Advances in Computational Mathematics (2006) 25: 237–258

Linearly constrained reconstruction of functions by kernels with applications to machine learning

R. Schaback and J. Werner

Göttingen, Germany

Received 25 March 2003; accepted 11 October 2003
Communicated by Joe Ward

Dedicated to C.A. Micchelli at the occasion of his 60th birthday

This paper investigates the approximation of multivariate functions from data via linear combinations of translates of a positive definite kernel from a reproducing kernel Hilbert space. If standard interpolation conditions are relaxed by Chebyshev-type constraints, one can minimize the norm of the approximant in the Hilbert space under these constraints. By standard arguments of optimization theory, the solutions will take a simple form, based on the data related to the active constraints, called support vectors in the context of machine learning. The corresponding quadratic programming problems are investigated to some extent. Using monotonicity results concerning the Hilbert space norm, iterative techniques based on small quadratic subproblems on active sets are shown to be finite, even if they drop part of their previous information and even if they are used for infinite data, e.g., in the context of online learning. Numerical experiments confirm the theoretical results.

Keywords: positive definite radial basis functions, quadratic programming, kernel machines, machine learning, support vector machines, regression.

Mathematics subject classifications (2000): 65D05, 65D10, 41A15, 41A17, 41A27, 41A30, 41A40, 41A63.

1. Introduction

This paper considers exact or approximate reconstruction of data $f_1, \dots, f_N$ on a large multivariate point set $X = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ by functions s that guarantee a uniform error bound

$$-\eta \leqslant s(x_j) - f_j \leqslant \eta, \quad 1 \leqslant j \leqslant N. \tag{1}$$

The functions s should be as simple as possible. However, asking for a small error bound η will blow up the complexity of the solution s. In fact, one might even get exact interpolation with $\eta = 0$, but at the expense of a rather complicated s. Conversely, there is the least complicated solution $s = 0$ with a large error $\eta = \|f\|_\infty$. This tradeoff between the complexity of the reconstructing function s and the error bound η is called the "*bias-variance-dilemma*" in regression theory. Without referring to probabilistic arguments at all, it will be the main concern of this paper, together with providing efficient techniques to find reasonably good solutions. It is a standard technique to pick out "good" solutions s of (1) with special properties by adding a penalty or objective function and solving the corresponding optimization problem. The additional objective function is interpreted as a risk or loss function, depending on the statistical context. But the latter is considered here as a background information supporting a particular choice of optimization problem, and it will be completely ignored in this paper.

In the context of machine learning [15, 17], the function s represents an unknown regression map from a vector input to a scalar output, and the pairs (x_j, f_j) are called "training data". The case of classification learning can be written in a similar form without significant loss or increase of complexity [12].

2. Kernels

Minimizing η under all s from a fixed finite-dimensional space $\mathcal{S}$ is just a standard linear Chebyshev approximation problem, solvable by linear programming, preferably via the revised technique applied to the dual problem [5]. But one would need to vary $\mathcal{S}$ in order to find the best solution with fixed complexity, i.e. to find the minimal value of η when s comes from any space $\mathcal{S}$ with $k = \dim \mathcal{S}$ fixed, but $\mathcal{S}$ itself allowed to vary. This is a multivariate generalization of Chebyshev approximation by splines with free knots [1]. One could also start from a fixed error level η and try to find a space $\mathcal{S}$ of smallest dimension such that there is some $s \in \mathcal{S}$ satisfying (1). In all of these cases, the problem turns nonlinear, if $\mathcal{S}$ is allowed to vary.

If some reasonable penalty function based on s and η is minimized over the feasible set described by (1), the corresponding Kuhn–Tucker conditions will introduce *active point sets*

$$Y_\pm := \{y \in X\colon\ s(y) - f(y) = \pm\eta\}, \tag{2}$$

where, for simplicity, f stands for the function with $f_j = f(x_j)$ for all $x_j \in X$. These sets will naturally be very important in any numerical technique that calculates a good solution to (1), because they pick "critical" or "support" points from X for which there is no leeway for arbitrary perturbations. They are selected after some possibly complicated optimization process, but they may be of considerably smaller size than X itself.

Thus it is a reasonable idea to link point sets like $Y_\pm$ to function spaces $\mathcal{S}$ for s, and this can be done by the "kernel trick" of learning theory that maps points y to functions $K(\cdot, y)$ in "feature space" via some kernel function K. In particular, one can define

$$\mathcal{K}_Y := \operatorname{span}\{K(\cdot, y)\colon\ y \in Y\} \tag{3}$$

for finite subsets $Y \subset \mathbb{R}^d$, provided that K maps $\mathbb{R}^d \times \mathbb{R}^d$ into $\mathbb{R}$. This is very useful, because it maps sets into spaces.

But there is still another viewpoint, especially for those readers who consider the kernel trick as something new or special. If a general quadratic penalty on s is minimized that can be written as (s, s) via some inner product in some function space S, and if point evaluation is continous with respect to this inner product, then one can form the Hilbert space closure of S with a reproducing kernel K. Thus the "kernel trick" situation turns out to be necessary whenever an inner product penalty with continuous point evaluation is minimized. Consequently, there is quite some history [16] of kernel techniques long before learning theory became popular.

3. Optimization problems

Motivated by

- the above discussion,
- interpolation by radial basis functions or conditionally positive kernels and
- regression problems solved by support vector machines in the context of learning theory,

this paper focuses on functions s from reproducing kernel Hilbert spaces on $\mathbb{R}^d$. These are generated by a symmetric positive definite kernel $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ such that an inner product $(\cdot, \cdot)_K$ on functions of the form $K(\cdot, x)$ is well-defined and satisfies

$$\big(K(\cdot, x), K(\cdot, y)\big)_K = K(x, y) \tag{4}$$

for all $x, y \in \mathbb{R}^d$. The closure of the space of linear combinations of such functions under the inner product forms a Hilbert space $\mathcal{K}$ with the reproduction property

$$s(x) = \big(s, K(\cdot, x)\big)_K \quad \text{for all } s \in \mathcal{K},\ x \in \mathbb{R}^d. \tag{5}$$

In the context of radial basis function techniques, this space was called the "*native*" space for K. The more general case of "*conditionally*" positive definite kernels (see Micchelli [6]) is omitted here for simplicity. There is a standard trick [11] to transform conditionally positive definite kernels to strictly positive definite ones.

The goal is to find *simple* functions s from $\mathcal{K}$ satisfying (1), and their complexity will be controlled by picking them from finite-dimensional subspaces of the form (3) for finite subsets $Y \subset \mathbb{R}^d$. Since it is well-known that exact interpolation of the data by a unique function s_X^* from the space $\mathcal{K}_X$ is always possible, inequalities (1) always have admissible functions for any choice of $\eta \geqslant 0$ with a complexity related to $N = |X|$ in the worst case. However, one should preferably find functions s that satisfy (1) for small positive η and come from subspaces $\mathcal{K}_Y$ with sets Y that are considerably smaller than X.

Using the inner product (4) one can ask for a function $s \in \mathcal{K}$ that satisfies (1) for fixed η and has minimal norm in $\mathcal{K}$. Or one can allow η to vary also, and then minimize the quadratic form

$$\frac{1}{2}(s, s)_K + C\eta \tag{6}$$

for some fixed positive constant C over all $s \in \mathcal{K}$ and $\eta \in \mathbb{R}$ under the constraints (1). Increasing C allows to put more emphasis on the error at the expense of complexity. This quadratic optimization problem with linear constraints will be called the *unrestricted* problem in this paper, while the *restricted* problem fixes η.

Note that for $C = \eta = 0$ we have an optimal recovery problem [7–10] with exact data, and this paper generalizes optimal recovery in the sense of Micchelli et al. for uncertain data with strict bounds of the form (1).

In practice, both problems will be hard to handle for large N because of their $\mathcal{O}(N)$ variables and $2N$ restrictions. The situation is not improving if we go over to the dual problem. Furthermore, it is not clear at the beginning how the trade-off between error and complexity turns out for a specific data set. Thus it is necessary to go somewhat deeper into the geometry of the feasible set and focus on efficient iterative solution techniques.

