Optimal Designs and Spline Regression*

by

W. J. Studden

Published in "Optimizing Methods in Statistics" (1971)
Ed. Rustagi

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series No. 257

June 1971

*This work was supported by the NSF grant GP 20306.

by

W. J. Studden*

Purdue University

This paper may be divided roughly into two parts. One part consists of an elementary discussion of splines. The second part considers two types of optimal designs: D and I_σ optimal designs. Section 1 is of an introductory nature.

The splines are considered in section 2 while sections 3 and 4 are devoted to the D and I_σ optimal designs. In section 5 & 6 we consider some examples and computational procedures for obtaining optimal designs. The paper is mainly expository, however it is restricted to a very small portion of design theory. The applications in section 5 are new.

§1. Introduction. The design problem under discussion is as follows. Let $f' = (f_1,\ldots,f_m)$ denote an m-vector of continuous functions defined on a compact set X. The points of X are referred to as the possible levels of feasible experiments and the variable $x \in X$ is sometimes called the control variable. For each level $x \in X$ some experiment may be performed whose outcome $Y(x)$ is a random observation with mean value

$$(1.1) \qquad E\,Y(x) = \sum_{i=1}^{m} \theta_i\, f_i\,(x)$$

and variance σ^2 independent of x. The simplest situation is, say, where X is an interval of the real line, $f_1(x) \equiv 1$, $f_2(x) = x$ and $E\,Y(x) = \theta_1 + \theta_2\,x$. The functions $f_1,\ldots,f_m$ are called the regression functions and are known to the experimenter. The regression coefficients or parameters $\theta_1,\ldots,\theta_m$ and σ^2 are unknown. On the basis of N uncorrelated observations we wish to estimate some function of the parameters $\theta_1,\ldots,\theta_m$.

*This work was supported by the NSF grant GP 20306. Reporduction is permitted in whole or in part for any purposes of the United States Government.

An experimental design specifies a probability measure μ (usually discrete) on X. The associated experiment involves taking observations at the level x proportional to μ. Thus if μ assigns mass $p_1, \ldots, p_r$ to $x_1, \ldots, x_r$ and $Np_i = n_i$ are integers the experimenter takes n_i observations at x_i. Designs with Np_i not equal to an integer can in practice only be approximated.

If the unknown parameter vector $\theta' = (\theta_1, \ldots, \theta_m)$ is estimated by least squares then the covariance matrix of the estimates $\hat{\theta}$ is given by

$$E(\hat{\theta} - \theta)(\hat{\theta} - \theta)' = \frac{\sigma^2}{N} M^{-1} (\mu) \tag{1.2}$$

The matrix $M(\mu) = \int f(x)\, f'(x)\, d\mu(x)$ is called the (normalized) information matrix of the design μ. The design problem we consider is to choose μ so that $M(\mu)$ is large in some sense.

Further discussion of the design problem considered here can be found in Karlin and Studden (1966), Kiefer (1959) and Kiefer and Wolfowitz (1960).

In the following we are interested in two problems 1) maximize the determinant of $M(\mu)$, denoted by $|M(\mu)|$, 2) minimize the quantity $L[M^{-1}(\mu)]$ where L is a non-negative linear functional on the set of positive semidefinite matrices, i.e.

$$\begin{aligned} &L(\alpha D_1 + (1-\alpha) D_2) = \alpha L(D_1) + (1-\alpha) L(D_2) \quad 0 \le \alpha \le 1 \\ &L(D) \ge 0 \text{ for } D \ge 0. \end{aligned} \tag{1.3}$$

The first of these problems was considered by Kiefer and Wolfowitz (1960); the second by Fedorov (1971).

§2. Spline Functions. The term spline usually refers to a "piecewise polynomial". Here the space X is an interval $[a,b]$. The interval $[a,b]$ is divided into k + 1 pieces by k "knots" $\xi_1, \ldots, \xi_k$ where $\xi_0 = a < \xi_1 < \ldots < \xi_k < b = \xi_{k+1}$. A function $s(x)$ on $[a,b]$ is called a spline if $s(x)$ is equal to a polynomial on (ξ_i, ξ_{i+1}) (different on each interval) and satisfies certain differentiability conditions at the points $\xi_1, \ldots, \xi_k$.

