EXPERIMENT DESIGN AND OPTIMIZATION IN COMPLEX SYSTEMS

Payman Sadegh

LYNGBY 1996
IMM-PHD-1996-23

IMM

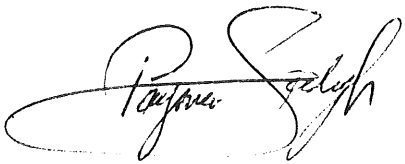
ISSN 0909-3192

Trykt af IMM - DTU
Bogbinder Hans Meyer

Preface

This work has been prepared at the Institute of Mathematical Modeling (IMM), the Technical University of Denmark (DTU), as one of the requirements for a Ph.D. degree in mathematics. Parts of this work were completed during my stay at the Applied Physics Laboratory (APL) of the Johns Hopkins University (JHU), in the period of April to December 1995.

The thesis concerns experiment design and optimization in complex systems. The attention is directed towards experiment design under partial or limited prior knowledge, and optimization in systems where no direct gradient information is available. The thesis contains independent research articles which present contributions to the field of interest.



Acknowledgements

I would like to express my gratitude to all the people who made it possible for me to complete this work.

I would like to thank my advisors, Prof. Henrik Madsen, DTU, and Prof. Jan Holst, Lund Institute of Technology. I am indebted to them not only professionally but also for their encouragement and support in all the aspects of this research.

I am grateful to Dr. James Spall who has been my advisor during my stay at JHU/APL. I would like to thank Dr. Spall for being both a great teacher, and a great friend of mine.

I would like to thank the administrative staff and the associates of IMM, especially the people of Time-Series Group: Henrik, Jan, Judith, Lars, and Torben for providing a pleasant research environment.

I would like to thank all the nice people of APL, and Dept. of Material Science at JHU, especially Mr. Mark Asher, Mr. Daniel Chin, and Dr. John Maryak, for their help during my stay at JHU/APL.

I am grateful to the Danish Research Academy (Forskerakademiet) for their financial support during my stay at JHU/APL.

Last but not least, I would like to thank my family and friends whose support and patience made the present thesis possible.

Summary

A basic problem arising often during the process of experimental modeling of a given system is the design of experiments. This stage of the modeling is of major interest in that the modeling results rely heavily upon the way the experiments are conducted. The present thesis is primarily concerned with maximization of the amount of information provided by experiments. The scope of the thesis spans areas such as system identification, statistical modeling, and system optimization. The thesis covers the theory of optimal experiments for regression models, containing input design for system identification as a special case. The thesis also studies design problems where no direct gradient of the objective function is available. The results are presented in two parts A and B.

Part A mainly contains contributions to input design for system identification. The developed techniques, however, can be generically applied to a large variety of regression models. Design under partial or very limited prior knowledge is a key word throughout this part. We use Bayesian techniques to design inputs for gray-box models, which are characterized by their physical significance and partial prior knowledge. These techniques enable us to design experiments that are robust to prior model uncertainty. We also study simple numerical procedures for the non-smooth problem of designing inputs that maximize the smallest eigenvalue of the information matrix in linear dynamic systems. This has applications in both parameter estimation problems, and in change detection problems where the direction of the parameter change is unknown.

In Part B, we study the algorithm simultaneous perturbation stochastic approximation (SPSA). The algorithm has recently attracted considerable attention in optimization problems where no

direct gradient information is available. The approach is based on a simultaneous perturbation approximation to the gradient based on (noisy) objective function measurements. SPSA is based on picking a simultaneous perturbation (random) vector in a Monte-Carlo fashion as part of generating the approximation to the gradient. The thesis presents two fundamental contributions to the generic optimization technique, and employs the algorithm to develop a unique approach for the experimental design problem of sensor configuration in complex systems. The contributions to the generic optimization include derivation of the optimal distribution for the Monte-Carlo process (optimal perturbation distribution), and a projection SPSA algorithm for constrained optimization. For the sensor configuration problem, we derive an appropriate optimality criterion as well as an efficient optimization technique based on the SPSA algorithm.

The contributions of the thesis are presented through independent research articles, each examining one aspect of interest. Numerical examples as well as real experimental results are included to illustrate theory and methods, and to introduce important application areas.

Sammenfatning

Et grundlæggende problem, som ofte opstår i forbindelse med eksperimentel modellering af et givet system er forsøgsplanlægning. Denne fase er af stor betydning for den endelige model, idet modelleringens resultater er stærkt afhængige af den måde, hvorpå forsøgene bliver gennemført. Denne afhandling beskæftiger sig primært med maksimering af informationsmængden ved udførelsen af et eksperiment. Afhandlingens omfang dækker fagområder som system identifikation, statistisk modellering, og optimering. Afhandlingen afdækker teorien for optimale forsøg for lineære og ikke-lineære regressionsmodeller. Dette omfatter input design for identifikation af dynamiske systemer som et specialtilfælde. Afhandlingen undersøger også design problemer, hvor gradienten af målfunktionen er ukendt. Afhandlingen er opdelt i to dele A og B.

Del A indeholder hovedsageligt bidrag til input design for system identifikation. De udviklede metoder kan dog generaliseres til en stor klasse af regressionsmodeller. Design under partiel eller meget begrænset apriori viden er et nøgleord for denne del af afhandlingen. Metodemæssigt anvendes bayesianske principper til input design for gray-box modeller, som er karakteriseret ved fysiske modelstrukturer og partiel apriori viden. De udviklede metoder tillader os at dimensionere forsøg, som er robuste overfor modellens apriori usikkerhed. Vi studerer også simple numeriske løsninger til det ikke-differentiable problem at maksimere den mindste egenværdi af informationsmatricen for lineære dynamiske systemer. Denne finder anvendelse ved både estimations-og detekteringsproblemer, hvor retningen af parameterændringen er ukendt.

I del B undersøger vi algoritmen "simultaneous perturbation stochastic approximation" (SPSA). Algoritmen har vundet stor op-

mærksomhed i forbindelse med optimeringsproblemer, hvor gradienten af målfunktionen er ukendt. Fremgangsmåden bygger på en samtidig perturbation (SP) gradient approksimation baseret på støjfyldte observationer af målfunktionen. SPSA er baseret på udvælgelse af en SP stokastisk vektor ved hjælp af Monte-Carlo teknikker. Afhandlingen præsenterer to fundamentale bidrag til SPSA-baseret optimering, og anvender algoritmen til at udvikle en enestående fremgangsmåde til optimal sensorkonfiguration i komplekse systemer. Afhandlingens bidrag til selve optimeringsmetoden består i at udlede den optimale fordeling for Monte-Carlo processen (optimal perturbation distribution) og at præsentere en projiceringsalgorime baseret på SPSA til optimering under bibetingelser. I forbindelse med sensorkonfigurationsproblemet, udleder vi et passende kriterium samt en SPSA-baseret optimeringsalgoritme.

Afhandlingens bidrag er præsenteret i uafhængige forskningsartikler. I alle artikler undersøges et aspekt af forskningsarbejdet. Numeriske eksempler samt eksperimentelle resultater er inkluderet for at illustrere teorien og metoderne. Samtidig introduceres vigtige anvendelsesområder.

Contents

Introduction

Overview

The remainder of the thesis contains two parts (A and B), and concluding remarks. Part A contains 3 and Part B contains 4 articles which study different aspects of optimal experiment design, and optimization in systems where no direct gradient of the objective function is available respectively. The concluding remarks present a summary of the contributions of the thesis, together with directions for future research.

1 Introduction to Part A

1.1 Background

Part A of the thesis concerns design of optimal experiments for regression models, including input design for system identification. Experiment design is perhaps one of the most fundamental issues one has to deal with during the process of experimental modeling of a system. The objective of the design is maximization of the information in data in some sense.

The present thesis has a statistical approach to the optimal design problem, in the sense that data are regarded as realizations of some random process. The experimental conditions affect the distribution of data, and the question is an optimal selection of the experimental conditions for a given design objective. The statistical theory of optimal experiments has been extensively studied and is still the focus of active research, see Kiefer (1959), Fedorov (1972), Kiefer (1974), Silvey (1980), Pukelsheim (1993), among many other references.

The approach in Part A is based on Fisher's information matrix which is similarly pursued in Goodwin and Payne (1977), Mehra (1974), and Zarrop (1979). The approach is suitable if the experimenter is interested in minimizing the covariance of the parameter estimate in some sense. This is due to the fact that the information matrix is related to the asymptotic covariance of the parameter estimate through the Cramér-Rao inequality. Minimizing the covariance of the parameter estimate is usually of interest in models that contain physically significant parameters. It should however be noted that this approach may not be very suitable for modeling problems where the model structure is of reduced order and the parameters

have no physical significance. Consider for instance an experimental modeling problem where the (reduced order) model is to be used for controller design purposes. Under such circumstances, a more suitable measure for the quality of the obtained model may be related to the error of the estimated transfer function resulting from modeling bias (due to the reduced model order), or the modeling bias together with the variance of the estimates, see e.g. Yuang and Ljung (1985), Wahlberg (1987), Gevers and Ljung (1986). In connection with change detection problems, on the other hand, experimental design approaches that are based on Kullback-Leibler information for optimal change detection in linear dynamic systems lead to problem formulations dual to those treated in this thesis, see e.g. Zhang (1989), and Hatanaka and Uosaki (1994). Analogous duality results for discrimination between regression models are treated in e.g. Fedorov and Khabarov (1986).

A point of primary concern in the thesis is the design of experiments under partial or sparse prior knowledge. The interest in such investigations arises from an inherent dichotomy that pertains to the calculations for optimal experiments. On the one hand, the experiments are often performed with the goal of identifying unknown system characteristics of interest. On the other hand, model based optimization techniques require *a priori* specifications of the system to be optimized. The amount of available prior knowledge and the way it is accounted for in the design, affect the relevant problem formulation and the optimization techniques to be applied.

1.2 Outline of Part A

Part A contains three articles (A1, A2, and A3). The articles A2 and A3 concern optimal input design for system identification while

A1 contains a theoretical study of the optimum design for regression models. The contents of these articles are summarized as follows.

In A1, we consider the problem of optimal experiment design for regression models. The objective of this article is two-fold: (1) to provide the necessary theoretical background for the rest of Part A, and more importantly (2) to illustrate that the developed techniques for input design (see A2 and A3) can be generically applied to regression design problems. The scope of A1 covers both linear and nonlinear regression models.

In A2, we study input design for identification of gray-box models. Gray-box models are characterized by their physical significance and partial prior knowledge. Both of these aspects are studied in connection with the optimal design of inputs. We adopt a Bayesian approach to deal with the partial prior knowledge that is expressed by probability distributions, and study the extension of the classical theory to the optimization of Bayesian criteria.

Finally in A3, we study maxmin design of input signals. The goal here is to maximize the smallest eigenvalue of the information matrix with respect to the designed input. This has applications in both parameter estimation problems, and change detection problems where the direction of the parameter change is unknown. The maxmin design criterion is in general non-smooth. We study numerical procedures for the optimization.

2 Introduction to Part B

2.1 Background

There are many experimental design problems that cannot be treated under any of the categories discussed in Part A. The research

in this part of the thesis was motivated by such a problem, concerning optimal sensor location for signal detection in complex structures. In Part B, we study a very powerful technique for optimization in complex systems, and use the technique to develop a unique approach to treat a wide range of sensor configuration problems.

Consider an optimization problem where no direct gradient of the objective function is available. The algorithm simultaneous perturbation stochastic approximation (SPSA) has recently attracted considerable attention in optimization problems of this type in areas such as experiment design, adaptive control, pattern recognition, discrete event systems, neural network training, and model parameter estimation, see e.g. Sadegh and Spall (1996) (appearing also as article B4 in the present thesis) and the relevant references therein. The SPSA algorithm is a variant of stochastic approximation in the Kiefer-Wolfowitz setting where optimization merely relies on noisy evaluations of the objective function. The essential feature of SPSA is its highly efficient gradient approximation that requires only two function evaluations per iteration regardless of the number of parameters being optimized, in contrast to the $2p$ evaluations required in classical Kiefer-Wolfowitz finite difference based approaches. Under reasonably general conditions, it was shown in Spall (1992) that the p-fold savings in function evaluations per gradient approximation can translate directly into a p-fold savings in total number of evaluations needed to achieve a given level of accuracy in the optimization process.

Part B includes both theoretical contributions to the generic optimization technique using SPSA, and a solution to the problem of optimal sensor configuration in complex systems where SPSA plays a key role.

2.2 Outline of Part B

Part B contains four articles B1, B2, B3, and B4. In B1, a brief overview of stochastic approximation is presented. The articles B2 and B3 solve two fundamental issues related to the optimization technique using SPSA, and B4 concerns optimal sensor configuration for complex systems.

The approximation to the gradient at each iteration of SPSA is computed by simultaneous perturbation of the parameter, using a random vector generated in a Monte-Carlo fashion. Since the user has full control over the probability distribution of the Monte-Carlo process, there is strong reason to select the perturbation distribution wisely in order to minimize the overall cost of optimization. In B2, we derive the optimal distribution for random perturbations. Section 2 of the article also contains an overview of the SPSA algorithm and its assumptions.

The original SPSA algorithm as given in Spall (1992) is an unconstrained algorithm. Constraints, on the other hand, are inseperable parts of almost all real world applications of optimization techniques. In B3, a constrained version of the SPSA algorithm using projections is presented.

Finally in B4, we are concerned with the experimental design problem of optimal sensor configuration in complex systems. The contribution of the paper is two-fold: (1) definition of a suitable performance measure for sensor configuration, and (2) description of an efficient and practical algorithm for achieving the optimality objective. The SPSA algorithm plays a central role in the optimization part.

References

Fedorov, V. and V. Khabarov (1986). Duality of optimal designs for model discrimination and parameter estimation. *Biometrika 73*, 183–190.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Academic, New York.

Gevers, M. and L. Ljung (1986). Optimal experiment design with respect to the intended model application. *Automatica 22*(5), 543–554.

Goodwin, G. C. and R. L. Payne (1977). *Dynamic System Identification: Experiment Design and Data Analysis.* Academic, New York.

Hatanaka, T. and K. Uosaki (1994). Optimal auxiliary input for fault detection, frequency domain approach. In *10th IFAC Symposium on System Identification, Copenhage, Denmark*, Volume 3, pp. 107–112.

Kiefer, J. (1959). Optimum experimental designs. *J. R. Statist. Soc. B. 21*, 272–319.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Stat. 2*(5), 849–879.

Mehra, R. K. (1974). Optimal input signals for parameter estimation in dynamic systems - survey and new results. *IEEE Transactions on Automatic Control AC-19*(6), 753–768.

Pukelsheim, F. (1993). *Optimal Design of Experiments.* John Wiley & Sons, New York.

Sadegh, P. and J. C. Spall (1996). Optimal sensor configuration for complex systems. In *Proc. of Test Technology Symposium.* U.S. Army Test and Evaluation Command. In press.

Silvey, S. D. (1980). *Optimal Design.* Chapman & Hall, London.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control 37*(3), 332–341.

Wahlberg, B. (1987). *On the Identification and Approximation of Linear Systems.* Ph. D. thesis, Department of Electrical Engineering, Linköping, Sweden.

Yuang, Z. D. and L. Ljung (1985). Unprejudiced optimal open loop input design for identification of transfer functions. *Automatica 21*(6), 708–967.

Zarrop, M. B. (1979). *Optimal Experiment Design for Dynamic System Identification*, Volume 21 of *Lecture Notes in Control and Information Sciences.* Springer-Verlag, Berlin.

Zhang, X. J. (1989). *Auxiliary Signal Design in Fault Detection and Diagnosis*, Volume 134 of *Lecture Notes in Control and Information Sciences.* Springer-Verlag, Berlin.

Theory of Optimal Experiments

Payman Sadegh

Institute of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

[A1]

Abstract

We study the theory of optimal experiments for linear and nonlinear regression models. In connection with nonlinear regression models, attention is restricted to the issue of robustness of experiments since the design criteria for nonlinear regression cases in general are dependent upon the unknown parameters. We adopt a Bayesian approach to deal with the robustness issue.

Key words: Experimental design, Linear theory, Nonlinear regression, Robust experiments, Bayesian criteria.

1 Introduction

1.1 Background and Organization of the Article

In many applications of science and technology, it is necessary to draw inference on a system based on experimental data. Examples of such a situation are estimation of a set of unknown parameters, detection of changes, hypothesis testing, etc. One of the basic issues in this connection is the design of optimal experimental conditions so as to maximize the amount of information obtained by the experiments. The approach of the article is statistical, i.e. data are considered as realizations of some random process whose distribution is affected by the experimental conditions.

Experiment design for parameter estimation in linear regression models provides the basis for the theory of statistical experiment design. Section 2 reviews the fundamental ingredients of the linear theory. This includes the general problem setting of experiment design for linear regression models, some widely used design criteria, the approximate theory of experiment design, main results, and simple algorithmic procedures for the design of optimal experiments. Section 3 is concerned with the design of experiments for parameter estimation in nonlinear regression models. The attention in Section 3 is directed towards the issue of robustness of experiments since, as we shall see, the design criteria in nonlinear cases turn out to be dependent upon the unknown parameters. We adopt a Bayesian approach to deal with the robustness issue. Finally, Section 4 offers concluding remarks.

1.2 General Problem Formulation

Assume that the distribution of data y depends upon:

- a control variable $u \in \mathcal{U} \subset \mathbb{R}^r$ which can be selected by the experimenter,
- a parameter $\theta \in \Theta \subset \mathbb{R}^p$ which is fixed and unknown to the experimenter, and whose value is of interest,
- a nuisance parameter $v \in \Upsilon \subset \mathbb{R}^l$ which is also fixed and unknown but it is not of primary interest.

The experimenter is allowed to take N independent observations on y at $u_{(1)}, ..., u_{(N)}$ chosen from $\mathcal{U}$ which is called an N-observation design. The basic problem is the selection of the N observations. The solution obviously requires an optimization approach. The approach discussed here is based on Fisher's information matrix. In Goodwin and Payne (1977), and in Sadegh, Holst, Madsen, and Melgaard (1995) (see article A2 for a revised version), it is illustrated how a decision theory approach is approximately equivalent to the information matrix approach.

Let ϑ in the following denote $(\theta^\top, v^\top)^\top$, U denote the set $\{u_{(i)}\}$, and $g(y|U, \vartheta)$ denote the probability density function of data y given U and ϑ.

Definition 1 Fisher's information matrix for the random variable y is defined by

$$M_F(U;\vartheta) = E\{(\frac{\partial g(y|U,\vartheta)}{\partial \vartheta})^\top (\frac{\partial g(y|U,\vartheta)}{\partial \vartheta})\} \tag{1.1}$$

where $E\{\cdot\}$ denotes mean value.

The importance of the information matrix is highlighted by the Cramér-Rao inequality which states that under certain regularity conditions (Rao (1973)), the lower bound on the asymptotic covariance

of any unbiased estimator of θ is given by the leading $p \times p$ matrix of $M_F^{-1}(U;\vartheta)$ which is equal to

$$V(U;\vartheta) = \{M_{\theta\theta}(U) - M_{\theta\upsilon}(U)M_{\upsilon\upsilon}^{-1}(U)M_{\upsilon\theta}(U)\}^{-1},$$

where

$$M_F(U;\vartheta) = \begin{bmatrix} M_{\theta\theta}(U) & M_{\theta\upsilon}(U) \\ M_{\upsilon\theta}(U) & M_{\upsilon\upsilon}(U) \end{bmatrix}.$$

In connection with estimating θ, we shall be interested in minimizing the covariance matrix $V(U;\vartheta)$ in some sense.

Now assume that it is of interest to estimate the real, vector valued, differentiable function $h(\theta) = (h_1(\theta), ..., h_s(\theta))^\top$. Let $Dh(\theta)$ denote the $p \times s$ matrix whose (i,j)th element is given by $\partial h_j(\theta)/\partial\theta_i$. Again, assuming sufficient regularity, the lower bound on the asymptotic covariance matrix of any unbiased estimator of $h(\theta)$ is

$$V_h(U;\vartheta) = \{Dh(\theta)\}^\top V(U;\vartheta)\{Dh(\theta)\}.$$

In connection with estimating the function $h(\theta)$, we shall be interested in minimizing the covariance matrix $V_h(U;\vartheta)$ in some sense.

It is now evident that because of the dependency of $V(U;\vartheta)$ and $V_h(U;\vartheta)$ upon ϑ, the design criterion is in general a function of the unknown parameter. We postpone a treatment of such problems to Section 3. First, we investigate the design problem for linear regression cases where the optimality criteria of interest are independent of the unknown parameter.

2 Linear Theory

2.1 Linear Theory Problem Setting

The linear theory of experiment design studies design problems where

- $E(y|u,\vartheta) = \theta_1 f_1(u) + ... + \theta_p f_p(u)$ where $\theta = (\theta_1, ..., \theta_p)^\top$, and $f_1, ..., f_p$ are known functions,
- $\mathrm{var}(y|u,\vartheta) = v$,
- the function $h(\theta)$ is linear in θ.

As we shall see in the following, the selected design criteria for the linear case do not depend upon ϑ. This eliminates a major theoretical and practical difficulty that pertains to the design problems in nonlinear cases.

Denoting $x = (f_1(u), ..., f_p(u))^\top$, the N-observation design problem can be formulated as the selection of vectors

$$x_{(i)} = (f_1(u_{(i)}), ..., f_p(u_{(i)}))^\top \in \mathcal{X} \subset \mathbb{R}^p,$$

(i.e. the $x_{(i)}$ are the design variables) where the set $\mathcal{X}$ is called the induced design space (the space induced by $(f_1, ..., f_p)$). The model is then given by

$$E(y|x,\vartheta) = x^\top \theta.$$

Simple calculations show that for Gaussian data, the covariance of the maximum likelihood estimator of θ is given by

$$V(x_{(1)}, x_{(2)}, ...; v) = v(\sum x_{(i)} x_{(i)}^\top)^{-1}. \tag{2.1}$$

Notice that the covariance simply is the product of v and $(\sum x_{(i)} x_{(i)}^\top)^{-1}$ which makes the following problem formulation appropriate. Select the vectors $x_{(i)} \in \mathcal{X}$ in order to maximize $\sum x_{(i)} x_{(i)}^\top$ in some sense. Notice that $\sum x_{(i)} x_{(i)}^\top$ does not depend upon the unknown parameter.

Remark 1 In case of non-Gaussian data, the expression (2.1) determines the covariance of the least squares estimate of the parameters. Hence, the above problem formulation is also suitable for the case of non-Gaussian data, if attention is restricted to the least squares estimator.

2.2 Design Criteria

In this subsection, we will be concerned with the selection of a *scalar* function of $\sum x_{(i)}x_{(i)}^{\top}$ as the design criterion. Our interest in optimizing a scalar function of $\sum x_{(i)}x_{(i)}^{\top}$ arises from the fact that it is in general not possible to design experiments that maximize $\sum x_{(i)}x_{(i)}^{\top}$ in a strong (matrix inequality) sense, see Silvey (1980). Choosing to work with minimization problems, it is reasonable to require that the considered scalar function $\phi(\cdot)$ be a non-increasing function of its matrix valued argument (since the original intent is to maximize the matrix valued argument in some sense). Hence, we require that

$$M_1 - M_2 \quad \text{nonnegative definite} \Rightarrow \phi(M_1) \leq \phi(M_2).$$

Various ϕ functions having the above property may be of interest.

D-optimality

The D-optimality criterion is defined by

$$\phi(\sum x_{(i)}x_{(i)}^{\top}) = -\log\det(\sum x_{(i)}x_{(i)}^{\top})$$

if $\sum x_{(i)}x_{(i)}^{\top}$ is nonsingular and $+\infty$ otherwise. This is perhaps the most widely studied design criterion in the area of experimental design. The criterion is related to minimizing the volume of the confidence ellipsoid of the parameter estimate, see e.g. Silvey (1980).

D_A-optimality

Assuming that one is interested in s linear combinations of θ defined by the linear transformation $K\theta$ where K is a $s \times p$ matrix,

then it follows that the covariance of $K\theta$ for nonsingular $\sum x_{(i)}x_{(i)}^\top$ is proportional to $K\{\sum x_{(i)}x_{(i)}^\top\}^{-1}K^\top$. Then it is natural to consider $\phi(\sum x_{(i)}x_{(i)}^\top) = \log\det(K\{\sum x_{(i)}x_{(i)}^\top\}^{-1}K^\top)$ for nonsingular $\sum x_{(i)}x_{(i)}^\top$ and $+\infty$ otherwise. This is the criterion of D_A-optimality, see e.g. Sibson (1974).

D_s-optimality

D_s-optimality is a special case of D_A-optimality where the s linear combinations of interest are simply s parameters of the total p parameters.

G-optimality

It may be of interest to predict $E(y)$ over some region $\mathcal{W}$ of $\mathbb{R}^p$, i.e. the quantity $w^\top\theta$, $w \in \mathcal{W}$, may be of interest. For any fixed $w \in \mathcal{W}$ (and nonsingular $\sum x_{(i)}x_{(i)}^\top$), the covariance of $w^\top\theta$ is proportional to $w^\top\{\sum x_{(i)}x_{(i)}^\top\}^{-1}w$. It is then a natural choice to consider minimization of $\phi(\sum x_{(i)}x_{(i)}^\top) = \max_{w\in\mathcal{W}} w^\top\{\sum x_{(i)}x_{(i)}^\top\}^{-1}w$. If $\mathcal{W} = \mathcal{X}$, this criterion is called the G-optimality criterion, see e.g. Kiefer and Wolfowitz (1960).

E-optimality

This is a special case of the criterion discussed above where $\mathcal{W}$ is the unit sphere. In this case $\phi(\sum x_{(i)}x_{(i)}^\top) = \max_{w^\top w=1} w^\top\{\sum x_{(i)}x_{(i)}^\top\}^{-1}w$ which is equivalent to maximizing the smallest eigenvalue of $\sum x_{(i)}x_{(i)}^\top$.

2.3 Approximate Theory of Experiment Design

Having specified a suitable criterion (ϕ function), one faces an optimization problem for specifying the optimal design variables. The problem can be formulated as the selection of n sets of observations where the total number of observations N is fixed and given. Let n_i denote the number of observations associated with the ith set, i.e. $\sum n_i = N$, and let $x_{(i)}$ denote the corresponding design variable. We seek a probability distribution specified by the support points $x_{(i)}$ and the probability weights $p_i = n_i/N$ (note that $0 \leq p_i \leq 1$, and $\sum p_i = 1$) in order to optimize a scalar function of $\sum_{i=1}^{n} x_{(i)} x_{(i)}^{\top} p_i$. It is very difficult to deal with the posed problem since it requires solving for probability distributions where the probability weights are restricted to integer multiples of $1/N$. The situation is quite analogous to the optimization problems where the optimization parameter can only take integer values. Disregarding the discreteness of the probability weights, which can in practice be justified if the total number of observations is large, the problem simplifies to selecting any probability distribution over the induced design space $\mathcal{X}$. This idea leads to what Kiefer termed as *the approximate theory* of experiment design (see Kiefer (1959)).

Let H denote the class of probability distributions on the Borel sets of $\mathcal{X}$ and $\mathcal{X}$ be a given compact subset of the Euclidean p-space. In the approximate theory, we seek a probability measure $\xi \in H$ minimizing a scalar function of

$$M(\xi) = \int_{\mathcal{X}} x x^{\top} \xi(dx). \tag{2.2}$$

Any $\xi \in H$ will be called a design measure. Note that the compactness of $\mathcal{X}$ ensures the existence of the integral in (2.2). In the

following, we denote $\mathcal{M} = \{M(\xi)|\xi \in H\}$.

2.4 Convexity of $\mathcal{M}$

It is easy to see that each $M \in \mathcal{M}$ is a real symmetric (nonnegative definite) matrix, and each element of $\mathcal{M}$ can be represented as a point in $\mathbb{R}^{p(p+1)/2}$. Moreover, the equation (2.2) and $\int_{\mathcal{X}} \xi(dx) = 1$ imply that the set $\mathcal{M}$ is the closed convex hull of $\{xx^\top|x \in \mathcal{X}\}$. Notice that $xx^\top$ is a M-matrix that corresponds to a design where all the observations are taken at x. We have the following results.

Theorem 1 For any measure $\xi \in H$, there exists a measure $\xi' \in H$ comprising at most $p(p+1)/2+1$ supporting points, such that $M(\xi) = M(\xi')$.

PROOF: Since $\mathcal{M}$ is the convex hull of the set $\{xx^\top|x \in \mathcal{X}\}$, then by Caratheodory's theorem (Rockafellar (1970)), any $M \in \mathcal{M}$ can be written as a convex combination of at most $p(p+1)/2+1$ points of the form $xx^\top$, $x \in \mathcal{X}$. The assertion of the theorem then follows since $xx^\top$ is the M-matrix corresponding to the design with the single supporting point x.

Corollary 1 Given a strictly decreasing ϕ function (i.e., if $M_1 - M_2$ is nonnegative definite and nonzero then $\phi(M_1) < \phi(M_2)$), then for any optimal measure $\xi^* \in H$, there exists another measure $\xi' \in H$ comprising at most $p(p+1)/2$ supporting points, such that $M(\xi^*) = M(\xi')$.

PROOF: If $M(\xi^*)$ belongs to the interior of $\mathcal{M}$, then there is an $a > 1$ such that $aM(\xi^*) \in \mathcal{M}$. But it holds that $\phi(aM(\xi^*)) < \phi(M(\xi^*))$ since the ϕ function is strictly decreasing with respect to its matrix valued argument, and $(a-1)M(\xi^*)$ is nonnegative definite and nonzero. Hence, $M(\xi^*)$ must belong to the boundary, and the corollary follows similarly to Theorem 1, noting that any point on the

boundary of $\mathcal{M}$ can be written as a convex combination of at most $p(p+1)/2$ elements of the set $\{xx^\top | x \in \mathcal{X}\}$.

The theorem and its corollary imply that an optimum can be found among measures with at most $p(p+1)/2+1$ and $p(p+1)/2$ supporting points respectively.

2.5 Main Results

In this subsection, we summarize the main results that enable us to derive conditions for optimality of a given design and devise simple algorithms for construction of optimal designs. These results are mainly due to Kiefer (1974).

From now on, we assume that $\phi(M)$ is a convex function of M. Many practically interesting criteria such as the D-optimality and E-optimality criteria possess this property, see e.g. Kiefer (1974).

In the following, we consider $\mathcal{M}$ as a convex subset of the Hilbert space of real symmetric matrices of order p, where scalar product is defined by $< A, B >= \text{trace}(AB)$ for any two matrices A, B, and consequently the norm is defined by $||A|| =< A, A >^{1/2}$.

