A penalty method for PDE-constrained optimization

Tristan van Leeuwen and Felix J. Herrmann

Dept. of Earth, Ocean and Atmospheric Sciences.
2207 Main Mall, Vancouver, BC Canada V6T 1Z4.

E-mail: tleeuwen@eos.ubc.ca

Abstract. We present a method for solving PDE constrained optimization problems based on a penalty formulation. This method aims to combine advantages of both full-space and reduced methods by exploiting a large search-space (consisting of both control and state variables) while allowing for an efficient implementation that avoids storing and updating the state-variables. This leads to a method that has roughly the same per-iteration complexity as conventional reduced approaches while defining an objective that is less non-linear in the control variable by implicitly relaxing the constraint. We apply the method to a seismic inverse problem where it leads to a particularly efficient implementation when compared to a conventional reduced approach as it avoids the use of adjoint state-variables. Numerical examples illustrate the approach and suggest that the proposed formulation can indeed mitigate some of the well-known problems with local minima in the seismic inverse problem.

1. Introduction

We consider PDE-constrained optimization problems of the form

$$\min_{m,u} \tfrac{1}{2}||r(u,d)||_2^2 \quad \text{s.t.} \quad c(m,u) = 0, \tag{1}$$

where m is the control variable, u is the state variable and d are the input data. The state-space constraint $c(m,u) = 0$ encodes the PDE while r measures the misfit between the input data and the modeled quantity u. These problems arise in many applications such as optimal control and design [1, 2], inverse problems in geophysics [3, 4], medical imaging [5] and non-destructive testing. Specifically for inverse problems, m is a (gridded) physical quantity of interest (e.g., soundspeed, density, conductivity) and u is a physical field (e.g, waves, electromagnetic) and $c(m,u)$ encodes the physics. The observed data is denoted by d and $r(u,d)$ measures the residual between the predicted field and the data. Oftentimes, measurements are made from multiple independent experiments, in which case u is a block vector containing the fields for different experiments. As a result, the size of u is typically much bigger than that of m.

A popular approach to solving these constrained problems is based on the corresponding Lagrangian formulation:

$$\min_{m,u,v} \mathcal{L}(m,u,v) = \tfrac{1}{2}||r(u,d)||_2^2 + \langle v, c(m,u)\rangle. \tag{2}$$

Solving this problem with a Newton-like method involves solving the KKT system [6]:

$$\begin{pmatrix} \star & \star & \nabla_m c \\ \star & \nabla_u r^T \nabla_u r + \star & \nabla_u c \\ \nabla_m c & \nabla_u c & 0 \end{pmatrix} \begin{pmatrix} \delta m \\ \delta u \\ \delta v \end{pmatrix} = - \begin{pmatrix} \nabla_m \mathcal{L} \\ \nabla_u \mathcal{L} \\ \nabla_v \mathcal{L} \end{pmatrix}, \tag{3}$$

where * denotes the complex-conjugate transpose and $\star$ denotes second-order derivatives, which are usually ignored, and

$$\nabla_m \mathcal{L} = \nabla_m c^* v, \tag{4}$$

$$\nabla_u \mathcal{L} = \nabla_u r^* r(u,d) + \nabla_u c^* v, \tag{5}$$

$$\nabla_v \mathcal{L} = c(m,u). \tag{6}$$

Advantages of this approach are that it eliminates the need to solve the PDEs explicitly. However, this approach is often unfeasible for large-scale applications because it involves simultaneously updating (and hence storing) all the variables.

Instead, one usually considers a *reduced* problem

$$\min_m \phi_{\text{red}}(m) = \tfrac{1}{2}||r(\overline{u}(m),d)||_2^2, \tag{7}$$

where $\overline{u}(m)$ is defined through $c(m,\overline{u}(m)) = 0$. The gradient of this objective is given by

$$\nabla_m \phi_{\text{red}}(m) = \nabla_m c(m,\overline{u})^* \overline{v}, \tag{8}$$

where $\overline{v}$ is solved from $\nabla_u \mathcal{L}(m,\overline{u},v) = 0$. The disadvantage of this approach is that it involves explicit elimination of the state-space constraints (which involve PDE solves) at each update. It also strictly enforces the constraint $c(m,u) = 0$ at each iteration, which might lead to a very nonlinear problem in m. Moreover, the corresponding reduced Hessian is typically a dense matrix that cannot be stored and computing its action involves additional PDE solves. Practical approaches are usually based on Quasi-Newton approximations of the reduced Hessian.

