Adaptive A-optimal experimental design for imaging dynamic targets

J. Fohring and E. Haber

April 2, 2015

Abstract

Efficiently monitoring subsurface flow using geophysical imaging techniques is important to several applications such as aquifer characterization or enhanced oil recovery projects. One method of increasing monitoring efficiency is to control the number of measurements required for a geophysical survey in an optimal way. To optimally design future geophysical experiments, we use A-optimal design methods while incorporating the flow dynamics.

A-optimal design methods minimize the posterior covariance matrix associated with estimating the target subsurface model. We reformulate the MAP estimate for the prior target in order to design for the future target, such that the problem is constrained by the flow dynamics. In this way, the prior experiments collected data and flow dynamics contribute to the design of the next experiment, thus incorporating updated information to the design algorithm.

The method is demonstrated for a two dimensional borehole seismic tomography experiment and a three dimensional seismic survey. An advection model is applied to describe subsurface flow and propagate the error covariance matricies forward in time.

1 Introduction

- Monitoring; Many problems in geophysics address the issue of monitoring subsurface flow, such as aquifer characterization, enhanced oil recovery, or saltwater intrusion. In general one would like to image the motion of certain subsurface parameters over a period of time. One method to do so is to collect measurements dependent on these parameters and invert through a mathematical relationship to recover parameter distributions. When monitoring for these dynamic parameters, the choice of where the data are collected becomes important as resolution is both temporally and spatially dependent.

- much has been done in experimental designs for linear and nonlinear problems...A optimal, etc.. appropriate references. As well as for large scale ill posed, refs. inverting

for fluid parameters given geophysical imaging...refs...talk about seismic, and helmoltz and simultaneous sources

- Determining the design of the data collection experiment requires knowledge of the target parameter distribution. Typically optimal experimental design methods rely on minimizing the posterior error covariance matrix, associated with estimating the target through the inversion of geophysical data. This poses a problem, particularly for static cases, since the point of the data collection is to image the unknown target. However, if the dynamics governing the motion of the unknown target are known, even to a small degree in accuracy, then it is possible to design the data collection experiment to be conducted in the future as the target moves. In order to design for a future experiment it is necessary to know the initial distribution of target parameters as well as the dynamics. If the initial condition is provided with the dynamics, then either an iterative approach or and all at once approach can be taken to determine date collection designs for a set of time points.

- ideally, incorporation of real data in the estimation of historic, but also future experimental designs would lead to the best estimates. Adaptively, chase the target, by recomputing the design after data is collected.

- We also consider what happens if we wish to recover/update the parameters governing the fluid flow. How does this change designs for the future time step?

- computational limitations: (constrained all at once, no error) no historic info included, (Kalman formulation, error) covariance matrices are large and dense, making factorization/ reduction techniques difficult. geophysical experiments are computationally expensive. (adaptive, constrained) no covariance matrices to worry about, includes last data set/design, can easily tack on regularization.

As a model problem we consider the following dynamical system of simple mass transport, where the dynamics is given by

$$m_t + \nabla \cdot \vec{\mathbf{u}} m = 0 \quad m(0, \vec{\mathbf{x}}) = m_0 \tag{1}$$

2 A-Optimal experimental design

The section outlines the optimal experimental design criteria chosen to compute optimal experimental designs for a given model estimation problem, along with the design optimization objective function, and sparsity constraints.

The general approach to compute optimal experimental designs for inverse problems is to minimize the error contained in the covariance matrix of the estimate of the unknown model parameters. There are many different optimization criteria for minimizing the error covariance matrix, such as A,D, or E- optimality to name a few. Here we choose Bayesian

A-optimal experimental design, which minimizes the trace of the error covariance matrix, as it is somewhat easier to deal with numerically.

Consider the discrete static ill-posed inverse problem of estimating model parameters $\mathbf{m} \in \mathbb{R}^N$ from noisy observation data $\mathbf{d} \in \mathbb{R}^M$, where the data are given by

$$\mathbf{d} = \mathbf{Fm} + \boldsymbol{\epsilon}.$$

The imaging forward operator $\mathbf{F} : \mathbb{R}^N \to \mathbb{R}^M$ is an $M \times N$ matrix with each row representing an experiment. The data space $M = ns \times nr$ is determined by the maximum number of measurements allowed.

Assuming a Bayesian inversion approach is taken, where the posterior estimate $\widehat{\mathbf{m}}$ is conditioned on a known prior distribution $\pi_m = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and that the error $\boldsymbol{\epsilon}$ is Gaussian $N(0, \vec{W}^{-1})$, the Bayesian posterior estimate is then obtained by minimizing the negative log-likelihood of Bayes rule

$$
\begin{aligned}
\widehat{\mathbf{m}}(\mathbf{w}) &= \operatorname{argmin} \frac{1}{2} \|\mathbf{Fm} - \mathbf{d}\|_{\mathbf{W}}^2 + \frac{1}{2} \|\mathbf{m} - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}^2 \\
&= (\mathbf{F}^\top \mathbf{WF} + \boldsymbol{\Sigma}^{-1})^{-1} (\mathbf{F}^\top \mathbf{Wd} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}).
\end{aligned}
$$

The weighting matrix $\mathbf{W} = \operatorname{diag}[\mathbf{w}]$ with $\mathbf{w} = (\omega_1, .., \omega_M)^\top$, relates to the data variance as $\operatorname{Var}(d_i) = 1/\omega_i$.

A-optimal experimental designs are defined by minimizing the average variance of the estimate $\widehat{\mathbf{m}}$, thus minimizing the trace of the posterior covariance matrix of $\widehat{\mathbf{m}}$ [?], which leads to the following A-optimal design objective function

$$\phi_{A_B}(\mathbf{w}) = \operatorname{tr}[(\mathbf{F}^\top \mathbf{WF} + \boldsymbol{\Sigma}^{-1})^{-1}]. \tag{2}$$

Given ϕ_{A_B}, an optimal design is computed by solving a sparsity controlled optimal design optimization problem [?]. The design weights ω_i are determined by minimizing $\phi(\mathbf{w})$ with an l_1 penalty on $\mathbf{w}$ to promote sparsity,

$$
\begin{aligned}
\widehat{\mathbf{w}} &= \operatorname{arg\,min} \phi(\mathbf{w}) + \beta \|\mathbf{w}\|_1 \\
\text{s.t.} \quad & 0 \geq \omega_i.
\end{aligned}
\tag{3}
$$

It is important to note that in this formulation $\widehat{\mathbf{m}}$ is chosen to be a linear (or affine) function of the data, although other estimators could be used, such as a Tikhonov estimator. Since the idea is to design experiments at future time steps for linear dynamic problems where the estimators covariance matrix can be estimated, the simple Bayesian estimator is sufficient.

3 Designing experiments for dynamic targets

Let us consider the case of design for a dynamical system. There are two main cases to be considered. First, if the dynamic is fast compared to the experimental time then, the design needs to be done prior to t_0, that is, prior to the evolution of the system. Such a design has been addressed in [**?**] and it is a direct use of the standard A optimal design given the dynamics. A second case that needs to be considered is when the dynamics is slow compared to the experimental time. In this case one has the time to collect the data and process it and then change the experimental design given the estimation of the model.

