# Maximum Entropy Sampling by Branch and Bound

Henry P. Wynn and Noha A. Youssef

London School Of Economics

# 1 Introduction

Two main topics are considered in experimental design first of them is how to construct a design and the second is how to analyze it. The main concern of this report is to select the design points at which we conduct the experiment and collect data. Many criteria have been introduced in the literature for selecting a design, where each criterion exists for a specific purposes of the design of experiment. Maximum entropy sampling is a commonly used criterion when the most informative design with respect to the parameters is of our interest. Also, it is considered a kind of combinatorial optimization.

This report presents in Section 2 a brief background about the kriging model. Section 3 explains the entropy measure and the resulting maximum entropy sampling criterion. Because of the noticeable role of the covariance structure while applying the maximum entropy sampling in the Gaussian case Section 4 introduces examples of covariance structures associated with both stationary and non-stationary Gaussian process. The Branch and Bound algorithm and its use in the maximum entropy sampling are described in Section 5. Examples of upper bounds are also introduced in Section 5. Section 6 explains the main steps of the algorithm and discusses the results of applying it. Conclusions are summarized in Section 7.

# 2 Kriging Model

Kriging model is a form of Gaussian process regression. The use of Kriging model in computer experiments has been introduced by Sacks et al. (1989). This model has some characteristics that fulfil the requirements of modeling the computer experiment of being a deterministic model. Deterministic model ensures that the output is always fixed for the same the inputs, i.e. there is no random error.

Kriging model consists of two components. The first component is a general linear model while the second is treated as a realization of a stationary Gaussian random function Koehler and Owen (1996).

Let $S = [0,1]^p$ be the design space and $x \in S$ be a scaled $p$ dimensional vector of input values. The Kriging approach models the associated response as

$$Y(x) = \sum_{j=1}^{k} \beta_j h_j(x) + Z(x) \tag{1}$$

where the $h_j$'s are known fixed functions, the $\beta_j$'s are unknown coefficients and $Z(x)$ is a stationary Gaussian random function with $E[Z(x)] = 0$ and covariance

$$\text{Cov } [Z(x_i), Z(x_j)] = \sigma^2 R(x_j, x_i)$$

where $\sigma^2$ is the process variance and $R(x_j, x_i)$ is the correlation between $Z(x_i), Z(x_j)$.

If the frequentist approach is adopted then we have

$$\begin{aligned} \mathrm{E}(Y(x)) &= f(x)^T\beta \\ \mathrm{cov}(Y(x_1), Y(x_2)) &= \sigma^2 R(x_j, x_i) \end{aligned}$$

if there is available information about $\beta$ so that $\mathrm{E}(\beta) = \mu$ and $\mathrm{cov}(\beta) = \Phi$, then according to the Bayesian approach we have

$$\begin{aligned} \mathrm{E}(Y(x)) &= \mu \\ \mathrm{cov}(Y(x_1), Y(x_2)) &= \sigma^2(H'\Phi H + R(x_j, x_i)) \end{aligned}$$

where $H$ is the vector of $h_j$, $j = 1, ...k$, and $\Phi$ is the variance-covariance of $\beta$ The Bayesian approach and the frequentist one yield identical estimators if the prior distribution of $Z(\cdot)$ is Gaussian and if the prior distribution of the $\beta_j$'s is diffuse.

# 3 Maximum Entropy Design

In the context of experimental design, there exist different criteria that match different goals, e.g. D optimality, A optimality, integrated mean squared error and entropy. Entropy criterion is used for the purpose of choosing the design that maximizes $\xi$ from the set of all possible designs $\Xi$ to acquire the maximum amount of information about $\Theta$ the model parameters.

Entropy is defined as the negative measure of information,

$$\mathrm{Ent}(Y(x)) = \mathrm{E}_{Y(x)}[-\log p(Y(x))] \tag{2}$$

where $Y(x)$ is a random vector, $p(\cdot)$ is a density function of $Y(x)$ and $x = x_i$. For simplicity we suppress $Y(x)$ to $Y$ .