Definition 2 Let F map a Banach space to $\mathbb{R}$. The directional derivative of F at a point A in direction B of the Banach space is defined by

$$D_F(A,B) = \lim_{\gamma \to 0^+} \frac{F(A+\gamma B) - F(A)}{\gamma}.$$

Definition 3 Let F be a real functional defined over a finite dimensional Hilbert space with scalar product $< .,. >$. Then F is said to

be differentiable at A if there exists a $\nabla F(A)$ such that

$$\lim_{B \to A, B \neq A} \frac{|F(B) - F(A) - < \nabla F(A), B - A > |}{||B - A||} = 0. \qquad (2.3)$$

Remark 2 The existence of the directional derivative at a point is different from differentiability at that point. In fact for a convex function, directional derivative at a point exists if the function is Lipschitz near that point, and the Lipschitz condition holds if the (convex) function is bounded from above near the point, see Clarke (1983). The differentiability, on the other hand, is a much stronger condition. In Case 2 which follows later in this subsection, we study a nondifferentiable example where the directional derivative exists and is well defined.

It can be easily verified that if F is differentiable at a point A, then the directional derivative $D_F(A, B)$ is equal to $< \nabla F(A), B >$ (see also Appendix 3 of Silvey (1980) or Rockafellar (1970)). This implies that differentiability at a point results in linearity of the directional derivative in the second argument. The linearity in the second argument is critical to the subsequent theory.

Theorem 2 If ϕ is convex on $\mathcal{M}$ and differentiable at $M(\xi^*)$, then ξ^* is ϕ-optimal if and only if $D_\phi(M(\xi^*), xx^\top - M(\xi^*)) \geq 0$ for all $x \in \mathcal{X}$.

PROOF: See Silvey (1980).

The necessity part of the above theorem is obvious and follows directly from the definition of the directional derivative. The nontrivial fact is the sufficiency part. It is perhaps easier to see that a necessary and sufficient condition for optimality of a design ξ^* is that $D_\phi(M(\xi^*), M' - M(\xi^*)) \geq 0$ for all $M' \in \mathcal{M}$ (see Silvey (1980), Theorem 3.6). This condition simply states that if we are standing on the

bottom of a convex valley, there is no point in the valley to look downwards to. The condition $D_\phi(M(\xi^*), M' - M(\xi^*)) \geq 0$ has little practical value, since to check the optimality using this condition, the directional derivative towards all the points of $\mathcal{M}$ should be computed. However, Theorem 2 states that in case ϕ is differentiable at $M(\xi^*)$, it is sufficient to check the condition $D_\phi(M(\xi^*), M' - M(\xi^*)) \geq 0$ only for those M' that are written as $M' = xx^\top$, $x \in \mathcal{X}$ (and not for all $M' \in \mathcal{M}$). The crucial point in proving the sufficiency of the theorem is the linearity of the directional derivative in the second argument. The following two cases clarify the matter.

Case 1 *A differentiable case:* Assume that a design ξ is given such that $M(\xi)$ is nonsingular. Straight forward calculations show that for $\phi(M) = -\log\det(M)$, one has (Silvey (1980), page 21)

$$\begin{aligned} D_\phi(M(\xi), M' - M(\xi)) &= \operatorname{trace}\{M(\xi)^{-1}[M(\xi) - M']\} \\ &= p - \operatorname{trace}\{M(\xi)^{-1}M'\}. \end{aligned}$$

Hence, a necessary and sufficient condition for the optimality of the design ξ is that $\operatorname{trace}\{M(\xi)^{-1}xx^\top\} \leq p$ or $x^\top M(\xi)^{-1}x \leq p$ for all $x \in \mathcal{X}$.

It follows then that the solution to $\max_{x\in\mathcal{X}} x^\top M(\xi)^{-1}x$ resolves whether or not a given design ξ is optimum. If the maximum value is smaller than or equal to p, then the design ξ is optimum. In case $\max_{x\in\mathcal{X}} x^\top M(\xi)^{-1}x > p$, then from the definition of the directional derivative, it is easy to see that $x^\circ x^{\circ\top} - M(\xi)$ where $x^\circ = \arg\max_{x\in\mathcal{X}} x^\top M(\xi)^{-1}x$ provides a *descent direction* for $\phi(\cdot)$ at $M(\xi)$. This direction is indeed the steepest descent direction. Note that any $M' \in \mathcal{M}$ can be written as a convex combination of at most $p(p+1)/2+1$ points of the set $\{xx^\top, x \in \mathcal{X}\}$, i.e. $M' = \sum a_i x_{(i)} x_{(i)}^\top$,

$x_{(i)} \in \mathcal{X}$, $\sum a_i = 1$. Now

$$\begin{aligned} D_\phi(M(\xi), M' - M(\xi)) &= D_\phi(M(\xi), \sum a_i x_{(i)} x_{(i)}^\top - M(\xi)) \\ &= D_\phi(M(\xi), \sum a_i (x_{(i)} x_{(i)}^\top - M(\xi))) \\ &= \sum a_i D_\phi(M(\xi), x_{(i)} x_{(i)}^\top - M(\xi)) \\ &\geq D_\phi(M(\xi), x^\circ x^{\circ\top} - M(\xi)) \end{aligned}$$

which proves our claim that $x^\circ x^{\circ\top}$ gives a steepest descent direction. This idea will be used later to devise simple optimization algorithms. It is obvious that similar discussions hold for any other differentiable convex ϕ function.

Case 2 *A nondifferentiable case:* Consider now $\phi(M) = -\lambda_{\min}(M)$ where $\lambda_{\min}(\cdot)$ denotes smallest eigenvalue. Suppose that a design ξ is given and that the smallest eigenvalue of $M(\xi)$ has multiplicity r. The directional derivative is given by (see Kiefer (1974))

$$D_\phi(M(\xi), M' - M(\xi)) = -\lambda_{\min}\{Q(\xi)^\top (M' - M(\xi)) Q(\xi)\},$$

where each column of the $p \times r$ matrix $Q(\xi)$ corresponds to an orthonormal eigenvector of the smallest eigenvalue of $M(\xi)$ (note that since $M(\xi)$ is symmetric, there exist r orthogonal eigenvectors corresponding to the smallest eigenvalue of $M(\xi)$). The differentiability follows only if $r = 1$ in which case

$$\begin{aligned} D_\phi(M(\xi), M' - M(\xi)) &= -Q(\xi)^\top (M' - M(\xi)) Q(\xi) \\ &= \text{trace}\{Q(\xi) Q(\xi)^\top (M(\xi) - M')\}. \end{aligned}$$

Here the directional derivative is obviously linear in the second argument and Theorem 2 states that a necessary and sufficient condition for optimality is that $\text{trace}\{Q(\xi) Q(\xi)^\top (M(\xi) - xx^\top)\} \geq 0$ for

all $x \in \mathcal{X}$. However, for $r \geq 2$, Theorem 2 can not be applied since $\lambda_{\min}(Q(\xi)^\top Z Q(\xi))$ will no longer be linear in Z.

Analogous to Case 1, the optimization $\max_{x \in \mathcal{X}} x^\top Q(\xi) Q(\xi)^\top x$ resolves whether or not a given design ξ is optimum, and determines the steepest descent direction (for non-optimal ξ) in case that the smallest eigenvalue of $M(\xi)$ is simple (i.e. $r = 1$).

In case $r \geq 2$, it is not possible to use Theorem 2. All one can say in this case is that a necessary and sufficient condition for optimality of a design ξ is that

$$\max_{M' \in \mathcal{M}} \lambda_{\min}\{Q(\xi)^\top (M' - M(\xi)) Q(\xi)\} \leq 0$$

which requires another maxmin optimization, but probably of smaller order. The order of the matrices for this maxmin optimization is r instead of p in the original problem. Sadegh, Hansen, Madsen, and Holst (1996) (see article A3) present a solution to maximization of $\lambda_{\min}(M)$ in the context of input design for dynamic system identification, using successive solutions to linear programs.

We continue with another useful theorem which establishes equivalence for design measures of different optimality criteria.

Theorem 3 If ϕ is differentiable at all points of $\mathcal{M}^+$, the subset of $\mathcal{M}$ where $\phi(M) < \infty$, and a ϕ-optimal design exists, then ξ^* is ϕ-optimal if and only if

$$\min_{x \in \mathcal{X}} D_\phi(M(\xi^*), xx^\top - M(\xi^*)) = \max_{\xi} \min_{x \in \mathcal{X}} D_\phi(M(\xi), xx^\top - M(\xi))$$

where the maximization is over the set $\{\xi | M(\xi) \in \mathcal{M}^+\}$.

PROOF: See Silvey (1980).

We can use the above theorem together with the result given by Case 1 to establish the equivalence for design measures of D- and G-

optimality. Consider a D-optimal measure $\xi^* \in \mathcal{M}^+$, and note that $D_\phi(M(\xi^*), xx^\top - M(\xi^*)) = p - x^\top M(\xi^*)^{-1}x$. Theorem 3 yields

$$\max_{x\in\mathcal{X}} x^\top M(\xi^*)^{-1}x = \min_{\xi} \max_{x\in\mathcal{X}} x^\top M(\xi)^{-1}x$$

which implies that ξ^* is G-optimal. In an identical way, we conclude that any G-optimal design is D-optimal. For further comments on this equivalence result, see Fedorov (1972), page 71.

Remark 3 Particular attention should be paid to the case where $M(\xi)$ is singular. If such case for example occurs for the D-optimality criterion, the directional derivative at $M(\xi)$ will be non-existent. Gaffke (1985) shows how singularity can be overcome. This problem is also treated in Pukelsheim (1980) and Silvey (1980).

2.6 Numerical Algorithms

Here, we use the developed theory to devise numerical algorithms for optimization. Consider Procedure 1 below.

Procedure 1 Start with a design $\xi^{(1)}$. Set the iteration number $k = 1$. Then iteratively repeat the following.

Step 1: Solve $\min_{x\in\mathcal{X}} D_\phi(M(\xi^{(k)}), xx^\top - M(\xi^{(k)}))$. If the minimum value is larger than or equal to zero then stop. Otherwise denote a solution by $x^{(k)}$ and the single point design concentrated at $x^{(k)}$ by $\bar{\xi}^{(k)}$, and continue.

Step2: Update $\xi^{(k+1)} = \xi^{(k)} + \gamma_k(\bar{\xi}^{(k)} - \xi^{(k)})$, $0 \le \gamma_k \le 1$, $\lim_{k\to\infty} \gamma_k = 0$, and $\sum_{k=1}^{\infty} \gamma_k = \infty$. Increment k by one.

The procedure can be explained as follows. The stop criterion in Step 1 according to Theorem 2 is a necessary and sufficient condition for optimality of $\xi^{(k)}$ if $\phi(\cdot)$ is differentiable at $M(\xi^{(k)})$, and

$x^{(k)}$ corresponds to the steepest descent direction in case that $\xi^{(k)}$ is non-optimal. The updating of the design measure in Step 2 is a consequence of the fact that for any two measures ξ_1 and ξ_2 and real numbers $a_1, a_2 \geq 0$, $a_1 + a_2 = 1$, we have

$$M(a_1\xi_1 + a_2\xi_2) = a_1 M(\xi_1) + a_2 M(\xi_2).$$

Now, note that $\xi^{(k+1)} = (1 - \gamma_k)\xi^{(k)} + \gamma_k \bar{\xi}^{(k)}$, and conclude that

$$M(\xi^{(k+1)}) = M(\xi^{(k)}) + \gamma_k (M(\bar{\xi}^{(k)}) - M(\xi^{(k)})),$$

or since $\bar{\xi}^{(k)}$ is a single support measure concentrated at $x^{(k)}$, then

$$M(\xi^{(k+1)}) = M(\xi^{(k)}) + \gamma_k (x^{(k)} x^{(k)^\top} - M(\xi^{(k)}))$$

which implies that $M(\xi^{(k+1)})$ is obtained by updating $M(\xi^{(k)})$ in the steepest descent direction.

Gaffke and Mathar (1992) treat a more general class of algorithms. Their problem setting concerns optimization over a compact convex subset of some finite dimensional Hilbert space. The interior of the convex set is assumed to be nonempty and the function to be optimized is assumed to be differentiable at all interior points. The setting of Gaffke and Mathar (1992) also considers the cases where the function is nondifferentiable at some boundary points, and it is pointed out that the algorithm may go unstable if a gradient near those points is computed. They also establish convergence of the algorithm for more general step lengths (γ_k) than those given in Procedure 1. The choice of step lengths as in Procedure 1 provides a basis for sequential experiment design, see e.g. Fedorov (1972), Chapter 4.

It should be clear by now that the necessary and sufficient condition for optimality of a design ξ^* is given as a condition on $M(\xi^*) \in \mathcal{M}$ (rather than a condition on ξ^* directly, see Theorem 2). The sufficient and necessary optimality condition also appears in Step 1 of

Procedure 1. The discussion following the procedure explains that the updating of $\xi^{(k)}$ in Step 2 indeed corresponds to an updating of $M(\xi^{(k)})$ in $\mathcal{M}$, in the steepest descent direction. In other words, the problem of finding an optimal design measure is translated into an optimization over a convex (compact) subset of a finite dimensional $p(p+1)/2$-space. This makes the connection between the setting here and the setting of Gaffke and Mathar (1992) clearer.

Theorem 4 Let ϕ be differentiable at all points of $\mathcal{M}$. Procedure 1 either stops at some iteration, or otherwise

$$\lim_{k\to\infty} \phi(M(\xi^{(k)})) = \inf_{M\in\mathcal{M}} \phi(M).$$

PROOF: See Gaffke and Mathar (1992), Theorem 3.2.

This concludes our brief summary of the linear theory of experimental design.

3 Nonlinear Regression

3.1 Problem Setting

Consider nonlinear regression models

$$E(y|\theta, u) = \eta(u;\theta),$$

where $\eta(u;\theta)$ is differentiable, and the observation errors are i.i.d. random variables with

$$\mathrm{var}(y|\theta, u) = v.$$

In case data are Gaussian, it follows from the definition of Fisher's information matrix that the same expression for the (asymptotic)

covariance of the maximum likelihood estimate of θ can be obtained as in the linear case after replacing x by $f(u;\theta) = \partial\eta(u;\theta)/\partial\theta$. We are therefore interested in optimizing a scalar function of the matrix

$$M(U;\theta) = \sum f(u_{(i)};\theta) f(u_{(i)};\theta)^{\mathsf{T}}.$$

Now, the problem of interest can be formulated as the optimization:

$$\min_U \phi(M(U;\theta)).$$

This optimization problem can not be solved since the unknown parameter θ is involved in the criterion. Three possible solutions may be suggested:

- fix θ to some *a priori* guessed value,
- define a probability distribution for the possible values of $\theta \in \Theta \subset \mathbb{R}^p$ and solve $\min_U E_\theta\{\phi(M(U;\theta)\}$ where $E_\theta\{\cdot\}$ denotes mean value with respect to θ (we refer to this type of criteria as *Bayesian* criteria),
- consider a minmax type of criterion, i.e. $\min_U \max_{\theta\in\Theta\subset\mathbb{R}^p} \phi(M(U;\theta))$.

In connection with the Bayesian criterion $\min_U E_\theta\{\phi(M(U;\theta)\}$, it is important to note that the approach is fundamentally different from the so called Bayesian experiment design (see e.g. Pilz (1991)). Bayesian experiment design refers to optimization of design criteria that are related to a Bayesian estimation of the parameters. As it is evident from the discussions so far, the optimization criteria of interest in this article are related to maximum likelihood (or least squares) estimators. Fedorov and Atkinson (1988) dub the approach of optimizing $\min_U E_\theta\{\phi(M(U;\theta))\}$, as "pseudo-Bayesian" instead.

The first possible solution of fixing θ to some *a priori* value is the easiest one. By fixing the parameter to, say $\theta_0 \in \Theta$, the problem

translates into a linear regression setting where $x = f(u;\theta_0)$. Despite this simplicity, the approach may not be useful since the obtained design may exhibit poor properties for other possible values of $\theta \in \Theta$. We are here interested in designs that are *robust* to prior uncertainty in the parameter value. Our approach for design of robust experiments is based on the Bayesian criteria. For a treatment of minmax criteria, see Fedorov and Atkinson (1988) where necessary and sufficient conditions for minmax optimality are given. See also Pronzato and Walter (1987).

3.2 Bayesian Criteria

In this subsection, we consider optimization of the Bayesian criteria, i.e. optimization of the type

$$\min_U E_\theta\{\phi(M(U;\theta))\},$$

where $M(U;\theta) = \sum f(u_{(i)};\theta)f(u_{(i)};\theta)^\top$. Transferring the idea of the approximate theory (see Subsection 2.3) directly, we obtain that $M(\xi;\theta) = \int_{\mathcal{U}} f(u;\theta)f(u;\theta)^\top\xi(du)$. The problem is to find a measure ξ among the set of all probability measures defined over $\mathcal{U}$, a compact subset of the Euclidean r-space, in order to minimize $E_\theta\{\phi(M(\xi;\theta))\}$.

First, we have to investigate whether or not the linear theory can be fully or partially employed for the optimization of Bayesian criteria. Unfortunately, the fact that the parameter θ is not fixed any longer is devastating to the theory developed in the previous section. The reason is that unlike the linear case, it is not possible to perform the optimization over a subset of some finite-dimensional space. Recall that the necessary and sufficient optimality condition (Theorem 2) and the numerical procedure are derived under the consideration

that one is able to perform the optimization over a convex compact subset of $p(p+1)/2$-space. In the nonlinear case, it is in general not possible to identify such convex subset since the parameter is not fixed. In the following, we present some (approximate) techniques for optimization of the Bayesian criteria.

Discretization

The idea is based on defining a discrete grid over the set $\mathcal{U}$. Let $G_u = \{u_{(i)}, i = 1, ..., N_G\}$ denote the selected grid. We will be interested in finding a set of probabilities defined over G_u. Despite the similarities of this approach to the N-observation design setting (see Subsection 2.3), the two problems are basically different since the restriction that the probabilities should be multiple integers of $1/N$ does not apply here.

Let $\alpha = \{\alpha_1, ..., \alpha_{N_G}\}$ denote a set of probability weights defined over G_u, i.e. the weights $\alpha_i \in \mathbb{R}$ satisfy $\sum \alpha_i = 1$, $\alpha_i \geq 0$. Let $\mathcal{A}$ denote the compact convex subset of $\mathbb{R}^{N_G}$ which is the convex hull of the points specified by the unit length vectors that have 1 in the entry i, $i = 1, 2, ..., N_G$. Denoting $\phi_B(\alpha) = E_\theta\{\phi(M(\alpha;\theta))\}$ where $\phi(M(\alpha;\theta)) = \sum \alpha_i M(u_{(i)};\theta))$, the problem transforms to $\min_{\alpha\in\mathcal{A}} \phi_B(\alpha)$. Now, we have

Theorem 5 The function $\phi_B(\alpha)$ is convex.

PROOF: See Sadegh, Holst, Madsen, and Melgaard (1995) (see article A2 for a revised version), Theorem 5.

In the following, we also assume differentiability of $\phi_B(\cdot)$ at the point of interest. Wets (1989), page 582, Proposition 2.7, gives regularity conditions under which an integral functional is differentiable. Similar to Theorem 2, we have

Theorem 6 If ϕ_B is differentiable at some $\alpha^* \in \mathcal{A}$, then α^* is optimal if and only if

$$E_\theta\{D_\phi(M(\alpha^*;\theta), f(u;\theta)f(u;\theta)^\top - M(\alpha^*;\theta))\} \geq 0$$

for all $u \in G_u$.

PROOF: Identical to the proof of Theorem 2, it can be shown that a necessary and sufficient condition for optimality of α^* is that for all $i = 1, ..., N_G$,

$$D_{\phi_B}(\alpha^*, \alpha^{(i)} - \alpha^*) \geq 0,$$

where $\alpha^{(i)}$ is a unit length vector in direction i. The theorem follows since

$$D_{\phi_B}(\alpha^*, \alpha^{(i)} - \alpha^*) = E_\theta\{D_\phi(M(\alpha^*;\theta), f(u_{(i)};\theta)f(u_{(i)};\theta)^\top - M(\alpha^*;\theta))\}$$

The theorem implies that one is required to solve

$$\min_{u \in G_u} E_\theta\{D_\phi(M(\alpha;\theta), f(u;\theta)f(u;\theta)^\top - M(\alpha;\theta))\},$$

in order to determine whether or not a given design α is optimal. This in general requires N_G integral evaluations. If the number of grid points is relatively small such that it is feasible to evaluate N_G integrals each time a steepest descent direction is to be computed, the theorem provides a basis for optimization algorithms similar to those described in Subsection 2.6. The only modification needed is to replace ξ by α throughout and replace Step 1 of Procedure 1 by

Step 1′ : Solve $\min_{u \in G_u} E_\theta\{D_\phi(M(\alpha^{(k)};\theta), f(u;\theta)f(u;\theta)^\top - M(\alpha^{(k)};\theta))\}$. If the minimum value is larger than or equal to zero then stop. Otherwise denote a solution by $u^{(k)}$, and the single point design concentrated at $u^{(k)}$ by $\bar{\alpha}^{(k)}$, and continue.

The problem setting of Gaffke and Mathar (1992) can be applied by noting that the optimization is performed over $\mathbb{R}^{N_G}$ where scalar product is defined by the usual scalar product between real vectors. Moreover, $\mathcal{A}$ is a convex compact subset of $\mathbb{R}^{N_G}$. Assuming differentiability of ϕ_B on open set containing $\mathcal{A}$, a result similar to Theorem 4 is obtained.

Theorem 7 Let ϕ_B be differentiable on an open set containing $\mathcal{A}$. Replace Step 1 of Procedure 1 by Step 1′, and replace ξ by α throughout. The procedure either stops at some iteration, or otherwise

$$\lim_{k\to\infty} \phi_B(\alpha^{(k)}) = \inf_{\alpha\in\mathcal{A}} \phi_B(\alpha).$$

In Sadegh, Holst, Madsen, and Melgaard (1995) (see article A2 for a revised version) two examples in the context of input design for dynamic system identification are given where such procedures are feasible since the number of grid points N_G is relatively small.

Remarks on Optimization of the Bayesian Criteria

In case the number of grid points N_G is large, it may not be feasible to evaluate N_G integrals in order to find a steepest descent direction at each iteration. A closely related approach in such cases is as follows (see Fedorov and Atkinson (1988)). The same procedure as before is applied but Step 1′ is modified to Step 1″ below.

Step 1″ : Solve

$$\min_{u\in\mathcal{U}} [E_\theta\{D_\phi(M(\alpha^{(k)};\theta), f(u;\theta)f(u;\theta)^\top - M(\alpha^{(k)};\theta))\}].$$

If the minimum value is larger than or equal to zero then stop. Otherwise denote a solution by $u^{(k)}$, add $u^{(k)}$ to the grid (if

it does not belong to the grid), denote the single point design concentrated at $u^{(k)}$ by $\bar{\alpha}^{(k)}$, and continue.

Notice that the only differences between this approach and the previous one are that the optimization here is over the (continuous) set $\mathcal{U}$ instead of the (discrete) set G_u, and the grid is expanded by one element at each iteration (if the solution to the minimization of Step 1'' does not belong to the grid).

Although the optimization over $\mathcal{U}$ is simpler than the optimization over G_u (for large number grid points), each function evaluation still requires integration which in general is time consuming. A suitable approach for optimization of integral based objective functions is stochastic approximation (SA), see e.g. Kushner and Clark (1978). SA algorithms in a Robbins-Monro setting estimate a stationary point of a function (i.e., a zero of the gradient equation for unconstrained optimization or a zero of the Lagrangian equation for constrained optimization) based on iterative updating of the optimization variable, using noisy evaluations of the gradient of the objective function at the current iterate where the noise on the gradient evaluations should satisfy certain conditions (see Kushner and Clark (1978)). Pronzato and Walter (1987) report a Robbins-Monro type SA approach in connection with optimization of Bayesian criteria arising in experimental design. In Fedorov and Atkinson (1988), the SA approach to the optimization above and its associated problem of selecting stop criterion are discussed.

4 Conclusion

We have studied the theory of experiment design for regression models. The study includes the linear theory where most of the theo-

retical results are available. Theoretical results for design in nonlinear cases, on the other hand, are sparse. We presented (approximative) numerical solutions for optimization of the Bayesian criteria arising in the experimental design for nonlinear cases.

References

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis.* Canadian Mathematical Society Series of Monographs and Advanced Texts. John Wiley and Sons, New York.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Academic, New York.

Fedorov, V. V. and A. C. Atkinson (1988). The optimum design of experiments in the presence of uncontrolled variability and prior information. In Y. Dodge, V. V. Fedorov, and H. P. Wynn (Eds.), *Optimal Design and Analysis of Experiments*, pp. 327–344. Elsevier Science Publishers B.V., Amsterdam.

Gaffke, N. (1985). Directional derivatives of optimality criteria at singular matrices in convex design theory. *Statistics 3*, 373–388.

Gaffke, N. and R. Mathar (1992). On a class of algorithms from experimental design theory. *Optimization 24*, 91–126.

Goodwin, G. C. and R. L. Payne (1977). *Dynamic System Identification: Experiment Design and Data Analysis.* Academic, New York.

Kiefer, J. (1959). Optimum experimental designs. *J. R. Statist. Soc. B. 21*, 272–319.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Stat. 2*(5), 849–879.

Kiefer, J. and J. Wolfowitz (1960). The equivalence of two extremum problems. *Canad. J. Math 12*, 363–366.

Kushner, H. J. and D. S. Clark (1978). *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer-Verlag, Berlin.

Pilz, J. (1991). *Bayesian Estimation and Experiment Design in Linear Regression Models.* John Wiley and Sons, New York.

Pronzato, L. and E. Walter (1987). Robust experiment design for nonlinear regression models. In V. Fedorov and H. Läuter (Eds.), *Lecture Notes in Economics and Mathematical Sciences,297*, pp. 72–86. Springer-Verlag, Berlin.

Pukelsheim, F. (1980). On linear regression designs which maximize information. *J. Statist. Plann. Infer 4*, 339–364.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* 2nd edition, Wiley.

Rockafellar, R. (1970). *Convex Analysis.* Princeton University Press.

Sadegh, P., L. H. Hansen, H. Madsen, and J. Holst (1996). Maxmin input design for linear dynamic systems. To be submitted.

Sadegh, P., J. Holst, H. Madsen, and H. Melgaard (1995). Experiment design for grey-box identification. *International Journal of Adaptive Control and Signal Processing 9*(6), 491–507.

Sibson, R. (1974). D_A-optimality and duality. *Colloq. Math. Soc. Janos Bolyai 9*, 677–692.

Silvey, S. D. (1980). *Optimal Design.* Chapman & Hall, London.

Wets, R. J. (1989). Stochastic programming. In G. L. Nemhauser, K. A. H. G. Rinnoo, and M. J. Todd (Eds.), *Handbooks in Operations Research and Management Science*, Volume 1, Chapter 2, pp. 573–629. Amsterdam: North-Holland.

Experiment Design for Gray-box Identification

Payman Sadegh* **Jan Holst+** **Henrik Madsen*** and **Henrik Melgaard***

*Institute of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

+ Institute of Mathematical Statistics, Lund Institute of Technology, Lund S-22100, Sweden.

This is a revised version of a paper published in *Int. Journal of Adaptive Control and Signal Processing*, Vol.9,491-507 (1995).

[A2]

Abstract

Gray-box models are characterized by their physical significance and partial prior information. These aspects of gray-box models affect the design of optimal excitations for identification, and we study the extension of the classical theory for experiment design to input design for identification of gray-box models. Partial prior information is expressed as a probability distribution for the parameter and is employed in the design of optimal excitations through optimization of Bayesian criteria.

Key words: Gray-box Identification, Experiment Design, Bayesian Criterion, Physical Models.

1 Introduction

This paper is concerned with the design of optimal identification experiments for dynamic gray-box models with focus on the design of best excitation signals. The question of how best to excite a system is often raised at one of the earliest stages of the whole system identification procedure. This involves designing excitations that maximize the informational value of data given the experimental constraints. An even more fundamental stage involves the question of model nature and model structure. At one end of the model spectrum, we find transparent models, i.e. models totally obtained from physics, chemistry, etc. At the other end, black-box models without physical background are found. Such models usually find application in areas where the system to be studied is too complex or the physical knowledge about the system is poor. In between are gray-box models obtained from both physics and experimental data. Even though the border-lines between black, gray, and transparent models are not well defined, it is possible to identify some general features of gray-box models. Physical significance and *partial* information about the system are the key-words we are concerned with in this paper. The latter has recently become the focus of systematic study and attention in the system identification community, see e.g. Bohlin (1984) and Bohlin (1989). Also see Blanke and Söderström (1994) for papers on gray-box modeling issues.

Suppose that a linear dynamic system is given. Assume that the system is a member of a general parametric model set, i.e. each member of the model set is totally characterized by a parameter. The parametrization is imposed by physics and is therefore allowed to be nonlinear. We wish to obtain accurate estimates of the parameter based on experimental data which are regarded as realizations of

some random process. Recalling that the lower bound on the asymptotic covariance of any unbiased estimator of the parameter is given by the inverse of Fisher's information matrix (Cramér-Rao inequality), it is reasonable to select the experimental condition so as to maximize the information matrix in some sense. As illustrated later, the approach can also be justified using decision theoretic considerations.

Since the information matrix is in general dependent upon the unknown parameter value, problem formulations based on maximizing the information matrix cannot be solved if no information about the parameter is available. We here consider the realistic situation that the value of the parameter is only partially known, and assume that the prior information about the parameter is embedded in a probability distribution. One easy approach to employ the available information involves evaluating the information matrix at some representative parameter value, say the prior mean of the given probability distribution. However, the optimal experiment derived hereby might exhibit poor properties if the true parameter value is different from the representative value. To solve this problem, we consider optimization of the mean value (with respect to the parameter) of some suitable scalar function of the information matrix. The corresponding criterion is referred to as Bayesian criterion.

The organization of the rest of the paper is as follows. First, a general theory is developed and presented to extend the classical experiment design to the design of optimal excitations for modeling of dynamic systems. All the nice theoretical properties of the classical experiment design are preserved if the design criteria are evaluated at some representative parameter point. Sections 2 and 3 focus on this topic, inspired by Goodwin and Payne (1977) and Silvey (1980). Then in Section 4, the application of the general theory to experi-

ment design for estimation of physical models is investigated. The design of experiments using Bayesian criteria is discussed in Section 5, which is followed by conclusions in Section 6.

It is important to note that the Bayesian approach of the present paper should not be confused with *Bayesian experiment design* (see Pilz (1991)) which is related to design for Bayesian estimators. The Bayesian approach here should be regarded as a useful and realistic way of employing the available partial knowledge for the design of *robust* experiments, i.e. experiments that are robust to prior model uncertainty.

2 The Information Matrix

Suppose that the distribution of data y is specified by a density function $g(y|\theta)$ where $\theta \in \Theta \subset \mathbb{R}^p$ is a parameter. We introduce the following definition.

Definition 1 Fisher's information matrix is defined by

$$M_F = E\{(\frac{\partial g(y|\theta)}{\partial \theta})^\top (\frac{\partial g(y|\theta)}{\partial \theta})\},$$

where $E\{\cdot\}$ denotes mean value. The average information matrix per sample is defined by $M = \lim_{N\to\infty} M_F/N$ where N is the total number of samples.

