RC24978 (W1004-055) April 16, 2010
Mathematics

IBM Research Report

Numerical Methods for the Design of Large-Scale Nonlinear Discrete Ill-Posed Inverse Problems

E. Haber
Department of Mathematics
University of British Colombia
Vancouver, Canada

L. Horesh
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

L. Tenorio
Department of Mathematical and Computer Science
Colorado School of Mines
Golden, CO USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems

E. Haber[*], L. Horesh[†] and L. Tenorio[‡]

November 20, 2009

Abstract

Design of experiments for discrete ill-posed problems is a relatively new area of research. While there has been some limited work concerning the linear case, little has been done to study design criteria and numerical methods for ill-posed nonlinear problems. We present an algorithmic framework for nonlinear experimental design with an efficient numerical implementation. The data are modeled as indirect noisy observations of the model collected via a set of plausible experiments.An inversion estimate based on these data is obtained by weighted Tikhonov regularization whose weights control the contribution of the different experiments to the data misfit term. These weights are selected by minimization of an empirical estimate of the Bayes risk that is penalized to promote sparsity. This formulation entails a bilevel optimization problem that is solved using a simple descent method. We demonstrate the viability of our design with a problem in electromagnetic imaging based on direct current resistivity and magnetotelluric data.

1 Introduction

Inverse problems have come to play a key role in a variety of fields such as computer vision, geophysics and medical imaging. Practical applications of inverse problems in these fields require consistent discretization schemes as well as numerical optimization methods and algorithms for solving the linear systems defined by their underlying PDEs. A great volume of work has already addressed some of these issues, see, for example, [7, 16, 3, 12, 17, 21] and references therein. Primarily, this work has addressed questions such as regularization, incorporation of a-priori information, sensitivity analysis, as well as the development of efficient inversion schemes for large-scale problems. In this paper we address important questions that arise in experimental design of discrete ill-posed nonlinear inverse problems.

[*] Department of Mathematics, University of British Colombia, Vancouver, Canada

[†] Business Analytic and Mathematical Sciences, IBM T J Watson Research Center, Yorktown Heights, NY.

[‡] Department of Mathematical and Computer Science, Colorado School of Mines, Golden, CO.

The basic setup is as follows: discrete data $d \in \mathbb{R}^n$ consist of noisy indirect observations of a discrete model (function) $m \in \mathbb{R}^s$ via a forward operator $F : \mathbb{R}^s \to \mathbb{R}^n$:

$$d(m;p) = F(m;p) + \epsilon, \tag{1.1}$$

where the noise $\epsilon \in \mathbb{R}^n$ is assumed random with a known probability distribution and p is a vector of experimental parameters that defines the experimental configurations. The operator F is assumed to be well-defined on a convex set $\mathcal{M}$ that contains m. It is also assumed that the parameter space is discretized so that p holds vector values $p_1, ..., p_n$. This is a common assumption in experimental design [5, 28, 8]. We shall write $F_k(m) = F(m; p_k)$ to refer to the forward problem in the k^{th} experimental configuration. Our approach proposes the consideration of all the experiments at once, while assigning weights that will be determined optimally. The forward mapping we consider is then $F(m) = [F_1(m), \ldots, F_n(m)]^\top$, where $F_k(m)$ is, in general, a nonlinear function of m. We shall also write $d(m) = [d_1(m), ..., d_n(m)]$ with $d_k(m) = F_k(m) + \epsilon_k$ (for simplicity we assume that the $d_i(m)$ are scalars but in the general case they may be vectors). In the forward problem, $F(m)$ is computed for a given model m and the data d can be simulated by the addition of noise realizations from the known noise distribution. The forward problem typically requires the solution of a large system of (possibly nonlinear) equations. On the other hand, the objective of the inverse problem is to obtain an estimate of a model m given the data d. In realistic settings, this question typically leads to a large-scale optimization problem where a combination of a data misfit and regularization terms is minimized. It is well known that the process of regularization may introduce a considerable amount of bias into the model estimates, even when the noise level is negligible [1, 10, 23]. We may regard this bias as a systematic error that depends on the experimental design and the regularization procedure. The key point is that such systematic error can be controlled by an appropriate choice of both. In this paper we address the former, which shall be henceforth called the *design problem*.

In the design problem we do not have (at least not formally) a model m or data d; instead we consider the question of choosing an optimal (in some sense) set of experiments to be conducted. That is, we would like to choose an optimal subset of the n potential experiments. The experimental design involves an inverse problem that we solve using penalized weighted least squares (i.e., generalized Tikhonov regularization):

$$\min_m \quad \mathcal{J}(m, w) = \frac{1}{2} \sum_{j=1}^{n} w_j [\, F_j(m) - d_j(m)\,]^2 + S(m), \tag{1.2}$$

where the regularization functional $S(m)$ is assumed to be twice differentiable [1], and $0 \leq w \in \mathbb{R}^{+^n}$ is a vector of non-negative weights (the regularization parameter α that multiplies $S(m)$ is integrated within the weights). Hence, the contribution of the selected experiments (those with $w_k \neq 0$) is weighted to control their overall influence in the data misfit. The weights are passed to the practitioner who will then conduct the chosen experiments and subsequently estimate the model using (1.2). Let us consider an example:

[1] The problem can be solved effectively also when this restrictive assumption does not hold, see for example [20] and references therein

Example 1 In magnetotelluric (MT) inversion (e.g., [25, 27]), the goal is to solve for the conductivity given measurements of the electric field at different frequencies. The forward functions $F_j(m)$ are derived by the discretization of a partial differential equation:

$$F_j(m) = q^\top [\, A + i\,\gamma_j G(m) \,]^{-1} b,$$

where q is the observation operator, b is the source term (both given), A is the discretization of the operator $\nabla \times \nabla \times$, $G(m)$ is a mass matrix weighted by the conductivity m, and γ_j is the source frequency. We have a choice of frequencies $\mathcal{G} = \{\, \gamma_1, \ldots, \gamma_n \,\}$ that can be used for data acquisition and the goal is to choose an optimal subset of $\mathcal{G}$.

