Journal of Computational Physics 232 (2013) 288–317

Contents lists available at SciVerse ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

# Simulation-based optimal Bayesian experimental design for nonlinear systems

Xun Huan, Youssef M. Marzouk *

*Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

ARTICLE INFO

*Article history:*
Received 4 August 2011
Received in revised form 12 August 2012
Accepted 13 August 2012
Available online 4 September 2012

*Keywords:*
Uncertainty quantification
Bayesian inference
Optimal experimental design
Nonlinear experimental design
Stochastic approximation
Shannon information
Chemical kinetics

ABSTRACT

The optimal selection of experimental conditions is essential to maximizing the value of data for inference and prediction, particularly in situations where experiments are time-consuming and expensive to conduct. We propose a general mathematical framework and an algorithmic approach for optimal experimental design with nonlinear simulation-based models; in particular, we focus on finding sets of experiments that provide the most information about targeted sets of parameters.

Our framework employs a Bayesian statistical setting, which provides a foundation for inference from noisy, indirect, and incomplete data, and a natural mechanism for incorporating heterogeneous sources of information. An objective function is constructed from information theoretic measures, reflecting expected information gain from proposed combinations of experiments. Polynomial chaos approximations and a two-stage Monte Carlo sampling method are used to evaluate the expected information gain. Stochastic approximation algorithms are then used to make optimization feasible in computationally intensive and high-dimensional settings. These algorithms are demonstrated on model problems and on nonlinear parameter inference problems arising in detailed combustion kinetics.



## 1. Introduction

Experimental data play an essential role in developing and refining models of physical systems. For example, data may be used to update knowledge of parameters in a model or to discriminate among competing models. Whether obtained through field observations or laboratory experiments, however, data may be difficult and expensive to acquire. Even controlled experiments can be time-consuming or delicate to perform. In this context, maximizing the value of experimental data—designing experiments to be "optimal" by some appropriate measure—can dramatically accelerate the modeling process. Experimental design thus encompasses questions of where and when to measure, which variables to interrogate, and what experimental conditions to employ.

These questions have received much attention in the statistics community and in many science and engineering applications. When observables depend linearly on parameters of interest, common solution criteria for the optimal experimental design problem are written as functionals of the information matrix [1]. These criteria include the well-known 'alphabetic optimality' conditions, e.g., *A*-optimality to minimize the average variance of parameter estimates, or *G*-optimality to minimize the maximum variance of model predictions. Bayesian analogues of alphabetic optimality, reflecting prior and posterior uncertainty in the model parameters, can be derived from a decision-theoretic point of view [2]. For instance, Bayesian

* Corresponding author. Tel.: +1 617 253 1337; fax: +1 617 253 7397.
*E-mail addresses:* xunhuan@mit.edu (X. Huan), ymarz@mit.edu (Y.M. Marzouk).

0021-9991/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcp.2012.08.013

*D*-optimality can be obtained from a utility function containing Shannon information while Bayesian *A*-optimality may be derived from a squared error loss. In the case of linear-Gaussian models, the criteria of Bayesian alphabetic optimality reduce to mathematical forms that parallel their non-Bayesian counterparts [2].

For nonlinear models, however, exact evaluation of optimal design criteria is much more challenging. More tractable design criteria can be obtained by imposing additional assumptions, effectively changing the form of the objective; these assumptions include linearizations of the forward model, Gaussian approximations of the posterior distribution, and additional assumptions on the marginal distribution of the data [2]. In the Bayesian setting, such assumptions lead to design criteria that may be understood as *approximations of an expected utility*. Most of these involve prior expectations of the Fisher information matrix [3]. Cruder "locally optimal" approximations require selecting a "best guess" value of the unknown model parameters and maximizing some functional of the Fisher information evaluated at this point [4]. None of these approximations, though, is suitable when the parameter distribution is broad or when it departs significantly from normality [5]. A more general design framework, free of these limiting assumptions, is preferred [6,7].

