Technical Report

TR-2008-001

Numerical methods for optimal experimental design of large-scale ill-posed problems

by

Eldad Haber

Mathematics and Computer Science

EMORY UNIVERSITY

Numerical methods for optimal experimental design of large-scale ill-posed problems

Eldad Haber *

Abstract

Experimental design for over-determined problems is a well studied topic where different criteria and optimization algorithms are explored. For ill-posed problems experimental design is rather new. In this paper we discuss optimal experimental design for ill-posed problems and suggest a numerical framework to efficiently achieve such a design. We demonstrate the effectiveness of our algorithm a common model problems.

keywords: Experimental design, ill-posed, constrained optimization.

1 Introduction

Inverse problems and methodologies are commonly used in order to solve under-determined ill-posed problems. Typically, one infers about the model m by conducting an experiment $F(m)$, collecting (noisy) data, d and use some inversion routine to estimate the model $\widehat{m}$ and its uncertainty. In this paper we deal only with linear inverse problems, that is $F(m) = Km$. The extension to nonlinear problems will be done in a sequential paper.

In past decades data collection and processing have been dramatically improved. Large amounts of "cleaner" data are now routinely collected. Recent advances in numerical PDE's and integral equations have enable us to better simulate the data. Finally, further advances in optimization algorithms allow us to invert the data harvesting high computational power. As a result, we are able to deal with inverse problems never solved before in many dimensions.

However, while it is possible to collect large amounts of data, it is not always clear how such data should be collected. In most cases data is collected based on a protocol that was developed decades ago. In many cases, such a protocol is neither optimal nor cost-effective. Furthermore, using a suboptimal experiment can reduces the overall resolution of the imaging method. As a result many imaging methods do not reach their full capabilities and important information is lost.

In some cases, poor experiments can be overcome by better algorithms (for example the Hubble telescope). However, a different way to achieve better (optimal) experiments is to

*Department of Mathematics and Computer Science, 400 Dowman Dr. Emory University, Atlanta, GA. phone: 404-727-4334, email haber@mathcs.emory.edu

properly design them. Obviously, the problem of experimental design is not new. It appears in many fields in physical, biological and social sciences. However, almost all the literature in the field treats the over-determine case where the problem is well posed (see e.g. [4, 5, 15, 1] and references within). Very little is done for the under-determined ill-posed case. In fact, the under-determine case is dismissed, for example, in the book by Pukelsheim [15] stating: "Clearly, for any reasonable inference, the number n of observation must be at least equal to $k+1$" where k is the number of unknowns. However, many practical problems, in fact, most geophysical and medical inverse problems, do have a smaller number of observations than the number of unknowns. Experimental design for such problems is mainly unexplored. For the under-determined case we are only aware of the work [6, 13] which uses techniques borrowed from the over determined case and the work of [16]. Non of these papers develop a systematic approach to experimental design of ill-posed problems. The only paper that treats ill-posed problems in a systematic way is the very recent paper of Bardow [3]. Our approach share some similarities to this approach.

A common character of many inverse problems is that they are large. That is, the number of parameters needs to be estimated can range from tens of thousands to millions. Even is the design space $\mathcal{Y}$ is small, the overall problem do be dealt with is very large. The computational techniques proposed in the literature above do not fit large scale problems. For ill-posed problems, the computational techniques proposed so-far relay on stochastic optimization. This can be prohibitively expensive for many inverse problems.

The goals of this paper are as follows. First, we intend to develop an appropriate methodology for the design of experiments for ill-posed problems. Such methodology should be broad and fit many inverse problems, including nonlinear inverse problems. Second, we intend to develop mathematical tools for the solution of the problem. We reformulate the problem such that standard constrained optimization methods can be used. Thirdly, we present an efficient algorithm for the solution of the design problem. The final goal of this paper is to apply the methodology for the design of a tomographic experiment and test its effectiveness.

The rest of the paper is organized as follows: In Section 2 the mathematical background for the problem is discussed. In Section 3 experimental design for linear inverse problems is proposed. In Sections 4 and 5 we present numerical optimization algorithms for the solution of the problem. In Section 6 we present numerical results for two different inverse problems. Finally, in Section 7 we summarize the paper and discuss future research.

2 Mathematical background

In this section we discuss problem setup and show the need for experimental design criteria for ill-posed problems.