(Shewry and Wynn, 1987)show that maximizing the information about the parameters is equivalent to minimizing the information for prediction at unsampled sites. In addition, Sebastiani and Wynn (2000) prove that the experiment which maximizes the entropy of the marginal distribution of $Y$ will be most informative for $\Theta$. Following the Bayesian approach, we will choose $\xi$ in $\Xi$ that minimizes the overall expected risk

$$\mathrm{E}_Y \mathrm{Ent}(\Theta|Y, \xi)$$

which will be the most informative for the Bayesian estimation of $\Theta$ or optimal.

Hence, $Y$ can be decomposed into $(Y_s, Y_{\overline{s}})$ where $s$ represents the selected index set $s \subseteq \{1, ..., N\}$ where $N$ is the design space size or in other words the number of the grid points we choose from, so we have

$$\mathrm{Ent}(Y) = \mathrm{Ent}(Y_s) + \mathrm{E}_{Y_s}\mathrm{Ent}(Y_{\overline{s}}|Y_s). \tag{3}$$

Throughout this report the Kriging model is used to model the response $Y(x)$ where $Y(x)$ is assumed to have Gaussian distribution,

$$\mathrm{Ent}(Y|\xi) = \frac{n}{2}[1 + \log 2\pi] + \frac{\log|\Sigma|}{2}, \tag{4}$$

hence, a priori information about the covariance structure is required since maximizing $\mathrm{Ent}(Y|\xi)$ is equivalent to maximizing the determinant of the variance of $Y|\xi$.

The maximum entropy problem in the Gaussian case can be formulated as

$$\begin{aligned} z &= \max_{S \subset N:|S|=s} \log\det C[S]; \\ \text{subject to} \sum_{j \in S} a_{ij} &\leq b_i, \forall i \in M. \end{aligned} \tag{5}$$

where $M$ is a finite index set, $a_{ij}$ and $b_i$ are real numbers represent the cost function associated with each point, for $i \in M$ and $j \in N$.

# 4 Covariance Structures

From the above section we notice the role the covariance structures play in constructing a design. Standard Gaussian process models use a stationary covariance, in which the covariance between any two points is a function of Euclidean distance. Sometimes functions may vary more in some parts of the input than in others (Paciorek and Schervish, 2006). This leads to the use of covariance structures associated with non-stationary Gaussian process. Examples of covariance structures are presented in this section for both stationary and non-stationary Gaussian process.

## 4.1 Covariance Structures For Stationary Gaussian Process

**Exponential Structure (SSc)** This structure is used to describe covariance for stationary Gaussian process. It takes the following form

$$\text{cov}(Z(x_i), Z(x_j)) = \sigma^2 \prod_{k=1}^{p} \exp(-\theta_k |x_{ki} - x_{kj}|^{q_k}) \tag{6}$$

where $(x_i, x_j)$ are any two sampled sites in $[0, 1]$, $0 < q_k \leq 2$ and $\theta_k \in (0, \infty)$. If $q_k < 2$ then these processes are not mean square differentiable.

**Gaussian Structure** This function is a special case of the exponential correlation function when $q = 2$. It takes the following form

$$\text{cov}(Z(x_i), Z(x_j) = \sigma^2 \prod_{k=1}^{p} \exp(-\theta_k |x_{ki} - x_{kj}|^2). \tag{7}$$

## 4.2 Covariance Structures For Non-stationary Gaussian Process

**Brownian Sheet Covariance Matrix** Brownian sheet is a one dimensional centered Gaussian process that is $B = \{B(t)\}_{t \in R_+^p}$, where $R_+^p$ is $p$ dimensional positive real numbers, whose covariance is given by

$$\text{Cov}(B(i), B(j)) = \prod_{k=1}^{p} \min(i_k, j_k) \tag{8}$$

We can rewrite this formula in the same notation we use in this report, so

$$\text{Cov}(Z(x_i), Z(x_j)) = \prod_{k=1}^{p} \min(x_{ik}, x_{jk}) \tag{9}$$

**Covariance Function as Linear Combination of Basis Functions** Any realization of a Gaussian process can be expressed as a fixed set of $m$ basis functions $Z(x) = \sum_{i=1}^{m} \lambda_i \psi_i(x)$. Hence every covariance function can be expressed as an expansion in terms of basis functions. Haar Wavelet basis function can be used for this purpose. It takes the following form

$$\psi(x) = \begin{cases} 1, & \text{if } 0 \leq x < \frac{1}{2}; \\ -1, & \text{if } \frac{1}{2} \leq x < 1; \\ 0, & otherwise. \end{cases} \tag{10}$$