The rest of the paper is organized as follows: In Section 2 we discuss optimality criteria for the design problem. These criteria lead to optimization problems whose numerical implementations are described in Section 3. In Section 4 we present numerical results obtained with an inverse problem based on direct current resistivity and magnetotelluric data. Section 5 summarizes the study and outlines questions for future research work.

2 Optimality criteria

Experimental designs require the definition of optimality criteria. In the case of linear models with a full-column-rank forward matrix operator, the least-squares estimate is unbiased and it is thus sensible to define optimality criteria based solely on the information matrix (inverse of the covariance matrix) [8, 28]. In our case such a measure is insufficient; the bias needs to be considered. We therefore define measures based on an average error of the inversion estimate.

For a fixed model $\boldsymbol{\mu} \in \mathcal{M}$, the data vector $d(\boldsymbol{\mu}) \in \mathbb{R}^n$ is acquired by solving the forward problem and adding noise [2]:

$$d(\boldsymbol{\mu}) = F(\boldsymbol{\mu}) + \epsilon.$$

Next, an estimate $\widehat{m}$ of the model $\boldsymbol{\mu}$ is obtained from the solution of (1.2) with a fixed w:

$$\widehat{m} = \widehat{m}(w, d(\boldsymbol{\mu})) = \arg\min_{m} \;\; \frac{1}{2} \| \, W^{1/2} \, [\, F(m) - d(\boldsymbol{\mu}) \,] \, \|^2 + S(m), \tag{2.3}$$

where $W^{1/2} = \operatorname{diag}(\, w^{1/2} \,)$. Clearly, the estimate $\widehat{m}$ depends on w and $d(\boldsymbol{\mu})$; its associated squared error is

$$\textbf{error} = \|\widehat{m} - \boldsymbol{\mu}\|^2.$$

In practice, the experiments are conducted with other unknown models; it is therefore important to consider the variability of the MSE over $\mathcal{M}$. To do so, we use an average measure of the MSE: the elements of $\mathcal{M}$ are modeled as random with some prior distribution π and the Bayes risk $R(w) = \mathbf{E}[\, \mathrm{MSE}(w, \boldsymbol{\mu}) \,]$ is minimized so as to obtain an estimate of the weights w that works well on average across $\mathcal{M}$. This is a natural way to proceed as the experimental design is carried out prior to data acquisition; the design parameters are chosen to minimize

[2] We use $\boldsymbol{\mu}$ to indicate a fixed model rather than m, which is a variable in the function to be optimized.

an expected value over noise and model distributions (this is an example of a frequentist calculation that is Bayesianly justifiable [29]). However, both expectations are difficult to compute for the nonlinear, ill-posed problems we consider. Hence, we use approximations based on sample estimates.

The MSE is estimated via a sample average over L independent noise vectors ϵ_j[3]

$$\mathrm{MSE}(w, \boldsymbol{\mu}) \approx \frac{1}{L} \sum_{j=1}^{L} \| \widehat{m}(w, d_j) - \boldsymbol{\mu} \|^2, \tag{2.4}$$

where $d_j = F(\boldsymbol{\mu}) + \epsilon_j$. For the expectation over π we use K independent samples $\boldsymbol{\mu}_i$ (the $\boldsymbol{\mu}_i$ are sometimes already available and thus we refer to them as 'training models'). An empirical estimate of $R(w)$ is then

$$\begin{aligned} \min_{w} \quad & \widehat{R}(w) = \frac{1}{LK} \sum_{i,j} \| \widehat{m}(w, d_j(\boldsymbol{\mu}_i)) - \boldsymbol{\mu}_i \|^2 \\ \text{s.t} \quad & 0 \le w. \end{aligned} \tag{2.5}$$

Clearly, trivial solutions of the optimization problems (2.5) can be obtained by simply choosing $w_k \gg 0$ for all experiments. This merely states that conducting more experiments yields better model reconstructions. To obtain efficient and effective designs we must penalize solutions that include many nonzero w_k. We therefore proceed as in our previous work on experimental design of linear inverse problems [15]; we use a sparsity-controlled experimental design where the sparsity of w is controlled by an ℓ_p penalty. Thus, the sparsity controlled design problem is:

$$\begin{aligned} \min_{w} \quad & \widehat{R}_\beta(w) = \widehat{R}(w) + \beta \, \|w\|_p \\ \text{s.t} \quad & 0 \le w. \end{aligned} \tag{2.6}$$

Here the parameter $\beta \ge 0$ is tuned to achieve a desired sparsity level. For the ℓ_p penalty we consider values of $p \in [0, 1]$. The ℓ_0 penalty leads to the sparsest design, however, such designs are combinatorially difficult to find. The ℓ_1-norm setup leads to a convex term in the objective function that can be tackled using a variety of methods [2, 30, 31, 34, 9], this choice will be discussed next. Some heuristics for the designs with $p = 0$ have been discussed in [15] and will be revisited below.

Note that (2.6) can be casted as a bilevel optimization problem:

$$\min_{w, \widehat{m}_{ij}} \quad \frac{1}{LK} \sum_{i,j} \| \widehat{m}_{ij} - \boldsymbol{\mu}_i \|^2 + \beta \|w\|_p \tag{2.7a}$$

$$\text{s.t} \quad 0 \le w \tag{2.7b}$$

$$\widehat{m}_{ij} = \arg\min \ \frac{1}{2} \left\| W^{1/2} [\, F(m) - d_j(\boldsymbol{\mu}_i) \,] \right\|^2 + S(m). \tag{2.7c}$$

Remarks:

[3] The subscript j (and later ij) is used here to indicate the j^{th} vector (matrix) and not the j^{th} entry of a vector (matrix). The dimensions of ϵ_j should be clear from the context.

- When F is linear the problem reduces to the one already considered in [15]; no bilevel optimization is necessary. This case is discussed in the next section.

- For simplicity, we have formulated (2.7c) as an *unconstrained optimization problem.* In many situations it may be better to cast this inverse problem using the PDEs as equality constraints. Such formulation is not considered here.

