Towards ASCM

Adaptive Sampler For Complex Models

Noha A. Youssef

London School Of Economics

1 Introduction

Optimal experimental design is used to obtain better inference about the unknown quantities in a statistical model. Obtaining an optimal design can be considered as a decision making problem. The need of sequential optimal design arises to improve the quality of inference by making use of the new data. Obtaining this design involves more than one stage. Each stage introduces a new design which in turn results in new information used to update the prior information we have before performing this stage. This means that we are adapting or adjusting the experiment at each stage. That is why we will call this procedure ASCM (Adaptive Sampler for Complex Models). The main steps of this procedure are illustrated in Figure 1. The Bayesian framework allows us to implement ASCM by providing a link between the before and after of an experiment, via prior and posterior distribution. This adaptation setting has different components,

- modeling
- integrating
- optimizing
- experimenting
- updating

This report provides the basics of Bayesian sequential experimental design for Gaussian processes. It can be taken as a rough outline of my PhD under the MUCM grant. It highlights the role of the Bayesian inference methods to generate the design with the support of optimization and numerical techniques. Section 2 describes the model used to express the response of the experiment and a method of reducing the Gaussian process. A brief overview of the Bayesian inference for the linear models is given in Section 3. Optimality implies a specific goal to be fulfilled, so that different utility functions corresponding to different goals

are introduced in Section 4. Section 5, shows numerical techniques might be needed to get the criterion, on which a design is chosen, in case of intractable integrations. Section 6, explains how the optimization search methods can be employed to get the optimal design at each stage. An illustrative example of getting the ASCM design for a maximum entropy criterion is given in Section 7. A summary of ASCM features is given in Section 8.

2 Modeling

Following the same notation given by (Lindley and Smith, 1972), the response of an experiment can be expressed using the following model

$$Y(x) = \eta(x, \theta) + \epsilon$$

where $Y(x)$ is the model response or the process, $\eta(x, \theta)$ is a function in the factors x and the parameters θ. The form of $\eta(x, \theta)$ determines the type of the model whether it is linear or nonlinear. Linearity is fulfilled when $\eta(x, \theta)$ can be written in the form $X\theta$, where X is called the design matrix, if not, the model is nonlinear.

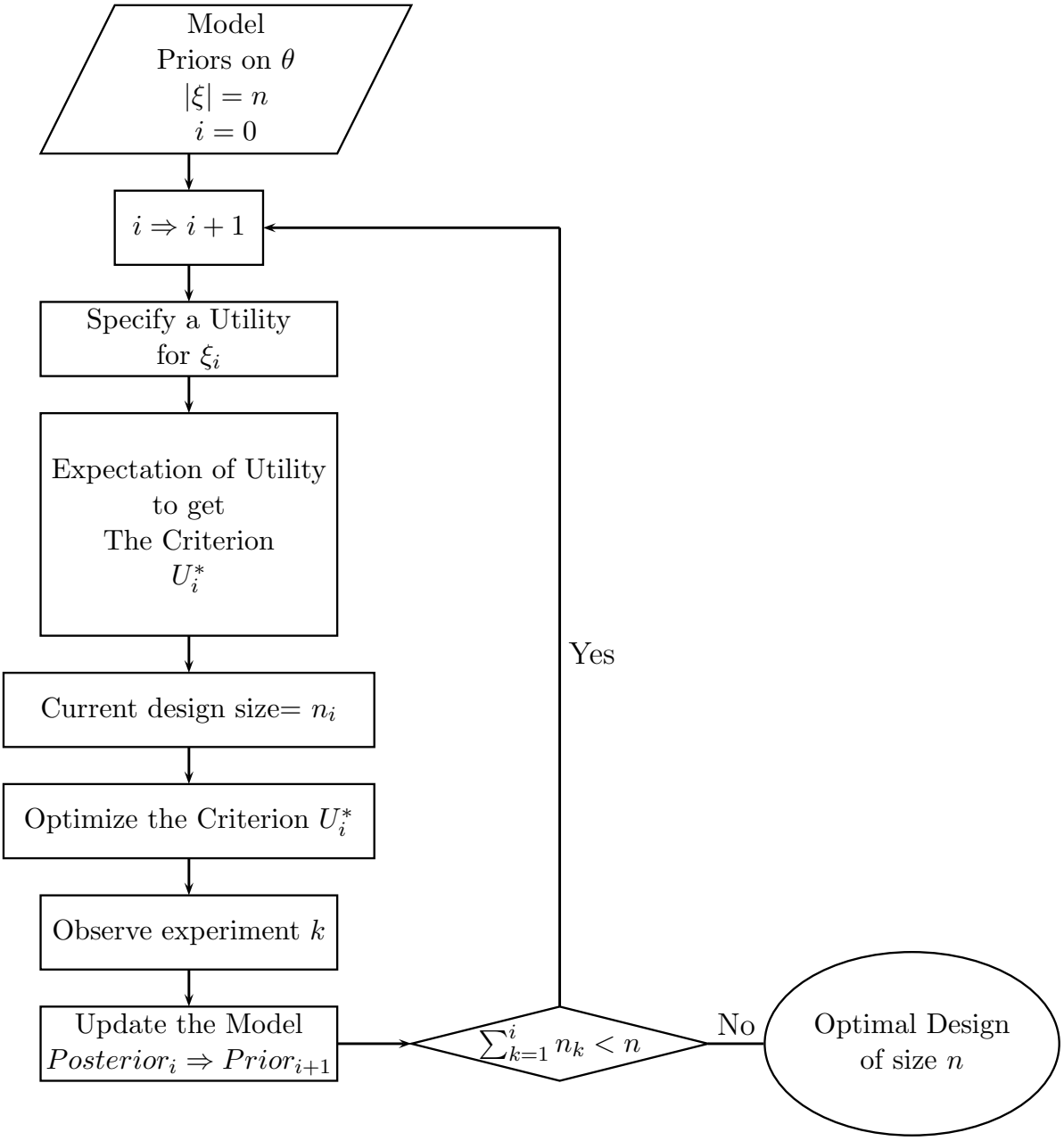
More generally, the model used in computer experiments is

