The Performance of Equidistributed Sequences in Nonparametric Regression Based on a Quasi Least Squares Method

Ewaryst Rafajłowicz
Institute of Engineering Cybernetics
Technical University of Wrocław

and

Rainer Schwabe
Institute of Mathematics I
Free University, Berlin

Abstract

In this paper a method of generating experimental designs for estimating a response function is proposed. This method is based on results for multivariate integration by means of quasi Monte Carlo methods. We start with sequences of (deterministic) uniformly distributed points. From these we select such sequences, which assure the best possible convergence rate of a quasi least squares nonparametric regression function estimator (in the integrated mean squared error sense). Such sequences can be generated by vectors for which the components are algebraically independent. Finally, in the class of those vectors we search for generators, which are well suited already for estimation by a moderate number of observations. The proposed design sequences are not necessarily optimal in the classical sense, which requires a model to be completely specified, but useful when the model is not completely specified and various sets of model spanning functions are to be considered.

1 Introduction

The aim of the present paper is to propose a method for generating experimental designs, which is based on the application of equidistributed (or uniform) sequences. The method is intended to be used in nonparametric regression settings when the model is expressed by an orthogonal expansion. But it is apparently also applicable to a broader class of problems which include cases when the regression model specification of the response function is incomplete or not reliable.

Parallel to that we investigate an estimator which is called a *quasi least squares estimator* or an *orthogonal series estimator*. Estimators of probability density functions based on orthogonal expansions were proposed by Schwartz [30] (see [39] and [34] for further discussions)

Estimators based on orthogonal expansions of the response function were also investigated in a number of papers, both in the fixed design as well as in the random design situation (see [10], [23], [7] for more recent contributions). However, to the best knowledge of the authors, the possibilities of applying equidistributed sequences in nonparametric regression and the corresponding consequences have not been investigated for the choice of an estimation method, so far. For the definition and fundamental properties of sequences, which are uniformly distributed in the deterministic sense the reader is referred to [5], [14] and [19].

Most of the existing theory on optimal design of experiments is based on the assumption that a linear regression model under study is correctly specified (see for example [9], [21], [22]). This fundamental assumption implies that the existing design criteria (as those related to D–, A– or G–optimality) take only the variance of the estimator for the parameters of the response function into account. When considering regression in a nonparametric setting, we are in a quite different position, since – in addition to the variance – there is also a bias term present. Thus, we have to measure the accuracy of estimation in terms of the mean squared error, which incorporates both the bias and the variance.

In the field of nonparametric regression (see [6], [16], [38], [13], [26] for monographs on this topic and the bibliography cited therein) one has to distinguish between two different statements of the problem: First, the so–called fixed design case, when the choice of the input sequence is at the disposal of the experimenter and, second, the random design case, when input–output pairs can only be observed while the inputs are assummed to be random. In this paper we focus on the fixed design case and we emphasize that only few papers are devoted to the question of choosing "good" input sequences in a nonparametric setting. Attempts to the choice of design points for kernel type nonparametric regeression estimators can be found in [17]. In [23], possibilities of using points provided by quadrature formulas for numerical integration in nonparametric estimation are discussed. In [28] experimental design problems are considered which are robust against an incomplete specification of the model in the context of parametric regression. A discussion ot these ideas in the functional setup of Fourier regression models can be found in [2]. Recently, experimental design problems for a regression estimation by the local least squares were considered in [18] and [3].

2 Statement of the problem

Let $f : R^d \to R$, $d \geq 1$ be an unknown response function which is to be estimated from the observations

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $(\epsilon_i)_{i=1}^n$ is a sequence of zero mean, homoscedastic and uncorrelated random variables, representing errors in the observations. Concerning the response function f, assumptions will be imposed only on the smoothness like continuity, Lipschitz continuity etc, later on.

The deterministic sequence $(x_i)_{i=1}^n$ of settings is assumed to be equidistributed (uniformly distributed) in $X = [0,1]^d$ (see, e. g. [14]), i. e., for every continuous function h it is assumed that

$$n^{-1} \sum_{i=1}^{n} h(x_i) \longrightarrow \int_X h(x)dx, \quad \text{as} \quad n \to \infty. \tag{2}$$

In particular, we consider *open type equidistributed sequences* in which the settings x_i are generated sequentially and which are well suited for nonparametric estimation.

The related *closed type equidistributed sequences* which have the form of lattice points (see [32] for a definition) were found to be optimal experimental designs in a trigonometric regression setup in [2] and [25].

The problem of nonparametric estimation of f from $((x_i, y_i))_{i=1}^n$ is to find a sequence of functions, f_n say, such that $f_n \to f$ in a suitable sense. As a measure of performance for the estimation accuracy we consider the integrated mean squared error

$$IMSE(f_n, f) = E \int_X (f_n(x) - f(x))^2 \, dx. \tag{3}$$

The present objective is to discuss the influence of the choice of $(x_i)_{i=1}^n$ on the rate of covergence of the integrated mean squared error $IMSE(f_n, f)$. Note that in the sequel $f_n = \hat{f}_n$ will be a random function which is chosen on the basis of the observations $(y_i)_{i=1}^n$. Accordingly the integrated mean squared error decomposes into a variance term for the regressor f_n and into a bias term due to approximation errors according to

$$IMSE(f_n, f) = \int_X \mathrm{Var}(f_n(x))\, dx + \int_X (Ef_n(x) - f(x))^2 \, dx. \tag{4}$$

The application of equidistributed sequences as experimental designs forces us to investigate asymptotic properties of the *quasi least squares estimator* which is based on an orthogonal expansion, but is, simultaneously, close to the classical least squares estimator. The motivations for the consideration of quasi least squares estimators are the following:

1. An extension of the classical least squares estimator to the nonparametric case (see [24], [4]) has theoretical advantages, but the necessity of inverting large matrices reduces its practical applicability when the dimension $d = \dim(x_i)$ of the input variable is larger than two or three.

2. When an appropriately chosen equidistributed sequence is applied together with an orthonormal sequence of regression approximation, then the matrix of normal equations for the classical least squares estimates has diagonal entries close to n^{-1} and off–diagonal entries close to zero.

The second fact motivates our choice of the shape of the estimator $\hat{f}_n$ which will be defined in the next section.

A further reason why the estimator $\hat{f}_n$ requires new investigations relies on the fact that the estimator proposed here has a form similar to estimators used in the random design case although it is used for a deterministic sequence $(x_i)_{i=1}^n$, here.