We assume that the forward problem has the form

$$K(y)m + \epsilon = d \tag{2.1}$$

where $K(y)$ is an $n \times k$ matrix of the forward operator with typically $n < k$, which operates on the model, $m \in \mathcal{M}$, and depends on the experimental parameters $y \in \mathcal{Y}$. The vector

ϵ is the noise assumed to be Gaussian iid with standard deviation σ and d is the observed data. The goal of inversion algorithms is to recover m or some of its properties given the experimental setting y and the data d. In experimental design we ask the question: how to pick the experimental parameters y such that we obtain a "better experiment" in some sense?

Most OED techniques use the Fisher information matrix (or its equivalence) as the main tool to design experiments. Let us quickly review a few of the approaches. Assume, for simplicity, a Tikhonov regularized least-square solution with some smoothing matrix L and a regularization parameter α. We can estimate m as

$$\widehat{m} = (K(y)^\top K(y) + \alpha L^\top L)^{-1} K(y)^\top d$$

where $0 \leq \alpha$ is assumed to be a **fixed** regularization parameter. Such an assumption is necessary for the (well known) analysis below.

Now, we can try and estimate the expectation of the error in our estimate by looking at the norm[1] $\|\widehat{m} - m\|^2$ that is

$$\begin{aligned} &\mathbf{E}\,\|\widehat{m} - m\|^2 = \mathbf{E}\,\|(C(y)^{-1}K(y)^\top K(y) - I)m - C(y)^{-1}K(y)^\top \epsilon\|^2 = \\ &\underbrace{\|(C(y)^{-1}K(y)^\top K(y) - I)m\|^2}_{\text{bias}} + \underbrace{\sigma^2\ \text{trace}\,(K(y)C(y)^{-2})K(y)^\top}_{\text{variance}} \end{aligned} \tag{2.2}$$

where:

$$C(y) = (K(y)^\top K(y) + \alpha L^\top L) \tag{2.3}$$

The decomposition of the solution to bias and variance is a classical one [8]. Note that the bias term can be also written as

$$\text{Bias}(y, m) \quad = \quad \|(C(y)^{-1}K(y)^\top K(y) - I)m\|^2 = \alpha^2 \|C(y)^{-1}L^\top L m\|^2$$

This expression can have some computational advantages which we discuss later.

For over determined full rank problems where $k < n$ one typically sets α to 0. In this case the bias is 0 and therefore the error is controlled by the variance alone. Thus, in this case, it is reasonable to define the OED as

$$\min_y\ \text{trace}\ (K(y)C(y)^{-2}K(y)^\top) \overbrace{=}^{\text{if } \alpha=0} \min_y\ \text{trace}\ (C(y)^{-1}) \tag{2.4}$$

This approach is often referred to as the A-optimal design and it is thoroughly discussed in the literature. Other approaches to "scalarize" the information matrix, $C(y)^{-1}$ involve its determinant, its largest eigenvalue and other related quantities [4]. These are known as the D and E optimal designs. For problems that are dominated by the variance such approaches make sense. They help to decrease the variance of the result such that we are able to better cope with errors in the data.

[1] we may look at any semi-norm or a different function based on the goal of the reconstruction

However, such methods have a major flaw when dealing with ill-posed problems. Since the bias is typically large for such problems and commonly dominates the variance, controlling the variance alone may do very little for the improvement of the final error in our estimate.

A different approach for experimental design, based on a Bayesian framework, is reviewed in [5]. In this approach one minimizes a utility function which leads to similar designs as the non Bayesian ones (but with different interpretations). Our approach is connected to the Bayesian design and can be viewed as a version of Bayesian or empirical Bayes design. The similarities and differences to Bayesian designs are explored in the next section. The goal of the next sections is to explore an approaches for OED that take into consideration the ill-posed nature of the problem.

3 Experimental design criteria

As discussed in the previous section using the variance alone as an objective function to be minimize for optimal experiments does not take into consideration the effect of the bias in the solution. In order to do that we look at the risk directly.

To obtain an optimal design we would like to decrease the overall risk and therefore choose y that decreases the bias as well as the variance. The problem is that the bias depends on the unknown model and thus it is impossible to evaluate it directly. Nevertheless, there are a number of approaches to obtain an estimate of the bias. We now discuss three such approaches.

- **Average optimal experimental design**

  Assuming the model is in a bounded convex set $\mathcal{M}$, we look at the "average" case. We can choose a design that minimizes the risk over all possible models in $\mathcal{M}$ that is

  $$\text{P1}: \min_{y} \quad \alpha^2 \frac{\int_{\mathcal{S}} \|C(y)^{-1}L^\top Lm\|^2 \, dm}{\int_{\mathcal{S}} \, dm} + \sigma^2 \text{ trace } \left(K(y)C^{-2}K(y)^\top\right) \tag{3.5}$$

  The problem with the average design is that it does not give any priority to "more reasonable" models in the set $\mathcal{M}$. Unless the set $\mathcal{M}$ is well constrained this design may be overly pessimistic.