The simplest case stipulates the $s(x)$ is linear on each interval (ξ_i, ξ_{i+1}) $i = 0, 1, \ldots, k$ and is continuous at each ξ_i, $i = 1, 2, \ldots, k$. We thus have a polygonal line segment. For the quadratic case we may consider $s(x)$ to be quadratic on each interval with possibly continuity and also differentiability at each ξ_i. Discontinuities of the second derivative of $g(x)$ are allowed to each ξ_i.

Generally a function which is equal to a polynomial of degree at most n on each interval and has n-1 continuous derivatives at ξ_i can be written in the form

$$\sum_{i=0}^{n} \alpha_i x^i + \sum_{i=1}^{k} \beta_i (x-\xi_i)_+^n \tag{2.1}$$

where $(x)_+^n = x^n$ for $x > 0$ and is zero for $x < 0$. Thus the term $(x-\xi_i)_+^n$ is zero unless $x > \xi_i$. Variable degrees of differentiability can be allowed at ξ_i by using terms $(x-\xi_i)_+^r$ for $r < n$, however we shall not consider these here.

Spline polynomials have received considerable attention from mathematicians, working especially in approximation theory. These functions seem to be extremely suitable for interpolating or approximating data in real world situations since in many cases the underlying functional from $s(x)$ is different on different parts of X. For example if $s(x)$ denotes the distance or path a rocket travels in time $t = x$, intermitant auxillary rockets may change the form of $s(x)$.

The spline polynomials are the least oscillatory functions for interpolating data. For example the differentiable function $s(x)$ which interpolates the values y_i at ξ_i (i.e. $s(\xi_i) = y_i$) and minimizes $\int (s''(x))^2 \, dx$ is a cubic spline of the form (2.1) with knots at ξ_i (and is linear below ξ_1 and above ξ_k).

In using the functions

$$(2.2) \qquad 1\ x_1, \ldots, x^n, (x-\xi_1)_+^n, \ldots (x-\xi_k)_+^n$$

one is sometimes interesting in knowing for which set of x values $x_1, x_2, \ldots, x_{n+k+1}$ can one interpolate an arbitrary set of ordinates $y_1, y_2, \ldots, y_{n+k+1}$ using a unique linear combination of the functions (2.2). This is the case if and only if

$$(2.3) \qquad x_i < \xi_i < x_{n+i+1} \qquad i = 1,2,\ldots, k$$

This results says that we cannot overburden any given interval with two many x_i values. For example, if $n = 1$ and $k = 2$, we use $1, x, (x-\xi_1)_+, (x-\xi_2)_+$ and the inequalities (2.3) say that x_1, x_2, x_3, x_4 must satisfy $x_1 < \xi_1 < x_3$ and $x_2 < \xi_2 < x_4$.

For further and more complete discussion of splines we refer the reader to Rice (1968), Rivlin (1969) and Schoenberg (1969).

§3. Optimal Designs. In this section we discuss two types of problems:

1. Maximize $|M(\mu)|$ with respect to μ.
2. Minimize $L[M^{-1}(\mu)]$ where L is a linear functional on the set of positive semidefinite matrices such that $L(A) \geq 0$ for $A \geq 0$; i.e. A positive semidefinite.

A solution to problem 1 is called a D-optimal design and a solution to problem 2 an L-optimal design. We are particularly interested in certain special L-optimal designs which we denote by $I_{\bar{\sigma}}$-optimal designs. These will be considered near the end of this section. For the D-optimal designs we have the following "Equivalence Theorem" of Kiefer and Wolfowitz (1960) which shows that the D-optimal and minimax designs are equivalent.

Theorem 3.1 The conditions

(i) μ^* maximizes $|M(\mu)|$

(ii) μ^* minimizes $\sup\limits_{x \in X} f'(x)\, M^{-1}(\mu)\, f(x)$

(iii) $\sup\limits_{x \in X} f'(x)\, M^{-1}(\mu)\, f(x) = m$

are equivalent. The set B of all μ satisfying these conditions is convex and closed and $M(\mu)$ is the same for all $\mu \in B$.

We recall here that $M(\mu)$ is the information matrix and $f'(x)\, M^{-1}(\mu)\, f(x)$ is proportional to the variance of the estimate of the regression function at the point x.

Recently V. V. Fedorov has generalized these results to problem 2 and has shown the following.