To begin, consider the discrete dynamic process and imaging described by the following set of linear equations

$$
\begin{align}
\mathbf{m}_k &= \mathbf{T}\mathbf{m}_{k-1} + \boldsymbol{\eta}_k, \tag{4a} \\
\mathbf{d}_k &= \mathbf{F}\mathbf{m}_k + \boldsymbol{\epsilon}_k, \tag{4b}
\end{align}
$$

where transport matrix $\mathbf{T} : \mathbb{R}^N \to \mathbb{R}^N$ is a discretization of the dynamic process and the noise vectors η and ϵ are assumed to be Gaussian normal, $\eta \sim N(\eta_0, \boldsymbol{\Sigma}_\eta)$ and $\boldsymbol{\epsilon} \sim N(\epsilon_0, \mathbf{W}^{-1})$, for $k = 1, .., n$ time points.

We further consider two cases. First, if the dynamical system is well understood and modeled in high accuracy then we can ignore $\boldsymbol{\eta}_k$. In this case, if $\mathbf{m}_0$ is known then the whole dynamics is determined and therefore, this case leads to the estimation of the initial condition $\mathbf{m}_0$ only. A second case is that we assume some noise in the dynamics. In this case we need to take the dynamical noise into consideration and need to design and experiment to estimate the model $\mathbf{m}_k$ given the the previous estimate $\mathbf{m}_{k-1}$ and the its variance. We now discuss these cases.

3.1 Design for no noise in dynamics

Beginning with the simplest case, where $\eta_k = 0$ and assuming n time steps we can write the dynamics as block system of the form

$$
\underbrace{\begin{pmatrix} -\mathbf{I} & & & & \\ \mathbf{T} & -\mathbf{I} & & & \\ & \mathbf{T} & -\mathbf{I} & & \\ & & \ddots & \ddots & \\ & & & \mathbf{T} & -\mathbf{I} \end{pmatrix}}_{=:\mathbf{A}_n} \underbrace{\begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \mathbf{m}_3 \\ \vdots \\ \mathbf{m}_n \end{pmatrix}}_{=:\mathbf{m}_{[1:n]}} + \underbrace{\begin{pmatrix} \mathbf{T} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{=:-\mathbf{B}_n} \mathbf{m}_0 = \mathbf{0}. \tag{5}
$$

Thus, the set of time dependent models can then be written as the solution to the simple linear system

$$
\mathbf{m}_{[1:n]} = \mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{m}_0. \tag{6}
$$

Now, consider that we have done $n - 1$ experiments, obtained $n - 1$ data vectors and that we are seeking to conduct the n-th experiment in order to better evaluate model, $\mathbf{m}_0$ at time

n. The forward problem for all times can be written as

$$
\begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_n \end{pmatrix} = \begin{pmatrix} \mathbf{F}_1 & & & \\ & \mathbf{F}_2 & & \\ & & \ddots & \\ & & & \mathbf{F}_n \end{pmatrix} \begin{pmatrix} -\mathbf{I} & & & \\ \mathbf{T} & -\mathbf{I} & & \\ & \mathbf{T} & -\mathbf{I} & \\ & & \ddots & \ddots & \\ & & & \mathbf{T} & -\mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{T} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \mathbf{m}_0 \tag{7}
$$

In the case that $\mathbf{F}_k = \mathbf{F}$ we can write the forward problem as

$$
\mathbf{d}_{[1:n]} = (\mathbf{I}_n \otimes \mathbf{F})\mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{m}_0 \tag{8}
$$

Let $\mathbf{W}_k$ be the variance of the data at time step k and let $\mathbf{W}_{[1:n]} = \mathsf{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_{n-1}, \mathbf{W}_n)$. Since the experiments at times $1, \ldots, n-1$ have already been conducted, the variances, $\mathbf{W}_1, \ldots, \mathbf{W}_{n-1}$ are set. For the n-th experiment we can control only the variance of the last time step, $\mathbf{W}_n = \mathsf{diag}(\mathbf{w})$.

In Bayesian design, assuming a covariance matrix of the form $\boldsymbol{\Sigma}_{\mathbf{m}_0}^{-1} = \mathbf{L}^\top\mathbf{L}$, the covariance matrix of the estimated model at time n can be directly written as

$$
\boldsymbol{\Sigma}_n = (\mathbf{B}_n^\top\mathbf{A}_n^{-\top}(\mathbf{I}_n \otimes \mathbf{F})^\top\mathbf{W}_{[1:n]}(\mathbf{w})(\mathbf{I}_n \otimes \mathbf{F})\mathbf{A}_n^{-1}\mathbf{B}_n + \mathbf{L}^\top\mathbf{L})^{-1}, \tag{9}
$$

where we defined $\mathbf{W}_{[1:n]}(\mathbf{w}) = \mathsf{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_{n-1}, \mathsf{diag}(\mathbf{w}))$. A optimal Bayesian design minimizes the trace of the covariance matrix at time n.

An interesting observation is to be made here. The A-optimal Bayesian design does not take the data that was collected from time 1 to $n-1$ in order to obtain the design for time n. This is a serious flaw of the Bayesian design. Consider that the data indicates that $\mathbf{m}_0$ is concentrated in a certain region, then, there is no reason to design an experiment that captures a general $\mathbf{m}_0$. The important point to be made here is that the data allows us to re-estimate the model $\mathbf{m}_0$ beyond our initial knowledge and focus our efforts its evaluation. indeed, given the data $\mathbf{d}_{[1:n-1]} = [\mathbf{d}_1^\top, \ldots, \mathbf{d}_{n-1}^\top]$ the estimated model at time $n-1$ given the data is

$$
\begin{aligned}
\widehat{\mathbf{m}}_0^{(n-1)} =\ & (\mathbf{B}_{n-1}^\top\mathbf{A}_{n-1}^{-\top}(\mathbf{I}_{n-1} \otimes \mathbf{F})^\top\mathbf{W}_{[1:n-1]}(\mathbf{I}_{n-1} \otimes \mathbf{F})\mathbf{A}_{n-1}^{-1}\mathbf{B}_{n-1} + \boldsymbol{\Sigma}_{\mathbf{m}_0})^{-1} \\
& \mathbf{B}_{n-1}^\top\mathbf{A}_{n-1}^{-\top}(\mathbf{I}_{n-1} \otimes \mathbf{F})^\top\mathbf{W}_{[1:n-1]}\mathbf{d}_{[1:n-1]}
\end{aligned} \tag{10}
$$

For time t_n we propose to modify the optimization problem, taking the estimated model into consideration and minimizing

$$
\min_{\mathbf{m}_0} \frac{1}{2}\|(\mathbf{I}_n \otimes \mathbf{F})\mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{m}_0 - \mathbf{d}_{[1:n]}\|_{\mathbf{W}_{[1:n]}}^2 + \frac{1}{2}\|\mathbf{L}(\mathbf{m}_0 - \mathbf{m}_0^{(n-1)})\|^2 \tag{11}
$$

5

Assuming that the imaging technique does not change in time, the negative log data likelihood is given by

$$\frac{1}{2}\left\|(\mathbf{I}\otimes\mathbf{F})\mathbf{A}(\mathbf{u})^{-1}\mathbf{B}(\mathbf{u})\mathbf{m}_0 - \bar{\mathbf{d}}\right\|_{\overline{\mathbf{W}}}^2 \tag{12}$$

where, $\overline{\mathbf{W}}$ is a block diagonal matrix with diagonal weight matrices $\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n$ on its main diagonal for time points $k = 1, 2, ..., n$.