4. Restricted problem

Theorem 1. Let $\eta > 0$ be given. The finite-dimensional restricted problem on X is given by

$$\begin{cases} \text{minimize} \quad F(s) := \frac{1}{2}(s, s)_K \quad \text{subject to} \\ s \in \mathcal{K}_X, \quad -\eta \leqslant s(x) - f(x) \leqslant \eta \quad (x \in X). \end{cases} \tag*{(P)$_{fr}$}$$

It has a unique solution $s^*(\cdot) = \sum_{x \in X} \alpha_x^* K(\cdot, x)$ where the unique coefficient vector $\alpha_X^* = (\alpha_x^*)_{x \in X}$ is characterized by the existence of nonnegative vectors λ_X^* and μ_X^* such that $\alpha_X^* = \mu_X^* - \lambda_X^*$ and

$$\left.\begin{aligned} \lambda_x^*\bigl(s^*(x) - f(x) - \eta\bigr) &= 0, \\ \mu_x^*\bigl(-s^*(x) + f(x) - \eta\bigr) &= 0, \end{aligned}\right. \quad \text{for all } x \in X.$$

There exist (possibly empty) sets $Y_+, Y_- \subset X$ such that $\alpha_x^* < 0$ for all $x \in Y_+$, $\alpha_x^* > 0$ for all $x \in Y_-$,

$$s^*(x) - f(x) = \eta \quad (x \in Y_+), \qquad s\bigl(x^*\bigr) - f(x) = -\eta \quad (x \in Y_-),$$

and $\alpha_x^* = 0$ for all $x \in X \setminus (Y_+ \cup Y_-)$.

Proof. The problem is equivalent to a feasible quadratic program with a strictly convex objective function. Thus existence and uniqueness of a solution are evident. An application of the Kuhn–Tucker theorem proves the characterization of a solution. Now define

$$Y_+ := \{x \in X\colon \lambda_x^* > 0,\ \mu_x^* = 0\}, \qquad Y_- := \{x \in X\colon \lambda_x^* = 0,\ \mu_x^* > 0\}.$$

Obviously Y_+ and Y_- have the desired properties. Here we used the fact that because of $\eta > 0$ no $x \in X$ exists such that $\lambda_x^* > 0$ and $\mu_x^* > 0$. □

Theorem 2. Let $\eta > 0$ be given. The infinite dimensional restricted problem on X is given by

$$\begin{cases} \text{minimize} \quad F(s) := \frac{1}{2}(s,s)_K \quad \text{subject to} \\ s \in \mathcal{K}, \quad -\eta \leqslant s(x) - f(x) \leqslant \eta \quad (x \in X). \end{cases} \qquad (\mathrm{P})_{ir}$$

It has a unique solution which is given by the unique solution s^* of the corresponding finite-dimensional restricted problem on X.

Proof. We show that the infinite problem $(\mathrm{P})_{ir}$ is solved by the solution s^* of the finite-dimensional problem. Then uniqueness will follow from strict convexity of the objective function. Since $(\mathrm{P})_{ir}$ is a convex program we have to show that $F'(s^*)(s - s^*) \geqslant 0$ for any $s \in \mathcal{K}$ that is feasible for $(\mathrm{P})_{ir}$. With Y_+ and Y_- as in theorem 1 the reproduction formula gives

$$\begin{aligned} F'(s^*)(s - s^*) &= (s^*, s - s^*)_K \\ &= \sum_{x\in X} \alpha_x^* \big(K(\cdot, x), s - s^*\big) \\ &= \sum_{x\in X} \alpha_x^* \big(s(x) - s(x^*)\big) \\ &= \sum_{x\in Y_+\cup Y_-} \alpha_x^* \big(s(x) - s^*(x)\big) \\ &= \sum_{x\in Y_+} \underbrace{\alpha_x^*}_{<0} \underbrace{\big(s(x) - f(x) - \eta\big)}_{\leqslant 0} + \sum_{x\in Y_-} \underbrace{\alpha_x^*}_{>0} \underbrace{\big(s(x) - f(x) + \eta\big)}_{\geqslant 0} \\ &\geqslant 0 \end{aligned}$$

proving the desired result. □

5. Unrestricted problem

Theorem 3. Let $C > 0$ be given. The finite-dimensional unrestricted problem

$$\begin{cases} \text{minimize} \quad F(s,\eta) := \frac{1}{2}(s,s)_K + C\eta \quad \text{subject to} \\ (s,\eta) \in \mathcal{K}_X \times \mathbb{R}, \quad -\eta \leqslant s(x) - f(x) \leqslant \eta \quad (x \in X) \end{cases} \qquad (\mathrm{P})_{fu}$$

has a unique solution (s^*, η^*). This solution can be uniquely represented as

$$s^*(\cdot) = \sum_{x\in X} \alpha_x^* K(\cdot, x), \qquad \eta^* = \frac{1}{C}\sum_{x\in X} \alpha_x^*\bigl(f(x) - s^*(x)\bigr).$$

The unique coefficient vector $\alpha_X^* = (\alpha_x^*)_{x\in X}$ is characterized by the existence of nonnegative vectors λ_X^* and μ_X^* such that $\alpha_X^* = \mu_X^* - \lambda_X^*$ and

$$\begin{aligned} \lambda_x^*\bigl(s^*(x) - f(x) - \eta^*\bigr) &= 0,\\ \mu_x^*\bigl(-s^*(x) + f(x) - \eta^*\bigr) &= 0, \end{aligned} \quad \text{for all } x \in X.$$

If $\eta^* > 0$ there exist (possibly empty) sets $Y_+, Y_- \subset X$ such that $\alpha_x^* < 0$ for all $x \in Y_+$, $\alpha_x^* > 0$ for all $x \in Y_-$, $\alpha_x^* = 0$ for all $x \in X \setminus (Y_+ \cup Y_-)$,

$$s^*(x) - f(x) = \eta^* \quad (x \in Y_+), \qquad s^*(x) - f(x) = -\eta^* \quad (x \in Y_-),$$

and

$$\sum_{x\in Y_-} \alpha_x^* - \sum_{x\in Y_+} \alpha_x^* = C.$$

Proof. Problem $(\mathrm{P})_{fu}$ is equivalent to the quadratic program

$$\begin{cases} \text{minimize} \quad \frac{1}{2}\alpha_X^{\mathrm{T}} K_{X,X}\alpha_X + C\eta \quad \text{subject to} \\ (\alpha_X, \eta) \in \mathbb{R}^{|X|} \times \mathbb{R}, \quad -\eta 1_X \leqslant K_{K,X}\alpha_X - f_X \leqslant \eta 1_X. \end{cases}$$

This is a feasible quadratic program whose objective function is bounded from below (by 0) on the set of feasible solutions. The existence of a solution (α_X^*, η^*) respectively a solution (s^*, η^*) is implied by a well known theorem of Frank and Wolfe [4, 18]. Uniqueness of a solution follows by a standard convexity argument. By the Kuhn–Tucker theorem a solution (α_X^*, η^*) is characterized by the existence of nonnegative multipliers λ_X^*, μ_X^* such that

$$K_{X,X}\alpha_X^* + K_{X,X}\bigl(\lambda_X^* - \mu_X^*\bigr) = 0$$

respectively $\alpha_X^* = \mu_X^* - \lambda_X^*$,

$$\bigl(\lambda_X^* + \mu_X^*\bigr)^{\mathrm{T}} 1_X = C$$

and

$$\bigl(\lambda_X^*\bigr)^{\mathrm{T}}\bigl(K_{X,X}\alpha_X^* - f_X - \eta^* 1_X\bigr) = 0, \qquad \bigl(\mu_X^*\bigr)^{\mathrm{T}}\bigl(-K_{X,X}\alpha_X^* + f_X - \eta^* 1_X\bigr) = 0.$$

Adding the last two equations leads to

$$-\underbrace{\bigl(\mu_X^* - \lambda_X^*\bigr)}_{=\alpha_X^*}{}^{\mathrm{T}} K_{X,X}\alpha_X^* + f_X^{\mathrm{T}}\underbrace{\bigl(\mu_X^* - \lambda_X^*\bigr)}_{=\alpha_X^*} - \eta^* \underbrace{\bigl(\lambda_X^* + \mu_X^*\bigr)^{\mathrm{T}} 1_X}_{=C} = 0.$$

Thus

$$\eta^* = \frac{1}{C}\big(\alpha_X^*\big)^{\mathrm{T}}\big(f_X - K_{X,X}\alpha_X^*\big) = \frac{1}{C}\sum_{x\in X}\alpha_x^*\big(f(x) - s^*(x)\big).$$

This gives the characterization part of the theorem. Furthermore Y_+ and Y_- can be defined just as in the proof of theorem 1 since we assumed that $\eta^* > 0$, i.e. by

$$Y_+ := \big\{x \in X\colon\ \lambda_x^* > 0,\ \mu_x^* = 0\big\}, \qquad Y_- := \big\{x \in X\colon\ \lambda_x^* = 0,\ \mu_x^* > 0\big\}.$$

The theorem has been proven. □

In the last part of theorem 3 we had to exclude the case $\eta^* = 0$. But one can exactly characterize this situation.