Theorem 3.2. The conditions

(i) μ^* minimizes $L[M^{-1}(\mu)]$

(ii) μ^* minimizes $\sup\limits_{x \in X} L[M^{-1}(\mu)\, f(x)\, f'(x)\, M^{-1}(\mu)]$

(iii) $\sup\limits_{x \in X} L[M^{-1}(\mu)\, f(x)\, f'(x)\, M^{-1}(\mu)] = L[M^{-1}(\mu^*)]$

are equivalent. The set of designs satisfying (i), (ii) or (iii) is convex.

The second problem is the problem of "quadratic loss". That is, any function L satisfying the required conditions can be put in the form $L(M^{-1}) = \operatorname{tr} M^{-1} C$ where C is positive semidefinite and tr denotes the trace of a matrix. The term quadratic loss is used here since the expected value of $(\hat{\theta} - \theta)'\, C(\hat{\theta} - \theta)$ is proportional to $\operatorname{tr} M^{-1} C$. The design giving information matrix M is used to obtain the estimates $\hat{\theta}$. Thus, minimizing the expected value of the quadratic form $(\hat{\theta} - \theta)' C(\hat{\theta} - \theta)$ is equivalent to minimizing $L(M^{-1}) = \operatorname{tr} M^{-1} C$.

Both of these theorems may be proven using a variational technique. We consider the design

$$\mu_0 = (1-\alpha)\ \mu^* + \alpha\mu_x$$

where μ_x concentrates all of its mass at the single point x. The proof consists of showing that μ^* minimizes $L(M^{-1}(\mu))$ if and only if the derivative of

$$g(\alpha) = L[\ M^{-1}((1-\alpha)\ \mu^* + \alpha\mu_x)]$$

is ≥ 0 for $\alpha = 0$ for all x. This in turn is equivalent to part (ii) of theorem .

We are particularly interested in those L which are invariant under a basis change of the regression functions $f_1, f_2, \ldots, f_m$. Further we would like explicit expressions or characterizations of the optimal designs for the case of spline regression.

One of the invariant functionals L for Theorem 3.2 is the integral of the variance of the response surface estimate. Thus if $d\sigma$ denotes a measure on X (or possibly on a larger domain) then

$$\begin{aligned} \int f'(x)\ M^{-1}(\mu)\ f(x)\ d\sigma(x) &= \int \operatorname{tr} M^{-1}(\mu)\ f(x)\ f'(x)\ d\sigma(x) \\ &= \operatorname{tr} M^{-1}(\mu)\ M(\sigma) \\ &= L[(M^{-1}(\mu)] \end{aligned}$$

For fixed σ we wish to minimize this quantity with respect to the design μ.

Note that the expression $\operatorname{tr} M^{-1}(\mu)\ M(\sigma)$ is invariant under basis change of the regression functions. In this case if we know that the minimizing

$\mu = \mu^*$ concentrates its mass on m points $x_1^*, \ldots, x_m^*$ (m = the number of regression functions) then we can use as a basis the Lagrange functions $\ell_i(x)$ defined by the conditions $\ell_i(x_j) = \delta_{ij}$ $i, j = 1, \ldots, m$.

Lemma (Fedorov) If for given σ the design μ^* minimizing $L[M^{-1}(\mu)]$ concentrates mass on $x_1^*, \ldots, x_m^*$ then the corresponding weights are proportional to $K_{ii}^{1/2}$ where $K_{ii} = \int \ell_i^2(x)\, d\sigma(x)$.

Proof This follows by noting that for the Lagrange basis

$$\operatorname{tr} M^{-1}(\mu)\, M(\sigma) = \sum_{i=1}^{m} \frac{K_{ii}}{p_i}$$

Schwarz's inequality then gives

$$\sum_i \frac{K_{ii}}{p_i} \geq \left(\sum K_{ii}^{1/2}\right)^2$$

with equality only if $p_i = c\, K_{ii}^{1/2}$.