Given a prior distribution for $\overline{\mathbf{m}} \sim N(\overline{\mathbf{m}}_0, \boldsymbol{\Sigma}_{\overline{\mathbf{m}}})$, with $\boldsymbol{\Sigma}_{\overline{\mathbf{m}}} = \text{blkdiag}(\boldsymbol{\Sigma}_{\eta_1}, .., \boldsymbol{\Sigma}_{\eta_n})$, the posterior covariance matrix for the Bayes estimate $\widehat{\mathbf{m}}_0$ is

$$\boldsymbol{\Sigma}_{\widehat{\mathbf{m}}_0} = \left(((\mathbf{I}\otimes\mathbf{F})\mathbf{A}^{-1}\mathbf{B})^\top \, \overline{\mathbf{W}} \, (\mathbf{I}\otimes\mathbf{F})\mathbf{A}^{-1}\mathbf{B} + \boldsymbol{\Sigma}_{\overline{\mathbf{m}}}^{-1}\right)^{-1} \tag{13}$$

Thus the Bayesian A-optimal design functional becomes

$$\phi_{A_B}(\overline{\mathbf{w}}) = \text{tr}\left[\left(((\mathbf{I}\otimes\mathbf{F})\mathbf{A}^{-1}\mathbf{B})^\top \, \overline{\mathbf{W}} \, (\mathbf{I}\otimes\mathbf{F})\mathbf{A}^{-1}\mathbf{B} + \boldsymbol{\Sigma}_{\overline{\mathbf{m}}}^{-1}\right)^{-1}\right], \tag{14}$$

where the designs recovered are for all times points $k = 1, ..., n$. This method was demonstrated in [?] for and advection diffusion dynamic flow problem.

Estimating designs in this way requires that we fully accept the accuracy of the dynamics, and that the prior distribution covariance matrix is known for each experiment to be conducted. Choosing a linear regularization operator $\mathbf{L}$ such that $\mathbf{L}^\top\mathbf{L} = \boldsymbol{\Sigma}_k^{-1}$ is an option to deal with the priors.

3.2   Noise in the dynamics

Again we assume that $\mathbf{u}$ is known in equations (4a) so that $\mathbf{T} \leftarrow \mathbf{T}(\mathbf{u})$, and that no initial condition is needed since there is error in the estimates of $\mathbf{m}$. Denoting $\overline{\mathbf{m}} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_n]^\top$ and a set of data $\bar{\mathbf{d}} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_n]^\top$ and assuming the initial condition is given by the noise $\eta_0 = \mathbf{m}_0$ and $\mathbf{T}_0 = \mathbf{I}$ we can write

$$\underbrace{\begin{pmatrix} \mathbf{I} & 0 & & & \\ -\mathbf{T} & \mathbf{I} & 0 & & \\ & -\mathbf{T} & \mathbf{I} & \ddots & \\ & & \ddots & \ddots & 0 \\ & & & -\mathbf{T} & \mathbf{I} \end{pmatrix}}_{=:\mathbf{A}} \underbrace{\begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \mathbf{m}_3 \\ \vdots \\ \mathbf{m}_n \end{pmatrix}}_{=:\overline{\mathbf{m}}} = \mathbf{0}. \tag{15}$$

For a time invariant imaging technique, the Bayesian negative log likelihood is given by

$$\mathcal{J}(\overline{\mathbf{m}}) = \frac{1}{2}\left((\mathbf{I}\otimes\mathbf{F})\overline{\mathbf{m}} - \bar{\mathbf{d}}\right)^\top \overline{\mathbf{W}} \left((\mathbf{I}\otimes\mathbf{F})\overline{\mathbf{m}} - \bar{\mathbf{d}}\right) + \frac{1}{2}\left(\mathbf{A}\overline{\mathbf{m}}\right)^\top \boldsymbol{\Sigma}_{\overline{\mathbf{m}}}^{-1}\left(\mathbf{A}\overline{\mathbf{m}}\right) \tag{16}$$

where, $\overline{\mathbf{W}}$ is a block diagonal matrix with diagonal weight matrices $\mathbf{W}_0, \mathbf{W}_1, ..., \mathbf{W}_n$ on its main diagonal, and $\boldsymbol{\Sigma}_{\overline{\mathbf{m}}} = \text{blkdiag}(\boldsymbol{\Sigma}_{\eta_1}, .., \boldsymbol{\Sigma}_{\eta_n})$.

The posterior covariance matrix obtained after minimizing $\mathcal{J}(\overline{\mathbf{m}})$ for the Bayes estimate $\overline{\mathbf{m}}$ is given by

$$\left((\mathbf{I} \otimes \mathbf{F})^\top \overline{\mathbf{W}} \, (\mathbf{I} \otimes \mathbf{F}) + \mathbf{A}^\top \Sigma_{\overline{\mathbf{m}}}^{-1} \mathbf{A} \right)^{-1} \tag{17}$$

Thus, computing an optimal design entails minimizing the trace of the posterior covariance matrix. Again,

$$\phi_{A_B}(\overline{\mathbf{w}}) = \mathrm{tr}\left[\left((\mathbf{I} \otimes \mathbf{F})^\top \overline{\mathbf{W}} \, (\mathbf{I} \otimes \mathbf{F}) + \mathbf{A}^\top \Sigma_{\overline{\mathbf{m}}}^{-1} \mathbf{A} \right)^{-1} \right], \tag{18}$$

This formulation provides another method of designing all of the experiments given the covariance matrices for each time point. The main difference in this formulation than the previous section is that the error is now propagated in time. However, this method still does not incorporate historic designs or collected data, and it requires storing and manipulating large dense covariance matrices.

3.3   Adaptive optimal design

In this section we present two reformulations of the model estimation problems such that historic information is included. In this way the future design will depend on information from the previous experiment and the dynamics of the problem. We will consider the formulation with a linear and a non-linear regularization term.

