Adaptive sequential design for regression on multi-resolution bases

Serge Cohen, Sébastien Déjean, Sébastien Gadat

June 8, 2011

Keywords: Active Learning, Optimal Designs, Stochastic Algorithms, Model Selection, Wavelets, MSC 62K05, 62L05, 93E35.

Abstract

This work investigates the problem of construction of designs for estimation and discrimination between competing linear models. In our framework, the unknown signal is observed with the addition of a noise and only a few evaluations of the noisy signal are available. The model selection is performed in a multi-resolution setting. In this setting, the locations of discrete sequential D and A designs are precisely contraint in a small number of explicit points. Hence, an efficient stochastic algorithm can be constructed that alternately improves the design and the model. Several numerical experiments illustrate the efficiency of our method for regression. One can also use this algorithm as a preliminary step to build response surfaces for sensitivity analysis.

1 Introduction

A new algorithm for building optimal designs is proposed in this paper to estimate an unknown signal η. We will use a one dimensional framework but all our results can be extended to multi-dimensional cases provided E is a finite union of finite dimensional compact and convex spaces. We discriminate competing models in a multi-resolution setting described by a set of functions $(\Lambda_{r,t})_{(r,t)}$. The signal is observed with the addition of a noise

$$\forall x \in E \qquad f(x) = \eta(x) + \sigma\zeta(x), \tag{1}$$

where $\zeta(x)$ will denote a normalized white noise and σ^2 is the variance of the model. To estimate η, we allow ourself as few observations of f as possible but we can choose the locations where f is evaluated to build our design. For any set I of pairs (r,t) where r is a resolution index and t a translations factor, and any design $\mathbf{x}$, we estimate η as

$$\hat{\eta}(x) = \sum_{(r,t)\in I} \hat{\theta}_{r,t}\Lambda_{r,t}(x).$$

Of course, $\hat{\theta}$ depends on I and $\mathbf{x}$ and we aim to obtain a good estimation of η with a small variance. We suppose that the cost of the evaluation of f is larger than any optimisation procedure to get I and $\mathbf{x}$ and our algorithm will have to select a few points among E to provide good estimations of η. A good example of such situations can be encountered in biology where experiments are often expensive and one cannot make many experiments. Another example is provided in the experimental section, dealing with the Motorcycle impact experiment introduced in [S85]. Hence, our motivation is to obtain a multi-resolution estimation with a few number of points in the design.

In the last decade, a large amount of works concerned by models selection based on stochastic or penalized approaches has been successfully developed. From a theoretical point of view, the recent ℓ^1 penalized approaches (Lasso of [T96], LARS of [EJHT03] or Dantzig Selector of [CT07]) propose to introduce some ℓ^1 penalties to constraint the estimator $\hat{\theta}$ to obtain sparse estimation $\hat{\eta}$ in the basis $(\Lambda_{r,t})_{(r,t)}$. These works reach satisfactory theoretical results for the selection of the model. But to the best of our knowledge, no exact optimization method for selecting a design dedicated to the sparse reconstruction of the model exists. Implicitly, the ℓ^1 penalized regression methods rely on properties of the design to obtain some good reconstruction but no work has been proposed to choose explicitly

a design in order to satisfy such properties. The work [CK03] uses an heuristic adaptive thresholding technique and propose a model selection procedure that deals with any (non-regular) design. One may also mention the interesting work [LO06]: their authors propose a construction of optimal and sequential designs for nonparametric regression models in the Daubechies basis with a simulated annealing strategy. This approach is also coupled with a model selection strategy. At last, the paper [KP04] considers a warping method of the space state and then thresholds the wavelet coefficients with non-regular designs. In [AAP06], some methods are provided to use wavelet bases with irregular designs using wavelet kernel penalized estimation. In the experimental section, we will provide comparisons of our method with ℓ^1 penalized approach (the LARS estimator of the Adaptive Lasso [Z06]) and the thresholding methods of [AAP06].

Regarding now the community of optimal designs, many works have been concerned by finding designs represented as a continuous measure on the state space E. But there exist a few amount of *explicit* results for discrete optimal designs. Moreover, a large amount of these numerical methods yield some continuous designs although from a practical point of view, discrete designs are easier to handle. Some results can be found in [DS97] but most of them are obtained by the optimization of numerical algorithms like simulated annealing and therefore are not explicit. Concerning the special case of multi-resolution bases and optimal designs, several authors have already used a simulated annealing procedure for building robust minimax integer-valued designs (see for instance [FW00], [LO06] or [OW03]). Note that simulated annealing may be a costly task when the dimension of the problem becomes large. Among these methods, few ones propose a model selection approach except [LO06] or [BC02] where the model selection strategy is implemented with an hypothesis testing approach.

We want to expand the signal η in a multi-resolution basis $(\Lambda_{r,t})_{r,t}$. We successively select some points x_i where f is evaluated to build a n-set $\mathbf{x} = \{x_1, \dots x_n\}$ and elements of $(\Lambda_{r,t})_{r,t}$ to reach a correct estimation $\hat{\eta}$ of η. But for numerical reasons, it is infeasible to explore all possible subsets of functions and all n-sets $\mathbf{x}$ to choose among them the best model and design. In our settings, the design $\mathbf{x} = \{x_1, \dots, x_n\}$ will be adapted to the sequential ealuations of f. We propose to develop a recursive strategy: we iteratively evaluate f with a sequence $(x_1, \dots x_n)$ of points in E and find both a design $\mathbf{x}_n = (x_1, \dots x_n)$ and a family of linearly independent functions in an optimal way to estimate η.

Our contribution is twofold, we first provide an algorithm based on multi-resolution bases and adaptive strategy. This algorithm is then applied with several bases. In the very special case of triangle Schauder basis, we provide also a new theoretical result about the locations of discrete optimal designs. This result will be useful for numerical implementations. and can be tensorized to dimensions larger than one. We will use some classical ideas of optimal design theory such as $D-$ or $A-$optimal designs (see e.g. [KC87, DS97] for general classical reminders on the subject). The model selection strategy will use ideas coming from the adaptive regression point of view developed with Multivariate Adaptive Regression Splines (MARS) described in [F91] and classical coarse to fine multi-resolution methods (see e.g. [M90, DJ94, BG05]). Given any initialization $(\mathbf{x_0}, I_0)$, our method will produce $\mathbf{x_{n+1}}$ from $\mathbf{x_n}$ by the addition of a new point $x_{n+1} \in E$. We will not remove an element of $\mathbf{x_n}$ at step $n+1$ as it would amount to loose some available information (which is costly!), thus $\mathbf{x}_n \subset \mathbf{x}_{n+1}$. The variance of our model will be controlled by the construction of a suitable design dedicated to I_n. The computation of I_n will rely on a stochastic approach that aims to optimize the bias of our linear reconstruction. The paper is organized as follows: next section presents our model and our algorithm. Section 3 recalls optimal design theory and the location properties of optimal designs. Section 4 precisely describes the stochastic algorithm for the model selection. At last, Section 5 provides some experimental comparisons, especially with ℓ^1-penalized approaches which are widely used now and the kernel penalized wavelet estimator described in [AAP06].

2 Model

2.1 Integrated Mean Square Error

Assume $E = [0;1]$, η is supposed to be expanded in a multi-resolution basis denoted by $(\Lambda_{r,t})_{r\geq 0, 0\leq t\leq 2^r-1}$:

$$\eta = \sum_{r,t} \theta_{r,t}\Lambda_{r,t}. \tag{2}$$

Note that this assumption is true as soon as η belongs to some Besov space. More precisely, we will assume that for some unknown $s > 0$, η belongs to some homogeneous Besov space $\dot{B}_2^{s,2}$. These spaces are described for instance in the chapter VI, paragraph 9 of [M90]. We will not technically detail this point as it is not in the scope of this work, but note that such spaces contain a very large class of functions. In the sequel, we will use a generic notation I to refer to a list of pairs $I = \{(r_1, t_1), \dots (r_p, t_p)\}$. I_n will thus describe the functions used at step n among the whole set $(\Lambda_{r,t})_{r\geq 0, 2^r>t\geq 0}$.

We recall that

$$f(x) := \eta(x) + \sigma\zeta(x), \tag{3}$$

where $\zeta(x)$ is a white noise. In the sequel, we will suppose σ is unknown and we will provide a procedure to estimate it.

Let us suppose that the design $\mathbf{x}$ and the set I are given. Let us denote by $\hat{\eta}_{\mathbf{x},I}$ the least square estimator of η with the linear model based on $(\Lambda_{(r,t)})_{(r,t)\in I}$ and the observations $(f(x_i))_{x_i\in\mathbf{x}}$. One can use a standard Integrated Mean Square Error (IMSE) and define

$$J(\mathbf{x}, I) = \int_E \mathbb{E}\left[\hat{\eta}_{\mathbf{x},I}(u) - \eta(u)\right]^2 du. \tag{4}$$

We also denote by $f(\mathbf{x})$ the column vector of the function f observed at the points of the design $\mathbf{x}$ and we use the notation $\bar{\Lambda}_I(\mathbf{x})$ for the rectangular $(p \times n)$ matrix:

$$\bar{\Lambda}_I(\mathbf{x}) = \begin{pmatrix} \Lambda_{(r_1,t_1)}(x_1) & \dots & \Lambda_{(r_1,t_1)}(x_n) \\ \vdots & \dots & \vdots \\ \Lambda_{(r_p,t_p)}(x_1) & \dots & \Lambda_{(r_p,t_p)}(x_n) \end{pmatrix}. \tag{5}$$

We can expand (4) as

$$J(\mathbf{x}, I) = \underbrace{\int_E \mathrm{Var}[\hat{\eta}_{\mathbf{x},I}(u)]du}_{:=V_{\mathbf{x},I}} + \underbrace{\int_E \left(\mathbb{E}[\hat{\eta}_{\mathbf{x},I}(u)] - \eta(u)\right)^2 du}_{:=B_{\mathbf{x},I}}. \tag{6}$$

With standard notations, $M_{\mathbf{x},I}$ denotes the information matrix of the design $\mathbf{x}$ and functions $\Lambda_{(r_1,t_1)}, \dots \Lambda_{(r_p,t_p)}$, *i.e.* $M_{\mathbf{x},I} = \bar{\Lambda}_I(\mathbf{x})^t\bar{\Lambda}_I(\mathbf{x})$. Let us denote by $\mu_{1,1}(I)$ the integral $\int_E {}^t\bar{\Lambda}_I(u)\bar{\Lambda}_I(u)du$. Some immediate computation (see *e.g.* [F69]) yields

$$V_{\mathbf{x},I} = \sigma^2 Tr\left(\mu_{1,1}(I)M_{\mathbf{x},I}^{-1}\right).$$

The effect of the design on the bias term is not explicit and thus it will not be possible to find analytic criteria which describe precisely the influence of $\mathbf{x}$ on $B_{\mathbf{x},I}$. To overcome this difficulty, it is however possible to adopt the minimax approach described for instance in [OW06]. Given a current set of pairs I, let us denote by I^c $\{(r,t) \quad | \quad r \geq 0, 0 \leq t \leq 2^r - 1\ , (r,t) \notin I\}$. We can decompose η in $\eta_I + \eta_{I^c}$ where

$$\eta_I = \sum_{(r,t)\in I} \theta_{r,t}\Lambda_{r,t} \quad \text{and} \quad \eta_{I^c} = \sum_{(r,t)\in I^c} \theta_{r,t}\Lambda_{r,t}.$$

Then, it is clear that η_{I^c} will not be estimated by our estimator constructed only with the set I. If $(\Lambda_{r,t})_{r\geq 0, 2^r>t\geq 0}$ is orthonormal, the bias can be decomposed as

$$B_{\mathbf{x},I} = \int_0^1 \left(\left[\mathbb{E}\hat{\eta}_{\mathbf{x},I}(u) - \eta_I(u)\right]^2 + \eta_{I^c}(u)^2\right) du.$$

Obviously, since η_{I^c} is unknown, $B_{\mathbf{x},I}$ remains untractable but it is possible to use a worst case approach with the minimax design theory. We follow the notation used in [OW06]. To ensure an equilibrium

between the contribution of the bias and the variance to the error, we can impose a bound on the magnitude of the remainder $\|\eta_{I^c}\|_2 \leq \tau$ for a constant $\tau > 0$. We denote the least favorable case $B^{\star}_{\mathbf{x},I,\tau}$:

$$B^{\star}_{\mathbf{x},I,\tau} = \sup_{\|\eta_{I^c}\|_2 \leq \tau} B_{\mathbf{x},I} = \sup_{\|\eta_{I^c}\|_2 \leq \tau} \int_0^1 \left([\mathbb{E}\hat{\eta}_{\mathbf{x},I}(u) - \eta_I(u)]^2 + \eta_{I^c}(u)^2 \right) du.$$

Since, $J(\mathbf{x}, I) \leq B^{\star}_{\mathbf{x},I,\tau} + \sigma^2 V_{\mathbf{x},I}$, our next algorithm will detail the choice of I and $\mathbf{x}$. Its dependence on the ratio $\nu = \sigma^2/\tau^2$ corresponds to a bias/variance tradeoff.

