})$ and information $\mathcal{I}(\mathcal{P}) = \int_{\mathcal{S}} I(\mathcal{A})\, d\mathcal{P}(\mathcal{A})$ via expectation with respect to that measure.The conditioned event $\mathcal{P}$ is synonomous with the measure itself and is a random choice of $\mathcal{A}$ with respect to the probability measure. Finding exact mathematical requirements for the validity of this proposal is an interesting question; we justify the method in our setting as follows: We put a finite discrete probability measure α on the subsets having probability α_i for event $\mathcal{A}_i$ (with $\alpha_i \geq 0$ and $\sum \alpha_i = 1$). Then $\mathcal{E}(Y \mid \alpha) = \sum \alpha_i E(Y \mid X \varepsilon \mathcal{A}_i)$. Now the degree of conditioning of (a "master" expert who estimates) this quantity is the extension of $I(\mathcal{A})$ to the space of probability measures specified given by $\mathcal{I}(\alpha) = \sum \alpha_i I(\mathcal{A}_i)$. The goal is to estimate a conditional expectation with high information and with high accuracy.

We consider the accuracy question first. Let $F = \sum \alpha_i F_i$ for given probabilities α_i. Write the mean squared error $E\{ (F - \mathcal{E}(Y \mid \alpha))^2 \}$ as follows-

$$E\left\{ \left(\sum_i \alpha_i \left(\sum_j w^i_j a^i \cdot x^i_j + \left(\sum_j w^i_j - 1 \right) f^i(0) + w^{*i} + \sum_j w^i_j \zeta^i_j + \sum_j w^i_j N^i_j \right) \right)^2 \right\}.$$

Now the sum of the first 4 summands (out of 5 total) in the coefficient of α_i above is bounded in absolute value by $B_i (w^i)^{1/2}$. This follows by examining the bracket term in (4) for the regression case in the proof of the appropriate Theorem I, II, III, IV or V. Taking the expectation above, using the mean 0 property of each N^i_j, one obtains a bound on this expectation given by

$$E\left\{ \left(\sum_i \sum_j \sum_r \sum_s \alpha_i \alpha_r w^i_j w^r_s N^i_j N^r_s \right) \right\} + \left(\sum \alpha_i B_i(w^i)^{1/2} \right)^2.$$

Apply the independence and only the terms with j =s remain. It is now routine to see that the expectation is bounded by

$E\{(\sum_i \sum_j \sum_r \alpha_i \alpha_r |w^i_j| |w^r_j| |N^i_j| |N^r_j|)\} + (\sum \alpha_i B_i(w^i)^{1/2})^2$ which by the Cauchy inequality is bounded by $(\sum_i \sum_j \sum_r \alpha_i \alpha_r |w^i_j| |w^r_j| \sigma^i_j \sigma^r_j) + (\sum \alpha_i B_i(w^i)^{1/2})^2$.

Hence the mean squared error for the estimate of $\mathcal{E}(Y \mid \alpha)$ is bounded by

$$G(w, \alpha) = \sum_j (\sum_i \alpha_i |w^i_j| \sigma^i_j)^2 + (\sum \alpha_i B_i(w^i)^{1/2})^2.$$

If expert i uses Theorem I the above result holds also when he uses a model involving nonlinear functions of x^i_j - i.e. $Y_j = f^i(0) + a^i \cdot h(x^i_j) + \zeta^i_j + N^i_j$; in fact it clearly holds for any such (linear or nonlinear) case where each expert i limits himself to affine estimators F_i for which the worst bound on squared bias has a known form $B_i(w^i)$, e.g. as will be with Theorem VI where $Y_j = g(x^i_j) + \zeta^i_j + N^i_j$ with $f^i(0) = g(0)$ for g in a ball of a reproducing kernel Hilbert space.

So one solution to the fusion problem would be to minimize the sum of $G(w, \alpha)$ and a penalty term which is a suitable convex increasing function of the quantity $1/\mathcal{I}(\alpha)$.

(B) $\min_{\alpha: \sum \alpha_i = 1,\ 0 \le \alpha_i} G(w, \alpha) + h(1/\mathcal{I}(\alpha))$

w^i constrained by appropriate theorem, $w^i_j = 0$ if j does not lie in $\mathcal{N}_i$

The objective function in **(B)** is not jointly convex in w, α but is convex in each with the other fixed so alternating convex minimization methods could be applied. For the two class probability of class 2 problem we propose a form for h which is a function of w and α (so **(B)** is no longer biconvex): Let F be truncated to always take values in [0, 1]. The bounds on mean square error clearly remain valid. Let λ be positive and β be a small positive number in (0, 1). Then the penalty h is given by

$$h = h\left(1/\mathcal{I}(\alpha), w\right) = \lambda\left(1/\mathcal{I}(\alpha) - 1/(d+1)\right) \quad 1/F^{\beta}(1-F)^{\beta}.$$

When F is very close to 0 or 1 the penalty becomes very small and $G(w, \alpha)$ dominates the minimization. This is justified by noticing that, if F is sufficiently close to 0 or 1 and the error bound $G(w, \alpha)$ is sufficiently close to 0, then F is close to f(0) with a probability nearly one. On the other hand, if F sufficiently far from 0 or 1 and β is sufficiently small, then the penalty term is essentially $\lambda\left(1/\mathcal{I}(\alpha) - 1/(d+1)\right)$ where λ represents the information-accuracy trade-off coefficient.

It is assumed in the analysis that each expert has a correct model. We now derive a solution to the fusion problem when each expert's model may be wrong with a small probability π_o (independent of other experts and the predictor data it is applied to). Assume each expert uses one of the models of Theorems I - V. Then a corrupted expert can be modeled simply by changing his $\varepsilon^i(x)$ to an appropriate default $\varepsilon^{i*}(x)$. For the two class problem this default $\varepsilon^{i*}(x)$ will be set equal to 1 except at x=0 where it is 0. In the following, if expert i is incorrect then use the above default $\varepsilon^{i*}(x)$ and write the bias term in his bound as $B_i^*(w^i)$. (These * bias terms are are explicit for Theorems I, IV and V but may be easily derived in the other two cases. Recall he chooses a model independent of the data x^i_j.) Then $G(w, \alpha)$ becomes

$$\sum_j \left(\sum_i \alpha_i |w^i_j| \sigma^i_j\right)^2 + \left(\sum \alpha_i B_i(w'^i)^{1/2} + \sum \alpha_i J_i \left(B_i^*(w'^i)^{1/2} - B_i(w'^i)^{1/2}\right)\right)^2$$

where the random variables J_i are i.i.d. Bernoulli (π_o).Taking the expectation we get the bound $G^*(w, \alpha)$ subject to corruption wth probability π_o:

$$G^*(w, \alpha) = \sum_j \left(\sum_i \alpha_i |w^i_j| \sigma^i_j\right)^2 + \pi_o(1-\pi_o) \sum \alpha_i^2 \left(B_i^*(w'^i)^{1/2} - B_i(w'^i)^{1/2}\right)^2$$
$$+ \left(\sum \alpha_i B_i(w'^i)^{1/2} + \pi_o \sum \alpha_i \left(B_i^*(w'^i)^{1/2} - B_i(w'^i)^{1/2}\right)\right)^2 .$$

Then $G^*(w, \alpha)$ is used in **(B)** to obtain the fusion bounds and estimators under

corruption, which may be computed as functions of π_o to provide an operating characteristic. Furthemore one may deduce easily from theorems I, IV,V that, if each $\varepsilon^i(x)$ is bounded above by the default $\varepsilon^{i*}(x)$, $B_i(w^i) \leq B_i^*(w^i)$ and $G^*(w, \alpha)$ is increasing in π_o. (For the two class problem this is the case provided that $\varepsilon^i(x) \leq 1$.) So a solution for π_o will produce a bound for all smaller π_o as well. Hence in practice we need only bound π_o above.