Note that if the I_σ-optimal design concentrates on m points $x_1^*, \ldots, x_m^*$ then the design problem reduces to minimizing

$$\sum_i \left(\int \ell_i^2(x)\, d\sigma\right)^{1/2} \quad \text{with respect to } x_1^*, \ldots, x_m^*.$$

For the minimax design the corresponding expression is

$$\max_{x \in X} \sum_i \ell_i^2(x) .$$

§4. Comparison of D and I_σ optimal designs. In comparing the two types of designs we shall assume that both designs are concentrated on the same set of points $x_1^*, \ldots, x_m^*$ and that $X = [-1, 1]$. If μ concentrates mass p_i on x_i^*, $i = 1, \ldots, m$ and $\ell_i(x)$ denotes the Lagrange functions corresponding to $x_1^*, \ldots, x_m^*$ then

$$f'(x)\, M_f^{-1}(\mu)\, f(x) = \ell'(x)\, M_\ell^{-1}(\mu)\, \ell(x)$$
$$= \sum_{i=1}^{m} \frac{\ell_i^2(x)}{p_i}$$

Here we denote the information matrix for the basis $f. = (f_1,\ldots, f_m)$ by M_f and for the Lagrange basis by M_ℓ. For the Lagrange basis the matrix M_ℓ is diagonal with diagonal elements $p_1, p_2,\ldots, p_m$. For the D-optimal design μ^* on $x_1^*,\ldots, x_m^*$ the weights are equal and

$$f'(x)\, M_f^{-1}(\mu^*)\, f(x) = m \sum_i \ell_i^2(x).$$

Moreover this function reaches its maximal value m at the points x_i^*.

If the weights p_i at x_i^* are not taken equal then $f'(x)\, M^{-1}(\mu)\, f(x)$ is raised at those x_i for which $p_i < \frac{1}{m}$ and lowered if $p_i > \frac{1}{m}$. For σ uniform on $[-1,1]$ the I_σ- design will produce variances $f'(x)\, M^{-1}\, f(x)$ which are low in the middle and higher at the ends in order to minimize the integral.

For a given σ the minimum value of $\int f'(x)\, M^{-1}\, f(x)\, d\sigma(x) = \operatorname{tr} M^{-1}(\mu)\, M(\sigma)$ will be $(\sum K_{ii}^{1/2}(\sigma))^2$ where $K_{ii}(\sigma) = \int \ell_i^2(x)\, d\sigma(x)$. The optimal weights are proportional to $K_{ii}^{1/2}$. Using the D-optimal design with equal weights gives a value $\operatorname{tr} M^{-1}(\mu)\, M(\sigma) = m \sum_i K_{ii}(\sigma)$. Thus we should compare $(\sum K_{ii}^{1/2}(\sigma))^2$ with $m \sum_i K_{ii}(\sigma)$, the former of course being the smaller of the two.

§ Examples. All of the examples below are concerned with polynomial or polynomial spline regression.

Example 1. Polynomial regression. It has been known for sometime [see Guest (1958)] that if $f_i(x) = x^{i-1}$, $i = 1,2,\ldots, n+1$ and $X = [-1,1]$ then the D-optimal design concentrates equal mass on the zeros $x_0^* = -1 < x_1^* < \ldots < x_{n-1}^* < x_n^* = 1$ of $(1-x^2)\, P_n'(x)$ where P_n is the nth Legendre polynomial, orthogonal on $[-1,1]$ to the uniform measure.

Recently Fedorov has shown that the I-optimal design for the uniform measure $d\sigma = dx$ is also supported by this same set of points. The weights in this case can be shown to be proportional to $|P_n(x_i^*)|^{-1}$. For the case $n = 2$ this gives weight .25, .50, .25 to -1, 0, 1. For $n = 3$ the design has weight .154, .346, .346, .154 at the points -1, -.447, +.447, +1. One can choose the measure σ with mass outside of X. For example if σ is uniform on $(-a, a)$ and $n = 2$ the resulting design is again on -1, 0, 1 with weights proportional to the square roots of

$$\frac{1}{4}\left(\frac{1}{5} + \frac{1}{3a^2}\right) \qquad \frac{1}{5} - \frac{2}{3a^2} + \frac{1}{a^4} \qquad \frac{1}{4}\left(\frac{1}{5} + \frac{1}{3a^2}\right)$$

Note that the weights are the same for $a = 1$ and $a \to \infty$.