3.3.1   Linear regularization

Referring again to equations (4a), setting $\eta_k = 0$ and assuming $\mathbf{u}$ is known, the model objective function is constructed such that the data misfit term for data collected at time $k-1$ and the data misfit term for time k are included,

$$\mathcal{G}(\mathbf{m}_{k-1}, \mathbf{w}_k) = \frac{1}{2}\|\mathbf{F}_{k-1}\mathbf{m}_{k-1} - \mathbf{d}_{k-1}^c\|_2^2 + \frac{1}{2}\|\mathbf{F}\mathbf{T}\mathbf{m}_{k-1} - \mathbf{d}_k\|_{\mathbf{W}_k}^2 + \frac{1}{2}\|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 \tag{19}$$

where $\mathbf{F}_{k-1} = \mathbf{W}_{k-1}\mathbf{F}$, and $\mathbf{m}_k = \mathbf{T}\mathbf{m}_{k-1}$. The operator $\mathbf{L}$ is a smoothing regularization operator. One could argue in this instance that the first term including the historic data can be interpreted as a prior such that the regularization term is not necessarily required. In this way, while it appears that this is a frequentist approach, it can still be interpreted as a Bayesian one if both the collected data misfit term and the regularization combine as a prior.

Minimizing equation (19) with respect to $\mathbf{m}_{k-1}$ yields the posterior covariance matrix

$$\Sigma_{k-1}(\mathbf{w}_k) = \left(\mathbf{F}_{k-1}^\top \mathbf{F}_{k-1} + \mathbf{T}^\top \mathbf{F}^\top \mathbf{W}_k \mathbf{F}\mathbf{T} + \mathbf{L}^\top \mathbf{L} \right)^{-1}. \tag{20}$$

By formulating the problem in this manner, one can adaptively update optimal designs after each experiment has been carried out.

The design functional to then be minimized is given by

$$\phi_k(\mathbf{w}_k) = \mathrm{tr}\left[\left(\mathbf{F}_{k-1}^\top \mathbf{F}_{k-1} + \mathbf{T}^\top \mathbf{F}^\top \mathbf{W}_k \mathbf{F}\mathbf{T} + \mathbf{L}^\top \mathbf{L} \right)^{-1} \right]. \tag{21}$$

A second benefit of building the design objective function to depend on an historic design and data set is that the design problem does not grow in size as time progresses. If we were to incorporate all historic data and designs, then we would end up with a rather large system. However, if all historic information is included, one could argue that this would provide a more accurate estimate and design.

The basic algorithm illustrating the adaptive design process for a linear observation method is outlined in Algorithm 1.

Algorithm 1 Iterative Optimal ϕ_{A_B} Design

1: Initialize: $\mathbf{F}_0$ ▷ Estimate a naive design
2: **for** $k = 1$ to n **do**   ▷ n time points
3:     $\widehat{\mathbf{w}}_k(\mathbf{F}_{k-1}) \rightarrow \mathrm{diag}(\widehat{\mathbf{w}}_k)\mathbf{F} = \mathbf{F}_k$ ▷ compute the optimal design by solving 3 with 21
4:     $\mathbf{d}_k = \mathbf{F}_k\mathbf{m}_k + \epsilon_k$ ▷ collect data
5:     $\mathbf{F}_{k-1} \leftarrow \mathbf{F}_k$ ▷ update F
6: **end for**

3.3.2 Non-linear regularization

At this point it is important to consider what information is available when estimating the optimal design. In equation (19), we added collected data misfit term, a linear regularization term promotes smoothness in the model $\mathbf{m}_k$, which of course appears in equation (20) as $\mathbf{L}^\top\mathbf{L}$. However, although the design $\mathbf{F}_k$ in the design objective function, the collected data does not. To include the collected data, we can instead impose a non-linear regularization function,

$$\mathcal{G}(\mathbf{m}, \mathbf{w}) = \sum_{k=1}^{n} \left(+ \frac{1}{2}\|\mathbf{FTm}_{k-1} - \mathbf{d}_k\|_{\mathbf{W}_k}^2 + R(\mathbf{m}_{k-1})\|_2^2 \right). \tag{22}$$

Since there is no closed form for the posterior covariance matrix for a non-linear inverse problem, we must apply an approximation as in [**?**]. Assigning the non-linear regularization functional results in an approximation to the inverse Hessian (or posterior covariance matrix) that depends on $\mathbf{m}$. Choosing a Gauss-Newton approximation of the Hessian, the optimal design optimization problem (3) becomes a bi-level optimization with an inequality constraint,

$$\min_{\mathbf{w}_k} \quad \mathrm{tr}\left[\left(\mathbf{T}^\top\mathbf{F}^\top\mathbf{W}_k\mathbf{FT} + \mathbf{S}^\top(\widehat{\mathbf{m}}_{k-1})\mathbf{S}(\widehat{\mathbf{m}}_{k-1}) \right)^{-1} \right] \tag{23a}$$

$$\text{s.t.} \quad \widehat{\mathbf{m}}_{k-1} = \mathrm{argmin}\frac{1}{2}\|\mathbf{F}_{k-1}\mathbf{m}_{k-1} - \mathbf{d}_{k-1}^c\|_2^2 + R(\mathbf{m}_{k-1}) \tag{23b}$$

$$0 \geq \omega_i, \tag{23c}$$

where $\mathbf{S}(\mathbf{m})$ is the gradient of the regularization functional $R(\mathbf{m})$. Typically solving the lower level optimization would require a stochastic approach of averaging over a set of possible models. However, in this case the historic data is available so that the best estimator $\widehat{\mathbf{m}}_{k-1}$ can be recovered so that it is not necessary to sample average over a set of possibilities.

4 Numerical optimization

In this section we present the computation of the gradients of the linear and non-linear design functionals.

4.1 Linear adaptive design

Finding an optimal design for the time step k involves solving the optimization problem 3 by minimizing the objective functional (21),

$$\widehat{\mathbf{w}_k} = \operatorname{argmin} \mathcal{J}_k := \left\{ \operatorname{tr}\left[\left(\mathbf{T}^\top\mathbf{F}^\top\mathbf{W}_k\mathbf{F}\mathbf{T} + \mathbf{F}_{k-1}^\top\mathbf{F}_{k-1} + \mathbf{L}^\top\mathbf{L}\right)^{-1}\right] + \beta\mathbf{e}^\top\mathbf{w}_k \right\}$$

$$\text{s.t.} \quad 0 \geq \omega_i.$$

where $\mathbf{e}$ is a vector of ones (the constraint allows us to write the 1-norm as a sum).

This is a difficult task for large problems, particularly due to the trace of a large dense matrix and the computation of the derivative of the trace. To this end, several actions can be taken to reduce the computational costs.

First, a stochastic Hutchinson trace estimator is used to approximate the trace [?] such that Equation (??) can be written as

$$\mathcal{J}_k(\mathbf{w}_k) = \mathbf{v}^\top(\mathbf{T}^\top\mathbf{F}^\top\mathbf{W}_k\mathbf{F}\mathbf{T} + \mathbf{G})^{-1}\mathbf{v} + \beta\mathbf{e}^\top\mathbf{w}_k, \tag{24}$$

where $\mathbf{G} = \mathbf{F}_{k-1}^\top\mathbf{F}_{k-1} + \mathbf{L}^\top\mathbf{L}$ and $\mathbf{v}$ is a random vector with values 1 and -1 with equal distribution. This approximation makes the computation of the derivative of $\mathcal{J}$ tractable for large problems.