- **Bayesian optimal experimental design**

  A different approach is to assume that m is associated with a probability density function (PDF). In this case one considers the Bayesian design and the goal is to find y that solves the following optimization problem

  $$\text{P2}: \min_{y} \quad \mathbf{E}_m \; \alpha^2 \|C(y)^{-1}L^\top Lm\|^2 + \sigma^2 \text{trace}\,(K(y)C^{-2}K(y)^\top) \tag{3.6}$$

If m is Gaussian with and covariance matrix Σ_m. Then the Bayes OED can be reduced to

$$\min_y \ \alpha^2 \text{trace}(B\Sigma_m B^\top) + \sigma^2 \text{trace}\,(K(y)C(y)^{-2}K(y)^\top) \tag{3.7}$$

where $B = C(y)^{-1}L^\top L$. This expression can be further reduced if we assume that $\Sigma_m = \frac{1}{\alpha}(L^\top L)^{-1}$ (as in the MAP estimate). By Using the generalized SVD of K and L it is possible to show that the problem is equivalent to minimizing

$$\min_y \text{trace}(C(y)^{-1})$$

which is the A optimal Bayesian design.

The difficulty with Bayesian risk estimators is that they require a PDF for m. For many problems it is difficult or impossible to obtain such a PDF. In a sense, the average and the Bayesian are two extreme cases. In the first we know nothing about the PDF besides the set $\mathcal{M}$ and in the other we have a complete knowledge of the PDF. For most problems non of these cases is realistic. We now explore a third option that we believe is the most realistic.

- **Design based on Empirical Risk**

  For many problems a PDF that describes the model is difficult or impossible to obtain however, it is possible to obtain examples of plausible models. For example, in geophysics, the Marmusi model [19] is often used to test inversion algorithms. In medical physics the Shepp-Logan model is often being used as a golden standard. Furthermore, recent developments in geostatistics introduce algorithms that are able to obtain different realizations of a given media from a single realization without resorting to a PDF (see for example [17] and reference within).

  Here, we assume that we are able to obtained s examples of plausible models $m_1, \ldots, m_s$. We refer to these models as *training models*. Given the training models we can evaluate the *experimental risk* by using an average of the models. That is we solve

$$\text{P3}: \min_y \quad \frac{\alpha^2}{s}\sum_{j=1}^{s} \|C(y)^{-1}L^\top L m_j\|^2 + \sigma^2 \text{trace}\,(K(y)C^{-2}K(y)^\top) \tag{3.8}$$

  The appeal of this approach is that it does not require a PDF. We have successfully used a similar approach for the evaluation of regularization operators [10]. The approach is commonly used in machine learning where a function needs to be evaluated based on some examples.

Although the approaches above have very different statistical meanings and should be interpreted differently they lead to very similar numerical problems. In the next section we explore numerical methods for the solution of these problems.

4 Numerical Optimization Techniques for OED

4.1 Problem reformulation

Solving either P1,P2 or P3 gives an optimal design of the form

$$\min_{y} \ \mathcal{J}(y) = \text{Bias}(y) + \text{Var}(y) \tag{4.9}$$

The difference between the problems is the way that the bias is estimated. In this section we discuss numerical techniques for the solution of the above problems.

The numerical solution of OED can be complicated. For many problems the matrix $K(y)$ is not continuously differentiable with respect to y or it is difficult to obtain such a derivative. This implies that typical optimization methods may not be suitable for the solution of such problems. We therefore reformulate the OED problem. Rather than assuming to have a continuous variable y we choose a discrete subset of experiments by discretizing the space of possible experiments $\mathcal{Y}$. The idea of discretizing the experimental space is not new and appears in [15].

Assume that y is discretized to have the set $\{y_1, \ldots, y_p\}$. After discretization we obtain p possible experiments

$$k_j^\top m + \epsilon_j = d_j \qquad j = 1, \ldots p \tag{4.10}$$

where $k_j^\top$ is a row vector that represents a single experiment. Here, for simplicity, we assume that each discrete y corresponds to a single experiment. In the general case, each discrete y corresponds to a matrix K_j. The random variable ϵ_j is the noise corresponds to the j^{th} measurement and d_j is a corresponding j^{th} datum. Consider now a weighted experiment

$$\sqrt{w_j}(k_j^\top m + \epsilon_j) = \sqrt{w_j} d_j \qquad j = 1, \ldots p \tag{4.11}$$

The weight w_j can be thought of as the inverse of the standard deviation of the noise ϵ_j. If the noise level is large then w_j is small and if the noise level is small then w_j is large.