1.1. Contributions

In this paper we present an alternative to the reduced approach that has a roughly equivalent per-iteration complexity in terms of PDE solves and storage but retains some of the characteristics of the all-at-once approach in the sense that it exploits a larger search space by not enforcing the constraints $c(m,u)=0$ at each iteration. The approach is based on the following *penalty* formulation of the constrained problem:

$$\min_{m,u} \phi_\lambda(m,u) = \tfrac{1}{2}||r(u,d)||_2^2 + \tfrac{\lambda^2}{2}||c(m,u)||_2^2, \tag{9}$$

the solution of which coincides with that of the constrained problem (1) when $\lambda \uparrow \infty$. For a fixed λ we can solve this problem in a number of ways. As shown in section 2, we can use a variational projection approach described by [7] to define a reduced problem:

$$\min_m \overline{\phi}_\lambda(m) = \phi_\lambda(m, \overline{u}(m)), \tag{10}$$

where $\overline{u}(m)$ satisfies $\nabla_u \phi_\lambda(m,\overline{u}) = 0$. If we can solve this problem as efficiently as the original PDE $c(m,u)=0$, we arrive at a simple approach that does not require the calculation of sensitivities of $\overline{u}$ w.r.t. m. Moreover, it turns out that the corresponding Hessian can be well-approximated by a sparse matrix, making it feasible to a employ a Newton method for the optimization.

The merit of this approach is illustrated for a simple toy problem defined by $r(u,d) = u - d$, $c(m,u) = (\mathrm{diag}(m) + B)u - q$ with

$$B = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 1 \end{pmatrix}, \quad q = \begin{pmatrix} 1\frac{3}{4} \\ 2\frac{1}{4} \end{pmatrix}.$$

The solution in this case is $\overline{m} = (1,1)$ and $\overline{u} = (1,1)$. We use a Gauss-Newton method to solve both (7) and (10), starting from $m = (2,2)$. We use $\lambda = 0.1$ for the penalty approach. Figure 1 (a) shows the solution paths and (c) shows the convergence in terms of $||r||_2$ and $||c||_2$ for both the reduced (black) and penalty approaches (red). The penalty approach gets very close to the optimal solution in one iteration while the reduced approach takes a detour because it is forced to satisfy the constraint $c(m,u)=0$ at each iteration. Another perspective is offered by plotting the objective functions corresponding to the reduced and penalty approaches as a function of m. Figure 1 (c) shows the reduced objective (7) and (d) shows the penalty objective (10). The reduced objective in this case is quite non-linear while the penalty objective is quadratic. These plots illustrate that the penalty formulation can indeed lead to an objective function that is much better behaved.

Application of the proposed approach to seismic inversion, where the PDE is a Helmholtz equation, is detailed in section 3. Here, we also compare the penalty approach to both the all-at-once and the reduced approaches in terms of algorithmic complexity. Some numerical examples on seismic inversion using both the penalty and reduced formulations are given in section 4.

Possible extensions and open problems are discussed in section 5 and section 6 gives the conclusions.

1.2. Related work

The proposed method is related to the *equation-error* approach, which is typically used to identify the control variable in diffusion problems given a *complete* measurement of the state: $d = u$ by solving $c(m,d)=0$ for m [8]. Given *partial* measurements of the

state $d = Pu$, the proposed method can be seen as a way of bootstrapping this by first attempting to reconstruct the complete state from the partial measurements. Popular methods for solving constrained problems of the form (1) include the Alternating Direction Method of Multipliers (ADMM) [9], which rely on a an augmented Lagrangian formulation and exact penalty methods [10], which use alternative, non-quadratic, penalty functions. The method discussed here does not fall in either of these classes; it does not introduce a Lagrangian but uses a form of alternating optimization to efficiently solve the penalty formulation.

2. Variational projection

The solution of the penalty formulation (9) coincides with the solution of the original problem (1) when $\lambda \uparrow \infty$. For a fixed λ, we can solve (9) with a full Newton method based on solving

$$\begin{pmatrix} \nabla_u^2 \phi_\lambda & \nabla_{u,m}^2 \phi_\lambda \\ \nabla_{u,m}^2 \phi_\lambda & \nabla_m^2 \phi_\lambda \end{pmatrix} \begin{pmatrix} \delta u \\ \delta m \end{pmatrix} = - \begin{pmatrix} \nabla_u \phi_\lambda \\ \nabla_m \phi_\lambda \end{pmatrix}, \tag{11}$$

and updating $u := u + \delta u$ and $m := m + \delta m$. A disadvantage of this approach is that we would still need to store the state variable in order to update it.