The remainder of the paper is organized as follows. In the next section the estimator $\hat{f}_n$ is defined and sufficient conditions are given for its convergence in the integrated mean squared error sense. Then, it is proved that the convergence rate of the estimator $\hat{f}_n$ attains its lower bound for a large class of equidistributed sequences $(x_i)_{i=1}^n$ under some suitable smoothness assumptions. The above mentioned results are of a general type and we apply them to estimators based on Legendre polynomials and, alternatively, on the system of trigonometric functions. Finally, a method of selecting "good" equidistributed sequences for a finite number of observations is proposed.

3 The quasi least squares estimators and their consistency

Let $v_1, v_2, \ldots$ be a complete sequence of orthonormal functions in the space $L_2(X)$ of square integrable functions, i. e. $\int v_k(x)^2\, dx = 1$ and $\int v_k(x) v_\ell(x)\, dx = 0$, for all $k \neq \ell$. Denote by $a_k = \int_X f(x) v_k(x)\, dx$, $k = 1, 2, \ldots$, the coefficients of the response function f expanded in the basis (v_k). We define the estimator $\hat{f}_n$ of f as follows:

$$\hat{f}_n(x) = \sum_{k=1}^{N} \hat{a}_{kn} v_k(x), \tag{5}$$

where

$$\hat{a}_{kn} = \frac{1}{n} \sum_{i=1}^{n} y_i v_k(x_i), \quad k = 1, 2, \ldots \tag{6}$$

is the estimator of the regression coefficients a_k. The estimators $\hat{a}_{kn}$ and $\hat{f}_n$ will be called *quasi least squares estimators* as they are obtained in analogy to the classical *least squares estimators* by replacing the difficult to compute inverse of the normalized Fisher information matrix $n^{-1}\sum_{i=1}^{n}\bar{v}_N(x_i)\bar{v}_N^T(x_i)$ by the $N \times N$ identity matrix, for a finite basis $\bar{v}_N = (v_1, v_2, \ldots, v_N)^T$.

In (5) the degree N of regression plays the role of the smoothing parameter. In the asymptotic considerations the approximation order $N = N(n)$ will depend on the number n of observations. Typically, the degree N will increase to infinity substantially slower than the sample size n.

Remark 1 *For further readings on techniques for choosing smoothing parameters in nonparametric regression we refer to the monographs mentioned in the introduction. This aspect is beyond the scope of this paper.*

Let the equidistributed sequence $(x_i)_{i=1}^n$ be chosen in such a way that the following condition holds for all functions h in a prespecified subset H of the set $C(X)$ of continuous functions on X:

(EQ) There exists a constant $c = c(H) > 0$ such that $|n^{-1}\sum_{i=1}^{n} h(x_i) - \int_X h(x)\,dx| \leq c\phi_n$ for all $h \in H$, where $\phi_n = n^{-1/2-\delta}$ for some $\delta = \delta(H) > 0$.

In (EQ) the constants c and δ depend only on the subset H but not on n. The condition (EQ) is apparently rather mild, since even i. i. d. random sequences give an order $n^{-1/2}$ of approximation, while equidistributed sequences are known to provide much faster convergence rates of the integration error (see [5]). The condition (EQ) holds, for example, if, in case $d = 1$, $(x_i)_{i=1}^n$ is the Van der Corput sequence and H is the class of functions having bounded variation (see [14], page 127). In the multidimensional case, the condition (EQ) is satisfied for functions of bounded variation (in the muldimiensional sense) and for "good" equidistributed sequences for which the definition and construction is given in [14], since for such sequences the distance $|n^{-1}\sum_{i=1}^{n} h(x_i) - \int_X h(x)\,dx|$ is of the order $\log(n)^\beta/n$, for some suitable constant $\beta > 0$ which may depend on the dimension d of x and on the properties of the equidistributed sequence $(x_i)_{i=1}^n$.

Theorem 1 *Let $(x_i)_{i=1}^n$ and $H \subset C(X)$ be such that (EQ) holds. Assume that $f \in H$, $fv_k \in H$ and $v_k^2 \in H$, $k = 1, 2 \ldots$. Let the order $N(n)$ of approximation satisfy*

$$N(n) \to \infty \quad \text{and} \quad N(n)/n \to 0 \quad \text{as} \quad n \to \infty. \tag{7}$$

Then $IMSE(\hat{f}_n, f) \to 0$ as $n \to \infty$.

Proof. Since (EQ) holds for $\{v_k^2,\ k = 1, 2, \ldots\}$ then $\mathrm{Var}(\hat{a}_{kn}) = O(n^{-1})$ and

$$n\mathrm{Var}(\hat{a}_{kn}) = \sigma^2 + O(\phi_n) \qquad k = 1, 2, \ldots \tag{8}$$

what can be verified by direct culculations.

From the orthonormality and completeness of the sequence (v_k) it follows that the systematical error $E\hat{f}_n - f$ and, hence, by (4), the integrated mean squared error $IMSE(\hat{f}_n, f)$ can be decomposed according to

$$IMSE(\hat{f}_n, f) = W_n + B_n^2 + R(N, f) \tag{9}$$

into a variance term

$$W_n = \sum_{k=1}^{N} \mathrm{Var}(\hat{a}_{kn}), \tag{10}$$

a bias term

$$B_n^2 = \sum_{k=1}^{N} (E\hat{a}_{kn} - a_k)^2 \tag{11}$$

and an approximation error

$$R(N, f) = \sum_{k=N+1}^{\infty} a_k^2, \tag{12}$$

where, for brevity, $N = N(n)$. Thus, from (8) we have

$$W_n = O(N/n) \tag{13}$$

From the definition of $\hat{a}_{kn}$ we obtain

$$(E\hat{a}_{kn} - a_k)^2 = \left(n^{-1} \sum_{i=1}^{n} f(x_i) v_k(x_i) - a_k \right)^2 \tag{14}$$

which is of the order $O(\phi_n^2)$, uniformly in k, in view of (EQ). Thus, from (9) we get

$$IMSE(\hat{f}_n, f) \leq c_1 \frac{N}{n} + c_2 N \phi_n^2 + R(N, f), \tag{15}$$

for some suitable constants $c_1 > 0$ and $c_2 > 0$. This finishes the proof by invoking (7), (EQ) and the completeness of (v_k) in $L_2(X)$. •

Remark 2 *Let us note that (13) is obtained without assuming that the sequence (v_k) is uniformly bounded. Thus, (13) holds, e. g., for the Legendre polynomials.*

As it will be shown next, the above theorem holds for a wide class of equidistributed sequences, orthonormal sequences (v_k), classes of unknown functions H and sequences $N(n)$. Imposing slightly more restrictive conditions on the above components, we can obtain more detailed information on the rate of convergence of $IMSE(\hat{f}_n, f)$, revealing the role of equidistributed sequences as experimental designs.