The importance of the information matrix is highlighted by the Cramér-Rao inequality which states that under certain regularity conditions, the lower bound on the (asymptotic) covariance of any unbiased estimator of θ is given by the inverse of the information matrix. In the paper, we will be concerned with maximizing the average information matrix in some sense.

Now consider the following linear time-invariant stochastic system

$$y_t = G_1(q^{-1})u_t + G_2(q^{-1})\epsilon_t, \quad t = ..., -\Delta T, 0, \Delta T, ... \tag{2.1}$$

where $\{u_t\}$, $\{y_t\}$ are input and output sequences, $\{\epsilon_t\}$ is a sequence of Gaussian i.i.d. random variables with covariance Σ, $G_1(q^{-1})$ and $G_2(q^{-1})$ are transfer functions in the backward shift operator q^{-1}, and ΔT is the sampling time. The transfer functions G_1, G_2, and the covariance Σ are dependent upon the parameter θ. We introduce the following assumptions:

A 1 the experiments are performed in open loop, i.e. the noise and the input sequences are uncorrelated,

A 2 the number of samples is large,

A 3 the input is restricted to the signals that admit a spectral representation,

A 4 the input power is restricted.

Without loss of generality, this last assumption implies

$$\int_0^{\frac{\pi}{\Delta T}} d\xi(\omega) = 1,$$

where $\xi(\omega)$ denotes the power spectral distribution of the input defined over the frequency range $[0, \pi/\Delta T]$. The distribution ξ will also be referred to as the *input design measure.* In the following, we present a frequency domain formula for computation of the average information matrix for the output of the system given by equation (2.1). The frequency domain formulation results in considerable simplicity, see Goodwin and Payne (1977). For convenience, we restrict attention to single-input single-output systems.

Theorem 1 Consider the system given by equation (2.1). The average information matrix per sample for the output data of the system is given by

$$M = \int_0^{\frac{\pi}{\Delta T}} \tilde{M}(\omega) d\xi(\omega) \tag{2.2}$$

where

$$\begin{aligned} \tilde{M}(\omega) &= \Re\{\frac{1}{\pi\Sigma}[\frac{\partial G_1(e^{\jmath\omega\Delta T})}{\partial\theta}]^\top G_2^{-1}(e^{\jmath\omega\Delta T}) \\ &\quad G_2^{-1}(e^{-\jmath\omega\Delta T})[\frac{\partial G_1(e^{-\jmath\omega\Delta T})}{\partial\theta}]\} + \bar{M}_c, \end{aligned}$$

($\Re$ denotes real part) and the constant (with respect to input) matrix $\bar{M}_c$ is given by

$$\begin{aligned} \bar{M}_c &= \frac{1}{2\pi}\int_{-\frac{\pi}{\Delta T}}^{\frac{\pi}{\Delta T}} [\frac{\partial G_2(e^{\jmath\omega\Delta T})}{\partial\theta}]^\top G_2^{-1}(e^{\jmath\omega\Delta T}) \\ &\quad G_2^{-1}(e^{-\jmath\omega\Delta T})[\frac{\partial G_2(e^{-\jmath\omega\Delta T})}{\partial\theta}]d\omega + \frac{1}{2\Sigma^2}(\frac{\partial\Sigma}{\partial\beta})^\top(\frac{\partial\Sigma}{\partial\beta}). \end{aligned}$$

PROOF: See Goodwin and Payne (1977).

Note that since M is a symmetric $p \times p$ matrix, it can be represented by a point in $\mathbb{R}^{\frac{p(p+1)}{2}}$. The matrix is moreover nonnegative definite. Denote the set of all admissible design measures by $\mathcal{X}$. We have the following theorem.

Theorem 2 Subject to the input power constraint, the set of all average information matrices

$$\mathcal{M} = \{M | M = \int_0^{\frac{\pi}{\Delta T}} \tilde{M}(\omega) d\xi(\omega), \int_0^{\frac{\pi}{\Delta T}} d\xi(\omega) = 1\},$$

is the convex hull of the average information matrices corresponding to single frequency design measures.

PROOF: It follows immediately from the definition of $\mathcal{M}$ that it is the convex hull of the set of points $\{\tilde{M}(\omega)|\omega \in [0, \pi/\Delta T]\}$. Besides, from equation (2.2), it is obvious that the average information matrix corresponding to a power restricted single frequency input measure is given by $\tilde{M}(\omega)$ where ω is that single frequency. The theorem then follows, see e.g. Goodwin and Payne (1977) and Silvey (1980).

For reasons that will be clear later, inputs comprising finitely many sinusoidal components (line spectra) are of particular interest. Such an input is totally characterized by the set $\{(\alpha_i, \omega_i)|i = 1, ..., l\}$ where $\alpha_i \geq 0$ is the power share of the input spectral content at ω_i and l is the finite number of sinusoidal components. Note that the input power constraint implies that for a line spectrum $\sum_i \alpha_i = 1$.

3 Optimality Criteria and Design of Experiments

As mentioned earlier, our original intent is to maximize the (average) information matrix in some sense. This is usually done by minimizing a suitably chosen scalar function of the information matrix. First, we illustrate how a decision theory approach including a loss function, which is related to the intended application of the model, and prior distribution for the parameter, which expresses the partial prior knowledge about the parameter value, leads to optimization of the mean (with respect to the parameter) of some scalar function of the information matrix.

Let $L(\theta, \hat{\theta})$ denote a suitably defined loss function where θ is the true parameter and $\hat{\theta}$ is an estimator for θ. Assume that $\hat{\theta}$ is unbiased and efficient (i.e., the asymptotic covariance reaches the lower bound

of the Cramér-Rao inequality). We consider minimization of

$$J = E_\theta E\{L(\theta, \hat{\theta})|\theta\}$$

where $E_\theta\{\cdot\}$ denotes mean value with respect to θ, i.e. the expectation is taken with respect to both data and θ. Using a second order approximation around $\hat{\theta} = \theta$, the criterion is asymptotically given by

$$\begin{aligned} J &= E_\theta E\{L(\theta, \hat{\theta})|\theta\} \\ &\simeq E_\theta E\{L(\theta,\theta) + (\frac{\partial L}{\partial \hat{\theta}})(\hat{\theta}-\theta) + \frac{1}{2}(\hat{\theta}-\theta)^\top(\frac{\partial^2 L}{\partial \hat{\theta}^2})(\hat{\theta}-\theta)|\theta\} \\ &= E_\theta\{L(\theta,\theta) + \frac{1}{2}\text{trace}(\frac{\partial^2 L}{\partial \hat{\theta}^2})M_F^{-1}\}, \end{aligned}$$

where we have replaced $E\{(\hat{\theta}-\theta)(\hat{\theta}-\theta)^\top|\theta\}$ by M_F^{-1}. Defining $L(\theta,\theta) = 0$ we conclude that the problem is reduced to that of minimizing

$$J = E_\theta\{\text{trace}(WM_F^{-1})\}$$

where $W = \partial^2 L/\partial\hat{\theta}^2$. All the derivatives are evaluated at $\hat{\theta} = \theta$.

Example 1 Consider the very simple system $y_t = \theta u_t + \epsilon_t$, $\theta \neq 0$, where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables. Assume that the modeling goal is to design a controller to keep the output as close as possible to the reference signal $y_{ref} = 1$. The minimum power of $(y_t - y_{ref})$ is clearly obtained by choosing $u_t = 1/\theta$. However, the true parameter θ is not known and it is assumed that it must be estimated during an open loop identification experiment prior to the application of the controller. It is desirable to derive an optimization criterion for designing the best identification input.

The extra output error introduced by using the unbiased efficient estimate $\hat{\theta}$ in the control computation is given by $(\hat{\theta}-\theta)/\hat{\theta}$ which contributes to the power of the error signal $(y_t - y_{ref})$ by

$[(\hat{\theta}-\theta)/\hat{\theta}]^2$. Thus, it is reasonable to define $L(\theta,\hat{\theta}) = (\hat{\theta}-\theta)^2$, and $E_\theta E\{L(\theta,\hat{\theta})|\theta\} = E_\theta\{\text{var}(\hat{\theta})\}$ where $\text{var}(\hat{\theta})$ denotes the variance of the estimate $\hat{\theta}$. This variance is asymptotically given by the lower bound of the Cramér-Rao inequality, i.e. the inverse of the information matrix. In this simple case the information matrix is independent of θ, which leads to optimization of M_F^{-1}.

Thus, we have related a problem formulation based on decision theoretic considerations to one concerning maximization of the mean (with respect to θ) of a scalar function of the information matrix. Throughout the rest of the paper, we consider optimization problems of the form $E_\theta\{\phi(M)\}$, where $\phi(\cdot)$ is a scalar function of its matrix valued argument. The optimization criterion is expressed in a Bayesian form which takes into account the fact that the prior information about θ is partial.

However, assume for the moment, and for the rest of this section, that the Bayesian criterion is approximated by replacing θ by its prior mean, θ_0, resulting in the non-Bayesian criterion $\phi(M)|_{\theta_0}$. We will return back to the Bayesian criterion later on, in Section 5.

Definition 2 *(ϕ-function)* Consider the mapping

$$\phi : \mathcal{M} \to \mathbb{R}.$$

If $\phi(\cdot)$ is non-increasing, i.e.

$$M_1 - M_2 \quad \text{nonnegative definite} \Rightarrow \phi(M_1) \leq \phi(M_2),$$

it is called a ϕ-function.

Definition 3 **(ϕ-optimality)** Consider a ϕ-function defined over $\mathcal{M}$. Then an input measure ξ^* is called ϕ-optimal if

$$\phi(M(\xi^*)) \leq \phi(M(\xi)), \quad \forall \xi \in \mathcal{X}$$

where $\mathcal{X}$ is the set of all admissible input measures and $M(\xi)$ is the average information matrix corresponding to the measure ξ. See Silvey (1980).

Example 2 $\phi(M) = -\log\det(M)$ is an example of a ϕ-function. This choice of $\phi(\cdot)$ is called D-optimality in the literature (see e.g. Fedorov (1972)).

Example 3 Another interesting choice is the D_s-optimality. Partition the parameter vector as $\theta = (\theta_1^\top, \theta_2^\top)^\top$ where θ_1 is the parameter of interest. Denote the corresponding information matrix by M and partition

$$M = \left\{ \begin{array}{cc} M_{11} & M_{12} \\ M_{12}^\top & M_{22} \end{array} \right\}$$

where M_{11} corresponds to the parameter of interest θ_1 and M_{22} corresponds to θ_2. Using the matrix inversion formula, the lower bound on the asymptotic covariance of $\hat{\theta}_1$ is given by $(M_{11} - M_{12}M_{22}^{-1}M_{12}^\top)^{-1}$, which leads to the ϕ-function corresponding to D_s-optimality

$$\phi(M) = -\log\det(M_{11} - M_{12}M_{22}^{-1}M_{12}^\top).$$

Many interesting practical design criteria satisfy the condition of Definition 2 and they are also *convex* on $\mathcal{M}$.

Example 4 Consider the D-optimality criterion again. The fact that (Fedorov (1972))

$$\det(\beta M_1 + (1-\beta)M_2) > [\det(M_1)]^\beta [\det(M_2)]^{(1-\beta)}, 0 < \beta < 1,$$

implies that $\phi(\cdot)$ is convex on $\mathcal{M}$.

From now on, we assume that the ϕ function is convex. Hence, solving for optimal input is equivalent to finding a point in the convex set $\mathcal{M}$ that minimizes the convex function $\phi(\cdot)$.

Theorem 3 ϕ-optimal designs exist comprising not more than $p(p+1)/2+1$ sinusoidal components. Furthermore, if the ϕ-function is strictly decreasing (i.e., nonnegative definite non-zero $M_1 - M_2$ yields $\phi(M_1) < \phi(M_2)$), the upper bound on the number of sinusoidal components that suffice for optimality is $p(p+1)/2$.

PROOF: See Goodwin and Payne (1977) and Silvey (1980).

The theorem implies that optimal design measures can be sought among line spectra with $l \leq p(p+1)/2+1$ or $l \leq p(p+1)/2$.

3.1 Algorithms for Constructing Optimal Design

One way of constructing optimal designs is by direct optimization. Theorem 3 implies that it is sufficient to write each candidate point in $\mathcal{M}$ as

$$\sum_{i=1}^{N_B} \alpha_i \tilde{M}(\omega_i),$$

where $N_B = p(p+1)/2+1$ and search over a $2N_B$ dimensional space to find the α's and ω's that minimize ϕ subject to

$$\sum_{i=1}^{N_B} \alpha_i = 1, \quad \forall i : \alpha_i > 0.$$

If $\phi(\cdot)$ is strictly decreasing then $N_B = p(p+1)/2$.

The other algorithm which is more conceptual is inspired by Silvey (1980), and based on the following theorem. We start by defining the Fréchet derivative

Definition 4 The Fréchet derivative of a ϕ-function at M_1 in the direction of M_2 is

$$F_\phi(M_1, M_2) = \lim_{\gamma \to 0^+} \frac{1}{\gamma}[\phi((1-\gamma)M_1 + \gamma M_2) - \phi(M_1)].$$

Theorem 4 Assume that a ϕ-function is convex and differentiable at $M(\xi^*)$. Then $\xi^* \in \mathcal{X}$ is ϕ-optimal iff $F_\phi(M(\xi^*), \tilde{M}(\omega)) \geq 0$ for all $\omega \in [0, \pi/\Delta T]$.

PROOF: Necessity: The fact that $\phi(\cdot)$ attains its minimum at $M(\xi^*)$ implies

$$\phi(M((1-\gamma)\xi^* + \gamma\xi)) - \phi(M(\xi^*)) \geq 0,$$

for all $\gamma \in [0,1]$ and all $\xi \in \mathcal{X}$. But, note that

$$(1-\gamma)M(\xi^*) + \gamma M(\xi) = M((1-\gamma)\xi^* + \gamma\xi).$$

Take ξ to be any single frequency design, i.e. $M(\xi) = \tilde{M}(\omega)$, $\omega \in [0, \pi/\Delta T]$. Then, the necessity follows from the definition of the Fréchet derivative.

Sufficiency: Each $M(\xi) \in \mathcal{M}$ is written as

$$M(\xi) = \sum_{i=1}^{l} \alpha_i \tilde{M}(\omega_i)$$

where $l \leq p(p+1)/2 + 1$. Since $\phi(\cdot)$ is differentiable at $M(\xi^*)$

$$F_\phi(M(\xi^*), M(\xi)) = \sum_i \alpha_i F_\phi(M(\xi^*), \tilde{M}(\omega_i))$$

where this last equality follows from the fact that the Fréchet derivative is linear in the second argument, see e.g. Rockafellar (1970). Then $F_\phi(M(\xi^*), \tilde{M}(\omega_i)) \geq 0$ for all ω_i implies that

$$F_\phi(M(\xi^*), M(\xi)) \geq 0.$$

However, convexity of $\phi(\cdot)$ implies that for any $M_1, M_2 \in \mathcal{M}$

$$\frac{1}{\gamma}[\phi((1-\gamma)M_1 + \gamma M_2) - \phi(M_1)]$$

is a non-decreasing function of γ, see Whittle (1971). Hence, setting $\gamma = 1$ yields

$$F_\phi\left(M(\xi^*), M(\xi)\right) \leq \phi(M(\xi)) - \phi(M(\xi^*)),$$

and the sufficiency follows.

Remark 1 The above theorem can be used to devise simple numerical algorithms for optimization. Consider $\phi(M) = -\log\det(M)$. For a square non-singular matrix M, we have $\partial \log\det M/\partial M = M^{-\mathsf{T}}$ (see e.g. Goodwin and Payne (1977)), then

$$\begin{aligned} F_\phi\{M(\xi^*), \tilde{M}(\omega)\} &= \lim_{\beta\to 0+} \frac{\partial}{\partial\beta}\{-\log\det[(1-\beta)M(\xi^*) + \beta\tilde{M}(\omega)]\} \\ &= -\mathrm{trace}([M(\xi^*)]^{-1}[\tilde{M}(\omega) - M(\xi^*)]) \\ &= p - \mathrm{trace}([M(\xi^*)^{-1}\tilde{M}(\omega)]). \end{aligned}$$

Theorem 4 then implies that ξ^* is optimal iff

$$v(\xi^*, \omega) = \mathrm{trace}([M(\xi^*)^{-1}\tilde{M}(\omega)]) \leq p, \quad \forall\omega \in [0, \pi/\Delta T].$$

The function $v(\xi^*, \omega)$ is called response dispersion, see Goodwin and Payne (1977). The algorithm then simply consists of starting with an initial measure and updating according to

$$\xi_{k+1} = (1-\beta_k)\xi_k + \beta_k\xi_k^\circ$$

where ξ_k° is a single frequency measure with the frequency of support $\omega_k^\circ = \arg\max_\omega v(\xi_k, \omega)$. Since $\xi_{k+1} = \xi_k$ for $\beta_k = 0$ and

$$\lim_{\beta_k\to 0+} \frac{\partial}{\partial\beta_k}\{-\log\det M(\xi_{k+1})\} = p - v(\xi_k, \omega_k^\circ) < 0$$

for any non-optimal ξ_k, this update leads to a decrease in the value of the criterion for sufficiently small β_k. Gaffke and Mathar (1992)

show that if $\{\beta_k\}$ is chosen such that $\sum_k \beta_k \to \infty$ and $\beta_k \to 0$ as $k \to \infty$, the algorithm converges to the optimum. Notice that the generic technique can be similarly applied to all differentiable convex ϕ functions.

Now, we are ready to study the problem of optimal excitations for identification of gray-box models.

4 Experiment Design and Physical Models

As mentioned in Section 1, physical significance and partial prior information are among the key features of gray-box models. Physical significance of the model is typically equivalent to nonlinearity of the model parametrization. Even, if attention is restricted to linear systems, the parametrization should be allowed to be nonlinear. In the present section, we briefly discuss the design of optimal experiments for models with physical significance. All the assumptions of the previous sections are also valid here. In the next section, we study the impact of the other feature of gray-box models, namely, the availability of partial prior information.

Suppose that a physical parameter θ_p is related to θ through $\theta = F(\theta_p)$. Then if F is differentiable, a simple application of the chain rule and Definition 1 yield that

$$M_p = (\frac{\partial F(\theta_p)}{\partial \theta_p})^\top M (\frac{\partial F(\theta_p)}{\partial \theta_p}),$$

where M_p is the average information matrix corresponding to parametrization θ_p. Once the single frequency average information matrices $\tilde{M}(\omega)$ for the parametrization θ are computed, it is possible to readily compute the single frequency average information matrices for the

reparametrized model using the above formula. However, it is not readily clear how a solution for the parametrization θ translates to a solution for the parametrization θ_p, and in general, the only way to find the optimal inputs for the reparametrized model is to solve a new optimization problem.

The other point is that models in continuous-time are more interesting in a gray-box setting since the physical knowledge is normally expressed as differential equations. Fortunately, only minor modifications are needed for using the developed theory to continuous-time design problems. Assume that a linear continuous-time model is given by

$$y(t) = G_1(\delta)u(t) + G_2(\delta)\epsilon(t)$$

where $\epsilon(t)$ is continuous-time white noise. Here, δ is the differentiation operator, $u(t)$ and $y(t)$ are input and output signals, respectively, and G_1 and G_2 are transfer functions. All the discussions concerning experimental design for discrete-time models remain valid if the following modifications are made (Goodwin and Payne (1977)):

- replace $e^{\jmath\omega\Delta T}$ by $\jmath\omega$,
- the integral limits are obtained by letting $\Delta T \to 0$.

The convexity of the information matrix and the rest of the theory remain intact.

Example 5 (Design for Estimation of the Heat Dynamics of a Building):

In this example, we investigate the design of optimal experiments for identification of a physical model that describes the heat transfer dynamics of a building, see Madsen and Holst (1995) for a further

discussion of this problem. The model is

$$\begin{bmatrix} dT_m \\ dT_i \end{bmatrix} = \begin{bmatrix} \frac{-1}{R_i C_m} & \frac{1}{R_i C_m} \\ \frac{1}{R_i C_i} & -(\frac{1}{R_a C_i} + \frac{1}{R_i C_i}) \end{bmatrix} \begin{bmatrix} T_m \\ T_i \end{bmatrix} dt$$
$$+ \begin{bmatrix} 0 & 0 & \frac{A_w p_s}{C_m} \\ \frac{1}{R_a C_i} & \frac{1}{C_i} & \frac{A_w(1-p_s)}{C_i} \end{bmatrix} \begin{bmatrix} T_a \\ Q_h \\ Q_s \end{bmatrix} dt + \begin{bmatrix} dw_m(t) \\ dw_i(t) \end{bmatrix} \tag{4.1}$$

where the following physical parameters are to be considered:

- R_i – resistance between the room air and the large heat accumulating medium,
- R_a – resistance between inner walls and the ambient air,
- C_i – capacity of the small heat accumulating medium (the indoor air and the inner part of the walls, see Madsen and Holst (1995)),
- C_m – capacity of the large heat accumulating medium.

The indoor air temperature is measured, and both process and measurement noise terms are considered. The remaining entities in the model (4.1) are

- T_i – temperature of the small heat accumulating medium (indoor air temperature),
- T_m – temperature of the large heat accumulating medium,
- A_w – effective window area,
- p_s – part of the solar radiation which reaches the large heat accumulating medium in the building,
- T_a – ambient air temperature,
- Q_h – heat supplied by heater (controlled input),
- Q_s – heat supplied by solar radiation,
- $w_m(t), w_i(t)$ – elements of a vector Wiener process introduced in order to describe the disturbances and errors in the system

model.

For the input design we only consider the influence on the temperatures in the building caused by the heater, Q_h, i.e. the effects of solar radiation, Q_s, and ambient air temperature, T_a, are neglected. Since only the indoor temperature, T_i, is measured the state-space model can be transformed into a single-input single-output transfer function model where the input is Q_h. We define the parameter vector $(R_i, R_a, C_i, C_m, K_1, K_2, \lambda)^\top$ where K_1, K_2, and λ are Kalman gains in the state estimator and the covariance of the innovation process respectively (see Madsen and Holst (1995)). We assume that the prior mean of the parameter is given by $(1, 1, 1, 4, 0.2, 0.5, 0.01)^\top$.

The D-optimal input, $u_o(t)$ comprises sinusoidal components with frequencies

$$\omega_1 = 0.088 \quad , \quad \omega_2 = 1.407,$$

and corresponding power shares

$$\alpha_1 = 0.5 \quad , \quad \alpha_2 = 0.5.$$

This design is composed of a low frequency and a high frequency component corresponding to the large and small capacities of the heat accumulating media.

Using the D_s-optimality criterion with the interesting parameter being the small capacitance C_i, the optimal input $u_{o1}(t)$ comprises sinusoids with

$$\omega_{11} = 0.143 \quad , \quad \omega_{21} = 2.644,$$

and corresponding power shares

$$\alpha_{11} = 0.052 \quad , \quad \alpha_{21} = 0.948.$$

Notice the considerable shift towards high frequencies.

The design with C_m as the interesting parameter is the signal $u_{o2}(t)$ with frequencies

$$\omega_{12} = 0.085 \quad , \quad \omega_{22} = 0.755,$$

and corresponding power shares

$$\alpha_{12} = 0.812 \quad , \quad \alpha_{22} = 0.188.$$

Notice the considerable shift towards low frequencies.

We then apply these input signals and estimate the parameters 100 times (cross-sections) for each input. The results are gathered in Table 1 where the numbers are obtained by averaging over the relevant values for the 100 cross-sections. While $u_o(t)$ is the D-optimal signal for estimating all the parameters (minimizes the determinant of the covariance matrix of the estimate), the other D_s-optimal inputs provide most information about the interesting parameters, C_i and C_m, respectively. The estimation for each cross-section is based on 5000 samples where the sampling time is $\Delta T = 0.05$.

5 Solution for Bayesian Criteria

As discussed earlier, the optimization criterion for input design is in general a function of the unknown parameter value. In this section, we consider optimization of the Bayesian criteria, i.e. criteria of the form

$$E_\theta\{\phi(M)\} \tag{5.1}$$

which take care of the fact that the prior knowledge about the parameter value is partial.

	$u_o(t)$	$u_{o1}(t)$	$u_{o2}(t)$
$\log\det(\cdot)$	*-33.13*	*-32.92*	*-32.85*
$\log\sigma^2_{\hat{C}_i}$	*-6.17*	*-7.42*	*-4.98*
$\log\sigma^2_{\hat{C}_m}$	*-0.11*	*0.80*	*-0.86*

Table 1: *Comparison between D-optimal and D_s-optimal designs. In the table,* $\det(\cdot)$ *denotes the determinant of the covariance of the estimate* $\hat{\theta}_p$, $\sigma^2_{\hat{C}_i}$ *the variance of the estimate* $\hat{C}_i$, *and* $\sigma^2_{\hat{C}_m}$ *the variance of the estimate* $\hat{C}_m$.

In this connection, it is of interest to study if the theory of the previous sections can be fully or partially employed. Especially, it is of interest to investigate whether or not:

- the Bayesian criteria admit the use of standard experimental design algorithms,
- the upper bound on the number of sinusoidal components that suffice for optimality (Theorem 3) still exists.

Attention will be restricted to discrete-time systems. The extension for continuous-time systems may be accomplished in the same way as in Section 4. Focus on the first question.

Theorem 5 Fix the input frequency set

$$\Omega = \{\omega_i | i = 1, ..., l, \omega_i \in [0, \pi/\Delta T]\}.$$

Define

$$f(\alpha) = E_\theta\{\phi(M(\xi))\}$$

where $\phi(\cdot)$ is convex and $\alpha = (\alpha_1, ..., \alpha_l)^\top$ gives the power share of the power restricted input measure $\xi \in \mathcal{X}$ on Ω. An input measure ξ^* characterized by α^* minimizes the Bayesian criterion in equation (5.1) among all other power restricted input measures whose frequencies are contained in Ω if and only if

$$\lim_{\beta \to 0+} \frac{\partial}{\partial \beta} f((1-\beta)\alpha^* + \beta\alpha^{(i)}) \geq 0$$

where $\alpha^{(i)}$ is a unit vector with 1 in the entry i.

PROOF: The key observation is that when $\phi(\cdot)$ is a convex function of M then $f(\cdot)$ is a convex function of α. To see why, take any two power shares α and α' and denote their corresponding measures by ξ and ξ'. Then for any $0 < \beta < 1$

$$\begin{aligned} f((1-\beta)\alpha + \beta\alpha') &= E_\theta\{\phi((1-\beta)M(\xi) + \beta M(\xi'))\} \\ &\leq (1-\beta)E_\theta\{\phi(M(\xi))\} + \beta E_\theta\{\phi(M(\xi'))\} \\ &= (1-\beta)f(\alpha) + \beta f(\alpha'). \end{aligned}$$

As a result of input power restriction

$$\sum_{i=1}^{l} \alpha_i = 1, \quad \alpha_1, ..., \alpha_l \geq 0,$$

the set of all feasible α's denoted by $\mathcal{A}$ is a convex set. $\mathcal{A}$ is in fact the convex hull of the points $\alpha^{(i)}$, $i = 1, ..., l$. The rest of the proof follows in a way similar to the proof of Theorem 4 after replacing $\mathcal{M}$ by $\mathcal{A}$ and $\phi : \mathcal{M} \to \mathbb{R}$ by $f : \mathcal{A} \to \mathbb{R}$.

Remark 2 Here, we use Theorem 5 to derive an algorithm for constructing optimal input measure for the criterion $E_\theta\{-\log\det(M)\}$. In fact, we wish to find the optimal power weight α^* on a *given* set of frequencies Ω. This set can for example be a fine grid on the frequency axis. Similar to Remark 1

$$\lim_{\beta\to 0+}\frac{\partial}{\partial\beta}f((1-\beta)\alpha^* + \beta\alpha^{(i)}) =$$
$$\lim_{\beta\to 0+}\frac{\partial}{\partial\beta}E_\theta\{-\log\det[(1-\beta)M(\xi^*) + \beta\tilde{M}(\omega_i)]\} =$$
$$p - E_\theta\{\mathrm{trace}([M(\xi^*)^{-1}\tilde{M}(\omega_i)])\}$$

where ξ^* denotes the input measure characterized by α^*. Then, ξ^* is optimum if and only if

$$v_B(\xi^*,\omega_i) = \mathrm{trace}(E_\theta\{[M(\xi^*)^{-1}\tilde{M}(\omega_i)]\}) \le p.$$

Thus, start with an initial power share on Ω and update exactly as in Remark 1 noting that $\omega_k^\circ = \arg\max_{\omega\in\Omega} v_B(\xi_k,\omega)$. The proof of convergence is analogous to the case described in Remark 1, see Gaffke and Mathar (1992). Again the generic technique can be applied to optimization of the criteria of equation (5.1) for all differentiable convex ϕ functions.

Remark 3 In Theorem 5 and Remark 2, the frequency set Ω is fixed. They state a necessary and sufficient condition for the optimal power weights on Ω and an algorithmic search to find these weights. However, Theorem 4 and Remark 1 required no fixed Ω. What was fixed was actually the convex set $\mathcal{M}$ where the ϕ-function was defined and that was exactly the convexity of $\mathcal{M}$ which led to the theorem on the upper bound on the number of sinusoidal components (Theorem 3). When solving for Bayesian criteria, this convex set is no longer

fixed since each point on the boundary of $\mathcal{M}$ is a function of the parameter and the parameter is not fixed any longer. Hence, the answer to the second question, in general nonlinear regression cases, is unfortunately negative. The reason is that no fixed convex hull of information matrices corresponding to single frequency design measures exist. However, the numerical search algorithms in many cases yield measures which essentially comprise finitely many sinusoidal components, see the examples below.

Example 6 Assume that the following system is given

$$y_t = \frac{bq^{-1}}{1 + aq^{-1}} u_t + \epsilon_t$$

where the unknown parameter is $\theta = (a, b)^\top$ and the covariance of the noise sequence is known.

Assume that the prior information about a is embedded in a Gaussian distribution with mean $m_a = 0.8$ and variance $\sigma_a^2 = (0.08)^2$. Fix $b = 1$ and use the algorithm given in Remark 2 to find the D-optimal power weights on a fine grid on the ω axis denoted by Ω. The set Ω is obtained by dividing [2.95,3.19] into 20 intervals of equal length. This procedure yields the design ξ^*, essentially comprising frequencies

$$\omega_1 = 3, \quad \omega_2 = 3.01, \quad \omega_3 = 3.14,$$

and power shares

$$\alpha_1 = 0.1079, \quad \alpha_2 = 0.7377, \quad \alpha_3 = 0.1544,$$

respectively.

To confirm that this design is indeed optimal, $v_B(\xi^*, \omega)$ is plotted as a function of $\omega \in \Omega$ in Figure 1. Since $v_B(\xi^*, \omega) \leq p$, Theorem 5 suggests that ξ^* is optimal.