A second issue arises when computing the derivative of $\mathcal{J}$, as this involves computing the derivative of an inverse matrix. From now on we drop the subscript k for clarity,

$$\nabla_\mathbf{w}\mathcal{J}(\mathbf{w}) = \mathbf{v}^\top\nabla_\mathbf{w}\left(((\mathbf{F}\mathbf{T})^\top\mathbf{W}\mathbf{F}\mathbf{T} + \mathbf{G})^{-1}\mathbf{v}\right) + \beta\mathbf{e}.$$

To compute the derivative of $\mathcal{J}$ we define

$$\mathbf{z} = ((\mathbf{F}\mathbf{T})^\top\mathbf{W}\mathbf{F}\mathbf{T} + \mathbf{G})^{-1}\mathbf{v} \quad \Leftrightarrow \quad ((\mathbf{F}\mathbf{T})^\top\mathbf{W}\mathbf{F}\mathbf{T} + \mathbf{G})\mathbf{z} = \mathbf{v}, \tag{25}$$

and differentiate implicitly to obtain

$$(\mathbf{F}\mathbf{T})^\top\operatorname{diag}(\mathbf{F}\mathbf{T}\mathbf{z}) + ((\mathbf{F}\mathbf{T})^\top\mathbf{W}\mathbf{F}\mathbf{T} + \mathbf{G})\nabla_\mathbf{w}\mathbf{z} = 0.$$

The gradient of $\mathbf{z}$ is then given by

$$\nabla_\mathbf{w}\mathbf{z} = -((\mathbf{F}\mathbf{T})^\top\mathbf{W}\mathbf{F}\mathbf{T} + \mathbf{G})^{-1}(\mathbf{F}\mathbf{T})^\top\operatorname{diag}(\mathbf{F}\mathbf{T}\mathbf{z}).$$

Thus, the first term in $\nabla_{\mathbf{w}} \mathcal{J}(\mathbf{w})$ is

$$-\mathbf{v}^{\top}((\mathbf{FT})^{\top}\mathbf{WFT} + \mathbf{G})^{-1}(\mathbf{FT})^{\top}\text{diag}(\mathbf{FTz}).$$

Taking the transpose

$$-\text{diag}(\mathbf{FTz})\mathbf{FT}((\mathbf{FT})^{\top}\mathbf{WFT} + \mathbf{G})^{-1}\mathbf{v} \quad \Leftrightarrow \quad -\text{diag}(\mathbf{FTz})(\mathbf{FTz}) = -\mathbf{FTz} \odot \mathbf{FTz},$$

so that the derivative of the objective function is then

$$\nabla_{\mathbf{w}} \mathcal{J}_k(\mathbf{w}) = -\mathbf{FTz} \odot \mathbf{FTz} + \beta \mathbf{e}. \tag{26}$$

Since the design problem (3) is nonlinear with respect to $\mathbf{w}$, a Newton type iterative method is used with a backtracking line search. This solution method requires both the evaluation of $\mathcal{J}$ and its gradient, which are conveniently obtained by solving for $\mathbf{z}$ only once. However, for large scale problems, the evaluation of $\mathbf{z}$ is not cheap.

To reduce computational costsit is possible to first preform a Cholesky factorization on $\mathbf{G} = \mathbf{R}^{\top}\mathbf{R}$, reduce equation (25) to standard form, and then compute the Lanczos bidiagonalization of $\mathbf{FTR}^{-1}$. This process requires calculating the Cholesky factorization and calculating $\mathbf{R}^{-1}$. However, this is evaluated only once outside of the design optimization, so depending on the problem size, it is computationally cheap.

Alternately, instead of the Lanczos approach Krylov subspace methods can be applied, such as Conjugate Gradients, allowing us to take advantage of matrix vector products and avoid ever forming $\mathbf{G}$.

4.2 Non-Linear adaptive design

5 Numerical example: Adaptive design

To demonstrate the adaptive optimal experimental design, a seismic borehole tomography survey is chosen to image tracer advection in the subsurface. In this synthetic example a spatial distribution of tracer is advected with the flow of water injected from one well and pumped out at another, and seismic tomography data are measured at each time point 90 days apart. In this section we present the experimental setup and a brief description of the discretization of governing equations, and the results of the synthetic example.

5.1 Experimental setup

The computational domain $\Omega = [0, 100] \times [0, 200]$ meters, is divided into $m = [50, 100]$ cells of width $h = [2, 2]$ meters and consists of a layered model of five rock types.
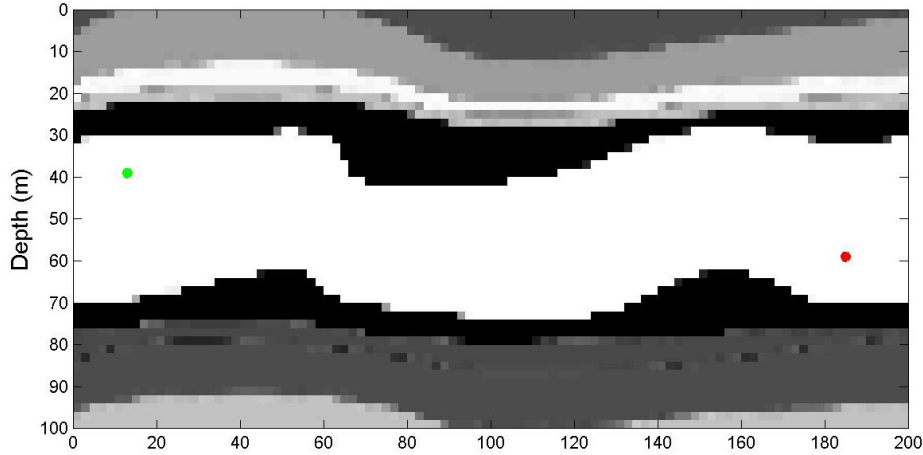
Figure 1: Initial layered Earth model.

5.1.1 Seismic borehole tomography

A simple borehole tomography experiment was chosen to demonstrate the adaptive design method. The experiment measures the travel times between sources n_s and receivers n_r. Travel times are given by integrating the slowness $m(t, \vec{\mathbf{x}})$, or inverse acoustic velocity of the media, along a ray path Γ,

$$\mathbf{d}_{j,k}(t) = \int_{\Gamma_{j,k}} m(t, \vec{\mathbf{x}}) \, d\ell.$$

Assuming the data are noisy and measured at times $(t_0, ...t_n)$ the tomography problem is given by

$$\mathcal{F}m(t_k, \vec{\mathbf{x}}) + \epsilon_k = \mathbf{d}(t_k).$$

11

Equation (5.1.1) is discretized on a standard staggered 2D grid using a finite volume scheme placing s and $\mathbf{d}$ at the cell centers, see figure 4 below.