- The search for an optimal experimental design is a large-scale bilevel optimization problem. While such problems are difficult to solve in general, it is important to remember that in many practical cases even a small improvement upon the current design (i.e., not necessarily the global optimum) is valuable.

- Our particular problem involves weakly coupled $L \times K$ lower level optimization problems (coupled through w). Any efficient algorithm for the solution of the experimental design problem should be able to handle this computational task.

3 Numerical solution of the optimization problem

3.1 The linearized case

We now review and propose some extensions for the linear/linearized case that has already been considered in [15]. If the regularization operator is quadratic of the form $S(m) = \|Bm\|_2^2$, then a number of simplifications can make the problem more tractable: First, consider a linearization of each nonlinear problem around the true model $\boldsymbol{\mu}$. This leads to an equation for the perturbation δm in terms of the data response $\tilde{d} = d - F(\boldsymbol{\mu})$:

$$\tilde{d} = J(\boldsymbol{\mu})\, \delta m + \mathcal{O}\left(\|\delta m\|^2 \right),$$

where $J(\boldsymbol{\mu}) \in \mathbb{R}^{n \times s}$ is the Fréchet derivative of F. By neglecting second order terms, the inverse problem for δm reduces to

$$\min_{\delta m} \; \frac{1}{2} \| \, W^{1/2} \left[\, J(\boldsymbol{\mu}) \delta m - \tilde{d} \, \right] \|^2 + \frac{1}{2} \| B(\boldsymbol{\mu} + \delta m) \|^2. \tag{3.8}$$

The best (true) solution of this problem is $\delta m = 0$. Nevertheless, it is obvious that such a solution is rarely obtained because the least square minimizer $\widehat{\delta m} = (J(\boldsymbol{\mu})^\top W J(\boldsymbol{\mu}) + B^\top B)^{-1} (J(\boldsymbol{\mu})^\top W \tilde{d} - B^\top B \boldsymbol{\mu})$ is generally nonzero.

The MSE of $\widehat{\delta m}$ can be decomposed into bias and variance components:

$$\begin{aligned} \mathrm{Bias}(w, \boldsymbol{\mu}) &= E\,(\widehat{\delta m}) - 0 = -[\, J(\boldsymbol{\mu})^\top W J(\boldsymbol{\mu}\,) + B^\top B \,]^{-1} B^\top B \boldsymbol{\mu} \\ \mathrm{Var}(w, \boldsymbol{\mu}) &= E \| \, \widehat{\delta m} \, - E\,(\widehat{\delta m}) \, \|^2 = \sigma^2 \, \mathrm{tr} \left[W J(\boldsymbol{\mu}) (J(\boldsymbol{\mu})^\top W J(\boldsymbol{\mu}) + B^\top B)^{-2} J(\boldsymbol{\mu})^\top W \, \right], \end{aligned}$$

where σ denotes the noise standard deviation. The linearized MSE for each fixed $\boldsymbol{\mu}$ is simply

$$\mathrm{MSE}(w, \boldsymbol{\mu}) = \| \, \mathrm{Bias}(w, \boldsymbol{\mu}) \, \|^2 + \mathrm{Var}(w, \boldsymbol{\mu}).$$

As discussed in Section 2, the variability of the MSE over the set of models $\mathcal{M}$ is controlled in an average sense. We use an empirical estimate based on an average of the linearized MSE of K training models $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$ from the prior π:

$$\widehat{R}_{\text{lin}}(w) \simeq \frac{1}{K} \sum_{j=1}^{K} \text{MSE}(w, \boldsymbol{\mu}_j). \tag{3.9}$$

An optimal design is then obtained by minimizing this average, linearized MSE with an ℓ_p penalty:

$$\begin{aligned} \min_{w} \quad & \widehat{R}_{\text{lin}}(w) + \beta \, \|w\|_p \\ \text{s.t} \quad & 0 \leq w. \end{aligned} \tag{3.10}$$

We have already designed algorithms for the solution of (3.10) for the case when F is linear [15]. Thus, it is straightforward to use our previous techniques to solve this problem.

3.2 The nonlinear case

Solving for w in the context of nonlinear experimental design requires the solution of bilevel optimization problems of the form (2.6). This task might be quite difficult because some of the components of the optimization problem may be non-differentiable or discontinuous, especially if the lower level problem is non-convex and includes inequality constraints. We therefore make some assumptions that significantly simplify the computations. Although these may seem to be strong assumptions, there is a broad range of inverse problems for which they are reasonable.

Assumptions: It is assumed that the lower level optimization problem

(A1) is convex and has a well defined minimum

(A2) includes no inequality constraints

The first assumption is rather restrictive; it implies that the underlying inverse problem has a unique solution for all possible designs. This assumption also implies that both F and S are twice differentiable. While this assumption is obviously false in general, it is valid in many practical situations. The second assumption prevents the use of bound constraints for the model recovery. Such constraints are important if one wishes to maintain continuous derivatives of the upper level optimization problem with respect to the weights. They are also important in applications as they provide means to include prior physical information.

Under assumptions A1 and A2, one can replace the bilevel optimization problem with the penalized empirical Bayes risk expression subject to an equality constraint as follows:

$$\min_{w, m_{ij}} \quad \widehat{R}_{\beta}(w, m_{ij}) = \frac{1}{LK} \sum_{i,j} \| \, m_{ij} - \boldsymbol{\mu}_i \|^2 + \beta \, \|w\|_p \tag{3.11a}$$

$$\text{s.t} \quad c_{ij} = c(m_{ij}, w) = J(m_{ij})^\top W[\, F(m_{ij}) - d_{ij} \,] + S'(m_{ij}) = 0 \tag{3.11b}$$

$$0 \leq w, \tag{3.11c}$$

where $d_{ij} = F(\boldsymbol{\mu}_i) + \epsilon_j$. In this subsection we use $p = 1$. An approximation for $p = 0$ is discussed in Section 3.3. There is a broad range of algorithms designed for nonlinear programming that could be used, at least in principle, to solve this problem. In fact, it can be posed as a control problem with m being the state and w the control. Nevertheless, two imperative characteristics of the problem should be addressed in the development of an efficient algorithmic framework:

- Realistically, the forward computation may be computationally intensive and therefore solving the constraint equation (the nonlinear inverse problem) $c(m_{ij}, w) = 0$ is typically a non-trivial task.
- The number of models and noise realizations can be quite large (in the thousands). Nevertheless, these problems are completely separable given w. This implies that a simple parallel or distributed implementation can be highly effective.