The fusion solution under corruption provides a method of potentially overcoming the curse of dimensionality: Imagine that each expert presents an analysis of a coresponding feature for a very large set of features. Suppose there are a moderately large number of truly predictive features and one feature that appears more predictive than the others in the training data but that does not generalize. Such spurious features will occur for a fixed sample size with higher frequency as the dimensionality(number of features) increases. An exhaustive analysis of all features with respect to the training data may yield the spurious feature (this feature might also dominate in **(B)**). But our corrupted version will prevent any feature from having too much influence. This is also a main idea in Breiman's random forests [6] where the weighting of the experts is uniform. Our theme is to find an optimal weighting while protecting against overweighting any one expert.

B. Random Forests for Microarray Classification

Tree learners [5,6] can be viewed as piecewise linear function estimators where a linear piece may be viewed as an expert: First a linear feature is chosen and the training data is divided optimally into two groups by binary thresholding the values under the feature mapping of the predictor vectors in the training set by optimizing some splitting criterion (such as the Gini criterion below). Each of the two groups is then split by choosing a new splitting feature and then optimally dividing, etc. A group is not split further when the responses of the members are sufficiently close (in some distance) to a mean, median or other linear fit of the responses of the whole group, where the fit uses the splitting features

which define the group. For instance, in our probability of class membership example, where the responses are 0 or 1 we might stop splitting only when the group (also called a node of the tree) has responses all 0's or all 1's. This occurs when the Gini criterion is used: For each node S_0, the Gini index $G(S_0) = 1- p_o^2 - p_1^2$ where p_i is the relative frequency of response i in $G(S_0)$. For a given threshold, yielding groups S_1 and S_2, the Gini criterion is $(n_1 G(S_1) + n_2 G(S_2))$ / n_o with $n_i = \#S_i$. This is then minimized over possible thresholds. Once the Gini criterion is not less than $G(S_0)$ for any threshold it can be proved that the node must have responses all 1 or all 0.

Once the tree is constructed (no more splitting possible) a test vector is run down the tree using the various features and thresholds applied to the test vector until it lands in a terminal node. The value for the test vector is the linear prediction at the test point furnished by the linear fit for the node. In our example using the Gini criterion below it is the common value for the group since the linear fit of the responses we actually used had weights summing to one.

Because trees partition the sample data into many nodes, each consisting of a much smaller subsample, they often generalize poorly since a locally linear prediction is based on the smaller subsample. Random forests [6] generate many trees(a forest) each constructed by choosing optimal splitting features at each node from a random subset of features(of a predeiermined size and structure for all of the trees in the forest). A test vector is run down the trees in the forest and the average of the terminal predictions(in our example the common nodal values) is used as the estimate. In the classification case one chooses class by taking the majority vote of the trees. Although the optimal classification is often achieved the average vote may inaccurately estimate the probability of correct classification, a quantity that is of primary interest in medical diagnosis and treatment of disease. Also in the general regression case Breiman has mentioned that more accurate estimators than a simple average need to be developed. (We believe that we have solved this problem with our fusion estimators (**B**) and versions for corrupted features.)

For a given query vector we may view the prediction furnished by the terminal node of tree i as that of expert i. Let d_i is the depth of the terminal node of tree i . $\mathcal{N}_i$ is the set of predictors

in its terminal node and A_i is the d_i dimensional subspace spanned by the features defining the node and U_i is the neighborhood in A_i characterized by the thresholds defining the node.

We now will apply the above **(B)** to our probability estimation setting. Since the responses and predictors have already been used in the trees' construction the results would, strictly speaking, hold only for new training predictors and responses which were independently generated. So if we divide the predictor data into two (equal) subsets,forming the trees with one and running the others down the trees, then form the various $\mathcal{N}_i$ from the latter, we may fuse these experts. With many dimensions and a large forest, spurious features may yield an occasional tree that incorrectly models in the terminal node for the query but that has a very low error bound. So we modify **(B)** for corruption.

In our preliminary experiment we use the full sample (with the test point left out), basing the prediction on the same x's used to grow the forest, in order to demonstrate the technique with a very small data set. We apply the techniques to microarray data from the University of Pittsburgh simulator. Sixteen patients, the first 8 in Group A (1) and the second 8 in Group B (0), provide arrays (120 dimensional feature vectors) each providing fluorescence measurements of the same 120 genes. Twenty of the genes were differentially expressed (had mean difference between groups, Table 1 gives the fluorescences for these 20 to 2 significant figures). The remaining were randomly generated with the same distribution for each group. For each of the 16 patients 12,500 trees were generated using the remaining 15 patients. At each node a random subset of features consisting of 8 components (genes) was used and the component yielding the smallest possible Gini criterion was chosen as splitting feature. The vote as a fraction of class 1 votes for the whole forest is given in column 1 of Table 2. (Software for random forests was developed by Len Russo.)

Table 1			**G roup A**					**Group B**							
3.9	-1.9	5.6	9.9	9.9	7.5	7.1	7.6	11.	11.	11.	12.	11.	11.	12.	12.
28.	24.	27.	31.	31.	20.	23.	25.	15.	18.	15.	15.	18.	17	17	17
15.	16.	12.	24.	18.	15.	26.	13.	12.	10.	9.6	9.9	9.6	10	10	10
34.	28.	28.	32.	34.	37.	25.	24.	22.	22.	22.	20.	21.	22	21	22
32.	29.	31.	32.	32.	25.	36.	25.	24.	25.	22.	24.	23.	24	24	24
26.	34.	21.	21.	29.	22.	20.	20.	18.	18.	20.	19.	19.	18	18	19
14.	5.0	11.	16.	4.8	11.	12.	4.7	15.	17.	16.	16.	17.	16	17	17
26.	27.	28.	19.	16.	26.	32.	19.	15.	16.	15.	15.	17.	15	16	15
25.	28.	19.	29.	27.	27.	24.	28.	28.	28.	29.	30.	30.	30	29	30
32.	31.	28.	26.	36.	26.	39.	36.	24.	25.	22.	23.	24.	24	24	25
16.	16.	25.	19.	29.	27.	20.	21.	14.	14.	14.	13.	14.	12	15	12
29.	19.	17.	18.	21.	20.	16.	30.	31.	31.	31.	32.	31.	32	31	29
26.	17.	17.	13.	27.	21.	27.	29.	15.	16.	15.	14.	17.	14	14	13
11.	15.	11.	19.	8.6	10.	12.	19.	22.	22.	20.	23.	21.	21	22	21
13.	8.5	9.1	18.	9.0	8.6	17.	13.	20.	20.	18.	23.	20.	20	20	20
28.	27.	33.	33.	29.	26.	23.	23.	20.	19.	19.	21.	20.	24	20	23
19.	17.	16.	24.	10.	18.	20.	24.	26.	25.	24.	24.	25.	26	26	26
17.	5.	9.	21.	5.	14.	14.	11.	20.	20.	19.	22.	20.	20	22	19
28.	23.	35.	31.	26.	29.	26.	31.	22.	22.	21.	23.	21.	21	23	22
22.	17.	17.	19.	28.	22.	20.	25.	16.	16.	16.	14.	16.	16	15	15

As is explained in [7] random forests work well when there is a “high” probability that a “strong” variable (in our example one of the 20 components with group mean difference) is chosen at some node while there is a “small” probability that only “weak” variables (the remaining 100 components) are selected at every node. We add to this reasoning that in small sample problems there is nonnegligible probability that one of the weak components will be spuriously strong ,i.e. exhibit a good but meaningless separation (of the 16 patients) by chance at some node. This could however only occur in a (nonnegligible but) relatively small fraction of the forest. By considering corruption of experts we account for this influence.