$$Y(x) = f(x)^T\theta + Z(x) \tag{1}$$

where $Y(x)$ is a realization of a stochastic process usually assumed to be a Gaussian process, $f(x)$ is a p vector of known regression functions, θ is a p vector of unknown parameters and $Z(x)$ is a stochastic process with mean 0, variance σ^2 and correlation function $R(x_i, x_j)$, so the covariance between $Z(x_i)$ and $Z(x_j)$ is given by

$$Cov(Z(x_i), Z(x_j)) = \sigma^2 R(x_i, x_j).$$

Figure1: An illustrative flowchart for ASCM procedure



Wang (2008) shows that under suitable conditions, we can produce the Karhunen-Loeve expansion of the process $Z(x)$

$$Z(x) = \Sigma_{k=1}^{\infty} \sqrt{\lambda}_k \varepsilon_k g_k(x) \tag{2}$$

where λ_k and $\{g_k(\cdot)\}$ are the eigenvalues and the eigenfunctions of the spatial covariance matrix of the process $Z(x)$ and $\{\varepsilon_k\}$ is an uncorrelated zero mean unit variance process. The expansion of $R(x_i, x_j)$ is then given by

$$R(x_i, x_j) = \Sigma_{k=1}^{\infty} \lambda_k g_k(x_i) g_k(x_j). \tag{3}$$

It is attractive to approximate $Z(x)$ by truncating the expansion of $Z(x)$ at a suitable index, so

$$Z(x) \approx \Sigma_{k=1}^{p} \sqrt{\lambda}_k \varepsilon_k g_k(x) \tag{4}$$

and

$$R(x_i, x_j) \approx \Sigma_{i=1}^{p} \lambda_k g_k(x_i) g_k(x_j). \tag{5}$$

Our intention now is to approximate $Y(x)$ by

$$Y(x) = f(x)^T \theta + g(x)^T \phi \tag{6}$$

where $g(x)^T = (g_1(x), g_2(x), ..., g_p(x))$ and ϕ are the new parameters resulting form the approximated expansion.

Sometimes, $g_k(x)$ cannot be derived in a closed form. Thus a good approximation is required for these function. Haar functions are computationally feasible and better, for example, than the approximation using Fourier series. This will lead to the following scheme,

1. Approximate the covariance function using the karhunen-Loeve expansion.

2. Give both θ and ϕ a hierarchical Bayesian structure.

This will enable us to combine the main features of Krigning and those of Bayesian hierarchical linear model.