2.2 The adaptive algorithm

In our adaptive framework, we need to choose successively new points of the design $\mathbf{x}$ and we decide to update the set I or to keep this set unchanged. As pointed in the introduction, we do not delete points of the design $\mathbf{x}$. Consequently, the algorithm will necessary be of the form:

1. **Step 0**

   - Fix any initial set of functions I_0. For instance in a one dimensional setting with $E = [0; 1]$, we choose $I_0 = \{(0, 0); (1, 0); (1, 1)\}$ as we do not have any prior on the support concerning η.
   - For a given integer n_0, compute the optimal design $\mathbf{x_0}$ of size n_0 which minimizes the bound of the IMSE:
   $$\mathbf{x_0} = \arg\min_{\mathbf{x}} \left\{ B^{\star}_{\mathbf{x},I_0,\tau} + \nu Tr\left(\mu_{1,1}(I_0) M_{\mathbf{x}}^{-1}\right\}, I_0\right). \tag{7}$$

2. **Step n**

$\diamond\mathcal{MS}\diamond$ **M**odel **S**election step: update the set I_n to "optimize" $J(\mathbf{x_n}, .)$. Our strategy will build I_{n+1} from I_n and will rely on a random choice of an element $(r, t) \in I_n$. We will decide if we delete (r, t) or if we add an element that does not belong to I_n. We will explain the stochastic model selection in section 4.

$\diamond\mathcal{OD}\diamond$ **O**ptimal **D**esign: choose $\mathbf{x_{n+1}}$ deduced from $\mathbf{x_n}$ with the addition of one point x_{n+1}

$$\mathbf{x_{n+1}} = \mathbf{x_n} \cup \{x_{n+1}\},$$

using the former set I_{n+1}. This optimization is described in section 3 and mimic equation (7).

3 Optimization of the design ($\diamond\mathcal{OD}\diamond$ step)

In this part, we explain how to obtain the point x_{n+1} at step $\diamond\mathcal{OD}\diamond$ of the iteration n in the algorithm above. We first provide some details when $\nu = +\infty$, the case where only the variance term is considered. Next, we describe how we can deal with $\nu < +\infty$.

3.1 Generalities on $\diamond\mathcal{OD}\diamond$ step when $\nu = +\infty$

We will consider in this part two optimality criteria for the design based on the information matrix $M_{\mathbf{x},I}$ (for more general classical criteria, one may refer to [W92]). D-optimal designs are defined by $\arg\min_{\mathbf{x}} \Phi_0(M_{\mathbf{x}})$ where we denote $\Phi_0(M_{\mathbf{x}}) := \det M_{\mathbf{x}}^{-1}$ although A-optimal designs are based on the minimization of $\Phi_{1,C}$ where $\Phi_{1,C}(M_{\mathbf{x}}) = Tr\left(CM_{\mathbf{x}}^{-1}\right)$. In most cases, these two criteria (and many others described in [W92]) do not yield equivalent designs but each of them aims to control the variance of the linear model. We denote by $\tilde{V}_1(\mathbf{x}, I) := \Phi_0(M_{\mathbf{x}})$ as well as $\tilde{V}_2(\mathbf{x}, I) = \Phi_{1,Id}(M_{\mathbf{x}}) = Tr(M_{\mathbf{x}}^{-1})$, and $\tilde{V}_3(\mathbf{x}, I) = \Phi_{1,\mu_{1,1}(I)}(M_{\mathbf{x}}) = Tr(\mu_{1,1}(I)M_{\mathbf{x}}^{-1})$. In our work, we have mainly focused on the D-optimal designs for multi-resolution bases except in a very special case where some link with A-optimal designs is established.

Remark 1 *Since we do not handle continuous measures, it is impossible to easily recover classical results for D and A optimal designs using some equivalence theorems stated in [KW59, F69, KS66]. In the sequel, we claim some properties for the location of optimal designs for the Schauder and Haar bases.*

Our adaptive strategy is based on solutions of

$$x_{n+1} \in \arg\min_x \tilde{V}_i(\mathbf{x} \cup x, I_{n+1}), \tag{8}$$

where $\tilde{V}_i$ is defined by one of the previous criteria. Note that this step does not require any evaluation of f and the simplest way to find x_{n+1} seems to use some standard optimization algorithm (gradient descent or simulated annealing).

3.2 Optimal designs with Schauder/Haar basis when $\nu = +\infty$

In the sequel, the Haar mother function is $H_{0,0}(x) = \mathbf{1}_{x\in[0;1/2[} - \mathbf{1}_{x\in[1/2;1]}$ and we define the Schauder triangle mother function as $S_{0,0}(x) = \int_0^x H_{0,0}(t)dt$. These functions $H_{0,0}$ and $S_{0,0}$ are examples of functions $\Lambda_{0,0}$. For instance, Haar and Schauder bases are then uniquely defined by classical dilatations and translations

$$\Lambda_{r,t}(x) = 2^{r/2}\Lambda_{0,0}(2^r x - t).$$

Some elements of the Schauder basis can be seen on Figure 1.

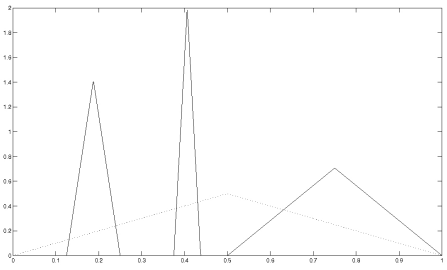

Figure 1: Several functions $\Lambda_{r,t}$, here (r,t) belongs to $\{(1,1)(3,1)(4,6)\}$.

A very important issue with Schauder basis is the fact that for every n, the argmin in (8) always occur on dyadic points $x = \frac{t}{2^r}$, whose resolutions r are bounded by the maximal resolution of the elements in I. This fact clearly speeds up the numerical resolution of equation (8). We will give a theoretical proof of this fact later. Figure 2 presents for instance the behaviour of $\tilde{V}_1$.

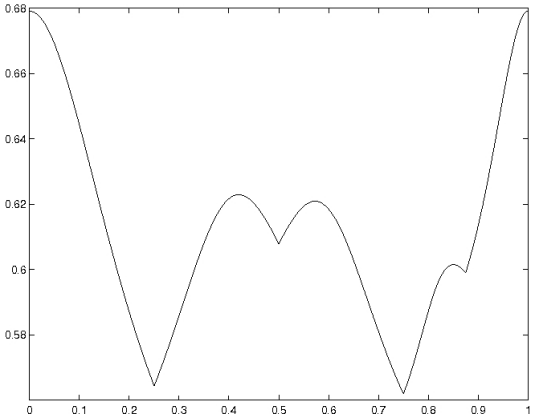

Figure 2: Evolution with respect to x of the variance $\tilde{V}_1$ when $I = \{(0,0);(1,0);(1,1);(2,3)\}$ for the Schauder triangle basis

Given any design $\mathbf{x}$, we are searching the location of $x \in E$ such that $\mathbf{x} \cup \{x\}$ is D-optimal. Next theorem shows that in fact, x is necessarily a dyadic point which is a singular point of functions described by I. If $Supp(\Psi)$ is the support of any function Ψ, we define $\mathcal{S}(I)$ as

$$\mathcal{S}(I) = \bigcup_{(r,t)\in I} \left\{ \arg\max_{x\in E} \Lambda_{r,t}(x) \right\} \bigcup_{(r,t)\in I} \{\partial Supp(\Lambda_{r,t})\}.$$

We obtain the following result whose proof is deferred to the appendix.

Theorem 1 *Let* $\mathbf{x}$ *be any fixed design and* $(\Lambda_{r,t})_{(r,t)\in I}$ *be any finite subset of functions of the Schauder triangle basis, then*

$$\arg\min_x \tilde{V}_1(\mathbf{x}\cup x, I) \in \mathcal{S}(I).$$

This result is very useful for our adaptive strategy to build $\mathbf{x}_{n+1}$ from the design $\mathbf{x}_n$ at step n. Indeed, it is sufficient to explore the small finite number of dyadic points $\mathcal{S}(I)$, described above and to select the point which minimizes the D-criterion. We also state a similar result for the case of ridge regession estimators.

Theorem 2 *Let* $\mathbf{x}$ *be any fixed design and* $(\Lambda_{r,t})_{(r,t)\in I}$ *be any finite subset of functions of the Schauder triangle basis, then for any* $\theta > 0$*:*

$$\arg\min_x \det\left(M_{\mathbf{x}\cup x,I} + \theta I_d\right)^{-1} \in \mathcal{S}(I).$$

Conjecture 1 *We conjecture also that for any symmetric positive matrix* C *and for any* $t > 0$*, the minima of* $\det\left(tC + M^{-1}_{\mathbf{x}\cup x,I}\right)$ *remain in* $\mathcal{S}(I)$*.*

Even if the proof of the conjecture is still open, it is supported by many numerical experiments. Concerning the other criteria, given the last observations $\mathbf{x}$, we search for the x which minimizes the criteria $\tilde{V}_2$ or $\tilde{V}_3$. One can re-write the A-optimal design criterium using the useful Schur factorization. For the sake of simplicity, let $\mu := \mu_{1,1}(I)$. Then

$$Tr\left(\mu M^{-1}_{\mathbf{x}\cup x}\right) = Tr\left(\mu M^{-1}_{\mathbf{x}} - \frac{\mu M^{-1}_{\mathbf{x}}\bar{\Lambda}^t_I\bar{\Lambda}_I M^{-1}_{\mathbf{x}}}{1+{}^t\bar{\Lambda}_I M^{-1}_{\mathbf{x}}\bar{\Lambda}_I}\right) = Tr\left(\mu M^{-1}_{\mathbf{x}}\right) - \frac{{}^t\bar{\Lambda}_I M^{-1}_{\mathbf{x}}\mu M^{-1}_{\mathbf{x}}\bar{\Lambda}_I}{1+{}^t\bar{\Lambda}_I M^{-1}_{\mathbf{x}}\bar{\Lambda}_I}.$$

Thus, the location of the optimal x at the step $n+1$ is deduced from the step n by maximizing the second term of the last equation. This can be performed easily in view of next theorem.

Theorem 3 *Let* $\mathbf{x}$ *be any fixed design and* $(\Lambda_{r,t})_{(r,t)\in I}$ *be any finite subset of functions of the Schauder triangle basis. Denote by* C *any symmetric non negative matrix, if the Conjecture 1 is true, then*

$$\arg\min_x Tr\left(CM^{-1}_{\mathbf{x}\cup x,I}\right) \subset \mathcal{S}(I).$$

We can deduce the $\diamond\mathcal{OD}\diamond$ step at iteration n from the last theorems for the Schauder triangle basis when the family of functions in I_{n+1} and the current design $\mathbf{x_n}$ are given:

- We consider the *finite* set $\mathcal{S}(I_{n+1})$.
- Compute the criterion for each element in $\mathcal{S}(I_{n+1})$ and choose x_{n+1} in $\arg\min_{x\in\mathcal{S}(I_{n+1})} \tilde{V}_i(\mathbf{x_n}\cup x, I_{n+1})$.

The study of Haar basis is much simple since the function $x \mapsto \tilde{V}_i(\mathbf{x}\cup x, I)$ is constant on the support of the elements of I which are intervals. Hence, it is sufficient to compute $\tilde{V}_i(\mathbf{x}\cup x, I)$ at the most $2\times|I|$ times to obtain its maximum value and then find an optimal point into this interval. From this simple remark, one can deduce the $\diamond\mathcal{OD}\diamond$ step at iteration n for the Haar basis when the family of functions I_{n+1} and the current design $\mathbf{x_n}$ are given:

- Rank by increasing order the elements of $\mathcal{S}(I_{n+1}) = \{\tilde{x}_1 < \tilde{x}_2 \cdots < \tilde{x}_l\}$.
- Sample $l+1$ elements $(\xi_j)_{j=1\ldots l+1}$ in each interval $]0;\tilde{x}_1[,\ldots]\tilde{x}_j,\tilde{x}_{j+1}[,\ldots]\tilde{x}_l;1[$ and choose x_{n+1} as a solution of the optimisation problem $\arg\min_{x\in\{\xi_1,\ldots\xi_{l+1}\}} \tilde{V}_i(\mathbf{x_n}\cup x, I_{n+1})$.

3.3 Optimal designs with general multi-resolution bases when $\nu = +\infty$

In the general case of multi-resolution wavelet bases such as Meyer basis, our approach consists in approximating solutions of several optimization problems. For D-optimal design, the algorithm uses the formula (15) provided in the appendix. The sequential D-optimal design is then equivalent to find at each step n the maximum of

$$x \mapsto^t \bar{\Lambda}_{I_{n+1}}(x) M_{\mathbf{x_n},I_{n+1}} \bar{\Lambda}_{I_{n+1}}(x).$$

For A-optimal design, we use the Schur factorization formula. Hence to find the sequential A-optimal design is equivalent to find at each step n the maximum of the function

$$x \mapsto \frac{1 +^t \bar{\Lambda}_{I_{n+1}}(x) M_{\mathbf{x_n} I_{n+1}} \bar{\Lambda}_{I_{n+1}}(x)}{^t\bar{\Lambda}_{I_{n+1}}(x) M^{-1}_{\mathbf{x_n},I_{n+1}} \mu M^{-1}_{\mathbf{x_n},I_{n+1}} \bar{\Lambda}_{I_{n+1}}(x)},$$

where μ is either Id or $\mu_{1,1}(I)$. This can be done either by using gradient descent algorithm or an exhaustive search when the set E is of small dimension (this will be the case in our simulations). Of course, the $\diamond\mathcal{OD}\diamond$ step in this case is much longer than the one in the Haar or Schauder case.