The variational projection approach, as described by [7], can be used to eliminate the state variable and define a reduced problem:

$$\min_m \overline{\phi}_\lambda(m) = \tfrac{1}{2}||r(\overline{u}, d)||_2^2 + \tfrac{\lambda^2}{2}||c(m, \overline{u}(m))||_2^2, \tag{12}$$

where $\overline{u}(m) = \operatorname{argmin}_u \tfrac{1}{2}||r(u,d)||_2^2 + \tfrac{\lambda^2}{2}||c(m,u)||_2^2$ (i.e., $\nabla_u \phi_\lambda(m, \overline{u}) = 0$). It is readily verified that the gradient and Hessian of the reduced objective are given by

$$\nabla \overline{\phi}_\lambda(m) = \nabla_m \phi_\lambda(m, \overline{u}), \tag{13}$$

$$\nabla^2 \overline{\phi}_\lambda(m) = \nabla_m^2 \phi_\lambda(m, \overline{u}) - \nabla_{m,u}^2 \phi_\lambda(m, \overline{u}) \left(\nabla_u^2 \phi_\lambda(m, \overline{u})\right)^{-1} \nabla_{u,m}^2 \phi_\lambda(m, \overline{u}). \tag{14}$$

It immediately follows that a stationary point $\overline{m}$ of the reduced objective (i.e., $\nabla \overline{\phi}_\lambda(\overline{m}) = 0$) together with the corresponding $\overline{u}$ satisfy the first order optimality conditions of the full objective and vice versa. Moreover, a point $\overline{m}$ that satisfies the second order optimality condition of the reduced objective $\nabla^2 \overline{\phi}_\lambda(\overline{m}) \succ 0$ together with a corresponding $\overline{u}$ that satisfies $\nabla_u^2 \phi_\lambda(\overline{m}) \succ 0$ (in case both $r(m,u)$ and $c(m,u)$ are linear in u this is satisfied automatically), satisfy the second order optimality conditions of the full objective. This can be verified by observing that the Hessian of the reduced objective is the Schur complement of the Hessian of the full objective and using the properties of the Schur complement of a symmetric positive definite matrix (cf. [11, prop. 14.1]). We can use these expressions for the gradient and Hessian to design a Newton-like method to solve (12).

3. Application to seismic inversion

In seismic waveform inversion, the goal is to obtain detailed estimates of subsurface medium parameters from seismic data by solving a PDE-constrained optimization problem [4, 12]. Such data are typically collected for a large number of sources, which we indicate by $l = [1, 2, \ldots, N_s]$. The response is recorded as a time-series from which

we extract a few relevant frequencies, indexed by $k = [1, 2, \ldots, N_f]$. The data for each source and frequency is organized in a vector $\mathbf{d}_{kl}$. The governing PDE in this case is taken to be the scalar Helmholtz equation:

$$A_k(\mathbf{m})\mathbf{u}_{kl} = \mathbf{q}_{kl}, \tag{15}$$

where $A_k(\mathbf{m})$ is a discretization of the Helmholtz operator $\left(\omega_k^2 m + \nabla^2\right)$ for angular frequency ω_k, $\mathbf{m}$ are the gridded medium parameters, $\mathbf{u}_{kl}$ denotes the wavefield, and $\mathbf{q}_{kl}$ is the source function. The resulting measurements are denoted by $P\mathbf{u}_{kl}$.

3.1. Penalty method

The objective (12) is now given by

$$\phi_\lambda(\mathbf{m}, \mathbf{u}) = \tfrac{1}{2} \sum_{k,l} ||P\mathbf{u}_{kl} - \mathbf{d}_{kl}||_2^2 + \lambda^2 ||A_k(\mathbf{m})\mathbf{u}_{kl} - \mathbf{q}_{kl}||_2^2, \tag{16}$$

where $\mathbf{u}$ is a vector containing all wavefields $\mathbf{u}_{kl}$. The wavefields satisfying $\nabla_{\mathbf{u}}\phi_\lambda(\mathbf{m}, \overline{\mathbf{u}}) = 0$ can be solved from

$$\begin{pmatrix} \lambda A_k(\mathbf{m}) \\ P \end{pmatrix} \mathbf{u}_{kl} = \begin{pmatrix} \lambda \mathbf{q}_{kl} \\ \mathbf{d}_{kl} \end{pmatrix}. \tag{17}$$

Given the wavefields $\overline{\mathbf{u}}_{kl}$ that solve (17), the gradient and Hessian of the reduced objective are now given by

$$\nabla\overline{\phi}_\lambda = \sum_{k,l} \lambda^2 G_{kl}^* \left(A_k \overline{\mathbf{u}}_{kl} - \mathbf{q}_{kl}\right), \tag{18}$$

$$\begin{aligned} \nabla^2\overline{\phi}_\lambda &= \sum_{k,l} \lambda^2 G_{kl}^* G_{kl} - \lambda^4 G_{kl}^* A_k \left(P^*P + \lambda^2 A_k^* A_k\right)^{-1} A_k^* G_{kl} \\ &= \lambda^2 \sum_{k,l} G_{kl}^* G_{kl} - G_{kl}^* \left(I + \lambda^{-2} A_k^{-*} P^* P A_k^{-1}\right)^{-1} G_{kl}, \end{aligned} \tag{19}$$