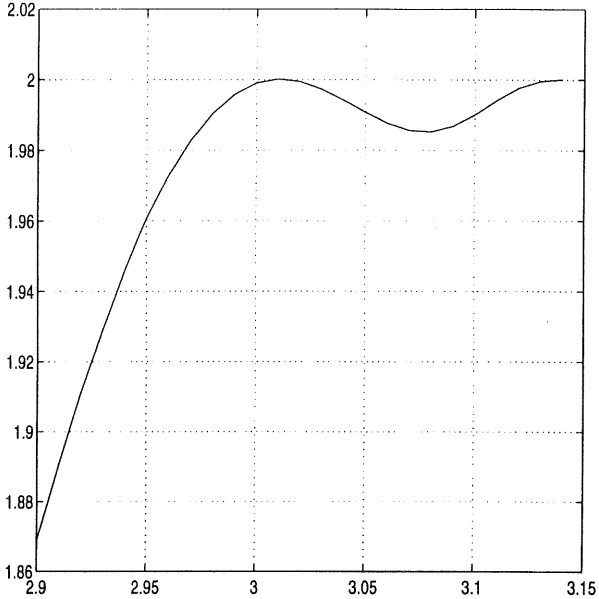


Figure 1: $v_B(\xi^*, \omega)$ *as a function of* $\omega \in \Omega$ *in Example 5.2.*

If also a is fixed in the design computation at $a = 0.8$, the design measure consists of a single frequency component $\omega^* = 3.01$.

Example 7 Now consider a second order continuous-time model with $G_1(s) = 1/[(s+a)(s+b)]$ and $G_2(s) = 1$. The unknown parameter is $\theta = (a, b)^\top$ and the noise is wide band white with known covariance. Fix $b = 2$ and assume that the prior information about a is embedded in a Gaussian distribution with mean $m_a = 1$ and variance $\sigma_a^2 = 0.4^2$. Similar to the previous example, we find the D-optimal measure ξ^* with

$$\omega_1 = 0, \quad \omega_2 = 0.43,$$

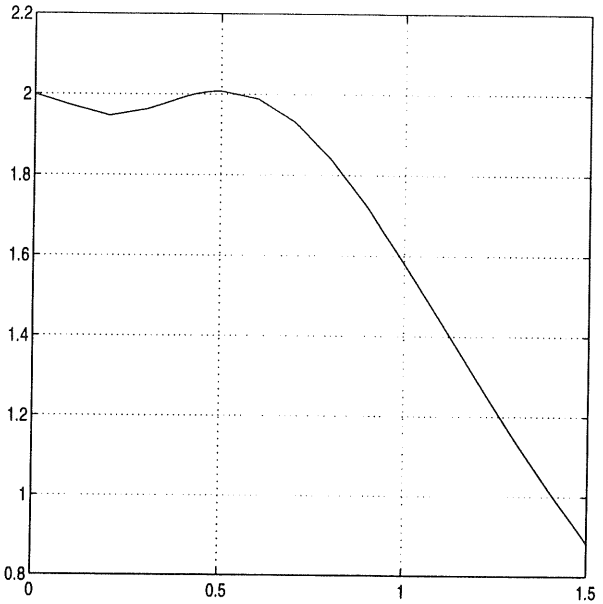


Figure 2: $v_B(\xi^*, \omega)$ *as a function of* $\omega \in \Omega$ *in Example 5.3.*

and

$$\alpha_1 = 0.0218, \quad \alpha_2 = 0.9782.$$

The frequency grid Ω is obtained by dividing [0,1.5] to intervals of length 0.1 and 0.01 adapted to the area around 0.5.

A plot of $v_B(\xi^*, \omega)$ as a function of $\omega \in \Omega$ is given in Figure 2 which confirms that ξ^* is indeed optimal ($v_B(\xi^*, \omega) \leq p$).

Fixing $a = 1$ yields a single frequency design measure with $\omega^* = 0.49$.

6 Conclusion

The design of optimal input signals for identification of gray-box models has been studied. Physical significance and partial prior information are identified as the key features of gray-box models. Considering them as part of the experimental design procedure necessitates extensions of the design theory, which are studied. Probability distributions are used as means for expressing partial prior information.

The connection between design for a model with some input-output parametrization and a physical model is conceptually as simple as computing the information matrix for the reparametrized model.

When a Bayesian criterion is considered as a means of taking the whole prior distribution of the parameter into account, the convex theory of experimental design which leads to the theorem on finite number of support points for the design measure can not be applied directly. The design problem is then solved over a finite grid of frequencies. The constructive, simple experimental design algorithms for the non-Bayesian criterion can be generalized to this class of problems.

Examples are included throughout the text to illustrate the concepts.

Acknowledgement

We are grateful to the reviewers for providing very thorough and inspiring discussions of a previous version of this paper.

References

Blanke, M. and T. Söderström (Eds.) (1994). *Preprints of the 10th IFAC Symposium on System Identification, Copenhagen, Denmark.*

Bohlin, T. (1984). Computer-aided grey-box validation. Technical Report TRITA-REG-8403, Department of Automatic Control, Royal Institute of Technology, Stockholm, Sweden.

Bohlin, T. (1989). The fundamentals of modelling and identification. Technical Report TRITA-REG-89/00002, Department of Automatic Control, Royal Institute of Technology, Stockholm, Sweden.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Academic, New York.

Gaffke, N. and R. Mathar (1992). On a class of algorithms from experimental design theorem. *Optimization 24*, 91–126.

Goodwin, G. C. and R. L. Payne (1977). *Dynamic System Identification: Experiment Design and Data Analysis.* Academic, New York.

Madsen, H. and J. Holst (1995). Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings 22*, 67–79.

Pilz, J. (1991). *Bayesian Estimation and Experiment Design in Linear Regression Models.* John Wiley and Sons, New York.

Rockafellar, R. (1970). *Convex Analysis.* Princeton University Press.

Silvey, S. D. (1980). *Optimal Design.* Chapman & Hall, London.

Whittle, P. (1971). *Optimization Under Constraints.* Wiley-Interscience, New York.

Maxmin Input Design for Linear Dynamic Systems

Payman Sadegh* **Lars H. Hansen*°** **Henrik Madsen*** and **Jan Holst+**

*Institute of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

+ Institute of Mathematical Statistics, Lund Institute of Technology, Lund S-22100, Sweden.

° Technology R&D, Grundfos A/S, DK-8850 Bjerringbro, Denmark.

To be submitted.

[A3]

Abstract

This paper considers the problem of input design for maximizing the smallest eigenvalue of the information matrix for linear dynamic systems. The optimization of the smallest eigenvalue is of interest in parameter estimation and parameter change detection problems. We describe a simple cutting plane algorithm to determine the optimal frequency power weights of the input, using successive solutions to linear programs. We present a case study related to estimation of thermal parameters of a building.

Key words: Maxmin optimization, Information matrix, Input design, Linear dynamic systems, Linear programming, Cutting plane, Domestic heating.

1 Introduction

Consider the problem of estimating a set of unknown parameters or detecting parametric changes in a dynamic system based on experimental input-output data. The accuracy of the estimates or the detectability of changes are often dependent upon the experimental conditions under which the data are collected. We regard the output data as a realization of some random process which is obviously affected by the controlled input to the system. We wish to determine the input sequence so as to maximize the amount of useful information in the output data. Similar to the usual approach of the statistical experiment design literature (see e.g. Silvey (1980)), we use Fisher's information matrix as a measure of quantifying the amount of information in data. Silvey (1980) argues that it is not possible to design an experiment to maximize the information matrix in a strong (matrix inequality) sense. Therefore, we consider instead the optimization of some suitable *scalar* function of the information matrix, see e.g. Silvey (1980) and Fedorov (1972) for a discussion of the widely used and statistically meaningful criteria. A particularly useful and important choice is the smallest eigenvalue of the information matrix or the so called E-optimality criterion (see e.g. Pazman (1986)). We shall discuss the role of the E-optimality criterion (maxmin design criterion) later in connection with parameter estimation and change detection problems.

The problem of input design has been extensively studied in the literature using different approaches. The statistical approach of the present work is similar to Goodwin and Payne (1977). Other selected references are Gevers and Ljung (1986), Zhang (1989), Tulleken (1990), Godfrey (1993), Sadegh, Holst, Madsen, and Melgaard (1995), which treat different aspects of the input design problem. The present

paper is distinguished by the fact that the considered optimality criterion is nonsmooth and special optimization techniques should be employed. The paper indeed shows that the maxmin input design can be addressed within the setting of another extensively studied problem that is maximizing (minimizing) the smallest (largest) eigenvalue of a linear combination of given symmetric matrices (see e.g. Boyd and Yang (1989), Fan and Nekooie (1992), Jarre (1993)). We discuss in some detail a cutting plane algorithm (see Kelley (1960), Zangwill (1969)) for the optimization of the criterion. The algorithm is both efficient and relatively simple, requiring only successive solutions to linear programs.

The rest of the paper is organized as follows. In Section 2, we state the problem formulation. Section 3 presents the solution. In Section 4, we study the design of optimal inputs for estimating thermal parameters of a building. Finally, Section 5 offers concluding remarks.

2 Problem Formulation

Consider a random variable y with the probability density function $h(y|\theta)$ where $\theta \in \mathbb{R}^p$ is a (p-dimensional) parameter. We introduce

Definition 1 Fisher's information matrix for the random variable y is defined by

$$M_F = E\{(\frac{\partial h(y|\theta)}{\partial \theta})^\top (\frac{\partial h(y|\theta)}{\partial \theta})\} \tag{2.1}$$

where $E\{\cdot\}$ denotes mean value. For a total number of samples N_T, the (average) information matrix per sample is defined by $M = \lim_{N_T \to \infty} M_F/N_T$.

We now consider the linear dynamic system given by

$$y_t = G_1(q^{-1})u_t + G_2(q^{-1})\epsilon_t, \quad t = ..., -1, 0, 1, ... \tag{2.2}$$

where $\{u_t\}$ and $\{y_t\}$ are the input and output sequences respectively, $\{\epsilon_t\}$ is a sequence of Gaussian i.i.d. random variables which without loss of generality can be assumed to have unit covariance, and G_1 and G_2 are transfer functions in the backward shift operator q^{-1}. The transfer functions G_1 and G_2 depend upon the parameter θ.

We shall be concerned with the problem of maximizing $\lambda_{\min}(M)$ with respect to the input sequence, where $\lambda_{\min}(\cdot)$ denotes smallest eigenvalue, and M is the information matrix per sample for the output data of the system (2.2). Note that all the eigenvalues are real since the information matrix is a real symmetric matrix (the information matrix is moreover nonnegative definite). The stated maxmin problem is of interest in a variety of areas such as parameter estimation and change detection.

Parameter Estimation: Consider the system (2.2). Let in the following $\hat{\theta}_{N_T}$ denote the maximum likelihood estimate of θ based on N_T observations, and $||\cdot||$ denote the usual Euclidean norm. It is well-known that under mild stationarity and regularity conditions (see e.g. Goodwin and Payne (1977))

$$\sqrt{N_T}(\hat{\theta}_{N_T} - \theta) \stackrel{\text{dist}}{\Rightarrow} \Delta\theta \sim N(0, M^{-1}). \tag{2.3}$$

We have that $\Delta\theta^\top M \Delta\theta \sim \chi^2_p$, and hence $Pr(\Delta\theta^\top M \Delta\theta \leq \chi^2_{1-\delta;p}) = 1-\delta$, $0 < \delta < 1$, where $\chi^2_{1-\delta;p}$ is the $1-\delta$ fractile of a χ^2-distribution with p degrees of freedom. Obviously

$$\Delta\theta^\top M \Delta\theta \geq \lambda_{\min}(M)||\Delta\theta||^2. \tag{2.4}$$

The equality in (2.4) is reached for $\Delta\theta$ being along any eigenvector corresponding to $\lambda_{\min}(M)$. It then follows that for any $0 \leq \delta \leq 1$, we

have $||\Delta\theta||^2 \leq \chi^2_{1-\delta;p}/\lambda_{\min}(M)$ with probability $1-\delta$. Upon maximizing $\lambda_{\min}(M)$ with respect to the experiment, we minimize the largest (probabilistic) uncertainty bound on the estimate.

Change Detection: Assume that a change in the system given by (2.2) is characterized by a change in the parameter from θ to $\theta+\Delta\theta_c$. Under general regularity conditions, the quantity $\Delta\theta_c^\top M_F \Delta\theta_c$ for $\Delta\theta_c \to 0$ tends to the divergence between the model under no change hypothesis and the model under the change hypothesis (Kullback (1959)) where M_F is the information matrix for the model given by (2.2). It is then obvious that the average value of the divergence per sample tends to $\Delta\theta_c^\top M \Delta\theta_c$. Divergence is a suitable measure of the detectability of a parametric change, see e.g. Basseville and Nikiforov (1993). We have $\Delta\theta_c^\top M \Delta\theta_c \geq \lambda_{\min}(M)||\Delta\theta_c||^2$ where the equality is reached for $\Delta\theta_c$ being along an eigenvector corresponding to $\lambda_{\min}(M)$. For any fixed change magnitude $||\Delta\theta_c||$, maximizing the smallest eigenvalue with respect to the input sequence is related to maximizing the smallest (with respect to the direction of $\Delta\theta_c$) divergence between the two models.

We introduce the following assumptions:

A1 : the input and the noise sequences are uncorrelated (i.e., the experiments are performed in open loop),

A2 : the input is generated by a finite register with length N, i.e. the input sequence repeats periodically with cycle N,

A3 : the total number of samples N_T is large,

A4 : the input power is constrained.

We further assume that the general regularity and stationarity conditions that ensure the convergence result (2.3) hold. For simplicity, we restrict attention to single input systems.

Consider the system given by (2.2) and denote the one step ahead

prediction error at time t by e_t. Using (2.1), it follows that the information matrix, M_F, for the system is given by (Goodwin and Payne (1977))

$$M_F = \sum_{t=0}^{N_T-1} \mu_t \mu_t^{\top} + M_c, \tag{2.5}$$

where

$$\mu_t = G_2^{-1}(q^{-1})(\frac{\partial G_1(q^{-1})}{\partial \theta})u_t, \tag{2.6}$$

and

$$\begin{aligned} M_c &= E\{\sum_{t=0}^{N_T-1}[-G_2^{-1}(q^{-1})(\frac{\partial G_2(q^{-1})}{\partial \theta})e_t] \\ &\quad [-G_2^{-1}(q^{-1})(\frac{\partial G_2(q^{-1})}{\partial \theta})e_t]^{\top}|\theta\}. \end{aligned}$$

This result is obtained from the definition of the information matrix.

Considerable simplicity is obtained if we represent the input sequence in the frequency domain. The assumption A2 implies that we can represent the input as

$$u_t = c_0 + \sum_{k=2}^{N}(\sqrt{2}c_{k-1})\sin(2\pi(k-1)t/N + \psi_k).$$

Without loss of generality, the input power restriction (see A4) can be expressed as $\sum_{k=1}^{N} c_{k-1}^2 = 1$. Now note that

$$\lim_{N_T\to\infty} \frac{1}{N_T}\sum_{N_T} \sin(2\pi k_1 t/N + \psi)\sin(2\pi k_2(t-T)/N + \psi') = 0$$

for all integer T, all ψ, ψ', and all $k_1, k_2 \in \{0, 1, ..., N-1\}$, $k_1 \neq k_2$. Denoting the information matrix per sample corresponding to the

input u_t by $M(u_t)$, it follows immediately that $M(\sin(\omega t + \psi)) = M(\sin(\omega t))$ for all ψ and ω. This result together with (2.5) yield

$$M = c_0^2 M(1) + \sum_{k=2}^{N} c_{k-1}^2 M(\sqrt{2}\sin(2\pi(k-1)t/N)). \tag{2.7}$$

Now, define $\alpha_k = c_{k-1}^2$, $k = 1, ..., N$. Then the input power restriction can be written as $\sum_{k=1}^{N} \alpha_k = 1$ and

$$M = \sum_{k=1}^{N} \alpha_k M_k \tag{2.8}$$

where $M_1 = M(1)$, and $M_k = M(\sqrt{2}\sin(2\pi(k-1)t/N)$ for $k \geq 2$. From the input power restriction $\sum_k \alpha_k = 1$ and the equation (2.8), it is evident that the symmetric nonnegative definite information matrix per sample, M, lies in the convex hull of the symmetric nonnegative definite matrices M_k (Goodwin and Payne (1977)).

Remark 1 It follows from (2.5) that the actual values of M_k (and M) are dependent upon the true parameter value θ. However, the true parameter is in general unknown at the experiment design stage, especially when the experiment concerns estimation of the parameters. In this paper, we assume that the M_k matrices are evaluated at an *a priori* value for the parameter, say its prior mean. The sensitivity and the robustness of the design to other parameter values should usually be checked, see Sadegh, Holst, Madsen, and Melgaard (1995) for the design of robust experiments using a Bayesian formulation.

Remark 2 Using a slightly different assumption than A2, we can obtain a result analogous to (2.8). Assume that the input can be represented as the linear combination $u_t = \sum_{k=0}^{N-1} c_k \phi_t^{(k)}$ where the $\phi_t^{(k)}$ are

given functions satisfying $\lim_{N_T\to\infty}\sum_{N_T}\phi_t^{(k_1)}\phi_{t-T}^{(k_2)}/N_T = 0$ for all integer T and all $k_1, k_2 \in \{0, 1, ..., N-1\}, k_1 \neq k_2$. Again, defining $\alpha_k = c_{k-1}^2$, $k = 1, ..., N$, it is straight forward to show that $M = \sum_{k=1}^{N} \alpha_k M_k$ where M_k is the information matrix per sample under application of the input $\phi_t^{(k-1)}$ to the system. The M_k can be easily obtained using e.g. simulations where the simulations involve application of the input $\phi_t^{(k-1)}$ to the system and calculation of the relevant quantities in (2.5). The numerical case study of the paper (Section 4) illustrates such procedures. Assuming that $\lim_{N_T\to\infty}\sum_{t=1}^{N_T}[\phi_t^{(k_1)}]^2/N_T = \lim_{N_T\to\infty}\sum_{t=1}^{N_T}[\phi_t^{(k_2)}]^2/N_T$ for all $k_1, k_2 \in \{0, 1, ..., N-1\}$, the input power constraint can without loss of generality be stated by $\sum_k \alpha_k = 1$, and the input design problem concerns optimal allocation of the input power among the $\phi_t^{(k)}$. Note that A2 simply implies that we can select $\phi_t^{(0)} = 1$, and $\phi_t^{(k)} = \sqrt{2}\sin(2\pi kt/N)$, $k \geq 1$.

Now, denoting $\alpha = (\alpha_1, ..., \alpha_N)$, the maxmin problem can be stated as

$$\max_{\alpha\in\mathcal{A}}\{\lambda_{\min}(\sum_{k=1}^{N}\alpha_k M_k)\} \tag{2.9}$$

where $\mathcal{A} = \{\alpha | \sum_{k=1}^{N} \alpha_k = 1, \alpha_k \geq 0\}$.

The optimization problem (2.9) can be equivalently formulated as a problem with linear objective function as follows. For convenience, we define $f(\alpha) = \lambda_{\min}(\sum_{k=1}^{N} \alpha_k M_k)$. It is also more convenient

to consider the equivalent optimization problem

$$\begin{aligned} &\max\{f(\alpha)\} \\ &\alpha \in \mathcal{A} \\ &f(\alpha) \leq \lambda_{\min}(\sum_{k=1}^{N} \alpha_k M_k). \end{aligned} \tag{2.10}$$

Without loss of generality, we can assume that all the M_k are positive definite implying that $f(\alpha) > 0$ for all $\alpha \in \mathcal{A}$. To ensure the positive definiteness of the M_k, we possibly need to add a constant matrix $\epsilon_0 I$ to each nonnegative definite M_k where I is the unity matrix of proper order and ϵ_0 is some positive number. This modifies the objective function of (2.9) to $\lambda_{\min}(\sum_{k=1}^{N} \alpha_k M_k + \epsilon_0 I)$. Recalling $\lambda_{\min}(M) = \min_{\|w\|=1} w^\top M w$, it follows that the addition of $\epsilon_0 I$ to the M_k merely adds the constant ϵ_0 to the objective function. Assuming the positive definiteness of the M_k, it is then allowable to define the variable $\beta = (\beta_1, ..., \beta_N)$, $\beta = \alpha / f(\alpha)$ since $f(\alpha) \geq \epsilon_0$. Since $\sum_{k=1}^{N} \alpha_k = 1$ for all $\alpha \in \mathcal{A}$, we have that $\sum_{k=1}^{N} \beta_k = 1/f(\alpha)$. Furthermore, it holds that $\lambda_{\min}(\sum_{k=1}^{N} \alpha_k M_k)/f(\alpha) = \lambda_{\min}(\sum_{k=1}^{N} \beta_k M_k)$, and consequently we obtain the equivalent optimization problem

$$\begin{aligned} &\min_{\beta} \sum_{k=1}^{N} \beta_k \\ &f(\beta) = \lambda_{\min}(\sum_{k=1}^{N} \beta_k M_k) \geq 1 \\ &\beta_k \geq 0 \qquad\qquad k = 1, ..., N. \end{aligned} \tag{2.11}$$

Provided that a solution to (2.11) is available, the solution to (2.9) is readily obtained using the simple transformation $\alpha = \beta / \sum_{k=1}^{N} \beta_k$. The

equivalence of (2.9) and (2.11) is quite analogous to the equivalence of the matrix games and linear programs in the game theory (Dantzig (1963)).

3 Optimization Procedure

The constraint function $f(\beta)$ in (2.11) is nondifferentiable at those values of β where the multiplicity of $\sum_{k=1}^{N_c} \beta_k M_k$ is larger than one (similarly, the objective function of (2.9) is in general nondifferentiable). However, it can be readily verified that $f(\beta)$ is concave, i.e. for $0 \leq \gamma \leq 1$ and any β', β'':

$$f(\gamma\beta' + (1-\gamma)\beta'') \geq \gamma f(\beta') + (1-\gamma) f(\beta'').$$

Subdifferentials of a nonsmooth concave function play the same important role as the gradients of a differentiable function. We therefore introduce the following definition.

Definition 2 The subdifferential of a concave function $F(x)$, $x \in \mathbb{R}^n$, is the set of all vectors $z \in \mathbb{R}^n$, such that $F(v) \leq F(x) + z^\top (v - x)$ for all $v \in \mathbb{R}^n$. The subdifferential of F at x is denoted by $\partial F(x)$.

We can compute the subdifferential for the concave function $f(\beta)$ using basic rules of subdifferential calculus, see Rockafellar (1970). Analogous to proposition 2.8.8 of Clarke (1983), we obtain that at a point β where the multiplicity of the smallest eigenvalue of $\sum_{k=1}^{N} \beta_k M_k$ is equal to r, the subdifferential of f is given by:

$$\partial f(\beta) = co\{(w^\top Q(\beta)^\top M_1 Q(\beta) w, ..., w^\top Q(\beta)^\top M_N Q(\beta) w)^\top : w \in \mathcal{S}_r\} \tag{3.1}$$

where each column of the $p \times r$ matrix $Q(\beta)$ is equal to one of the r orthonormal eigenvectors of $\sum_{k=1}^{N} \beta_k M_k$ corresponding to the smallest eigenvalue (recall that $\sum_{k=1}^{N} \beta_k M_k$ is symmetric), $\mathcal{S}_r$ is the r-dimensional unit sphere, and $co\{\cdot\}$ denotes convex hull.

The optimization problem (2.9) can be addressed within the setting of maximizing the smallest eigenvalue of a linear combination of symmetric matrices. Boyd and Yang (1989) review some standard techniques for solving similar nondifferentiable problems. Methods based on a smooth approach to nondifferentiable optimization have been recently reported, see e.g. Shapiro and Fan (1995), Jarre (1993). A particularly simple and efficient method which is suitable for the maxmin optimization of (2.11) is the cutting plane method (Kelley's cutting plane method, see Kelley (1960)). In the following, we describe the method is some detail.

3.1 Cutting Plane Algorithm

Kelley (1960) considers a cutting plane algorithm for optimization problems of the form:

$$\begin{aligned} &\min_{\beta} q^{\top}\beta \\ &g(\beta) \geq 0, \end{aligned} \tag{3.2}$$

with the assumptions that the scalar valued function $g(\beta)$ is real, continuous, concave, and the set $B = \{\beta : g(\beta) \geq 0\}$ is compact. Moreover, the elements of $\partial g(\beta)$ are assumed to be uniformly bounded on some compact polytope containing B. The general form of the cutting plane algorithm is as follows (Luenberger (1984)):

Procedure 1 *Cutting Plane Algorithm:* Select a polytope P_i containing B.

Step 1 : Minimize $q^\top \beta$ over P_i to obtain $\beta^{(i)}$. If $\beta^{(i)} \in B$, then $\beta^{(i)}$ is optimal. Otherwise,

Step 2 : Add the hyperplane $a^{(i)\top}(\beta - \beta^{(i)}) + g(\beta^{(i)}) \geq 0$ where $a^{(i)}$ is any element of $\partial g(\beta^{(i)})$ to obtain a new polytope (update P_i) and go to Step 1.

Recalling that $\sum_{k=1}^{N} \beta_k \leq 1/\epsilon_0$ (ϵ_0 is a number such that $f(\alpha) \geq \epsilon_0$), the optimization problem (2.11) can be reformulated as (3.2) by letting $q = (1, ..., 1)^\top$, and

$$g(\beta) = \min(1/\epsilon_0 - \sum_{k=1}^{N} \beta_k, f(\beta) - 1, \beta_1, ..., \beta_N).$$

It is easy to check that $g(\beta)$, as defined above, is continuous and concave, and the restriction defined by $g(\beta) \geq 0$ is compact. As the start polytope for solving (2.11) using the cutting plane algorithm, we select a polytope defined by the restrictions

$$1/\epsilon_0 - \sum_{k=1}^{N} \beta_k, \beta_1, ..., \beta_N \geq 0.$$

At any iteration i, it holds that either the algorithm stops or $f(\beta^{(i)}) < 1$, while $1/\epsilon_0 - \sum_{k=1}^{N} \beta_k^{(i)}, \beta_1^{(i)}, ..., \beta_N^{(i)} \geq 0$. This implies that at all the iterations where the algorithm does not stop $g(\beta^{(i)}) = f(\beta^{(i)}) - 1$. Since the convergence proof of Kelley's cutting plane algorithm is based on the asymptotic behavior of a limit sequence of $g(\beta^{(i)})$ (see Kelley (1960)), the hyperplanes for optimization of (2.11) can be selected as $a^{(i)} \in \partial f(\beta^{(i)})$ (notice that $\partial(f(\beta) - 1) = \partial f(\beta)$). Since,

the elements of $\partial f(\beta)$ are uniformly bounded on any compact set, the assumptions for applying Kelley's cutting plane algorithm to the optimization problem (2.11) hold. Furthermore, it can be readily verified that for any $a^{(i)} \in \partial f(\beta^{(i)})$, we have $a^{(i)\top}\beta^{(i)} = f(\beta^{(i)})$. Therefore, the hyperplanes at Step 2 of Procedure (1) are selected as $a^{(i)\top}\beta \geq 1$.

Notice that each iteration of the algorithm requires solution to a linear program. The draw back of the method is that the number of constraints of the linear program at each iteration grows with the number of iterations. Simple devices may be used to circumvent this problem, e.g. by deleting the inactive constraints at the end of each iteration (see Luenberger (1984)). Different numerical experimentations indicate the efficiency of the algorithm for the maxmin problem of interest. For a detailed treatment of the cutting plane algorithm where convergence is established under more general conditions, see Zangwill (1969), Chapter 14.

4 Case Study: Domestic Heating of a House

An exemplification of the described theory is given in this case study, which is concerned with the domestic heating of a house. This case study is inspired by a low energy test house at the Department of Buildings and Energy, the Technical University of Denmark. A water based central heating system is used as the domestic heating system.

The low energy house and the central heating system are modelled and implemented in Matlab®. The goal is to find an optimal sequence of pump pressures in order to obtain accurate estimates of some thermal capacities in the house, using (indoor) room temperature measurements.

4.1 The Model

This subsection presents the model for the heat transfer dynamics. The house has a ground floor of approximately 120 m^2, and a wooden outer wall which is insulated with 300 mm mineral wool. The power needed to maintain 20 °C at an ambient temperature of -12 °C is about 2.5 kW. For details, see e.g. Rasmussen and Saxhof (1982), Madsen, Melgaard, and Holst (1990), and Madsen, Nielsen, and Saxhof (1992). The house contains two separate rooms A and B each of 60 m^2.

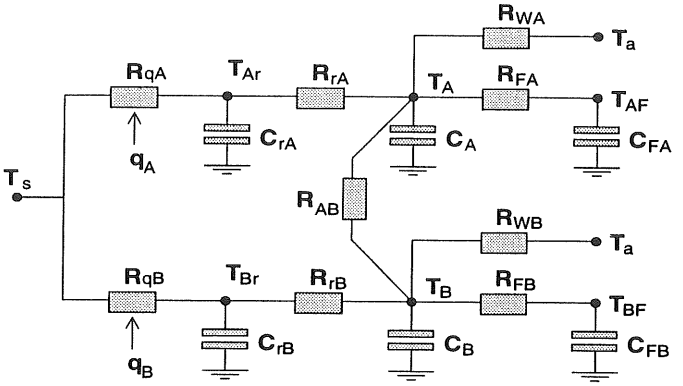


Figure 1: *Thermal network equivalent model of the house.*

The modeling objective here is to obtain accurate estimates of the parameters that are related to dominant time constants of the system. Therefore, lumped modeling of the heat transfer will be appropriate, provided that certain conditions hold (see Hansen (1996)). Based on a second order lumped model for the heat transfer in each room, we obtain the thermal network model illustrated in Figure 1.

In Figure 1, R and C generically denote thermal resistance and thermal heat capacity, respectively. The indices A and B refer to the

rooms A and B, and the indices F, W, r and q refer to the floor, the outer wall, the radiator, and the flow in the radiator, respectively. As indicated in Figure 1, R_{qA} and R_{qB} are dependent upon the actual flows. This makes the model nonlinear in q_A and q_B. Finally, T_s denotes the temperature of the supply water from the boiler and T_a denotes the ambient temperature. The outputs (measurements) are the two room temperatures T_A and T_B. The measurements are taken in the presence of mutually uncorrelated i.i.d. Gaussian noise with unit covariance.