4 Asymptotic optimality of equidistributed experimental design sequences – general results

To obtain an upper bound for the covergence rate of $IMSE(\hat{f}_n, f)$, we have to impose additional smoothness conditions on the response f. The following assumption is sufficiently general for a given sequence (v_k) of basis functions and a given subset $H \subset C(X)$ of response functions of interest.

(SM) There exist a constant $\mu = \mu(H) > 0$ such that for all $f \in H$ we have $R(N, f) \leq c(\mu, f)/N^{2\mu/d}$ for some $c(\mu, f) > 0$, uniformly in N.

From (15) and (SM) it follows

$$IMSE(\hat{f}_n, f) \leq c_1 \frac{N}{n} + c_2 \frac{N}{n^{1+2\delta}} + \frac{c(\mu, f)}{N^{2\mu/d}}. \tag{16}$$

Let us note that the second term in (16) converges faster to zero than the first one. Hence, $IMSE(\hat{f}_n, f) = O(\frac{N}{n} + N^{-2\mu/d})$ and the rate of convergence becomes optimal if $N(n)$ grows approximately like $n^{d/(d+2\mu)}$ which proves

Corollary 1 *Let the assumptions of Theorem 1 and, additionally, (SM) be satisfied. Let $N(n)$ fulfil the growth condition $\gamma_1 n^{1/(1+2\mu/d)} \leq N(n) \leq \gamma_2 n^{1/(1+2\mu/d)}$ for some constants $0 < \gamma_1 \leq \gamma_2$ Then $IMSE(\hat{f}_n, f) = O(n^{-2\mu/(d+2\mu)})$.*

As we shall see, the above convergence rate is sharp in the sense that it is attained for some function classes.

Let $W^\mu(X)$ denote the Sobolev space of functions on X, where $\mu = p + \alpha$, $0 \leq \alpha < 1$, i. e. $w \in W^\mu$ is p times differentiable ($p = 0$ means continuity) and its pth derivative is Lipschitz continuous with exponent α. When $d > 1$, then the above requirements are assumed to hold for all corresponding partial derivatives.

It was proved by Stone [33] that the best rate of convergence in the integrated mean squared error is $O(n^{2\mu/(d+2\mu)})$, for every estimator $\tilde{f}_n$ of $f \in W^\mu(X)$ provided that f can be any function from $W^\mu(X)$ (see also [36], [37], [11], [7], [12], [27], [8] for further references on convergence rates). As the conditions (EQ) and (SM) hold for $H = W^\mu(X)$ we can draw the following conclusion from Corollary 1:

Corollary 2 *Assume that $H = W^\mu(X)$ and let the assumptions of Corollary 1 hold. Then, uniformly in $f \in W^\mu(X)$, the sequence $\hat{f}_n$ converges to f with the best possible rate in the integrated mean squared error sense.*

All the above results are kept as general as possible. As we shall see in the next section, they allow to obtain convergence rates relatively easily for concrete classes of functions and design sequences.

5 Experimental designs based on equidistributed sequences generated by algebraic numbers

A design sequence $(x_i)_{i=1}^n$ is said to be asymptotically optimal for estimating $f \in W^\mu(X)$ by an estimator $\tilde{f}_n$, if $\tilde{f}_n$ is consistent and attains the optimal convergence rate in the integrated mean squared error sense.

In the present section our aim is to point out asymptotically optimal sequences, when for the estimator $\hat{f}_n$ either the Legendre polynomials or the system of trigonometric functions are used.

Remark 3 *To verify condition (SM) one can utilize either of the following approaches:*

a) *Relate directly some a–priori knowledge on the smoothness of f to the rate of decay of $a_k = \int_X f(x)v_k(x)\,dx$ and evaluate $\sum_{k=N+1}^{\infty} a_k^2$.*

b) *Denote by Γ_N the subspace spanned by the linear combinations of $v_1, v_2, \ldots, v_N$. Use the fact that*

$$R(N, f) = \inf_{f \in \Gamma_N} ||f - f_N||_2^2 \tag{17}$$

and apply available results on the best approximation in the L_2–norm $||\cdot||_2$.

c) *When b) is not applicable, (17) can be bounded by*

$$d_N(f) = \sup_{g_N \in \Gamma_N} ||f - g_N||_\infty \tag{18}$$

(conditions for the existence of the best approximation in the supremum norm $||\cdot||_\infty$ can be found for various systems $v_1, v_2, \ldots$, e. g., in [15], [35]). Then, for $X = [0,1]^d$, it follows from (17) that

$$R(N, f) = \inf_{f_N \in \Gamma_N} ||f - f_N||_2^2 \le \int_X (f(x) - g_N(x))^2 dx \le d_N^2(f). \tag{19}$$

Now, one can use known bounds on $d_N(f)$, e. g., Jackson's theorems for polynomial and trigonometric approximations.

5.1 Estimation on the basis of Legendre polynomials

In the one–dimensional case, $X = [0,1]$, assume that $f \in W^\mu(X)$, $0 < \mu \le 1$. As basis functions $v_1, v_2, \ldots$ we take the Legendre polynomials, which are shifted and rescaled to be orthonormal over $[0,1]$.

From Jackson's theorem (see [35]) it follows that for $f \in W^\mu(X)$ we have $d_N(f) = O(N^{-\mu})$. Thus, according to Remark 3 c) above, condition (SM) holds.

Experimental design. As design points we consider

$$x_i = \{\{i\theta\}\}, \qquad i = 1, 2, \ldots, \tag{20}$$

where θ is an irrational number and $\{\{t\}\}$ denotes the fractional part of t. Additionally, assume θ is an algebraic number. From [14] (Theorem 5.4, page 146) it follows that for any $h \in W^{\mu}(X)$ we have

$$\left| n^{-1} \sum_{i=1}^{n} h(x_i) - \int_0^1 h(x)\, dx \right| = O((D_n^\star)^{\mu}), \tag{21}$$

where $D_n^\star$ is the discrepancy of the sequence (x_i). When θ is an algebraic number the discrepancy $D_n^\star$ is of the order $O(N^{-1+\epsilon})$ for every $\epsilon > 0$. From this fact and (21) it follows that condition (EQ) is satisfied both for $f \in W^{\mu}(X)$, $0 < \mu \leq 1$, and, hence, for the functions fv_k and $v_k v_j$, $k, j = 1, 2, \ldots$, since the v_k are continuously differentiable. Thus, from Corollaries 1 and 2 we obtain:

Corollary 3 *Let $f \in W^{\mu}(X)$, $0 < \mu \leq 1$ be estimated by $\hat{f}_n$ based on the Legendre polynomials orthonormalized to $[0,1]$. Let the observations be taken at the points $x_i = \{\{i\theta\}\}$, $i = 1, 2, \ldots n$ with θ being an irrational number of the algebraic type. Assume that $N(n)$ satisfies the growth condition of Corollary 1. Then, $\hat{f}_n$ attains the best convergence rate in in the integrated mean squared error sense and $\{\{i\theta\}\}$ is an asymptotically optimal design sequence.*

We mention that the sequence $x_i = \{\{i\theta\}\}$ is of the open type, i. e., adding new points does not influence the positions of the preceding points in the sequence. Some particular examples of θ will be given in the next section.