where $G_{kl} = \frac{\partial A_k(\mathbf{m})\overline{\mathbf{u}}_{kl}}{\partial \mathbf{m}}$, which is typically a sparse diagonally dominant matrix. For large λ we can expand the inverse as $(I + B)^{-1} = I - B + B^2 + \ldots$, and approximate the Hessian as

$$\nabla^2\overline{\phi}_\lambda \approx \sum_{k,l} G_{kl}^* A_k^{-*} P^* P A_k^{-1} G_{kl}, \tag{20}$$

which is the Gauss-Newton Hessian of the reduced approach. For small λ, we approximate the Hessian as

$$\nabla^2\overline{\phi}_\lambda \approx \lambda^2 \sum_{k,l} G_{kl}^* G_{kl}, \tag{21}$$

thus effectively ignoring the second term in the Hessian. Algorithm 1 gives a Gauss-Newton algorithm based on this approximation.

From the expressions for the gradient and Hessian we see that, for small λ, the leading order dependency of $\overline{\phi}_\lambda$ on $\mathbf{m}$ is through $A_k(\mathbf{m})$. The sensitivity of $\overline{\mathbf{u}}_{kl}$ does not appear in the gradient and appears in the Hessian only as higher order term. This confirms that this approach leads to a less non-linear problem than the traditional reduced approach, where the main dependency is through $A_k(\mathbf{m})^{-1}$, as we will see below.

3.2. Reduced Lagrangian approach

The objective for the reduced approach is given by

$$\phi_{\mathbf{red}}(\mathbf{m}) = \tfrac{1}{2}\sum_{k,l} ||PA_k(\mathbf{m})^{-1}\mathbf{q}_{kl} - \mathbf{d}_{kl}||_2^2. \tag{22}$$

The gradient and Hessian of the reduced objective are given by

$$\nabla\phi_{\mathbf{red}} \;= \sum_{kl} G_{kl}^{*}\overline{\mathbf{v}}_{kl} \tag{23}$$

$$\nabla^2\phi_{\mathbf{red}} = \sum_{kl} G_{kl}^{*}A_k^{-*}P^{*}PA_k^{-1}G_{kl}, \tag{24}$$

where

$$A_k(\mathbf{m})\overline{\mathbf{u}}_{kl} \;= \mathbf{q}_{kl}, \tag{25}$$

$$A_k(\mathbf{m})^{*}\overline{\mathbf{v}}_{kl} = \;-P^{*}(P\overline{\mathbf{u}}_{kl} - \mathbf{d}_{kl}). \tag{26}$$

The Hessian in this case cannot be easily approximated by a sparse matrix and its application would require additional PDE solves [13]. In contrast to the penalty objective, the dependency on $\mathbf{m}$ of the reduced objective is solely through $A_k(\mathbf{m})^{-1}$.

Algorithm 2 gives a Quasi-Newton algorithm for the solution of the unconstrained optimization problem.

3.3. Complexity estimates

Assuming we can solve equations (17) and (25) equally efficient, the penalty-based method requires a factor of 2 less computation and storage. Note, however, that for the penalty-based formulation we get an (approximate) Newton method while the reduced method only uses a Quasi-Newton approach. A Gauss-Newton method for the reduced approach would require far more PDE-solves and is not considered here. A summary of the leading order computational costs of the penalty, reduced and all-at-once approaches is given in table 1.

4. Examples

For the examples, we discretize the 2D Helmholtz operator using a 5-point finite-difference stencil with absorbing boundary conditions. The Quasi-Netwon algorithm 2 is augmented with a simple back-tracking linesearch to ensure descent.

4.1. Camembert

We present an example on a simple toy model, depicted in figure 2 (a). 20 Sources and 94 receivers are located in vertical wells on either side of the model. T The convergence histories in terms of the data misfit $||r(u,d)||_2$ and constraint $||c(m,u)||_2$ are shown in figures 2 (a). This illustrates the difference in solution paths between the reduced and penalty methods. he reconstructions after inverting the data at 10 Hz with 10 iterations are shown in figures 2 (c-d). The results are very similar, but inversion using the penalty approach was roughly twice as fast because there was no need to solve an adjoint PDE at each iteration.

4.2. Marmousi

A slightly more realistic model is depicted in figure 3 (a). Here, 51 sources and 101 receivers are located at the top of the model. The data are generated on a finer grid (20 m spacing) than used for the inversion (50 m spacing). We apply the outlined procedure (5 iterations) to data for frequencies 1,2,3,4,5 Hz consecutively each time using the end result of one frequency as initial model for the next. The initial model for the first frequency is depicted in 3 (b). This continuation from low to high frequencies is common practice in seismic inversion as it helps to avoid some of the problems with local minima [14]. The results on data without noise are depicted in 4. The penalty method converges much faster in terms of the relative model error and gives a better final result. Moreover, the penalty method was roughly twice as fast because there was no need to solve an adjoint PDE at each iteration.