Based on Figure 1, the following coupled first order differential equations for the room A can be derived

$$\begin{aligned} C_{rA}\frac{dT_{Ar}}{dt} &= \frac{1}{R_{qA}}(T_s - T_{Ar}) + \frac{1}{R_{rA}}(T_A - T_{Ar}) \\ C_{FA}\frac{dT_{AF}}{dt} &= \frac{1}{R_{FA}}(T_A - T_{AF}) \\ C_A\frac{dT_A}{dt} &= \frac{1}{R_{rA}}(T_{Ar} - T_A) + \frac{1}{R_{FA}}(T_{AF} - T_A) + \frac{1}{R_{WA}}(T_a - T_A) \\ &+ \frac{1}{R_{AB}}(T_B - T_A) \end{aligned} \tag{4.1}$$

The relationship between the resistance R_{qA} and the flow q_A is given by $R_{qA} = 1/c_p\rho q_A$ where c_p and ρ denote the specific heat capacity and density of water, respectively. Identical equations hold for the room B. The total hydraulic flow to the radiator, q, is obviously the sum of q_A and q_B. The relationship between q_A and q (q_B and q) is in general nonlinear. However, assuming small flow perturbations for q_A and q_B around some nominal values allows the linearizations $\Delta q_A = k_A \Delta q$ and $\Delta q_B = k_B \Delta q$, where Δq_A, Δq_B, and Δq denote perturbations around the nominal values of q_A, q_B, and q respectively, and $k_A + k_B = 1$. The small perturbation assumption also allows linearization of the system of equations (4.1)

and the similar equations corresponding to the room B. We therefore obtain a total linear model from the pump pressure perturbations to the indoor temperatures. The smallest time constant for the total linearized model is approximately 4 minutes (see Appendix for numerical values) which allows a sampling time of 1 minute.

4.2 Optimal Design of Inputs

The pump pressure around the nominal value (used for linearization) is the designed input to the system. We denote the designed input by Δp_t. The input power restriction is a realistic constraint in this case study, implying restricted pump power.

A complete design of inputs should be based on including all the unknown physical constants in the parameter vector θ. However, the physical knowledge of the system confirmed by numerical experimentation shows that the worst estimable parameters are related to the slow dynamics of the system (due to the large floor capacities). Therefore, we select $\theta = (C_{FA}, C_{FB})^\top$ as the parameter vector.

In order to design optimal inputs, it is required to specify the functions $\phi_t^{(k)}$ and compute the corresponding matrices M_k (see Remark 2). We select $\phi_t^{(k-1)}$ as $25c_k \sin(\omega_k t)$, i.e. the input is represented as

$$\Delta p_t = 25 \sum_k c_k \sin(\omega_k t) \quad [\text{mBar}] \tag{4.2}$$

where $\sum c_k^2 = 1$, and the frequencies ω_k, $k \in \{1, \ldots, 20\}$, are selected as $\omega_k = \frac{2\pi}{60 \times 50 \times k}$. The reason for this selection is that each sinusoid $\sin(\omega_k t)$ has a period of $50 \times k$ hours, and the largest time constant of the system ($\simeq 180$ hours) is included within the time range [50,1000] hours. The above values for the frequencies ω_k are used throughout

the case study. The M_k matrices are computed by numerical differentiation (with respect to θ) of the simulated noise free output, under the application of the input $25\sin(\omega_k t)$, see (2.6) and note that in this example $G_2(q^{-1}) = I$.

4.3 Simulation results

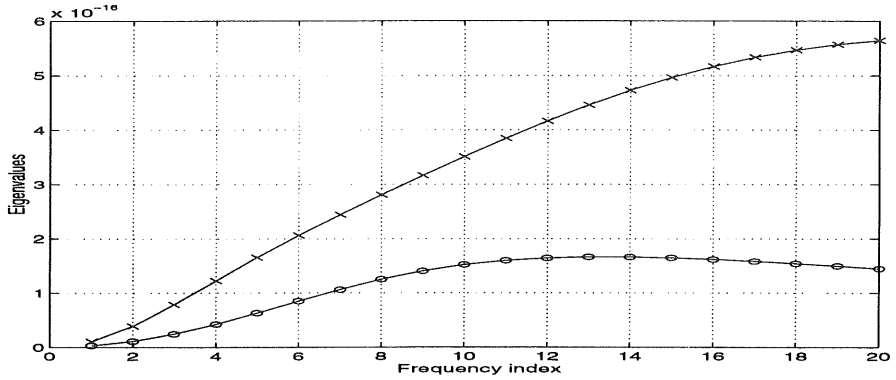


Figure 2: *Eigenvalues of M_k versus k.*

In Figure 2, the two eigenvalues of M_k versus the (frequency) index k are shown and at $k \simeq 13$ the smallest eigenvalue has a maximum with multiplicity one.

The behavior of the cutting plane algorithm is illustrated in Figure 3. The top figure shows $f(\beta^{(i)})$ as a function of iteration number i, while the bottom figure shows $\lambda_{\min}(\sum_k \alpha_k^{(i)} M_k)$ as a function of i. Notice that the value of $f(\beta)$ can be used as a stop criterion for the algorithm (convergence follows if $f(\beta) \geq 1$). The optimal input is computed to be

$$\Delta p_t = 25\left(0.74\sin\left(\frac{2\pi t}{60\times 50\times 13}\right) + 0.67\sin\left(\frac{2\pi t}{60\times 50\times 14}\right)\right).$$

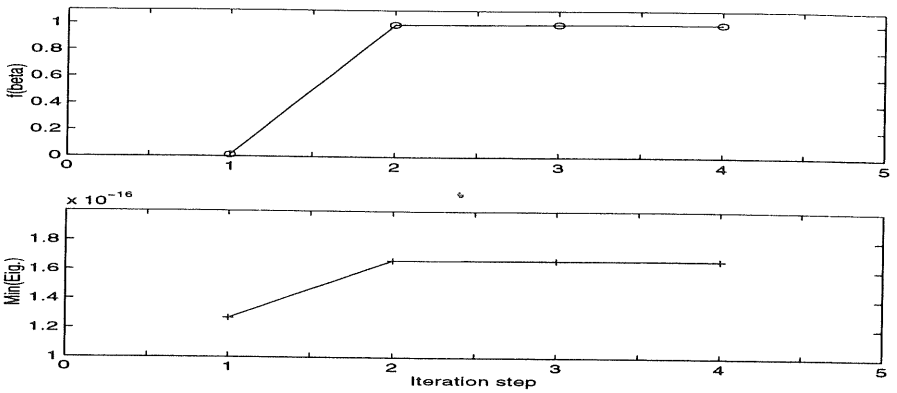


Figure 3: *Convergence of the cutting plane algorithm.*

The result is in agreement with the optimal input suggested by Figure 2.

Let us now examine the case where R_{AB} tends to infinity which means that the two rooms become thermally independent. The eigenvalues of M_k are plotted as a function of k in Figure 4. In this case the maximum of the smallest eigenvalue has multiplicity 2 at $k \simeq 16$.

Figure 5 illustrates $f(\beta^{(i)})$ (top figure) and $\lambda_{\min}(\sum_k \alpha_k^{(i)} M_k)$ (bottom figure) as a function of i. The optimal solution given by the cutting plane algorithm is

$$\Delta p_t = 25\left(0.89\sin\left(\frac{2\pi t}{60\times 50\times 16}\right) + 0.45\sin\left(\frac{2\pi t}{60\times 50\times 17}\right)\right)$$

which is in agreement with Figure 4.

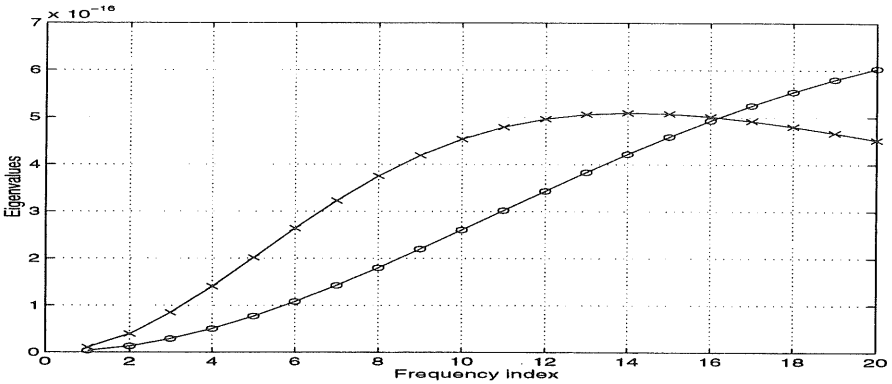


Figure 4: *Eigenvalues of M_k versus k.*

5 Concluding Remarks

We have studied the problem of input design for maximizing the smallest eigenvalue of the information matrix. It is established that the design problem can be addressed within the setting of maximizing the smallest eigenvalue of a linear (indeed convex) combination of given symmetric (nonnegative definite) matrices. We have presented a cutting plane algorithm for the optimization, requiring only successive solutions to linear programs. Numerical experience indicates the efficiency of the algorithm for input design problem. The method is illustrated by a case study related to domestic heating of a house.

Appendix

Here we list the thermal data of the test house which are used throughout the case study.

From measurements on the central heating system it is known

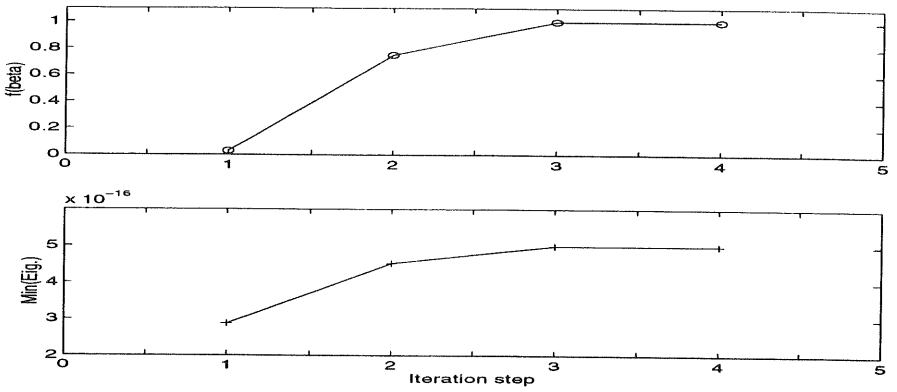


Figure 5: *Convergence of the cutting plane algorithm.*

that a pump pressure of 0.1 Bar yields $q_0 = 28.7$ l/h. These values are used as nominal values. At $q_0 = 28.7$ l/h, the flow fractions are given by $k_A = 0.25$ and $k_B = 0.75$.

Parameter	value	unit
C_{rA}, C_{rB}	6.9	kJ/K
C_A, C_B	158	kJ/K
C_{FA}	11.6	MJ/K
C_{FB}	5.80	MJ/K
R_{rA}, R_{rB}	0.333	K/W
R_{AB}	0.15	K/W
R_{WA}, R_{WB}	0.186	K/W
R_{FA}, R_{FB}	5.56	mK/W
c_p	4.2	kJ/(kg K)
ρ	992	kg/m^3

References

Basseville, M. and I. V. Nikiforov (1993). *Detection of Abrupt Changes.* Prentice-Hall, Englewood Cliffs, New Jersey.

Boyd, S. and Q. Yang (1989). Structured and simultaneous Lyapunov functions for stability problems. *International Journal of Control 49*, 2215–2240.

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis.* Canadian Mathematical Society Series of Monographs and Advanced Texts. John Wiley and Sons, New York.

Dantzig, G. B. (1963). *Linear Programming and Extensions.* Princeton University Press.

Fan, M. K. H. and B. Nekooie (1992). On minimizing the largest eigenvalue of a symmetric matrix. *Linear Algebra and its Applications 214*, 225.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Academic, New York.

Gevers, M. and L. Ljung (1986). Optimal experiment design with respect to the intended model application. *Automatica 22*(5), 543–554.

Godfrey, K. R. (1993). *Perturbation Signals for System Identification.* Prentice Hall, Englewood Cliffs, New Jersey.

Goodwin, G. C. and R. L. Payne (1977). *Dynamic System Identification: Experiment Design and Data Analysis.* Academic, New York.

Hansen, L. H. (1996). Preliminary analysis of a low energy test house. IMM and Grundfos A/S. To appear as an IMM technical report.

Jarre, F. (1993). An interior point method for minimizing the maximum eigenvalue of a linear combination of matrices. *SIAM*

Journal on Control and Optimization 31(5), 1360–1377.

Kelley, J. E. (1960). The cutting plane method for solving convex problems. *J. Soc. Indus. Appl. Math. VIII*(4), 703–712.

Kullback, S. (1959). *Information Theory and Statistics.* John Wiley & Sons, New York.

Luenberger, D. (1984). *Linear and Nonlinear Programming.* 2nd Edition, Addison Wesley, Reading.

Madsen, H., H. Melgaard, and J. Holst (1990). Identification of building performance parameters. In J. Bloem (Ed.), *Workshop on Advanced Identification Tools in Solar Energy Research*, Non Nuclear Energy, pp. 37–60. Commission of the European Communities, DG XII.

Madsen, H., A. A. Nielsen, and B. Saxhof (1992). Identification of models for the heat dynamics of building. Institute of Mathematical Modeling, DTU.

Pazman, A. (1986). *Foundations of Optimum Experimental Design.* D. Reidel Publishing Company, Dordrecht.

Rasmussen, N. H. and B. Saxhof (1982). Simultaneous testing of heating systems. Technical Report 128, Thermal Insulation Laboratory, The Technical University of Denmark.

Rockafellar, R. (1970). *Convex Analysis.* Princeton University Press.

Sadegh, P., J. Holst, H. Madsen, and H. Melgaard (1995). Experiment design for grey-box identification. *International Journal of Adaptive Control and Signal Processing 9*(6), 491–507. See article A2 for a revised version.

Shapiro, A. and M. K. H. Fan (1995). On eigenvalue optimization. *SIAM Journal on Optimization 5*(3), 552–569.

Silvey, S. D. (1980). *Optimal Design.* Chapman & Hall, London.

Tulleken, H. J. A. F. (1990). Generalized binary noise test-signal concept for improved identification-experiment design. *Automatica 26*(1), 37–49.

Zangwill, W. I. (1969). *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey.

Zhang, X. J. (1989). *Auxiliary Signal Design in Fault Detection and Diagnosis*, Volume 134 of *Lecture Notes in Control and Information Sciences.* Springer-Verlag, Berlin.

An Overview of Stochastic Approximation

Payman Sadegh

Institute of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

[B1]

Abstract

In this brief overview, we give a general introduction to stochastic approximation. We present three variants of the Kiefer-Wolfowitz stochastic approximation algorithms, namely, finite difference stochastic approximation (FDSA), random perturbation stochastic approximation (RDSA), and simultaneous perturbation stochastic approximation (SPSA).

Key words: Stochastic approximation, Kiefer-Wolfowitz algorithms, FDSA, RDSA, SPSA.

1 Introduction

In the usual numerical problems of nonlinear programming and optimization theory (see e.g. Zangwill (1969)), one is given a scalar function, say $L(\theta)$, that explicitly relates the p-dimensional design parameter θ to the performance criterion or design objective of interest. Moreover, one is given a set of functions $q_i(\theta)$ that determine restrictions on the design variables through $q_i(\theta) \leq 0$. However, in many optimization problems, the explicit form for $L(\theta)$ (and/or the restriction functions $q_i(\theta)$) is not known, or the available information about the functional relationship of interest is not sufficient to *readily* offer a value for these functions at a given θ.

Consider an unconstrained optimization problem and suppose it is of interest to obtain a solution to the gradient equation $g(\theta) \equiv \partial L/\partial\theta = 0$. Assume that it is possible to do experimentations with the system and for each parameter value, depending on the quantities available for measurement, record either the gradient or the value of the objective function at that parameter value. Unfortunately, noise is an inherent part of all measurement systems. Stochastic approximation (SA) deals with techniques that incorporate the noisy observed values of the gradient or the objective function within iterative procedures to estimate the optimum, see e.g. Kushner and Clark (1978), Ljung (1978), Ljung (1977), Ljung, Pflug, and Walk (1992), among many other references. Stochastic approximation techniques that deal with noisy gradient measurements, are referred to as Robbins-Monro processes. The procedures concerned with the observed values of the objective function are referred to as Kiefer-Wolfowitz algorithms.

Robbins-Monro processes are most easily understood if one considers the problem of finding a zero of some (unknown) function of a

design parameter where the function can be observed in the presence of noise. For example, assume that an engineer is interested in finding the flow rate of cooling water (the design parameter θ) in a chemical reaction process so as the production of a scalar output, say $f(\theta)$, attains a target value a (this example is taken from Kushner and Clark (1978), page 4). Let us assume that $f(\theta)$ is a non-decreasing differentiable function. We consider the situation that the explicit relation between the output product and water flow is unknown. However, the engineer can run an experiment with the reactor for different values of θ which during each experiment run, say run k, are kept equal to $\hat{\theta}_k$, and observe a noisy value of the output, say $y_k(\hat{\theta}_k)$, at the end of the experiment run. Denoting the measurement noise by ϵ_k, then $y_k(\hat{\theta}_k) = f(\hat{\theta}_k) + \epsilon_k$. If the function $f(\theta)$ were known, it would be possible to apply the following classical Newton iteration

$$\hat{\theta}_{k+1} = \hat{\theta}_k - [(\frac{\partial f}{\partial \theta})|_{\theta=\hat{\theta}_k}]^{-1}[f(\hat{\theta}_k) - a],$$

to obtain the zero of $f(\theta) - a$. Now that the function is not known, the Robbins-Monro procedure replaces the above iteration by

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k[y(\hat{\theta}_k) - a]$$

where the gain sequence $\{a_k\}$ satisfies $\sum_{k=1}^{\infty} a_k = \infty$, and $a_k \to 0$ for $k \to \infty$. The condition $a_k \to 0$ asymptotically cancels the noise effect, and $\sum_{k=1}^{\infty} a_k = \infty$ asymptotically ensures convergence to the true zero. Note the identical problem setting of finding the zeros of a function when the function can be measured, and finding the zeros of the gradient equation when the gradient can be measured.

2 Kiefer-Wolfowitz Algorithms

As mentioned earlier, stochastic approximation in the Kiefer-Wolfowitz setting, on the other hand, relies on noisy evaluations of the objective (say, loss) function for gradient approximations. This type of stochastic algorithms has the general form

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)$$

where $\hat{g}_k(\hat{\theta}_k)$ is an approximation to $g(\hat{\theta}_k)$ at iteration k. The gain sequence $\{a_k\}$ satisfies the same conditions as in Robbins-Monro processes. The basic advantage of the Kiefer-Wolfowitz algorithms is that they do not require direct gradient information, which in many applications may be very difficult to obtain. Furthermore, there are areas such as Monte-Carlo based optimization where it may be computationally more efficient to obtain a value of the objective function, relative to computation of a gradient.

Stochastic algorithms in the Kiefer-Wolfowitz setting may be characterized by the method a gradient is approximated at each iteration. In the following, we consider three general variants of the Kiefer-Wolfowitz algorithms.

- *Finite Difference Stochastic Approximation (FDSA).*

  In this type of algorithms, the gradient is approximated using basic (one or two-sided) finite differencing. Denoting measurement of the loss function at the design parameter θ by $y(\theta)$, the form for the two-sided finite difference gradient approximation at iteration k is given by

$$[\hat{g}(\hat{\theta}_k)]_i = \frac{y(\hat{\theta}_k + c_k e_i) - y(\hat{\theta}_k - c_k e_i)}{2c_k}$$

where $[\hat{g}(\hat{\theta}_k)]_i$ denotes the component i of the approximated gradient, e_i is a unit vector in direction i, and $\{c_k\}$ is a suitable gain sequence. Both here and in the following, the sequence c_k satisfies, $c_k \to 0$ for $k \to \infty$, and $\sum_{k=1}^{\infty} (a_k/c_k)^2 < \infty$. The one-sided approximation can be written analogously. The algorithm (which laid the foundation for the Kiefer-Wolfowitz type algorithms) was introduced in Kiefer and Wolfowitz (1952) where the study was limited to the scalar case and strong convergence of the iterate was shown. Blum (1954) extends the results of Kiefer and Wolfowitz (1952) to the multivariate case and shows strong convergence of the one-sided FDSA algorithm. Sacks (1958) considers two-sided FDSA and shows asymptotic normality of the iterates.

- *Random Directions Stochastic Approximation (RDSA).*

  In contrast to FDSA, all components of the parameter are randomly perturbed as part of generating the approximation to the gradient. The basic form for the gradient approximation is

$$\hat{g}(\hat{\theta}_k) = \frac{y(\hat{\theta}_k + c_k d_k) - y(\hat{\theta}_k - c_k d_k)}{2c_k} d_k$$

  where d_k is a sequence of user specified random vectors satisfying certain conditions. RDSA was apparently first introduced in Ermoliev (1969), and more thoroughly analyzed in Polyak and Tsypkin (1973), Kushner and Clark (1978), and Ermoliev (1983).

- *Simultaneous Perturbation Stochastic Approximation (SPSA).*

  Similar to RDSA, all components of the parameter vector are perturbed simultaneously using a random vector Δ_k. Denoting the element i of the random vector Δ_k by Δ_{ki}, the

simultaneous perturbation gradient approximation is written as

$$[\hat{g}(\hat{\theta}_k)]_i = \frac{y(\hat{\theta}_k + c_k\Delta_k) - y(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_{ki}}.$$

SPSA was first introduced in Spall (1987) and more thoroughly analyzed in Spall (1992) where strong convergence and asymptotic normality of the iterate were shown under reasonably general conditions. Spall (1992) also contains theoretical as well as numerical comparison results to FDSA. Spall (1995) extends SPSA to include second order effects with the purpose of enhancing the convergence rate of the algorithm.

Finally, we note that both Robbins-Monro and Kiefer-Wolfowitz algorithms may be considered within the more general setting

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k(g(\hat{\theta}_k) + b_k + \gamma_k)$$

where $g(\cdot)$ is continuous, $b_k \to 0$ a.s. for $k \to \infty$ and the random sequence $\{\gamma_k\}$ satisfies certain regularity conditions. For details see Kushner and Clark (1978).

3 Conclusion

We have presented an overview of stochastic approximation algorithms. These algorithms are divided to two general types, Robbins-Monro and Kiefer-Wolfowitz processes. Robbins-Monro processes require noisy measurements of the gradient for optimization, while Kiefer-Wolfowitz algorithms only require noisy measurements of the objective function. We presented three variants of the Kiefer-Wolfowitz type algorithms, namely, FDSA, RDSA, and SPSA. The SPSA algorithm, which relative to FDSA requires $1/p$ the number of measurements per gradient approximation, is of special interest. The reason

is that the p-fold savings per gradient approximation can under reasonably general conditions translate to a p-fold savings in the total number of measurements in order to achieve a given level of accuracy in the optimization process.

References

Blum, J. R. (1954). Multidimensional stochastic approximation methods. *Ann. Math. Stat 25*, 737–744.

Ermoliev, Y. (1969). On the method of generalized stochastic gradients and quasi-fejer sequences. *Cybernetics 5*, 208–220.

Ermoliev, Y. (1983). Stochastic quasigradient methods and their application to system optimization. *Stochastics 9*, 1–36.

Kiefer, J. and J. Wolfowitz (1952). Stochastic estimation of a regression function. *Ann. Math. Stat. 23*, 462–466.

Kushner, H. J. and D. S. Clark (1978). *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer-Verlag, Berlin.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control AC-22*, 551–575.

Ljung, L. (1978). Strong convergence of a stochastic approximation algorithm. *Annals of Statistics 6*, 680–696.

Ljung, L., G. Pflug, and H. Walk (1992). *Stochastic Approximation and Optimization of Random Systems.* Birkhäuser, Berlin.

Polyak, B. T. and Y. Z. Tsypkin (1973). Pseudogradient adaptation and training algorithms. *Automation and Remote Control 34*, 377–397.

Sacks, J. (1958). Asymptotic distributions of stochastic approximation procedures. *Ann. Math. Stat 26*, 373–405.

Spall, J. C. (1987). A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proc. American Control Conference*, pp. 1161–1167.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control 37*(3), 332–341.

Spall, J. C. (1995). Stochastic version of second-order (Newton-Raphson) optimization using only two function measurements. In C. Alexopolos and K. Kang (Eds.), *Proc. of the 1995 Winter Simulation Conference.*

Zangwill, W. I. (1969). *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey.

Optimal Random Perturbations for Stochastic Approximation using a Simultaneous Perturbation Gradient Approximation [1]

Payman Sadegh* and **James C. Spall+**

*Institute of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

+ The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099, USA.

Resubmitted to *IEEE Transactions on Automatic Control.*

[B2]

[1]The first author's work was partly supported by the Danish Research Academy, grant S950029, during his stay at JHU/APL. James Spall's work is supported by U.S. Navy contract N00039-95-C-0002 and the JHU/APL IRAD program.

Abstract

The simultaneous perturbation stochastic approximation (SPSA) algorithm has recently attracted considerable attention for challenging optimization problems where it is difficult or impossible to obtain a direct gradient of the objective (say, loss) function. The approach is based on a highly efficient simultaneous perturbation approximation to the gradient based on loss function measurements. SPSA is based on picking a simultaneous perturbation (random) vector in a Monte Carlo fashion as part of generating the approximation to the gradient. This paper derives the optimal distribution for the Monte Carlo process. The objective is to minimize the mean square error of the estimate. We also consider maximization of the likelihood that the estimate be confined within a bounded symmetric region of the true parameter. The optimal distribution for the components of the simultaneous perturbation vector is found to be a symmetric Bernoulli in both cases. We end the paper with a numerical study related to the area of experiment design.

Key words: Optimization, Stochastic approximation, SPSA, Optimal probability distribution, Experiment design.

1 Introduction

Consider the problem of determining the value of a p-dimensional parameter vector to minimize a loss function $L(\theta)$, where only measurements of the loss function are available (i.e., no gradient information is directly available). The simultaneous perturbation stochastic approximation (SPSA) algorithm has recently attracted considerable attention for challenging optimization problems of this type in application areas such as adaptive control, pattern recognition, discrete event systems, neural network training, and model parameter estimation, see, e.g., Rezayat (1995), Maeda, Hirano, and Kanata (1995), Hill and Fu (1995), Cauwenberghs (1994), Chin (1994), and Parisini and Alessandri (1995).

SPSA was introduced in Spall (1987) and more thoroughly analyzed in Spall (1992). The essential feature of SPSA– which accounts for its power and relative ease of use in challenging multivariate optimization problems– is the underlying gradient approximation that requires only two loss function measurements regardless of the number of parameters being optimized. Note the contrast of two function measurements with the $2p$ measurements required in classical finite difference based approaches (i.e., the Kiefer-Wolfowitz SA algorithm). Under reasonably general conditions, it was shown in Spall (1992) that the p-fold savings in function measurements per gradient approximation can translate directly into a p-fold savings in total number of measurements needed to achieve a given level of accuracy in the optimization process. This means that the SPSA approach uses the same number of iterations as the finite difference approach to achieve a given level of mean square error in the optimization process, but each iteration of SPSA uses only $1/p$ the number of function measurements.

An essential part of the gradient approximation is a simultaneous (random) perturbation relative to the current estimate of θ. This perturbation is generated in a Monte Carlo fashion as part of the optimization process. Since the user has complete control over the perturbation distribution, there is strong reason to choose a distribution as a means of minimizing the number of (potentially costly) function measurements needed in the optimization process. These function measurements may involve physical experiments involving labor or material costs as well as computer related costs associated with simulations or data processing.

The aim of this paper is to determine the form of the optimal distribution for the simultaneous perturbations. This will involve both analytical analysis based on the asymptotic properties of the parameter iterate and numerical finite sample experimentation. The related objectives considered here are to minimize the mean square error of the estimate and to maximize the likelihood that the parameter iterate is restricted to a symmetric bounded region around the true parameter.

The rest of the paper is organized as follows. In Section 2, we briefly review the SPSA algorithm, and present the problem formulation. Section 3 considers the choice of random perturbations. In Section 4, we study an optimization problem from the area of statistical experiment design for dynamic system identification. Section 5 offers concluding remarks.

2 Problem Formulation

Consider the problem of finding a root θ^* of $g(\theta) \equiv \partial L(\theta)/\partial\theta = 0$ for some differentiable loss function $L : \mathbb{R}^p \to \mathbb{R}$. In the case where

the dependence of the loss function upon θ is unknown, but the loss function is observed in the presence of noise, an SA algorithm of the generic Kiefer-Wolfowitz type (Ruppert (1983)) is appropriate.

Let us now briefly review the SPSA algorithm (Spall (1992)) for the problem posed above. Let $\hat{\theta}_k$ denote the estimate for θ at the kth iteration. The SPSA algorithm has the form

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)$$

where $\{a_k\}$ is a gain sequence and $\hat{g}_k(\hat{\theta}_k)$ is a simultaneous perturbation approximation to $g(\hat{\theta}_k)$ at iteration k. The simultaneous perturbation approximation is defined as follows. Let $\Delta_k \in \mathbb{R}^p$ be a vector of p mutually independent mean zero random variables $\{\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kp}\}$. Consistent with the usual framework of stochastic approximations, we have noisy measurements of the loss function at specified "design levels". In particular, at the kth iteration

$$\begin{aligned} y_k^{(+)} &= L(\hat{\theta}_k + c_k\Delta_k) + \epsilon_k^{(+)} \\ y_k^{(-)} &= L(\hat{\theta}_k - c_k\Delta_k) + \epsilon_k^{(-)} \end{aligned}$$

where $\{c_k\}$ is a gain sequence and $\epsilon_k^{(+)}$ and $\epsilon_k^{(-)}$ represent measurement noise terms. The basic simultaneous perturbation form for the estimate of $g(\cdot)$ at the kth iteration is then

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{k1}} \\ \vdots \\ \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{kp}} \end{bmatrix}. \tag{2.1}$$

Note that at each iteration, only *two* measurements are needed to form the estimate. To help mitigate noise effects in high noise environments, it is sometimes useful to consider gradient averaging at

each iteration $\hat{g}_k(\hat{\theta}_k) = q^{-1} \sum_{j=1}^{q} \hat{g}_k^{(j)}(\hat{\theta}_k)$, where each $\hat{g}_k^{(j)}$ is generated as in Eq(2.1) based on a new pair of measurements that are conditionally (on $\hat{\theta}_k$) independent of the other measurement pairs, and q is some integer ≥ 2; this is examined in Spall (1992) but will not be examined further here. Throughout the paper, we assume that:

A1 : $a_k = a/k^{\alpha}$, and $c_k = 1/k^{\gamma}$ where $a > 0$, $0 < \alpha \leq 1$, $\gamma > 0$, $\alpha - \gamma > 0.5$, $\alpha - 2\gamma > 0$, and $3\gamma - \alpha/2 \geq 0$ (since c_k and Δ_k always appear together as $c_k\Delta_k$, we fix the numerator in c_k to unity and let Δ_k vary freely).

A2 : $E\{\epsilon_k^{(+)} - \epsilon_k^{(-)} | \hat{\theta}_k, \Delta_k\} = 0$, and for some $\alpha_0, \delta > 0$ and $\forall k$, $E\{\epsilon_k^{\pm(2+\delta)}\} < \alpha_0$. Moreover, there is a σ^2 such that $E\{(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 | \hat{\theta}_k, \Delta_k\} \to \sigma^2$ as $k \to \infty$.