To address these issues we propose a solution based on a reduced space method in which m is eliminated from the equations and viewed as a function of w. Computation of the objective function can be performed by solving the constraints for a given w in parallel. Implicit differentiation can then be used to compute the gradient of the objective function with respect to w and, since the objective function is separable, these computations can also be performed in parallel.

A straightforward calculation shows that

$$\begin{aligned}
\frac{\partial c_{ij}}{\partial m_{ij}} &= J(m_{ij})^\top W J(m_{ij}) + S''(m_{ij}) + K_{ij} && \text{(3.12a)}\\
\frac{\partial c_{ij}}{\partial w} &= J(m_{ij})^\top \operatorname{diag}[\, F(m_{ij}) - d_{ij}\,], && \text{(3.12b)}
\end{aligned}$$

where K_{ij} stands for second order terms. The derivative of m_{ij} with respect to w can thereby be expressed as

$$\begin{aligned}
M_{ij} &:= \frac{\partial m_{ij}}{\partial w} = -\left(\frac{\partial c_{ij}}{\partial m_{ij}}\right)^{-1} \frac{\partial c_{ij}}{\partial w} && \text{(3.13)}\\
&= -\left[\, J(m_{ij})^\top W J(m_{ij}) + S''(m_{ij}) + K_{ij}\,\right]^{-1} J(m_{ij}\,)^\top \operatorname{diag}\left[\, F(m_{ij}) - d_{ij}\,\right].
\end{aligned}$$

It is important to note that the matrix $\partial c_{ij}/\partial m_{ij} \in \mathbb{R}^{LK\times LK}$ in (3.12a) is invertible because of our convexity assumption of the lower problem with a well defined minimum (A1).

Given the derivatives of m_{ij}, w can be easily updated using a variety of available methods. The gradient required for the update is obtained by differentiation of the reduced objective function (3.11a)

$$\nabla_w R_\beta(w, m_{ij}(w)) = \frac{1}{LK} \sum_{i,j} M_{ij}^\top\, (\, m_{ij} - \boldsymbol{\mu}_i) + \beta\, e, \tag{3.14}$$

where e denotes a vector of ones.

The computation of the model $m_{ij}(w)$ requires the solution of a nonlinear problem, while the calculation of the gradient requires the computation of products of the form $M_{ij}^\top z$, with

$z = m_{ij} - \boldsymbol{\mu}_i$. It is worth noting that this calculation does not require the explicit computation or construction of M_{ij}. Instead, it is possible to compute the product in three steps: (i) solve the system

$$[\, J(m_{ij})^\top W J(m_{ij}) + S''(m_{ij}) + K_{ij} \,]\, y = z; \tag{3.15}$$

(ii) Compute the product $J(m_{ij})\, y$, and (iii) pointwise multiply $J(m_{ij})y$ by the vector $F(m_{ij}) - d_{ij}$. Efficient methods for inversion of the matrix (3.15) are based on Krylov methods [16], which do not require the matrix $J(m_{ij})$ but rather the computation of products of $J(m_{ij})$ and $J(m_{ij})^\top$ with a vector. Typically, the matrix $S''(m_{ij})$ is used as a preconditioner to accelerate convergence [16], although other preconditioners may also be considered [18].

In contrast to our previous work on experimental design for non-linear inversion [19], one of the advantages of the formulation presented here is that the sensitivity relations of the inverse problem with respect to the design parameters are merely accessed in the form of matrix-vector products. This property is particularly valuable for large-scale problems and for any other situation where the sensitivity relations are not given explicitly by the code employed. Thus, it is possible to use the proposed design algorithm with any currently available inverse solver that complies with the assumptions A1 and A2.

Once the gradient is obtained, the facilitation of the projected gradient method for solving the design optimization problem is straightforward [24]. The algorithm for the solution of the problem is described in Algorithm 1.

Algorithm 1 Optimal design

```
(Initialization)
  Set $w = 1, \quad k = 1$
while (non convergence) do
    Solve equation (3.11b) for all $m_{ij}(w)$ in parallel
    Compute $r_{ij} = F(m_{ij}) - d_{ij}$
    Solve $[\, J(m_{ij})^\top W J(m_{ij}) + S''(m_{ij}) + K_{ij} \,]^{-1} y_{ij} = m_{ij} - \boldsymbol{\mu}_i$ in parallel
    set $\nabla_w R_\beta = \sum_{i,j} r_{ij} \odot (J(m_{ij}) y_{ij}) + \beta e$
    Project the update pointwise: $w_{k+1} = \max(w_k - \gamma \nabla_w R_\beta, 0)$
end while
```

It is important to note that the main computational effort in execution of this algorithm is devoted to the solution of the $L \times K$ nonlinear equations for the models m_{ij} and the $L \times K$ linear equations of the gradient. However, these computations can be easily parallelized. Given a multi-processor architecture, the computation time for the solution is dominated by our ability to solve the (nonlinear) inverse problem.