The results for data with 10 % additive Gaussian noise are depicted in figure 5. These experiments indicate that the penalty method is robust against a moderate amount of noise.

Finally, results for inversion of frequencies 2,3,4,5 Hz are shown in 6. This experiment highlights a key issue in seismic inversion; the lack of low frequencies in the data. This causes local minima in the reduced objective which may lead gradient-based methods to converge to a wrong solution. While not producing as good a result as when starting from 1 Hz, the penalty approach does a lot better than the conventional approach. This confirms that the penalty formulation does indeed mitigate some of the problems with local minima.

5. Discussion

This paper lays out the basics of an efficient implementation of the penalty method for PDE-constrained optimization. A few remaining issues and possible extensions are listed below.

Solving large, sparse, mildly overdetermined systems A key step in the penalty method is the solution of the augmented PDE (17). While we can make use of factorization techniques for small-scale applications, industry-scale applications will typically require (preconditioned) iterative solution of such systems. A promising candidate is a generic accelerated row-projected method described by [15, 16] which proved successful in seismic applications and can be easily extended to deal with overdetermined systems [17].

Time-domain formulation We have described application of the penalty method to a formulation in the frequency domain, in which case we can hope to store a complete wavefield $\mathbf{u}_{kl}$ for one source and frequency. In a time-domain formulation, the PDE (after spatial discretization) can be written as

$$M(\mathbf{m})\mathbf{u}'(t) + S\mathbf{u}(t) = \mathbf{q}(t),$$

which can be solved via some form of time-stepping. The augmented wave-equation involves an extra equation of the form $P\mathbf{u}(t) = \mathbf{d}(t)$. While we can in principle form a large overdetermined system for the wavefield at all timesteps, this is not feasible and a suitable time-stepping strategy will have to be developed to solve the augmented PDE without storing the wavefields for all timesteps.

Other PDE-constrained optimization problems The penalty formulation can be applied to any PDE-constrained optimization problem. We can only expect

an efficient implementation along the lines described in this paper, however, if the PDE is linear in the state variable. In particular, application to multi-parameter seismic inversion (e.g., visco-acoustic, variable density, visco-elastic) is straightfoward.

Dimensionality reduction The computational cost may be reduced by using recently proposed dimensionality reduction techniques that essentially reduce the number of right-hand-sides of the PDE by random subselection or aggregration [18, 19, 20]. These techniques can be directly included in the penalty formulation by introducing new sources and their corresponding data;

$$\widetilde{\mathbf{q}}_{kl'} = \sum_l w_{ll'} \mathbf{q}_{kl}, \quad \widetilde{\mathbf{d}}_{kl'} = \sum_l w_{ll'} \mathbf{d}_{kl},$$

where $l' = [1, 2, \ldots, N_s']$ ($N_s' << N_s$) and $w_{l'l}$ are suitably chosen random weights.

Regularization Regularization penalties on the model $\mathbf{m}$ can be included in a trivial manner. Also, regularization techniques on the (Gauss-Newton) subproblems can be included. This includes for example ℓ_1 regularization as proposed by [21] and ℓ_2 regularization as is used in trust-region methods.

Penalties Penalties other than the ℓ_2 norm on either the data-misfit or the constraint can be included in the formulation (9). However, this will most likely prevent us from solving for the wavefields efficiently via system of equations (17). Still, most alternative penalties may be approximated by a weighted ℓ_2 norm, in which case a system like (17) *can* be formed and solved via an iteratively re-weighted least-squares approach [15].

Nuisance parameters Formulations of the inverse problem may include additional parameters that need to be estimated. An example of this are the source-weights in the seismic applications. The penalty objective, in this case becomes

$$\phi_\lambda(\mathbf{m}, \mathbf{u}, \theta) = \tfrac{1}{2} \sum_{k,l} ||\theta_{kl} P \mathbf{u}_{kl} - \mathbf{d}_{kl}||_2^2 + \lambda^2 ||\theta_{kl} A_k(\mathbf{m}) \mathbf{u}_{kl} - \mathbf{q}_{kl}||_2^2,$$

The wavefield may be solved first from (17) and subsequent minimization over θ for the given wavefields is trivial as a closed-form solution is available:

$$\theta_{kl} = \frac{\mathbf{u}_{kl}^* \left(P^* \mathbf{d}_{kl} + \lambda^2 A_k^* \mathbf{q}_{kl} \right)}{||P\mathbf{u}_{kl}||_2^2 + \lambda^2 ||A_k \mathbf{u}_{kl}||_2^2}.$$

This approach can be generalized to more complicated situations in which a closed-form solution is not available [7].