Theorem 4. The second component η^* of a solution (s^*, η^*) to $(\mathrm{P})_{fu}$ is zero if and only if $C \geqslant \|K_{X,X}^{-1}f_X\|_1$.

Proof. Because of the constraints in $(\mathrm{P})_{fu}$ the case $\eta^* = 0$ can only occur if s^* interpolates f on X, i.e. $s^*(\cdot) = \sum_{x\in X}\alpha_x^* K(\cdot, x)$ where $\alpha_X^* = K_{X,X}^{-1}f_X$. The arguments in the proof of the characterization in theorem 3 show that $(K_{X,X}^{-1}f_X, 0)$ solves the quadratic program equivalent to $(\mathrm{P})_{fu}$ if and only if there exist nonnegative vectors $\lambda_X^*, \mu_X^* \in \mathbb{R}^{|X|}$ such that

$$K_{X,X}^{-1}f_X = \mu_X^* - \lambda_X^*, \qquad \big(\lambda_X^* + \mu_X^*\big)^{\mathrm{T}}1_X = C.$$

Thus the solution of $(\mathrm{P})_{fu}$ has a positive second component if and only if the system

$$\begin{pmatrix} -I & I \\ 1_X^{\mathrm{T}} & 1_X^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} \lambda_X \\ \mu_X \end{pmatrix} = \begin{pmatrix} K_{X,X}^{-1}f_X \\ C \end{pmatrix}, \qquad \begin{pmatrix} \lambda_X \\ \mu_X \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

is not solvable. By the Farkas lemma [3, 18] this is equivalent to the existence of $(z, \delta) \in \mathbb{R}^{|X|} \times \mathbb{R}$ such that

$$\begin{pmatrix} -I & 1_X \\ I & 1_X \end{pmatrix}\begin{pmatrix} z \\ \delta \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} K_{X,X}^{-1}f_X \\ C \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} z \\ \delta \end{pmatrix} < 0$$

respectively

$$-\delta 1_X \leqslant z \leqslant \delta 1_X, \qquad \big(K_{X,X}^{-1}f_X\big)^{\mathrm{T}}z + \delta C < 0. \tag{7}$$

The following lemma will finish the proof of the theorem. □

Lemma 1. Inequalities (7) have a solution (z, δ) iff $C < \|K_{X,X}^{-1}f_X\|_1$.

Proof. Let us first assume that $C < \|K_{X,X}^{-1}f_X\|_1$. Take an arbitrary $\delta > 0$ and define $z := -\delta\,\mathrm{sign}(K_{X,X}^{-1}f_X)$ where the sign function acts componentwise on a vector. Obviously (z, δ) solves (7). On the other hand, suppose that (z, δ) solves (7). Necessarily $\delta > 0$. Since

$$\delta\big(C - \|K_{XX}^{-1} f_X\|_1\big) = \min_{-\delta 1_X \leqslant y \leqslant \delta 1_X} \big(K_{X,X}^{-1} f_X\big)^{\mathrm{T}} y + \delta C \leqslant \big(K_{X,X}^{-1} f_X\big)^{\mathrm{T}} z + \delta C < 0$$

we get the reverse direction. □

We now consider the infinite-dimensional unrestricted problem.

Theorem 5. Let $C > 0$ be given and (s^*, η^*) be the unique solution of the finite-dimensional unrestricted problem $(\mathrm{P})_{fu}$ formulated in theorem 3. We assume η^* to be positive respectively C to be sufficiently small. Then (s^*, η^*) is the unique solution of the infinite-dimensional unrestricted problem

$$\begin{cases} \text{minimize} \quad F(s, \eta) := \frac{1}{2}(s, s)_K + C\eta \quad \text{subject to} \\ (s, \eta) \in \mathcal{K} \times \mathbb{R}, \quad -\eta \leqslant s(x) - f(x) \leqslant \eta \quad (x \in X). \end{cases} \qquad (\mathrm{P})_{iu}$$

Proof. At the beginning we show that (s^*, η^*) solves $(\mathrm{P})_{iu}$ by showing that

$$F'\big(s^*, \eta^*\big)\big[(s, \eta) - \big(s^*, \eta^*\big)\big] \geqslant 0$$

for any pair $(s, \eta) \in \mathcal{K} \times \mathbb{R}$ feasible for $(\mathrm{P})_{iu}$. Since η^* is supposed to be positive there exist subsets $Y_+, Y_- \subset X$ such that

$$s^*(\cdot) = \sum_{x \in X} \alpha_x^* K(\cdot, x)$$

with $\alpha_x^* < 0$ for $x \in Y_+$, $\alpha_x^* > 0$ for $x \in Y_-$, $\alpha_x^* = 0$ for all $x \in X \setminus (Y_+ \cup Y_-)$,

$$s^*(x) - f(x) = \eta^* \quad (x \in Y_+), \qquad s^*(x) - f(x) = -\eta^* \quad (x \in Y_-),$$

and

$$\sum_{x \in Y_-} \alpha_x^* - \sum_{x \in Y_+} \alpha_x^* = C.$$

The reproduction formula implies

$$\begin{aligned} F'\big(s^*, \eta^*\big)\big[(s, \eta) - \big(s^*, \eta^*\big)\big] &= \big(s^*, s - s^*\big)_K + C\big(\eta - \eta^*\big) \\ &= \sum_{x \in Y_+ \cup Y_-} \alpha_x^* \big(K(\cdot, x), s - s^*\big) + C\big(\eta - \eta^*\big) \\ &= \sum_{x \in Y_+ \cup Y_-} \alpha_x^* \big(s(x) - s^*(x)\big) + C\big(\eta - \eta^*\big) \\ &= \sum_{x \in Y_+} \alpha_x^* \big(s(x) - f(x) - \eta^*\big) + \sum_{x \in Y_-} \alpha_x^* \big(s(x) - f(x) + \eta^*\big) \\ &\quad + C\big(\eta - \eta^*\big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x\in Y_+} \underbrace{\alpha_x^*}_{<0} \underbrace{\big(s(x) - f(x) - \eta\big)}_{\leqslant 0} \\
&\quad + \sum_{x\in Y_-} \underbrace{\alpha_x^*}_{>0} \underbrace{\big(s(x) - f(x) + \eta\big)}_{\geqslant 0} \\
&\quad + \underbrace{\Bigg[\bigg(\sum_{x\in Y_+} \alpha_x^* - \sum_{x\in X_-} \alpha_x^*\bigg) + C\Bigg]}_{=0}(\eta - \eta^*) \\
&\geqslant 0
\end{aligned}
$$

which gives the desired result. The uniqueness again follows from a standard convexity argument. □

6. Monotonicity results

Before iterative techniques are considered, it will pay off to look at algorithms that just add a single point to the active set. The following monotonicity result will turn out to be very useful.

Lemma 2. Let s_Z be the unique solution to the restricted problem for given $\eta > 0$ on a subset $Z \subset X$.

(1) If $|s_Z(x) - f(x)| \leqslant \eta$ for all $x \in X$ then s_Z solves the restricted problem on X.