*Example:* Let $y = \beta + \theta_1\psi(x_1) + \theta_3\psi(x_2) + \theta_3\psi(x_1)\psi(x_2)$, hence the

$$\text{cov}(y(x_i), y(x_j)) = \sigma^2_\beta + \sigma^2_1\psi(x_{1i})\psi(x_{2i}) + \sigma^2_2\psi(x_{2i})\psi(x_{2j}) + \sigma^2_3\psi(x_{1i})\psi(x_{2i})\psi(x_{1j})\psi(x_{2j})$$

where $\psi$ is the Haar basis function of form(10), $\sigma^2_\beta$ is the variance of $\beta$, $\sigma^2_i$, $i = 1, 2, 3$, are the corresponding variances to the Haar coefficients.

# 5 Branch and Bound for MES

## 5.1 Principles of Branch and Bound

Branch and Bound algorithm is a commonly used technique in several mathematical programming applications especially in combinatorial problem when a problem is difficult to be solved directly (Lawler and Wood, 1966). It is preferred over many other algorithms because it reduces the amount of search needed to find the optimal solution.

Branch and Bound creates a set of subproblems by dividing the space of current problem into unexplored subspaces represented as nodes in a dynamically generated search tree, which initially contains the root (the original problem). Performing one iteration of this procedure is based on three main components: selection of the node to process, bound calculation, and branching, where branching is the partitioning process at each node of the tree and bounding means finding lower and upper bounds to construct a proof of optimality without exhaustive research. The algorithm can be terminated whenever the difference between the upper bound and the lower bound is smaller than the chosen $\epsilon$ or if the set of live nodes is empty, i.e. there is no unexplored parts of the solution space left, and the optimal solution is then the one recorded as "current best".

The key point of this algorithm is that if a sub-solution does not improve the current best solution, then it is pruned from the set of live problems. The steps of the algorithm can be reduced by choosing a good bounding function which gives values close to the optimal value for the subproblem bounded.

In the context of experimental design, Branch and Bound has been used to search for the designs based on $\mathcal{D}$ optimality (Welch, 1982). Assuming the Gaussian case, Ko et al. (1995) apply Branch and Bound algorithm for MES as an exact algorithm. As mentioned above, in the Gaussian case, the problem is reduced to the maximization of the determinant of the covariance matrix of the chosen design or the logarithm of this determinant.

Given a set of $N$ of $n$ points called the design space or the candidate set, a set $F \subset N$ of $f$ forced points to be in the required design, $E$ is a set of points we select from and a design of size $s$, such that $f \leq s \leq n$.

We aim at choosing a set $S$ of $s$ points satisfying $F \subset S \subset N$, such that the selected design is the most informative among the all possible designs.

All computations for the entropy criterion are based on the covariance matrix $C$ of order $n \times n$ which is the covariance for all points in the design space. The goal is to find the sub-matrix $C[S]$ of order $s \times s$, whose rows and columns are indexed by the indices of $S$, with the largest determinant. Finding such a matrix lies in the category of combinatorial optimization. Using Branch and Bound algorithm requires finding good upper bounds for our target function $log|C[S]|$.

## 5.2 Upper Bounds For MES

To overcome the difficulty of finding the maximum entropy, many authors search for different types of bounding functions that speed up the implementation of the algorithm. Upper bounds for Branch and Bound are defined whether the problem is with side constraints or not. This section reviews some upper bounds introduced in the literature.