<u>Example 2</u>. <u>Linear-splines</u>. Here are considered the set of regression functions

$$1,\ x,\ (x - \xi_1)_+ \ \ldots\ (x - \xi_k)_+$$

on the interval $[-1,1]$ where $-1 = \xi_0 < \xi_1 < \ldots < \xi_k < \xi_{k+1} = 1$. This example is fairly easy to work with since we may restrict ourselves to designs concentrating mass only on the $k+2$ points $\xi_0, \xi_1, \ldots, \xi_k, \xi_{k+1}$. With the aid of the Lemma in §3 it is fairly easy to show that the I_σ-optimal design for σ uniform on $[-1,1]$ has weights proportional to

$$(\xi_1-\xi_0)^{1/2},\ (\xi_2-\xi_0)^{1/2},\ (\xi_3-\xi_1)^{1/2}, \ldots,\ (\xi_{k+1}-\xi_{k-1})^{1/2},\ (\xi_{k+1}-\xi_k)^{1/2}$$

Referring to the discussion in §4 we note that if we have a pair of intervals close together the regression variance at the middle ξ_i will increase considerably. For the case where the ξ_i are equally spaced the weights are proportional to $1 : \sqrt{2} : \sqrt{2} : \ldots : \sqrt{2} : 1$. For $k = 1$ this gives weights .293, .414 and .293 (approx.) on -1, 0, 1.

<u>Example 3</u>. <u>Quadratic and Cubic Splines</u>. For quadratic splines we consider the regression functions

$$1,\ x,\ x^2,\ (x-\xi_1)_+^2, \ldots,\ (x-\xi_k)_+^2$$

on the interval [-1,1]. It is known (see Studden and Van Arman (1969)) that we can restrict ourselves to designs concentrating mass at $m = k+3$ points; two of which are the endpoints and one in each of the intervals $(\xi_i,\ \xi_{i+1})$, $i=0,1,\ldots,k$. For $k=1$ the following table indicates the optimal points of allocation for the D-optimal design. The design is on -1, x_1, x_2, 1 and the weights are all equal.

ξ	0	.2	.4	.6	.8
x_1	-.390	-.312	-.239	-.166	-.089
x_2	+.390	+.476	.573	.687	.825

Further calculations show that x_1 is always less than zero and approaches zero as $\xi \to 1$.

For $k=1$ the I-optimal design for the uniform measure is given below. The design is on points $x_0 = -1$, x_1, x_2, $x_3 = 1$ with weights p_0, p_1, p_2, p_3.

ξ	x_1	x_2	p_0	p_1	p_2	p_3
0	-.400	-.400	.164	.336	.336	.164
.2	-.325	.481	.176	356	317	.151
.4	-.253	.574	.187	.378	.298	.137
.6	-.180	.684	.200	.403	.280	.117
.8	-.099	.822	.217	.435	.260	.088

Generally speaking the points x_1 and x_2 are about the same. For ξ less than about .5 they shift slightly towards the endpoints ± 1. The weights p_i however on the interior points x_1 and x_2 become considerably heavier than on the endpoints.

The results for the quadratic case using $k \geq 1$ are somewhat similar. For example if $\xi_1 = -.3$ and $\xi_2 = .3$ the D-optimal design has equal mass on -1, -.569, 0, 569, 1.

For the cubic we take

$$1,\ x,\ x^2,\ x^3,\ (x-\xi)_+^3 \qquad -1 \leq x \leq 1$$

and consider only the D-optimal design. Our computations again show that the number of points used in the D-optimal design is the same as the number of regression functions, in this case m=5. (This has not been proven and does not follow as in the quadratic case). Here we label the points -1, x_1, x_2, x_3, 1.

ξ	0	.2	.4	.6	.8
x_1	-.629	-.584	-.547	-.515	-.484
x_2	0	.104	.193	.273	.352
x_3	.629	.679	.733	.796	.877

We note, as expected, that the middle points x_2 is not equal to ξ.