Carrying out the optimization with **(B)** for practical datasets will be continued in further work. Here we omly give an initial set of weights for **(B)** (corrupt version) for the constrained optimization and compute the terms in the objective function:Let $\pi_o = .02$ be probability of corrupt expert, take $\sigma = .25\mathbf{I}$ and (for fast computation) use the linear solution $\mathbf{w}^i$ of Thm. II (M=.5) ; so $B_i(w^i)^{1/2} = (2r_i)^{-1} ||\Sigma\ w^i_j\ x^i_j\ ||$ and it is easy to get the bound $B_i^*(w^i)^{1/2} = B_i(w^i)^{1/2} + \Sigma |w_j^i|$. (r_i is the distance to the furthest x^i_j , π_o was an estimate of the probability of a near perfect separation for at least one noise gene in two random subsets of 8 genes). For thresholds .06,

.055,.05,.045,.04,.035, we thin the forest grown for each patient q (forming an "orchard") by removing those trees for which the full bound in Thm.II applied to the terminal node exceeded the threshold. We find .04 to be the smallest threshold such that each corresponding orchard contained at least 100 trees. Then, for the orchards corresponding to .04, we use the uniform weighting for the α's in each bound as starting point for the optimization. (The next step would be to perform alternating steepest descent using all of the trees in the forest.) The resulting estimates and bounds appear in Table 2. The "forest" column is Breiman's average estimate.The "fusion" columns represent F, the probability estimate, and G*(w, α), the associated square error bound under corruption in (**B**). h is the penalty value based on the degree of conditioning and F.

Table 2

	Probability of Group A membership			Local Sq. Error Bound			Patient
λ = .1, β = .25	forest	fusion	kernel	fusion	h	kernel	
	0.86	1.00	0.87	0.042	0.000	0.064	1
	0.91	0.88	1.00	0.034	0.029	0.066	2
	0.89	1.00	1.00	0.043	0.000	0.061	3
	0.86	1.00	0.87	0.042	0.000	0.086	4
	0.79	0.89	0.96	0.035	0.028	0.063	5
	0.89	0.89	0.86	0.035	0.028	0.054	6
	0.86	1.00	0.90	0.042	0.000	0.056	7
	0.95	1.00	1.00	0.042	0.000	0.070	8
	0.18	0.20	0.10	0.031	0.033	0.038	9
	0.15	0.08	0.08	0.037	0.026	0.039	10
	0.17	0.00	0.05	0.042	0.000	0.040	11
	0.09	0.00	0.06	0.042	0.000	0.039	12
	0.13	0.00	0.01	0.042	0.000	0.042	13
	0.17	0.00	0.02	0.042	0.000	0.042	14
	0.11	0.00	0.00	0.042	0.000	0.042	15
	0.07	0.07	0.01	0.037	0.025	0.041	16

VII. Estimation for General Nonlinear Functions: Error Bounds and Improved Estimators for Kernel Vector Machines

Although many naturally ocurring situations can be handled by the contextual estimators of section IV, there are cases that do not fit those described in section IV-C. So we apply our proposal to the more general vector machine model: If a dictionary of functions is the set of translates of a kernel $K_{x'} = K(x',)$ which generates a reproducing kernel Hilbert space (denoted by RKHS; treated rigorously in subsection A) and if f is within $\varepsilon(x)$ of a

(possibly infinite) weighted sum of dictionary elements which is bounded by M in RKHS norm, then, as we shall show, a dimensionality reduction occurs for the minimax analysis. Hence a query-based local minimax counterpart to Tikhonov's regularized kernel and Vapnik's global support vector surface estimation is derived.

The vector machine(VM) set up may be described as follows.(See [32].) Let K(u,v) be a positive semidefinite, piecewise continuous, bounded, nonnegative, symmetric function on the cartesian product of a compact subdomain V of R^d with itself. Assume further that K(u,v) is positive at diagonal points (u,u). Let us map each x in our predictor space to $\Phi(x) = (\lambda_1^{1/2}\phi_1(x), \lambda_2^{1/2}\phi_2(x), ..., \lambda_i^{1/2}\phi_i(x), ...) \varepsilon \Phi$ (maybe infinite dimensional) where λ_i , ϕ_i are the eigenvalues and orthonormal eigenfunctions of the integral operator with kernel K(u,v) : $\int_V$ K(u,v)g(v)dv. The task is now to employ affine estimation using the linear span of the mapped predictor data $\{\Phi(x_j)\}$, i.e.use functions of the form $c_o + (c_1\Phi(x_1) + c_2\Phi(x_2) + ...) * \Phi(x)$ for x in R^d. By Mercer's theorem $\Phi(x)\cdot\Phi(y) = K(x,y)$ so that all computations in Φ can easily be done using the kernel function K(x,y).

The vector machine methodology can be equivalently presented by staying in R^d and using functions which lie in the RKHS . In fact the $\Phi(x)$ above will correspond to the function K(x, -). We adopt this approach here since it makes the treatment of $\varepsilon(x)$ and some of the context assumptions more natural and it is indeed functions on R^d that we are trying to estimate. (See [37].) The various loss functions used in that setting correspond to various approaches-hinge loss-Vapnik's support vector machine [43], squared error loss-least squares vector machine [32],etc.... The details begin in our first subsection.

A. Finite Sample Minimax Bounds for Local Estimation by Sums of Kernels

Consider the pre-hilbert space of models $f(x; a) = \sum a_{x'} K(x', x)$ where the sums are initially over finitely many x' and where K(u, v) is a piecewise continuous, bounded, symmetric, nonnegative kernel function on V x V, positive at diagonal points (u,u), and for

which the matrix $K(x_i, xj)$ is positive (semi) definite for any finite (non) distinct $\{x_i\} \subseteq V$. Define an inner product $[f (x; a) , f (x; b)] = \sum \sum a_{x'} b_{x''} K(x', x'')$. Now extend this to form a real Hilbert space by completion. For any g in the constructed Hilbert space g can be identified with the pointwise limit of a sequence of models in the pre-hilbert space which converges to g in the constructed Hilbert space. It can easily be shown that $[g , K(u, -)] = g(u)$ where g(u) is the value of the associated pointwise limit at u. Hence the space is called a reproducing kernel Hilbert space (RKHS). Consider the set of models f (x; a) in this space with RKHS norm $|| \ ||$ bounded by M (a now varies in an abstract infinite dimensional space). We assume f(x) is within $\varepsilon(x)$ of one of these. Unlike some of the dictionaries described previously ([20],[3]) one can not reasonably assume that f is exactly a weighted sum of the form (#) of kernel translates since the kernel width remains fixed and the norms are bounded by M ; hence $\varepsilon(x)$ enters into the analysis.

One of the main techniques in the minimax derivation in this setting is the simplification of the problem using functional analysis in Hilbert space. This is best motivated by seeing what it does for the global penalized estimation problem with known diagonal noise matrix. Here (as in [37a]) we minimize empirical loss with a Tikhonov regularization penalty term and use the minimizing model as estimate (see section II-C.) i.e. find minimizing g's (if they exist) in the RKHS for

$$Q(g) = 1/k \sum \sigma_j^{-2} L (Y_j , g(x_j)) + \gamma || g ||^2$$ with (throughout section VII) $|| \ ||$ equal RKHS norm and k distinct predictors $\{x_i\} \subseteq V$. Now consider any g in the RKHS and write

$$g(x) = \sum a_i K(x_i, x) + p(x)$$ where p(x) is orthogonal to each $K(x_j, x)$.

Let $g^+(x) = \sum a_i K(x_i, x)$. By the reproducing property $g(x_j) = g^+(x_j)$ for each j.

Also $|| g^+ ||^2 \leq || g ||^2$. Therefore $Q(g) \geq Q(g^+)$.

Hence finding a solution (if it exists) to the global penalized estimation problem reduces to searching for the best (minimizing Q(g)) estimator g which is a linear sum of k kernels each centered at one of the sample predictors. This is called the "Representer" Principle.

In fact the above proof shows that this principle holds more generally when $Q(g) = H(g(x_1), g(x_2),...g(x_k)) + \gamma \| g \|^2$.

