CHAPTER 7

Finite-Dimensional Approximating Subspaces

7.1 Quadrature formulae, computing with basis functions.

Suppose we wish to compute the minimizer in $\mathcal{H}_R$ of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - L_i f)^2 + \lambda\|P_1 f\|^2 \tag{7.1.1}$$

where either n is very large, and/or, we do not have a closed form expression for

$$\xi_i(t) = L_{i(u)}R^1(t,u).$$

If, for example,

$$L_i f = \int_\Omega K(t_i,u)f(u)\,du,$$

then

$$\xi_i(t) = \int_\Omega K(t_i,u)R^1(t,u)\,du, \tag{7.1.2}$$

and, if a closed form expression is not available for $\xi_i(t)$, it would appear that a quadrature or other approximation to (7.1.2) would be necessary.

A quadrature formula in the context of $\mathcal{H}_R$ can be obtained as follows. Let $s_1,\ldots,s_N$ be N (distinct) points in $\mathcal{T}$ such that the $N \times M$ matrix with $l\nu$th entry $\phi_\nu(s_l)$ is of rank M, and, for any $f \in \mathcal{H}_R$, let f_0 be that element in $\mathcal{H}_R$ that minimizes $\|P_1 f_0\|^2$ subject to $f_0(s_l) = f(s_l)$, $l = 1,\ldots,N$. Then, if η is the representer of integration in $\mathcal{H}_R$

$$< \eta, f >= \int f(s)\,ds, \tag{7.1.3}$$

we approximate $< \eta, f >$ by $< \eta, f_0 >$, which is a linear combination of the values of f at $s_1,\ldots,s_N$ (i.e., a quadrature formula). Certain classical quadrature formulae can be obtained this way (see Schoenberg (1968)). Now define $\hat{\eta}$ by the relationship $< \hat{\eta}, f >= < \eta, f_0 >$, all $f \in \mathcal{H}_R$. Since $< \hat{\eta}, f >$ depends on f only through $f(s_1),\ldots,f(s_N)$, it follows that $\hat{\eta}$ is in span $\{R_{s_1},\ldots,R_{s_N}\}$, where R_{s_l} is the representer of evaluation at s_l. Thus $< \hat{\eta}, f - f_0 >= 0$, for any $f \in \mathcal{H}_R$. It can

also be shown that $< \eta - \hat{\eta}, f >= 0$ for any $f \in \mathcal{H}_0$, that is, the quadrature approximation is exact for $f \in \mathcal{H}_0$. This is equivalent to $\eta - \hat{\eta} \perp \mathcal{H}_0$. These facts result in the so-called hypercircle inequality (see Golomb and Weinberger (1959), Wahba (1969)), which goes as follows. Since $< \hat{\eta}, f >=< \eta, f_0 >$, we have

$$< \eta - \hat{\eta}, f >=< \eta, f - f_0 >=< \eta - \hat{\eta}, f - f_0 >=< \eta - \hat{\eta}, P_1(f - f_0) >$$

$$=< P_1(\eta - \hat{\eta}),\ P_1(f - f_0) >$$

so that

$$| < \eta, f > - < \hat{\eta}, f > | \leq \|P_1(\eta - \hat{\eta})\| \cdot \|P_1(f - f_0)\|.$$

High-order convergence rates for $\|P_1(f - f_0)\|$ are available in the context of univariate splines (see Schultz (1973a), Schumaker (1981)), and thin-plate splines (Duchon (1978)). The famous "optimal quadrature" problem can be formulated as the problem of choosing $s_1, \ldots, s_N$ to minimize $\|P_1(\eta - \hat{\eta})\|$ or $\sup_{f \in \mathcal{E}} \|P_1(f - f_0)\|$ for some class $\mathcal{E}$ (see Schoenberg (1968) and Section 12.2).

This kind of quadrature was discussed in Nychka et al. (1984) in the context of numerically minimizing (7.1.1) and it was noted that essentially the same accuracy can be obtained in a computationally much simpler way by minimizing (7.1.1) in a certain subspace $\mathcal{H}_N$ spanned by N suitably chosen basis functions. Given basis functions $B_1, \ldots, B_N$, one sets

$$f = \sum_{k=1}^{N} c_k B_k,$$

substitutes this expression into (7.1.1), and solves for the coefficients $c_1, \ldots, c_N$. We next discuss the choice of basis functions.