5.2 Estimating a multivariate response function using a trigonometric system

As a sequence of basis functions $v_1, v_2, \ldots, v_N$ we take all the trigonometric functions of the form $\{1, \sqrt{2}\sin(2\pi t), \sqrt{2}\cos(2\pi t), \ldots \sqrt{2}\sin(2\pi M t), \sqrt{2}\cos(2\pi M t)\}$, $t \in [0,1]$, with respect to each variable in $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})^T$, including also all the interactions of the form $\sqrt{2}\sin(2\pi(m^{(1)}x^{(1)} \pm \ldots \pm m^{(d)}x^{(d)}))$ and $\sqrt{2}\cos(2\pi(m^{(1)}x^{(1)} \pm \ldots \pm m^{(d)}x^{(d)}))$ where $m^{(1)}, \ldots, m^{(d)} = 0, 1, \ldots, M$. Thus, $\Gamma_N^{trig} = span[v_1(x), v_2(x), \ldots, v_N(x)]$ with $N = (2M+1)^d$.

From [20] (Theorem 5.5.7) it follows that for $f \in W^{\mu}(X)$ we have $d_N(f) = O(N^{-\mu/d})$ provided that the periodicity assumption specified below also holds. Thus, condition (SM) is fulfilled in this case.

Design sequence. We generate a sequence of experimental points as follows

$$x_i = \{\{i \cdot \theta\}\}, \qquad i = 1, 2, \ldots, \tag{22}$$

where $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(d)})^T$ is a vector of distinct irrational numbers satisfying linear independence over the rationals, i. e., for every set of rational numbers $\alpha_0, \alpha_1, \ldots, \alpha_d$ not all of them being 0 we have

$$\alpha_0 + \sum_{j=1}^{d} \alpha_j \theta^{(j)} \neq 0. \tag{23}$$

Furthermore, it is required that each component $\theta^{(j)}$ of θ is an algebraic number, $j = 1, 2, \ldots, d$. We will call θ the generator of a sequence of experimental designs. Here, the fractional part $\{\{\cdot\}\}$ is applied to each component of the vector.

Now, our aim is to show that the sequence generated by equation (22) satisfies condition (EQ) for an appropriately chosen set of functions $H^\mu \subset W^\mu(X)$, $X = [0,1]^d$. For simplicity assume that $\mu > 1$ is an integer. Define H^μ as the class of all one–periodic functions on R^d for which the partial derivatives

$$\frac{\partial^{|q|} f(x)}{\partial(x^{(1)})^{q_1} \ldots \partial(x^{(d)})^{q_d}}, \quad 0 \leq q_j \leq \mu,$$

are continuous on X, where the norm $|q|$ of the multi–index q is defined as $|q| = q_1 + \ldots + q_d$. For $m = (m^{(1)}, m^{(2)}, \ldots, m^{(d)})^T$ denote by a_m the Fourier coefficient of f corresponding to the frequencies $m^{(1)}, m^{(2)}, \ldots, m^{(d)}$. From the well–known result of Zaremba (see [32], page 71) it follows that for each $f \in H^\mu$ there exists $c(f) > 0$ such that

$$|a_h| \leq \frac{c(f)}{(\bar{m}^{(1)} \cdot \bar{m}^{(2)} \cdot \ldots \cdot \bar{m}^{(d)})^\mu}, \tag{24}$$

where $\bar{m}^{(j)} = \max(1, m^{(j)})$. Now, we are in the position to invoke the result from [5] (page 402) that, in case $\mu > 2$, there is a suitable constant $c_\mu > 0$ such that

$$\left| n^{-1} \sum_{i=1}^{n} h(\{\{i \cdot \theta\}\}) - \int_X h(x)\, dx \right| \leq \frac{c_\mu}{n} \tag{25}$$

for every $h \in H^\mu$. Thus, from (25) it follows that for H^μ and design points generated according to (22) the condition (EQ) holds.

Under various conditions imposed on h error bounds similar to (25) hold with a right hand side of the form $c \log(n)^\beta / n$, for some constant $\beta > 0$ (see [32], page 21 and 28, as well, the bibliography cited therein). Obviously, the condition (EQ) is satisfied also in such cases.

Summarizing the above considerations we obtain:

Corollary 4 *Let $f \in H^\mu = W^\mu(X)$, $\mu \in \{2, 3, 4 \ldots\}$ be estimated by $\hat{f}_n$ based on Γ_N^{trig} and let $N(n)$ fulfil the growth condition of Corollary 1. If $\theta^{(j)}$, $j = 1, 2, \ldots d$ are algebraic numbers, linearly independent over the rationals, then (22) is asymptotically optimal and $\hat{f}_n$ attains the best convergence rate in the integrated mean squared error sense, uniformly on $H^\mu = W^\mu(X)$.*

Remark 4 **a)** *Corollary 4 does not ensure that the best convergence rate is attained for every periodic $f \in H^\mu$, since H^μ contains only functions with smooth mixed partial derivatives. One can extend the class of functions for which (25) holds to the class of functions with bounded variation in the sense of Hardy and Krauze (see [14]).*

b) *To ensure that for $N(n)$ the conditions of Corollary 1 hold the degree M of the trigonometric fit has to depend on n. In particular, it is sufficient to choose $M = M(n)$ according to*

$$c_1 n^{1/(2\mu+d)} \leq M(n) \leq c_2 n^{1/(2\mu+d)}, \tag{26}$$

for certain constants $0 < c_1 < c_2$.

c) *When n is of moderate size a proper choice of θ can be assured by the algorithm, which will be proposed in the next section. The case when components of g are rational numbers is extensively discussed in [25], where necessary and sufficient conditions for the orthogonality of the design (22) in estimating a trigonometric regression by the classical least squares are discussed. (Note also that in a general case $\hat{f}_n$ coincides with the classical least squares estimator, when x_i, $i = 1, 2, \ldots, n$ is an orthogonal design.)*

d) *When there is no a–priori knowledge concerning the smoothness of f (expressed in μ), then $M(n)$ cannot be chosen according to (26), but under– or over–estimation of μ influences only the convergence rate, while $\hat{f}_n$ is still convergent (see Theorem 1). A data driven choice of M (or N) is beyond the scope of this paper.*

We also emphasize that the choice of θ as the generator of "good" design points does not depend on the smoothness parameter μ.