With some abuse of notation let

$$K = \begin{pmatrix} k_1^\top \\ \vdots \\ k_p^\top \end{pmatrix}.$$

Define the weight matrix $\sqrt{W}$ be

$$\sqrt{W} = \text{diag}\,(\sqrt{w}).$$

We rewrite the weighted design as

$$\sqrt{W}(Km + \epsilon) = \sqrt{W} d \tag{4.12}$$

If many of the w's are zeros, then only a small set of the experiments is carried out, that is we assume that some of the data has infinite standard deviation. The idea therefore is to replace the problem of finding the vector y with finding the weights w. These weights yield more information then just what experiment should be carried out. They also guide us to the appropriate standard deviation that should be obtained at each measurement.

Typically, the vector w is very large because it contains all (discrete) possible experiments. Furthermore, it is clear that the risk can be decreased if we allow using most or all the experiments. This is, of-course, not necessarily a good idea because performing more experiments adds to the cost and our goal is to choose a small set of experiments, not all possible ones. We therefore modify the optimization problem by adding a regularization term that penalize w. Since we are interested in w's that are mainly 0's, that is, a sparse w, we use the L_1 regularization. Ideally, one would like to use the L_0 regularization to obtain the sparsest solution, however, since the L_0 solution yields a combinatorially difficult problem, it is common to use the L_1 solution instead. A justification for this approach is given in [7]. Using the L_1 solution one could then approximate the L_0 solution (see [4]). We explain this heuristic in Section 5

It is obvious that w must be non-negative. For most practical problems we also must have an upper bound for w, $w_{\max}$. The reason is that very large w implies that we need to collect the data with very high accuracy (small standard deviation). Such standard deviation may not be attainable from practical reasons.

To summarize, we suggest to solve the following optimization problem

$$\begin{aligned} \min_{w} \quad & \mathcal{J}(w) = \mathrm{Bias}(w) + \mathrm{Var}(w) + \beta\|w\|_1 \\ \text{s.t.} \quad & 0 \le w \le w_{\max} \end{aligned} \tag{4.13}$$

where in general

$$\mathrm{Bias}(w) = \mathrm{E}_m \; \alpha^2 \|C(w)^{-1} L^\top L m\|^2 \tag{4.14a}$$

$$\mathrm{Var}(w) = \mathrm{trace}\left(W\, K C(w)^{-2} K^\top W\right) \tag{4.14b}$$

$$C(w) = K^\top W K + \alpha L^\top L \tag{4.14c}$$

where E_m is an average with respect to m.

The formulation above is a new approach for OED of under-determined problems and it has a few main advantages. Mainly, many linear and linearized OED problems can be solved using similar tools. Also, it is easy to see that the problem is continuously differentiable and convex with respect to w. Since w is non-negative the non-differentiable one-norm can be replaced by the sum of w. Thus we obtain a large but tractable problem.

Computing the trace and the expected value of a large dense matrices which involves the inverse cannot be effectively done for large scale problems. Instead, we seek to approximate the objective function with estimates that do not require the direct computation of the trace or the inverse. In the next subsections we suggest a method to obtain such approximations and to solve the problem using modern optimization tools.

4.2 Estimation of the variance

In order to estimate the variance we need to estimate the trace of a possible large scale matrix. In previous years stochastic trace estimators were successfully used for the estimation of the trace of similar problems. In particular, Golub and von Matt [9] have used the method proposed by Hutchnison [11] for trace estimation. For a SPD matrix, H, the trace is estimated by

$$\text{trace}(H) \approx \sum_{i=1}^{s} v_i^\top H v_i \tag{4.15}$$

where v_i is a random vector of ± 1. The accuracy of the estimation was numerically studied in [2] with a surprising result that the best compromise between accuracy and computational cost is achieved for $s = 1$. Our numerical experiments yields similar results. We therefore replace the trace by the approximation (4.15). Using this approximation we rewrite the approximation to the variance as

$$\widehat{\text{Var}}(w) = v^\top W\, KC(w)^{-2} K^\top\, Wv = w^\top V(w)^\top V(w) w \tag{4.16}$$