### 5.2.1 Upper Bounds For Unconstrained MES

**Diagonal Bounds** Diagonal bounds are defined as

$$\phi = \max \log |C[S]| \leq \sum_{l=1}^{s} \log(C_{[ll]}) \tag{11}$$

where $C_{[ll]}$ is the $l^{th}$ greatest diagonal element of $C$. So, we can calculate $\phi$ in the theoretical sense.

**Spectral Bounds for MES (Ko et al., 1995)** Spectral bounds are defined as

$$\phi = \max \log |C[S]| \leq \sum_{l=1}^{s} \log \lambda_l(C) \tag{12}$$

where $\lambda_l$ denotes the $l^{th}$ greatest eigenvalue.

Ko et al. (1995) develop an algorithm finding an upper bound function based on the interlacing property of the eigenvalues of a symmetric matrix.

**Lemma 1** *given a symmetric matrix $B$ with rows and columns indexed by $N$, $R = \{1, 2, ..., r\}$ and $B_r = B[R]$, for $1 \leq r \leq n-1$, then the interlacing property holds*

$$\lambda_{r+1}(B_{r+1}) \leq \lambda_r(B_r) \leq \lambda_r(B_{r+1}) \leq ... \leq \lambda_2(B_{r+1}) \leq \lambda_1(B_r) \leq \lambda_1(B_{r+1}). \tag{13}$$

It follows that if $F = \phi$ then the upper bound function is

$$Ub(C, F, E, s) = (\prod_{i=1}^{s} \lambda_i(C[E]) \tag{14}$$

but if not then the upper bound is

$$Ub(C, F, E, s) = \det C[F] \prod_{i=1}^{s-f} \lambda_i(B(F, E)). \tag{15}$$

#### 5.2.2 Upper Bounds For Constrained MES

In the case of constrained entropy, which is

$$\begin{aligned} z &= \max_{S \subset N : |S| = s} \log |\Sigma| \\ \text{s.t} \sum_{j \in S} a_{ij} &\leq b_i, \forall i \in M. \end{aligned}$$

the problem becomes more difficult because the upper bound in this case is also the solution of an optimization problem, which means that at every iteration the upper bound is obtained by solving another optimization problem. Most of these bounds are based on Fischer's inequality,

**Lemma 2** *If $B$ is a square symmetric positive semidefinite matrix with rows columns indexed from the cardinality set $S^*$ and $S_1, S_2, ..., S_s$ is a partition of $S^*$, then* $\det B \leq \prod_{k=1}^{s} \det B[S_k, S_k]$.

**Partition Bounds** Hoffman et al. (2001) suggest partition bounds as upper bounds for constrained MES.

$$\begin{aligned} \phi = \max \log |C[S]| &\leq \max \sum_{l=1}^{s} \ln \det C[S_l]; \\ \text{subject to } S_k &\subset N, \forall\ k = 1, 2, ..., s; |S_k| = \pi_k, \forall\ k = 1, 2, ..., s; \\ S_k \bigcap S_{k'} &= \Phi, \forall\ 1 \leq k \leq k' \leq s; \\ \sum_{k=1}^{s} \sum_{j \in S_k} a_{ij} &\leq b_i, \forall\ i \in M. \end{aligned} \tag{16}$$

where $\prod = \pi_1, \pi_2, ..., \pi_s$ be a partition of $s$, that is a multiset of non-negative integers such that $\sum_{k=1}^{s} \pi_k = s$, $S_1, S_2, ..., S_s$ is a partition of $S$.

**Spectral Partition Bound** Hoffman et al. (2001) also define a new upper bound called spectral partition bound as

$$\phi = \max \log |C[S]| \leq \sum_{l=1}^{s} \ln \Lambda_l(\mathcal{N}) \tag{17}$$

where $\mathcal{N} = \{N_1, N_2, ..., N_n\}$ denotes any partition of $N$, $\Lambda(N_k)$ is the multiset of $|N_k|$ eigenvalues of $C[N_k]$. $\Lambda_0(\mathcal{N})$ denotes the multiset union of $|N| = n$ elements from the sets $\Lambda(N_k)$. For a multiset $\Lambda(.)$, $\Lambda_l(.)$ denotes $l^{th}$ greatest element. MES problem turns to be finding the best partition $\mathcal{N}$ that minimize the upper bound. It is easy to show that $\phi(\mathcal{N})$ is an upper bound for $z$, i.e. $z \leq \phi(\mathcal{N})$.