3.4 $\diamond\mathcal{OD}\diamond$ step with other regressors when $\nu = +\infty$

One may naturally wonder what can be done with other regressor. Especially we think of penalized regressors. Penalized methods can be viewed as the minimization of some penalized likelihood $\mathcal{L}$

$$\hat{\theta} := \arg\min_{\theta} \left\{ \log \mathcal{L}(\mathbf{x_n}, f(\mathbf{x_n}), \theta)) + \lambda \mathcal{N}(\theta) \right\},$$

where $\mathcal{N}(\theta)$ denotes the norm of penalization.
Ridge regression: For $\mathcal{N} = \|.\|_2^2$, we obtain the ridge regressor. For a fixed λ, the covariance of the ridge regression is $\left(M_{\mathbf{x_n},I_{n+1}} + \lambda Id\right)^{-1}$. Theorem 2 yields that optimal designs for the Schauder basis are also located in $\mathcal{S}(I_n)$. It is thus possible to infer some $\diamond\mathcal{OD}\diamond$ step for the Schauder basis using Ridge regression at step n: for each $x \in \mathcal{S}(I_n)$, compute the covariance matrix and one of the associated criterion $\tilde{V}_i$ where $M_{\mathbf{x_n},I_{n+1}}$ is replaced by $M_{\mathbf{x_n},I_{n+1}} + \lambda Id$. Then just pick the argmin x in $\mathcal{S}(I_n)$ and update the design $\mathbf{x_n}$ with it. Note that this method can be obviously extended to the simpler case of Haar basis. For more general wavelet bases, the price to pay is an intensive use of optimisation algorithm to find in E the argmin points.
Lasso regression: The choice of the penalization terms $\mathcal{N} = \|.\|_1$ yields the lasso regressor. The effect of such a penalization on the covariance of $\hat{\theta}$ is more complicated. Following the work [K01] on the soft thresholding estimator, the estimate of the covariance is approached using a Delta method by the matrix

$$Var(\hat{\theta}_{Gsoft}) = \left(H(\hat{\theta}) + \lambda \Gamma G(\hat{\theta}, \sigma)\right)^{-1} \Sigma(\hat{\theta}) \left(H(\hat{\theta}) + \lambda \Gamma G(\hat{\theta}, \sigma)\right)^{-1},$$

where H (resp. Σ) is the (resp. expected) Hessian matrix of the likelihood, where G is a gaussian kernel smoothed with bandwith σ and where Γ is the diagonal matrix $\Gamma = diag(I\{\theta < \sigma\}/\sigma)$. Indeed, $Var(\hat{\theta}_{Gsoft})$ depends on the observation x_{n+1} in the design which is unknown at the step n since the thresholding matrix Γ depends on $f(\mathbf{x_{n+1}})$. It is thus to the best of our knowledge very complicated to obtain this sequential design strategy.
Choice of λ: In the penalized approach (lasso, ridge, elastic net, ...), an efficient way to choose λ is to use a data-driven calibration cross-validation. Indeed, such a choice increases $Var(\hat{\theta}_{Gsoft})$ which depends on the x selected to update $\mathbf{x_n}$. Thus, we chose not to further investigate this question.

3.5 Minimax Optimal Design ($\nu < \infty$)

In the general case $\nu = \frac{\sigma^2}{\tau^2} < \infty$, we take into account the effect of the least favourable bias to choose our design. We have to find at step n the point x_n that optimizes the criterion Φ

$$\forall x \in E \qquad \Phi(x) = B^{\star}_{\mathbf{x}_n \cup x, I_n, \tau} + \sigma^2 Tr\left(\mu_{1,1}(I_n) M^{-1}_{\mathbf{x}_n \cup x, I_n}\right).$$

The optimization of Φ is a very difficult task since $B^{\star}_{\mathbf{x}_n \cup x, I_n, \tau}$ is not explicit. We adopt a suboptimal approach and deal with discrete designs located on dyadic points. Given a sufficiently large

maximal resolution $r_{max} \in \mathbb{N}$, we restrict our space of admissible x's to the set of dyadic points $E_{max} = \{k2^{-r_{max}}, k = 0, \dots, 2^{r_{max}}\}$ and we denote by $0 := \xi_1 < \cdots < \xi_{2^{r_{max}}+1} = 1$ the points in E_{max}. One can describe an approximation $\hat{\Phi}(x)$ of $\Phi(x)$ using E_{max} instead of E as in [OW06]. For the sake of simplicity, denote $d = 2^{r_{max}} + 1$ and $p = |I_{n+1}|$ the number of elements of I_{n+1}. Given any dyadic design $\mathbf{x} = \{x_1, \dots, x_m\}$ of length m, let us denote by $P(\mathbf{x})$ the diagonal matrix of occupation measure of E_{max} among $\mathbf{x}$:

$$\forall (k,l) \in \{1 \dots d\} \times \{1 \dots d\} \qquad P_{k,l}(\mathbf{x}) = \frac{|\{j | x_j = \xi_k\}|}{m} \mathbf{1}_{k=l}.$$

The application $\eta_{I^c} \in L^2(E)$ is replaced by its approximation vector $\bar{\eta}_{I^c} \in \ell^2(E_{max}) = (\eta_{I^c}(\xi_i))_{i=1\dots d}$ and we replace the constraint $\|\eta_{I^c}\|_2 \leq 1$ by $\|\bar{\eta}_{I^c}\|_2^2 \leq d$. If we denote by $Q_{I_{n+1}} = \left(\Lambda_{I_{n+1}}(\xi_1), \dots, \Lambda_{I_{n+1}}(\xi_d)\right)$ of size $p \times d$, the Average Squared Bias (ASB) is then approximated by

$$ASB \simeq \frac{1}{d} \left[\|Q_{I_{n+1}} (Q_{I_{n+1}}^t P(\mathbf{x}) Q_{I_{n+1}})^{-1} Q_{I_{n+1}}^t P(\mathbf{x}) \bar{\eta}_{I^c}\|_2^2 + \|\bar{\eta}_{I^c}\|_2^2 \right].$$

We assume $Q_{I_{n+1}}$ to have a full rank and we denote its the singular value decomposition by

$$Q_{I_{n+1}} = U_{d\times p}(I_{n+1}) \Delta_{p\times p}(I_{n+1}) \, {}^t V_{p\times p}(I_{n+1}),$$

where ${}^t U(I_{n+1}) U(I_{n+1}) = {}^t V(I_{n+1}) V(I_{n+1}) = I_p$ (the identity matrix of size p). Let us define the matrix $P_j(\mathbf{x}, I_{n+1}) = {}^t U(I_{n+1}) P^j(\mathbf{x}) U(I_{n+1})$, we know from [OW06] that the least favourable approximated bias is the largest eigenvalue of $P_1^{-1}(\mathbf{x}, I_{n+1}) P_2(\mathbf{x}, I_{n+1}) P_1^{-1}(\mathbf{x}, I_{n+1})$ denoted by $\lambda(\mathbf{x}, I_{n+1})$. The function Φ is then approximated by

$$\hat{\Phi}_{\mathbf{x}_n, I_{n+1}}(x) = \tau^2 \left[\lambda(\mathbf{x}_n \cup x, I_{n+1}) + \nu Tr \left(\mu_{1,1}(I_n) M_{\mathbf{x}_n \cup x, I_n}^{-1} \right) \right].$$

The above $\hat{\Phi}$ is used to select the optimal design at step $n+1$ *via* $x_{n+1} = \arg\min_x \hat{\Phi}_{\mathbf{x}_n, I_{n+1}}(x)$.

4 Stochastic Model Selection ($\diamond \mathcal{MS} \diamond$ step)

In this section, a stochastic algorithm which updates I_n is described. We provide first a method to estimate the "importance" of each point of I_n with respect to the bias term in (6) and then we deduce a ranking criterion. This ranking will be plugged into the acceptation/rejection step of a stochastic Simulated Annealing (S.A.). In a sense, this strategy is a kind of forward/backward stepwise model selection strategy. We try to add points which reduce the bias term but do not increase the variance of the estimation. Thus, this idea is similar to the MARS approach in [F91], the adaptive strategy of [CK03], or the algorithm of [BFI07].

4.1 Ranking criterion

Recall first that if $\hat{\eta}_{\mathbf{x}_n, I_n}(x)$ denotes the estimation of η with model I_n and design $\mathbf{x}_n$, then B is given by:

$$B_{\mathbf{x_n}, I_n} = \int_E \left(\mathbb{E}[\hat{\eta}_{\mathbf{x}_n, I_n}(x)] - \eta(x) \right)^2 dx.$$

Let us recall that we can decompose $\mathbb{E}\hat{\eta}_{\mathbf{x}_n, I_n}(x)$ in the basis

$$\mathbb{E}\hat{\eta}_{\mathbf{x}_n, I_n}(x) = \sum_{(r,t) \in I_n} \theta_{r,t} \Lambda_{r,t}.$$

Our idea is that the importance of each function $\Lambda_{r,t}$ for the bias can be deduced from the sensitivity of the bias to a small perturbation of the coefficient $\theta_{r,t}$. This idea is comparable to the Recursive Feature Elimination (RFE) algorithm of [GWBV02]: $\Lambda_{r,t}$ is important when $|\partial B_{\mathbf{x_n}, I_n} / \partial \theta_{r,t}|$ is large.

Now, we can formally compute each partial derivative of $B_{\mathbf{x_n}, I_n}$ with respect to $\theta_{r,t}$ to measure the importance of each $\Lambda_{r,t}$ as

$$\left| \frac{\partial B_{\mathbf{x_n}, I_n}}{\partial \theta_{r,t}} \right| = \left| 2 \int_E \Lambda_{r,t}(x) \left[\mathbb{E}\hat{\eta}_{\mathbf{x}_n, I_n}(x) - \eta(x) \right] dx \right|.$$

As pointed above, the exact computation of this last term is impossible as it requires to know η. Indeed, one can first estimate the bias pointwise $\mathbb{E}\hat{\eta}_{\mathbf{x}_n,I_n}(\xi) - \eta(\xi)$ for each point of the design using a standard "leave one out" cross validation coupled with a bootstrap sampling strategy. Then we use a classical method of interpolation to estimate the bias as a function on E. Let us denote $\epsilon := \hat{\eta}_{\mathbf{x}_n,I_n} - \eta$.

1. Fix the number of bootstrap samplers b (for instance $b = 10$).
2. For each point of the design x_k:
   (a) Build b bootstrap samples without the observation x_k: $(\mathcal{D}_1, \dots \mathcal{D}_b)$.
   (b) Compute for each bootstrap sample the estimators $\hat{\eta}_{\mathcal{D}_1,I_n}, \dots, \hat{\eta}_{\mathcal{D}_b,I_n}$.
   (c) Compute the "leave one out" bootstrap bias at x_k by
$$\hat{\epsilon}(x_k) = \frac{1}{b}\sum_{m=1}^{b} \hat{\eta}_{\mathcal{D}_m,I_n}(x_k) - \hat{\eta}_{\mathbf{x_n},I_n}(x_k).$$
3. Estimate the bias with $(\hat{\epsilon}(\xi), \xi \in \mathbf{x_n})$ and a simple spline smoothing estimator using a Gaussian kernel whose bandwith parameter is chosen by the Generalized Cross Validation criterion. For more details, one may refer to [BT04] for instance. This step provides a bias interpolation ϵ defined on E.
4. Estimate the derivative of the bias by a simple integration
$$\widehat{\partial_{(r,t)}B} := \left| 2\int_E \Lambda_{r,t}(x)\epsilon(x)dx \right|.$$

4.2 Stochastic Learning of I_n with a Simulated Annealing dynamic

We propose a method to update I_n. This algorithm is largely inspired by Metropolis-Hastings methods. The Simulated Annealing strategy is classically decomposed in a proposition step and an acceptation rule.

4.2.1 Reversible Jump proposal

Recall formally some elements of the S.A. procedure. Let Ω be a measurable set with a measure m and let μ be a measure on Ω with a density (also denoted by μ) with respect to m. We aim to minimize some cost C. The S.A. involves a simulation of a non-homogeneous Markov chain whose invariant distribution at iteration n is $\mu_n \propto \mu^{-C/T_n}$ where $(T_n)_{n\geq 0}$ is a temperature with $T_n \to 0$. Under classical conditions (see [H88, GG84] for instance), μ_∞ concentrates on the set of minima of C. The S.A. method with transition distribution $q(I, I')$ works as follow:

- from state $I \in \Omega$, first propose a state I' with probability $q(I, I')$
- then, accept the transition with a probability which is adjusted so that μ is invariant.