A3 : For all $k < \infty$, $\{\Delta_{ki}\}$ $(i = 1, ..., p)$ are i.i.d. and symmetrically distributed about 0 with $|\Delta_{ki}| \leq \alpha_0$ a.s. and $E|\Delta_{ki}^{-1}| \leq \alpha_1$ a.s. for some $\alpha_0, \alpha_1 > 0$. For some $\alpha_2, \alpha_3, \delta > 0$, it holds that $E\{|L(\hat{\theta}_k \pm c_k\Delta_k)|^{2+\delta}\} \leq \alpha_2$ and $E(\Delta_{ki}^{-2-\delta}) \leq \alpha_3$, $i = 1, ..., p$. Moreover, there are ρ^2, ξ^2 such that as $k \to \infty$, $E(\Delta_{ki}^2) \to \rho^2$ and $E(\Delta_{ki}^{-2}) \to \xi^2$ for all $i = 1, ..., p$.

A4 : $\sup_k ||\hat{\theta}_k|| < \infty$ a.s. where $||\cdot||$ denotes usual Euclidean norm.

A5 : θ^* is an asymptotically stable solution of the differential equation $dx/dt = -g(x)$.

A6 : Let $D(\theta^*) = \{x_0 : \lim_{t\to\infty} x(t|x_0) = \theta^*\}$ where $x(t|t_0)$ denotes solution to the differential equation of A5 based on initial condition x_0. There exists a compact set $S \subset D(\theta^*)$ such that $\hat{\theta}_k \in S$ infinitely often for almost all $\hat{\theta}_k$.

A7 : For almost all $\hat{\theta}_k$, there is an open ball about $\hat{\theta}_k$ whose radius is independent of k or $\hat{\theta}_k$, where the third derivative of the loss function exists continuously and is uniformly bounded.

The reader is referred to Spall (1992) for remarks on the assumptions.

The problem of selecting random perturbations is formulated as selecting a sequence of probability distributions for Δ_{ki}, $k = 1, 2, ...$, each from the set of allowable probability distributions for the random perturbations (see A3). The objective is to optimize a suitable criterion related to the parameter estimate.

For small k, the exact distribution of $\hat{\theta}_k$ is dependent upon the (unknown) joint probability distribution of the noise sequence. Therefore, we solve the optimal random perturbation problem using the asymptotic distribution of the estimate. It follows from Proposition 2 of Spall (1992) that as $k \to \infty$:

$$k^{\frac{\beta}{2}}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} Z \sim N(\xi^2 d, \rho^2 D) \tag{2.2}$$

where β is a positive constant, and d and D are quantities *not* dependent upon the random perturbations. The matrix D depends on the Hessian of $L(\theta)$ at θ^* and σ^2, and d depends on the third order derivative of $L(\theta)$ at θ^*. Both d and D are dependent upon a, α, and γ. The reader is referred to Spall (1992) for the detailed forms of d and D. In the following, we let ψ denote the factors affecting d and D that may be unknown to the user.

From Eq(2.2), it is evident that the distribution of Z is affected by the random perturbations only through ρ^2 and ξ^2 (see A3). Hence, using the asymptotic result for sufficiently large number of iterations, the problem simplifies to selection of a *single* probability distribution for Δ_{ki}, for *all* $k = 1, 2, ...$, optimizing some criterion related to Z.

3 Optimal Choice of Random Perturbations

As mentioned in the previous section, the analysis here is based on the asymptotic distribution of the parameter iterate; the authors are unaware of any corresponding finite sample result that would be useful in such calculations.

Subsections 3.1 and 3.2 provide the optimal distribution with the goal of minimizing the trace of mean square error of the estimate, and maximizing the probability of restricting the estimation error within some bounded symmetric about zero region, respectively.

3.1 Mean Square Error Criterion

In this subsection, we seek a probability distribution that minimizes the expression $MSE \triangleq E\{\text{trace}[ZZ^\top]\}$. We refer to this criterion as the mean square error criterion. As we shall see, the solution to this problem can be readily obtained using a simple formula.

Proposition 1 below assumes that ψ is completely known (see Section 2 for definition of ψ). Corollary 1, on the other hand, presents the optimal *generic* form of the distribution, assuming no knowledge of ψ.

Using Eq(2.2)

$$MSE = \rho^2 \text{trace}\{D\} + \xi^4 d^\top d. \tag{3.1}$$

Denote $K_1 = \text{trace}\{D\}$ and $K_2 = d^\top d$ (the numbers K_1 and K_2 do not depend upon the random perturbations). In the following, we let $Pr(\cdot)$ denote probability.

Proposition 1 For all $k = 1, 2, ...,$ and $i = 1, ..., p$, the symmetric Bernoulli distribution

$$Pr(\Delta_{ki} = \pm(\frac{K_1}{2K_2})^{\frac{1}{6}}) = \frac{1}{2} \tag{3.2}$$

is the unique single allowable distribution for Δ_{ki}, minimizing the mean square error criterion.

PROOF: First, we show that the optimal solution is necessarily a symmetric Bernoulli distribution. In order to prove this, we show that for any distribution P to qualify as optimum, it should necessarily hold that $\rho^2\xi^2 = 1$. From the Schwarz inequality $\rho^2\xi^2 \geq 1$. Now suppose $\rho^2\xi^2 > 1$. Reduce ξ^2 slightly to ξ'^2 such that $\rho^2\xi'^2 > 1$ and find a distribution P' with ρ^2 and ξ'^2 as inverse second and second moments (such distribution can always be found since $\rho^2\xi'^2 > 1$). Referring to Eq(3.1), it is obvious that the mean square error under P' is smaller than the mean square error under P (note that this result holds *regardless* of the values of K_1 and K_2). Hence, P can not be an optimal distribution and it should necessarily hold that $\rho^2\xi^2 = 1$. However, $E(\Delta_{ki}^2)E(1/\Delta_{ki}^2) = 1$ iff Δ_{ki} and $1/\Delta_{ki}$ are constant multiples of one another, and for symmetrically distributed variables that holds iff Δ_{ki} is symmetric Bernoulli distributed.

The proof of Eq(3.2) and the uniqueness follows easily; the distribution given by Eq(3.2) is allowable, and among all the symmetric Bernoulli distributions, the one with ρ^2 being the unique minimizer of

$$[K_1\rho^2 + K_2\xi^4]_{\xi^2=1/\rho^2} = K_1\rho^2 + K_2(1/\rho^2)^2$$

is the unique optimal distribution. Q.E.D.

From a practical point of view, Corollary 1 below is important in showing that a Bernoulli distribution with given ρ^2 and ξ^2 will

always improve upon any other distribution with the the same ρ^2 and ξ^2. This result requires *no knowledge* of K_1 and K_2.

Corollary 1 For a given ρ^2 (or ξ^2), the Bernoulli distribution $\Delta_{ki} = \pm\rho$ (or $\Delta_{ki} = \pm\xi^{-1}$) provides a lower value of Eq(3.1) than any other distribution with the same ρ^2 (or ξ^2).

PROOF: Follows immediately from the necessity part of Proposition 1.

Remark 1 To invoke the full optimality of the result in Proposition 1, we require complete knowledge of ψ. This is analogous to the calculations for the optimal gain sequences of stochastic algorithms, see e.g. Fabian (1971) and Chin (1997). The result in Corollary 1 partially mitigates this situation in that it implies that no matter how a given perturbation distribution is determined, there is a Bernoulli distribution that yields a lower MSE, for any ρ (or ξ) of the given distribution. Another frequently encountered situation is the case where an *a priori* model for $L(\theta)$ is given, but only as an implicit function (i.e., it is only possible to compute $L(\theta)$ for each θ). In such cases, it is often difficult to accurately evaluate the second and third order derivatives or the noise variance σ^2 to be used for the calculation of the optimal distribution. The following procedure may be useful in such situations. By applying SPSA to the available model using very large number of iterations K, we obtain the estimate $\hat{\theta}_K$ which we use as the true optimum in our calculations. We then approximate ψ using the given model and $\hat{\theta}_K$, and use Eq(3.2) to find an approximation to the optimal perturbation magnitude which will be used as the initial guess for a numerical search. Corollary 1 implies that the optimal perturbation distribution should be sought among symmetric Bernoulli distributions. We sample the Δ_{ki} from Bernoulli distributions with varying magnitudes around the initial

guess. For each magnitude, we apply SPSA a number of times (cross sections), obtain $\hat{\theta}_k$ for each cross section to find $||\hat{\theta}_k - \hat{\theta}_K||^2$ where $k << K$ is some large iteration number of interest, and average over the computed values of $||\hat{\theta}_k - \hat{\theta}_K||^2$ to numerically evaluate the mean square error for each one of the Bernoulli distributions respectively. The numerical study of the paper illustrates such a procedure.

Remark 2 There might be situations where the value of ψ is only partially known, and the partial knowledge is embedded in a probability distribution. It then follows that

$$MSE = E_\psi E\{ZZ^\top|\psi\} = E_\psi(K_1)\rho^2 + E_\psi(K_2)\xi^4$$

where $E_\psi(\cdot)$ denotes expectation with respect to the partially known factor ψ. The solution is readily found using Eq(3.2) after replacing K_1 and K_2 by $E_\psi(K_1)$ and $E_\psi(K_2)$.

It is important to note that the SPSA algorithm only requires noisy evaluations of the loss function, and this is typically done by real experimentations on the system. The given *a priori* models are then only used for the optimal design of random perturbations.

3.2 Probability Criterion

Our objective in this subsection is to maximize the likelihood of restricting the error Z within some bounded symmetric (about zero) region. We denote such a region by V_θ in the following. A similar approach is pursued in Gusev and Krasulina (1995) to determine the constants of a Robbins-Monro type stochastic approximation algorithm. The optimality criterion is written as

$$J = Pr\{Z \in V_\theta\}. \tag{3.3}$$

An important special case for J is where V_θ is the closed unit ball. Then the criterion to be maximized is $Pr\{||Z|| \leq A\}$, where as usual, $||\cdot||$ denotes Euclidean norm and A is a positive number chosen by the user. It reflects the user's tolerable amount of error.

The criterion can be rewritten as

$$J = E_\psi(Pr\{Z \in V_\theta|\psi\}) \tag{3.4}$$

where we have also considered the case where ψ is only partially known and the prior knowledge of ψ is embedded in a probability distribution. The following proposition shows that the choice of probability distribution function to maximize J which in general requires a functional optimization is reduced to optimization of a function of only one variable.

Proposition 2 For all $k = 1, 2, ...,$ and $i = 1, ..., p$, the single distribution for Δ_{ki}, maximizing the probability criterion J (see Eq(3.3) or Eq(3.4)) is necessarily a symmetric Bernoulli distribution.

PROOF: Identical to the proof of Proposition 1, build P' in order to compare to P with $\rho^2\xi^2 > 1$. Compare the asymptotic distributions (see Eq(2.2)) and notice the relation between the bias terms, $||d\xi'^2|| < ||d\xi^2||$, which holds *regardless* of the value of ψ. However, the covariance term $\rho^2 D$ remains the same, leading to the conclusion that for any ψ and any V_θ, $Pr\{Z \in V_\theta|\psi\}$ is larger under P'. This proves the required necessity. Q.E.D.

It easily follows that a result identical to Corollary 1 holds for the probability criterion. In case an implicit *a priori* model for $L(\theta)$ is given, we can apply a procedure similar to the one explained in Remark 1 for the design of random perturbations. The required modification is as follows. For each Bernoulli distribution, we compute the number of times (among the cross sections) $\hat{\theta}_k - \hat{\theta}_K$ lies within

the region of interest and divide this number by the total number of cross sections to assess J numerically for the varying Bernoulli distributions.

Remark 3 Consider the degenerate case $d = 0$. This for example occurs when the third order derivatives of the loss function at θ^* are zero, see Spall (1992). Then, clearly the optimal solution according to both the mean square error and probability criteria will be a distribution with $\rho \to 0$, forcing the covariance $\rho^2 D$ to zero. This implies that $\Delta_{ki} \to \pm\infty$ is the optimal choice for random perturbations. However, $\lim_{k\to\infty} c_k = 0$, meaning that it is not possible to draw any definitive conclusion about the optimal size of $c_k\Delta_k$ based on the asymptotic properties. In finite sample cases, c_k does not get infinitesimally small, and it is obviously not allowed to let $|\Delta_{ki}| \to \infty$, either. However, a practical guideline in $d = 0$ situations is to select the magnitude of Δ_{ki} as large as the algorithm does not go unstable. This example shows that the results based on the asymptotic distribution must be interpreted and used with some care in finite sample cases.

4 Numerical Study

In this section, we apply SPSA to a statistical experiment design problem for parameter estimation in a dynamic model, see e.g. Goodwin and Payne (1977). Consider the following autoregressive model with exogenous inputs (ARX(2,1)):

$$y_t = h_1 y_{t-1} + h_2 y_{t-2} + u_t + e_t \tag{4.1}$$

where $\{u_t\}$ and $\{y_t\}$ are input and output sequences and $\{e_t\}$ is a sequence of mean zero i.i.d. Gaussian random variables. We assume

that the input sequence is generated by a *finite* register with length 10, meaning that the input repeats periodically with cycle 10. We wish to compute the input sequence parameter $(u_1, ..., u_{10})^\top$ which starting from zero initial condition minimizes

$$J = -E\{\log\det M_F\} + 0.5\sum u_t^2 \tag{4.2}$$

where

$$M_F = \begin{bmatrix} \sum_{i=n_1}^{n_2} y_{t-1}^2 & \sum_{i=n_1}^{n_2} y_{t-1}y_{t-2} \\ \sum_{i=n_1}^{n_2} y_{t-1}y_{t-2} & \sum_{i=n_1}^{n_2} y_{t-1}^2 \end{bmatrix}.$$

Notice that such a problem formulation implies that we deal with a static optimization problem and not a dynamic one since we consider the whole sequence of data $\{y_t\}$ in batch mode within the loss function and a fixed number of parameters independent of the size of the data set. We explain Eq(4.2) as follows. Assuming that we are interested in estimating $\Lambda = (h_1, h_2)^\top$, the basic least squares estimate is given by (see, e.g. Ljung (1987))

$$\hat{\Lambda} = M_F^{-1} \begin{bmatrix} \sum_{i=n_1}^{n_2} (y_t - u_t)y_{t-1} \\ \sum_{i=n_1}^{n_2} (y_t - u_t)y_{t-2} \end{bmatrix}.$$

Hence, by selecting the input sequence to maximize the expected value of the (logarithm) of the determinant of M_F, we wish to avoid the problem of the singularity of M_F. Indeed, for large values of sample size, the matrix M_F is (approximately) proportional to Fisher's information matrix for the model given by Eq(4.1) (Goodwin and Payne (1977), Chapter 6). This choice of optimality criterion is called D-optimality in the statistical experiment design literature, see e.g.

Fedorov (1972). Since the positive semi-definite matrix M_F is an increasing function of the input power $\sum u_t^2$, the second term of the criterion penalizes signals with large power. For a detailed treatment of the problem of input design for dynamic system identification, see Goodwin and Payne (1977). In a large part of the literature on experiment design, the solution is obtained by fixing an *a priori* model for the data and calculation of the information matrix as a function of input. Since the design of experiments precedes the parameter estimation stage (the goal of the design is indeed to find the experiments that yield good identification of the system), the knowledge of the system prior to the identification experiment may be very poor and is mostly based on *preliminary* investigations and parameter estimations. The sensitivity of the solution to *a priori* fixed model parameters, and solutions for the case of imperfect prior knowledge using e.g. Bayesian formulation has attracted some attention, see Pronzato and Walter (1987) and Sadegh, Holst, Madsen, and Melgaard (1995). The solution presented in this example only requires knowledge of the model structure; no knowledge of the values of the model parameters is needed. The optimization is solely based on real experimentations which consist of applying input sequences to the system (at initial rest) and approximation of the gradient with respect to the input sequence. Therefore, we directly find the optimal inputs without requiring a preliminary parameter estimation stage.

Let us assume that the model parameters are given by $h_1 = 1.45$, $h_2 = -0.475$ (which correspond to poles 0.5 and 0.95), and the standard deviation of e_t is 0.05. Note that these values are used for data generation purpose, and to (approximately) determine the optimal distribution of the random perturbations. The SPSA algorithm requires no knowledge of these values. In the following, we select $n_1 = 9$, $n_2 = 64$, $a_k = 0.1/k^{0.9}$, and $c_k = 1/k^{0.15}$.

We first apply SPSA with 50000 iterations, $q = 1$, and $\Delta_{ki} = \pm 0.1$ (Bernoulli distributed) in order to obtain an estimate of the (uncomputable) optimal sequence $\{u_t^*\}$ for later reference. This value will be used as the true optimum for the rest of the paper since the number of iterations for all later estimation is $1200 << 50000$. Then, we assess the second and third order derivatives of the loss function at the optimum, $\{u_t^*\}$, by numerical finite difference method for the noise free case. Also, we approximate σ^2 by simulation of 1000 realizations of $[\log\det(M_F)]$ at $\{u_t^*\}$. Inserting these estimates in Eq(3.2), the following distribution is obtained

$$Pr(\Delta_{ki} = \pm 0.19) = \frac{1}{2}.$$

This distribution shall only be used as an initial guess for a numerical search to find the optimizer for the mean square error and probability criteria since only rough estimates of K_1 and K_2 (see Eq(3.2)) are available.

We apply Bernoulli distributions with magnitude of the outcome around 0.19, estimate the optimal input sequence a number of times (100 times in this example), and assess the values of the mean square error and probability criteria numerically. The optimal distribution, according to both the mean square error and probability criteria, is found to be a ± 0.25 Bernoulli distribution. We use the same procedure as above to compare the optimal distribution against other choices of distribution. In Table 1, all the distributions correspond to Bernoulli distributed variables. The top row of the table provides the relevant Bernoulli distributions. For the probability criterion, we have chosen the special case below Eq(3.3) with $A = 4 \times 10^{-3}$. The results indicate that an inappropriate choice of random perturbations (e.g. ± 1 in this numerical study) would lead to very poor estimation properties.

	± 0.15	± 0.25	± 0.4	± 1
MSE	0.0063	0.0052	0.0073	0.1061
J	0.36	0.51	0.35	0.0

Table 1: *Performance of SPSA under varying Bernoulli distributions.*

We also apply a random variable uniformly distributed over $[-0.3, -0.2] \cup [0.2, 0.3]$. This choice is interesting since the support of the distribution includes the optimal support point of the optimal Bernoulli (± 0.25). The numerical evaluations of MSE and J yield 0.0062 and 0.39, respectively, which are noticeably worse than the results for the optimal Bernoulli distribution.

Finally, notice that in Table 1, the number of iterations have been chosen relatively large (1200) in order to let the iterates reach the asymptotic condition. In order to investigate the performance of the asymptotic solution for small sample cases and large initial deviations from the true optimum, consider a case of 10 iterations with a 17.5% initial deviation for all components of $\{u_t\}$. We are particularly interested in numerically determining whether or not a *distribution form* other than symmetric Bernoulli yields better results than the asymptotically optimal solution. Therefore, we test the optimal Bernoulli (± 0.25) distribution against two bimodal distributions. One is chosen to be a random variable uniformly distributed over $[-0.3, -0.2] \cup [0.2, 0.3]$. The other corresponds to a random variable triangular distributed over both $[0.2, 0.3]$ and $[-0.3, -0.2]$. The corresponding MSE values are 0.0756, 0.0789, 0.0764, respectively. This comparison obviously does not establish the optimality of the Bernoulli distribution in all finite sample cases. However, it indicates that the asymptotic solution may perform reasonably well even for very small sample sizes.

5 Concluding Remarks

The paper deals with the optimal choice of random perturbations for the SPSA algorithm. Since the user has full control over this choice, there is strong reason to pick this distribution wisely in order to reduce the overall costs of optimization. We have shown that for the mean square error and probability criteria, the optimal random perturbations should be sampled from a symmetric Bernoulli distribution. The choice of the optimal Bernoulli distribution (i.e. the magnitude of its outcome) is dependent upon the prior information about the loss function. However, in the usual case where such information is unavailable, this paper shows that the Bernoulli *distribution form* is the (asymptotically) optimal form *regardless* of the value of the variance of the perturbation distribution. This has significant practical implication as the perturbation distribution is typically determined based on small scale experimentation and/or limited prior knowledge about the form of the loss function. All the results are based on the asymptotic theory. Investigating the choice of random perturbations for finite sample cases is of significant theoretical and practical interest and represents a possible topic for future research on the subject.

References

Cauwenberghs, G. (1994). *Analog VLSI Autonomous Systems for Learning and Optimization.* Ph. D. thesis, Dept of Electrical Engineering, California Institute of Technology.

Chin, D. C. (1994). A more efficient global optimization based on Styblinski and Tang. *Neural Nets.* 7, 573–574.

Chin, D. C. (1997). Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Transactions on Systems, Man, and Cybernetics 27*. In press.

Fabian, V. (1971). Stochastic approximation. In J. J. Rustagi (Ed.), *Optimizing Methods in Statistics*, pp. 439–470. Academic, New York.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic, New York.

Goodwin, G. C. and R. L. Payne (1977). *Dynamic System Identification: Experiment Design and Data Analysis*. Academic, New York.

Gusev, S. V. and T. P. Krasulina (1995). An algorithm for stochastic approximation with a preassigned probability of not exceeding a required threshold. *Journal of Computer and Systems Sciences International 33*, 39–41.

Hill, S. D. and M. C. Fu (1995). Transfer optimization via simulation perturbation stochastic approximation. In *Proc. Winter Simulation Conference*, pp. 242–249.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, New Jersey.

Maeda, Y., H. Hirano, and Y. Kanata (1995). A learning rule of neural networks via simultaneous perturbation and its hardware implementation. *Neural Nets 8*, 251–259.

Parisini, T. and A. Alessandri (1995). Non-linear modeling and state estimation in a real power plant using neural networks and stochastic approximation. In *Proc. American Control Conference*, pp. 1561–1567.

Pronzato, L. and E. Walter (1987). Robust experiment design for non-linear regression models. In V. Fedorov and H. Läuter (Eds.), *Lecture Notes in Economics and Mathematical Sci-*

ences,297, pp. 72–86. Springer-Verlag, New York.

Rezayat, F. (1995). On the use of an SPSA-based model free controller in quality improvement. *Automatica 31*, 913–915.

Ruppert, D. (1983). Kiefer-wolfowitz procedure. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, pp. 379–381. Wiley.

Sadegh, P., J. Holst, H. Madsen, and H. Melgaard (1995). Experiment design for grey-box identification. *International Journal of Adaptive Control and Signal Processing 9*(6), 491–507. See article A2 for a revised version.

Spall, J. C. (1987). A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proc. American Control Conference*, pp. 1161–1167.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control 37*(3), 332–341.

Constrained Optimization via Stochastic Approximation with a Simultaneous Perturbation Gradient Approximation[1]

Payman Sadegh

Institute of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

Accepted for publication in *Automatica.*

[B3]

[1]This work is partly supported by the Danish Research Academy, grant S950029. I also would like to thank Prof. Jan Holst, Lund Institute of Technology, for his valuable comments.

Abstract

The paper deals with a projection algorithm for stochastic approximation using simultaneous perturbation gradient approximation for optimization under inequality constraints where no direct gradient of the loss function is available and the inequality constraints are given as explicit functions of the optimization parameters. It is shown that under application of the projection algorithm, the parameter iterate converges almost surely to a Kuhn-Tucker point. The procedure is illustrated by a numerical example.

Key words: Optimization, Stochastic approximation, SPSA, Constrained optimization, Inequality constraints, Kuhn-Tucker point.

1 Introduction

The simultaneous perturbation stochastic approximation (SPSA) algorithm has recently attracted considerable attention for multivariate optimization problems where only noisy measurements of the loss function are available (i.e., no gradient information is directly available), see e.g., Rezayat (1995), Maeda, Hirano, and Kanata (1995), Cauwenberghs (1994), Chin (1994), and Parisini and Alessandri (1995).

SPSA was introduced in Spall (1987) and more thoroughly analyzed in Spall (1992). The algorithm is a variant of the stochastic approximation (SA) in a Kiefer-Wolfowitz setting (Kushner and Clark (1978)) where only noisy measurements of the loss function are available (used for gradient approximations). The essential feature of SPSA is its highly efficient gradient approximation that requires only *two* loss function measurements regardless of the number of optimization parameters. The gradient approximation is generated by simultaneous (random) perturbation relative to the current estimate of the parameter θ. Note the contrast of two function measurements with the $2p$ measurements required in the classical finite-difference based Kiefer-Wolfowitz SA algorithm, where p is the number of parameters. Under reasonably general conditions, it was shown in Spall (1992) that the p-fold savings in function measurements per gradient approximation translates directly into a p-fold savings in the total number of measurements needed to achieve a given level of accuracy in the optimization process.

The original SPSA algorithm as presented in Spall (1992) is an unconstrained algorithm. Constraints, on the other hand, are essential parts of almost all real world optimization applications. The present work may be regarded as the extension of the convergence

result of Spall (1992) to constrained optimization problems. This paper presents a projection SPSA algorithm that can handle *inequality* constraints. A similar approach is pursued in L'Ecuyer and Glynn (1994) for optimization of queuing systems using stochastic approximation. Here, we focus on SPSA and treat more general constraints. However, we restrict attention to the constraints that are given as explicit functions of the optimization parameter. Common to the Kiefer-Wolfowitz stochastic approximations, function evaluations often mean *real measurements* on the system. We are interested in situations where the constraints are determined by the feasible operating conditions of the system. Hence, we assume that function evaluations at the points where the constraints are violated are not feasible. This is stronger than the requirement of restricting the solution to the feasible domain (as in constrained versions of Robbins-Monro type SA algorithms, see Kushner and Clark (1978)). In this regard, the projection algorithm is advantageous relative to other constrained SA optimization techniques such as the Lagrangian method (see Kushner and Clark (1978)) where the parameter iterate only asymptotically lies in the feasible set. We establish the almost sure convergence of the parameter iterate to a Kuhn-Tucker point under application of the projection algorithm.

The organization of the rest of the paper is as follows. Section 2 studies the projection SPSA algorithm and the convergence result. Section 3 presents a numerical example where the procedure is illustrated and tested using (finite sample) numerical experimentations. Finally, Section 4 offers concluding remarks.

2 Projection SPSA Algorithm and Strong Convergence

In this section, we treat a projection SPSA algorithm for minimization under constraints, i.e. the problem of

$$\min_{\theta \in G} L(\theta)$$

where similar to the regularity conditions for the unconstrained case (Spall (1992)), the loss function $L(\theta)$ is continuously differentiable on an open set containing G. The reader is referred to Spall (1992) for a detailed treatment of the (unconstrained) SPSA algorithm. We deal with *inequality* constraints and introduce

Assumption 1 The set $G = \{\theta : q_1(\theta), ..., q_s(\theta) \leq 0\}$ is non-empty, bounded, and the functions $q_i(\theta)$, $i = 1, ..., s$, are continuously differentiable. At each $\theta \in \partial G$, where ∂ denotes boundary, the gradients of the active constraints are linearly independent. Furthermore, there exists an $\epsilon < 0$ such that the set $G^- = \{\theta : q_1(\theta), ..., q_s(\theta) \leq r\}$ is non-empty for $\epsilon \leq r < 0$ (i.e., the set G has non-empty interior).

The proof of convergence to a Kuhn-Tucker point as follows later is based on Theorem 5.3.1 of Kushner and Clark (1978) where the assumption on G (Kushner and Clark (1978), page 190, A5.3.1) states that G is the closure of its interior rather than the non-emptiness of G^- in Assumption 1. It is easy to see that because of the continuity of the $q_i(\theta)$, the set $\{\theta : q_i(\theta) < 0, i = 1, ..., s\}$ is open and indeed equal to $intG$ where int denotes interior. This together with the non-emptiness of G^- yields $G = \overline{intG}$. Assumption 1 is formulated with the goal of easing later presentation.

Another type of constrained problems involves constraint functions that can only be observed in the presence of noise, see e.g.

Ljung, Pflug, and Walk (1992). Such constraints will not be examined here.

Let $\hat{\theta}_k$ denote the estimate for θ at the kth iteration, and for all $\theta \in \mathbb{R}^p$, let $P(\theta)$ be the nearest point to θ on G where the norm is defined as the usual Euclidean norm. The projection algorithm has the general form

$$\hat{\theta}_{k+1} = P(\hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)) \tag{2.1}$$

where the gain sequence $\{a_k\}$ shall satisfy certain conditions (as follows) and $\hat{g}_k(\hat{\theta}_k)$ is an approximation to the gradient at $\hat{\theta}_k$. The simultaneous perturbation estimate for the gradient at θ, $g(\theta)$, is defined as follows. Let $\Delta_k \in \mathbb{R}^p$ be a vector of p mutually independent mean-zero random variables $\{\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kp}\}$ satisfying certain conditions (Spall (1992)). A condition on random perturbations is norm boundedness, i.e. $||\Delta_k|| \leq \alpha_0$ for some $\alpha_0 > 0$. In Spall (1992), the boundedness condition is $||\Delta_k|| \leq \alpha_0$ a.s. Noting that the user has full control over random perturbations, for simplicity we follow the strict boundedness assumption. Consistent with the usual framework of stochastic approximations, we have noisy measurements of the loss function. In particular, at the kth iteration

$$\begin{aligned} y_k^{(+)} &= L(\theta + c_k\Delta_k) + \epsilon_k^{(+)} \\ y_k^{(-)} &= L(\theta - c_k\Delta_k) + \epsilon_k^{(-)} \end{aligned}$$

where $\{c_k\}$ is a gain sequence and $\epsilon_k^{(+)}$ and $\epsilon_k^{(-)}$ represent measurement noise terms that satisfy $E\{\epsilon_k^{(+)} - \epsilon_k^{(-)}|\theta, \Delta_k\} = 0$. The gain sequences $\{a_k\}$ and $\{c_k\}$ are positive for all k and tend to zero as $k \to \infty$. Moreover, $\sum_{k=0}^{\infty} a_k = \infty$, and $\sum_{k=0}^{\infty} (a_k/c_k)^2 < \infty$. For convenience, we take $c_k = c/k^\gamma$, $\gamma \geq 0$.