*Proof*

$$\begin{aligned}
&\ln \det C[S] \\
&\leq \quad \sum_{k=1}^{n} \ln \det C[S \bigcap N_k] \\
&= \quad \sum_{k=1}^{n} \sum_{\lambda \in \Lambda(S \bigcap N_k)} \ln \lambda \\
&\leq \quad \sum_{k=1}^{n} \sum_{l=1}^{|S \bigcap N_k|} \ln \Lambda_l(N_k) \\
&\leq \quad \sum_{l=1}^{s} \ln \Lambda_l(\mathcal{N}).
\end{aligned} \tag{18}$$

**Masked Spectral Bound** Burer and Lee (2007) introduce another bound based on Oppenheim's inequality which is defined as

$$\det A \leq \frac{\det A \circ B}{\prod_{j=1}^{n} B_{jj}}$$

called masked spectral bound defined as

$$\phi(X) = \min \sum_{l=1}^{s} \ln(\lambda_i(C \circ X)). \tag{19}$$

where $X$ is the mask matrix which is symmetric and positive semidefinite and $\circ$ denotes the element-wise product (Hadamard product). It is considered a generalization to diagonal, eigenvalue, and spectral partition bounds. The aim is to minimize the masked spectral bound over all masks. This can be done using affine scaling algorithm.

# 6 Application of Branch and Bound for MES

## 6.1 Structure of The Algorithm

In this report the Branch and Bound algorithm is applied to find the maximum entropy design by using the spectral bounds.

Let $k$ stand for the iteration number, $U_k$ stand for the upper bound at $k^{th}$ iteration, $\mathcal{I}$, the incumbent,i.e. for best current value of the target function and $\mathcal{L}$ be the set of all unexplored subspaces. The algorithm can be summarized as follows;

**Inputs** $N$ the candidate set of points, known $\sigma_Y^2$, known $\sigma_\beta^2$, $C[N]$ known covariance matrix of all candidate points, $E$ the set of all eligible points, $F$ the set of all points forced to be in the design and $\epsilon > 0$ a small chosen number.

**Initialization** $k = 0$, initial design $S_0$, $\mathcal{I} = \log\det C[S_0]$, $U_0 = Ub(C, F_0, E_0, s)$, $\mathcal{L} = \{(C, F, E, s, Ub)\}$.

**Step 1** Remove the problem, tuple, with $\max Ub$ from $\mathcal{L}$ in order to be explored.

**Step 2** Branch the problem according to these conditions

- If $|F| + |E| - 1 > s$, compute $Ub(C, F, E \setminus i, s)$ and if $Ub(C, F, E \setminus i, s) > U_k$, where $i$ is any index selected to be removed from $E$, then add $(C, F, E \setminus i, s)$ to $\mathcal{L}$, else if $|F| + |E| - 1 = s$, then set $S = F \cup E \setminus i$ and compute $\log\det C[S]$, and if $\log\det C[S] > \mathcal{I}$ then set $\mathcal{I} = \log\det C[S]$ and the current design is $S$.
- If $|F|+1 < s$, compute $Ub(C, F\cup i, E\setminus i, s)$ and if $Ub(C, F\cup i, E\setminus i, s) > U_k$, then add $(C, F, E\setminus i, s)$ to $\mathcal{L}$ else if $|F| + 1 < s$, then set $S = F \cup i$ and compute $\log\det S$ and if $\log\det C[S] > \mathcal{I}$ then set $\mathcal{I} = \log\det S$ and the current design is $S$.