(2) Let $Y := \{x \in Z\colon\ |s_Z(x) - f(x)| = \eta\}$ be the active set for s_Z. Assume $|s_Z(x) - f(x)| > \eta$ for some $x \in X \setminus Z$, and let $s_{Y\cup\{x\}}$ be the unique solution of the restricted problem with the same η on $Y \cup \{x\}$. Then

$$\|s_{Y\cup\{x\}}\|_K > \|s_Z\|_K.$$

Proof. First we show that s_Z solves

$$\text{minimize} \quad F(s) := \frac{1}{2}(s,s)_K \quad \text{subject to} \quad s \in \mathcal{K}_X, \quad \|s - f\|_{\infty,Y} \leqslant \eta \qquad (\mathrm{P})_Y$$

for *any* subset $Y \subset X$ such that $\|s_Z - f\|_{\infty,Y} \leqslant \eta$. This will prove the first part of the theorem. Let $Y_+, Y_- \subset Z$ be two sets as in theorem 1, i.e. $s_Z(\cdot) = \sum_{x\in Z} \alpha_x K(\cdot, x)$ with $\alpha_x < 0$ for all $x \in Y_+$, $\alpha_x > 0$ for all $x \in Y_-$,

$$s_Z(x) - f(x) = \eta \quad (x \in Y_+), \qquad s_Z(x) - f(x) = -\eta \quad (x \in Y_-)$$

and $\alpha_x = 0$ for all $x \in Z \setminus (Y_+ \cup Y_-)$. For any $s \in \mathcal{K}_X$ with $\|s - f\|_{\infty,Y} \leqslant \eta$ we then conclude that

$$F'(s_Z)(s - s_Z) = (s_Z, s - s_Z)_K$$

$$
\begin{aligned}
&= \sum_{x\in Y_+\cup Y_-} \alpha_x \big(K(\cdot,x), s - s_Z\big)\\
&= \sum_{x\in Y_+} \underbrace{\alpha_x}_{<0} \underbrace{\big(s(x) - f(x) - \eta\big)}_{\leqslant 0} + \sum_{x\in Y_-} \underbrace{\alpha_x}_{>0} \underbrace{\big(s(x) - f(x) + \eta\big)}_{\geqslant 0}\\
&\geqslant 0.
\end{aligned}
$$

This shows that s_Z solves the problem $(\mathrm{P})_Y$. Now let Y be the set of active constraints for s_Z defined in the second part of the lemma. From the first part we have

$$
\begin{aligned}
\|s_Z\|_K &= \inf_{\|s-f\|_{\infty,Z}\leqslant\eta} \|s\|_K = \inf_{\|s-f\|_{\infty,Y}\leqslant\eta} \|s\|_K\\
&\leqslant \inf_{\|s-f\|_{\infty,Y\cup\{x\}}} \|s\|_K = \|s_{Y\cup\{x\}}\|_K.
\end{aligned}
$$

If $\|s_Z\|_K = \|s_{Y\cup\{x\}}\|_K$ then s_Z and $s_{Y\cup\{x\}}$ are both solutions to $(\mathrm{P})_Y$, so $s_Z = s_{Y\cup\{x\}}$ due to uniqueness in theorem 1. This contradicts the fact that $|s_Z(x) - f(x)| > \eta$ and $|s_{Y\cup\{x\}}(x) - f(x)| \leqslant \eta$. □

The above result is qualitative, but it can be quantified using an idea from [2] in the interpolation case.

Theorem 6. (1) Let $Y \subset X$ and $x \in X \setminus Y$. If $s_{Y\cup\{x\}} \in \mathcal{K}_{Y\cup\{x\}}$ and $\hat{s} \in \mathcal{K}_Y$ with $s_{Y\cup\{x\}} = \hat{s}(y)$ for all $y \in Y$, then

$$
\|s_{Y\cup\{x\}}\|_K^2 = \|\hat{s}\|_K^2 + \frac{(s_{Y\cup\{x\}} - \hat{s}(x))^2}{K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}}. \tag{8}
$$

(2) Let s_Z be the unique solution to the restricted problem for given $\eta > 0$ and a subset $Z \subset X$, $Y := \{x \in Z\colon |s_Z(x) - f(x)| = \eta\}$ be the active set for s_Z. Assume $|s_Z(x) - f(x)| > \eta$ for some $x \in X \setminus Z$, and let $s_{Y\cup\{x\}}$ be the unique solution of the restricted problem with the same η on $Y \cup \{x\}$. If $\hat{s} \in \mathcal{K}_Y$ is any function coinciding with $s_{Y\cup\{x\}}$ on Y, we have

$$
\|s_{Y\cup\{x\}}\|_K^2 \geqslant \|s_Z\|_K^2 + \frac{(s_{Y\cup\{x\}} - \hat{s}(x))^2}{K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}}. \tag{9}
$$

The denominator on the right-hand side is positive.

Proof. We have

$$
s_{Y\cup\{x\}}(\cdot) = \sum_{y\in Y} \alpha_y K(\cdot,y) + \alpha_x K(\cdot,x), \qquad \hat{s}(\cdot) = \sum_{y\in Y} \hat{\alpha}_y K(\cdot,y).
$$

Since $s_{Y\cup\{x\}}$ and $\hat{s}$ coincide on Y, we have

$$
K_{Y,Y}\hat{\alpha}_Y = K_{Y,Y}\alpha_Y + \alpha_x K_{Y,\{x\}}
$$

respectively

$$\alpha_Y = \hat{\alpha}_Y - \alpha_x K_{Y,Y}^{-1} K_{Y,\{x\}}.$$

Thus

$$\begin{aligned}\|s_{Y\cup\{x\}}\|_K^2 &= \alpha_Y^{\mathrm{T}} K_{Y,Y}\alpha_Y + 2\alpha_x\alpha_Y^{\mathrm{T}} K_{Y,\{x\}} + \alpha_x^2 K(x,x)\\ &= \hat{\alpha}_Y^{\mathrm{T}} K_{Y,Y}\hat{\alpha}_Y + \alpha_x^2\big[K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}\big]\\ &= \|\hat{s}\|_K^2 + \alpha_x^2\big[K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}\big].\end{aligned}$$

On the other hand

$$\begin{aligned}s_{Y\cup\{x\}}(x) - \hat{s}(x) &= \alpha_Y^{\mathrm{T}} K_{Y,\{x\}} + \alpha_x K(x,x) - \hat{\alpha}_Y^{\mathrm{T}} K_{Y,\{x\}}\\ &= \alpha_x\big[K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}\big].\end{aligned}$$

This leads to

$$\|s_{Y\cup\{x\}}\|_K^2 = \|\hat{s}\|_K^2 + \frac{(s_{Y\cup\{x\}} - \hat{s}(x))^2}{K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}}.$$

Since

$$K_{Y\cup\{x\},Y\cup\{x\}} = \begin{pmatrix} K_{Y,Y} & K_{Y,\{x\}}\\ K_{Y,\{x\}}^{\mathrm{T}} & K(x,x)\end{pmatrix}$$

is positive definite and

$$\begin{aligned}0 &< \begin{pmatrix} K_{Y,Y}^{-1} K_{Y,\{x\}}\\ -1\end{pmatrix}^{\mathrm{T}} \begin{pmatrix} K_{Y,Y} & K_{Y,\{x\}}\\ K_{Y,\{x\}}^{\mathrm{T}} & K(x,x)\end{pmatrix} \begin{pmatrix} K_{Y,Y}^{-1} K_{Y,\{x\}}\\ -1\end{pmatrix}\\ &= K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}\end{aligned}$$

the denominator above is positive.

In the second part of the theorem $\hat{s}$ is feasible for the restricted problem on Y with the solution s_Z. The theorem has been proven. □

Readers familiar with the theory of interpolation by positive definite kernels will note that the denominators in (8) and (9) coincide with the square of the power function for interpolation by K on Y evaluated at x.

Theorem 6 shows that optimal progress in the Hilbert space norm is made when the additional fractions in (8) and (9) are maximized. Current greedy techniques [13] have only considered the numerator of these fractions. Thus it may improve greedy techniques to use the full fractions, provided that the overall behaviour can be shown to improve when the Hilbert space norm of the approximant is boosted up. This turns out to be true. More precisely, if a function $s \in \mathcal{K}_Y$ is constructed that satisfies (1) on Y with a small η and has $\|s\|_K$ close to $\|f\|_K$, then s is close to f everywhere. This can be quantified by the following result, though we have no idea about the size of the occurring constants.

Lemma 3. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, $f \in \mathcal{K} \setminus \{0\}$, $Y \subset \Omega$ finite, $s \in \mathcal{K}_Y$ with $\|s - f\|_{\infty,Y} \leqslant \eta$ for some $\eta > 0$ and $\|s\|_K \geqslant \|f\|_K - \epsilon$ for some $\epsilon > 0$. Then

$$\|f - s\|_K^2 \leqslant 2\|f\|_K\bigl(\epsilon + C_K(Y)\eta\bigr)$$

and

$$\|f - s\|_{\infty,\Omega} \leqslant \|P_Y\|_{\infty,\Omega}\sqrt{2\|f\|_K\bigl(\epsilon + C_K(Y)\eta\bigr)} + C_\infty(Y)\eta$$

for all sufficiently small positive ϵ, η. Here, the constants $C_K(Y)$ and $C_\infty(Y)$ depend only on Ω, K and Y, but not on f, s, ϵ or η. Furthermore, the symbol P_Y stands for the power function of standard kernel interpolation on Y.