3.3 Approximating the ℓ_0 solution

The minimization of $\widehat{R}_\beta(w) = \widehat{R}(w) + \beta \|w\|_1$ was discussed in Section 3.2 and justified by the well known property of the ℓ_1-penalty in promoting sparsity of the solution. Alternatively, one may consider a penalty that leads to the sparsest design; for example, the minimization of $\widehat{R}_\beta(w) = \widehat{R}(w) + \beta \|w\|_0$, where the ℓ_0 penalty is defined as the number of nonzero entries of w. Since the solution of this problem is NP hard, we use a strategy suggested in [5]. The basic idea is to first solve the ℓ_1-norm problem, while keeping only the estimated nonzero entries in w. Let I_Z and I_N be, respectively, the sets of indices corresponding to zero and nonzero entries of w. The following optimization problem is used to approximate the ℓ_0 solution:

$$\begin{aligned} \min \quad & \widehat{R}(w(I_N)) \\ \text{s.t} \quad & w(I_Z) = 0. \end{aligned} \tag{3.16}$$

By construction, this ℓ_0 approximation has the same sparsity pattern as the ℓ_1 solution but with possibly different values. To implement the ℓ_0 approximation we define the matrix P_N such that

$$P_N w = w(I_N);$$

that is, the matrix P_N selects the set of indices that are non-zero at the solution. We then proceed by minimizing the empirical Bayes risk without any explicit regularization term

$$\begin{aligned} \min \quad & \widehat{R}(P_N w) \\ \text{s.t} \quad & 0 \leq w. \end{aligned}$$

From a technical standpoint, it means that the algorithms and software used for the ℓ_1 design problem can be used to solve the approximated ℓ_0 design. The only required modifications are: setting the regularization parameter to zero (i.e., $\beta = 0$) and reducing the experiment set to only include the active ones.

3.4 Controlling the sparsity of w

Since the choice of the ℓ_1 regularization parameter β controls the sparsity of w, it is natural to look for ways to select β that lead to sparse optimal designs. Here we discuss an approach that utilizes our objective function.

In principle, it is possible to determine the sparsity of w a-priori. For example, one may decide ahead of time that only t experiments are to be conducted. While this approach has been used in the past, we have found it to be rather limited. To see why, Figure 1 shows a typical curve of the empirical Bayes risk as a function of the non-zero entries in w. Since this curve represent a tradeoff between the number of nonzeros of the solution and the empirical Bayes risk as a function of β, it is often referred to as a *Pareto curve*. Now assume that one wishes to conduct only a small number of experiments and that optimal weights for these experiments have been obtained. When this number is small, then it is evident from the curve that a small increase in the number of experiments can improve the results substantially. On the other hand, if the number of experiments to be conducted is relatively large, then it is possible to select fewer experiments with no significant loss in the quality of the recovered models. It stands to reason that a reasonable tradeoff between multiple experiments and minimal risk is achieved when the number of experiments is close to the value corresponding to the curve's corner; in this case increasing the number of experiments does not significantly improve the outcome of the experiment while a reduction

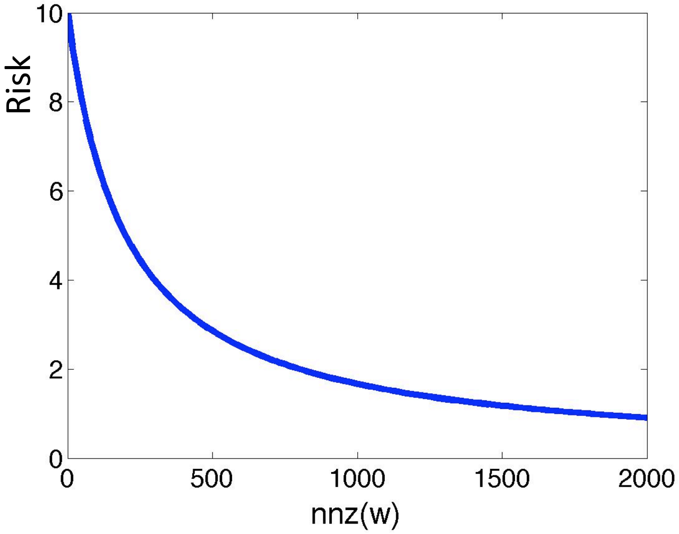


Figure 1: A Pareto curve: The empirical Bayes risk as a function of the number of nonzeros in w.

in the number of experiments significantly deteriorates the quality. The drop of the curve depends on various factors, such as the amount of information that each experiment adds compared to its noise contribution. Thus, a plot of the risk as a function of the sparsity of w can be highly informative in designing efficient experiments.

The Pareto curve is typically obtained using a continuation process [13, 26]. One starts by solving the problem with a large β to obtain the optimal solution $w^*(\beta)$; β is then decreased and the optimal value is updated.

In essence, one can spare the repeated computations of the design problem for a range of β values and instead attempt to recover an optimal value of β by means of parameter estimation. This can be done by computing the sensitivity of the regularized empirical Bayes risk to β and deriving the maximal curvature value as a sensible trade-off point. While such approach may seem appealing at first glance, the computation of such sensitivities is not trivial, especially when the inner optimization level includes realistic, non-linear, PDE-based problems.

4 Applications to joint inversion of electromagnetic data

In this section we describe an application of experimental design to geophysical electromagnetic experiments. The forward problem for a typical direct current (DC) resistivity experiment is given by the partial differential equation

$$\nabla \cdot m \nabla \phi = q(\mathbf{x}) \quad \phi \in \Omega; \qquad \nabla \phi \cdot \mathbf{n} = 0 \quad \phi \in \partial\Omega,$$

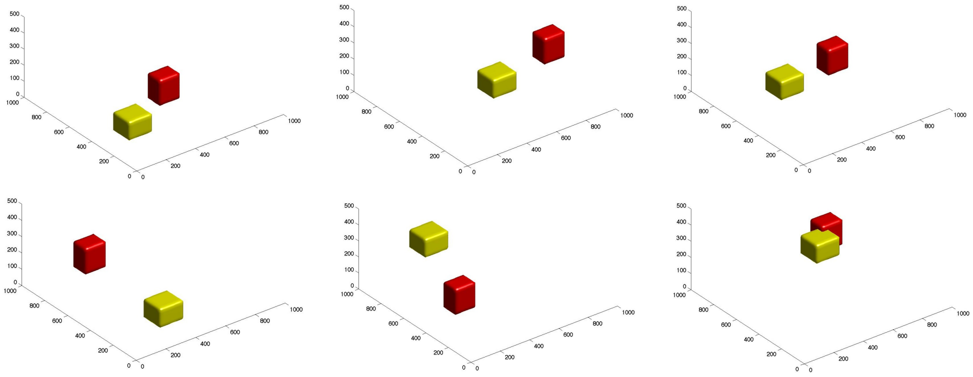


Figure 2: The training models used in the AMT/DC experiment.

where m is the conductivity, ϕ is the potential field and q is a vector of sources. The data are typically gradient measurements of the potential field. Upon discretization of the PDE, the forward model can be written as

$$F(m;\beta) = QA(m)^{-1}q,$$

where Q projects the electric potentials onto the measurement locations and A is the discretization of the operator $\nabla\cdot m\nabla\phi$ given by a sparse matrix that depends on the conductivity m.