The basic simultaneous perturbation (SP) form for the estimate

of $g(\theta)$ at iteration k is defined by

$$g_k^{SP}(\theta) = \begin{bmatrix} \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{k1}} \\ \vdots \\ \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{kp}} \end{bmatrix}.$$

Note that at each iteration, only *two* measurements are needed to form the estimate. The main features of our proposed solution relative to the unconstrained SPSA algorithm are as follows. Firstly, the projection $P(\cdot)$ always restricts the iterates $\hat{\theta}_k$ within G which is obviously not needed for the unconstrained case. The projection is indeed an essential feature of the constrained algorithm; eliminating $P(\cdot)$, the iterates may vary anywhere in $\mathbb{R}^p$ as a result of noisy observations, no matter how the gain or random perturbation sequences of the algorithm are selected. Secondly, in the unconstrained algorithm, we have $\hat{g}_k(\hat{\theta}_k) = g_k^{SP}(\hat{\theta}_k)$. Such approximation cannot be directly used here since it may occur that $\hat{\theta}_k \in G$ but $\hat{\theta}_k \pm c_k\Delta_k \notin G$. Especially, in case $\hat{\theta}_k \in \partial G$, there is always a (random) direction Δ_k such that $\hat{\theta}_k \pm c_k\Delta_k \notin G$, no matter how small the gain c_k is selected. Notice that the case $\hat{\theta}_k \in \partial G$ is expected to occur frequently for the very relevant situation that the true optimum belongs to the boundary of the feasible domain. Except for simulation based optimization cases, function evaluations involve real measurements on the system and it is usually not allowed to take measurements outside the feasible domain. To overcome this problem, we further project $\hat{\theta}_k$ onto a (closed) set G_k contained within G to obtain $P_k(\hat{\theta}_k)$ which shall (only) be used for computing an SP gradient approximation at the kth iteration. If the distance d_k between the nearest points on ∂G and ∂G_k is equal to or larger than $c_k\alpha_0$, then $P_k(\hat{\theta}_k) \pm c_k\Delta_k \in G$, ensuring that the SP approximation to the gradient at $P_k(\hat{\theta}_k)$ (instead of $\hat{\theta}_k$) requires

no function measurement outside G. The SP gradient approximation at $P_k(\hat{\theta}_k)$ obviously introduces an (extra) error term relative to the SP gradient approximation at $\hat{\theta}_k$. However, if $G_k \to G$ for $k \to \infty$, then continuous differentiability of $L(\theta)$ yields that the extra error term tends to zero. This line of argument will be used in the proof of convergence later. But first, we describe a procedure for selecting the G_k (a simple case of this is given in the illustrative example of the paper). Define $G_k \subset G$ by $G_k = \{\theta : q_i(\theta) \leq r_k < 0, i = 1, ..., s\}$ where $r_k \to 0$ as $k \to \infty$. Assumption 1 states that there exists an $\epsilon < 0$ such that G_k is non-empty for $\epsilon \leq r_k < 0$, $k = 1, 2, ...$ Hence, select $\epsilon \leq r_1 < 0$ and select c such that $d_1 \geq c_1\alpha_0$. Once c is selected, choose $r_k \to 0$ such that $d_k \geq c_k\alpha_0$ (note that $c_k \to 0$ as $k \to \infty$).

Remark 1 It follows from above that the bottom-line in computing an SP gradient approximation at $P_k(\hat{\theta}_k)$ is to ensure the feasibility of function evaluations. There may therefore exist different methods to obtain a point $\theta'_k \in G$ for the SP gradient approximation at iteration k such that $\theta'_k \pm c_k\Delta_k \in G$, and in some sense, the magnitude of $\theta'_k - \hat{\theta}_k$ is small for all k (to avoid large error terms on the gradient approximations) and becomes infinitesimally small as $k \to \infty$. The proposed solution of the paper provides a suitable technique which can be generically applied to all types of constrained problems where Assumption 1 holds.

Finally, it should be noted that projections in general are unfortunately not very easy to compute unless linear approximations to $q_i(\theta)$ at the current iterate are obtained first. Such approximations can often be justified in practice, since $a_k \to 0$.

Proposition 1 Let Assumption 1, and assumptions A1-A5 and conditions of Lemma 1 (for simplicity, replace the a.s. boundedness of Δ_k by strict boundedness) of Spall (1992) hold where all the regular-

ity conditions on $L(\cdot)$ hold on an open set containing G. Then under the projection algorithm (see Eq(2.1)) where $\hat{g}_k(\hat{\theta}_k) = g_k^{SP}(P_k(\hat{\theta}_k))$, as $k \to \infty$

$$\hat{\theta}_k \to KT \quad \text{a.s.},$$

where KT is the set of Kuhn-Tucker points (i.e. the set of points θ where there are $\lambda_i \geq 0$ such that $g(\theta) + \sum_{i:q_i(\theta)=0} \lambda_i dq_i(\theta)/d\theta = 0$).

Proof: Decompose the error $\hat{g}_k(\hat{\theta}_k) - g(\hat{\theta}_k) = g_k^{SP}(P_k(\hat{\theta}_k)) - g(\hat{\theta}_k)$ into a sum of $b_k^I = E(g_k^{SP}(P_k(\hat{\theta}_k))|\hat{\theta}_k) - g(P_k(\hat{\theta}_k))$, $e_k = g_k^{SP}(P_k(\hat{\theta}_k)) - E(g_k^{SP}(P_k(\hat{\theta}_k))|\hat{\theta}_k)$, and $b_k^{II} = g(P_k(\hat{\theta}_k)) - g(\hat{\theta}_k)$. Identical to the proof of Lemma 1 and Proposition 1 of Spall (1992), it can be shown that

(i) $\sup_k |b_k^I| < \infty$ and $b_k^I \to 0$ a.s. for $k \to \infty$,

(ii) $\lim_{k\to\infty} Pr(\sup_{m\geq k} |\sum_{i=k}^{m} a_i e_i| \geq \eta) = 0$ for any $\eta > 0$,

where $Pr(\cdot)$ denotes probability. Moreover, since G is bounded, $G_k \to G$, and $L(\theta)$ is continuously differentiable at all $\theta \in G$,

(iii) $\sup_k |b_k^{II}| < \infty$ and $b_k^{II} \to 0$ for $k \to \infty$.

Then the assumptions of Theorem 5.3.1 of Kushner and Clark (1978) are satisfied and the proposition follows. Q.E.D.

Remark 2 In the proof of the above proposition, we have used (ii) rather than A5.3.2 of Kushner and Clark (1978), page 191, which states that for some $T_0 > 0$ and any $\eta > 0$

$$\lim_{n\to\infty} Pr(\sup_{j\geq n} \max_{t\leq T_0} |\sum_{i=m(jT_0)}^{m(jT_0+t)-1} a_i e_i| \geq \eta) = 0 \qquad (2.2)$$

where $m(t) = \max\{n : \sum_{i=0}^{n-1} a_i \leq t\}$ for $t \geq 0$ and $m(t) = 0$ otherwise. Eq(2.2) is indeed the assumption used to prove Theorem 5.3.1 of Kushner and Clark (1978). However, (ii) is a stronger condition and implies Eq(2.2), see Kushner and Clark (1978), pp 28-29.

Remark 3 Referring to Kushner and Clark (1978), page 51, conditions (i) and (ii) of Proposition 1 hold also for the basic (two sided) finite difference stochastic approximation (FDSA). Adjusting the G_k to the component-wise perturbation of parameters for gradient approximations, it then follows that the same convergence proof holds for the projection FDSA.

3 Illustrative Example

We study a simple numerical example of finding the optimal temperature profile in a tubular reactor for two first-order irreversible consecutive reactions. See Fan (1966) for details. The first-order reactions $A \longrightarrow B \longrightarrow C$ take place in the reactor. The reaction $A \longrightarrow B$ has the specific rate $k_1(t)$ and $B \longrightarrow C$ has the rate $k_2(t)$ at time t. Denoting the concentration of A by $x_1(t)$ and the concentration of B by $x_2(t)$, we arrive at the following state-space equation which describes the dynamics of the reactions (Fan (1966))

$$\begin{aligned} \dot{x}_1(t) &= -k_1(t)x_1(t) \\ \dot{x}_2(t) &= k_1(t)x_1(t) - k_2(t)x_2(t). \end{aligned} \tag{3.1}$$

The specific rates are given by $k_1(t) = k_{10}e^{-E_1/RT(t)}$ and $k_2(t) = k_{20}e^{-E_2/RT(t)}$ where $T(t)$ is the temperature profile (the control variable) and k_{10}, k_{20}, E_1, E_2, and R are constants.

We wish to find the temperature profile that (starting from time

$t_0 = 0$) maximizes $x_2(t_f)$, i.e. the concentration of the product B at $t = t_f$. By selecting a sampling time, the problem becomes a multivariate optimization problem where the temperature values at discrete time points should be determined such that the final concentration of B is maximized. We present solutions both under no constraints and under the situation that the applied profiles should satisfy $335 \leq T(t) \leq 342$. Unlike the (unconstrained) solution given in Fan (1966), our solution is in principle based on trials and experimentations on the system. The trials consist of applying temperature profiles to the system and doing measurements on $x_2(t_f)$ for each applied profile. We'll require no model in order to find the optimum. Neither do we require knowledge of the values of the constants. In this example, we use the presented model for (and only for) simulation, data generation, and testing our procedure.

Let us assume the following numerical values (Fan (1966)): $k_{10} = 0.534 \times 10^{11}$ /min, $k_{20} = 0.461 \times 10^{18}$ /min, $E_1 = 18000$ cal/mole, $E_2 = 30000$ cal/mole, $R = 2$ cal/mole-K°, $t_f = 8$ min, $x_1(0) = 0.8160$ mole/liter, $x_2(0) = 0.2260$ mole/liter.

Let us further assume that the temperature $T(t)$ is constant for $i - 1 \leq t < i$, $i = 1, 2, ..., 8$ (i.e., a piecewise constant input). The ith element of the 8-dimensional optimization parameter θ is equal to $T(t)$ for $i - 1 \leq t < i$.

The following constants are used throughout the example unless otherwise specified. The number of iterations for the SPSA algorithm is 250, the random perturbations are Bernoulli distributed with magnitude one, i.e. $Pr(\Delta_{ki} = \pm 1) = 0.5$ for $i = 1, ..., 8$, and all $k = 1, 2, ...$, and the gain sequences are selected as $a_k = 1000/k^{0.602}$, $c_k = 1/k^{0.101}$. These decay rates for a_k and c_k are empirically found to yield optimal performance for the unconstrained SPSA algorithm in finite sample cases, see Spall (1995). It should however be noted

that the optimal sequences for the constrained case may be quite different, and finding good gain sequences may in general be a difficulty. It is moreover assumed throughout the example that the measured values of $x_2(t_f)$ are corrupted with additive i.i.d. Gaussian noise with standard deviation 0.0005, and finally, the initial temperature profile, $T_{in}(t)$, for the optimizations is chosen to be 342K° at $t = 0$ and to drop 1K° per minute.

In order to determine the true optimal profiles, we use standard techniques which unlike SPSA make use of the model given by Eq(3.1) and assume noise free data. Eq(3.1) can be written as $\dot{x}(t) = A(t)x(t)$ where $x(t)$ is the state vector and $A(t)$ is a piecewise constant matrix ($A(t)$ is constant in the interval $i - 1 \leq t < i$, $i = 1, ..., 8$). The explicit relation between the objective function and the control variable is obtained using $x(t_f = 8) = \exp\{\sum_{j=0}^{t_f-1} A(j)\}x(0)$, and standard optimization algorithms can be applied to find the optimal profiles. We use MATLAB® optimization toolbox functions CONSTR and FMINS (see Grace (1994)) for the constrained and unconstrained cases, respectively.

Now, let us try both the constrained and unconstrained SPSA algorithms to *estimate* the optimal temperature profiles. We define the sets $G_k = \{\theta : 335+c_k \leq \theta_i \leq 342-c_k, i = 1, ..., 8\}$ for the constrained case. For each of the constrained and unconstrained cases, we estimate the optimal profile 500 times (i.e., 500 cross-sections for each algorithm). The obtained estimates are denoted by $\hat{T}_c(t)$ and $\hat{T}_u(t)$, respectively (notice the randomness in the iterates due to measurement noise for SPSA). As expected, all the 500 realizations of $\hat{T}_c(t)$ are restricted within [335,342] (for all $0 \leq t < 8$) while the largest value (among 500 realizations) of $\max_t \hat{T}_u(t)$ is 345.5. For each realization of $\hat{T}_c(t)$ and $\hat{T}_u(t)$, we compute (1) the relative error defined by

$\{\int_0^{t_f}[T^*(t)-T_r(t)]^2dt/\int_0^{t_f}[T^*(t)-T_{in}(t)]^2dt\}^{\frac{1}{2}}$ where $T_r(t)$ and $T^*(t)$ are the relevant realization and true optimal profile (as computed previously), and (2) the noise free value of $x_2(t_f)$ corresponding to the realization (hence randomness in this computed value is only due to randomness in $\hat{T}_c(t)$ and $\hat{T}_u(t)$). By averaging over these computed values, we obtain an average relative error (ARE) and an average final product value (AFP) for both the constrained and unconstrained cases. The results are summarized in Table 1 where OFP denotes the relevant optimal final product value as given by the true optimal profiles.

In order to investigate the effect of the extra error on the gradient approximation (introduced to make the measurements feasible), we estimate the constrained optimal profile 500 times using the projection SPSA algorithm, but we use $\hat{g}_k(\hat{\theta}_k) = g_k^{SP}(\hat{\theta}_k)$. The corresponding ARE and AFP values for this case are 0.1561 and 0.6988 respectively. Comparing the obtained ARE to 0.1819 (see Table 1) indicates improvement, but at the expense of infeasibility of the measurements.

It is also of interest to assess the convergence rate of the constrained algorithm. We estimate the optimal profile 500 times using constrained SPSA with 1000 iterations (same algorithm constants as before) for each cross-section which yields an ARE value of 0.1139. We then use $-\log(0.1819/0.1139)/\log(250/1000) = 0.338$ as an assessment for the convergence rate. Using Proposition 2 of Spall (1992), the (asymptotic) convergence rate of the unconstrained algorithm for the gain sequences of this example is equal to 0.2 which is considerably less than the computed rate 0.338.

Finally, we apply the constrained two-sided FDSA algorithm 500 times with the same algorithm constants as for the constrained SPSA, but 32 iterations for each cross-section. Notice that the total num-

ber of measurements for the FDSA algorithm with 32 iterations is equal to $32 \times 2 \times 8 = 512$ which is slightly larger than the total number of measurements for the SPSA algorithm with 250 iterations ($250 \times 2 = 500$). The ARE and AFP values become 0.2117 and 0.6988. The ARE value for the constrained FDSA is noticeably larger than 0.1819 obtained for the constrained SPSA algorithm for (almost) the same number of measurements.

It should be noted that a formal treatment of the convergence rate and accuracy of the estimate of the constrained SPSA algorithm is required before one is able to draw any definitive conclusion about the behavior of the algorithm.

Constrained			*Unconstrained*		
ARE	AFP	OFP	ARE	AFP	OFP
0.1819	*0.6988*	*0.6989*	*0.3291*	*0.6996*	*0.6999*

Table 1: *Constrained and unconstrained optimization using 500 cross-sections of SPSA. All the final product values are based on noise free evaluations of* $x_2(t_f)$.

4 Concluding Remarks

The paper presents a projection algorithm for constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation where no gradient information is directly available. The algorithm can handle inequality constraints given as explicit functions of the parameter. The constraints should define a set with non-empty interior. We have considered the case where measurements outside the constraint set are not feasible which is stronger than restricting the solution to the feasible domain. We have estab-

lished almost sure convergence of the iterate to a Kuhn-Tucker point.

Possible directions for future study are the performance of the algorithm, distribution or convergence rate of the iterate, possible error bounds on the estimate, and optimal tuning of the algorithm constants, i.e. optimal selection of gain sequences. Finally, an identical proof of convergence can be applied to a projection FDSA algorithm (see Remark 3). It will be of interest to compare the number of measurements that constrained SPSA and constrained FDSA need to reach a certain level of accuracy (see Section 1 and Spall (1992) for a similar comparison in the unconstrained case).

References

Cauwenberghs, G. (1994). *Analog VLSI Autonomous Systems for Learning and Optimization.* Ph. D. thesis, Dept of Electrical Engineering, California Institute of Technology.

Chin, D. C. (1994). A more efficient global optimization based on Styblinski and Tang. *Neural Nets. 7*, 573–574.

Fan, L. T. (1966). *The Continuous Maximum Principle.* John Wiley & Sons, New York.

Grace, A. (1994). *Optimization Toolbox for Use with MATLAB®.* The Math Works Inc.

Kushner, H. J. and D. S. Clark (1978). *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer-Verlag, Berlin.

L'Ecuyer, P. and P. W. Glynn (1994). Stochastic optimization by simulation: Convergence proofs for the GI/G/1 queue in steady-state. *Management Science 40*(11), 1562–1578.

Ljung, L., G. Pflug, and H. Walk (1992). *Stochastic Approximation and Optimization of Random Systems.* Birkhäuser, Berlin.

Maeda, Y., H. Hirano, and Y. Kanata (1995). A learning rule of neural networks via simultaneous perturbation and its hardware implementation. *Neural Nets. 8*, 251–259.

Parisini, T. and A. Alessandri (1995). Non-linear modeling and state estimation in a real power plant using neural networks and stochastic approximation. In *Proc. American Control Conference*, pp. 1561–1567.

Rezayat, F. (1995). On the use of an SPSA-based model free controller in quality improvement. *Automatica 31*, 913–915.

Spall, J. C. (1987). A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proc. American Control Conference*, pp. 1161–1167.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control 37*(3), 332–341.

Spall, J. C. (1995, September). Implementation of simultaneous perturbation algorithm for stochastic optimization. Submitted to *American Statistician.*

Optimal Sensor Configuration for Complex Systems [1]

Payman Sadegh* and **James C. Spall+**

*Institute of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

+ The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099, USA.

In *Proc. of the Test Technology Symposium* (1996). U.S. Army Test and Evaluation Command. In press.

[B4]

[1]The first author's work was partly supported by the Danish Research Academy, grant S950029, during his stay at JHU/APL. James Spall's work is supported by U.S. Navy contract N00039-95-C-0002 and the JHU/APL IRAD program. The authors would like to thank Prof. James Spicer, JHU, Dept. of Material Science, for his valuable comments and assistance regarding the I-beam experiments, and Dr. Jane Spicer, JHU/APL, for her valuable comments.

Abstract

The paper considers the problem of sensor configuration for complex systems with the aim of maximizing the useful information about certain quantities of interest. Our approach involves two fundamental contributions: (1) definition of an appropriate optimality criterion or performance measure, and (2) description of an efficient and practical algorithm for achieving the optimality objective. The criterion for optimal sensor configuration is based on maximizing the overall sensor response while minimizing the correlation among the sensor outputs, so as to minimize the redundant information being provided by the multiple sensors. The procedure for sensor configuration is based on the powerful simultaneous perturbation stochastic approximation (SPSA) algorithm. SPSA avoids the need for detailed modeling of the sensor response by simply relying on observed responses as obtained by limited experimentation with test sensor configurations. We will illustrate the approach with the optimal placement of acoustic sensors for signal detection in structures. This includes both a computer simulation study for an aluminum plate, and real experimentations on a steel I-beam.

Key words: Optimal sensor configuration, Detection, Complex systems, Stochastic approximation, SPSA, Acoustic sensors and signals.

1 Introduction

In many applications of science and technology, it is necessary to draw inference on a system based on data collected by experimentation with the system. Examples of such situation are estimation of a set of unknown parameters, detection of abrupt changes, and hypothesis testing. In this connection, it is often of interest to investigate the optimal experimental conditions so as to maximize the amount of useful information in the collected data. In this paper, we are concerned with the optimal configuration of a number of sensors for a system, so as to maximize the obtained useful information about an operating condition of interest. Configuration of sensors encompasses placement of sensors on an object, adjustment of sensor operating conditions (such as frequency response, potentiometer settings, pressure sensitivities), sensor orientations at fixed locations, etc. Similar to any other general optimization problem, our solution requires the following two steps. (1) Selection of a criterion or performance measure, and (2) selection of an algorithm to carry out the optimization task.

The choice of criterion is often affected by the amount of available *a priori* knowledge of the system and the design objectives. Also, the choice of optimization technique relies heavily upon the level of available *a priori* knowledge. Here, we are interested in complex systems where the prior knowledge is sparse or too complex to readily offer useful models for solving the sensor configuration problem. More specifically, for any sensor configuration, the sensors responses under the operating condition of interest are generated by an *unknown* random process. Instead, we assume that it is possible to fix the sensors at a configuration of interest, and experiment with the real system or a prototype to generate a realization of the process underlying

the sensors responses for that sensor configuration. There are many situations where such experimentations are possible. Here are two important examples:

Nondestructive evaluation of structures: Acoustic emission (AE) signal recording has been widely acknowledged as a powerful tool for nondestructive evaluation of structures (see e.g. Miller and McIntire (1988), Grabec, Sachse, and Govekar (1991)). A number of acoustic sensors are mounted on the surface of a structure to provide measurements of acoustic emission signals caused by crack formation in the structure. The sensor data can therefore be used to detect a forming crack. It is of interest to locate the acoustic sensors so as to maximize the information about the generated AE signals. Obviously, the received AE signals for any sensor placement are dependent upon the crack location, and there are infinitely many locations in the structure where a crack may form. Regarding the location for a possible forming crack as a random component, the AE signals received by any sensor configuration may be regarded as realizations of some random process. The process is moreover unknown since it is not readily clear how the randomness in crack location translates into the randomness in response (generated AE signal). A possible experimental setup for sensor placement consists of laser induced nondestructive simulation of AE cracking events at possible locations, and recordings of the generated AE signals. We discuss this application further in Section 4.

Biomedicine: Many physiological experiments involve the application of repeated stimuli to a system whose response is recorded and determined. For example, consider the recording of evoked potentials on the scalp where the evoked potentials are due to the application of stimuli. The recorded response amplitude is comparable to the noise level (EEG background noise, electrode noise, electri-

cal noise, etc) which makes response determination difficult, see e.g. Challis and Kitney (1990). It may then be of interest to study an appropriate placement and calibration of the electrodes on the scalp in order to obtain good measurements of the evoked potentials that are related to certain types of diseases. The evoked potentials may be regarded as realizations of some random process because of response variabilities due to the patient effect, the disease effect, etc. These factors affect the response in a way unknown to the experimenter. The relevant experimentations for optimal electrode configuration may involve recordings of the evoked potentials on the scalp of selected patients having the diseases of interest.

The relevant experimentations may be possible in a variety of other applications such as routine monitoring of a system, vision based quality monitoring of a production, and vibration monitoring of an object.

Assuming the availability of relevant data, one may alternatively suggest experimental modeling of the response prior to the design of optimal sensor configuration. This may include (a) to obtain a stochastic model to explain the sensor response for each sensor configuration, or (b) to build up a stochastic model for the sensor response as a function of the sensor configuration. The second approach is similar to response surface methods in the area of experiment design, see e.g. Montgomery (1990).

The obvious advantage of the above approaches relative to those relying upon complex or oversimplified *a priori* models for response determination, is their reliance upon data. However, both (a) and (b) attempt to solve the sensor configuration problem *indirectly*, using demanding prior experimental modeling. In particular, the approach (a) requires prior experimentations for all possible sensor configurations. However, in many systems of interest, possible measurement

configurations comprise an infinite set. Even a finite approximation to this set leaves the experimenter with the demanding task of data collection for many configurations, and the resulting nontrivial tasks of model building and sensor configuration optimization based on the fitted models. Also, the modeling task of the approach (b) is by no means trivial. It may in general involve nonlinear function approximators, and upon completion of demanding experimental procedures, it remains to employ the obtained model within some optimization procedure to obtain the best sensor configuration. Such approaches are practically prohibitive in many complex systems. Moreover, the detailed modeling procedures for these approaches are problem dependent, and therefore can not be treated within a unified framework.

The approach here on the other hand, is based on a very efficient experimental scheme, and provides a framework to address a large variety of sensor configuration problems. It resembles the experimental approaches as described above, from the point of view of relying upon data, rather than complex, uncertain, or oversimplified *a priori* models. However, the approach seeks to solve the sensor configuration problem *directly* without requiring modeling as an intermediate stage. For any sensor configuration, the direct approach requires the possibility of computing a value for the criterion from the recorded response. The first contribution of the paper involves selection of a suitable criterion to be used within the direct approach. Because of the inherent response randomness, any function evaluation provides a *noisy* value of the criterion. The next step involves the selection of a suitable optimization algorithm which is able to yield the optimal configuration, solely on the basis of the noisy evaluations of the criterion.

The concept of optimization using techniques that only rely upon (noisy) evaluations of the criterion is not new. The traditional finite

difference stochastic approximation (FDSA) which is a stochastic approximation (SA) technique in the Kiefer-Wolfowitz setting (Kiefer and Wolfowitz (1952)), uses finite differencing of the noisy values of the criterion for gradient approximations (Kushner and Clark (1978)). In this way, the optimization parameter (in the context of sensor configuration design, the parameter is the vector of spatial coordinates of the sensors and/or other sensor parameters) is *estimated* using an iterative procedure that consists of gradient approximation and parameter updating. Although FDSA is flexible in the respect that no gradient information is required, the practical implementability of the technique is questionable in high dimensional problems. Recall that at each iteration, the basic two-sided FDSA requires $2p$ function evaluations for gradient approximation where p is the number of parameters being optimized. For example if the location of the sensors is of interest, the number of parameters is equal to the number of sensors times 2 or 3 depending on the relative spatial coordinates. For high dimensional cases, the large number of function evaluations, which involve *real measurements* on the system, limits the usefulness of the FDSA technique.

A relatively new SA algorithm in the Kiefer-Wolfowitz setting, simultaneous perturbation stochastic approximation or SPSA (see Spall (1992)), has made solution to problems of this complexity possible. Similar to FDSA, SPSA approximates the gradient using only noisy evaluations of the criterion. However, in contrast to FDSA, only *two* function evaluations are needed at each iteration to form the gradient approximation where the gradient is approximated by *simultaneous* random perturbation of the parameters. Notice the difference between the simultaneous perturbation gradient approximation and the component-wise perturbations in the finite difference based methods. In Spall (1992), it is shown that under reasonably

general conditions, SPSA and FDSA require the same number of iterations to reach a certain level of accuracy for the parameter estimate. This means that under general conditions, the p-fold savings in the number of measurements per iteration translates directly into a p-fold savings in the *total* number of measurements. Since the SPSA approach requires no detailed modeling information and no demanding experimental procedures, it provides a powerful tool for solving the sensor configuration problems in complex systems.

It is possible to formulate the sensor configuration problem as a problem of experimental design where the design variables are the possible configurations. Fedorov and Muller (1989) for instance treat a problem related to the design of observation networks using experimental design techniques. More rigorous mathematical treatment for such problems can be found in e.g. Ylvisaker (1987). However, the starting point for applying these methods is the availability or construction of a model relating the sensor response to the sensor configuration.

The rest of the paper is organized as follows. Section 2 discusses the choice of optimality criterion. Section 3 offers a brief review of the SPSA algorithm which is central to the experimental methodology of the paper, together with a step by step guide to implementation of the algorithm for the sensor configuration problem. Section 4 concerns optimal sensor location for signal detection in structures. This includes both computer simulation and real experimental results for an aluminum plate and a steel I-beam respectively. The study has application in the nondestructive evaluation of structures. Finally, Section 5 offers concluding remarks.

2 Criterion for Sensor Configuration

The fact that the sensor response for any sensor configuration is an unknown process, together with the availability of relevant data as discussed in Section 1, is the motivation for exploring experimental approaches for the optimization. Among the experimental procedures, the direct optimization approach using the SPSA algorithm offers a uniquely efficient alternative. The realizability and the relevance of the technique, on the other hand, is conditioned on (and only on) the existence of an appropriate criterion which is computable from response recording. In this section, we discuss a suitable selection for the criterion. It should however be noted that the choice of a criterion to be used within the SPSA based optimization approach is not unique and may be influenced by particular design objectives and further available information.

The formal setting to be considered here is as follows. Let $\{X_\theta(t)\}$, $t = 1, 2, ...$, denote a set of responses received by the sensors with configuration θ under the operating condition of interest. Each element of the sequence, $X_\theta(t)$, is a N-dimensional vector where N is the number of sensors. The sequences $\{X_\theta(t)\}$ are realizations of some unknown random process, and a realization of the process can be generated by experimentation with the system. We seek a configuration θ that provides good measurements of all possible realizations of the process. We introduce the following criterion:

$$J_0(\theta) = E\{(\det\{\sum_t X_\theta(t)X_\theta(t)^\top\})^{1/N}\}$$

where the mean value $E\{\cdot\}$ is with respect to the process generating $\{X_\theta(t)\}$, and the summation is taken over a time window of interest.

The intuitive rationale for selecting the criterion is based on the

properties of its mean argument, i.e. $(\det\{\sum_t X_\theta(t)X_\theta(t)^\top\})^{1/N}$, as follows.

1. Defining the overall response of a sensor as the sum of squares of the sensor responses over the time window of interest, it is evident that the diagonal elements of $\sum_t X_\theta(t)X_\theta(t)^\top$ account for the overall response of the sensors. The determinant of a positive semi-definite matrix increases with its diagonal elements.
2. The off-diagonal elements of $\sum_t X_\theta(t)X_\theta(t)^\top$ account for the correlation among the sensors responses. The determinant of a positive semi-definite matrix decreases with large in magnitude off-diagonal elements.
3. The $1/N$ exponent scales the units properly such that the criterion is measured in the same physical units as the overall response.
4. The quantity $(\det\{\sum_t X_\theta(t)X_\theta(t)^\top\})^{1/N}$ is readily computable given that a response realization $\{X_\theta(t)\}$ is available. This computed value obviously gives a *noisy* unbiased evaluation of $J_0(\theta)$.

Further Motivation for the Criterion: Assume that the measurement of $X_\theta(t)$ (response) takes place in the presence of additive mean zero noise which is uncorrelated with the process generating $\{X_\theta(t)\}$. In many applications, it is of interest to consider the quadratic expression

$$\sum_t X_\theta(t)^\top \Sigma^{-1} X_\theta(t) \tag{2.1}$$

where the summation is taken over a time window of interest and Σ is the covariance of the measurement noise which is assumed to be

constant within the time window. For instance, consider the problem where it is of interest to detect occurrence of a realization of the process generating $\{X_\theta(t)\}$ by testing the null hypothesis $X_\theta(t) = 0$ for all t within the time window of interest. Given that the measurement noise terms are (temporally) uncorrelated and Gaussian, then the quantity $\sum_t X_\theta(t)^\top \Sigma^{-1} X_\theta(t)$ is related to the noncentrality parameter of the relevant χ^2-distribution for testing the null hypothesis (see e.g. Kendall and Stuart (1973)).

Despite both the theoretical and the intuitive appeal of the quadratic expression, it can not be directly used for the sensor configuration problem. This is due to the fact that the measurement noise covariance is unknown and time varying over long time periods, and any estimate provided for the covariance can only be expected to hold over certain time intervals. However, we seek a configuration that provides appropriate measurements of all future realizations of the process generating $\{X_\theta(t)\}$, and the measurements may be taken under totally different (measurement) noise environments.