Proof. Let $s' \in \mathcal{K}_Y$ be the exact interpolant to f on Y. Since both s and s' are based on Y and

$$\|s - s'\|_{\infty,Y} = \|s - f\|_{\infty,Y} \leqslant \eta,$$

there are constants $C_K(Y)$ and $C_\infty(Y)$ depending only on Ω, K and Y such that

$$\begin{aligned}\|s - s'\|_K &\leqslant C_K(Y)\|s - s'\|_{\infty,Y} \leqslant C_K(Y)\eta,\\ \|s - s'\|_{\infty,\Omega} &\leqslant C_\infty(Y)\|s - s'\|_{\infty,Y} \leqslant C_\infty(Y)\eta,\end{aligned}$$

due to standard norm-equivalence arguments. We thus have

$$\|s'\|_K \geqslant \|s\|_K - \|s - s'\|_K \geqslant \|f\|_K - \bigl(\epsilon + C_K(Y)\eta\bigr).$$

Since any $t \in \mathcal{K}_Y$ can be represented as $t(\cdot) = \sum_{y\in Y} \beta_y K(\cdot, y)$ the reproduction formula leads to the standard orthogonality result

$$\bigl(f - s', t\bigr)_K = \sum_{y\in Y} \beta_y\bigl(f - s', K(\cdot, y)\bigr) = \sum_{y\in Y} \alpha_y\bigl[\underbrace{f(y) - s'(y)}_{=0}\bigr] = 0.$$

From this we get

$$\|f\|_K^2 = \bigl(f - s', f\bigr)_K + \bigl(f, s'\bigr)_K = \bigl(f - s', f - s'\bigr)_K + \bigl(f, s'\bigr)_K = \|f - s'\|_K^2 + \|s'\|_K^2,$$

the standard minimum norm property. Now let the positive numbers ϵ, η be so small, that

$$\epsilon + C_K(Y)\eta \leqslant \|f\|_K. \tag{10}$$

Then we have

$$\begin{aligned}\|f - s\|_K^2 &= \|f - s'\|_K^2 + \|s - s'\|_K^2\\ &= \|f\|_K^2 - \|s'\|_K^2 + \|s - s'\|_K^2\\ &\leqslant \|f\|_K^2 - \bigl[\|f\|_K - \bigl(\epsilon + C_K(Y)\eta\bigr)\bigr]^2 + C_K(Y)^2\eta^2\\ &= 2\|f\|_K\bigl(\epsilon + C_K(Y)\eta\bigr) - \bigl(\epsilon + C_K(Y)\eta\bigr)^2 + C_K(Y)^2\eta^2\end{aligned}$$

$$
\begin{aligned}
&= 2\|f\|_K\big(\epsilon + C_K(Y)\eta\big) - \epsilon^2 - 2C_K(Y)\epsilon\eta \\
&\leqslant 2\|f\|_K\big(\epsilon + C_K(Y)\eta\big).
\end{aligned}
$$

For the last part of the theorem we still assume ϵ, η to be so small that (10) holds. Then we get

$$
\begin{aligned}
\|f - s\|_{\infty,\Omega} &\leqslant \|f - s'\|_{\infty,\Omega} + \|s - s'\|_{\infty,\Omega} \\
&\leqslant \|P_Y\|_{\infty,\Omega}\|f - s'\|_K + C_\infty(Y)\eta \\
&\leqslant \|P_Y\|_{\infty,\Omega}\sqrt{2\|f\|_K\big(\epsilon + C_K(Y)\eta\big)} + C_\infty(Y)\eta,
\end{aligned}
$$

where we used the standard error bound

$$\|f - s'\|_{\infty,\Omega} \leqslant \|P_Y\|_{\infty,\Omega}\|f - s'\|_K$$

for kernel interpolation of all $f \in \mathcal{K}$ on Y. The factor P_Y is the power function with

$$P_Y^2(x) = K(x,x) - K_{Y,\{x\}}^{\mathrm{T}} K_{Y,Y}^{-1} K_{Y,\{x\}}$$

in terms of (8). □

Note that the function s in lemma 3 is fairly general and need not be an optimal solution to the restricted problem on Y. For the latter, there is a stronger result.

Lemma 4. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, $f \in \mathcal{K} \setminus \{0\}$, $Y \subset \Omega$ finite, $s \in \mathcal{K}_Y$ the unique solution of the restricted problem on Y for given $\eta > 0$. Then

$$\|f - s\|_K^2 \leqslant \|f\|_K^2 - \|s\|_K^2 + C_K(Y)^2\eta^2$$

and

$$\|f - s\|_{\infty,\Omega} \leqslant \|P_Y\|_{\infty,\Omega}\sqrt{\|f\|_K^2 - \|s\|_K^2} + C_\infty(Y)\eta$$

with the constants $C_K(Y)$, $C_\infty(Y)$ defined in the proof of the previous lemma.

Proof. Again let $s' \in \mathcal{K}_Y$ be the exact interpolant to f on Y. Then

$$
\begin{aligned}
\|f - s\|_K^2 &= \|f - s'\|_K^2 + \|s - s'\|_K^2 \\
&= \|f\|_K^2 - \|s'\|_K^2 + \|s - s'\|_K^2 \\
&\leqslant \|f\|_K^2 - \|s'\|_K^2 + C_K(Y)^2\eta^2 \\
&\leqslant \|f\|_K^2 - \|s\|_K^2 + C_K(Y)^2\eta^2,
\end{aligned}
$$

since $\|s'\|_K \geqslant \|s\|_K$ due to the optimality of s. Similarly

$$
\begin{aligned}
\|f - s\|_{\infty,\Omega} &\leqslant \|P_Y\|_{\infty,\Omega}\|f - s'\|_K + C_\infty(Y)\eta \\
&= \|P_Y\|_{\infty,\Omega}\sqrt{\|f\|_K^2 - \|s'\|_K^2} + C_\infty(Y)\eta \\
&\leqslant \|P_Y\|_{\infty,\Omega}\sqrt{\|f\|_K^2 - \|s\|_K^2} + C_\infty(Y)\eta.
\end{aligned}
$$

□

7. Iterative techniques

We now turn to the numerical solution of restricted or unrestricted quadratic optimization problems as considered in the preceding sections. If the problems are large, we want to use iterative methods based on rather small subproblems. We start with the restricted problem for a given $f \in \mathcal{K}$. Following lemma 3 we work on functions s_k that are based on small active sets Y_k in the sense of (2) and try to make $\|s_k\|_K$ large while keeping $\|f - s_k\|_{Y_k} \leqslant \eta$ small.

The following iterative technique is an adaptation of the greedy interpolation algorithm in [13, 14].

Algorithm 1. Let $X \subseteq \mathbb{R}^d$, K, f, and $\eta > 0$ be given. The *greedy algorithm* for the solution of the restricted problem starts from some finite set $Z_0 \subseteq X$ and iterates for $k = 0, 1, \ldots$ as follows:

1. Solve the restricted problem for fixed η on Z_k by some function s_k. Let $Y_k := \{x \in Z_k\colon |s_k(x) - f(x)| = \eta\}$.
2. If there is no $x \in X$ with $|s_k(x) - f(x)| > \eta$, the iteration stops and s_k solves the restricted problem for all of X.
3. Otherwise the iteration is repeated for $Z_{k+1} := Y_k \cup \{x\}$.

The efficient implementation of step 1 is treated later.

Theorem 7. If X is finite, the greedy algorithm stops after a finite number of steps with the solution of the restricted problem on X.

Proof. Because of lemma 2 we have $\|s_{k+1}\|_K < \|s_k\|_K$, $k = 0, 1, \ldots,$ as long as the algorithm does not stop. Thus no subset in the sequence $\{Z_k\}$ can occur twice. Since there are only finitely many subsets of X the algorithm stops. Thus the algorithm stops with a set $Z \subset X$ and a solution s_Z of the restricted problem on Z with $|s_Z(x) - f(x)| \leqslant \eta$ for all $x \in X$. The first part of lemma 2 shows that s_Z solves the restricted problem on X. □

Note that the greedy algorithm discards all points of $Z_k \setminus Y_k$ at step k. In the language of learning machines, the algorithm only memorizes those training samples that are "*currently active support vectors*" and forgets the others. After recognizing and digesting a new sample x, one or several of the older active samples may be discarded, even if they may be needed again some time later. The monotonicity of $\|s_k\|_K$ still guarantees progress. If the total number of training samples is finite, the full sample is mastered after a finite and usually rather small number of learning steps based on active samples only.