We assume the following reversible property

$$q(I, I') > 0 \iff q(I', I) > 0.$$

The probability to accept the transition I to I' at iteration n is then defined as:

$$\forall I' \neq I \quad Q_n(I, I') = \frac{\mu_n(I')q(I', I)}{\mu_n(I)q(I, I')} \wedge 1. \tag{9}$$

When μ corresponds to a Gibbs field associated to a cost function C (J defined in (6) in our case), this ratio is in fact given by

$$\forall I' \neq I \quad Q_n(I, I') = \left\{ e^{\frac{C(I)-C(I')}{T_n}} \frac{q(I', I)}{q(I, I')} \right\} \wedge 1. \tag{10}$$

The main difficulty is to ensure the weak reversibility condition given in the former paragraph: $q(I, I') > 0 \iff q(I', I) > 0$.

4.2.2 Birth and Deletion transition

In our framework, we start with $\{(0,0);(1,0);(1,1)\}$ and we decide to use the following dynamic for the iteration $I_n \mapsto I_{n+1}$:

$\mathcal{B}$: Birth of any element $i \notin I_n$ associated to a son or a father already in I_n.

$\mathcal{R}$: Rebirth of the element $\Lambda_{0,0}$ if $(0,0) \notin I_n$.

$\mathcal{D}$: Deletion of any element $i \in I_n$ provided that one son or father is still in I_n.

Given any iteration n with a design $\mathbf{x}_n$ and a basis I_n, we fix:

- $p_{\mathcal{R}}$ the probability to add $\Lambda_{0,0}$, if $\Lambda_{0,0} \in I_n$, we set $p_{\mathcal{R}} = 0$ otherwise $p_{\mathcal{R}} = 0.1$.
- $p_{\mathcal{B}} \in]0;1[$ the probability to add a function to I_n.
- $p_{\mathcal{D}}$ the probability to delete one element of I_n.

We state $p_{\mathcal{B}} = 5p_{\mathcal{D}}$ and $p_{\mathcal{B}} + p_{\mathcal{D}} + p_{\mathcal{R}} = 1$.

- In the birth case, denote by $I_n^{\mathcal{B}}$ the set of elements in I_n such that one of their sons or father is not in I_n. Then, propose the birth of the ascendant or descendant of some element $\Lambda_i, i \in I_n^{\mathcal{B}}$ where we sample i with a discrete probability $r_{\mathcal{B}} \propto \widehat{\partial B}_i$.
- In the deletion case, denote $I_n^{\mathcal{D}}$ the set of elements in I_n such that one of their descendants or ascendant is in I_n and propose the deletion of one element of $I_n^{\mathcal{D}}$ following the distribution $r_{\mathcal{D}} \propto \widehat{\partial B}_i^{-1}$.

The resulting transition kernel of the simulated Markov chain is then a mixture of the different transition kernels associated with the moves described above. We choose now classically $T_n = \frac{C_1}{C_2 + \log(n)}$ and this yields the transition kernel q.

Remark 2 *Please remark that the reachable vertices in I_n are not a priori connected in the tree representation space, it is a consequence of the reversibility condition. These moves are defined by heuristic considerations, the only condition to be fulfilled is to maintain the correct invariant distribution defined by (10).*

These moves are not the classical dynamic of dyadic trees for sets I_n. The classical evolution would generate connected trees but such trees are not sparse. At last, the necessary reversible jump condition is fulfilled by the definition of $\mathcal{B}$, $\mathcal{R}$ and $\mathcal{D}$.

4.3 Variance estimate and summary of the $\diamond\mathcal{MS}\diamond$ step

In order to obtain the complete transition kernel defined in (10), it is now sufficient to estimate ΔC which is the difference of the IMSE before and after the proposed transition. It is thus necessary to estimate both the bias and the variance of the model. The quantity $\Delta B_{\mathbf{x}_n, I_n}$ can be estimated using a bootstrap "leave one out" procedure as described in the paragraph 4.1. To compute $\Delta V_{\mathbf{x}_n, I_n}$, one just need to have a good estimation $\hat{\sigma}$ of σ defined in (1). This can be done using a maximum likelihood estimator of σ as pointed in [BFI07]. Note that in the resulting algorithm, $\hat{\sigma}$ is just needed to control the acceptation / rejection rule. Thus, the practical transition rule is given by a slightly modified equation (replace σ by its estimate $\hat{\sigma}$ in the control term C).

To sum up, the $\diamond\mathcal{MS}\diamond$ step can be described as

1. Choose a jump $\mathcal{B}, \mathcal{D}$ or $\mathcal{R}$ according to the distribution $(p_{\mathcal{B}}, p_{\mathcal{D}}, p_{\mathcal{R}})$.
2. Compute each ranking coefficient $\widehat{\partial B}(i), \forall i \in I_n$.
3. Compute the sampling distribution $r_{\mathcal{B}}$ or $r_{\mathcal{D}}$ and propose a new state I_{n+1} according to this discrete distribution.

4. If $\mathbf{x}_n = \{x_1, \dots x_n\}$ is the design, compute $\hat{\sigma}$ with the M.L.E. of σ given by:

$$\hat{\sigma}_n^2 = \frac{1}{|\mathbf{x}_n|}{}^t \left(f(\mathbf{x}_n) - (\Lambda_{I_n}(x_1), \dots, \Lambda_{I_n}(x_n))\hat{\theta}_n \right) M_{\mathbf{x}_n, I_n}^{-1} \left(f(\mathbf{x}_n) - (\Lambda_{I_n}(x_1), \dots, \Lambda_{I_n}(x_n))\hat{\theta}_n \right).$$

5. Compute the approximated differential cost $\Delta B_{(\mathbf{x}_n, I_n) -> (\mathbf{x}_n, I_{n+1})} + \hat{\sigma}_n^2 \Delta V_{(\mathbf{x}_n, I_n) -> (\mathbf{x}_n, I_{n+1})}$ and accept the transition with the probability

$$Q_n(I, I') = \left\{ e^{-[\Delta B + \hat{\sigma}^2 \Delta V]} \right\} \wedge 1.$$

   In the acceptation case, update I_{n+1} with the proposed transition. Otherwise keep I_n unchanged: $I_{n+1} = I_n$.

Since f is decomposed on I_n at step n, the M.L.E. $\hat{\sigma}_n$ is in general larger than the true σ. Hence, the acceptation/rejection rule Q_n prevents to overfit the models I_n.

5 Experimental results

In the sequel, we call a deterministic equispaced design a regular design whereas uniformly sampled designs will denote random designs obtained with a random uniform sampling over E. For all experiments, we set $E = [0; 1]$ and we aim to compare:

1. Least square linear model with
   - Optimal design coupled with model selection for the Schauder basis.
   - Optimal design coupled with model selection for the Haar basis.
   - Optimal design coupled with model selection for the Meyer basis.
   - Optimal minimax design coupled with model selection for the Schauder basis.
   - Optimal minimax design coupled with model selection for the Meyer basis.
2. Sparse Penalized regressor Adaptive Lasso (Ada-Lasso) [Z06] solved by the LARS implementation [EJHT03] (Schauder, Haar or Meyer basis).
   - Ada-Lasso regressor with design uniformly sampled in E.
   - Ada-Lasso regressor with regular design.
3. Wavelet approach of [AAP06] (Daubechies (D6) and Symmlets (S6) basis):
   - Kernel penalized estimation with uniformly sampled design.
   - Kernel penalized estimation with regular sampled design .

This section presents three examples where the unknown signal η must be recovered from as few observations as possible. The first synthetic example deals with the approximation of some unknown function that does not belong to finite dimensional vector spaces spanned by the triangle Schauder basis. The second example illustrates the database of Motorcycle impact experiment ([S85]). We provide also few results in a 2-dimensional setting. We consider the evolution of the IMSE with respect to the number of evaluations of the signal. Each implementation of [EJHT03] and [AAP06] will be used with uniformly sampled design and regular design. Note that both the Adaptive Lasso implementation and the Wavelet Penalized Kernel will be optimized using a cross-validation criterion. For the Adaptive Lasso, the bias of the penalization is removed once the model is determined using a least square estimation procedure based on the selected model. At last, the calibration of ν for minimax design is important. This parameter must be chosen sufficiently large to include the effect of the variance in our model. We have set $\nu = 20$ in our experiments, which seems a reasonnable value according to the size of the bias and variance in our experiments.

5.1 Description of the data

We investigate first the efficiency of estimation of the methods described above. We initialize the multi-resolution basis functions I_0 as $\bar{\Lambda}_{I_0} = \{\Lambda_{0,0}; \Lambda_{1,0}; \Lambda_{1;1}\}$.

5.1.1 Definition of the synthetic η

The synthetic function η to recover is a mixture of localized Gaussian kernel with different scaling parameters. For this, we set η to be localized around some values in E with different amplitudes and frequencies:

$$\forall x \in [0;1] \quad \eta(x) = 5e^{-1000(x-0.25)^2} + 5e^{-100(x-0.75)^2} + 20e^{-100(x-0.5)^2}.$$

The signal f is thus given by $f(x) = \eta(x) + \sigma\zeta(x)$ and some realizations of f are shown on Figure 3.

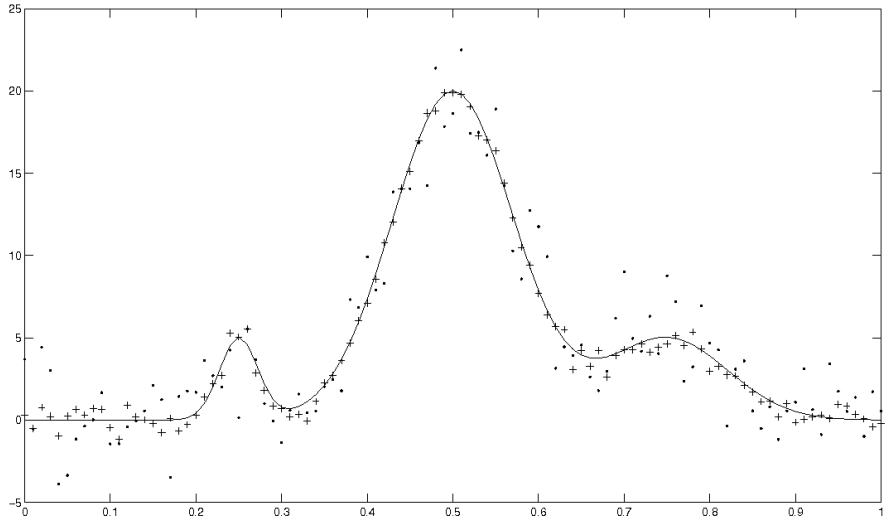

Figure 3: Synthetic function η and some realizations of $f(x)$ with $\sigma = 0.5$ (crossed points) or $\sigma = 2$ (dashed points).

5.1.2 Motorcycle impact experiment

In the Motorcycle Impact Experiment (see [S85] for a brief description of the data), the efficiency of crash helmets it studied. In [S85], the author uses 133 observations and a spline smoothing approach to estimate a curve deduced from the discrete observations. We scale the 133 observations between 0 and 1 and we compute a kernel smoothing interpolation described in [S85] to have a signal η to compare with our estimator. At last, we randomize the kernel interpolation by the addition of a white noise which yields an homoscedastic regression problem.

Remark 3 *Indeed, the Motorcycle experiment belongs to the more general class of problems of heteroscedastic regressions. Note that we use this dataset as a benchmark to validate our algorithm.*

5.2 Results

In order to obtain a reliable estimation of the IMSE for each algorithm, we repeat our experiments 50 times for each different method. Each experiment is reproduced also setting $\sigma = 0.5$ (low noise) and then $\sigma = 2$ (high noise). The Figure 4 shows the performance of our non-minimax sequential algorithm described in the beginning of this section on the synthetic dataset. One can see that in the low noise setting, there is no clear advantage to use Meyer basis which performs as Haar and Schauder bases. In contrast, in the high noise setting, the Meyer wavelet basis provides the best results. One can infer from this simulation that the linear model built with the Meyer wavelets is much more stable than the other ones. This is certainly due to the fact that Haar and Schauder mother functions are compactly supported and the information matrix of the least square estimator is "more singular" than the one constructed with the Meyer basis.

Moreover, Figure 11 provides the detailed numerical list of IMSE performance for each of the methods listed above in the case $\sigma = 0.5$ and Figure 12 provides the same results for the large noise case $\sigma = 2$ for the synthetic dataset. Best results for each number of points in the designs are underlined in the tables.