**Step 3** Set $k = k + 1$.

**Step 4** Update $U_{k+1}$ to be the highest upper bound in $\mathcal{L}$.

Repeat steps $2, 3, 4$ while $U_k - \mathcal{I} > \epsilon$.

**Outputs** A set $S$ with largest value of entropy.

Figure 6.1 summarizes the steps of the Branch and Bound algorithm for maximum entropy sampling.

### 6.2 Results

For implementing the Branch and Bound algorithm, a program has been written in MATLAB. This code has been applied to choose a 6 point design from a grid of 16 points over the interval $[0, 1]^2$. Different covariance structures have been used to illustrate how the designs can differ by using different covariance structures. Figure 2 shows a maximum entropy design of six points with exponential covariance structure with $q = 1$. The design reflects the reality that the entropy aims at maximizing the determinant of the variance, where some of the points occupy the four corners of the grid and the remaining points lay on the diagonal of the grid. While for the Gaussian covariance structure, see Figure 3, we can observe that most of the design points are pushed towards the borders of the grids, and very few points are pushed to the center of the grid. Some computational difficulties have been encountered when using the Brownian sheet covariance matrix. It has been noticed that when the grid points we choose from begin from zero, this will lead to the singularity of the covariance matrix. To overcome this problem we have changed the starting point of the grid to be a number other than zero. However, this is not a solution to the problem. Different solutions can be suggested here, one of them is to remove the zero eigenvalues that causes the singularity problem. Figure 4 shows a design of six points which most of the design points are concentrated in the upper right corner of the grid. This design is an expected design because the largest variance of the Brownian sheet process exists at the very end points. Also, a linear combination of Haar basis are used to express a covariance structure corresponds to a stochastic process of linear combination of Haar basis. We can observe from Figure 5 that the design corresponding to this type of covariance is uniformly distributed over the grid.

## 7 Conclusions

Finding a model based optimal design is a crucial issue in context of experimental design. The most informative design can be obtained using entropy criterion. Though, it is not easy to find this design due the computational difficulties we can encountered. Full search for this design is time consuming. Cheap algorithms such as greedy or exchange algorithms can help finding this design, but they do not guarantee the optimal solutions. Using the Branch and Bound algorithm to find such a design is helpful, although it has some drawbacks especially when we are searching for a design from high cardinality design space. Search for sharper upper bounds functions, i.e. its values are closer the target function, is needed in order to speed up the algorithm. Lots of work is needed to improve the performance of the Branch and Bound algorithm for maximum entropy sampling. Also, the use of Branch and Bound can be extended to include other criteria.

# References


Burer, S. & Lee, J. (2007). Solving maximum-entropy sampling problems using factored masks. *Math. Program.*, 109(2-3, Ser. B):263–281.

Hoffman, A., Lee, J., & Williams, J. (2001). New upper bounds for maximum-entropy sampling. In: *mODa 6—advances in model-oriented design and analysis (Puchberg/Schneeberg, 2001)*, Contrib. Statist., pages 143–153. Physica.

Ko, C.-W., Lee, J., & Queyranne, M. (1995). An exact algorithm for maximum entropy sampling. *Oper. Res.*, 43(4):684–691.

Koehler, J. R. & Owen, A. B. (1996). Computer experiments. In: *Design and analysis of experiments*, volume 13 of *Handbook of Statist.*, pages 261–308. North-Holland.

Lawler, E. L. & Wood, D. E. (1966). Branch-and-bound method: A survey. *Operations Res.*, 14:699–719.

Paciorek, C. J. & Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–435. With comments and a rejoinder by the authors.

Sebastiani, P. & Wynn, H. P. (2000). Maximum entropy sampling and optimal Bayesian experimental design. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(1):145–157.

Shewry, M. C. & Wynn, H. P. (1987). Maximum entropy sampling. *Journal of Apply Statistics*, 14:165–170.

Welch, W. J. (1982). Branch and bound search for experimental designs based on d optimality and other criteria. *Technometrics*, 24(1):41–48.