The actual progress of the algorithm crucially depends on the presented samples, like the success of a learning student depends on the quality of the teacher or trainer. If in step k the new training sample $(x, f(x))$ realizes the maximum of the error $\|s_k(x) - f(x)\|_{\infty,X}$, one can speak of a "maximum error" trainer, because the training always uses the case in which the student would commit the largest error. In view of (8) and (9) this is not necessarily the sample with the best progress in the sense of boosting up $\|s_k\|_K$. One should rather pick x to maximize the right-hand terms there. But this is somewhat more complicated, and it is not clear whether there is a substantial improvement justifying the additional work.

It is interesting to look at an infinite version of the above, simulating a learning system with strongly limited memory and capabilities, while exposed to a possibly infinite set of training samples presented sequentially one after the other. This is called *online learning* in learning theory [15]. In such a case, the greedy algorithm at stage k just ignores samples $x \in X$ with $|s_k(x) - f(x)| \leqslant \eta$ and waits for an x that does not satisfy this inequality.

Theorem 8. If X is infinite, the greedy algorithm does not cycle and generates a sequence of functions s_k with

$$\|s_k\|_K < \|s_{k+1}\|_K < \cdots \leqslant \|f\|_K.$$

Proof. Follows from the monotonicity lemma. □

Due to these results, the greedy algorithm always makes some progress in a weak sense, though it forgets everything except its critical observation sets Y_k and the current function s_k based on that set. Note that no assumption about the actually presented samples is made so far. It seems to be hard to prove that the progress of the algorithm is substantial, unless one makes such an assumption.

But there is a simple modification to the algorithm that makes it finite even for infinite X. It suffices to overdo its "training" somewhat, resulting in a sharpened monotonicity behavior.

Algorithm 2. Let $\Omega \subseteq \mathbb{R}^d$, positive numbers $\eta \geqslant \epsilon > 0$ and a kernel K be given, and consider an unknown function $f : \Omega \to \mathbb{R}$ which is in the native space for K. The *regularized* greedy algorithm coincides with the standard greedy algorithm, except that it uses the tolerance $\eta - \epsilon$ in step 1, while using η in step 2.

Theorem 9. Irrespective of the training technique and the presented sample set $X \subseteq \Omega \subset \mathbb{R}^d$, the regularized greedy algorithm performs only a finite number of actual learning steps if Ω is compact and the kernel is continuous. The maximal number of learning steps is determined by $\Omega \subseteq \mathbb{R}^d$, $\|f\|_K$, η, ϵ, and K only. The sets Y occurring in the algorithm have a separation distance

$$q(Y) := \frac{1}{2} \min_{y,z \in Y, y \neq z} \|y - z\|_2$$

that is uniformly bounded below by a positive constant.

Proof. Since all s_k have $\|s_k\|_K \leqslant \|f\|_K$, they are equicontinuous, and so are the $f - s_k$. This is a standard fact from reproducing kernel Hilbert space theory, if the kernel is continuous and the domain is compact. The proof uses the reproduction property and considers

$$\begin{aligned}\big|f(x) - f(y)\big|^2 &= \big|\big(K(x,\cdot) - K(y,\cdot), f\big)_K\big|^2\\ &\leqslant \big\|K(x,\cdot) - K(y,\cdot)\big\|_K^2 \|f\|_K^2\\ &= \big(K(x,x) - K(y,x) + K(y,y) - K(x,y)\big)\|f\|_K^2\end{aligned}$$

to prove that all bounded sets of functions consist of equicontinuous functions. Because $|f - s| \leqslant \eta - \epsilon$ on Y and $|(f - s)(x)| > \eta$, the uniform equicontinuity of all $f - s$ occurring in the algorithm implies $\operatorname{dist}(x, Y) \geqslant \delta > 0$ uniformly. Thus all Y occurring in the iteration have a separation distance uniformly bounded from below. Then there is a uniform upper limit M to the size of Y, and one can use a standard topology on Ω^M to conclude that the set of all such Y is compact. If the difference $\|\tilde{s}\|_K - \|s\|_K$ is positive and uniformly bounded below by a positive constant throughout the algorithm, the theorem follows.

It remains to prove that the difference $\|\tilde{s}\|_K - \|s\|_K$ is positive and uniformly bounded below by a positive constant. Assume a sequence of Y_k, s_k and x_k such that the above difference goes to zero. We can extract a subsequence such that the Y_k converge to some Y with a fixed number of points, the x_k converge to some x and the values of s_k on Y_k converge. In fact, the latter differ from values of f only by at most η, and the values of f and all s_k are uniformly bounded due to the standard bound

$$u(x)^2 = \big(u, K(x,\cdot)\big)_K^2 \leqslant K(x,x)\|u\|_K^2 \leqslant K(x,x)\|f\|_K^2$$

for all u with $\|u\|_K \leqslant \|f\|_K$.

Let s be the interpolant defined on Y that attains these values on Y. Since Y is nondegenerate and since in this case the solution of a kernel interpolation problem depends continuously on the data locations and the data, we know that $s - s_k$ converges to zero everywhere, including the point x. In fact, convergence of the data on Y implies convergence of coefficients, and this in turn implies convergence everywhere. Furthermore, we know that the $\|s_k\|_K$ converge to $\|s\|_K$ and that s satisfies the necessary and sufficient conditions for being the solution of the restricted problem on Y. In fact, the sign conditions on the coefficients will not be violated in the limit, when the values of the s_k converge on Y.

Now let $\tilde{s_k}$ solve the restricted problem on $Y_k \cup \{x_k\}$. Due to convergence of the Y_k to Y and x_k to x with the values of the s_k converging to those of s on Y and x, we have convergence of $\tilde{s_k}$ to the solution $\tilde{s}$ of the restricted problem on $Y \cup \{x\}$, and $\|\tilde{s_k}\|_K$ converges to $\|\tilde{s}\|_K$. Thus

$$\big\|\tilde{s}\big\|_K = \lim\big\|\tilde{s_k}\big\|_K = \lim \|s_k\|_K = \|s\|_K$$

which contradicts the monotonicity lemma, since we have $|f - s| \leqslant \eta - \epsilon$ on Y and $|s(x) - f(x)| \geqslant \eta$. □

The regularized greedy algorithm will always discard training data that are close to its active data. Thus it feeds only on new and interesting cases, and after a finite number of actual learning steps and with a compact set of conceivable inputs, there is nothing it can learn to improve its performance.

The regularized greedy algorithm still works if the "real" function f is not constant during the possibly infinite sampling and learning process. It will always adapt to any stationary situation after a finite number of steps.

Corollary 1. If the regularized greedy algorithm is executed with $\epsilon = \eta$, then it works with exact interpolation on small subsets and still shares the same properties.

The single steps then are computationally much more efficient, but they do not guarantee optimal complexity. because there will be no data points discarded through the process. Therefore this strategy only pays off at startup time, when the process is not anywhere near its complexity limits.

8. An efficient quadratic solver

We now exploit the special features of the finite-dimensional restricted problem on active sets in order to carry out the iteration steps of the greedy method efficiently.

Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite, $b \in \mathbb{R}^n$ and $\eta > 0$. Consider the quadratic programming problem

$$\left\{\begin{array}{l} \text{minimize} \quad f(x) := \frac{1}{2}x^{\mathrm{T}}Ax \quad \text{on} \\ M := \left\{x \in \mathbb{R}^n \colon \begin{pmatrix} A \\ -A \end{pmatrix} x \geqslant \begin{pmatrix} b - \eta \cdot 1 \\ -b - \eta \cdot 1 \end{pmatrix}\right\}. \end{array}\right. \tag{P}$$

We would like to adapt the active set method of quadratic programming to this problem. In particular we will show that it is only necessary to solve a small linear system of equations at each iteration, usually the main work to be done.

Let a_i^{T} denote the ith row of A and b_i the ith component of b, furthermore let a_{ij} be the entry in row i and column j. Thus

$$A = (a_{ij}) = \begin{pmatrix} a_1^{\mathrm{T}} \\ \vdots \\ a_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}.$$

For $J \subseteq I := \{1, \dots, n\}$ the matrix A_J is the $|J| \times n$-matrix with rows a_j^{T}, $j \in J$. Because of the symmetry of A the matrix A_J^{T} ist the $n \times |J|$-matrix with columns a_j, $j \in J$. Similarly, for $J, K \subseteq I$ let $A_{J,K}$ be the $|J| \times |K|$-matrix with entries $(a_{jk})_{(j,k)\in J\times K}$.