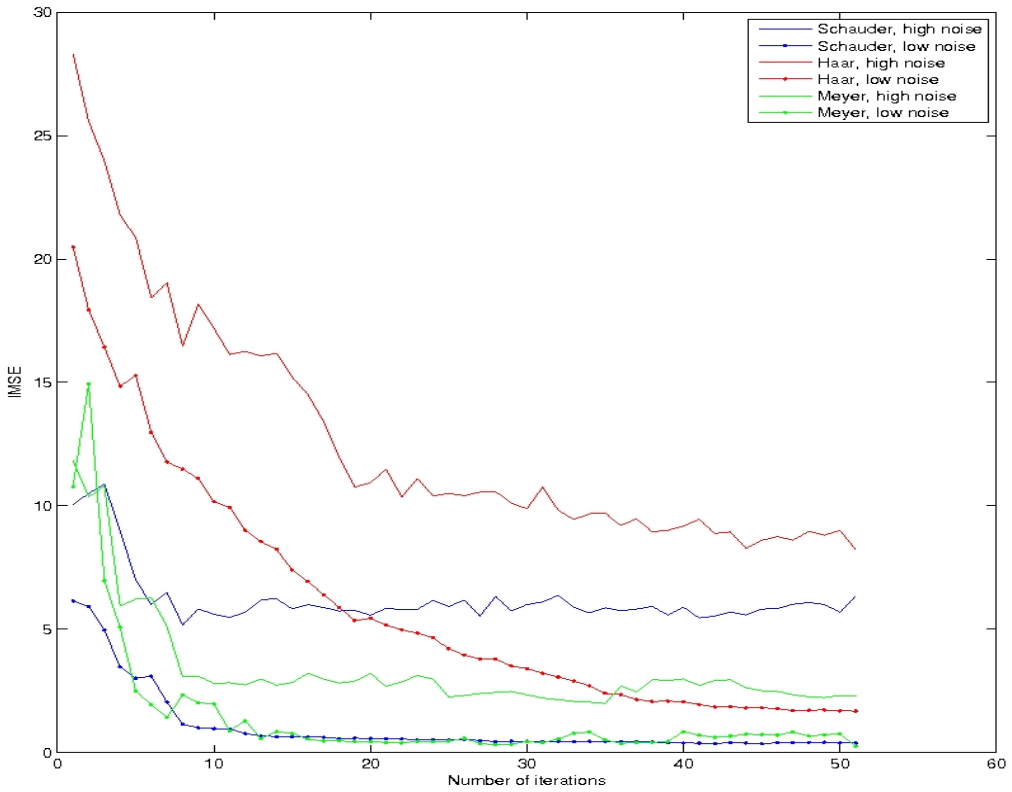

Figure 4: Integrated Mean Square Error for the synthetic example with low (dashed curves) and high noise (continuous curves) for the Haar (red), Schauder (blue) and Meyer (green) multi-resolution families. The x-axis represents the number of iterations of our method (number of points in the optimal designs).

These numerical results confirm that the use of the Meyer wavelet basis is always better than the other multi-resolution setting in our sequential strategy and this phenomena is much more important when the variance is important. The other striking fact is the poor performance of the Adaptive Lasso with the several multi-resolution families we used. This point is mainly due to the limited size of the set of experiments and the Adaptive Lasso is completely inapropriate in this framework. Moreover, in some rare cases (depending on the basis used), some very conservative choice are made by Ada-Lasso (see for instance the case of the Schauder basis in the high noise setting).

One can also remark the very good behaviour of the thresholding procedure as soon as there are enough points in the design. Since such methods are specifically dedicated to the multi-resolution framework, it is not surprising that they behave better than Adaptive Lasso in these cases. However, note that in a very few cases of measurements, a simple linear model with the design learned by our sequential method is more efficient than non-linear thresholding procedures.

At last, the minimax sequential design strategy performs the best. But the method is costly compared to the non-minimax Schauder one and there is no clear advantage to use it for such small IMSE in the small noise setting. For the large noise case, the minimax sequential design is justified since it improves the IMSE of 50 % compared to the non-minimax Schauder sequential approach.

The Figure 8 shows the performance of the sequential algorithm for the Motorcycle dataset. We display some estimation results obtained with the Schauder basis (Figure 9) and the Meyer wavelet basis (Figure 10). Note that the conclusions for the synthetic data still hold for the Motorcycle dataset since the Meyer basis is clearly better than the other ones as pointed in the Figure 8. For the sake of simplicity, we have omitted the results of the Adaptive Lasso algorithm on Figures 11 and 12. One may remark that with a very small number of points in the design, the sequential method always outperforms the wavelet thresholding procedure. But for $n = 50$, thresholding with Daubechies or Symlets bases are equivalent to our procedure with non-minimax design. One can remark that the minimax sequential design achieves again the best performances. This is especially the case for a very small number of experiments ($n = 10$): the minimax criterion tends to scatter the points of the design among E at the beginning of our strategy. It is explained by the fact that the variance term is smaller than the worst bias for small number of experiments and low resolution elements in I_n. This phenomenon is reversed

Method	IMSE (n=10)	IMSE (n=30)	IMSE (n=50)
Sequential Haar	10.1	3.4	1.7
Sequential Schauder	1.0	0.45	0.4
Sequential Meyer	0.9	0.4	0.38
Sequential minimax Schauder	0.7	0.4	0.4
Sequential minimax Meyer	0.65	0.35	0.3
Ada-Lasso Haar Random	70.8	75.7	56.2
Ada-Lasso Haar Regular	69	42.9	31
Ada-Lasso Schauder Random	50.2	20.8	14.3
Ada-Lasso Schauder Regular	13.6	13.9	12.3
Ada-Lasso Meyer Random	116.4	66.8	72.6
Ada-Lasso Meyer Regular	290	47.8	45.2
Wavelet Kernel Penalized D6 Random	8.2	10.3	1.8
Wavelet Kernel Penalized D6 Regular	4.9	1.0	0.9
Wavelet Kernel Penalized S6 Random	5.2	2.1	0.4
Wavelet Kernel Penalized S6 Regular	83.5	27.7	0.4

Figure 5: Integrated Mean Square Error for the synthetic data with low noise.

Method	IMSE (n=10)	IMSE (n=30)	IMSE (n=50)
Sequential Haar	17.2	9.9	9.0
Sequential Schauder	5.6	5.9	5.6
Sequential Meyer	2.8	2.3	2.3
Sequential minimax Schauder	3.1	1.5	1.2
Sequential minimax Meyer	1.7	1.4	1.1
Ada-Lasso Haar Random	85	71.6	71.5
Ada-Lasso Haar Regular	71.1	50.6	43.1
Ada-Lasso Schauder Random	24.3	37.3	24.1
Ada-Lasso Schauder Regular	16.9	17.1	12.2
Ada-Lasso Meyer Random	155	195	301
Ada-Lasso Meyer Regular	282	49	43
Wavelet Kernel Penalized D6 Random	21.4	2.5	22.9
Wavelet Kernel Penalized D6 Regular	15.5	11.9	2.7
Wavelet Kernel Penalized S6 Random	8.5	4.1	2.4
Wavelet Kernel Penalized S6 Regular	4.0	3.9	2.2

Figure 6: Integrated Mean Square Error for the synthetic data with high noise.

when the number of iterations is growing.

Figure 7 shows the good behaviour of our estimation of the bias as described in the paragraph 4.1 with a few number of functions in I_n. Note that this estimation deteriorates a little bit when the number of functions in I_n is growing. This last point is not very important since we mainly use this estimation as an indicator in the proposition step of our stochastic algorithm.

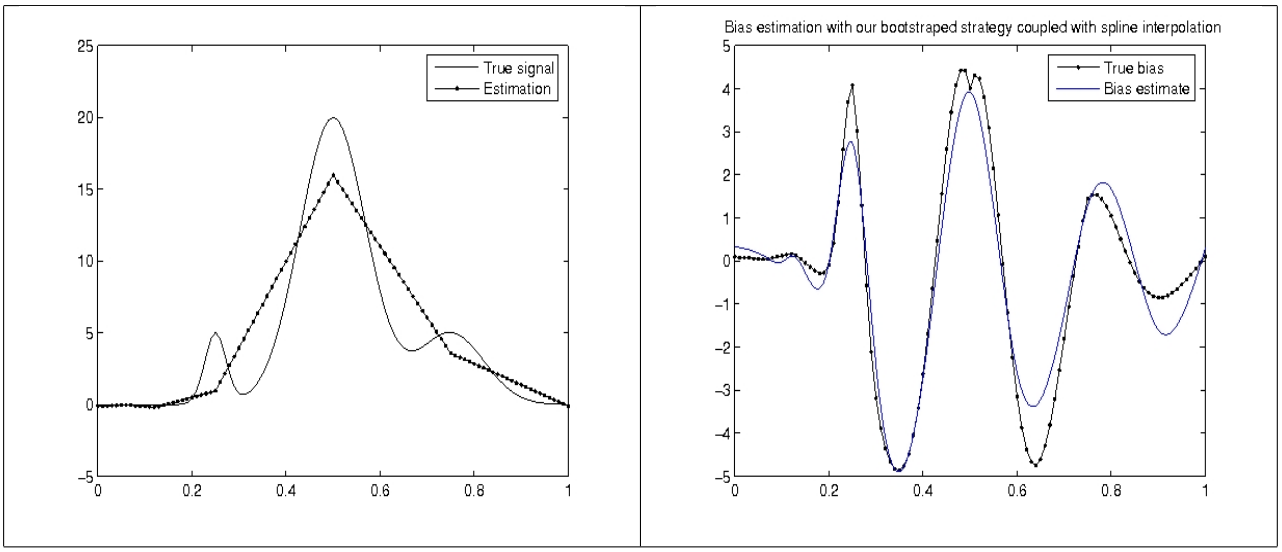

Figure 7: Estimation of the bias by our bootstraped procedure with 10 experiments and 5 functions in the family I on the synthetic example.

5.3 Computation Time

At last, concerning the computational costs, since the sequential design for Meyer basis is obtained by a simulated annealing algorithm, it may be interesting to switch to other multi resolution bases such as the Schauder one. Indeed, an iteration $n \to n+1$ with the Schauder basis requires less than 5 seconds although the same iteration with the Meyer wavelet basis needs for large n approximately 1 minute.

Although both the design and the model are estimated with our algorithm, it is as fast as the penalized approach or a thresholding method. However, the application of our method with the Meyer basis is much more costly. The minimax approach is obviously more costly than the non-minimax one when the set of functions is the Schauder one. But the cost of the minimax approach is equivalent to the one of the non-minimax approach with Meyer basis.

5.4 2-Dimensional experiments

We provide a short illustrating example in 2 dimensions with the Schauder triangle basis. Let us suppose

$$\forall (x,y) \in [0;1]^2 \qquad \eta(x,y) = 10xy(x-1)(y-1)e^{5(x-3/4)^2+(y-3/4)^2-5(x-1/4)^2+(y-1/4)^2}. \tag{11}$$

We present in Figure 13 the estimation of η with the tensorized Schauder basis with a small number of iterations. Such basis are deduced from the bases in the 1 dimensional case as follows:

$$\mathcal{F}_n = \{(x,y) \mapsto \Lambda_{r_1,t_1}(x); (x,y) \mapsto \Lambda_{r_2,t_2}(y); (x,y) \mapsto \Lambda_{r_1,t_1}(x)\Lambda_{r_2,t_2}(y)\},$$

where (r_1,t_1) and (r_2,t_2) belong to a finite set I_n. Roughly speaking, one can adapt the conclusions of Theorem 1 to obtain some properties for the location of the designs. The simple adaptation of the $\diamond\mathcal{OD}\diamond$ step consists in:

1. Compute all singular points $\mathcal{S}_x(I_n)$ of elements in $\Lambda_{(r,t)}$ on the x-coordinate.
2. Compute all singular points $\mathcal{S}_y(I_n)$ of elements in $\Lambda_{(r,t)}$ on the y-coordinate.
3. Then $\mathcal{S}(I_n) = \mathcal{S}_x(I_n) \times \mathcal{S}_y(I_n)$.

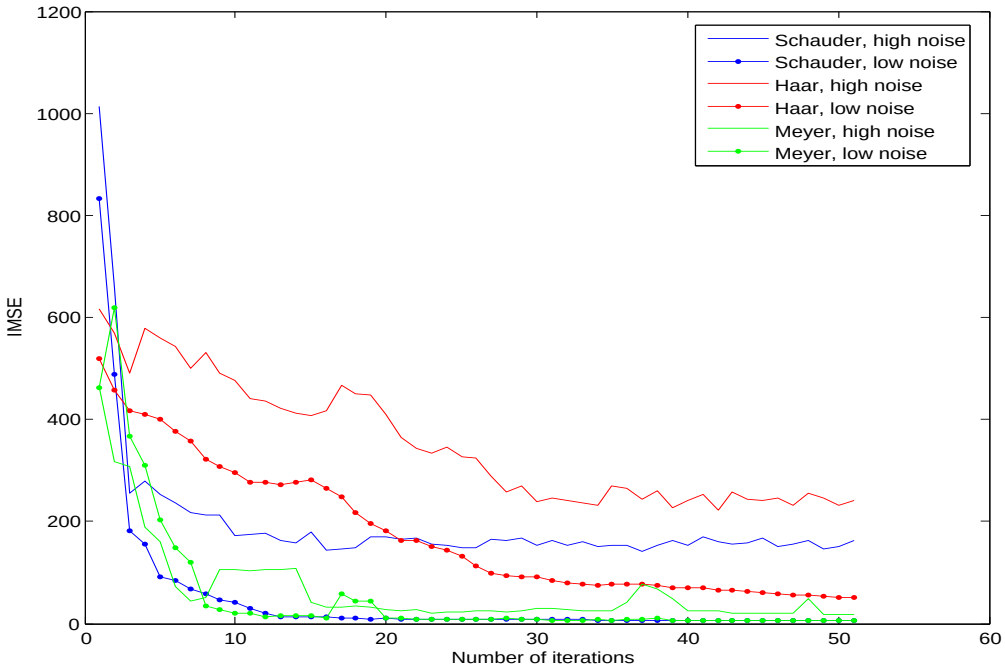

Figure 8: Integrated Mean Square Error of our sequential methods for the Motorcycle example with low (dashed curves) and high noise (continuous curves) for the Haar (red), Schauder (blue) and Meyer (green) multi-resolution families. The x-axis represents the number of iterations of our method (number of points in the optimal designs).