§6. Computational Procedures. Let the design μ concentrate mass $p_0, p_1, \ldots, p_{r+1}$ at the points $x_0, x_1, \ldots, x_{r+1}$. We shall restrict attention again to polynomial or polynomial spline regression on $[-1,1]$. In this case we know that the optimal designs have $x_0 = -1$ and $x_{r+1} = +1$. We wish to maximize $|M|$ or minimize $\operatorname{tr} M^{-1} C$ with respect to $x_1, x_2, \ldots, x_r$ and $p_0, p_1, \ldots, p_{r+1}$ subject to the condition $\Sigma p_i = 1$. If $f(x) = (f_1(x), \ldots, f_m(x))$ and $g(x)$ denotes the corresponding vector of derivatives one can easily show that for minimizing $\operatorname{tr} M^{-1} C$ the x_i and p_i are solutions of the equations

$$\begin{aligned} & f'(x_i)\, M^{-1}(\mu)\, C\, M^{-1}(\mu)\, f(x_i) = \lambda \qquad i = 0, \ldots, r+1 \\ (6.1) \qquad & f'(x_i)\, M^{-1}(\mu)\, C\, M^{-1}(\mu)\, g(x_i) = 0 \qquad i = 1, \ldots, r \\ & \Sigma p_i - 1 = 0 \end{aligned}$$

Here λ is an unknown multiplier equal to $\lambda = \operatorname{tr} M^{-1}(\mu^*)\, C$. For maximizing $|M|$ we have the additional information that $f'(x)\, M^{-1}(\mu^*)\, f(x) \leq m$ and the corresponding equations are

$$\begin{aligned} & f'(x_i)\, M^{-1}(\mu)\, f(x_i) = m \qquad i = 0, 1, \ldots, r+1 \\ (6.2) \qquad & f'(x_i)\, M^{-1}(\mu)\, g(x_i) = 0 \qquad i = 1, \ldots, r \end{aligned}$$

These equations are for the most part impossible to solve by hand, however standard computer routines seem to give answers fairly quickly at least for r up to four or five.

In obtaining optimal designs the main difficulty is in choosing the points $x_1, \ldots, x_m$. The weights $p_1, \ldots, p_m$ are usually easy to determine. Recently Fedorov (1971) (see also Wynn (1970)) has investigated the problem of determining the optimal μ^* by iterative methods. We shall indicate the procedure for the D-optimal designs. At the kth stage we have a design μ_k. We determine x_k so that

$$f'(x_k)\ M^{-1}(\mu_k)\ f(x_k) = \sup_{x \in X}\ f'(x)\ M^{-1}(\mu_k)\ f(x)$$

The new μ_{k+1} is then

$$\mu_{k+1} = \alpha_k\ \mu_k + (1-\alpha_k)\mu_{x_k}$$

where μ_{x_k} concentrates all of its mass at x_k. The scalar α_k is given by

$$\alpha_k = \frac{d_k - m}{[d_k - 1]\ m} \quad \text{where} \quad d_k = f'(x_k)\ M^{-1}(\mu_k)\ f(x_k)\ .$$

The procedure converges rather slowly at the rate k^{-1}. However in carrying out the procedure one can proceed from $M^{-1}(\mu_k)$ to $M^{-1}(\mu_{k+1})$ without recalculating this $m \times m$ inverse. Moreover the problem is reduced essentially to finding $\sup_{x \in X} f'(x)\ M^{-1}(\mu_k)\ f(x)$.

References

Fedorov, V. V. (1971) Theory of Optimal Experiments (to appear), Academic Press, N. Y.

Guest, P. G. (1958) The spacing of observations in polynomial regression, Ann. Math. Statist. Vol. 29, p. 294-299.

Karlin, S. and Studden, W. J. (1966) Optimal experimental designs, Ann. Math. Statist. 37, p. 783-815.

Kiefer, J. (1959) Optimum experimental designs, J. Roy. Statist. Soc. Ser. B. 21, p. 273-319.

Kiefer, J. and Wolfowitz, J. (1960) The equivalences of two extremum problems, Can. J. Math. 12, p. 363-366.

Rice, J. R. (1968) The Approximation of Functions, Vol. 2, Addison-Wesley, Reading, Mass.

Rivlin, T. J. (1969) An Introduction to the Approximation of Functions, Blaisdell Pub. Co., Waltham, Mass.

Schoenberg, I. J. (1969) Approximations with Special Emphasis on Spline Functions, Academic Press, N. Y.

Studden, W. J. and D. J. Van Arman (1969) Admissible designs for polynomial spline regression, Ann. Math. Statist. 40, 1557-1569.

Wynn, H. P. (1970) The Sequential generation of D-optimal experimental designs, Ann. Math. Statist., Vol. 41, p. 1655-1664.