6 Selecting generators of experimental designs

From Corollaries 3 and 4 we know that sequences of designs $x_i = \{\{i \cdot \theta\}\}$, $i = 1, 2, \ldots$ are asymptotically optimal if the components of θ are irrational numbers of the algebraic type and (23) holds. This essentially reduces the area of search for "good" design generators θ, but we still have some freedom. Our aim in this section is to utilize this freedom to discuss the choice of a suitable θ, which would be satisfactory for finite n.

6.1 How to measure the quality of a design?

As it was mentioned in the introduction, the classical criterions of experimental designs, such as D-optimality, can not be directly applied in the nonparametric setting for the following reasons:

1) In the classical regression setting the exact knowledge of a model structure is assumed. For linear models this implies that only variances of estimators are considered as sources of errors.

2) The consequence of 1) is that designs, which are optimal in the classical setting are frequently clustered in a relatively small number of design points.

In view of Theorem 1 it would be desirable to consider $IMSE(\hat{f}_n, f)$ as an experimental design criterion. We cannot, however, follow this way directly, since it is not possible to calculate the bias of $\hat{f}_n$. From Corollary 1 it follows that the bias approaches zero sufficiently fast, but to ensure this convergence we have to impose a side condition on the uniformity of the sequence of designs.

Let $\frac{1}{n}\sum_{i=1}^{n} \bar{v}_N(x_i)\bar{v}_N^T(x_i)$ be the normalized Fisher information matrix for estimating a response function spanned by $\bar{v}_N = (v_1, v_2, \ldots, v_N)^T$ when the design $(x_i)_{i=1}^{n}$ is used. As a common base for comparisons of design sequences for different values of n and N we take the ideal (generalized) uniform design, for which

$$tr\left[\left(\int_X \bar{v}_N(x)\bar{v}_N^T(x)dx\right)^{-1}\right] = N,$$

where tr denotes the trace of a matrix. Note that for the uniform design and the response f spanned by $\bar{v}_N$ the trace is proportional to the integrated mean squared error criterion. Thus, a design sequence x_i, $i = 1, 2, \ldots n$ is of a sufficient quality, if $q = \frac{1}{N} tr\left[\left(\frac{1}{n}\sum_{i=1}^{n} \bar{v}_N(x_i)\bar{v}_N^T(x_i)\right)^{-1}\right] - 1 \leq \delta$, where $\delta > 0$ is a given tolerance. For computational reasons we have to allow larger tolerances for larger values of N and n.

The main difference between this and the classical A-optimal design problem statement is in the additional requirement that the search is restricted to equidistributed sequences. This side condition assures that the bias term is small for sufficiently large n.

6.2 Sieve method of selecting "good" generators of experimental designs

We are not able to minimize $tr\left[\left(\frac{1}{n}\sum_{i=1}^{n} \bar{v}_N(x_i)\bar{v}_N^T(x_i)\right)^{-1}\right]$ over the whole class of equidistributed sequences, since it is too rich. Therefore, we further confine the search for generators to $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots \theta^{(d)})^T$ of the form

$$\theta^{(j)} = \sqrt{\nu_j}, \quad \nu_j = 2, \ldots, m, \quad j = 1, 2, \ldots d, \tag{27}$$

for some sufficiently large integer m, provided that the components $\theta^{(j)}$ are irrational numbers satisfying (23).

Denote by

$$q(\theta, n, N) = \frac{1}{N} tr\left[\left(\frac{1}{n}\sum_{i=1}^{n} \bar{v}_N(x_i)\bar{v}_N^T(x_i)\right)^{-1}\right] - 1 \tag{28}$$

the deviation of the design from uniformity when the design sequence $x_i = \{\{i \cdot \theta\}\}$, $i = 1, 2, \ldots n$ is used.

The minimization of $tr\left[\left(\int_X \bar{v}_N(x)\bar{v}_N^T(x)dx\right)^{-1}\right]$ and, hence, of the deviation $q(\theta, n, N)$ with respect to all possible sequences (27) is still a time consuming task. For this reason we propose to use the following algorithm for searching "good" (but not necessarily optimal) design generators θ. The algorithm is called a sieve method, since it is based on the following idea: select a number of "good" one–dimensional generators $\theta = \theta_1$, then form pairs of them and select "good" two–dimensional generators $\theta = \theta_2$, which are then extended to triples, $\theta = \theta_3$, and selection is repeated. This way of selection is not only advantageous by a large reduction of the computational burden, but also because of obtaining design sequence with uniformly distributed projections onto lower–dimensional hypercubes. The last feature is of particular importance for situations in which not all input variables are essential for fitting the model. It is worthwhile noting that the order N of approximation will also depend on the dimension d, $N = N(n, d)$.

Remark 5 *When presenting the algorithm we treat the sample size n, the degree of approximation N as fixed, but our aim is to obtain generators θ, which are useful for various values of these parameters. For this reason, the quality of a generator θ obtained for one setting will be checked for other values of n and N.*

The details of the proposed algorithm are as follows:

Sieve algorithm for selecting design generators

Step 0 Specify the range m of search being a sufficiently large positive integer. Denote by $d_{\max}$ the largest dimension $d \geq 1$, for which the search will be performed. Choose a set $(\delta_d)_{d=1}^{d_{\max}}$ of real numbers, $\delta_d > 0$, which are interpreted as bounds for acceptable values of (28) for given dimension d.

Step 1 Set $d = 1$ and select the finite set of one–dimensional generators $\theta^{(1)}$, which are of the form $\theta^{(1)} = \sqrt{\nu}$, $\nu = 2, \ldots, m$ and for which the corresponding sequence $x_i = \{\{i\sqrt{\nu}\}\}$, $i = 1, 2, \ldots, n$ satisfies $q(\theta_1, n, N) \leq \delta_1$. Denote by Θ_1 the set of all such generators.

Step 2 If $d < d_{max}$ set $d := d+1$ and go to Step 3, otherwise STOP the algorithm and provide the sets $\Theta_1, \Theta_2, \ldots, \Theta_{d_{\max}}$ as the result.

Step 3 Form the set of candidate generators, $\mathcal{G}_d$ say, as follows: $\mathcal{G}_d = \Theta_{d-1} \times \Theta_1$, where "$\times$" denotes the Cartesian product. For each $\theta_d \in \mathcal{G}_d$ calculate $q(\theta_d, n, N)$ and form the set Θ_d by selecting those θ_d for which $q(\theta_d, n, N) \leq \delta_d$ and condition (23) holds. Go to Step 2.

If while executing Step 1 or Step 3, it occurs that $\Theta_d = \emptyset$, then one has to increase m or increase the values δ_d and restart the algorithm, again.