where the matrix $V(w)$ is

$$V(w) = C(w)^{-1} K^\top \text{diag}\,(v)$$

4.3 Estimation of the bias

The estimation of the bias requires the estimation of the expected value of (4.14a). If the Bayesian approach is taken (which includes the average design) then it is possible to use a stochastic estimator and replace the bias with an approximate

$$\widehat{\text{Bias}}(w) = \frac{\alpha^2}{s} \sum_{j=1}^{s} m_j^\top B(w)^\top B(w) m_j \tag{4.17}$$

where

$$B(w) = C(w)^{-1} L^\top L$$

and m_j are independently drawn from the probability density function of m. Since the average design is a special case of the Bayesian design (where the prior in flat) it can be computed in a similar way. It is interesting to note that this has the same form as empirical risk design. Thus the average, Bayesian and empirical risk designs can be computed using the same computational tools. It is important to note though that the final result has a different interpretation depending on the approach taken.

5 Solving the optimization problem

5.1 Numerical solution of the bound constraint optimization problem

Using stochastic approximations to the bias and variance we obtain an optimization problem of the form

$$\begin{aligned} \min \quad & \mathcal{J} = \frac{\alpha^2}{s} \sum_{j=1}^{s} m_j^\top B(w)^\top B(w) m_j + w^\top V(w)^\top V(w) w + \beta e^\top w \\ \text{s.t} \quad & 0 \leq w \leq w_{\max} \end{aligned} \tag{5.18}$$

This is a large scale constrained least-squares problem which is smooth in w.

To solve the problem we have used the projected gradient and a projected Gauss-Newton methods. In order to use any gradient descent method we require the computation of the gradients. It is easy to note that both the bias and the variance have a nonlinear least-squares form thus we obtain

$$\begin{aligned} \nabla_w \left(w^\top V(w)^\top V(w) w \right) &= J_v(w)^\top V(w) w \\ \nabla_w \left(m^\top B(w)^\top B(w) m \right) &= J_b(w)^\top B(w) m \end{aligned}$$

where

$$J_v(w) = \frac{\partial (V(w)w)}{\partial w} \quad \text{and} \quad J_b(w) = \frac{\partial (B(w)m)}{\partial w}$$

To obtain (an expression for) the matrices J_v and J_b we use implicit differentiation. First, we write

$$(K^\top W K + \alpha L^\top L)^{-1} (K^\top W K) m = r_b \quad \leftrightarrow \quad (K^\top W K) m = (K^\top W K + \alpha L^\top L) r_b$$

Note that the matrix J_b is nothing but $\nabla_w r_b$. Differentiating both sides we obtain

$$K^\top \operatorname{diag}(Km) = (K^\top W K + \alpha L^\top L) \frac{\partial r_b}{\partial w} + K^\top \operatorname{diag}(K r_b)$$

which implies that

$$J_b = C^{-1} \left(K^\top \operatorname{diag}\left(K(m - r_b) \right) \right) \tag{5.19}$$

To differentiate the variance we use a similar trick. First note that

$$\frac{\partial (V(w)w)}{\partial w} = V(w) + \frac{\partial (V(w) w_{\text{fixed}})}{\partial w}$$

To compute the second term in the above sum we write $V(w) w_{\text{fixed}} = r_v$ which implies

$$(K^\top W K + \alpha L^\top L)^{-1} K^\top \operatorname{diag}(v)\, w_{\text{fixed}} = r_v \quad \leftrightarrow \quad K^\top \operatorname{diag}(v)\, w_{\text{fixed}} = (K^\top W K + \alpha L^\top L) r_v$$

Differentiating both sides with respect to w we obtain

$$0 = (K^\top W K + \alpha L^\top L)\frac{\partial r_v}{\partial w} + K^\top \mathrm{diag}\,(K r_v)$$

which implies that

$$J_v = V - C^{-1} K^\top \,\mathrm{diag}\,(K r_v) \tag{5.20}$$

This completes the evaluation of derivatives to the objective function. It is important to note that neither the matrix C nor its inverse are needed explicitly in order to evaluate the objective function and the gradients. Whenever we need to evaluate a product of the form $C^{-1}u$ we simply solve the system $Cx = u$. To solve such system only matrix-vector products of the form Cv are needed.