We show in Figure 13 some estimations obtained with a very small number of iterations of our algorithms. The red surface is the true response and the meshed surface is our estimation. One can remark that the algorithm still takes into account the singularities of the signal as it was the case in the one dimensional examples. Moreover, the approximation of the extrema is pretty good and this example shows that our method is promising for response surface applications.

6 Conclusion

The adaptive method developed in this paper is numerically competitive for synthetic and real examples compared to thresholding wavelet or ℓ^1 penalized methods. The iterative scheme is fast and may be very fast if it does not require any complicated optimization step as it is the case in the special case of the Schauder basis. The approximation properties of both the global signal or its maxima are satisfying. At last, the model selection ability could be of great interest for variable selection motivations and are meaningful. High resolution functions are added when needed or not used when the signal is well approximated.

But, on the theoretical side, many questions remain open. First, it would be very fruitful to generalize the result that localizes the optimal designs on dyadic points for other multi-resolution sets of smooth functions. But it is a difficult task owing to the underlying non-linear nature of the optimization problem in the case of general wavelet bases.

Second, the proof of the convergence of the stochastic coupled algorithm on $(I_n, \mathbf{x}_n)$ remains a difficult task. To do so, it is necessary to fix a precise cooling strategy and to use the consistency result of theorem

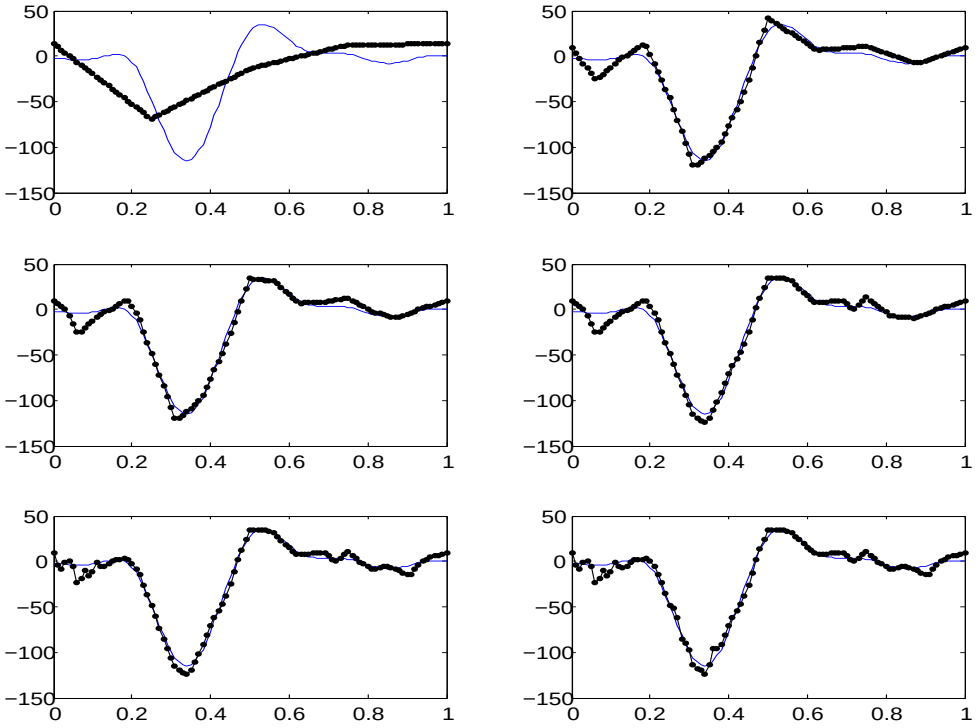

Figure 9: Estimations on the Motorcycle example with high noise ($\sigma = 10$) at iteration 0 (a), 10 (b), 20 (c), 30 (d), 40 (e), 50 (f). Continuous curve: true signal, Dashed curve: interpolation of our sequential procedure with the Schauder wavelet basis.

4 provided in the appendix as a first step.

One may also consider a modified energy criterion based on a discrepancy term to bound the bias B which is unknown in our sequential framework. This would certainly enable some theoretical extensions of the consistency result (Theorem 4) in the fixed basis setting

At last, it would be very fruitful to infer a sequential optimal design approach for the ℓ^1 penalized approaches. The optimization of the variance with respect to the design seems difficult, mostly because the choice of the penalty parameter as discussed in paragraph 3.4 is not explicit.

Aknowledgements We would like to gratefully acknowledge Jean-Marc Azaïs, Fabrice Gamboa and Anestis Antoniadis for their helpful comments on an early version of the manuscript and their warm support. We are very much indebted to the anonymous referees and the associate editor for their constructive comments that resulted in a major revision of the original manuscript.

7 Appendix

7.1 Location properties of the design

We will denote by n the number of points in a fixed design $\mathbf{x}$ and by p the cardinal of I. Recall that $\bar{\Lambda}_I$ is the rectangular matrix defined by equation (5). Since I is fixed, we drop the indice I in this appendix

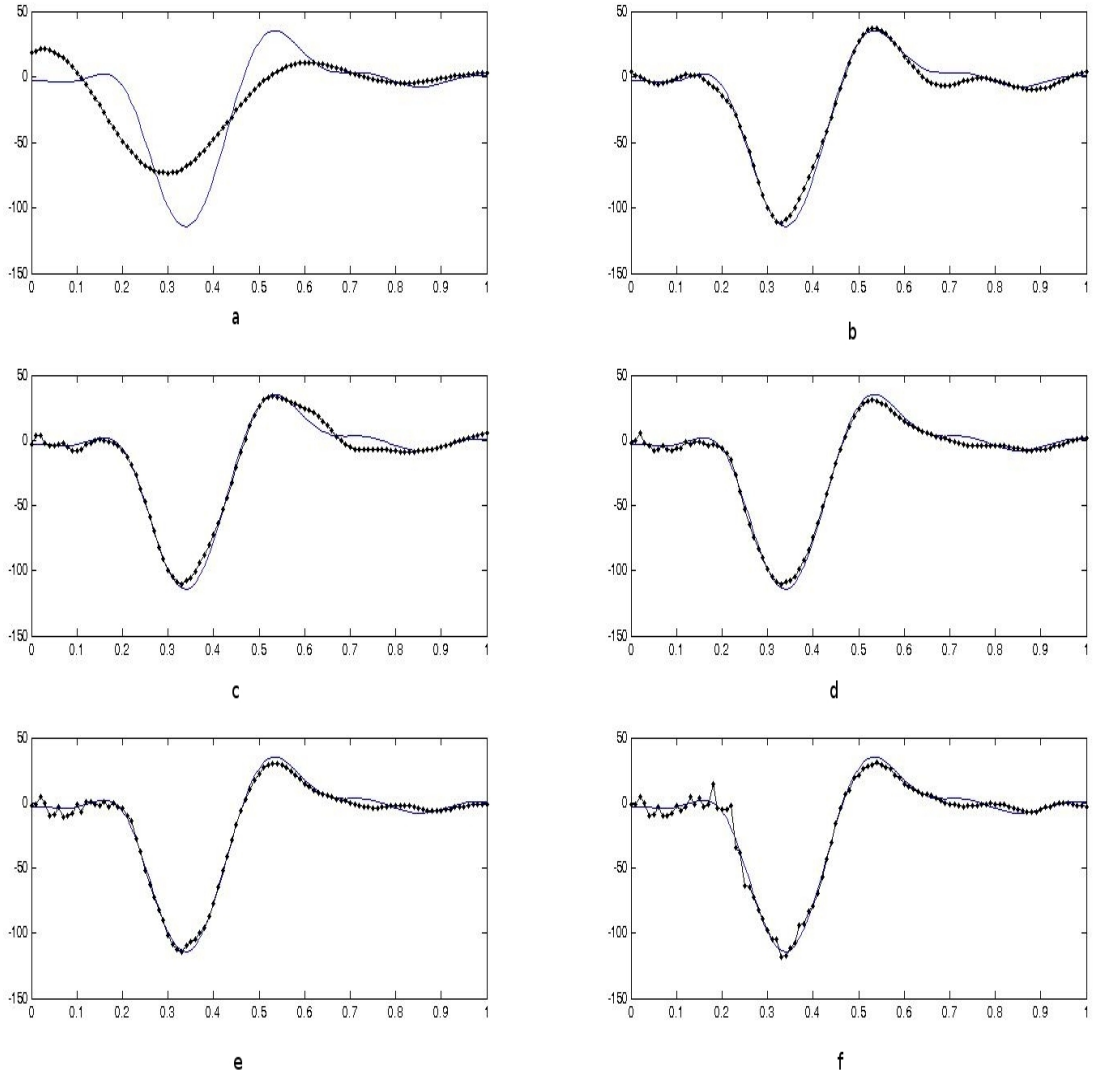

Figure 10: Estimations on the Motorcycle example with high noise ($\sigma = 10$) at iteration 0 (a), 10 (b), 20 (c), 30 (d), 40 (e), 50 (f). Continuous curve: true signal, Dashed curve: interpolation of our sequential procedure with the Meyer wavelet basis.

for the notation of $M_{\mathbf{x}} =^t \bar{\Lambda}_I \bar{\Lambda}_I$.

We suppose that $n + 1 \geq p$ (which is rather trivial in our context) and denote by F the map given by $F(x) = \det(M_{\mathbf{x}\cup x})$. We will show that F is a convex map on every interval where it is differentiable. Assuming x to be suitably chosen among differentiable points of $\bar{\Lambda}_I$, we will note $\bar{\Lambda}'_I(x)$ the vector composed of the differentiable maps of $\bar{\Lambda}_I$ computed at point x and the matrix

$$M'_x = \left((\bar{\Lambda}_{I_1}\bar{\Lambda}'_{I_2} + \bar{\Lambda}'_{I_1}\bar{\Lambda}_{I_2})(x)\right)_{i_1,i_2\in I} = \frac{\mathrm{d}}{\mathrm{d}x}\left(M_{\mathbf{x}\cup x}\right).$$

Using the standard Euclidean scalar product on $\mathbb{R}^p$, one can check immediately that

$$\forall U \in \mathbb{R}^p, \quad M'_x U = \langle \bar{\Lambda}_I(x); U\rangle \Lambda'_I(x) + \langle \bar{\Lambda}'_I(x); U\rangle \bar{\Lambda}_I(x).$$

First, we state some classical results on matrices whose proofs are based on standard arguments on matrices of rank 1. Some details can be found in [MN95] and in chapter one of [F69].

Proposition 1 *If $M^{-1}_{\mathbf{x}\cup x}$ and $M^{-1}_{\mathbf{x}}$ are non-singular,*

$$M^{-1}_{\mathbf{x}} = M^{-1}_{\mathbf{x}\cup x} + \frac{M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I(x)^t\bar{\Lambda}_I(x)M^{-1}_{\mathbf{x}\cup x}}{1 -^t \bar{\Lambda}_I(x)M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I(x)} \tag{12}$$

$$M^{-1}_{\mathbf{x}\cup x} = M^{-1}_{\mathbf{x}} - \frac{M^{-1}_{\mathbf{x}}\bar{\Lambda}_I(x)^t\bar{\Lambda}_I(x)M^{-1}_{\mathbf{x}}}{1 +^t \bar{\Lambda}_I(x)M^{-1}_{\mathbf{x}}\bar{\Lambda}_I(x)}. \tag{13}$$

Method	IMSE (n=10)	IMSE (n=30)	IMSE (n=50)
Sequential Haar	296	91	50
Sequential Schauder	41.9	7.5	6.5
Sequential Meyer	19.7	7.4	6.0
Sequential minimax Schauder	23	9.5	5.5
Sequential minimax Meyer	15.7	5.2	5.1
Wavelet Kernel Penalized D6 Random	1549	26.4	9
Wavelet Kernel Penalized D6 Regular	458	15	12
Wavelet Kernel Penalized S6 Random	188	154	8.9
Wavelet Kernel Penalized S6 Regular	28.4	11.3	9.5

Figure 11: Integrated Mean Square Error for the Motorcycle experiment and low noise.

Method	IMSE (n=10)	IMSE (n=30)	IMSE (n=50)
Sequential Haar	477	239	232
Sequential Schauder	171	153	152
Sequential Meyer	104	28.5	18.3
Sequential minimax Schauder	94	23	14
Sequential minimax Meyer	88	21	12
Wavelet Kernel Penalized D6 Random	1074	158	93
Wavelet Kernel Penalized D6 Regular	556	115.7	135
Wavelet Kernel Penalized S6 Random	180	129	30
Wavelet Kernel Penalized S6 Regular	122	59	18

Figure 12: Integrated Mean Square Error for the Motorcycle experiment and high noise.