We now describe one step of the active set method of quadratic programming applied to the special problem (P). Note that we fix the index set $I := \{1, \dots, n\}$.

(0) Let $(x, I^-, I^+) \in M \times I \times I$ be a triple such that[⋆]

$$a_i^{\mathrm{T}} x - b_i = -\eta \quad (i \in I^-), \qquad a_i^{\mathrm{T}} x - b_i = +\eta \quad (i \in I^+).$$

Obviously (due to $\eta > 0$) we have:

- $I^- \cap I^+ = \emptyset$,
- $\{a_i\}_{i \in I^- \cup I^+}$ are linearly independent.

(1) Let $(p, y) \in \mathbb{R}^n \times \mathbb{R}^n$ be a pair consisting of the solution p and and a vector y such that y_{I^-}, y_{I^+} are optimal multipliers for the equality constrained quadratic program

$$\begin{cases} \text{minimize} \quad f(x+p) = f(x) + (Ax)^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} A p \quad \text{subject to} \\ A_{I^-} p = 0, \qquad A_{I^+} p = 0. \end{cases}$$

Thus one has to solve the linear system of equations

$$\begin{pmatrix} A & -A_{I^-}^{\mathrm{T}} & A_{I^+}^{\mathrm{T}} \\ -A_{I^-} & 0 & 0 \\ A_{I^+} & 0 & 0 \end{pmatrix} \begin{pmatrix} p \\ y_{I^-} \\ y_{I^+} \end{pmatrix} = - \begin{pmatrix} Ax \\ 0 \\ 0 \end{pmatrix}.$$

Let $J := I \setminus (I^- \cup I^+)$. The first equation leads to

$$p_J = -x_J, \qquad p_{I^-} = y_{I^-} - x_{I^-}, \qquad p_{I^+} = -x_{I^+} - y_{I^+}. \tag{11}$$

The last two equations $A_{I^-} p = 0$, $A_{I^+} p = 0$ can be written as

$$\begin{pmatrix} A_{I^-,I^-} & A_{I^-,I^+} & A_{I^-,J} \\ A_{I^+,I^-} & A_{I^+,I^+} & A_{I^+,J} \end{pmatrix} \begin{pmatrix} p_{I^-} \\ p_{I^+} \\ p_J \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

So we get (y_{I^-}, y_{I^+}) as a solution to

$$\begin{pmatrix} A_{I^-,I^-} & A_{I^-,I^+} \\ A_{I^+,I^-} & A_{I^+,I^+} \end{pmatrix} \begin{pmatrix} y_{I^-} \\ -y_{I^+} \end{pmatrix} = \begin{pmatrix} A_{I^-} x \\ A_{I^+} x \end{pmatrix}.$$

Observe that the coefficient matrix is symmetric and positive definite. When (y_{I^-}, y_{I^+}) is known, the direction p can be computed from (11). Formally we define $y_J := 0$.

(2) If $x + p \in M$:

– Compute $x_+ := x + p$.

– Determine
$l^- \in I^-$ with $y_{l^-} = \min_{i \in I^-} y_i$,
$l^+ \in I^+$ with $y_{l^+} = \min_{i \in I^+} y_i$.

– If $\min(y_{l^-}, y_{l^+}) \geqslant 0$ respectively $y_{I^-} \geqslant 0$ and $y_{I^+} \geqslant 0$, then:

[⋆] We do not require that $|a_i^{\mathrm{T}} x - b_i| < \eta$ for $i \in I \setminus (I^- \cup I^+)$. Thus, at the beginning, $I^- = \emptyset$ or $I^+ = \emptyset$ are possible.

  * STOP: $x^* := x_+$ solves (P), $I^* := I^- \cup I^+$ is the corresponding set of active constraints and y_{I^-}, y_{I^+} the corresponding nonnegative multipliers.

- Else:

  * If $y_{l^-} < y_{l^+}$ then $I_+^- := I^- \setminus \{l^-\}$, $I_+^+ := I^+$ else $I_+^- := I^-$, $I_+^+ := I^+ \setminus \{l^+\}$.

Else (i.e. $x + p \notin M$):

- Compute the maximal steplength $s(x, p) := \sup\{t \geqslant 0\colon x + tp \in M\}$. This can be done in the following way. Compute

$$s^-(x,p) := \min\left\{\frac{b_i - \eta - a_i^{\mathrm T}x}{a_i^{\mathrm T}p}\colon\ i \in I \setminus I^-,\ a_i^{\mathrm T}p < 0\right\},$$

$$s^+(x,p) := \min\left\{\frac{b_i + \eta - a_i^{\mathrm T}x}{a_i^{\mathrm T}p}\colon\ i \in I \setminus I^+,\ a_i^{\mathrm T}p > 0\right\},$$

and $s(x, p) := \min(s^-(x, p), s^+(x, p))$. Note that $s(x, p) \in [0, 1)$ (since $x \in M$ and $x + p \notin M$).

- If $s(x, p) = s^-(x, p)$:

  * Let

$$r^- \in R^- := \left\{i \in I \setminus I^-\colon\ a_i^{\mathrm T}p < 0,\ \frac{b_i - \eta - a_i^{\mathrm T}x}{a_i^{\mathrm T}p} = s^-(x,p)\right\}$$

and put $I_+^- := I^- \cup \{r^-\}$, $I_+^+ := I^+$.

- Else (i.e. $s(x, p) = s^+(x, p)$):

  * Let

$$r^+ \in R^+ := \left\{i \in I \setminus I^+\colon\ a_i^{\mathrm T}p > 0,\ \frac{b_i + \eta - a_i^{\mathrm T}x}{a_i^{\mathrm T}p} = s^+(x,p)\right\}$$

and put $I_+^- := I^-$, $I_+^+ := I^+ \cup \{r^+\}$.

- Compute $x_+ := x + s(x, p)p$.

(3) Make the update $(x, I^-, I^+) := (x_+, I_+^-, I_+^+)$ and go to (1).

Remark. We have

$$\begin{aligned} f(x+tp) &= f(x) + t(Ax)^{\mathrm T}p + \frac{1}{2}t^2p^{\mathrm T}Ap \\ &= f(x) + t\big[-Ap + A_{I^-}^{\mathrm T}y_{I^-} - A_{I^+}^{\mathrm T}y_{I^+}\big]^{\mathrm T}p + \frac{1}{2}t^2p^{\mathrm T}Ap \\ &= f(x) + t\left(\frac{1}{2}t - 1\right)p^{\mathrm T}Ap. \end{aligned}$$

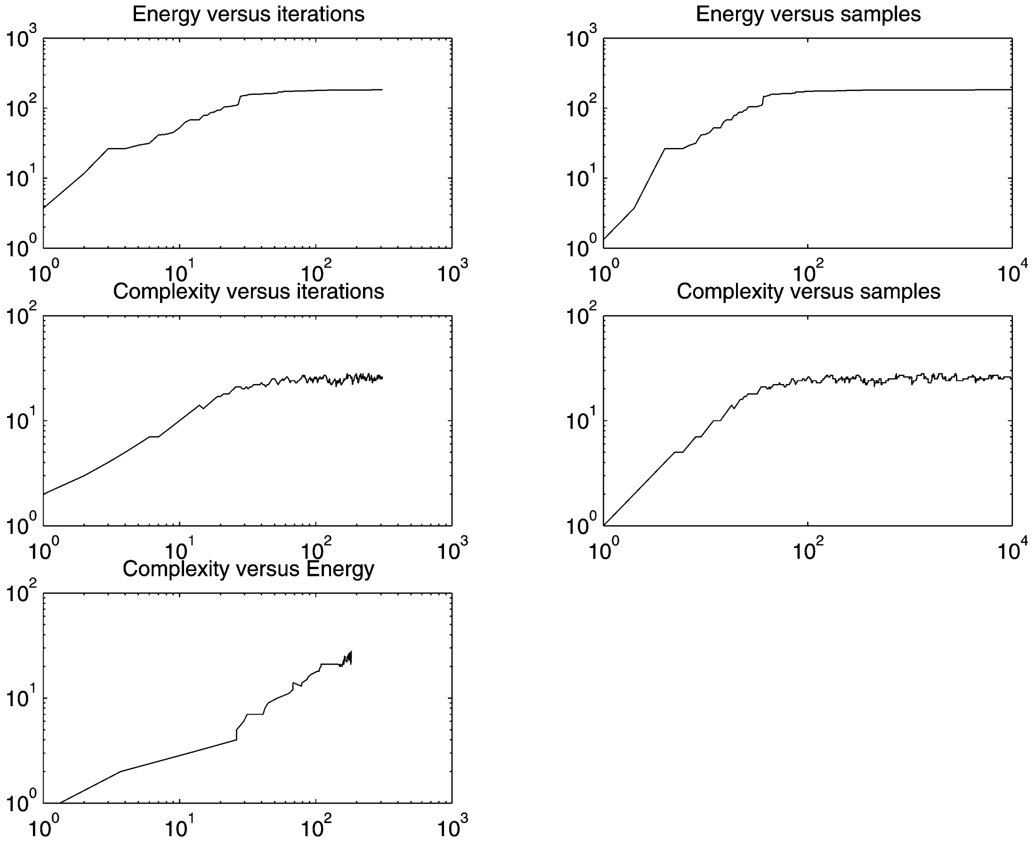

Figure 1. Typical online learning algorithm behavior.