A complementary way to achieve a similar goal is using a magnetotelluric (MT) experiment in which electromagnetic waves at different frequencies interact with the earth generating electric and magnetic fields that can be recorded. In this case the forward model is obtained from Maxwell's equations in the quasi-static regime:

$$\nabla\times\,\mu^{-1}\nabla\times\,\vec{E} - i\beta m\vec{E} = i\beta\vec{s} \quad \vec{E}\in\Omega; \qquad \nabla\times\,\vec{E}\times\mathbf{n} = 0 \quad \vec{E}\in\partial\Omega,$$

where $\vec{E}$ is the electric field, $\vec{s}$ are sources, μ the magnetic permeability, m is the conductivity, β the frequency and $\mathbf{n}$ is a unit vector in the normal direction. Upon discretization of Maxwell's equations using finite volume or finite elements [12, 22], the forward model can be written as

$$F(m;\beta) = Q_\beta A_\beta(m;\beta)^{-1}(i\beta s),$$

where A_β is a sparse matrix that depends on the conductivity and the frequency. The matrix Q_β projects the electric field everywhere to the measurement locations [13].

Equipment that can record both DC and audio magneto-telluric (AMT) signals has been recently developed. The experimental design question is to determine how many and what frequencies should be recorded to obtain the best possible experiment. In particular, we try to determine if the DC measurements provide important information for the the recovery process.

To setup the problem we assume six training models in the domain $[0, 10^3]\times[0, 10^3]\times[0, 5\times10^2]$. The models are plotted in Figure 2. Each model is composed of 2 conductive cubes randomly embedded within the media. We then divide the domain into $64\times64\times32$

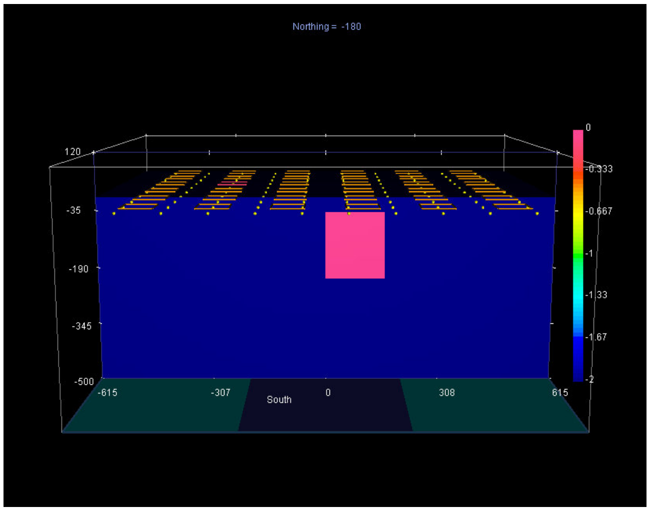


Figure 3: A schematic description of the locations of the receivers for the AMT/DC experiment

cells and solve the forward problem on this mesh. A thorough discussion on the forward modeling and inversion codes used for this experiment can be found in [14]. For each model we computed two sets of data with two different noise realizations. Although the number of noise realizations is low, such number is often used to evaluate the variance of linear inverse problems [11]. The standard deviation of the noise was set to 1% of the average of $|F(m)|$. For the data, we discretized the frequency range $[10^0, 10^4]$ using 129 points evenly spaced in log scale. The goal is to pick a subset of frequencies in this range. For each frequency, we record 900 measurements of the z component of the magnetic field on top of the earth, as shown in Figure 3. We then use these data and models to compute the Pareto curve of empirical Bayes risk vs the number of experiments. This curve is shown in Figure 4.

Among all the ‘kinks’ observed in the curve, the one that makes the most sense happens when the number of nonzeros in w is around 42. At this point, conducting more experiments leads to a mild amelioration of the empirical Bayes risk whilst performing fewer experiments increases the empirical Bayes risk dramatically. We therefore choose this to be the optimal design. The frequencies and corresponding weights for this design are shown in Figure 5. It is clear that the zero frequency has a nonzero weight and therefore an optimal design includes the DC response. To show that our design indeed yields better results when applied to a model that was not used for determining optimal experimental design, we compare the latter to a commonly used design; one where the frequencies are equally sampled in a log scale. As the results in Figure 6 show, better results can be obtained from a well designed survey; the recovered model is evidently closer to the ‘true’ model in the linearized and nonlinear designs. The results obtained with the linearized design are also shown in the figure. This design does lead to an improvement over the traditional one but the results of the nonlinear design are still better.

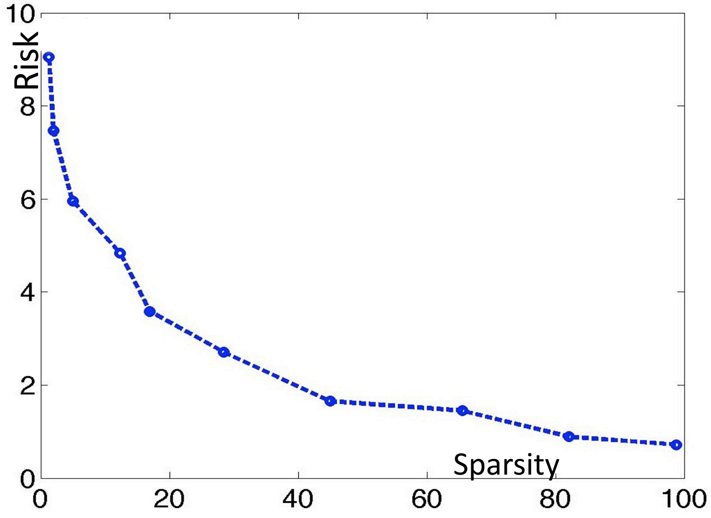


Figure 4: The Pareto curve for the AMT/DC experiment described in Section 4. It shows the empirical Bayes risk as a function of the number of frequencies to be recorded.