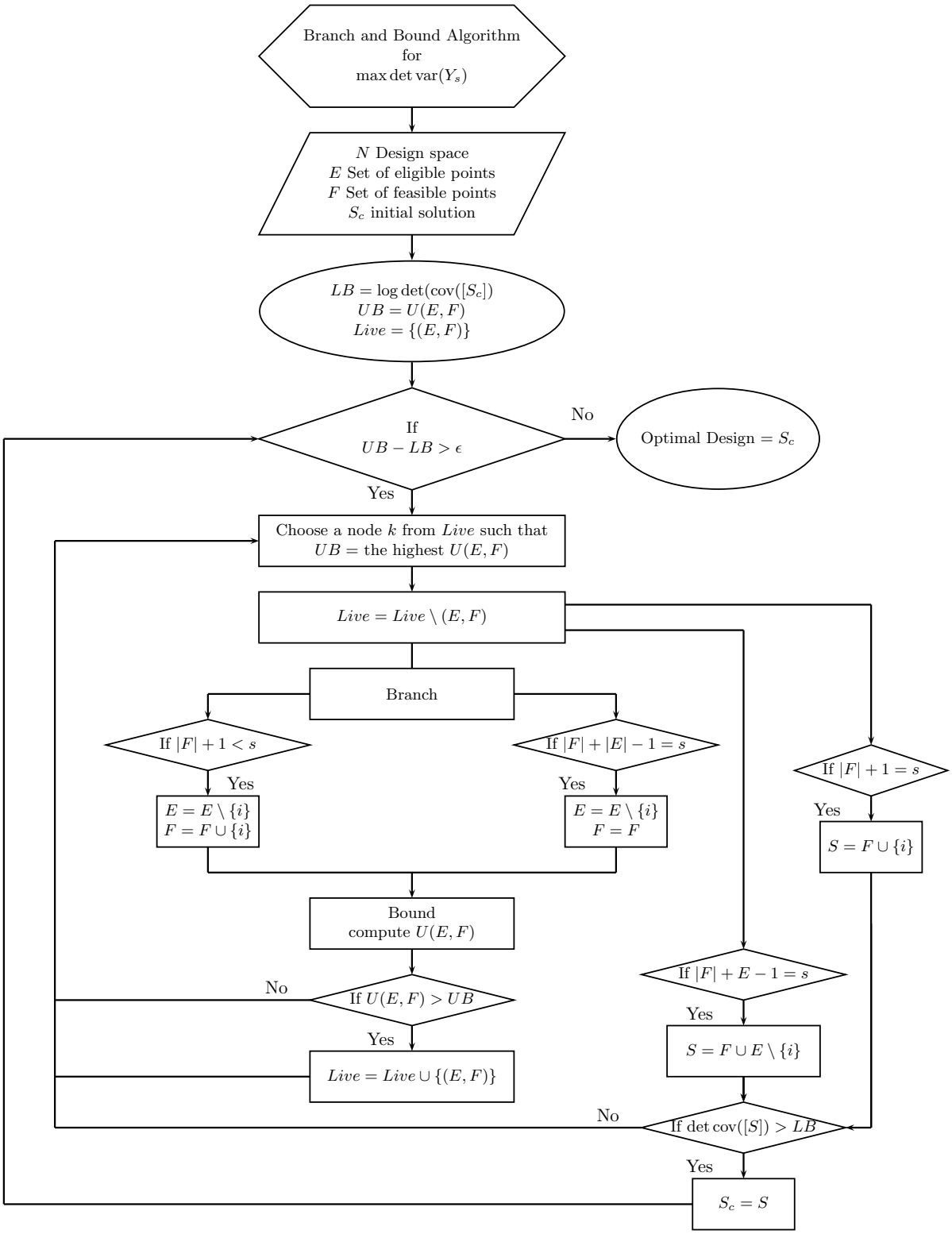


Figure 1: A Flow Chart for Branch and Bound Algorithm for MES

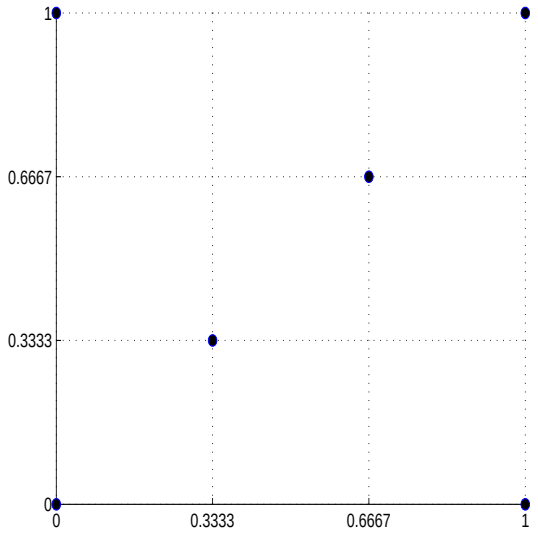


Figure 2: A six-point maximum entropy design for the exponential covariance structure

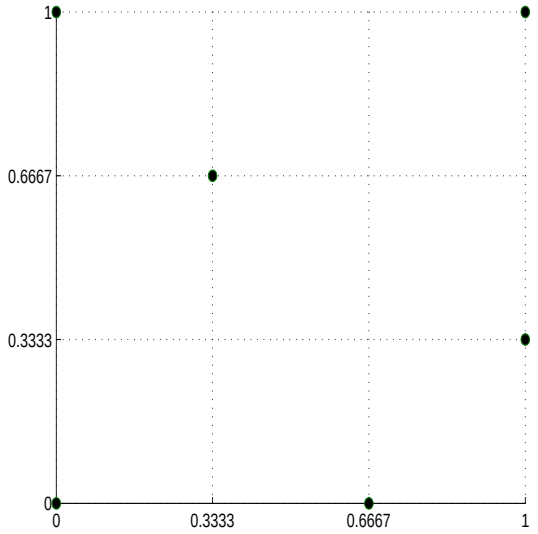


Figure 3: A six-point maximum entropy design for the Gaussian covariance structure