Contination strategies for λ The experiments presented in this paper were done for a fixed value of λ. To ensure that one really solves the original constrained problem, however, a suitable strategy for increasing λ needs to be developed.

Optimization strategies In this paper we described a Gauss-Newton method to minimize the reduced penalty objective (12). Using the expressions for the gradient and Hessian presented in section 3.1, we can design other Newton variants.

Recently proposed stochastic algorithms aimed at dramatically reducing the cost of PDE-constrained optimization problems by working on a small subsets of the right-hand-sides [22, 19, 23, 20] are directly applicable to this formulation.

6. Conclusions

We have presented a new method for PDE-constrained optimization based on a penalty formulation of the constrained problem. The method relies on solving for the state variables from an augmented system that is comprised of the original discretized PDE and the measurements. The resulting estimates of the state variables can be used to directly estimate the control variable from the PDE via an equation-error approach. The main benefits of this method are: *i)* The state variables for each experiment can be obtained independently and do not have to be stored or updated as part of an iterative optimization procedure, *ii)* the penalty formulation leads to a less non-linear formulation than the reduced approach where the PDE-constraint is eliminated explicitly, and *iii)* the gradient of the objective with respect to the control variable can be computed directly from the state variables, without the need to solve adjoint PDEs. We illustrate the approach on a non-linear seismic inverse problem, showing that the reduced non-linearity leads to significantly better results than the reduced approach at roughly half the computational costs (due to the fact that there is no adjoint equation to solve). Moreover, the penalty approach succesfully mitigates some of the the issues with local minima making the procedure less sensitive to the initial model.

Acknowledgments

The authors wish to thank Dan and Rachel Gordon for valuable discussions on the CARP-CG method. This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BGP, BP, Chevron, CGG, ConocoPhillips, ION, Petrobras, PGS, Total SA, WesternGeco and Woodside.

	# PDE's	Storage	Gauss-Newton update
penalty	$N_s \times N_f$	N_g	solve sparse SPSD system in N_g unknowns
reduced	$2(N_s \times N_f)$	$2N_g$	solve matrix-free linear system in N_g unknowns, requires $N_f \times N_s$ per mat-vec
all-at-once	0	$N_f \times N_s \times N_g$	solve sparse symmetric, possibly indefinite system in $(N_f + N_s + 1) \times N_g$ unknowns

Table 1. Leading order computation and storage costs per iteration of different methods; N_s denotes the number of sources, N_f denotes the number of frequencies and N_g denotes the number of gridpoints. for large-scale seismic inverse problems we typically have $N_s = \mathcal{O}(10^6)$, $N_f = \mathcal{O}(10^1)$ and $N_g = \mathcal{O}(10^9)$

Algorithm 1 Gauss-Newton algorithm based on the penalty formulation

for $t = 0$ to N **do**
  $\mathbf{g}_t = 0$
  $H_t = 0$
  for $k = 1$ to N_f **do**
    for $l = 1$ to N_s **do**
      solve $\begin{pmatrix} \lambda_t A_k(\mathbf{m}_t) \\ P \end{pmatrix} \mathbf{u}_{kl} = \begin{pmatrix} \lambda_t \mathbf{q}_{kl} \\ \mathbf{d}_{kl} \end{pmatrix}$
      $\mathbf{g}_t = \mathbf{g}_t + \lambda_t^2 G_k(\mathbf{m}_t, \mathbf{u}_{kl})^* (A_k(\mathbf{m}_t)\mathbf{u}_{kl} - \mathbf{q}_{kl})$
      $H_t = H_t + (\lambda_t^2 - 1) G_k(\mathbf{m}_t, \mathbf{u}_{kl})^* G_k(\mathbf{m}_t, \mathbf{u}_{kl})$
    end for
  end for
  solve $H_t \Delta \mathbf{m}_t = -\mathbf{g}_t$
  update $\mathbf{m}_{t+1} = \mathbf{m}_t + \Delta \mathbf{m}_t$
  update λ_t.
end for

Algorithm 2 Quasi-Newton algorithm based on the reduced formulation

for $t = 0$ to N **do**
  $\mathbf{g}_t = 0$
  for $k = 1$ to N_f **do**
    for $l = 1$ to N_s **do**
      solve $A_k(\mathbf{m}_t)\mathbf{u}_{kl} = \mathbf{q}_{kl}$
      solve $A_k(\mathbf{m}_t)^* \mathbf{v}_{kl} = -P^*(P\mathbf{u}_{kl} - \mathbf{d}_{kl})$
      $\mathbf{g}_t = \mathbf{g}_t + G_k(\mathbf{m}_t, \mathbf{u}_{kl})^* \mathbf{v}_{kl}$
    end for
  end for
  apply L-BFGS Hessian $\Delta \mathbf{m}_t = \mathsf{lbfgs}(-\mathbf{g}_t, \{\mathbf{m}_{t-M}, \ldots, \mathbf{m}_t\}, \{\mathbf{g}_{t-M}, \ldots, \mathbf{g}_t\})$
  find a suitable steplength α
  update $\mathbf{m}_{t+1} = \mathbf{m}_t + \alpha \Delta \mathbf{m}_t$
end for