For squared loss the solution exists and is called the Tikhonov regularization ([37]). In matrix notation this is the minimum of $k^{-1}(K\mathbf{a} - \mathbf{Y})^t \sigma^{-1}(K\mathbf{a} - \mathbf{Y}) + \gamma\, \mathbf{a}^t K\, \mathbf{a}$ wrt. $\mathbf{a}$, the vector of coefficients of the k kernels, where σ is the diagonal matrix of $\{\sigma_j^2\}$ and $K = K(x_i, x_j)$. The solution may be written $\mathbf{a} = (k\gamma \mathbf{I} + \sigma^{-1} K)^{-1} \sigma^{-1} \mathbf{Y}$. Note this formula is also the associated Tikhonov regularization for any positive definite σ since the representer theorem clearly holds for $Q(g) = k^{-1}(\mathbf{g} - \mathbf{Y})^t \sigma^{-1}(\mathbf{g} - \mathbf{Y}) + \gamma \| g \|^2$ with $\mathbf{g} = (g(x_1), g(x_2),...g(x_k))^t$. (If we include a constant in the estimator $g(x) = b + \sum a_i K(x_i, x)$, let $L (y, g(x)) = (1 - y\, g(x))^+$ and consider classification problems with responses 1 or -1, one obtains a quadratic programming problem when minimizing Q(g) with ||g|| defined as the RKHS norm of g-b. The solution corresponds to Vapnik's support vector hyperplane. In most cases the results of classifying by either Vapnik or Tikhonov methods is similar. See [37], [39]. In these references +1,-1 are used for the classes instead of 0,1 as we use. Their formulas are obtained easily from ours; eg. when $\sigma_j = 1/2$ making $\sigma = \mathbf{I}$ above gives their formula for the kernel coefficients.)

Now in the proof of our following Theorem VI below the key step will be to maximize

(7) $L (g(0), \sum w_j g(x_j))$ subject to $\| g \| \leq M$.

Now consider any g in the RKHS with $\| g \| \leq M$ and write (with $x_o = 0$)

$g(x) = a_o K(x_o, x) + \sum a_i K(x_i, x) + p(x)$ where p(x) is orthogonal to each $K(x_j, x)$.

Let $g^+(x) = a_o K(x_o, x) + \sum a_i K(x_i, x)$. By the reproducing property $g(x_j) = g^+(x_j)$ for each j and $\| g^+ \| \leq M$. So $L (g(0), \sum w_j g(x_j)) = L (g^+(0), \sum w_j g^+(x_j))$.

Hence (if a solution exists) the constrained loss maximization problem, where the loss

is measured between the target at the query and a smoother applied to the target at the training predictors, reduces to searching over linear sums of k+1 kernels each centered at one of the training predictors or at the query. Thus we have established what we call the maximum "Representer" Principle. Clearly this principle holds more generally for $Q(g) = H(g(x_o),g(x_1),...g(x_k))$ subject to $|| g || \leq M$.

We now show how to solve (3) (approximately for general $\varepsilon(x)$ and exactly with an explicit formula for $\varepsilon(x) = 0$) when a bound M on $|| f(x;a) ||$ is assumed for the approximating model. In fact, for the exact case, the optimal weight vector w is of the form of a Tikhonov regularization where our minimax theory has determined the regularization parameter γ as a function of M and hence our minimax error bound provides a local error estimate for the appropriate Tikhonov regularization. For simplicity we state and prove the result for distinct predictors. It holds more generally for non-distinct predictors with the objective function modification mentioned in Thm. I.

Theorem VI (Minimax Query-Based Vector Machine) Let f(x) be within $\varepsilon(x)$ in V of some member of the family { f (x ; a) generated by K(x',x): the RKHS norm of f(x;a) is less than or equal to M}. Assume distinct predictors and mean zero covariance upperbounded noise. Use squared error loss. Consider the matrix $K = ((K(x_i, x_j))) : i,j = 0,1, 2,.....k$. (V is compact and contains the query point x_o which we are taking as 0 but the results obtained are the same for any query point.). Set $w_o = -1$ (w has now k+1 components) , $\sigma_{oj}^2 = \sigma_{io}^2 = 0$, $\boldsymbol{\sigma}$ equal the k+1 by k+1 matrix formed by adding a 0'th row and 0'th column of 0's to the noise covariance matrix upper bound, and the k+1 dimensional vector $\mathbf{u} = (1,0,0...0)^t$. Let

$$L (\mathbf{w}) = w^t \boldsymbol{\sigma} w + B(w) \quad \text{where} \quad B(w) = \left(M (w^t K w)^{1/2} + \sum |w_j| \varepsilon(x_j) \right)^2$$

and let the local complexity

$$\mathcal{L} = 1 / [\mathbf{u}^t (\boldsymbol{\sigma} + M^2 K)^{-1} \mathbf{u}] .$$

Then the mean squared error of F (**w**), where w* = 0, is bounded by L (**w**) which is greater than or equal $\mathcal{L}$. For $\varepsilon(x) = 0$, if w = arg min L(**w**) and w* = 0, then L (**w**) is just the local complexity $\mathcal{L}$ and this is the best possible bound (solution to (3)) on mean squared error under this assumption. Finally, in the latter case,

$$w = -[(\sigma + M^2K)^{-1}\mathbf{u}] / [\mathbf{u}^t(\sigma + M^2K)^{-1}\mathbf{u}]$$

and, for known noise covariance σ, this just gives the Tikhonov regularization at x_o for the global estimation problem with the regularization paremeter $\gamma = k^{-1}M^{-2}$. Hence $\mathcal{L}$ is a best bound on the mean squared sampling error of the global vector machine estimator at x_o. Also $\mathcal{L}$ may be computed as a function of of M and w may be chosen using the bound $\mathcal{L}(M)$ as operating characteristic. In summary just as Bayesian justifications for Tikhonov regularization exist (see[37]) justifications via classical minimax statistical theory have here been established.

---

proof of Theorem VI: Let all sums be from 0 to k . Adding 0's to **N** as with σ we write

$$(8) \quad E((F(\mathbf{w})-f(0))^2 \mid x_1,..) = w^t \mathbf{N} w + \{ \sum w_j g(x_j) + w^* + \sum w_j \zeta(x_j) \}^2$$

with $|\zeta(x)| \leq \varepsilon(x)$ and where g lies in the RKHS and has norm less than or equal M. Since the set of such g's is invariant under negation, the bracket term above will be bounded in absolute value by maximizing $|\sum w_j g(x_j)|$ subject to the given condition on g and adding $|w^*| + \sum |w_j| \varepsilon(x_j)$ to it inside the brackets. It now is clear that the optimal w* = 0. Such g exists because it maximizes (7) when L is squared error loss since we may obtain the maximum value of $|\sum w_j g(x_j)|$ by maximizing $|\sum a_i (\sum w_j K_{ij})| = |\mathbf{a}^t K w|$ subject to $\mathbf{a}^t K \mathbf{a} \leq M^2$ as follows: we may restrict **a** to lie in the range of the map defined by the matrix K. The constraint region is now nondegenerately ellipsoidal. Then the simple optimization problem may be solved by noting that the maximizing **a** will also minimize $\mathbf{a}^t K \mathbf{a}$ subject to $\mathbf{a}^t K w = s$ for some value of s. So **a** is a critical point of $\mathbf{a}^t K \mathbf{a} - 2\lambda(\mathbf{a}^t K w - s)$ where λ is a Lagrange multiplier which

implies $K\mathbf{a} = \lambda K w$ and $\mathbf{a}^t K = \lambda w^t K$. Since $\mathbf{a}^t K \mathbf{a} = M^2$ at the optimizing $\mathbf{a}$, $\lambda^2 (w^t K w) = M^2$ and the maximum $| \mathbf{a}^t K w | = M (w^t K w)^{1/2}$. This yields the upper bound $L(\mathbf{w})$. For $\varepsilon(x) = 0$ we minimize $w^t (\sigma + M^2 K) w$ subject to $w_o = -1$. Since $K_{0\,0}$ is positive it is easy to see that $\sigma + M^2 K$ is nonsingular. A routine Lagrange argument now yields $w = -[(\sigma + M^2 K)^{-1} \mathbf{u}] / [\mathbf{u}^t (\sigma + M^2 K)^{-1} \mathbf{u}]$ and the minimum is just $\mathcal{L} = 1 / [\mathbf{u}^t (\sigma + M^2 K)^{-1} \mathbf{u}]$.