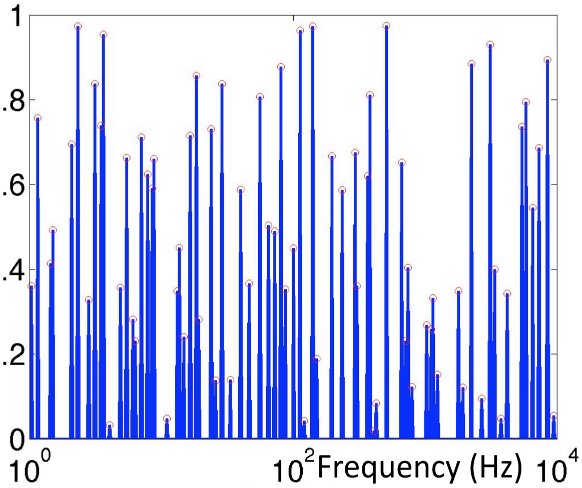
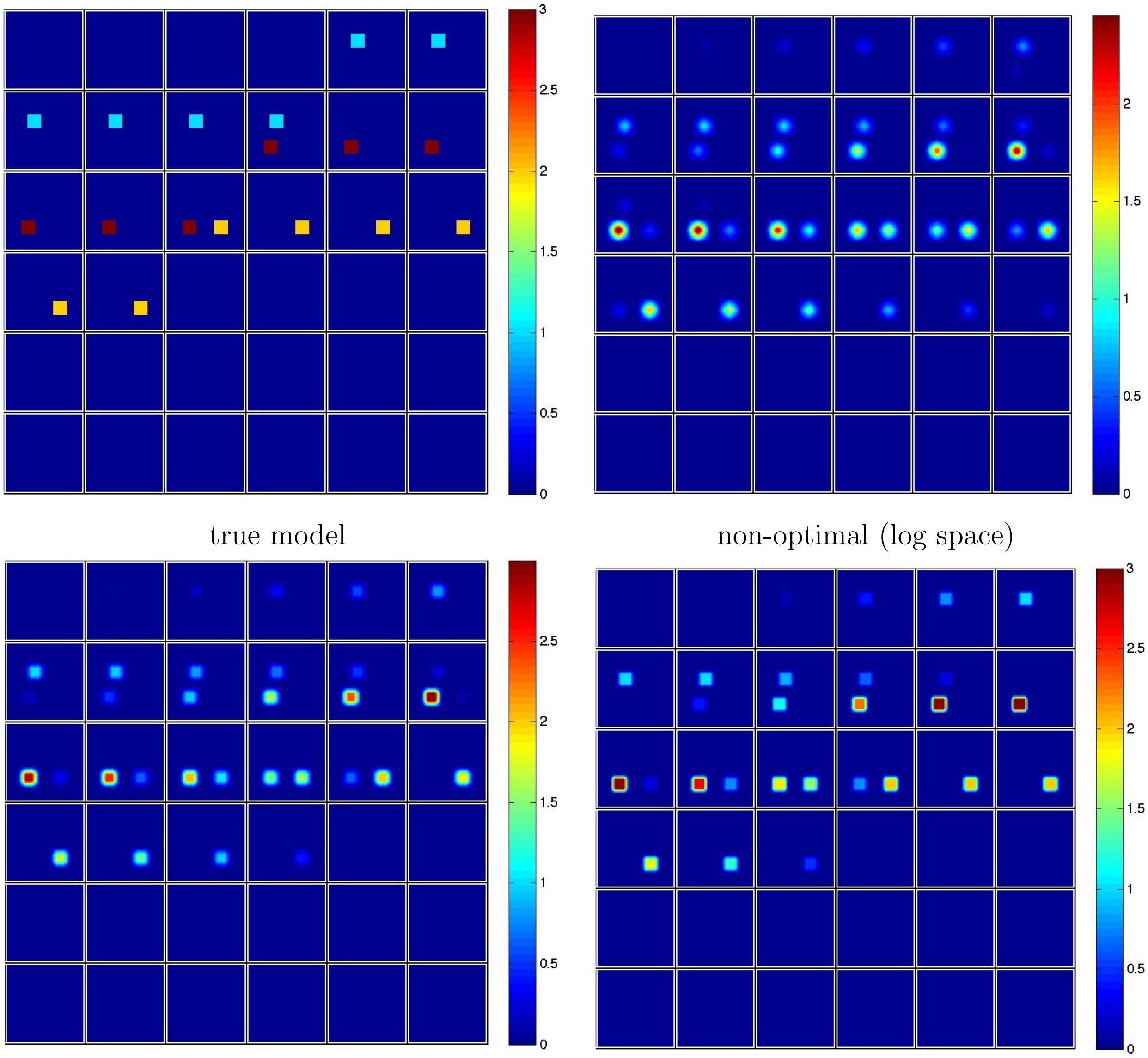


Figure 5: Optimal set of 42 out of 129 frequencies, the y axis is the magnitude of w_i.



true model

non-optimal (log space)

linearized optimal

optimal

Figure 6: Thirty two slices in depth of the true model (top left) and the results of the non-optimal, linearized optimal and optimal designs.

5 Summary

We have presented a methodology for experimental design of discrete nonlinear ill-posed problems that can be summarized as follows: The data consist of indirect, noisy observations of the model, provided by a set of plausible experiments under consideration. An inversion estimate based on these data is obtained by a weighted Tikhonov regularization whose weights control the contribution of the different experiments to the data misfit. These weights are selected by minimizing an empirical estimate of the Bayes risk subject to an ℓ_p penalty that promotes sparsity of the collection of chosen experiments. The problem so defined is a difficult bilevel optimization problem. However, we have shown that a descent method based on the sensitivities can be efficiently used to tackle the problem in practice. We have demonstrated the viability of our design to a problem in electromagnetic imaging based on direct current resistivity and magnetotelluric data.