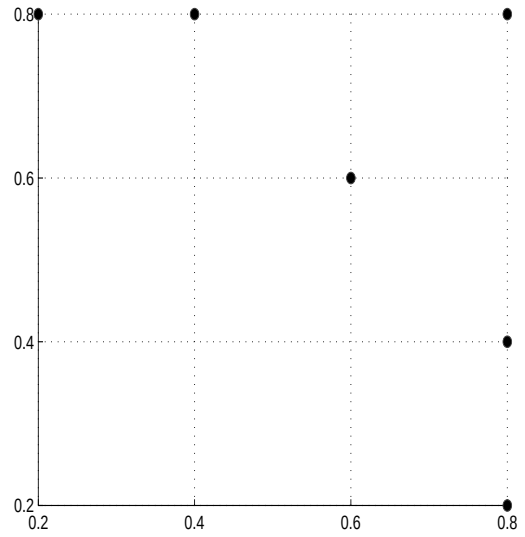


Figure 4: A six-point maximum entropy design for the Brownian sheet covariance structure

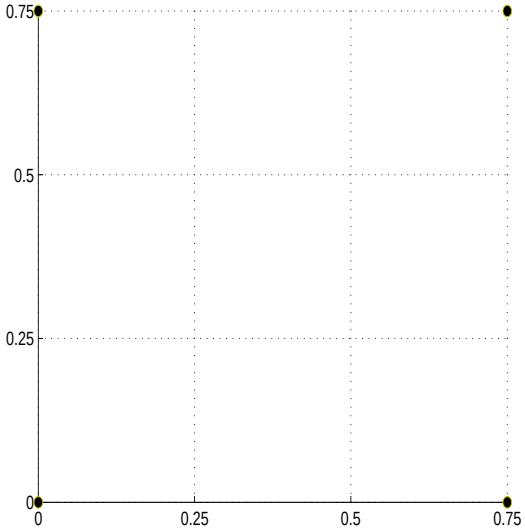


Figure 5: A four-point maximum entropy design when the covariance structure is a linear combination of Haar basis