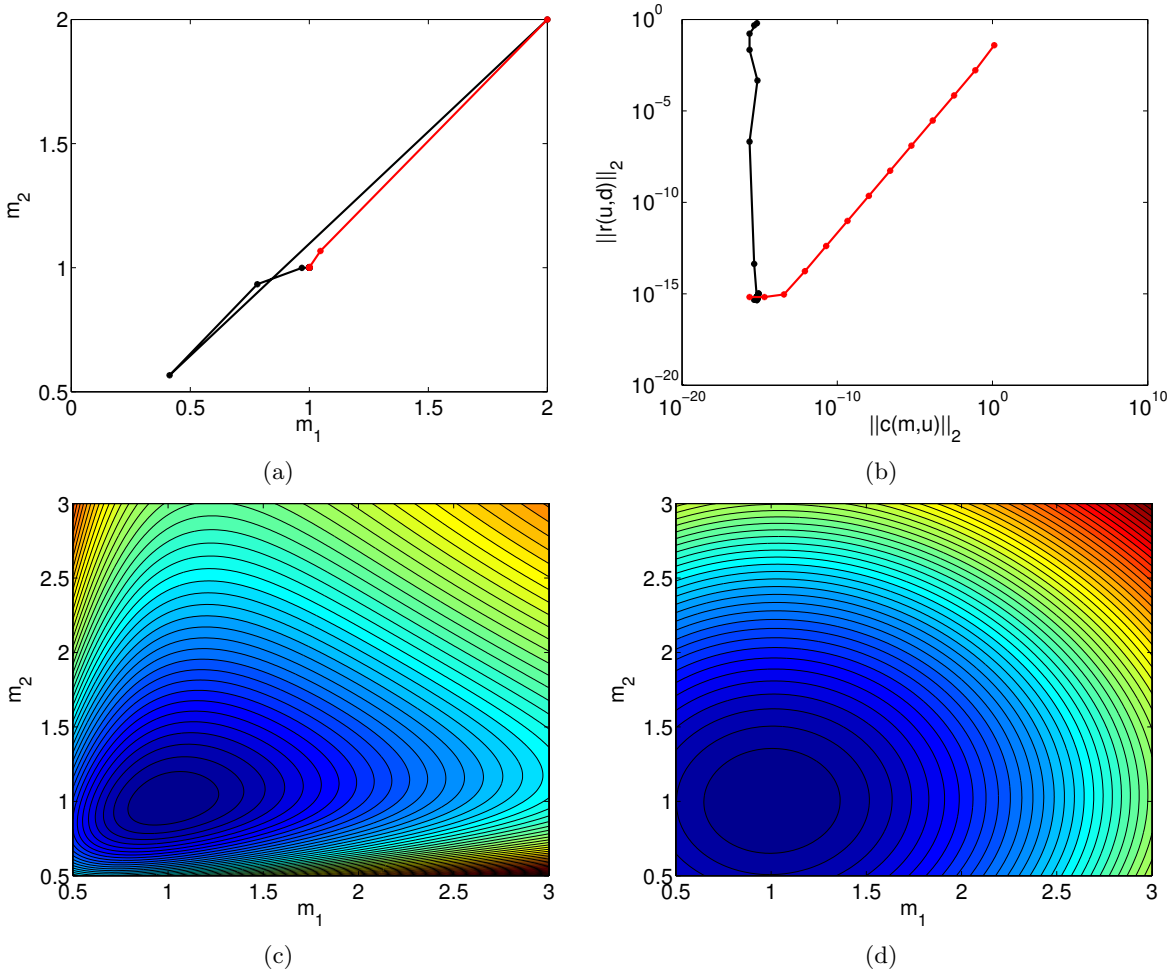


Figure 1. Illustration of the reduced and penalty approach on a simple 4 dimensional test problem. (a) Solution paths and (b) the convergence in terms of $||r||_2$ and $||c||_2$ for both the reduced (black) and penalty approaches (red). The objective functions corresponding to the reduced and penalty approaches are shown in (c) and (d) respectively.