To show that the solution when $\varepsilon(x) = 0$ is just the global regularization solution for a particular regularization parameter value rewrite $L(\mathbf{w})$ where the vector w and matrices σ and K are now k-dimensional (i.e. remove the 0th rows,columns, etc.) as follows

$$L(\mathbf{w}) = w^t \sigma w + M^2(w^t K w + K(x_o,x_o) - 2\, w^t \mathbf{k_o})$$

where $\mathbf{k_o}$ is the vector with components $K(x_o,x_i)$. Now set the gradient wrt. w to 0 obtaining

$$w = M^2 (\sigma + M^2 K)^{-1} \mathbf{k_o} \quad \text{or} \quad F(\mathbf{w}) = w^t \mathbf{Y} = \mathbf{k_o}^t (M^{-2} \mathbf{I} + \sigma^{-1} K)^{-1} \sigma^{-1} \mathbf{Y} .$$

The global estimator at x_o is $\mathbf{k_o}^t \mathbf{a}$ where $\mathbf{a} = (k\gamma \mathbf{I} + \sigma^{-1} K)^{-1} \sigma^{-1} \mathbf{Y}$ is the associated Tikhonov solution obtained earlier. But $\mathbf{k_o}^t \mathbf{a}$ is the minimax estimator F when $\gamma = k^{-1} M^{-2}$. **QED.**

From the theorem we see that, as in the linear estimation case, the local estimate is the same as the global estimate when there are no context assumptions and $\varepsilon(x)$ is 0. Researchers who use vector machines and have heuristics for determining the regularization paremeter value γ (and hence equivalently M) as a function of the kernel and the data now have a mean squared error bound at any query point for the global estimator. This bound could then be used to determine the kernel K(,) for a local analysis using an information measure as in section C. The local bound on error appears to be new and is indeed local since it varies with the query point.

The choice of M remains a key challenge. But with context assumptions appropriate choices of M are possible as we now demonstrate: We apply our local estimation method to functions f(x) which take values between 0 and 1 and which are within $\varepsilon(x)$, in the full given

neighborhood V, of an h(x) which is a member of the RKHS associated with the fixed kernel K(x',x). Denote by RKHS(A) the RKHS of functions on A generated by kernels centered at some a in A. We also want f(x) to possess an approximate appropriate degree of smoothness depending on the kernel bandwidth. We will incorporate the smoothness and the range (of f in [0,1]) context into the family { f(x;a) } and remove the restriction that the RKHS norm is bounded. For many kernels (e.g.rectangular, gaussian) RKHS(A) is dense in L_2 (A) so we need to restrict the admissible linear combinations of kernels. A natural class of functions is the restrictions to V of the cone of positive finite linear combinations, PK(V), of kernels centered at points in V. So let { f(x;a) } be PKB(V) = {g(x) in PK(V):g(x) $\leq$ 1 in V } . This is a special case ($\alpha = 0$) of a more general classes of approximands, which admit tight error bounds with our approach and are given by {f(x;a)} = $PK_\alpha(V)$ = { α + (1-α)g_1 -αg_2 : g_i in PKB(V) } where α lies in [0, 1].

A direct method to get a bound for (3) with such {f(x;a)} would require a maximization step which is 2 linear programs with arbitrarily many variables: indeed, for $\alpha = 0$, from (8) for given **w** we would maximize | Σ w_j g(x_j) + w* | for g in PKV(B) or equivalently a sup of | $\mathbf{a}^t K_1$ w + w* | over abritrarily long positive vectors a = (a_r) and corresponding arbitrary points z_r in V where K_1 = ((K(z_r, x_j))), K_2 = ((K(z_r, z_q))) and subject to the additional constraints that $\mathbf{a}^t K_2$ have components not exceeding one. (The general α case is similar.)This would then have to be approximately solved for varying **w** to get a near minimum. Furthermore the procedures have to be repeated for various neighborhoods and kernels according to our upcoming proposals.

Hence we consider an approximation with a closed form answer which will lend itself to our proposals. We seek the smallest ball centered at 0 in RKHS(V) that contains $PK_\alpha(V)$ - α. Let F*(**w**) be our estimator. We will get a bound on local estimation error by applying Theorem VI using a dimension free bound M_V on the norm of the g's in PK_α (V) - α in the present context. (We can also extend the analysis easily when locally estimating f(x) which takes values in [v,y].)

Theorem VII (Vector Machine with Context) Assume the hypotheses of Theorem VI except that

f(x) takes values in [0,1] and is within $\varepsilon(x)$ of g(x) in V where g is in $PK_\alpha(V)$. Then the estimator F*(**w**), which equals F(**w**) except $w^* = -\alpha \sum w_j$, has mean squared error bounded by L(**w**) of TheoremVI (which equals $\mathcal{L}$ in the exact case provided w = arg min L(**w**) and $w^* = -\alpha \sum w_j$) where we take

$$M = M_V = \left(\alpha^2 + (1-\alpha)^2 \right)^{1/2} \left(\max_{x \varepsilon V} \min_{y \varepsilon V} K(y, x) \right)^{-1/2} \geq \sup_{PK\alpha(V)} \left(\| g(x) - \alpha \| \right)$$

For such M we call F*(**w**) the contextual Tikhonov estimator. A good choice for α is .5 (which is assumed in Section B.) since it minimizes M_V as a function of α. In fact, for any M greater than or equal the right hand side of the above inequality, the same result holds.

---

proof of Theorem VII : Rewrite the right hand side of (8), where g is in $PK_\alpha(V)$, as

$$w^t \mathbf{N}\, w + \left\{ \sum w_j \left((1-\alpha) g_1(x_j) - \alpha g_2(x_j) \right) + w^* + \alpha \sum w_j + \sum w_j \zeta(x_j) \right\}^2$$

where g_i is in PKB(V). Now write $(1-\alpha)g_1(x_j) - \alpha g_2(x_j)$ as $h(x_j)$ above where $h = g - \alpha$. Then h has norm bounded by any $M \geq \sup_{PK\alpha(V)} \left(\| g(x) - \alpha \| \right)$. By the proof of TheoremVI, for any such M, a bound on the error is L(**w**) provided the constant term in { " } $= w^* + \alpha \sum w_j = 0$. This holds also for the minimizing **w**. To prove the inequality first note that, for g in $PK_\alpha(V)$,

$$\| g(x) - \alpha \| = \{ (1-\alpha)^2 \| g_1(x) \|^2 + \alpha^2 \| g_2(x) \|^2 - 2\alpha (1-\alpha)(g_1, g_2) \}^{1/2} \leq \{ (1-\alpha)^2 \| g_1(x) \|^2$$

$$+ \alpha^2 \| g_2(x) \|^2 \}^{1/2} \leq \left(\alpha^2 + (1-\alpha)^2 \right)^{1/2} \sup_{PKB(V)} \left(\| g(x) \| \right)$$ since (g_1, g_2) is non-negative by the non-negativity of both the kernel and the kernel coefficients for g's in PKB(V). Next we bound ||g|| for g in PKB(V): Note that $g(x) = \sum a_i K(z_i, x) \leq 1$ and hence

$\| g(x) \|^2 = \sum a_j \left(\sum a_i K(z_i, z_j) \right) \leq \sum a_j$. Now both the kernel and the a_j's are nonnegative so $1 \geq g(x) = \sum a_j K(z_j, x) \geq \left(\sum a_j \right) \left(\min_j K(z_j, x) \right) \geq \left(\sum a_j \right) \left(\min_y K(y, x) \right)$. So $1 \geq \left(\sum a_j \right) \left(\max_x \min_y K(y, x) \right)$ or $\| g(x) \| \leq \left(\sum a_j \right)^{1/2} \leq \left(\max_x \min_y K(y, x) \right)^{-1/2}$.