6.3 Selected experimental design generators

Below we report the results of using the sieve algorithm of selecting design generators for regressions spanned by the Legendre polynomials. In addition to "good" generators we display technical parameters of the algorithm in order to ensure repeatability of the calculations. In all the results of computations reported below the value $m = 1000$ was used. The sieve algorithm was run for $d_{max} = 3$.

Table 1 contains generators selected by the sieve algorithm for estimating parameters in univariate regressions of orders $N = 3, 4, 5, 6$ for different values of n. Each entry of this table was selected with tolerance $\delta_1 = 0.05$. Note that $\sqrt{577}$ is present in most of the entries, what allows to recommend it as a generator, which is universal to some extent. The sets of generators contained in Table 1 were then treated as sets Θ_1 in computing design generators for $d = 2$ and $d = 3$, which are reported in Table 2 for various n. The following coding is used in this table: $N = k \times k$ means that the generators were selected for estimating a bivariate regression spanned by the tensor products of the univariate Legendre polynomials up to the degree k. Empty cells of the table correspond to the cases when the number of observations is too small for obtaining a design with suffieciently good quality.

Figures 1, 2, 3 and 4 present the best design found by the sieve algorithm for bivariate regressions of various degrees. Analogously, Fig. 5 presents the design for estimating a regression with three inputs together with its two dimensional projections, which can be interpreted as designs for the corresponding bivariate regressions.

Remark 6 *We do not provide generators for regression functions spanned by the trigonometric system described in Section 5.2, since the optimal design sequence (providing $q = 0$) is the design on the equidistant grid, the same, which is IMSE (and also D- , A-) optimal in the classical setting (see, e.g., [9], [21]). The main drawback of these designs is in the rapidly growing number of points comprising them, when we need a number of design points larger than the number of estimated parameters. In such cases it may be useful to apply the sieve algorithm to select good generators.*

Acknowledgement. A large part of this work was done during a stay of the first author at the Free University Berlin. He wishes to express his sincere thanks for the warm hospitality and stimulating discussions there and for financial support from the *Deutsche Forschungsgemeinschaft* and *The Council for Scientific Research of Poland*. A substantial part of the present work was done while the second author was visiting the University of Technology at Darmstadt. He wishes to express his sincere thanks for the hospitality he met there. His research was also partly supported by the grant Ku 719/2–2 of the *Deutsche Forschungsgemeinschaft* at the Free University, Berlin.

References

[1] Atkinson A. C. and Donev A. N. *Optimum Experimental Design.* Clarendon Press, Oxford, 1992.

[2] Bates R. A., Riccomagno E., Schwabe R. and Wynn H. P. Lattices and dual lattices in optimal experimental design for Fourier models. *Comput. Statist. Data Anal.* (in print).

[3] Cheng M., Hall P. and Titterington D. M. Optimal design for curve estimation by local linear smoothing. *Bernoulli*, 4:3–14, 1998.

[4] Cox D. D. Approximation of least squares regression on nested subspaces. *Ann. Statist.*, 16:713–732, 1988.

[5] Davis P. J. and Rabinowitz P. *Methods of Numerical Integration.* Academic Press, Orlando, 1984.

[6] Eubank R. L. *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York, 1988.

[7] Eubank R. L. and Speckman P. Convergence rates for trigonometric and polynomial–trigonometric regression estimators. *Statist. Probab. Letters*, 11:119–124, 1991.

[8] Fan J. Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, 21:196–216, 1993.

[9] Fedorov V. V. *Theory of Optimal Experiments.* Academic Press, New York, 1972.

[10] Greblicki W. and Pawlak M. Fourier and Hermite series estimates of regression functions. *Ann. Inst. Statist. Math.*, 37:443–454, 1985.

[11] Hall P. On the rate of convergence of orthogonal series density estimators. *J. R. Statist. Soc., Ser. B*, B48:115–122, 1986.

[12] Hall P. On convergence rates in nonparametric problems. *Int. Statist. Rev.*, 57:45–58, 1989.

[13] Härdle W. *Applied Nonparametric Regression.* Cambridge University Press, Boston, 1990.

[14] Kuipers L. and Niederreiter H. *Unfiorm Distribution of Sequences.* Wiley, New York, 1974.

[15] Laurent P. J. *Approximation et Optimization.* Hermann, Paris, 1972.

[16] Müller H.-G. *Nonparametric Regression Analysis of Longitudinal Data.* Springer, Berlin, 1988.

[17] Müller H.-G. Optimal designs for nonparametric kernel regression. *Statist. Probab. Letters*, 2:285–290, 1984.

[18] Müller W. G., Optimal design for local fitting. *J. Statist. Plann. Inference*, 55: 389–397, 1996.

[19] Niederreiter H. *Random Number Generation and Quasi–Monte Carlo Methods.* SIAM, Philadelphia, 1992.

[20] Nikolskij S. M. *Approximation of Multivariate Functions and Embedding Theorems*, (in Russian). Nauka, Moscow, 1977.

[21] Pázman A. *Foundations of Optimum Experimental Design.* Reidel, Dordrecht, 1986.

[22] Pukelsheim F. *Optimal Design of Experiments.* Wiley, New York, 1993.

[23] Rafajłowicz E. Nonparametric orthogonal series estimators of regression: a class attaining the optimal convergence rate in l_2. *Statist. Probab. Letters*, 5:219–224, 1987.

[24] Rafajłowicz E. Nonparametric least squares estimation of a regression function. *Statistics*, 19:349–358, 1988.

[25] Riccomagno E., Schwabe, R. and Wynn H. P. Lattice–based optimum design for Fourier regression. *Ann. Statist.*, 25:2313–2327, 1997.

[26] Rosenblatt M. *Stochastic Curve Estimation.* IMS, Hayward, 1991.

[27] Rutkowski L. and Rafajłowicz E. On optimal global rate of convergence of some nonparametric identification procedures. *IEEE Trans. Autom. Control*, AC–34:1089–1092, 1989.

[28] Schwabe R. Model robust experimental design in the presence of interactions: The orthogonal case. In *PROBASTAT'94, Proc. Int. Conf. Math. Statist., Smolenice, 1994, Tatra Mountains Math. Publ.*, 7:97–104, 1996.

[29] Schwabe R. *Optimum Designs for Multi-Factor Models.* Lecture Notes in Statistics, Vol. 113, Springer, New York, 1996.

[30] Schwartz S. C. Estimation of probability density by an orthogonal series. *Ann. Math. Statist.*, 38:1261–1265, 1967.

[31] Silvey S. D. *Optimal Design.* Chapman and Hall, London, 1980.

[32] Sloan I. H. and Joe S. *Lattice Methods for Multiple Integration.* Clarendon Press, Oxford, 1994.

[33] Stone C. J. Optimal global rate of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053, 1982.