Given the gradients we can now use any gradient based method for the solution of the problem. We have experimented with the projected gradient which requires only gradient evaluation and the projected Gauss-Newton method. For the Gauss-Newton method one can approximate the Hessian of the objective function by

$$\nabla^2 \mathcal{J} = J_v^\top J_v + J_b^\top J_b$$

Given the Jacobian it is possible to use the method suggested by Lin and More [12]. The active set is (approximately) identified by the gradient projection method and a truncated Gauss-Newton iteration is performed on the rest of the variables. Again, it is important to note that the matrices J_v and J_b need not be calculated. A product of either with a vector involves a solution of the system $Cx = u$. Thus, it is possible to use conjugate gradient to compute an approximation of a Gauss-Newton step.

5.2 Approximating the L_0 problem

Although it is straight-forward to solve the L_1 regularization problem one can attempt to approximate the L_0 solution. The L_0 solution is the vector w that has the least number of non-zero entries. Obtaining this solution is a combinatorial difficult problem. It was proven in [7] that the L_1 solution is often the L_0 solution. However, this is not always the case. Nevertheless, it is possible to obtain solutions that approximate the L_0 solution by using the L_1 solution. To do this we note that the L_0 solution is obtained by the following approach. First, assume that w is divided into two sets I_0 and I_A. The set I_0 contains all the indices for which $w_{I_0} = 0$ and the set w_{I_A} are the rest. If we know a-priori which indices are zero then we could simply solve a convex optimization problem

$$\begin{aligned} \min_{w_{I_A}} \quad & \mathcal{J}(w_{I_A}, w_{I_0} = 0) = \frac{\alpha^2}{s}\sum_{j=1}^{s} m_j^\top B(w_{I_A})^\top B(w_{I_A}) m_j + w_{I_A}^\top V(w_{I_A})^\top V(w_{I_A}) w_{I_A} \\ \text{s.t} \quad & 0 \le w_{I_A} \le w_{\max} \end{aligned} \tag{5.21}$$

Note that this problem does not require any regularization term since the non-zero set is assumed known. The combinatorial nature of the L_0 problem arise because it is difficult to

identify the set I_A. Nevertheless, one can approximate the set I_A by the non-zero set obtained from the L_1 minimization. This idea is discussed in [4] where numerical experiments verify that this approximate L_0 solution can be different from the L_1 solution. In this work we use this approach as well and set the the final weights to the ones that solve (5.21) with the set I_A obtained from the solution of the L_1 problem.

6 Numerical results

In the section we present numerical results which demonstrate the ideas above and demonstrate that experimental design can be important to the application. As a test problem we use the ray tomography example. This is a common model problem for geophysical inverse problems and the one used in [6, 16].

The goal of borehole ray tomography is to determine the slowness of a medium. Sources and receivers are placed in boreholes or on the surface of the earth and travel times from sources to receivers are recorded. The forward model is

$$d_j = \int_{\Gamma_j} m(\mathbf{x})\, d\ell + \varepsilon_j \quad j = 1, \ldots, n \tag{6.22}$$

where Γ_j is the ray path that connects source to receiver. In the linearized case we consider here, the ray path does not change as a function of m (see [14, 18]). The goal of experimental design in this case is to choose the optimal placement of sources and receivers.

For our numerical simulation we assume that each borehole covers the interval $[0, 1]$ and that the distance between the holes is 1 as well. The model is discretized using 64^2 cells. A sketch of the experiment is presented in Figure 1.

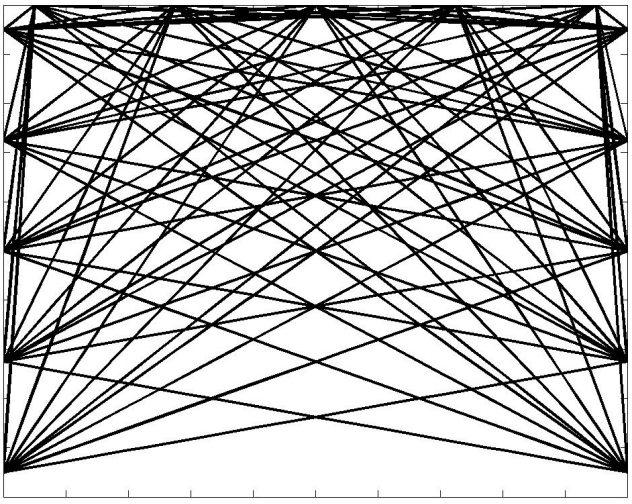

Figure 1: A sketch of the borehole tomography experiment.