Thus the decrease along the ray $\{x + tp\colon\ t \geqslant 0\}$ is maximal for $t = 1$. Therefore it is reasonable to take $x_+ := x + p$ as the next iterate if it is feasible. Since the steplength is either $t = 1$ or $t = s(x, p) \in [0, 1)$ we have $f(x_+) \leqslant f(x)$. There is no progress if and only if $p = 0$ or $s(x, p) = 0$. Let us suppose that $p = 0$ and one of the multipliers y_i, $i \in I^- \cup I^+$, is negative (otherwise the algorithm stops). An index with a negative multiplier will be removed from $I^- \cup I^+$. In the next step the direction p_+ will not be equal to zero due to the linear independence of $\{a_i\}_{i \in I^- \cup I^+}$. If $s(x, p) = 0$, either $s^-(x, p) = 0$ or $s^+(x, p) = 0$. If for instance $s^-(x, p) = 0$ there exists an index $i \in I \setminus I^-$ with $a_i^{\mathrm{T}} x - b_i = -\eta$. So in this case I^- does not contain all indices for which the first set of constraints is active. □

A simple Matlab function implementing the above algorithm for the solution of (P) is available on request from the second author.

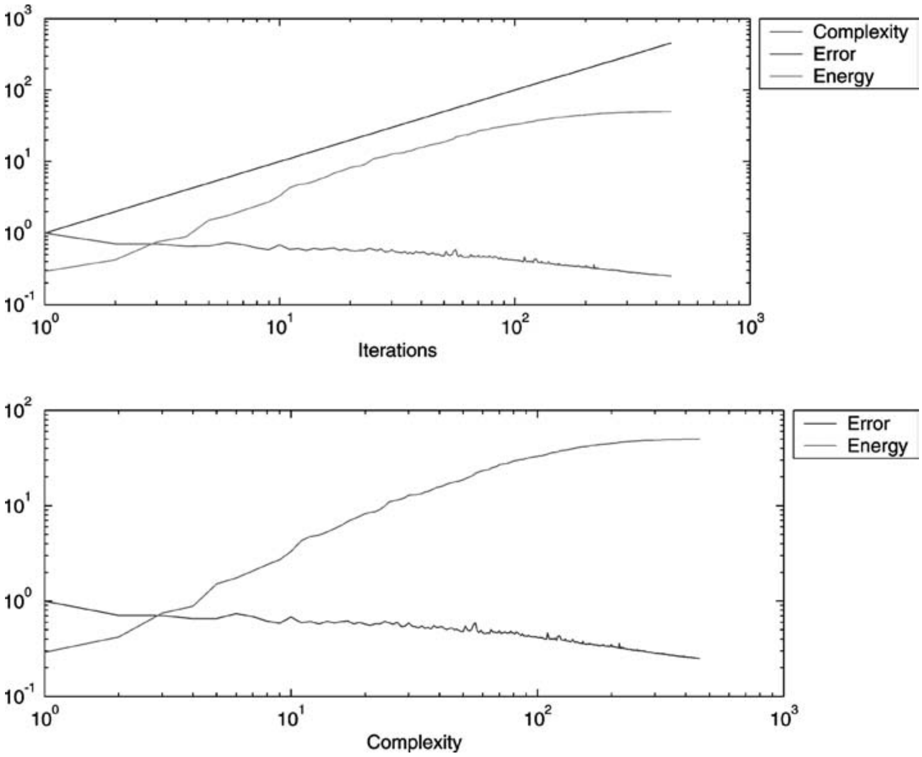

Figure 2. Greedy algorithm for census data.

9. Examples

Example. Figure 1 shows the behavior of the greedy algorithm for approximating the peaks function in MATLAB at error level $\eta = 0.01$ using up to 10000 random data. The algorithm is run in "online" form, ignoring data where the error is already below η, and performing an update step otherwise. *Energy* is measured as $(s, s)_K$, while *complexity* stands for the number of points in $Y_{\pm}$. The term *iteration* stands for an update step of the learning algorithm when confronted with an undiscardable sample. Note the strong correspondence between energy and complexity, while the algorithm builds up its knowledge. After some 100 samples, the algorithm tends to discard the majority of the new samples. It performs about 400 actual learning steps to master the function at that error level. From that point on it only very rarely accepts further samples it cannot deal with, and it works at a complexity of about 35 support vectors throughout.

Example. Figure 2 shows the behaviour of the technique of the previous chapter when applied iteratively on 22784 training samples with 16 variables describing input data for estimating the price of houses from the 1990 US census⋆ data.

⋆ `http://www.cs.toronto.edu/~delve/data/census-house/desc.html (16L dataset)`

Acknowledgement

The authors thank the referees for several valuable comments and corrections.

References

[1] D. Braess, Chebyshev approximation by splines wuth free knots, Numer. Math. (1974) 357–366.

[2] S. De Marchi, R. Schaback and H. Wendland, Optimal point locations for radial basis interpolation, in preparation (2002).

[3] R. Fletcher, *Practical Methods of Optimization* (Wiley, 1987).

[4] M. Frank and P. Wolfe, An algorithm for quadratic programming, Naval Res. Logist. Quart. 3 (1956) 95–110.

[5] C.I. Barrodale and Phillips, Algorithm 495: Solution of an overdetermined system of linear equations in the Chebychev norm, ACM Trans. Math. Software (TOMS) 1 (1975) 264–270.

[6] C.A. Micchelli, Interpolation of scattered data: Distance matrices and conditionally positive definite functions, Constr. Approx. 2 (1986) 11–22.

[7] C.A. Micchelli and T.J. Rivlin, A survey of optimal recovery, in: *Optimal Estimation in Approximation Theory*, eds. C.A. Micchelli and T.J. Rivlin (Plenum Press, 1977) pp. 1–54.

[8] C.A. Micchelli and T.J. Rivlin, Optimal recovery of best approximations, Results Math. 3 (1978) 25–32.

[9] C.A. Micchelli and T.J. Rivlin, Lectures on optimal recovery, in: *Numerical Analysis, Lancaster 1984*, ed. P.R. Turner, Lecture Notes in Math. 1129 (Springer-Verlag, 1984) pp. 12–93.

[10] C.A. Micchelli, T.J. Rivlin and S. Winograd, Optimal recovery of smooth function approximations, Numer. Math. 260 (1976) 191–200.

[11] R. Schaback, Native Hilbert spaces for radial basis functions I, in: *New Developments in Approximation Theory*, eds. M.D. Buhmann, D.H. Mache, M. Felten and M.W. Müller, Internat. Ser. Numer. Math. 132 (Birkhäuser, 1999) pp. 255–282.

[12] R. Schaback, Mathematical results concerning kernel techniques, Manuscript (2002).

[13] R. Schaback and H. Wendland, Adaptive greedy techniques for approximate solution of large RBF systems, Numer. Algorithms 24 (2000) 239–254.

[14] R. Schaback and H. Wendland, Numerical techniques based on radial basis functions, in: *Curve and Surface Fitting*, eds. A. Cohen, C. Rabut and L. Schumaker (Vanderbilt University Press, Nashville, TN, 2000).

[15] B. Schölkopf and A.J. Smola, *Learning with Kernels* (MIT Press, 2002).

[16] J. Stewart, Positive definite functions and generalizations, an historical survey, Rocky Mountain J. Math. 6 (1976) 409–434.

[17] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).

[18] J. Werner, *Optimization Theory and Applications* (Vieweg, 1984).