The discrete tomography experiment in matrix form is then,

$$\mathbf{Fm}_k = \mathbf{d}_k, \text{ for } k = 1, \ldots, n. \tag{27}$$

The initial survey setup is pictured in figure 2, and includes 4 boreholes with 12 sources in the east borehole, 10 receivers in each center borehole, 20 receivers in the west borehole, and 20 surface receivers. The initial design was chose to overly cover all the area in the flow domain, with the goal of significantly reducing the number of data required.
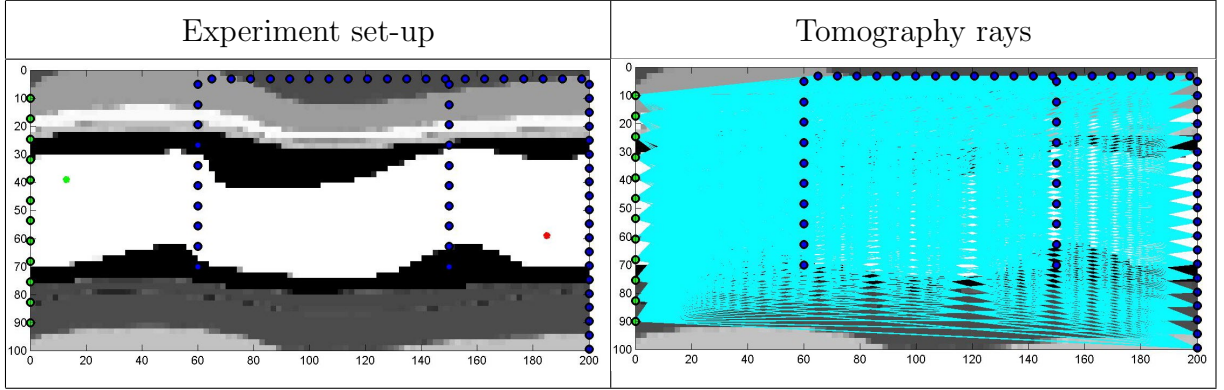
Figure 2: Initial experiment setup. Sources are pictured in green on the left and receivers in blue on the surface, two 75m boreholes, and the right 100m borehole. The initial setup covers the entire flow domain with 12 sources and 60 receivers.

5.1.2 Tracer dynamics

The subsurface dynamics are described by the tracer advection equations as presented in [?], which characterize the transport of a solute in a fully saturated porous media,

$$\frac{\partial(c\rho)}{\partial t} + \nabla \cdot c\rho\vec{u} = 0, \tag{28a}$$
$$\text{subject to } c\rho(0, \vec{x}) = c_0\rho(\vec{x}),$$

with hydraulic conductivity $\mathbf{K}$, and tracer concentration c in a fluid of density ρ. The fluid velocity field $\vec{u}$ satisfies a simple source-sink (q) flow model

$$\nabla \cdot \vec{u} = q \tag{28b}$$
$$\vec{u} = \mathbf{K}(\vec{x})\nabla p \tag{28c}$$
$$p(0, \vec{x}) = p_0(\vec{x}), \tag{28d}$$

with either Dirichlet or Neumann boundary conditions on the pressure p.

12

With the goal being to image the motion of the tracer, $c(t, \vec{\mathbf{x}})$, we assume that changes in c amount to changes in geophysical properties. This implies that $m(c)$ exists, and that the motion of m is governed by Equation (28a). From now on we will refer only to $\mathbf{m}$.
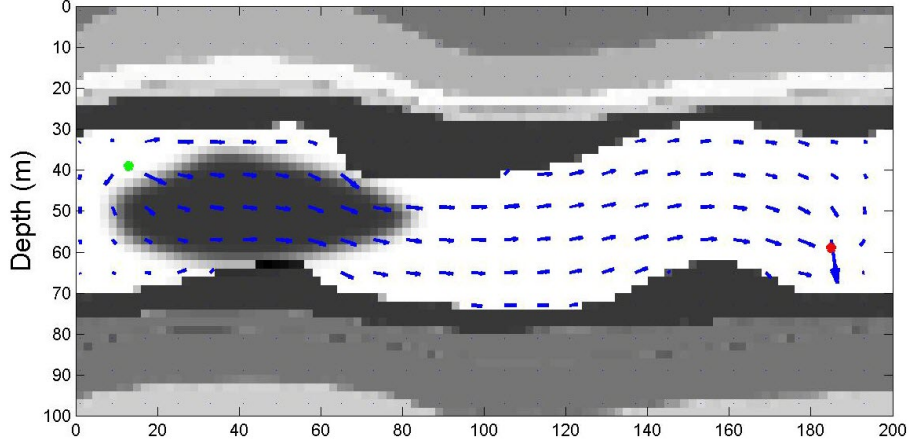

Figure 3: Velocity field(arrows) and the initial tracer concentration used to generate data

To discetize the advection of the tracer, a particle in cell discretization was chosen such that equation (28a) is then given by,

$$\mathbf{m}_k = \mathbf{T}(\mathbf{u})\mathbf{m}_{k-1}.$$

The remainder of the flow equations were disretized on a staggered grid using the finite volume method, with the components of the velocity field located on the cell edges and the tracer concentration at the cell centers.

The fluid velocity field $\mathbf{u}$ was computed by solving equations (**??**) with hydraulic conductivities for $\mathbf{K}$ ranging from $3e-1$ to $6e-06 kg/s^2 m^2$, and was time invariant.

5.1.3 Regularization

Next, the regularization functionals for the tracer estimation are discretized. For the linear regularization term $\frac{1}{2}\|\mathbf{Lm}\|_2^2$, the operator $\mathbf{L} = \mathsf{GRAD}$ is a standard 2-point discretization of the gradient of a cell-centered variable, and has the effect of smoothing the solution.

A nice choice for a non-linear regularization operator for this problem is smoothed Total Variation [**?**], since the tracer experiencing advective flow with no diffusion can have discontinuities.

$$\mathcal{R}^{\mathrm{s}}(s_0) = \beta \int \phi(|\nabla s_0|) \, d\vec{\mathbf{x}}. \tag{29}$$

Again, noting that $\mathbf{m}$ is discretized on a cell-centered grid, a standard discrete approximation for the smoothed total variation regularization is applied

$$R^m(\mathbf{m}) = h^2 \mathbf{e}^\top \sqrt{\mathbf{A}_f^c \left((\mathsf{GRAD} \ \mathbf{m}) \odot (\mathsf{GRAD} \ \mathbf{m}) \right) + \epsilon},$$

13

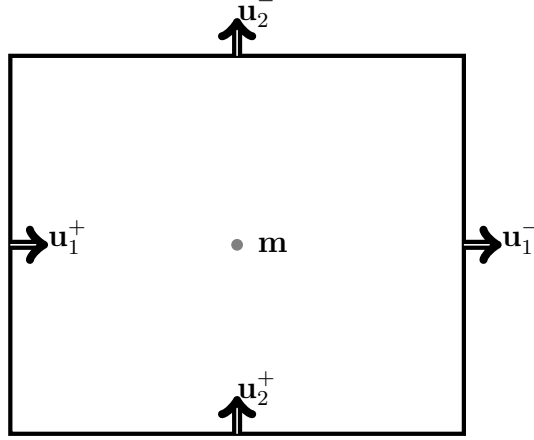Figure 4: An example of a stagered grid cell with model parameters at the cell center and fluid velocity field on the edges.

where GRAD is a standard 2-point discretization of the gradient of a cell-centered variable, which maps from cell-centers to faces, as described in [?, ?]. $\mathbf{A}_f^c$ is an averaging matrix from cell-faces to cell-centers. Note that we somewhat abuse the notation, and that the square root of a vector is the point-wise square root.