To solve the problem using our formulation consider first all possible source-receiver position $[r, s]$. Our design space, $\mathcal{Y}$, is comprised of three line segments where the sources

β	Bias (L_1)	Bias (L_0)	Variance (L_1)	Variance (L_0)	nnz(w)
1.0e+001	4.3e+002	3.1e+002	5.1e-002	9.8e-002	249
1.3e+000	2.4e+002	2.1e+002	2.7e-001	1.5e-001	310
1.6e-001	2.1e+002	2.1e+002	2.8e-001	1.7e-001	423
2.0e-002	2.1e+002	2.1e+002	2.9e-001	1.4e+000	491
2.4e-003	2.1e+002	2.1e+002	2.9e-001	3.4e-001	507
3.1e-004	2.1e+002	2.1e+002	2.9e-001	3.6e-001	519

Table 1: Bias and variance and the sparsity obtained for different β's.

and receivers can be placed. The lines $I_1 = \{x_1 = 0 \text{ and } 0 \leq x_2 \leq 1\}$, $I_2 = \{x_1 = 1 \text{ and } 0 \leq x_2 \leq 1\}$ and $I_3 = \{x_2 = 0 \text{ and } 0 \leq x_1 \leq 1\}$. We are free to choose rays that connect any point on I_k to a point on I_j $j, k = 1, \ldots, 3;\ j \neq k$. We descretize each line segments using 32 equally spaced points. This gives $32^2 \times 3 = 3072$ possible experiments. Our goal is to choose the roughly 500 optimal experiments.

Next, we need to choose a method to estimate the bias. Rather than assuming we have a probability density function, we divided the Marmousi model [19] into 4 yielding 3 training models and a single testing model. We use the three training models to obtain the optimal experiment and then use the testing model to asses our performance.

The models are plotted in Figure 2.

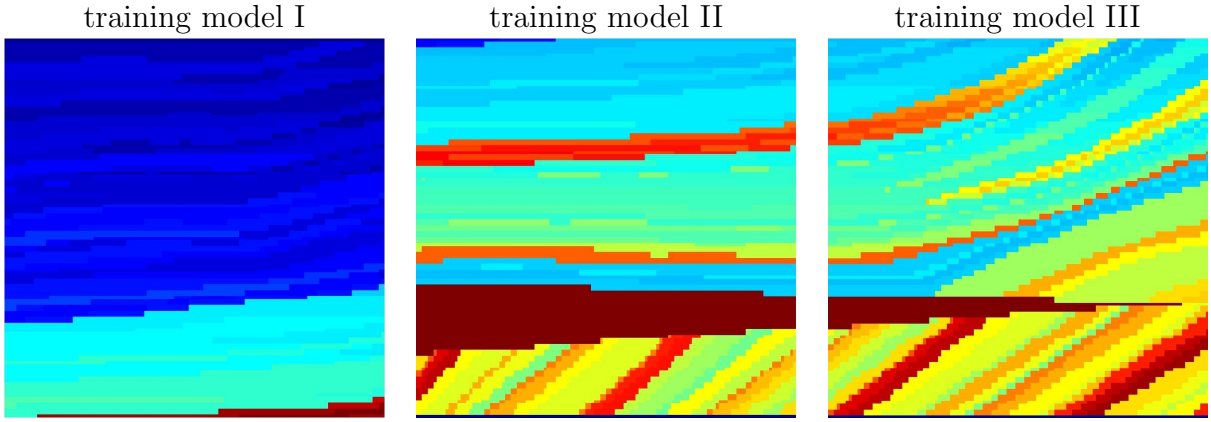

Figure 2: The training set made by the divided Marmusi model.

For an upper bound on our w we choose $w_{\max} = 1$. We then run our code for different β's obtaining different number of experiments and different fit to the training set for each β. The sparsity as well as the bias and variance for each optimal result is presented in Table 1. When looking at the table it is important to note that at the minima of the objective function the bias is substantially larger than the variance. It is thus demonstrated how methods for optimal experimental design for ill-posed problems must take the bias into consideration since it may play a much larger role that the variance.

To test our design we use the optimal design obtained and a design made by uniformly placing the rays we perform the experiment on the forth part of the Marmousi model. The norm $\|\hat{m}(w) - m\|$ was 1.7×10^3 for the optimal design and 3.1×10^3 for the one with uniform placement of the sources and receivers. In Figure 3 we construct the reconstructions and the "true" model. It is evident that optimally designing the survey yield substantial improvement in the recovered model.