QED. (We thank Alex Kheifets for pointing out two inequalities which lead to the above bound.)

It is interesting to see the geometry of RKHS(V) behind this theorem: For $\alpha = 0$ we are imbedding the truncated cone PKB(V) in a ball centered at the origin. For $\alpha = 1/2$ we are imbedding the differences of 2 truncated cones, (1/2)(PKB(V) - PKB(V)), in a ball yielding a better bound since the latter ball has radius at most $2^{-1/2}$ times that of the former ball. Finally, according to the proofs of Theorem VI and Theorem VII, the maximum Representer Principle yields a simple strategy with the given bounds. Research is still ongoing to determine how well M_V approximates $\sup_{PK\alpha(V)} (\| g(x) - \alpha \|)$. It is conjectured to be very accurate in the Gaussian cases proposed below when the kernel bandwith is at least half the radius of V.

One way to apply the above theorem is to use it for different neighborhoods V and approximands with RKHS norm bounded by M_V ,which is a function of the neighborhood V, and then minimize the upper bound of the theorem $L(\mathbf{w})$ as a function of V . For class 2 probability functions in an RKHS on a bounded neighborhood of the query point we now present this approach in section B for any given kernel K(,). Then, in section C, we use the associated minimax error bound to determine K(,) by optimizing an information measure.

B. An Improved Estimator for Learning Class Membership Probabilities on a Vector Machine with a Given Kernel

We first remark that the two theorems in the previous section are valid when the target function is defined only at the points $x_o, x_1,...x_k$ but there exists an extension f(x) to all of V which satisfies the hypotheses of the theorem. If the kernel has a sufficiently small bandwidth then a reasonable assumption is that the extension is in $PK_{1/2}(V)$ or in other words the approximand can be assumed to take the same values as the target at $x_o, x_1,...x_k$ and hence in all of V. (The more general case of extensions within $\varepsilon(x)$ of the approximands is straightforwardly similar. We do not present it to keep notation simple.) Note that overlapping neighborhoods may require extensions which don't agree on the intersection although they must agree at the common predictors and query. With this in mind we proceed to apply Theorem VII to estimating class 2 probability. (A similar analysis applies when the target is known to take

values in any bounded interval.)

Assume the probability function we are estimating is defined only at the sample points and query but, for each V_i in a class of compact neighborhoods $\{V_i\}$, it has an extension to V_i of the exact form h_i in $PK_{1/2}(V_i)$. The h_i's may be different on the intersection of different neighborhoods as long as they agree at the predictors and query common to that intersection. Now apply the local Theorem VII on V_i , i.e. use only the predictors in V_i forming kernel matrix K_i . This determines a bound $\mathcal{L}_i$ on mean squared error

$$\mathcal{L}_i \quad = \quad 1 \,/\, [\, \mathbf{u}^t (\sigma \; + \; M_{V_i}^{\,2} K_i)^{-1} \mathbf{u} \,]$$

One expects that the M_{V_i} will increase as the neighborhoods increase (this is a conjecture; we only have a proof for Gaussian or rectangular kernels) so that there will be a tradeoff between sample size and function complexity as V_i increases. The <u>improved Tikhonov estimator</u> of $f(x_o)$ is the estimator described above for $V = V_r$ corresponding to the index r for which $\mathcal{L}_r = \mathcal{L}^* = \min \mathcal{L}_i$ and the latter quantity is a bound on the mean squared error for any class 2 probability function whose extension to V_r is in $PK_{1/2}(V_r)$. This bound should often be much better that that of the <u>contextual Tichonov estimator</u> for which all of the predictors are used but for which the ball of approximands is much bigger than that for a smaller neighborhood.

In the case of a Gaussian kernel with covariance Σ , $K(x', x) = \exp\{-.5(x - x')^t \Sigma^{-1} (x - x')\}$, a good choice of neighborhoods is-

$$V_i \;=\; V_i(\Sigma) \;=\; \{\, x : (x - x_o)^t \Sigma^{-1} (x - x_o) \;\le\; \nu_i \,\} \qquad i=1,2,....k.$$

where the order statistic ν_i is the i'th smallest value in the set $\{(x_j - x_o)^t \Sigma^{-1} (x_j - x_o): j=1,...k\}$.

Then, by a straightforward calculation, we obtain $M_{V_i} \;=\; 2^{-1/2} \exp\{ .25\nu_i \}$.

C. Determination of Optimal Local Kernel Shape for Learning Class Membership Probabilities on Vector Machines without Cross Validation

Consider a kernel K and collection $\mathcal{V}_K$ of compact neighborhoods $\{V_i(K)\}$ of x_O which may vary with K. We assume that K (x', x) has the form of a probability density as x varies in R^d with parameter x' and K is defined on all of $R^d \times R^d$. Such kernels include the Gaussian density, $(\det(2\pi\Sigma))^{-1/2} \exp\{-.5(x - x')^t \Sigma^{-1}(x - x')\}$, where x' represents the location parameter. We define the local information criterion of any $PK_{1/2}(V_i(K))$ by the Fisher information of the kernel K which we denote by $\mathcal{I}(K)$:

$$\mathcal{I}(K) = \int_S \text{trace}\{(\text{grad}(\ln K (x', x)) (\text{grad}(\ln K (x', x)))^t\} K (x', x) \, dx$$

where grad represents the x' -gradient. This is just the trace of the Fisher information matrix for the d-parameter family of densities K (x', x). In fact for the Gaussian density it reduces to $\mathcal{I}(K)$ = trace $\{ \Sigma^{-1} \}$ which we denote also by $\mathcal{I}(\Sigma)$.

Let us continue the analysis assuming the density is Gaussian, using Σ instead of K in the notation. Assume Σ and $\mathcal{V}_\Sigma$ varies over bandwidths(covariances) for which the probability function has the appropriate extensions to members of $\mathcal{V}_\Sigma$. Let $\mathcal{L}^*_\Sigma$ be the error bound of the improved Tickonov estimator when the covariance is Σ. (Use the same neighborhoods as in B.)

We can now state a local kernel shape strategy of Neyman-Pearson type as

$$(\mathcal{E}_g) \qquad \Sigma = \arg \min_{\Sigma' : \mathcal{I}(\Sigma') \geq \eta} \mathcal{L}^*_{\Sigma'},$$

(minimum error for given information)

where η is a minimum information bound.

In carrying out the paradigm in the Gaussian density kernel case for a particular application it may be important to further limit the domain of possible covariances Σ. For example if we want to keep bandwidth above some value in all directions we may restrict Σ to

have the form $\Sigma + \mu \boldsymbol{I}$. This will prevent the estimate from overly depending on the projection of the data in a single direction. Also we see that carrying out this strategy in the Gaussian case by finding a near solution to (3) at each step would require enormous linear programs as **w** varies while using the contextual bounds requires only calculating M_{V_i} which in the Gaussian density case is given by $M_{V_i} = 2^{-1/2} \exp\{ .25\nu_i \} (\det(2\pi\Sigma))^{1/4}$.

There are several challenges in implementing this strategy with data. Further research work is necessary before practical implementations and evaluations can be produced of the algorithms proposed here. We do present a preliminary application of the contextual Tikhonov estimator and bound to bioinformatics in the next section where we have just used a heuristic procedure to determine Σ.