5.2   Results

To generate data, an initial tracer concentration was marched along in time, for time steps of 90 days, with Gaussian noise added. In total 6 experiments were conducted every 90 days.

The initial design for time point 1 did not include the dynamics. This is basically the static case, or when there has been no data collected. In algorithm 1 this amounts to setting $\mathbf{F}_k = 0$ and the transport matrix $\mathbf{T} = \mathbf{I}$. Figure ?? shows the variance, equation (21), as a function of the number of non zero weights (nnz). We chose $\beta = 1$ which corresponds with about 399 of the 720 data being used. The new design and the recovered model of the tracer are presented below.

6   Numerical example 2: Helmhotz maybe

This depends on how far ahead in time we design...The all at once approach assumes the velocity field is accurate(ish)...could still solve the all at once problem for updated velocity field too.
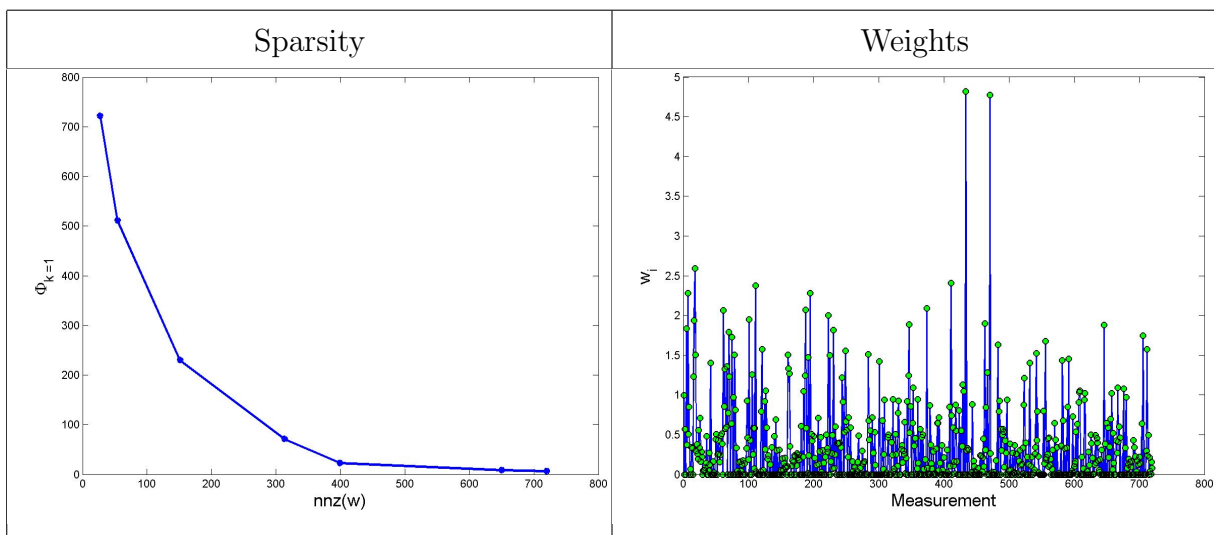
7   Concluding Remarks

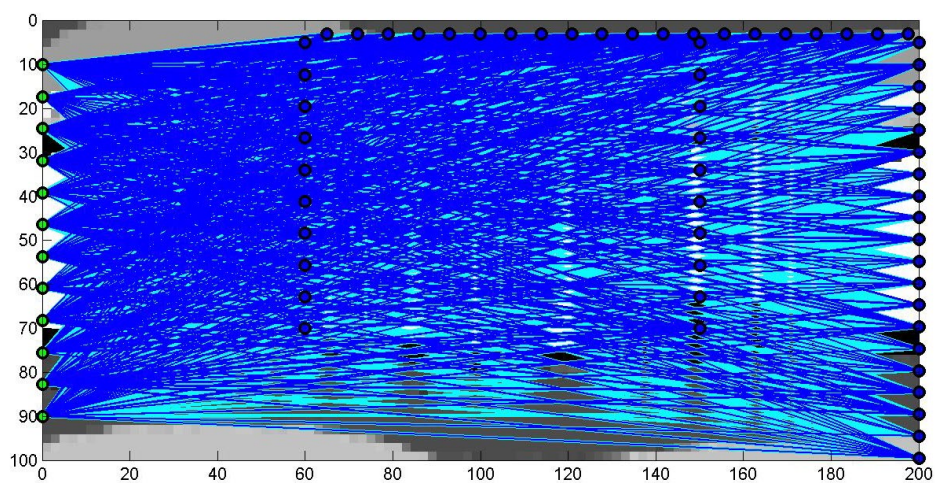Figure 5: Plot of variance $\phi_1(\mathbf{w})$ versus $\mathbf{w}$, and the weights for $\beta = 1$