3 Bayesian Inference for Normal Linear Regression Model

For simplicity of explanation, we will assume that $Y \sim \mathcal{N}(X_{n\times p}\beta_{p\times 1}, \sigma^2 I_{n\times n})$. Assuming that the mean parameters are unknown, Bayesian inference for normal linear model has two common cases, the first is when the error variance is known and the second is when the error variance is unknown. The main target here is to obtain the posterior distribution that matches the criterion of our design, whether it is for the inference about the parameters or for prediction about the response.

3.1 The Case of Known Variance

Note that the number of factors is d not p, where p is the number of functions in d factors. Bayesian approach suggests assigning prior distribution to the unknown parameters. The prior for the coefficients β is assumed to be,

$$\beta \sim \mathcal{N}(m, V)$$

where V is assumed to be positive definite. Also, assuming that $(XX)^{-1}$ exists then the LSE $\hat{\beta} \triangleq (XX)^{-1}X^T y$ can be used to derive the Bayesian estimate for the coefficient. The posterior distribution of β would be normal distribution with mean

$$m^* = (V^{-1} + \sigma^2 X^T X)^{-1}(V^{-1}m + \sigma^2 X^T y)$$

and variance

$$V^* = (V^{-1} + \sigma^2 X^T X)^{-1}$$

If we re-scale the prior coefficient β and set

$$V = \sigma^2 \Sigma$$

then the posterior of the coefficient β is still normal with a posterior distribution of the form

$$f(\beta \mid y) \sim \mathcal{N}(m^*, V^*)$$

where

$$\begin{aligned} m^* &= (\Sigma^{-1} + X^T X)^{-1}(\Sigma^{-1} m + X^T y) \\ V^* &= \sigma^2 (\Sigma^{-1} + X^T X)^{-1}. \end{aligned}$$

The predictive distribution with X_0 design matrix in that case would be a Normal distribution with mean $X_0 m^*$ and variance $\sigma^2(I + X_0(V^*)X_0^T)$.

3.2 The Case of Unknown variance

In the case of the unknown variance σ^2, the normal inverse Gamma could be chosen to be the joint prior for (β, σ^2)

$$f(\beta, \sigma^2) = \frac{(\frac{a}{2})^{\frac{d}{2}}}{(2\pi)^{\frac{p}{2}} \mid V \mid^{\frac{1}{2}} \Gamma(\frac{d}{2})} (\sigma^2)^{-\frac{d+p+2}{2}} \exp(-\frac{(\beta - m)^T V^{-1}(\beta - m) + a}{2\sigma^2}).$$

In this case the joint posterior distribution is given by

$$f(\beta, \sigma^2 | y) = \frac{(\frac{a*}{2})^{\frac{d+n}{2}}}{(2\pi)^{\frac{n}{2}} |V^*|^{\frac{1}{2}} \Gamma \frac{d+n}{2}} (\sigma^2)^{-\frac{n+d+p+2}{2}} \exp(-\frac{(y - m^*)^T (V^*)^{-1}(y - m^*)}{2\sigma^2} + a^*) \quad (7)$$

where

$$\begin{aligned} m^* &= (V^{-1} + X^T X)^{-1}(V^{-1}m + X^T y) \\ V^* &= (V^{-1} + X^T X)^{-1} \\ a^* &= a + m^T V^{-1} m + y^T y - (m^*)^T (V^*)^{-1}(m^*) \end{aligned}$$

and the predictive distribution of $y_{0_{r\times 1}}$ is

$$f(y_0|y) = \frac{(a^*)^{(\frac{n+d}{2})}\Gamma\frac{n+d+r}{2}}{|I_r + X_0 V^* X_0^T|^{\frac{1}{2}}(\pi)^{\frac{r}{2}}\Gamma\frac{n+d}{2}}(a^* + (y_0 - X_0 m^*)^T(I_r + X_0 V^* X_0^T)^{-1}(y_0 - X_0 m^*))^{-\frac{n+d+r}{2}} \quad (8)$$

which is a multivariate t distribution.

3.3 Approximations for posterior distributions

It becomes clear when the distributions are different from normal, computational difficulties arise to get a closed form to the posterior distributions. Thus, approximations to the posterior distributions have been investigated through the literature of the Bayesian inference (Gelman et al., 2004). This subsection goes through some of these approximations.

- Finding a crude or rough parameter estimation for the hyperparameters which is helpful in the case of hierarchical models.

- Finding the posterior mode by maximization methods like Newton's method, EM algorithm and its extensions. However, the simplest method to find the mode is to use the conditional maximization by changing the estimated value of one parameter till achieving the maximum and fixing the values of the other parameters at the rough estimation and repeating the procedure for all parameters.