More rigorous information theoretic criteria have been proposed throughout the literature. The seminal paper of Lindley [8] suggests using expected gain in Shannon information, from prior to posterior, as a measure of the information provided by an experiment; the same objective can be justified from a decision theoretic perspective [9,10]. Sebastiani and Wynn [11] propose selecting experiments for which the marginal distribution of the data has maximum Shannon entropy; this may be understood as a special case of Lindley's criterion. Maximum entropy sampling (MES) has seen use in applications ranging from astronomy [12] to geophysics [13], and is well suited to nonlinear models. Reverting to Lindley's criterion, Ryan [14] introduces a Monte Carlo estimator of expected information gain to design experiments for a model of material fatigue. Terejanu et al. [15] use a kernel estimator of mutual information (equivalent to expected information gain) to identify parameters in chemical kinetic model. The latter two studies evaluate their criteria on every element of a finite set of possible designs (on the order of ten designs in these examples), and thus sidestep the challenge of *optimizing* the design criterion over general design spaces. And both report significant limitations due to computation expense; [14] concludes that "full blown search" over the design space is infeasible, and that two order-of-magnitude gains in computational efficiency would be required even to discriminate among the enumerated designs.

The application of optimization methods to experimental design has thus favored simpler design objectives. The chemical engineering community, for example, has tended to use linearized and locally optimal [16] design criteria or other objectives [17] for which deterministic optimization strategies are suitable. But in the broader context of decision theoretic design formulations, sampling is required. [18] proposes a curve fitting scheme wherein the expected utility was fit with a regression model, using Monte Carlo samples over the design space. This scheme relies on problem-specific intuition about the character of the expected utility surface. Clyde et al. [19] explore the joint design, parameter, and data space with a Markov chain Monte Carlo (MCMC) sampler; this strategy combines integration with optimization, such that the marginal distribution of sampled designs is proportional to the expected utility. This idea is extended with simulated annealing in [20] to achieve more efficient maximization of the expected utility. [19,20] use expected utilities as design criteria but do not pursue information theoretic design metrics. Indeed, direct optimization of information theoretic metrics has seen much less development. Building on the enumeration approaches of [13–15] and the one-dimensional design space considered in [12], [7] iteratively finds MES designs in multi-dimensional spaces by greedily choosing one component of the design vector at a time. Hamada et al. [21] also find "near-optimal" designs for linear and nonlinear regression problems by maximizing expected information gain via genetic algorithms. But the coupling of rigorous information theoretic design criteria, complex physics-based models, and efficient optimization strategies remains an open challenge.

This paper addresses exactly these issues. Our interest is in physically realistic and hence *computationally intensive* models. We advance the state of the art by introducing flexible approximation and optimization strategies that yield optimal experimental designs for nonlinear systems, using a full information theoretic formalism, efficiently and with few limiting assumptions.

In particular, we employ a Bayesian statistical approach and focus on the case of parameter inference. Expected Shannon information gain is taken as our design criterion; this objective naturally incorporates prior information about the model parameters and accommodates very general probabilistic relationships among the experimental observables, model parameters, and design conditions. The need for such generality is illustrated in the numerical examples (Sections 5 and 6). To make evaluations of expected information gain computationally tractable, we introduce a generalized polynomial chaos surrogate [22,23] that captures smooth dependence of the observables jointly on parameters and design conditions. The surrogate carries no *a priori* restrictions on the degree of nonlinearity and uses dimension-adaptive sparse quadrature [24] to identify and exploit anisotropic parameter and design dependencies for efficiency in high dimensions. We link the surrogate with stochastic approximation algorithms and use the resulting scheme to maximize the design objective. This formulation allows us to plan single experiments without discretizing the design space, and to rigorously identify optimal "batch" designs of multiple experiments over the product space of design conditions.

Fig. 1 shows the key components of our approach, embedded in a flowchart describing a design–experimentation–model improvement cycle. The upper boxes focus on experimental design: the design criterion is formulated in Sections 2.1 and 2.2; estimation of the objective function is described in Section 2.3; and stochastic optimization approaches are described in Section 2.4. The construction of polynomial chaos surrogates for computationally intensive models is presented in Section 3. Section 4 briefly reviews computational approaches for Bayesian parameter inference, which come into play after the selected experiments have been performed and data have been collected. All of these tools are demonstrated on two example problems: a simple nonlinear model in Section 5 and a shock tube autoignition experiment with detailed chemical kinetics in Section 6.