In Wahba (1980b) it was proposed, in the context of $L_i f = f(t_i)$ with very large n, that a good subspace $\mathcal{H}_N$ of basis functions can be chosen as follows. Choose $s_1, \ldots, s_N$ points distributed "nicely" over $\mathcal{T}$, and let the basis functions $B_1, \ldots, B_N$ be chosen as follows. Let $B_1, \ldots, B_M$ be $\phi_1, \ldots, \phi_M$. Let $u_l = (u_{1l}, \ldots, u_{Nl})'$, $l = 1, 2, \ldots, N - M$ be $N - M$ linearly independent vectors with the property

$$\sum_{k=1}^{N} u_{kl} \phi_\nu(s_k) = 0,\ l = 1, \ldots, N - M,\ \nu = 1, \ldots, M \tag{7.1.4}$$

and let

$$B_{M+l} = \sum_{k=1}^{N} u_{kl} P_1 R_{s_k},\ l = 1, \ldots, N - M. \tag{7.1.5}$$

We have that for any $f \in \mathcal{H}_R$, f_0, that element in $\mathcal{H}_R$, which minimizes $\|P_1 f_0\|^2$ subject to $f_0(s_l) = f(s_l)$, $l = 1, \ldots, N$, is in $\mathcal{H}_N$. It so happens in $W_m[0, 1]$ (but *not* in general), that there exist coefficients u_{kl} in (7.1.5) so that the B_{M+l}, $l = 1, 2, \ldots, N - M$ have compact support. This special case is important,

so we will describe it in some detail. The coefficients are those that characterize ordinary divided differences. We show this next. Here, $M = m$ and we let $s_1 < s_2 < \ldots < s_N$. Using the reproducing kernel for $W_m[0,1]$ given in Section 1.2 we have, for any fixed s,

$$P_1 R_s(t) = \int_0^1 \frac{(s-u)_+^{m-1}(t-u)_+^{m-1}}{[(m-1)!]^2}\,du = \xi_s(t), \text{ say.}$$

Recall that, for fixed s, ξ_s, considered as a function of t, satisfies

$$\xi_s \in \pi^{2m-1},\ t \in [0,s],$$

$$\xi_s \in \pi^{m-1},\ t \in [s,1],$$

ξ_s has $2m-2$ continuous derivatives, and $\xi_s(t)$ is symmetric in s and t. Let $[s_l,\ldots,s_{l+2m}]\xi_s$ denote the $2m$th divided difference of ξ_s with respect to s, for example, a first divided difference is $[s_1,s_2]\xi_s = (\xi_{s_2} - \xi_{s_1})/(s_2 - s_1)$. Let

$$B_{m+l} = [s_l,\ldots s_{l+2m}]\xi_s,\ l = 1,\ldots,N-2m.$$

Then B_{m+l} (considered as a function of t) is a linear combination of $\xi_{s_l}, \xi_{s_{l+1}}, \ldots, \xi_{s_{l+2m}}$. B_{m+l} is hence a piecewise polynomial of degree at most $2m-1$ with knots at $s_l,\ldots,s_{l+2m}$, and possessing $2m-2$ continuous derivatives. We next show that $B_{m+l}(t) = 0$ for $t \notin [s_l, s_{l+2m}]$. For any fixed $t \le s_l \le s$ we may write

$$\xi_s(t) = \sum_{\nu=0}^{m-1} s^\nu f_\nu(t), \quad s \ge s_l \ge t,$$

for some f_ν's, since $\xi_s(t)$ is a polynomial of degree $m-1$ in s for $s \ge t$. Similarly, for $t \ge s_{l+2m} \ge s$ we may write

$$\xi_s(t) = \sum_{\nu=0}^{2m-1} s^\nu \tilde{f}_\nu(t)$$

for some $\tilde{f}_\nu$. Since $[s_l,\ldots,s_{l+2m}]s^r = 0$ for any $r = 1,2,\ldots,2m-1$, it follows that

$$[s_l,\ldots,s_{l+2m}]\xi_s(t) \equiv B_{m+l}(t) = 0,\quad t \notin [s_l, s_{l+2m}]$$

This gives $N-2m$ basis functions with compact support; the remaining m may be obtained, e.g., as

$$B_{N-m+k} = [s_{N-2m+k},\ldots,s_N]\xi_s,\quad k = 1,\ldots,m,$$

and then $B_{N-m+k(t)}(t) = 0$, for $t \le s_{N-2m+k}$.