- Finding a normal mixture or related approximation to the joint posterior distribution or to a particular marginal or conditional posterior distributions based on the curvature

of the distributions around its mode using standard distributions.

- Separate approximations to the marginal and conditional posterior densities in high dimensional problems such as hierarchical models in which the normal and t distributions do not fit.

4 Utility Functions

Let $\mathcal{X}$ be the candidate set of N points, $\mathcal{Y}$ be the sample space, y_s be the vector of observations corresponding to the selected design ξ of size n.

As mentioned before finding a Bayesian optimal design is quite similar to finding a Bayesian decision. The aim now is to find the design ξ that maximizes the expected utility of the best decision, i.e.,

$$U(\xi^*) = \max_{\xi} \quad \int_{\mathcal{Y}} \max_{d\in\mathcal{Y}} \int_{\Theta} U(d, \theta, \xi, y_s) p(\theta, y_s, \xi) d\theta dy_s \tag{9}$$

where $U(d, \theta, \xi, y_s)$ is the utility function chosen to satisfy certain goal, y_s is the observed data from a sample space $\mathcal{Y}$, d is the decision which consists of two parts: first selection of the design ξ and then the choice of a terminal decision d. By rearranging the integrand,

$$U(\xi^*) = \max_{\xi} \quad \int_{\mathcal{Y}} \max_{d\in\mathcal{Y}} [\int_{\Theta} U(d, \theta, \xi, y_s) p(\theta|y_s, \xi) d\theta] p(y_s|\xi) dy_s, \tag{10}$$

we are able to find the decision that maximizes the expected posterior utility (minimizing the posterior risk) and then find the design that maximizes the expected pre-posterior utility of the best decision.

Different utility functions have been used in the context of the optimal design. The optimality criteria in the Gaussian case can accomplish many goals relevant to the whole set of parameters, subset of the parameters, prediction or others like optimization or sensitivity. In this section we review some of the common utility functions used in the normal linear

model with known variance, for more details see (Chalenor and Verdinelli, 1995). These utility functions will be applied in the frame of the sequential design for the Gaussian process.

***A*-Optimality:** If the he aim of the experiment is to estimate more than one linear function of the parameters then the criterion can be obtained by maximizing the following expected utility

$$U(\xi) = -\int_{\mathcal{Y}}\int_{\Theta}(\theta-\hat{\theta})^T A(\theta-\hat{\theta})p(y_s,\theta|\xi)d\theta dy_s \tag{11}$$

where A is a symmetric nonnegative definite matrix. So by integrating over θ the corresponding criterion is $\phi_2(\xi) = -\text{trace}A(nM(\xi)+\Sigma)^{-1}$, where $nM(\xi)$ refers to the information matrix. This criterion is called Bayes A-optimality while the corresponding non-Bayes A-optimality is $-\text{trace}AnM(\xi)^{-1}$. If rank A is 1 which means that $A = cc^T$ we have the C-optimality. If this linear combination is normalized we call it Bayes E-optimality, which minimizes the maximum posterior variance of all possible combinations of the parameters.

***D*-Optimality:** It is used when the main concern is to minimize the covariance of the estimators of $\hat{\theta}$. The Bayesain D-optimality is obtained when the gain in Shannon information is used as the utility function so that we choose the design that maximize the expected gain of Shannon information

$$U(\xi) = \int_{\mathcal{Y}}\int_{\Theta}\log p(\theta|y_s,\xi)p(y_s,\theta|n)d\theta dy_s. \tag{12}$$

This function takes the following form in the normal linear model

$$U(\xi) = -\frac{k}{2}\log(2\pi) - \frac{k}{2} + \frac{1}{2}\log\det\sigma^2(nM(\xi)+\Sigma) \tag{13}$$

and this reduces to maximize $\det(nM(\xi)+\Sigma)$ which is Bayes D-optimality, while non-Bayes D-optimality is just $\det nM(\xi)$. D optimality can be obtained using other utility

functions that aim to different goals other than inference about θ, like prediction or discriminating between two models.

D_s-Optimality: This criterion can be considered as a special case from D-optimality, where the inference is about a subset of s parameters. In order to apply this criterion we need to partition the information matrix according to the set of parameters of main interest and get the conditional information matrix, for more details see (Karlin and Studden, 1966). One idea is apply this criterion to $g(x)^T\phi$ in equation(6).