D. Estimation of Class Membership Probabilities with Error Bounds for the Microarray Example

Kernel vector machines have already been used to classify microarray data (see [33]). We reanalyze the microarray data from the University of Pittsburgh simulator using the local minimax kernel estimation bound obtaining estimates of probability of correct classification.(Phil O'Neil developed and implemented the software for this project.) Since research is still ongoing to determine the accuracy for M_V and to practically implement the improved estimator and the shape finding algorithms, we present only an application of Theorem VII. Local minimax estimation of each patient's probability of membership in Group A was performed using linear combinations of Gaussian kernels centered at the predictors of the 15 other patients (leave-one-out method).The kernels were degenerate with a one dimensional distribution in the direction of the eigenvector with the largest eigenvalue of the (120 dimensional) sample covariance of the other 15 patients. The standard deviation σ was taken to be 1/3 the diameter of the projection of the 15- predictor dataset in that direction. Using the bound of Theorem VII with $\alpha = .5$, this would correspond approximately to using $M_V = 2^{-1/2} e^{9/16} = 1.23$ for the contextual Tikhonov estimator with $\alpha = .5$.

The probability estimates are in Table 2. The fifth column represents the (worst case) bound on the mean squared error of the probability estimated. The probability estimates were obtained after the data mining phase. The correct classes entered the calculation only after weights were furnished relative to each patient not left out. These weights were summed over those in Group A and the result equaled the probability estimate. Finally we note that the vector machine estimation procedure, as we have just applied it to the microarray data, is rotation invariant to the training set.

VIII. Remarks on Solutions for General Loss Functions :

For the global estimation using the methods reviewed in sections II A. and II B. but without the "K_h (D(0,x_j)) " factors and where the sum is over all the predictors,more general loss functions have (and should have) been used. For instance, with noises belonging to the exponential family, maximum likelihood is equivalent to a particular loss function which may be non quadratic. The same need exists to consider more general loss functions for the local application (with the " K_h (D(0,x_j)) " factors) . But with our method the loss is only needed at the point (f(0) , F(**w**)). In the case of independent responses and bounded noise, F(**w**) is near normal or at least more bell-shaped distributionally than any of the Yj's which are the first components in the loss function contributions in the other methods. Hence, in this case (which occurs for just modest sample sizes), the squared error loss is appropriate since our method "superimposes" all information at the query point . Nevertheless our results need to be extended to more general loss functions when unbounded noise and hence extreme outliers are present and then nonlinear modification of the estimators is appropriate using sensitivity and influence.

IX. Locally Quadratic and higher order Models: Problems of Real Algebraic Geometry with further Applications to Learning and Inverse Problems

Let us consider equation (4) and the notation of section III. which can be understood in both the inverse problem or regression setting (using δ functions for the θ_j 's and Σ $|w_j|$ $\varepsilon(x_j)$ instead of R(w) in deriving the bounds). The bounds(using our method) would be determined

by first characterizing the set of (a_o,a_1, ..., a_q) for which there is an f satisfying $\mathcal{C}$ and (2) (or a slightly larger set of a's) . This is the real geometry problem (real algebraic geometry if the { $h_i(x)$ } are algebraic functions). Next we maximize the objective function $\left| \sum a_i \left(\sum w_j H_{ji} \right) + C\, a_o + w^* \right|$ with respect to such a's. This maximum value is then added to R(w) inside the brackets to get the bound for each **w**. In all the cases we consider the set of a's is convex and the maximization problem is one of determining the tangent hyperplanes for the boundary of this set which are also level surfaces for the objective function.

For learning applications, where one wants to estimate class 2 posterior probability, bounds for locally quadratic models in d-dimensions would be desirable. Suppose $h_1(x),.....,h_q(x)$ are the $q = 2d + d(d-1)/2$ monomials of orders 1 and 2. Can we characterize the set of a's for which $a_o\, h_o(x) + a_1\, h_1(x) + ... + a_q h_q(x)$ is in the unit interval whenever x is in a given ball about 0? Furthermore can we characterize the hyperplanes tangent to its boundary and perform the appropriate maximization? We have done this for (the less challenging learning case) d = 1and will publish the details elsewhere. This has applications in Markov chain Monte Carlo computation of P-values for exact tests of model validity in multifactor experiments (see[25]).

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:

We reconstruct the value of f at 0 from data consisting of noisy integrals of f over V = [m, m+1] which contains 0. (Reconstruction at an arbitrary point in an arbitrary interval can be done by a simple linear reparameterization.) All integrals are over V. Write (in a different paremetric form with $h_i(0)$ nonzero)

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

Then

$$E\{(F(\mathbf{w}) - f(0))^2 \mid \theta_1,..\} = w^t \mathbf{N} w + \left\{a_1 \mathbf{Q}_1 + a_2 \mathbf{Q}_2 + C a_o + w^* + \mathbf{S}\Sigma w_j \theta_j(t)\, \zeta(t)\, dt\right\}^2$$

where $C = \Sigma w_j \mathbf{S} \theta_j(t)\, dt - 1$, $\mathbf{Q}_1 = \Sigma w_j \mathbf{S} (t - m)^2 \theta_j(t)\, dt - m^2$ and

$\mathbf{Q}_2 = \Sigma w_j \mathbf{S} (t - m)\theta_j(t)\, dt + m$.

Since a_o 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) space for which $u(x) = a_1x^2 + a_2x$ has oscillation bounded by M on the unit interval. (This involves solving the simultaneous inequalities: $| a_1x_i^2 + a_2x_i - a_1x_k^2 - a_2x_k | \le M$ for pairs (i,k) where x_i,x_k are 2 of the(2 or 3) critical and end points for u(x) in [0,1].) In general we get the bound by using the same set with M replaced by M + 2e where e is the maximum of $\varepsilon(x)$ in V.

We continue the analysis for the exact case. In Fig. 2 the set 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 a_1, a_2 space. So in maximizing $|a_1 \mathbf{Q}_1 + a_2 \mathbf{Q}_2 + w^*|$ we can choose $a_1\mathbf{Q}_1 + a_2\mathbf{Q}_2$ to have the same sign as w*. Hence w* =0 will minimize the bound for any w'. So we need only maximize $|a_1 \mathbf{Q}_1 + a_2 \mathbf{Q}_2|$.

For this we determine a tangent hyperplane to the boundary of the form $a_1\mathbf{Q}_1 + a_2 \mathbf{Q}_2 = \pm t$ and use $w^t \sigma w + |t|^2$ for the best bound for fixed w'. (In general case we add R(w) 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 = $-(M/-a_1)^{1/2}$

(-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 Φ_1, Φ_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|. So by a straightforward but extremely tedious calculation we find that the maximum of $|a_1 \Phi_1 + a_2 \Phi_2|$ is given by $B(M, \Phi_1, \Phi_2) =$

$|M \Phi_2|$ if $\Phi_1 = \Phi_2$

$|M \Phi_2^2 / \Phi_1|$ if $.5 < \Phi_1 / \Phi_2 < 1$

$|2M \Phi_2 - M \Phi_1| / (2 - (\Phi_1 / \Phi_2))^2$ if $1 < \Phi_1 / \Phi_2 < 1.5$

$4M | \Phi_1 - \Phi_2 |$ otherwise

For the upper bound on the minimax value we use $B(M+2e, \Phi_1, \Phi_2) + R(\mathbf{w}')$. We restate this as

Theorem VIII (Recovery of f(0) from integral data on [m,m + 1]) Let the parametric family be given by $f(x ; a) = a_o + a_1(x - m)^2 + a_2(x - m)$. Assume that,in [m,m + 1] , f(x) is within $\varepsilon(x)$ of some family member f (x ; a). Use squared error loss. Assume $\mathcal{C}$ is the condition that f has oscillation at most M on [m,m + 1]. Assume

$$\Sigma w_j \int \theta_j(t)\, dt = 1 \quad , \quad w^* = 0,$$

$$L_{\mathcal{C}}(\mathbf{w}) = w^t \sigma w + \left\{ B(M+2e, \Phi_1, \Phi_2) + \int \Sigma | w_j \theta_j(t) | \varepsilon(t)\, dt \right\}^2$$

and

$$l_{\mathcal{C}}(\mathbf{w}) = w^t \sigma w + \left\{ B(M, \Phi_1, \Phi_2) \right\}^2 .$$

Then the mean squared error of F (**w**) is bounded above by $L_{\mathcal{C}}(\mathbf{w})$ and

$$\min_{w} l_{\mathcal{C}}(\mathbf{w}) \quad = \quad \mathcal{L}_{\mathcal{C}} \quad (= \text{minmax value for (3) if } f \text{ is quadratic}).$$

(To get $f(x_0)$ from data on [v,y] ($v< x_0 <y$) change the coordinate using $x' = (x-x_o)/(y-v)$ and apply the method for $m = (v-x_o)/(y-v)$ to the data using weight functions $(y-v)\theta_j((y-v)x' + x_o)$.)