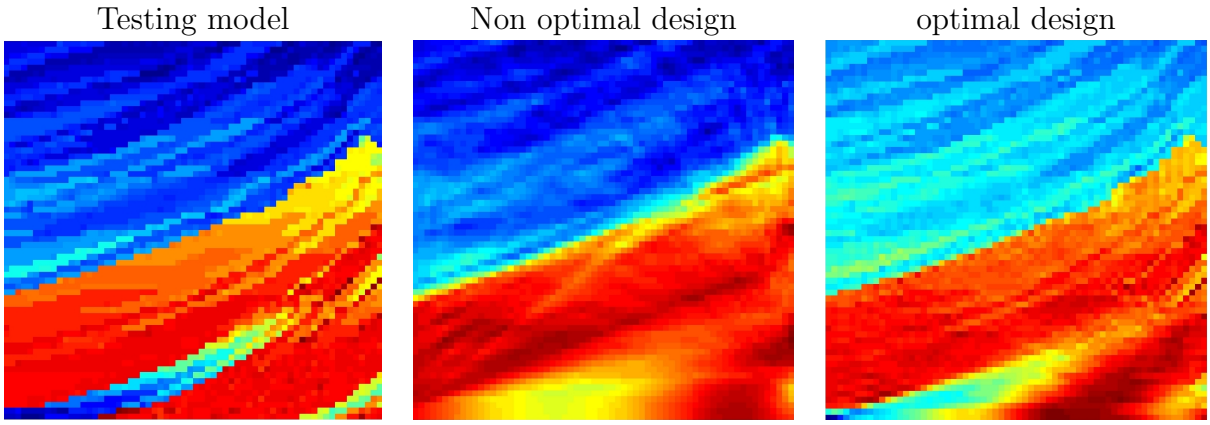

Figure 3: Testing the design, optimal vs nonoptimal designs.

7 Summary

In this paper we have discussed a method for the design of experiments for linear ill-posed problems. The main difference between the underdetermined and the over-determined case is that the bias has to be considered and typically plays an important role in the design process.

We have explored an efficient computational strategy for the solution of the problem. The strategy reformulates the optimization problem and solves for the weights of the discretized experiments. The reformulation leads to a convex, continuously differentiable optimization problem that can be treated using conventional optimization techniques.

We have tested our methodology on a typical inverse problems and have demonstrated that the approach improves naive designs. In a sequential work we intend to extend the above framework to nonlinear experimental design.

8 Acknowledgement

The author would like to thank Luis Tenorio for the many discussions and the reading of the manuscript.

References

[1] A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs*. Oxford University Press, 1992.

[2] Z. Bai, M. Fahey, and G. Golub. Some large scale matrix computation problems. *J. Computational & Applied Math*, 74:71–89, 1996.

[3] A. Bardow. Optimal experimental design for ill-posed problems, the meter approach. *Computers and chemical engineering*, 32:115–124, 2008.

[4] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University press, 2004.

[5] K. Chaloner and I Verdinelli. Bayesian experimental design: A review. *Statis. Sci.*, 10:237–304, 1995.

[6] A. Curtis. Optimal experimental design: cross borehole tomographic example. *Geophysics J. Int.*, 136:205–215, 1999.

[7] David D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[8] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[9] G. Golub and U. Von Matt. Tikhonov regularization for large scale problems. *Technical report SCCM 4-79*, 1997.

[10] E. Haber and L. Tenorio. Learing regularization functionals a suprvised training approach. *Inverse Problems*, 19:611–626, 2003. n3.

[11] M.F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *J. Commun. Statist. Simul*, 19:433–450, 1990.

[12] C.J Lin and J. More. Newton's method for large bound-constrained optimization problems. *SIAM Journal on Optimization*, 9:1100–1127, 1999.

[13] H. Maurer, D. Boerner, and A. Curtis. Design strategies for electromagnetic geophysical surveys. *Inverse Problems*, 16:1097–1117, 2000.

[14] R. L. Parker. *Geophysical Inverse Theory*. Princeton University Press, Princeton NJ, 1994.

[15] F. Pukelsheim. *Optimal design of experiments*. John Wiley & Sons, 1993.

[16] P. Routh, G. Oldenborger, and D. Oldenberg. Optimal survey design using the point spread function measure of resolution. In *Proc. SEG*, Huston, TX, 2005.

[17] P. Sarma, L.J. Durlofsky, K. Aziz, and W. Chen. A new approach to automatic history matching using kernel pca. *SPE Reservoir Simulation Symposium, Houston, Texas*, 2007.

[18] A. Tarantola. *Inverse problem theory*. Elsevier, Amsterdam, 1987.

[19] R.J. Versteeg. *Analysis of the problem of the velocity model determination for seismic imaging*. Ph.d. dissertation, University of Paris, 1991. France.