Moreover,

$$\frac{\det M_{\mathbf{x}\cup x}}{\det M_{\mathbf{x}}} = \frac{1}{1 -^t \bar{\Lambda}_I M_{\mathbf{x}\cup x} \bar{\Lambda}_I}, \tag{14}$$

$$\frac{\det M_{\mathbf{x}}}{\det M_{\mathbf{x}\cup x}} = \frac{1}{1 +^t \bar{\Lambda}_I M_{\mathbf{x}} \bar{\Lambda}_I}. \tag{15}$$

We now establish two technical lemma useful for our location theorem.

Lemma 1 *For any symmetric matrix S, we have the relation*

$$Tr(M'_x S) = Tr(S M'_x) = 2\langle S\bar{\Lambda}_I(x); \Lambda'_I(x)\rangle. \tag{16}$$

Proof: Consider first the case where $\Lambda_I(x), \Lambda'_I(x)$ are linearly independent in $\mathbb{R}^p$. One can show that

$$M'_x S\bar{\Lambda}_I(x) = \langle S\bar{\Lambda}_I(x); \bar{\Lambda}_I(x)\rangle \Lambda'_I(x) + \langle S\Lambda'_I(x); \bar{\Lambda}_I(x)\rangle \bar{\Lambda}_I(x),$$

and

$$M'_x S\Lambda'_I(x) = \langle S\bar{\Lambda}_I(x); \Lambda'_I(x)\rangle \Lambda'_I(x) + \langle S\Lambda'_I(x); \Lambda'_I(x)\rangle \bar{\Lambda}_I(x).$$

Since the rank of M'_x is 2, we can find a basis adapted to the family $(\bar{\Lambda}_I(x); \Lambda'_I(x))$ such that the endomorphism described by $M'_x S$ in the basis is

$$\begin{pmatrix} \langle S\Lambda'_I(x); \bar{\Lambda}_I(x)\rangle & \langle S\bar{\Lambda}_I(x); \bar{\Lambda}_I(x)\rangle & 0 & \dots & 0 \\ \langle S\Lambda'_I(x); \Lambda'_I(x)\rangle & \langle S\bar{\Lambda}_I(x); \bar{\Lambda}'_I(x)\rangle & 0 & \vdots & 0 \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & \dots & 0 & \dots & 0 \end{pmatrix}.$$

Thus in this case

$$Tr(M'_x S) = 2\langle S\bar{\Lambda}_I(x); \Lambda'_I(x)\rangle.$$

Suppose now that $\Lambda_I(x), \Lambda'_I(x)$ are linearly dependent, we get

$$\langle S\bar{\Lambda}_I(x); \bar{\Lambda}_I(x)\rangle \Lambda'_I(x) = \langle S\Lambda'_I(x); \bar{\Lambda}_I(x)\rangle \bar{\Lambda}_I(x),$$

and applying the same argument as above with the endomorphism $M'_x S$ the rank of which is one in this case, we also obtain

$$Tr(M'_x S) = 2\langle S\bar{\Lambda}_I(x); \Lambda'_I(x)\rangle. \qquad \square$$

If we denote by $Com(M)$ the matrix ${}^t cof(M)$, where $cof(M)$ is the matrix of cofactors of A, we have the following result.

Lemma 2 *Assume x to be a regular point for the map $\bar{\Lambda}_I$, and that $M_{\mathbf{x}}$, $M_{\mathbf{x}\cup x}$ are non-singular, then*

$$Tr\left({}^t Com(M_{\mathbf{x}\cup x}) M'_x\right) = Tr\left({}^t Com(M_{\mathbf{x}}) M'_x\right).$$

Proof: Apply lemma 1 first to $S = M^{-1}_{\mathbf{x}\cup x}$, we get

$$Tr(M^{-1}_{\mathbf{x}\cup x} M'_x) = 2\langle M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I; \bar{\Lambda}'_I\rangle. \tag{17}$$

Moreover, lemma 1 applied now to $S = M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}^t_I \bar{\Lambda}_I M^{-1}_{\mathbf{x}\cup x}$ yields

$$Tr\left(M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}^t_I\bar{\Lambda}_I M^{-1}_{\mathbf{x}\cup x} M'_x\right) = 2\langle M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I \underbrace{{}^t\bar{\Lambda}_I M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I}_{=\langle \bar{\Lambda}_I; M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I\rangle}; \bar{\Lambda}'_I\rangle.$$

Thus

$$Tr\left(M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}^t_I\bar{\Lambda}_I M^{-1}_{\mathbf{x}\cup x} M'_x\right) = 2\langle \bar{\Lambda}_I; M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I\rangle\langle \Lambda'_I; M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I\rangle. \tag{18}$$

From (12),(17), and (18), we get

$$\begin{aligned}
Tr(M^{-1}_{\mathbf{x}} M'_x) &= 2\langle M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I; \bar{\Lambda}'_I\rangle + \frac{2\langle \bar{\Lambda}_I; M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I\rangle\langle \Lambda'_I; M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I\rangle}{1 - {}^t\bar{\Lambda}_I M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I} \\
&= 2\langle M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I; \bar{\Lambda}'_I\rangle \times \left(1 + \frac{\langle \bar{\Lambda}_I; M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I\rangle}{1 - {}^t\bar{\Lambda}_I M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I}\right) \\
Tr(M^{-1}_{\mathbf{x}} M'_x) &= \frac{Tr(M^{-1}_{\mathbf{x}\cup x} M'_x)}{1 - {}^t\bar{\Lambda}_I M^{-1}_{\mathbf{x}\cup x}\bar{\Lambda}_I}.
\end{aligned}$$

Now, use (13) and the relation $A^{-1} = \frac{{}^t(Com(A))}{\det(A)}$ to reach the conclusion of the lemma:

$$Tr\left({}^t Com(M_{\mathbf{x}\cup x}) M'_x\right) = Tr\left({}^t Com(M_{\mathbf{x}}) M'_x\right). \quad \square$$

Proof of theorem 1: We will note $F(x) = \det(M_{\mathbf{x}\cup x})$. Suppose first that $M_{\mathbf{x}}$ is non-singular and x is not a dyadic point described by the set $\mathcal{S}$. In this case, classical differentiation used with lemma 2 yields

$$F'(x) = Tr\left({}^t Com(M_{\mathbf{x}\cup x}) M'_x\right) = Tr\left({}^t Com(M_{\mathbf{x}}) M'_x\right).$$

Finally, since Tr is a linear map, we immediately get

$$\begin{aligned}
F"(x) &= Tr\left({}^t Com(M_{\mathbf{x}}) M"_x\right) \\
&= Tr\left({}^t Com(M_{\mathbf{x}})\bar{\Lambda}'_I\,{}^t\bar{\Lambda}'_I\right) \\
&= {}^t\bar{\Lambda}'_I\,{}^t Com(M_{\mathbf{x}})\bar{\Lambda}'_I \geq 0.
\end{aligned}$$

Thus F is a convex function on each interval outside of $\mathcal{S}$ Consequently, its maximum are located on some dyadic points of $\mathcal{S}$. This is equivalent to the assertion of the proposition.

Suppose now $M_{\mathbf{x}}$ is singular, we can find a sequence $M_{\mathbf{x},\epsilon_n} = M_{\mathbf{x}} + \epsilon_n Id$ which is non-singular such that

$$\lim_{n\mapsto+\infty} M_{\mathbf{x}} + \epsilon_n Id = M_{\mathbf{x}}.$$

Consider now the function $F_{\epsilon_n}(x)$ defined as

$$F_{\epsilon_n}(x) = \det(M_{\mathbf{x},\epsilon_n\cup x}).$$

We can use the same arguments as before to conclude that $\arg\max F_{\epsilon_n} \subset \mathcal{S}$ since these arguments only rely on a slight modification of lemma 2 which becomes:

$$\begin{aligned} F'_{\epsilon_n}(x) &= Tr\left({}^tCom(M_{\mathbf{x},\epsilon_n \cup x}) M'_x\right) \\ &= Tr\left({}^tCom(M_{\mathbf{x},\epsilon_n}) M'_x\right). \end{aligned}$$

Now, remark that $\mathcal{E}$ is a finite set which is not varying with ϵ_n and

$$\forall x \qquad F_{\epsilon_n}(x) \leq \max_{x \in \mathcal{E}} F_{\epsilon_n}(x).$$

Taking the limit in the relation above yields the conclusion of the proof. □

Proof of theorem 3: Remark first that $t \mapsto \det\left(tId + M^{-1}_{\mathbf{x} \cup x}\right)$ is a polynomial function of t whose degree p is the size of $\bar{\Lambda}_I$. This polynomial function is expanded in

$$\det\left(tId + M^{-1}_{\mathbf{x} \cup x}\right) = t^p - Tr\left(M^{-1}_{\mathbf{x} \cup x_1}\right) t^{p-1} + Q_x(t)$$

where $deg(Q_x) \leq p-2$. Now for $x_1, x_2 \in E$ such that

$$Tr\left(M^{-1}_{\mathbf{x} \cup x_1}\right) \geq Tr\left(M^{-1}_{\mathbf{x} \cup x_2}\right),$$

we can immediately check that for sufficiently large t, we have

$$\det\left(tId + M^{-1}_{\mathbf{x} \cup x_1}\right) \leq \det\left(tId + M^{-1}_{\mathbf{x} \cup x_2}\right).$$

Consequently, the solutions of the trace maximization problem are the same as the one deduced from the determinant minimization problem and this remark ends the proof. □

Remark 4 *To extend now the proof to dimensions d higher than one with some tensorized family of Schauder functions, one just have to remark that both lemma 1 and 2 are still valid for $x \in \mathbb{R}^d$. Then a similar argument to the one used in the proof of theorem 1 shows the convexity of F except in the neighbourhood of points of the form $(2^{-j_1}k_1, \ldots, 2^{-j_d}k_d)$.*

7.2 Convergence in the case of fixed basis

We detail here the convergence of the estimation of the parameter $\hat{\theta}$ with the strategy of sequential optimal design *when the basis I remains fixed*. As both previous criteria yield the same optimal design, we are only concerned with the study of the sequential strategy:

$$x_{n+1} = \arg\max_x \det(M_{\mathbf{x_n} \cup x})$$

and

$$\mathbf{x_{n+1}} = \mathbf{x_n} \cup x_{n+1},$$

while $\hat{\theta}$ is classically given by

$$\hat{\theta}_n = M^{-1}_{\mathbf{x_n}} \bar{\Lambda}_I(\mathbf{x_n}) f(\mathbf{x_n}).$$

Theorem 4 *Let f and η be defined by equations (2) and (3) with a fixed basis I, and suppose that $\eta \in Span(I)$. Then the sequential optimal design is consistent: $\hat{\theta}_n \to \theta$ a.s. Moreover, there exists a positive constant C such that*

$$\|\hat{\theta}_n - \theta\|_\infty \leq C\sqrt{\frac{\log n}{n}}.$$

Remark 5 *The previous theorem ensures the consistency of $\hat{\theta}_n$ if the signal η is a linear combination of the functions $(\Lambda_i)_{i \in I}$. Note that when $\eta \notin Span((\Lambda_i)_{i \in I})$, the convergence to the orthogonal projection of η into $Span((\Lambda_i)_{i \in I})$ also holds.*

Proof of theorem 4: This proof is inspired by [P00] which states the almost sure convergence of $\hat{\theta}_n$ to θ provided the two conditions

C1 $\lambda_{\min}[M_{\mathbf{x_n}}] \to \infty$ *a.s.*

C2 $\log\left(\lambda_{\max}\left[M_{\mathbf{x_n}}\right]\right) = o\left(\lambda_{\min}\left[M_{\mathbf{x_n}}\right]\right) a.s.$

where $\lambda_{\min}(M)$ denotes the minimum eigenvalue of M and $\lambda_{\max}(M)$ the maximum eigenvalue of M. We establish first the condition C1. Remark that as the functions $(\Lambda_i)_{i\in I}$ are linearly independent, we can find $\rho > 0$ such that

$$B(0,\rho) \subset \overline{Conv(\bar{\Lambda}_I(t), t \in E)} \cup \overline{-Conv(\bar{\Lambda}_I(t), t \in E)},$$

where $Conv$ denotes the convex hull of a set. Now, we have for any symmetric positive definite M

$$\max_{y\in B(0,\rho)} {}^t y M^{-1} y = \lambda_{min}(M)^{-1}\rho^2,$$

and since $y \mapsto {}^t y M^{-1} y$ is convex, we can state that

$$\max_{x\in E} {}^t\bar{\Lambda}_I(x) M^{-1} \bar{\Lambda}_I(x) \geq \frac{\rho^2}{\lambda_{min}(M)}. \tag{19}$$

Remark that all maps in $\bar{\Lambda}_I$ are continuous and E is compact, thus

$$\exists\, L > 0 \quad \forall t \in E \qquad \left\| {}^t\bar{\Lambda}_I(t)\bar{\Lambda}_I(t)\right\|_2 \leq L,$$

where $\|A\|_2 := \sup_{x\in B(0,1)} \|Ax\|$, where we take the Euclidean norm in the last definition. Now, the spectral radius satisfies the triangular inequality and

$$\lambda_{\max}\left(\frac{M_{\mathbf{x_k}}}{k}\right) \leq \frac{\sum_{i=1}^{k} \lambda_{\max}(\bar{\Lambda}_I(x_i)^t\bar{\Lambda}_I(x_i))}{k} \leq L.$$

If $I_k = M_{\mathbf{x_k}}/k$, the last inequality yields

$$\lambda_{\max}(I_k) \leq L. \tag{20}$$

Next define $\rho_k = \det(I_k)$ and $d_k(t) = {}^t\bar{\Lambda}_I(t) I_k^{-1} \bar{\Lambda}_I(t)$, from Proposition 1 equation (15), we have

$$\rho_{k+1} = \left(\frac{k}{k+1}\right)^p \left(1 + \frac{d_k(x_{k+1})}{k}\right)\rho_k \geq \rho_k \left(\frac{k}{k+1}\right)^p.$$

Thus, for any $\epsilon > 0$, we can find $K_1 \geq 1$ such that

$$\forall k \geq K_1 \qquad \rho_{k+1} \geq (1-\epsilon)\rho_k, \tag{21}$$

and a simple induction shows that $\rho_k \geq (1-\epsilon)^{k-K_1}\rho_{K_1}$. Let $A_k = (1-\epsilon)^{k-K_1}\rho_{K_1}$, since $A_k \to 0$ as $k \to \infty$, we can find $K_2 \geq K_1$ such that $\forall k \geq K_2$

$$\frac{\rho^2}{A_k^{1/p}} > 2p \qquad \text{and} \quad \left(\frac{k+1}{k}\right)^p \leq 1 + \frac{2p}{k}. \tag{22}$$

We show now by induction that ρ_k is bounded from below by $(1-\epsilon)A_{K_2}$ for sufficiently big k. This is obviously true for $k = K_2 + 1$.