We use the following simple result to derive a lower bound on the value of the quadratic expression with respect to the unknown noise covariance. For positive definite $N \times N$ matrices A and M, it holds that

$$\min_{\det A \geq d} \operatorname{trace}(AM) = N d^{1/N} (\det(M))^{1/N}. \tag{2.2}$$

The result is obtained in a straight forward way by writing the Lagrangian for the optimization problem (see Fedorov and Khabarov (1986)). Now note $\sum_t X_\theta(t)^\top \Sigma^{-1} X_\theta(t) = \operatorname{trace}[\Sigma^{-1} \sum_t X_\theta(t) X_\theta(t)^\top]$ which together with (2.2) yield that under the assumption of finite positive definite Σ, the lower bound (with respect to Σ) on the quadratic expression is proportional to $\det[\sum_t X_\theta(t) X_\theta(t)^\top]^{1/N}$. This

further motivates the choice of $J_0(\theta)$.

Although we have assumed that response measurements are typically taken in the presence of noise, it might be possible to perform limited experiments for selecting the optimal sensor configuration under controlled (large signal to noise ratio or measurement noise free) conditions. In these cases, a noisy evaluation of $J_0(\theta)$ is obtained by fixing the sensors at configuration θ, generating a realization $\{X_\theta(t)\}$, and computing $(\det\{\sum_t X_\theta(t)X_\theta(t)^\top\})^{1/N}$ using the measured responses. However, when it is not even possible to perform such controlled experiments for sensor configuration design, the matrix $\sum_t X_\theta(t)X_\theta(t)^\top$ (to be used for function evaluations) can only be estimated from the response recordings. An estimate is easily obtained by $\sum_t (Y_\theta(t)Y_\theta(t)^\top - \hat{\Sigma})$ where $\{Y_\theta(t)\}$ denotes the obtained sensor data under the generation of $\{X_\theta(t)\}$ and $\hat{\Sigma}$ denotes an estimate for the *current* measurement noise covariance. The covariance estimate can for example be computed based on data collected prior to the generation of $\{X_\theta(t)\}$. Computing $[\det\{\sum_t (Y_\theta(t)Y_\theta(t)^\top - \hat{\Sigma})]^{1/N}$ gives a noisy unbiased evaluation of

$$E\{[\det(\sum_t Y_\theta(t)Y_\theta(t)^\top - \hat{\Sigma})]^{1/N}\} \tag{2.3}$$

where the mean value $E\{\cdot\}$ is with respect to the recorded data. The expression (2.3) approximates $J_0(\theta)$ since a realization of $\sum_t X_\theta(t)X_\theta(t)^\top$ is estimated rather than being directly computed.

Finally, we wish to emphasize again that the direct approach using the SPSA algorithm *is not* uniquely related to the criterion derived in this section. Any criterion which is a suitable performance measure, and is computable from response recordings can be used within the experimental methodology of the paper as follows in the

next section.

3 Overview of the SPSA algorithm and the Experimental Methodology

3.1 Overview of the SPSA algorithm

Since the SPSA algorithm plays a central role in the experimental methodology, this subsection provides a brief overview of the generic technique.

Consider the problem of determining a parameter $\theta \in \mathbb{R}^p$ that maximizes a differentiable objective function $J(\theta)$, where the explicit dependence of the loss function upon θ is unknown, but for each θ a noisy value for the objective function can be obtained (i.e., no gradient information is directly available). The SPSA algorithm has recently attracted considerable attention for challenging optimization problems of this type in application areas such as adaptive control, pattern recognition, discrete event systems, neural network training, and model parameter estimation, see e.g., Rezayat (1995), Maeda, Hirano, and Kanata (1995), Hill and Fu (1995), Cauwenberghs (1994), Chin (1994), and Parisini and Alessandri (1995). In this section, we briefly present the SPSA algorithm. The reader is referred to Spall (1992) for a detailed treatment.

Let $\hat{\theta}_k$ denote the estimate for the parameter θ at the kth iteration and $g(\theta)$ denote $\partial J/\partial\theta$. The SPSA algorithm has the form

$$\hat{\theta}_{k+1} = \hat{\theta}_k + a_k \hat{g}_k(\hat{\theta}_k)$$

where the gain sequence $\{a_k\}$ satisfies certain conditions (as follows) and $\hat{g}_k(\hat{\theta}_k)$ is a simultaneous perturbation approximation to $g(\hat{\theta}_k)$ at

iteration k. We define the simultaneous perturbation estimate for the gradient as follows. Let $\Delta_k \in \mathbb{R}^p$ be a vector of p mutually independent mean zero random variables $\{\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kp}\}$ satisfying certain conditions (see Lemma 1 and assumption A2 of Spall (1992)). Consistent with the usual framework of stochastic approximation, we have noisy evaluations of the objective function. In particular, at the kth iteration, consider the two noisy function evaluations

$$\begin{aligned} j_k^{(+)} &= J(\hat{\theta}_k + c_k\Delta_k) + \epsilon_k^{(+)} \\ j_k^{(-)} &= J(\hat{\theta}_k - c_k\Delta_k) + \epsilon_k^{(-)} \end{aligned}$$

where $\{c_k\}$ is a gain sequence, and $\epsilon_k^{(+)}$ and $\epsilon_k^{(-)}$ represent noise terms that satisfy $E\{\epsilon_k^{(+)} - \epsilon_k^{(-)} | \hat{\theta}_k, \Delta_k\} = 0$. The gain sequences $\{a_k\}$ and $\{c_k\}$ are positive for all k and tend to zero as $k \to \infty$. Moreover, $\sum_{k=0}^{\infty} a_k = \infty$, $\sum_{k=0}^{\infty} (a_k/c_k)^2 < \infty$.

The basic simultaneous perturbation form for the estimate of $g(\cdot)$ at the kth iteration is then

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} \frac{j_k^{(+)} - j_k^{(-)}}{2c_k\Delta_{k1}} \\ \vdots \\ \frac{j_k^{(+)} - j_k^{(-)}}{2c_k\Delta_{kp}} \end{bmatrix}. \tag{3.1}$$

Note that at each iteration, only *two* function evaluations are needed to form the estimate regardless of the number of parameters.

In Spall (1992), it is shown that under fairly general conditions, the SPSA iterates converge almost surely to the true optimum. The same reference derives the asymptotic distribution of the iterate, which under reasonably general conditions is shown to be Gaussian. This result is important for quantifying the accuracy of the estimate. Spall (1992) also compares the asymptotic behavior of SPSA and

FDSA both theoretically and through numerical studies. The results indicate that under fairly general conditions, SPSA requires $1/p$ the number of measurements needed by FDSA, in order to achieve a specified level of accuracy (in terms of the asymptotic mean square error of the estimate). Finally, Sadegh (1996) treats constrained optimization via SPSA and establishes almost sure convergence of a projection SPSA algorithm to a Kuhn-Tucker point.

3.2 The Experimental Methodology for Sensor Configuration

We have earlier in Section 1 discussed benefits and drawbacks of possible solutions to the sensor configuration problem, and discussed the efficiency of a direct SPSA based optimization approach. The key words we used to advocate the SPSA approach in Section 1, namely, no need for detailed modeling information and no need for demanding experimental procedures, are brought into a more transparent form in Subsection 3.1. Firstly, SPSA requires no direct gradient information. Secondly, the comparison results to FDSA indicate no loss of accuracy despite the p-fold reduction in the total number of measurements. A third argument to the benefit of the SPSA approach is its ease of use. To illustrate this feature, we here outline the general experimental methodology of optimal sensor configuration design.

In the following, we let $J(\theta)$ refer to $J_0(\theta)$ when a realization of $J_0(\theta)$ can be directly computed under sensor configuration experiments, or the expression given by (2.3) when a realization of $J_0(\theta)$ can only be estimated (see Section 2). Our aim is to maximize $J(\theta)$ with respect to sensor configuration θ, where for each θ a noisy unbiased evaluation of $J(\theta)$ is available. Prior to the implementation of the procedure, we have to select a set of algorithm constants. This in-

cludes selecting the gain sequences $\{a_k\}$ and $\{c_k\}$, and the probability distribution for random perturbations $\mathcal{P}_\Delta$. For guidelines regarding the choice of gain sequences see Spall (1995), and for the choice of probability distribution, see Sadegh and Spall (1995).

Initialization: Select a sensor configuration $\hat{\theta}_0$ using prior knowledge or simply by randomization. For $k = 0, 1, 2, ...$ repeat:

Step 1: By Monte-Carlo simulation, generate a sequence of independent random numbers $\Delta_{ki} \in \mathcal{P}_\Delta$ and form the random perturbation vector $\Delta_k = (\Delta_{k1}, ..., \Delta_{kp})^\top$.

Step 2: Configure the sensors at $\hat{\theta}_k \pm c_k\Delta_k$ respectively. For each one of the two configurations, obtain a noisy evaluation of $J(\theta)$ (see Section 2).

Step 3: Use the two noisy evaluations of Step 2, and the random perturbation vector Δ_k of Step 1, to obtain $\hat{g}_k(\hat{\theta}_k)$, i.e. the simultaneous perturbation approximation to $g(\hat{\theta}_k)$ at iteration k, see (3.1).

Step 4: Update to a new configuration $\hat{\theta}_{k+1} = \hat{\theta}_k + a_k\hat{g}_k(\hat{\theta}_k)$.

The above iterations are typically repeated until a suitable configuration is obtained.

In many practical situations, possible sensor configurations are subject to restrictions. Sadegh (1996) presents a constrained SPSA algorithm which can be used for these cases. The modifications for implementing this type of algorithm are minor, including projections to enforce the given constraints.

4 Application: Signal Detection in Complex Structures

In this section, we consider an application of the procedure to sensor placement for signal detection in complex structures. A very important problem in this connection is the nondestructive evaluation of structures by acoustic signal sensing. The use of acoustic emission (AE) signal sensing has recently received considerable attention in problems related to the nondestructive evaluation, see e.g. Miller and McIntire (1988). As a result of crack formation in a structure, acoustic emission signals are generated and propagated throughout the structure. It is then possible to detect a forming crack by doing acoustic signal measurements using a number of acoustic sensors placed on the surface of the structure. As a result of structure complexity and material and geometric irregularities in many practically important objects such as highway bridges, it is very difficult or impossible to develop mathematical models to determine the sensors responses. The lack of reliable mathematical models together with convenient availability of data, including laser induced simulations of AE cracking events point to the experimental methodology of the paper as a promising tool for solving this challenging sensor placement problem (see also the discussions in Section 1). In this section, we present an application of the methodology of the previous sections to a simulated aluminum plate (Subsection 4.1) and an actual steel I-beam (Subsection 4.2).

4.1 Simulation Results for Sensor Placement on a Plate

Here, we present a computer simulation based optimization of the sensor location for detection of impulse pressure inputs to a plate.

We apply the developed procedure and find the optimal Cartesian coordinates of 10 sensors placed on the surface of a $1m \times 1m$ plate (hence the number of optimization parameters is equal to 20). Notice that in a real application, the experimental procedure typically involves *real* experimentations and not computer simulations. In this spirit, the optimization procedure will have no knowledge of the plate equations implemented within the simulation. For the simulation purpose, we use the following partial differential equation (see Graff (1975))

$$(Eh^3/12(1-\nu^2))\nabla^4 w(x,y,t) + \rho h \frac{\partial^2 w(x,y,t)}{\partial t^2} = q(x,y,t), \quad (4.1)$$

where x and y denote respectively the horizontal coordinate and the vertical coordinate, $\nabla^4 = (\frac{\partial^4}{\partial x^4} + 2\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4})$ is the biharmonic operator, $w(x,y,t)$ denotes the deflection, $q(x,y,t)$ denotes the input pressure, E is the Young module, ρ is the density of the plate, h is the thickness of the plate, and finally $0 < \nu < 1$ is the Poisson ratio (it relates the orthogonal strain-stress couples). The following constants are used in the simulations: $E = 69.0 \times 10^9$ N/m^2, $\rho = 2.7 \times 10^3$ kg/m^3 (these values of E and ρ are typical for an aluminum plate, see any relevant standard table), $\nu = 0.01$, $h = 2.0 \times 10^{-5}$ m. The plate is subject to clamped edge boundary conditions, i.e. for all t, $w(x,y,t) = 0$ and $\partial w(x,y,t)/\partial t = 0$ along $x = 0$, $x = 1$, $y = 0$, and $y = 1$.

The input pressure is taken to be

$$q(x,y,t) = \begin{cases} 10^4 \delta(t) \text{ N/m}^2 & x = x_0, y = y_0 \\ 0 & \text{Otherwise} \end{cases}$$

The prior knowledge of $(x_0, y_0)^\top$ is embedded in a distribution obtained by truncating a $N((0.5,0.5)^\top, 0.025^2 I)$ distribution (where I as usual denotes the unity matrix) such that $Pr(0.1 \leq x, y \leq 0.9) = 1$.

Finally, each sensor records the deflection of the structure at the location point. The sampling time of data is 0.05 seconds for all the sensors.

For the SPSA algorithm, we select $a_k = 0.1/k^{0.602}$, $c_k = 0.1/k^{0.101}$. The initial placement, $\hat{\theta}_0$, for implementing the algorithm is obtained by randomly distributing the sensors over the square $0.1 \leq x, y \leq 0.9$.

For each placement of the sensors, a function evaluation includes the following. We randomly draw an input location according to the probability distribution for the possible input locations, solve the partial differential equation (4.1) numerically to find the sensor response for each sensor, and form the response (deflection) sequence $\{X_\theta(t)\}$, $t = 0.05, ..., 50 \times 0.05$, where the element i of $X_\theta(t)$ is equal to the response of sensor i at time t. A realization of $J_0(\theta)$ is then obtained by computing $(\det\{\sum_t X_\theta(t)X_\theta(t)^\top\})^{1/N}$. This case corresponds to a situation where it is possible to obtain noise free response recordings under limited controlled experiments for sensor configuration design.

We implement the experimental procedure of Subsection 3.2 with 250 iterations and denote the resulting sensor placement by $\hat{\theta}_{250}$. The initial and final placement of the sensors are plotted in Figure 1 for comparison. Note the shift of the sensors towards the mean value for (x_0, y_0) (signal generation center).

We fix the sensors at $\hat{\theta}_{250}$ and $\hat{\theta}_0$ respectively, do 100 function evaluations for each one of the sensor placements, and average over the 100 computed values. The obtained average values are equal to 0.0517 and 0.0309 respectively. We repeat the same procedure for two randomly selected sensor placements, θ_1 and θ_2. The placement θ_1 is obtained by randomization over $0.4 \leq x, y \leq 0.6$, and θ_2 is obtained by randomization over $0.2 \leq x, y \leq 0.8$. The reason we select θ_1 and θ_2 is to investigate sensor placements where the sensors are close

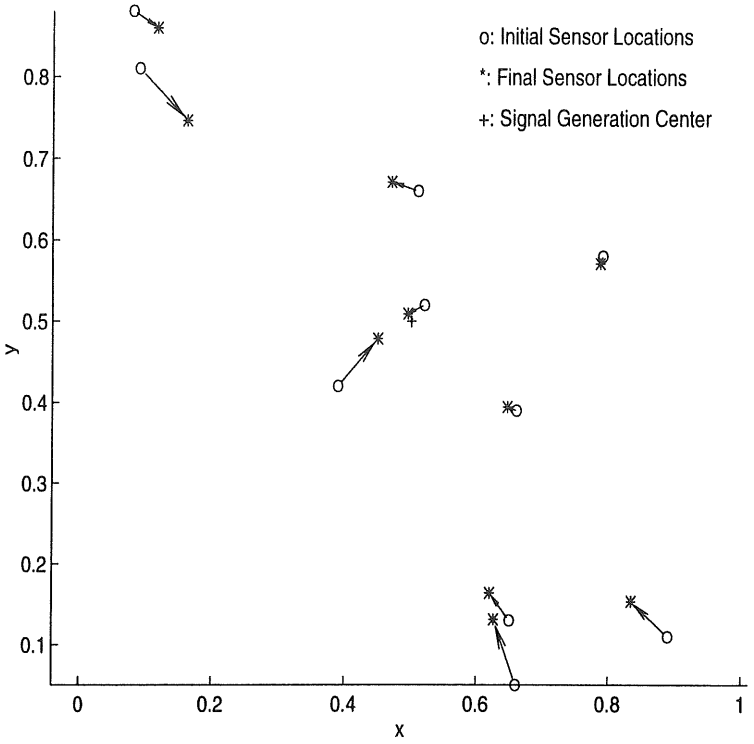


Figure 1: *Initial and final placement of the sensors on the plate.*

to the signal generation center but provide redundant information due to being densely located. This is especially the case for θ_1. The average function values obtained for θ_1 and θ_2 are $\simeq 0$ and 0.0138 respectively.

Firstly, notice that a considerable increase in the value of the criterion is obtained using the SPSA algorithm with relatively small number of function evaluations. Secondly, we illustrate what this increased value of the criterion means in terms of achieving good

hypothesis testing results. Consider a situation where the future observations of the process generating $\{X_\theta(t)\}$ are taken in the presence of additive noise uncorrelated with the process. We randomly select 25 locations, $L_1, ..., L_{25}$, according to the probability distribution for the input location. Our goal here is to compare the quality of the four sensor placements $\hat{\theta}_{250}$, $\hat{\theta}_0$, θ_1, and θ_2, in terms of detecting the signals generated from the 25 selected locations under different noise environments. We let the noise covariance Σ vary. For each noise covariance, we carry out the following procedure for each one of the sensor placements $\hat{\theta}_{250}$, $\hat{\theta}_0$, θ_1, and θ_2:

- For the locations $L_1, ..., L_{25}$, simulate the PDE given by (4.1) to obtain $\{X_\theta(t)\}$, $t = 0.05, ..., 50 \times 0.05$. For each one of the 25 simulated sequences $\{X_\theta(t)\}$, repeat ten times:
  - Generate 50 independent realizations of a $N(0, \Sigma)$ variable to obtain the noise sequence. Add the noise sequence to the response sequence to obtain the data sequence.
  - Use the data sequence to compute the T^2 test quantity to test the null hypothesis $X_\theta(t) = 0$, $t = 0.05, ..., 50 \times 0.05$ (see Kendall and Stuart (1973)). Make a binary decision upon acceptance or rejection of the null hypothesis, based on the calculated test quantity where the significance level of the tests is 5% (rejection of the hypothesis implies detection of the signal).
- Determine the number of detections among the ten repetitions for each input location.
- Average over the 25 computed values obtained in the previous step to obtain an average number of detections.

The results indicate robust performance for $\hat{\theta}_{250}$ while θ_1 and θ_2 perform well for some noise covariances and poorly for others. For two

specific covariances Σ_1 and Σ_2 where Σ_1 is nearly diagonal and Σ_2 has large positive off-diagonal elements, we have the following results. For Σ_1, the average number of detections for $\hat{\theta}_{250}$, $\hat{\theta}_0$, θ_1, and θ_2, are equal to 9.1, 7.5, 10, and 8.5. The corresponding numbers for Σ_2 are 6.2, 4.8, 0.8, and 3.2. These results confirm the relationship between the criterion $J_0(\theta)$ and the lower bound on the quadratic expression (2.1) as discussed in Section 2.

4.2 Small Scale I-beam Experiments

Here, we present results obtained by real experimentations on a steel I-beam. Because of practical laboratory constraints, our experimentation is limited to locating 3 acoustic sensors on the center line of the I-beam, i.e. the number of optimization parameters is equal to 3. The I-beam has a length of approximately 120 cm and a height of approximately 15 cm. The acoustic sensors transform a mechanical deflection to an electrical voltage.

To simulate AE cracking events, we use high energy laser with an energy varying within approximately 10% of the tuned energy level. We consider a situation where AE events occur with equal likelihood within an approximately 5 cm long line piece along the center line and around the center of the I-beam (i.e., a uniform distribution for the location of AE events is used). Our data collection/processing equipments consist of an oscilloscope where the outputs of the transducers are recorded (in mv), and a computer where the oscilloscope data are down-loaded and SPSA iterations are performed using MATLAB® software.

We apply the experimental procedure using 20 iterations of SPSA and obtain the placement $\hat{\theta}_{20}$. The SPSA constants are selected as $a_k = 0.1/k^{0.602}$ cm/mv^2, $c_k = 0.01/k^{0.101}$ cm. Figure 2 shows the ini-

tial and final placements of the sensors. It is interesting to note that the final placement is further away from the mean AE source but the sensors are also further apart, reducing redundant information. This can be justified considering the fact that steel is known to have small signal attenuation coefficient (see ASM (1989)). Therefore, it is more important to reduce the redundant information rather than to locate the sensors close to the mean AE source.

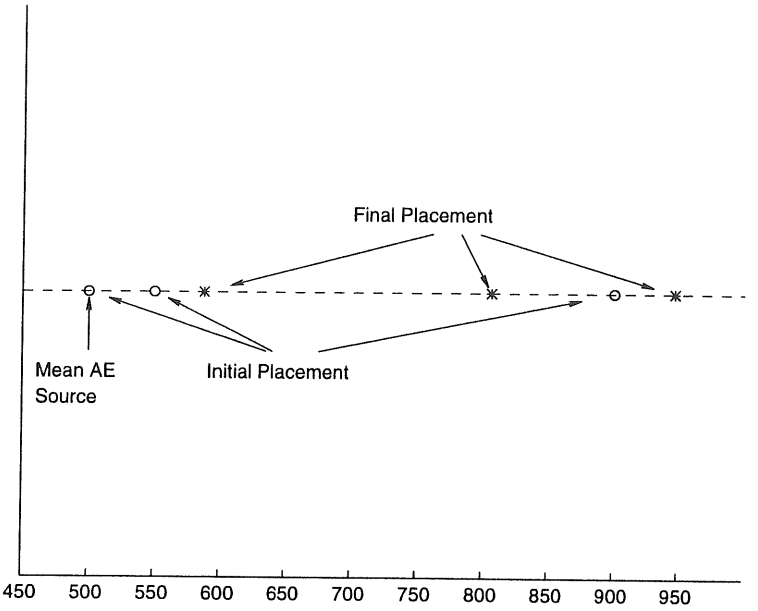


Figure 2: *Initial and final placement of the sensors on the steel I-beam.*

5 Conclusion

The paper studies the problem of optimal configuration of a number of sensors for a system. The presented approach is direct and uses experimental data without an intermediate step of modeling the sensor response. The approach is easily implementable and does not suffer from modeling inaccuracies. The SPSA algorithm provides a uniquely powerful tool and plays a central role within the experimental methodology of sensor configuration design. We have demonstrated the approach on the problem of sensor placement for signal detection in structures, using both computer simulations and real experiments.

References

ASM (1989). Nondestructive evaluation and quality control. In *Metals Handbook*, Volume 17. American Society for Metals.

Cauwenberghs, G. (1994). *Analog VLSI Autonomous Systems for Learning and Optimization.* Ph. D. thesis, Dept of Electrical Engineering, California Institute of Technology.

Challis, R. E. and K. I. Kitney (1990, November). Biomedical signal processing, Part I. *Medical & Biological Engineering & Computation 28*, 509.

Chin, D. C. (1994). A more efficient global optimization based on Styblinski and Tang. *Neural Nets. 7*, 573–574.

Fedorov, V. and V. Khabarov (1986). Duality of optimal designs for model discrimination and parameter estimation. *Biometrika 73*, 183–190.

Fedorov, V. and Muller (1989). Comparison of two approaches in the optimal design of an observation network. *Statistics 20*(3),

339.

Grabec, I., W. Sachse, and E. Govekar (1991). Solving AE problems by neural networks. In W. Sache, J. Roget, and K. Yamaguchi (Eds.), *Acoustic Emission: Current Practice and Future Directions*, pp. 165–182. American Society for Testing and Materials, Philadelphia. ASTM STP 1077.

Graff, K. F. (1975). *Wave Motion in Elastic Solids.* Oxford Engineering Science Series. Oxford, Claredon.

Hill, S. D. and M. C. Fu (1995). Transfer optimization via simulation perturbation stochastic approximation. In *Proc. Winter Simulation Conference*, pp. 242–249.

Kendall, M. G. and A. Stuart (1973). *The Advanced Theory of Statistics*, Volume 2. Charles Griffin & Co., London.

Kiefer, J. and J. Wolfowitz (1952). Stochastic estimation of a regression function. *Ann. Math. Stat. 23*, 462–466.

Kushner, H. J. and D. S. Clark (1978). *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer-Verlag, Berlin.

Maeda, Y., H. Hirano, and Y. Kanata (1995). A learning rule of neural networks via simultaneous perturbation and its hardware implementation. *Neural Nets. 8*, 251–259.

Miller, R. K. and P. McIntire (1988). Acoustic emission handbook. In *Nondestructive Testing Handbook.* ASNT- Columbus- Ohio.

Montgomery, D. C. (1990). *Design and Analysis of Experiments.* John Wiley and Sons, New York.

Parisini, T. and A. Alessandri (1995). Non-linear modeling and state estimation in a real power plant using neural networks and stochastic approximation. In *Proc. American Control Conference*, pp. 1561–1567.

Rezayat, F. (1995). On the use of SPSA-based model free controller in quality improvement. *Automatica 31*, 913–915.

Sadegh, P. (1996). Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation. Technical Report 3/96, Institute of Mathematical Modeling, Technical University of Denmark. See article B3 for a revised version.

Sadegh, P. and J. C. Spall (1995). Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. Submitted to *IEEE Transaction on Automatic Control.* See article B2 for a revised version.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control 37*(3), 332–341.

Spall, J. C. (1995, September). Implementation of simultaneous perturbation algorithm for stochastic optimization. Submitted to *American Statistician.*

Ylvisaker, D. (1987). Prediction and design. *AMS 15*, 1–19.

Concluding Remarks

The contributions of the thesis may be summarized as follows.

- A2 studies the design of optimal input signals for gray-box models, which are characterized by their physical significance and partial prior knowledge. The extension of the classical experiment design theory to optimization of Bayesian criteria is investigated.
- A3 presents a solution to input design for maximizing the smallest eigenvalue of the information matrix in linear dynamic systems. A cutting plane algorithm, based on successive solutions to linear programs is suggested for finding the optimal frequency power weights of the input.
- B2 derives the optimal perturbation distribution for the SPSA algorithm based on the asymptotic distribution of the parameter iterate.
- B3 introduces a constrained SPSA algorithm using projections and proves the convergence of the algorithm.
- B4 presents a solution for optimal sensor configuration in complex systems, together with an application from the area of nondestructive evaluation of structures.

Both A2 and A3 present relatively simple optimization procedures for experiment design. It will be of interest to study further application areas of the techniques and possible modifications of the algorithms under other (practically inspired) experimental constraints.

The result presented in B2 has a major practical implication in that it shows the Bernoulli perturbation distribution form is (asymptotically) optimal *regardless* of the prior knowledge about the form of the loss function. The paper gives formulas and procedures for de-

termining the optimal perturbation variance under different levels of prior knowledge about the system. It will be of major practical and theoretical interest to study the selection of optimal perturbations for finite-sample cases. Also, optimal perturbation for constrained optimization (see B3) is of major interest.

The result presented in B3 is of importance since almost all real world applications of optimization techniques pertain to the class of constrained problems. Possible directions for future study on this subject are the performance of the algorithm, distribution or convergence rate of the iterate, possible error bounds on the estimate, and optimal tuning of the algorithm constants, i.e. optimal selection of gain sequences. It will also be of interest to compare the properties of the constrained SPSA and FDSA algorithms, both formally and through numerical studies.

Finally, the solution presented in B4 (sensor configuration design) may have vast practical implications for sensor configuration problems where no reliable model of the system is available. Similar approaches based on noisy measurements of the objective function have been previously applied to other types of problems such as feedback control, but the approach to the best of my knowledge is new for sensor configuration design. It is of major interest to investigate the application of the procedure to sensor configuration problems arising in areas such as biomedicine, routine monitoring, and quality control. Since the constraints on possible sensor configurations typically arise from geometrical considerations, the resulting problem may be appropriately accommodated within the setting considered in B3 where the inequality constraints are given as explicit functions of the parameter.

Ph. D. theses from IMM

1. **Larsen, Rasmus.** (1994). *Estimation of visual motion in image sequences. xiv* + 143 pp.
2. **Rygard, Jens Moberg.** (1994). *Design and optimization of flexible manufacturing systems. xiii* + 232 pp.
3. **Lassen, Niels Christian Krieger.** (1994). *Automated determination of crystal orientations from electron backscattering patterns. xv* + 136 pp.
4. **Melgaard, Henrik.** (1994). *Identification of physical models. xvii* + 246 pp.
5. **Wang, Chunyan.** (1994). *Stochastic differential equations and a biological system. xxii* + 153 pp.
6. **Nielsen, Allan Aasbjerg.** (1994). *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data. xxiv* + 213 pp.
7. **Ersbøll, Annette Kjær.** (1994). *On the spatial and temporal correlations in experimentation with agricultural applications. xviii* + 345 pp.
8. **Møller, Dorte.** (1994). *Methods for analysis and design of heterogeneous telecommunication networks.* Volume 1-2, *xxxviii* + 282 pp., 283-569 pp.
9. **Jensen, Jens Christian.** (1995). *Teoretiske og eksperimentelle dynamiske undersøgelser af jernbanekøretøjer.* ATV Erhvervsforskerprojekt EF 435. *viii* + 174 pp.
10. **Kuhlmann, Lionel.** (1995). *On automatic visual inspection of reflective surfaces.* ATV Erhvervsforskerprojekt EF 385. Volume 1, *xviii* + 220 pp., (Volume 2, *vi* + 54 pp., fortrolig).
11. **Lazarides, Nikolaos.** (1995). *Nonlinearity in superconductivity and Josephson Junctions. iv* + 154 pp.
12. **Rostgaard, Morten.** (1995). *Modelling, estimation and control of fast sampled dynamical systems. xiv* + 348 pp.
13. **Schultz, Nette.** (1995). *Segmentation and classification of biological objects. xiv* + 194 pp.

14. **Jørgensen, Michael Finn.** (1995). *Nonlinear Hamiltonian systems. xiv* + 120 pp.
15. **Balle, Susanne M.** (1995). *Distributed-memory matrix computations. iii* + 101 pp.
16. **Kohl, Niklas.** (1995). *Exact methods for time constrained routing and related scheduling problems. xviii* + 234 pp.
17. **Rogon, Thomas.** (1995). *Porous media: Analysis, reconstruction and percolation. xiv* + 165 pp.
18. **Andersen, Allan Theodor.** (1995). *Modelling of packet traffic with matrix analytic methods. xvi* + 242 pp.
19. **Hesthaven, Jan.** (1995). *Numerical studies of unsteady coherent structures and transport in two-dimensional flows.* Risø-R-835(EN) 203 pp.
20. **Slivsgaard, Eva Charlotte.** (l995). *On the interaction between wheels and rails in railway dynamics. viii* + 196 pp.
21. **Hartelius, Karsten.** (1996). *Analysis of irregularly distributed points. xvi* + 260 pp.
22. **Hansen, Anca Daniela.** (1996). *Predictive control and identification - Applications to steering dynamics. xviii* + 307 pp.
23. **Sadegh, Payman.** (1996). *Experiment design and optimization in complex systems. xiv* + 162 pp.