We have made an effort to use as few new functions and parameters as possible; that is, to implement our algorithm one only requires functions and parameters already computed for the inverse problem. This makes our approach ideal for many nonlinear design problems. We further presented and explored the tradeoff curve between cost (number of experiments) and empirical Bayes risk. We believe that this curve is important in realistic applications where the number of essential experiments needs to be determined and the final design is intended to aid in the decision making process.

There is still work to be done. In some applications one may wish to take into account other factors like operational costs and availability of resources. In practical scenarios one can obtain an actual cost function by assessing the real cost (in time or money) of each experiment. For example, in geophysics the cost of a borehole experiment may be dictated by the depth of the borehole. Since deeper boreholes cost more, a cost function that can account for such expense can and should be used if practical solutions are to be obtained from the experimental design.

Finally, we have not considered the quality of the estimate obtained by minimizing the empirical Bayes risk. The selection of the weights is based on the minimization of empirical averages over noise and model samples. This is a typical application of empirical risk minimization [6, 32, 33]. Important considerations include the behavior of estimates as a function of sample size, for example, consistency and existence of exponential bounds for fixed samples, as well as more practical validation procedures [32]. Implementing such procedures in a computationally reliable way is a challenge that we intend to address in future work.

Acknowledgements. Our work has been partially supported by the National Science Foundation (DMS 0724717, 0724715, 0914987, 0724759, 0915121 and CCF 0728877).

References

[1] J. Bee Bednar, L.R. Lines, R.H. Stolt & A.B. Weglein Geophysical Inversion, SIAM, Philadelphia, PA, USA, 1992.

[2] E. van den Berg & M.P. Friedlander Probing the Pareto frontier for basis pursuit solutions. *SIAM J. on Scientific Computing*, 31:890–912, 2008.

[3] G. Biros & O. Ghattas. Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization Parts I & II. SIAM J. on Scientific Computing, 27, 2:687–739, 2005.

[4] J.P. Boyd. *Chebyshev & Fourier Spectral Methods.* Springer, 1989.

[5] S. Boyd & L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[6] L. Devroye, L. Györfi & G. Lugosi *A Probabilistic Theory of Pattern Recognition* Springer, 1996.

[7] H. W. Engl, M. Hanke & A. Neubauer *Regularization of Inverse Problems* Kluwer, 2000.

[8] V.V. Fedorov & P. Hackl. *Model-Oriented Design of Experiments.* Springer (Lecture Notes in Statistics), 1997.

[9] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright. Gradient projection for sparse reconstruction: Applications to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1:586–597, 2007.

[10] W.P. Gouveia & J.A. Scales. Resolution of seismic waveform inversion: Bayes versus Occam. *Inverse Problems*, 13:323–349, 1997.

[11] G. Golub & U. von Matt. Quadratically constrained least squares and quadratic problems. *Numer. Math.*, 59:561–580, 1991.

[12] E. Haber & U. Ascher. Preconditioned all-at-one methods for large, sparse parameter estimation problems. *Inverse Problems*, 17:1847–1864, 2001.

[13] E. Haber, U. Ascher & D. Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse problems*, 16:1263–1280, 2000.

[14] E. Haber, U. Ascher & D. Oldenburg. Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. *Geophysics*, 69:1216-1228, 2004.

[15] E. Haber, L. Horesh & L. Tenorio. Numerical methods for experimental design of large-scale linear ill-posed inverse problems. *Inverse Problems*, 24:055012, 2008.

[16] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems.* SIAM, Philadelphia, 1997.

[17] M. Heinkenschloss & L.N. Vicente. Analysis of inexact trust region SQP algorithms. *SIAM J. Optimization*, 12:283–302, 2001.

[18] L. Horesh, M. Schweiger, M. Bollhfer, A Douiri, A., S. Arridge & D.S. Holder. Multilevel preconditioning for 3D large-scale soft field medical applications modelling. *Information and Systems Sciences*, 532–556, 2006.

[19] L. Horesh, E. Haber & L. Tenorio. Optimal experimental design for the large-scale nonlinear ill-posed problem of impedance imaging. In *Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty.* Wiley, 2009 (To appear).

[20] L. Horesh and E. Haber. Sensitivity computation of the ℓ_1 minimization problem and its application to dictionary design of ill-posed problems. *Inverse Problems*, 25:095009, 2009.

[21] V. Isakov *Inverse Problems for Partial Differential Equations.* Springer, 2003.

[22] J. Jin. *The Finite Element Method in Electromagnetics.* Wiley, 1993.

[23] T.A. Johansen. On Tikhonov regularization, bias and variance in nonlinear system identification. Automatica, 33:441–446, 1997.

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

[25] T. Madden & R. Mackie. Three-dimensional magnetotelluric modeling and inversion. *Proceedings of the IEEE*, 77:318–321, 1989.

[26] J. Nocedal & S. Wright. *Numerical Optimization.* Springer, New York, 1999.

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

[28] F. Pukelsheim. *Optimal Design of Experiments.* Wiley, 1993.

[29] D.B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12:1151–1172, 1984.

[30] M.D. Sacchi & T.J Ulrych Improving resolution of Radon operators using a model re-weighted least squares procedure. *Journal of Seismic Exploration*, 4:315–328, 1995.

[31] M.A. Saunders S.S. Chen & D.L. Donoho. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1996.

[32] A. Shapiro, Dentcheva D. & A. Ruszczynski. Lectures on Stochastic Programming: Modeling and Theory SIAM, 2009.

[33] V.N. Vapnik. *Statistical Learning Theory.* Wiley, 1998.

[34] K.P. Whittall & D.W. Oldenburg. *Inversion of Magnetotelluric Data for a One Dimensional Conductivity.* Society of Exploration Geophysicists, Geophysical Monograph Series, 5, 1992.