REFERENCES

[1] Ahlswede, R. personal communication concerning handwritten letters of Karl Gauss.

[2] Atkeson, C. G. , Moore, A.W. and Schaal, S. - Locally Weighted Learning, Artificial Intelligence Review, 11: 11-73, 1997.

[3] Barron,A.R. Universal approximation bounds for superpositions of a sigmoidal function. I.E.E.E. Trans. on Information Theory 40-2 (1993), 930-945

[4] ----------Complexity Regularization with Applications to Artificial Neural Networks, Nonparametric Functional Estimation and Related Topics (1991), Kluwer.

[5] Breiman, L. et al.. Classification and Regression Trees, 1986 Wadsworth

[6] Breiman, L. Random Forests, in Machine Learning 2001.

[7] ---------- Consistency for a Simplified Model of Random Forests, Tech. Rep. Statistics U.C.Berkeley, 2004.

[8] Candes, E.J. (1998) Harmonic analysis of neural networks, Applied and Computational Harmonic Analysis.

[9] Cleveland, W.S. and Loader, C.- Computational Methods for Local Regression. Tech. Rep.11, ATT Bell Labs Stat. Dept.(1994).

[10] Cybenko, G.- Approximation by Superpositions of a Sigmoidal Function (1989), Math. Control,Systems and Signals,2, 304-314..

[11] Darken C.,Donahue M., Gurvits L.,Sontag E.: Rate of Approximation Results Motivated by Robust Neural Network Learning. COLT 1993: 303-309

[12] DeVore,R.A., and Temlyakov,V.N. (1995) Some remarks on greedy algorithms. Adv. in Comput. Math. 5 173-187.

.

[13] Donoho, D. L. - Nonlinear Wavelet Methods for Recovery of Signals, Densities, and Spectra from Indirect and Noisy Data, Different Perspectives on Wavelets, Proc.of Symp.in Appl. Math. v.27, Amer. Math. Soc. 1993, 173-205

[14] Freund, Y. and Schapire, R. Experiments with a New Boosting Algorithm, Machine Learning: Proc. of the 13'th int. conf. 1996, 148-156.

[15] Friedman,J.H. - Greedy Function Approximation: A Gradient Boosting Machine, Stanford Univ. Tech. Rep. ,Feb. 1999.

[16] Friedman, J. H. and Stuetzle,W.(1981) Projection pursuit regression J. Amer. Statist. Assoc. 76, 817-823.

[17] Goldberger,A.S.- Best Linear Unbiased Prediction in the Generalized Linear Regression Model,J.Am. Stat. Assoc.57 (1962),369-375.

[18] Host, G. -Kriging by Local Polynomials,Comp. Statistics and Data Anal. 29 (1999) 295-312.

[19] Hull, J.C. Options, Futures, and Other Derivatives (2006) Prentice Hall 6'th.ed.

[20] Jones,L.K. -A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training, Annals of Statistics, Vol. 20, No. 1, 1992, pp. 608-613.

[21] ----------Constructive Approximations for Neural Networks by Sigmoidal Functions, Proceedings of I.E.E.E., October 1990,vol. 78, no.10., pp.1586-1589.

[22] ----------Good Weights and Hyperbolic Kernels for Neural Networks, Projection Pursuit, and Pattern Classification: Fourier Strategies for Extracting Information from High-Dimensional Data, I.E.E.E. Trans. on Info. Th., vol. 40, no. 2, March 1994, pp. 439-454.

[23] ---------- Local Greedy Approximation for Nonlinear Regression and Neural Network Training, Annals of Statistics, 2000, Vol.28, No.5, 1379-1389.

[24] ----------. On a conjecture of Huber concerning the convergence of projection pursuit regression.1987 , Ann. Statist. 15, 880-882.

[25] Jones, L.K. and Byrne, C 1990 - General Entropy Criteria for Inverse Problems,with Applications to Data Compression,Pattern Classification and Cluster Analysis, IEEE Trans. Info. Th.-36- p.23-30.

[26] Jones, L.K. and O'Neil, P.J. -Markov chain Monte Carlo algorithms for computing conditional expectations based on sufficient statistics, JCGS, vol.11, no. 3 Sept.2002, 660-677.

[27] Jorgensen, B.-The Theory of Linear Models, Chapman & Hall, New York, 1993.

[28] Juditsky,A. and Nemirovski,A.--Functional Aggregation for Nonparametric Regression, Ann. of Stat.(2000), 28, n. 3, 681-712.

[29] Lee,W.S.,Bartlett,P.L., and Williamson,R.C. (1996) Efficient agnostic learning of neural networks with bounded fan-in. I.E.E.E Trans. on Info.Theory 42-6 , 2118-2132

[30] Makovoz Y., Random approximants and neural networks , 1996, J. Approx.Th..98-109

[31] Mallat,S. and Zhang,Z- Matching Pursuits with Time Frequency Dictionaries. IEEE Trans. on Signal Processing 41 1993, 3397-3415.

[32] Mangasarian O.L. and D.R.Musicant -Successive overrelaxation for support vector machines, Math. Prog. Tech. Rep. 98-14 , C.S. Dept. U. Wisc.Madison,1998.

[33] Mueller K. et al . An introduction to kernel based learning algorithms, IEEE Trans. on Neural Networks, vol.12, no.2, Mar. 2001,181-202.

[34] Mukherjee S. et al. Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures, Proc. Natl. Acad. of S., Dec. 2001

[35] Pinkus,A., Approximation theory of the MLP model in neural networks, Acta Numerica 8 (1999), 143-196.

[36] Poggio,T. and Girosi, F.-Networks for Approximation and Learning, Proc. IEEE,78(9), 1990, 1484-1487.

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

[37a] Kimeldorf,G. and Wahba, G. Some results on Tchebycheffian spline functions, J. Math. Anal.and Appl.,33,1(1971)82-95.

[38] Rejto, L and Walter, G.G. (1992). Remarks on projection pursuit regression and density estimation. Stochastic Anal. Appl. 10, 213-222.

[39] Rifkin, R.M. Everything Old is New Again: A Fresh Look at Historical Approaches to Machine Learning, Ph.D. thesis, M.I.T., 2002.

[40] Sacks, J. and Ylvisaker,D. (1978) Linear Estimation for Approximately Linear Models. Annals of Statistics, vol.6,no.5, 1122-1137.

[41] Stone, M and Brooks,R.J. (1990) Continuum Regression:cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principle components regression (with discussion), Jour. Royal Statist. Soc., Series B 52, 237-269.

[42] Sundberg, R. (2002) Shrinkage Regression, Encycl. of Environ., Vol. 4, 1994-1998, Wiley.

[43] Vapnik,V.N. -The Nature of Statistical Learning, Springer, 1995.

[44] Vapnik,V. and Bottou,L. (1993) Local algorithms for pattern recognition and dependencies estimation, Neural Computation 5(6): 893-909.