[34] Tarter M. E. and Kronmal R. A. On multivariate density estimates based on orthogonal expansions. *Ann. Math. Statist.*, 41:718–722, 1970.

[35] Timan A. F. *Theory of Approximation of Functions of a Real Variable.* Pergamon, Oxford, 1963.

[36] Wagner T. J. The rate of convergence of an algorithm for recovering functions from noisy measurements taken at randomly selected points. *IEEE Trans. Sys. Sci. Cybern.*, SSC–4:151–154, 1968.

[37] Wahba G. Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statist.*, 3:15–29, 1975.

[38] Wahba G. *Spline Models for Observational Data.* SIAM, Philadelphia, 1990.

[39] Watson G. S. Density estimation by orthogonal series. *Ann. Math. Statist.*, 40:1496–1498, 1969.

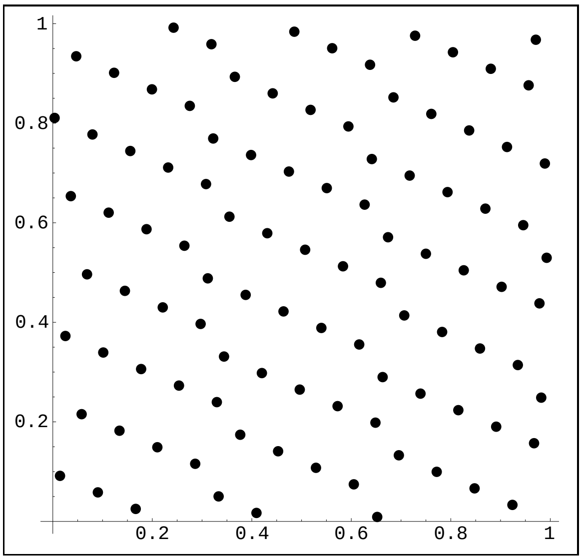

Figure 1: Experimental design for bivariate regression spanned by $N = 4 \times 4$ Legendre polynomials; design generated by $\theta = (\sqrt{235}, \sqrt{742})$, for $n = 100$ points; $q = 0.03$

	n=30	n=50	n=100	n=250
N=3	$\sqrt{197}, \sqrt{678}, \sqrt{731}, \sqrt{786}, \sqrt{843}$	$\sqrt{380}, \sqrt{577}, \sqrt{812}, \sqrt{870}, \sqrt{901}$	$\sqrt{197}, \sqrt{678}, \sqrt{731}, \sqrt{786}, \sqrt{843}$	$\sqrt{197}, \sqrt{678}, \sqrt{731}, \sqrt{786}, \sqrt{843}$
N=4	$\sqrt{197}, \sqrt{678}, \sqrt{731}, \sqrt{786}, \sqrt{843}$	$\sqrt{485}, \sqrt{530}, \sqrt{575}, \sqrt{577}, \sqrt{765}$	$\sqrt{235}, \sqrt{257}, \sqrt{552}, \sqrt{577}, \sqrt{822}$	$\sqrt{689}, \sqrt{798}, \sqrt{901}, \sqrt{930}, \sqrt{962}$
N=5	$\sqrt{197}, \sqrt{421}, \sqrt{731}, \sqrt{786}, \sqrt{843}$	$\sqrt{132}, \sqrt{442}, \sqrt{485}, \sqrt{530}, \sqrt{577}$	$\sqrt{235}, \sqrt{506}, \sqrt{530}, \sqrt{552}, \sqrt{577}$	$\sqrt{458}, \sqrt{689}$ $\sqrt{770}, \sqrt{798}, \sqrt{915}$
N=6		$\sqrt{485}, \sqrt{530}, \sqrt{577}, \sqrt{632}, \sqrt{718}$	$\sqrt{158}, \sqrt{235}$ $\sqrt{530}, \sqrt{552}, \sqrt{577}$	$\sqrt{330}, \sqrt{458}, \sqrt{557}, \sqrt{798}, \sqrt{915}$

Table 1: "Good" experimental design generators selected by the sieve algorithm for estimating the parameters in univariate regression spanned by the Legendre polynomials of degree N for n experiments (see the text for details)

	n=100	n=250
$N = 4 \times 4$	$(\sqrt{742},\sqrt{822})$, $(\sqrt{235},\sqrt{742})$, $(\sqrt{513},\sqrt{577})$	$(\sqrt{798},\sqrt{962})$, $(\sqrt{698},\sqrt{962})$, $(\sqrt{659},\sqrt{915})$
$N = 5 \times 5$	$(\sqrt{74},\sqrt{257})$, $(\sqrt{96},\sqrt{114})$, $(\sqrt{114},\sqrt{138})$	$(\sqrt{616},\sqrt{798})$, $(\sqrt{332},\sqrt{798})$, $(\sqrt{616},\sqrt{689})$
$N = 6 \times 6$		$(\sqrt{616},\sqrt{798})$, $(\sqrt{332},\sqrt{798})$, $(\sqrt{387},\sqrt{408})$
$N = 4 \times 4 \times 4$		$(\sqrt{577},\sqrt{659},\sqrt{915})$, $(\sqrt{458},\sqrt{659},\sqrt{930})$, $(\sqrt{458},\sqrt{659},\sqrt{870})$

Table 2: "Good" generators for estimating bi- and trivariate regressions spanned by tensor products of Legendre polynomials for different number of observations n; the entry $N = 4 \times 4$ means, for example, that a regression is spanned by 16 terms, which are obtained by multiplying all possible univariate Legendre polynomials up to the 4th order

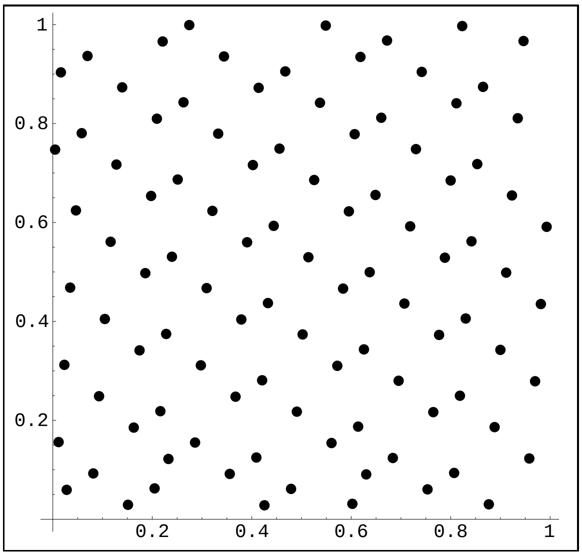

Figure 2: Experimental design for bivariate regression spanned by $N = 5 \times 5$ Legendre polynomials; design generated by $\theta = (\sqrt{74}, \sqrt{257})$, for $n = 100$ points; $q = 0.23$