G-Optimality: It is chosen to minimize the maximum prediction variance, it takes the following form in the case of known variance

$$\min_{\xi} \sup_{x \in \mathcal{X}} x^T (nM(\xi) + \Sigma)^{-1} x. \tag{14}$$

It has been showed in the literature that G and D optimality are equivalent under certain conditions since the general equivalence theorem (Kiefer and Wolfowitz, 1959).

Entropy: This criterion is used when the aim is to minimize the negative information about the unsampled points $y_{\bar{s}}$. Shewry and Wynn (1987) show that minimizing the entropy of the unsampled points is equivalent to maximizing the entropy of the sampled points y_s. Entropy takes the form below

$$\mathrm{Ent}(Y_s) = \mathrm{E}(-\log \pi(y_s|\xi)) = \int -\log(\pi(y_s|\xi))\pi(y_s|\xi) dy_s \tag{15}$$

Others: Minimizing the average of the posterior predictive variance

$$\min_{\xi}(\Sigma_{x_0} var(Y_{x_0}|Y_s)). \tag{16}$$

This criterion selects the design points that minimize the predictive variance at pre-specified set of points $\{x_0\}$. This set could be the whole candidate set or as with

entropy the unsampled points only.

5 Numerical Techniques

During the implementation of the ASCM, these criteria are updated at each stage. This updating could be easy updating, so there is no need for numerical techniques, as in the case of normal distribution. Also, some updating might require solving some intractable integrals. Thus, the need to numerical techniques is essential to perform the sequential experiments. Numerical techniques in use would be the quadrature rule which is a numerical approximation of an integral using a weighted sum of some values of the integrand. Gaussian quadrature rule is anticipated to be used because of its careful choice of quadrature points and the use of weight functions that improve the accuracy of the integral. In addition, the adaptive Gaussian quadrature is a more accurate choice as it is adaptively refining the intervals of the quadrature rule. Monte carlo Markov Chain techniques would be a possible alternative to evaluate the integrals of interest.

In summary, the following methods will be considered,

- MCMC techniques.
- Quadrature rules.
- Adaptive quadrature rules in addition to low-discrepancy sequences.

6 Optimization Algorithms

The main point in finding the optimal design is to select the design that optimizes the criterion of our interest among all other possible designs. This leads to use an algorithm searching most of the designs to pick the best. Such an algorithm can be found in the literature of combinatorial optimization. Two algorithms are suggested to find these designs, first of them is the Branch and Bound algorithm (Ko et al., 1995) which gives the global

optimal design but at the same time it requires extra work finding upper or lower bounds functions that help in accelerating the algorithm. The other one is the exchange algorithm which is commonly used in computer experiments. This algorithm allows to exchange the points in the current design by other ones in the candidate list if it improves the performance of the design. However, it does not guarantee a global optimum design and it might be trapped by local ones. In summary, the following methods are being considered

- Branch and Bound algorithm.
- Exchange algorithms.
- Global optimization.

7 Illustrative Example: ASCM Entropy Design

Let Y_{s_i} be the observations corresponding to the selected design ξ_i of size n_i at stage i, $i = 1, ..., k$, where k is the number of required stages, so $\Sigma_i^k n_i = n$.

The sequential scheme is used here to obtain the maximum entropy design. Which means, at each stage the conditional distribution of the response Y_{s_i} given the observations and designs chosen from the previous stages is of our interest instead of the marginal distribution of the observations. This can be framed in the following steps;

Stage 1

$$E(-\log \pi(y_{s_1}|\xi_1)) = \int -\log(\pi(y_{s_1}|\xi_1))\pi(y_{s_1}|\xi_1)dy_{s_1}$$

Stage 2

$$E(-\log \pi(y_{s_2}|y_{s_1}, \xi_1, \xi_2)) = \int -\log(\pi(y_{s_2}|y_{s_1}, \xi_1, \xi_2))\pi(y_{s_2}|y_{s_1}, \xi_1, \xi_2)dy_{s_2}$$

where $\pi(y_{s_2}|y_{s_1},\xi_1,\xi_2)$ is the predictive posterior distribution given by

$$\begin{aligned}\pi(y_{s_2}|y_{s_1},\xi_1,\xi_2) &= \int \pi(y_{s_2}|y_{s_1},\theta,\xi_1,\xi_2)\pi(\theta|y_{s_1},\xi_1,\xi_2)d\theta \\ &= \int \pi(y_{s_2}|\theta,\xi_1,\xi_2)\pi(\theta|y_{s_1})d\theta \end{aligned} \tag{17}$$