Figure 6: An image of the optimal design data, or rays plotted over the full set of rays The number of data required was decreased from 720 to 399
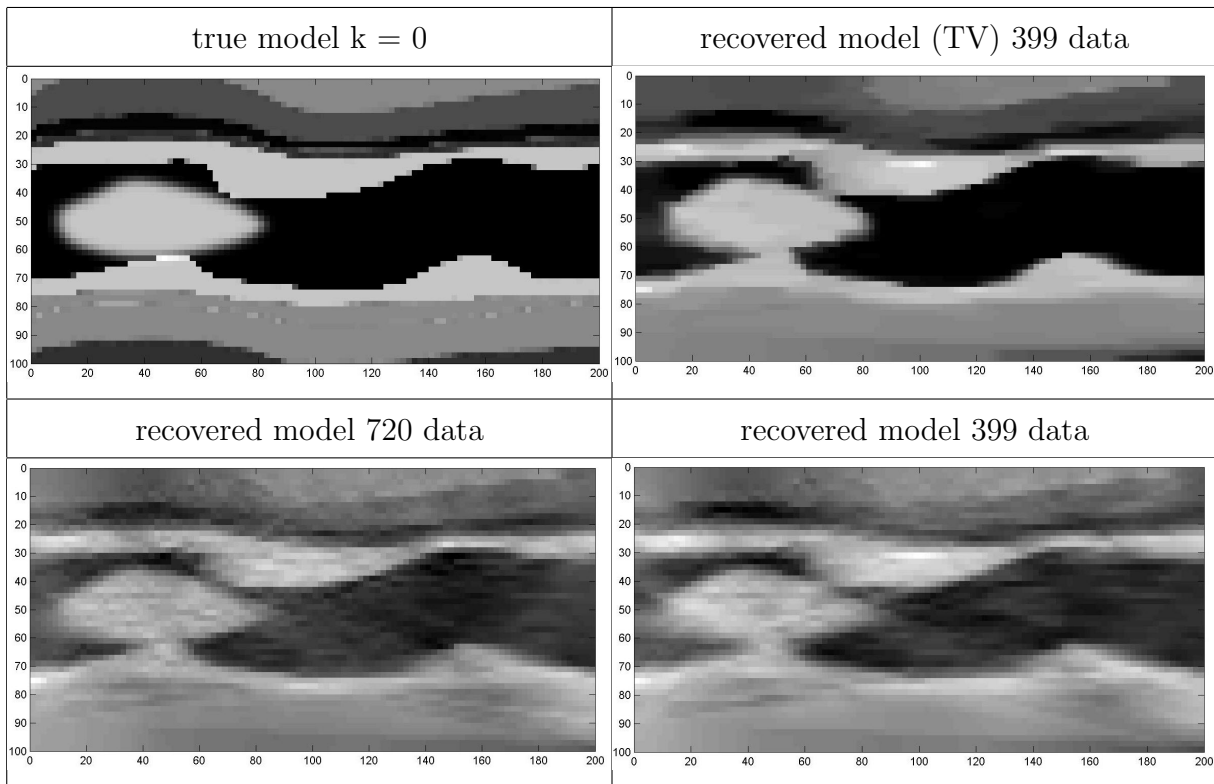
Figure 7: Static optimal design, $k = 0$. The 4 panels show the true tracer and subsurface, the image recovered using 399 data points and total variation for the regularization, and recoveries using all data and the optimal data for the linear gradient regularization
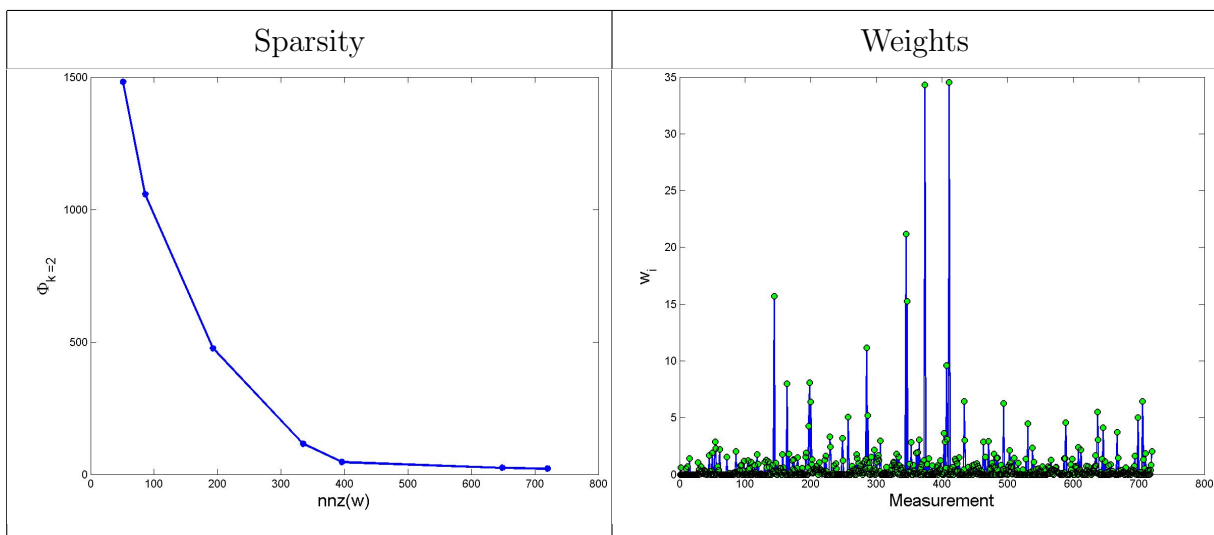

Figure 8: Experiment k = 1. Plot of variance $\phi_1(\mathbf{w})$ versus $\mathbf{w}$, and the weights for $\beta = 1$
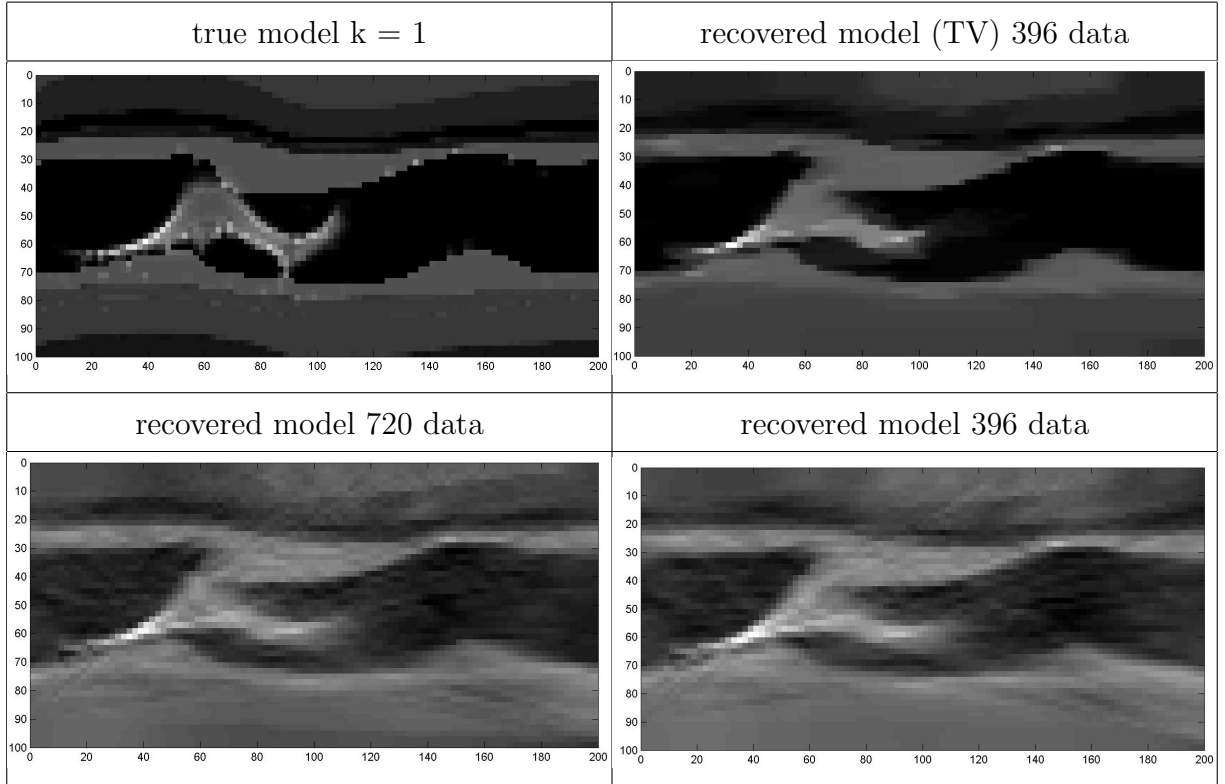
16

Figure 9: Adaptive optimal design, $k = 1$. The 4 panels show the true tracer in the subsurface after 90 days. The images were recovered using 396 data points for both the total variation and gradient regularization