Basis functions with compact support that span the space of natural polynomial splines with knots $s_1,\ldots,s_n, s_i \in [0,1]$ are studied in some detail in Schumaker (1981, §8.2). $n-2m$ of these basis functions are so-called B splines. These B splines are piecewise polynomials of degree $2m-1$, with $2m-2$

continuous derivatives, have exactly $2m+1$ knots, $s_l, \ldots, s_{l+2m}$, and are zero outside $[s_l, s_{l+2m}]$. It is known that (nontrivial) piecewise polynomials of the given degree and order of continuity cannot be supported on fewer knots. For equally spaced knots, with spacing h, the B splines are translated and scaled versions of the convolution of $2m$ uniform distributions on $[0, h]$. In general, they are nonnegative hill-shaped functions. They are very popular as basis functions both for their good approximation theoretic properties and their ease of computation. Simple recursion relations are available to generate them directly (see Lyche and Schumaker (1973), deBoor (1978), Schumaker (1981)). Software for generating B-spline bases given arbitrary knots s_l is publicly available (see Chapter 11).

Given basis functions $B_1, \ldots, B_N$ we now seek $f_{N,\lambda}$ of the form

$$f_{N,\lambda} = \sum_{k=1}^{N} c_k B_k$$

to minimize

$$\frac{1}{N}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{N} x_{ik}c_k\right)^2 + \lambda \sum_{k,l=1}^{N} c_k c_l \sigma_{kl},$$

where

$$x_{ik} = \int K(t_i, s)B_k(s)\,ds$$

and

$$\sigma_{kl} = < P_1 B_k, P_1 B_l > .$$

For $\mathcal{H}_R = W_m[0,1]$,

$$\sigma_{kl} = \int_0^1 \frac{d^m}{dx^m}B_k(x)\frac{d^m}{dx^m}B_l(x)\,dx$$

and σ_{kl} will be zero if B_k and B_l have no common support. Then

$$c = c_\lambda = (X'X + n\lambda\Sigma)^{-1}X'y$$

and

$$A(\lambda) = X(X'X + n\lambda\Sigma)^{-1}X'. \tag{7.1.6}$$

Here, we really have two smoothing parameters, namely, λ and N, the number of basis functions, assuming the distribution of $s_1, \ldots, s_N$ for each N is fixed. In principle one could compute $V(N, \lambda)$ and minimize over both N and λ. We think that, to minimize errors due to numerical approximation, one wants to make N as large as is convenient or feasible on the computing equipment, or at least large enough to ensure that no "resolution" is lost at this stage, and then choose λ to minimize V.

7.2 Regression splines.

A number of authors have suggested using regression splines (i.e., $\lambda = 0$), particularly in the case $L_i f = f(t_i)$. Then N is the smoothing parameter. In this case V is particularly easy to compute, since trace $A(N) = N/n$. von Golitschek and Schumaker (1990) show, under general circumstances, that if $f \notin \mathcal{H}_N$, then it is always better to do smoothing as opposed to regression. They show that there is always some $\lambda > 0$ that is better than $\lambda = 0$ in the sense that, if $f \notin \mathcal{H}_N$, then $ET(\lambda)$ defined by

$$E\sum_{i=1}^{n}(L_i f - L_i f_\lambda)^2$$

satisfies $(d/d)\lambda ET(\lambda)|_{\lambda=0} < 0$.

Furthermore, for $L_i f = f(t_i)$, and $\mathcal{H}_R = W_m$, Agarwal and Studden (1980) have shown that the optimal N is $O(1/n^{1/2m+1}) = O(1/n^{1/5})$ for $m = 2$, say. For n of the order of, say 100, the optimal N for B-spline regression will be quite small and the estimate will not "resolve" multiple peaks that could easily be resolved by a smoothing spline. Nevertheless, arguments have been made for the use of regression splines in the case of extremely large data sets, and/or in situations where one does not expect to recover much "fine structure" in the estimates. (That is, when the true f is believed to be approximately in $\mathcal{H}_N$ for small N.) If the knots, that is the s_k, are also considered free variables, then more flexibility is available. If the knots are chosen to minimize the residual sum of squares, then trace $A(N)$ can be expected to be an underestimate of the degrees of freedom for signal in the denominator of the GCV function. Friedman and Silverman (1987) and Friedman (1991) have proposed correction factors for this.