Stage 3 To compute the entropy now

$$E(-\log(\pi(y_{s_3}|y_{s_2},y_{s_1},\xi_1,\xi_2))) = \int \log(\pi(y_{s_3}|y_{s_2},y_{s_1},\xi_1,\xi_2))\pi(y_{s_3}|y_{s_2},y_1,\xi_1,\xi_2)dy$$

where

$$\begin{aligned}\pi(y_{s_3}|y_{s_2},y_{s_1},\xi_1,\xi_2,\xi_3) &= \int \pi(y_{s_3}|y_{s_2},y_{s_1},\theta,\xi_1,\xi_2,\xi_3)\pi(\theta|y_{s_1},y_{s_2},\xi_1,\xi_2,\xi_3)d\theta \\ &= \int \pi(y_{s_3}|\theta,\xi_1,\xi_2,\xi_3)\pi(\theta|y_{s_1},y_{s_2},\xi_1,\xi_2,\xi_3)d\theta \end{aligned} \tag{18}$$

Stage k

$$\begin{aligned}\pi(y_{s_k}|y_{s_{k-1}},...,y_{s_1},\xi_1,...,\xi_k) &= \int \pi(y_{s_k}|y_{s_{k-1}},...,y_{s_1},\theta,\xi_1,...,\xi_k)\pi(\theta|y_{s_1},...,y_{s_{k-1}},\xi_1,...,\xi_k)d\theta \\ &= \int \pi(y_{s_k}|\theta,\xi_1,...,\xi_k)\pi(\theta|y_{s_1},...,y_{s_{n-1}},\xi_1,...,\xi_k)d\theta \end{aligned} \tag{19}$$

7.1 Maximum Entropy Design (in the case of known variance)

Since the entropy at each stage depends on the predictive distribution of the response y_{s_i}, which is a normal distribution in the current case. This allows us to have the privilege of the normal distribution that is

$$Ent(Y_s) = \log(\det(Var(Y_s)))$$

except for some constants. So, we just need to maximize $\log|\sigma^2(I_r + X_0(V^*)X^T)|$ as in Section 3.1 where V^* is changed by conducting the former design. Also, we can notice that the argument we want to maximize does not depend on y_s, the observations. Technically, what we need to perform at stage i is to find the conditional covariance matrix corresponding to the design space given the former designs at each stage which is given by

$$Cov(Y_{\bar{s}}, Y_{\bar{s}}|Y_{s_{i-1}}) = \Sigma_{\bar{s}\bar{s}} - \Sigma_{\bar{s}s_{i-1}}\Sigma^{-1}_{s_{i-1}s_{i-1}}\Sigma_{s_{i-1}\bar{s}} \tag{20}$$

where Σ is the variance-covariance matrix of $\mathcal{Y}$, s are the sampled points and $\bar{s}$ are the unsampled ones. Then optimization algorithms are used to find the design that best maximizes the determinant of a sub matrix from this conditional covariance one. The optimal design is obtained by updating the conditional covariance matrix till we get the n points design.

7.2 Maximum Entropy Design (in the case of unknown variance)

According to the Bayesian inference in the case of unknown variance, the predictive distribution would be a Student t distribution. Although, it is still presenting integration problems, the use of multivariate t is a considerable advantage of the use of Bayesian linear models approach of Section 3.

8 Conclusions

At the end we summarize the basic features of ASCM

- The use of Bayesian linear theory.
- Reduction of the complex covariance function to the Bayesian linear model case.
- Development of utility functions
    1. For parameters.

2. For prediction.

3. Others.

- Use of the Bayesian linear sequential methods for sequential design (updating).

- Use of a variety of numerical methods for

1. Integration.

2. Optimization.

References

Chalenor, K. & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*, (second ed.). Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL.

Karlin, S. & Studden, W. J. (1966). *Tchebycheff systems: With applications in analysis and statistics*. Pure and Applied Mathematics, Vol. XV. Interscience Publishers John Wiley & Sons, New York-London-Sydney.

Kiefer, J. & Wolfowitz, J. (1959). Optimum designs in regression problems. *Ann. Math. Statist.*, 30:271–294.

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

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B*, 34:1–41.

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

Wang, L. (2008). *Karhunen Loeve expansions and their applicatins*. PhD thesis, London School of Econimcs.