Suppose now that $\rho_k \geq (1-\epsilon)A_{K_2}$. If $\rho_k \geq A_{K_2}$, in view of (21) we immediately obtain $\rho_{k+1} \geq (1-\epsilon)A_{K_2}$. We must thus study the case $A_{K_2} > \rho_k \geq (1-\epsilon)A_{K_2}$. From the definition of d_k and (19), we have

$$\max_{x\in E} d_k(x) \geq k\frac{\rho^2}{\lambda_{\min}(M_{\mathbf{x_k}})} \geq k\frac{\rho^2}{\det(M_{\mathbf{x_k}})^{1/p}} \geq \frac{\rho^2}{\rho_k^{1/p}}.$$

From equation (22) and our assumption on ρ_k, we obtain

$$\max_{x\in E} d_k(x) \geq \frac{\rho^2}{A_{K_2}^{1/p}} > 2p.$$

Finally, the definition of x_{k+1} yields

$$\rho_{k+1} = \rho_k \left(\frac{k}{k+1}\right)^p \left(1 + \frac{d_k(x_{k+1})}{k}\right) = \rho_k \left(\frac{k}{k+1}\right)^p \left(1 + \frac{\max_{x\in E} d_k(x)}{k}\right) \geq \rho_k.$$

This last inequality concludes the induction and ρ_k is bounded from below by a constant Γ. Now, remark that

$$\lambda_{\min}(I_k)\lambda_{\max}(I_k)^{p-1} \geq \det(I_k) \geq \Gamma,$$

and we obtain from equation (20) as $k \to +\infty$:

$$\lambda_{\min}(M_{\mathbf{x_k}}) \geq k\frac{\Gamma}{L^{p-1}} \to +\infty.$$

This last equation proves condition (C1).

For (C2), simple algebra yields as $k \to +\infty$:

$$\frac{\lambda_{\min}(M_{\mathbf{x_k}})}{\log\left(\lambda_{\max}(M_{\mathbf{x_k}})\right)} \geq \frac{k\Gamma}{L^{p-1}\log\left(kL\right)} \to \infty,$$

and this last equation proves condition (C2).

With notation of theorem 1 of [LW82], take $\delta = 0$ and apply now this theorem to conclude that

$$\|\hat{\theta}_n - \theta\|_\infty = O\left(\left[\frac{\log\left(\lambda_{\max}(M_{\mathbf{x_k}})\right)}{\lambda_{\min}(M_{\mathbf{x_k}})}\right]^{1/2}\right) = O\left(\sqrt{\frac{\log n}{n}}\right) \qquad \square$$

References

[AAP06] Amato U. and Antoniadis A. and Pensky M. (2006) Wavelet kernel penalized estimation for non-equispaced design regression. *Statistics and Computing*, 16, 37 – 55.

[BC02] Biswas A. and Chaudhuri P. (2002). An efficient design for model discrimination and parameter estimation in linear models. *Biometrika* 89(3), 709–718.

[BD59] Box G. E. P. and Draper N. R. (1959). A Basis for the Selection of a Response Surface Design. *Journal of the American Statistical Association*, 54(287) 622–654.

[BFI07] Busby D. and Farmer C. and Iske A. (2007) Hierarchical nonlinear approximation for experimental design and statistical data fitting. *SIAM J. Sci. Comput.* 29(1), 49–69.

[BG05] Blanchard G. and Geman D. (2005). Sequential testing designs for pattern recognition. *Annals of Statistics*, 33, 1155–1202.

[BT04] Berlinet A. and Thomas-Agnan C. (2004). Reproducing Kernel Hilbert Spaces in Probability and Statistics, Kluwer Academic Publishers, Boston.

[CK03] Castaño D. and Kunoth A. (2003). Adaptive fitting of scattered data by spline wavelets *Curve and surface fitting (Saint-Malo, 2002), Nashboro Press, Brentwood, TN*, 65–78.

[CT04] Candès E. and Tao T. (2004). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51, 4203 – 4215.

[CT07] Candès E. and Tao T. (2007). The Dantzig selector: Statistical estimation when p is much smaller than n. *The Annals of Statistics*, 35(6), 2313–2351.

[DJ94] Donoho D. L. and Johnstone I. M. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika.*

[DS97] Dette H. and Studden W. J. (1997). The theory of canonical moments with applications in statistics, probability, and analysis. *Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1997.*

[EJHT03] Efron B. and Johnstone I. and Hastie T. and Tibshirani R. (2003). Least angle regression. *The Annals of Statistics*, 32(2), 407-499.

[F69] Fedorov V.V. (1969), Theory of Optimal Experiments. *Academic Press, New York.*

[F91] Friedman J. H (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19,1–141.

[FW00] Fang Z. and Wiens D. P. (2000). Integer-valued, minimax robust designs for estimation and extrapolation in heteroscedastic, approximately linear models. *J. Amer. Statist. Assoc.*, 95 (451), 807–818.

[G95] Green P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

[GG84] Geman S. and Geman D. (1984). Stochastic Relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721–741.

[GWBV02] Guyon I. and Weston J. and Barnhill S. and Vapnik V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389-422.

[H88] Hajeck B. (1988), Cooling schedules for optimal annealing. *Math. Oper. Res.*, 13, 311–329.

[HM87] Hall P. and Marron J. (1987). Extent to which Least-square Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation. *Prob. Theory and Related Fields*, 74, 567–581.

[H95] Hwang R. (1995). Asymptotic properties of locally weighted regression. *Journal of Nonparametric Statistics*, 5 (3), 1029-0311.

[K01] Klinger A. (2001) Inference in high dimensional generalized linear models based on soft thresholding. *J. R. Statist. Soc. B*, 63(2), 377–392.

[KP04] Kerkyacharian G. and Picard D. (2004). Regression in random design and warped wavelets. *Bernoulli* 10(6), 1053–1105.

[KC87] Khuri A.I. and Cornell J.A. (1987), Response Surfaces: Designs and Analyses . *New York : Marcel Dekker, Inc.*

[KS66] Karlin S. and Studden W. J. (1966). Optimal experimental designs. *Annals of Mathematical Statistics*, 37, 783–815.

[KW59] Kiefer J. and Wolfowitz J. (1959). Optimum designs in regression problem. *Annals of Mathematical Statistics*, 30, 271–294.

[LW82] Lai T.L. and Wei C.Z. (1982). Least square estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10 (1), 154–166.

[LO06] Liu Y. and Oyet A. J. (2006). Exact Wavelet Designs for Discrimination. *Sankhya*, 68 (4), 569–586.

[M90] Meyer Y. (1990). Ondelettes et Oprateurs I. *Hermann.*

[MN95] Meyer R. K. and Nachtsheim C. J. (1995). The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs. *Technometrics*, 37(1), 60–69.

[OW03] Oyet A.J. and Wiens D.P. (2003). On exact minimax wavelet designs obtained by simulated annealing. *Statistics & Probability Letters*, 61, 111–121.

[OW06] Oyet A.J. and Wiens D.P. (2006). Exact Wavelet Designs for Discrimination. *Sankhya, The Indian Journal of Statistics*, 68 (4), 569–586.

[P00] Pronzato L. (2000). Adaptive optimization and D-optimum experimental design. *Ann. Statist.* 28 (6), 1743–1761.

[S85] Silverman B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, 47, 1–52.

[T96] Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (B)*, 58 (1), 267–288.

[W92] Wiens D.P. (1992). Minimax designs for approximately linear regression. *Journal of Statistical Planning and Inference*, 31, 353–371.

[Z06] Zhou H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Society*, 101, 1418–1429.

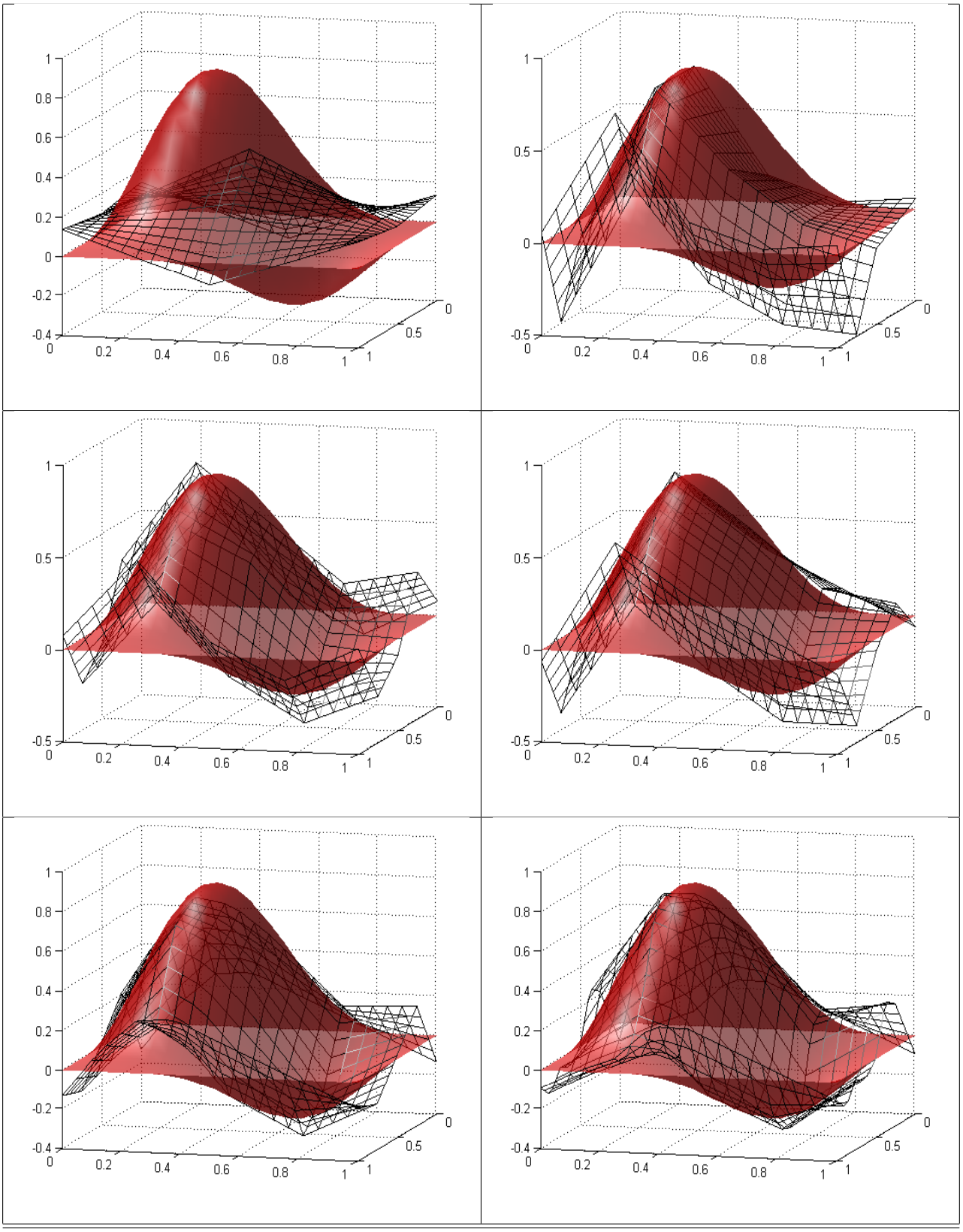

Figure 13: Estimations of the 2 dimensional signal with our adaptive sequential design method for 1, 6 (left and right top), 12, 18 (left and right middle), 24 and 30 (left and right bottom) iterations.