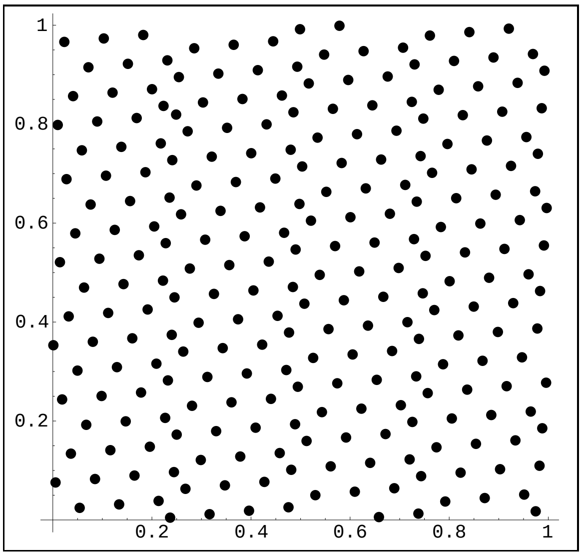

Figure 3: Experimental design for bivariate regression spanned by $N = 5 \times 5$ Legendre polynomials; design generated by $\theta = (\sqrt{798}, \sqrt{616})$, for $n = 250$ points; $q = 0.002$

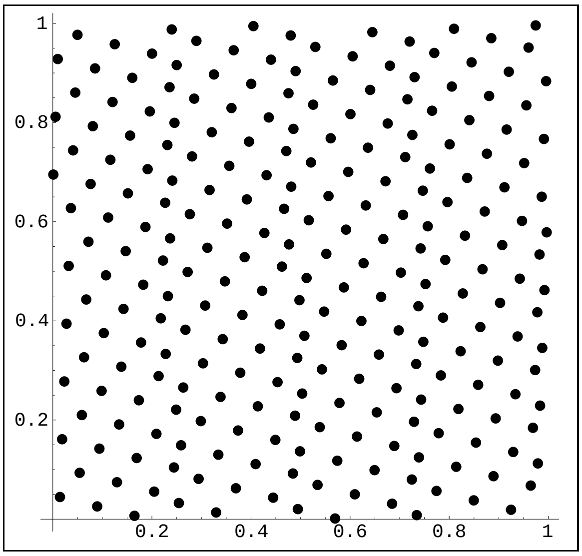

Figure 4: Experimental design generated by $\theta = (\sqrt{798}, \sqrt{332})$, $n = 250$ points, for $N = 6 \times 6$ Legendre polynomial regression; $q = 0.09$ (Note that the design from Fig. 3 has a similar efficiency in this case: $q = 0.092$)

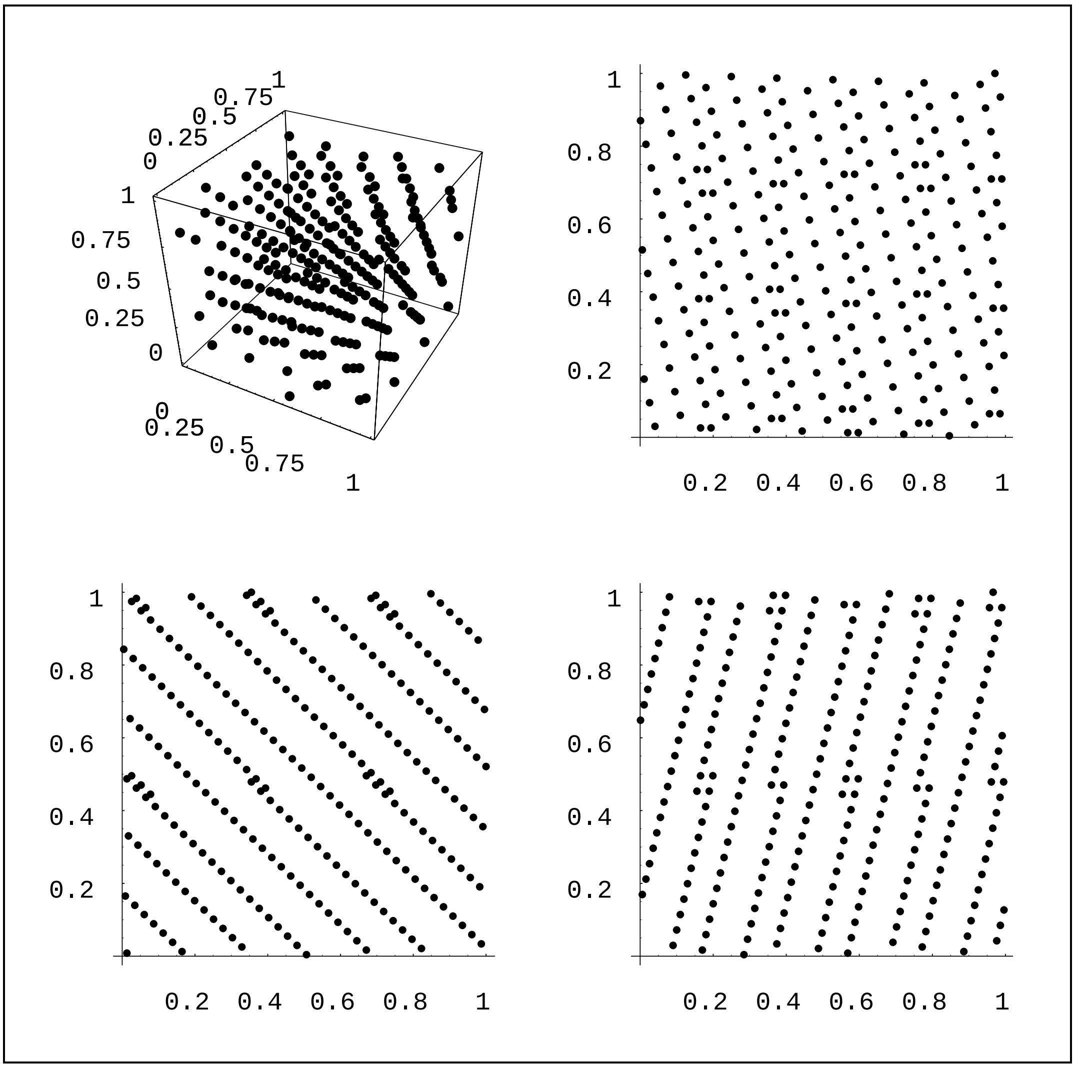

Figure 5: Experimental design generated by $\theta = (\sqrt{408}, \sqrt{659}, \sqrt{870})$, $n = 250$ points, for $N = 4\times4\times4$ regression (three input variables) and its two dimensional projections