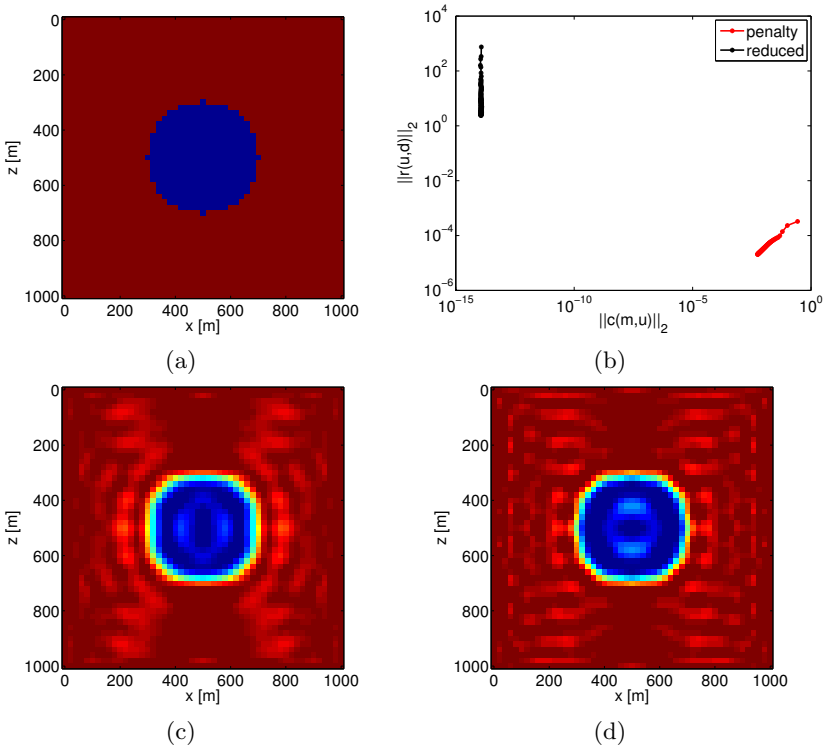


Figure 2. Camembert example. (a) True model, (b) convergence history, (c) reconstruction by penalty approach (d) reconstruction by reduced approach.

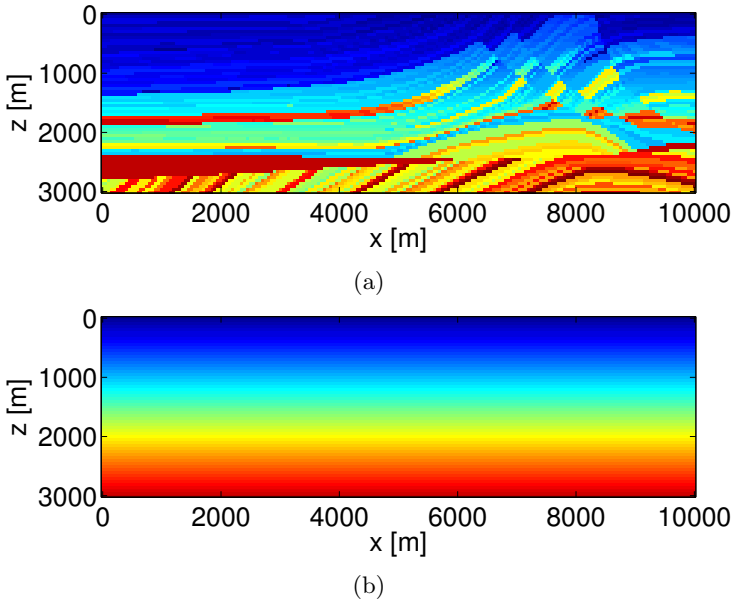
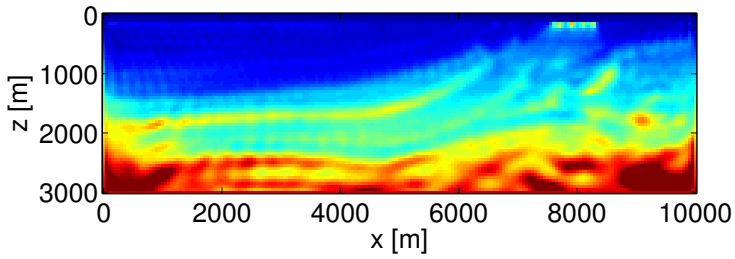
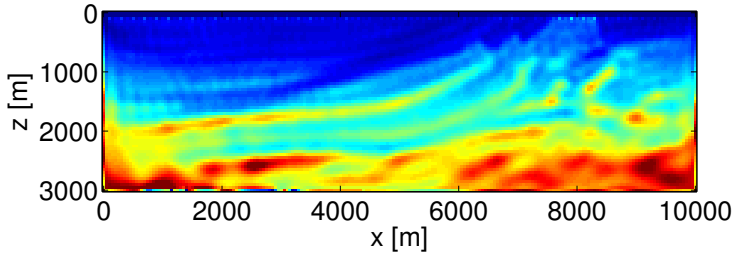


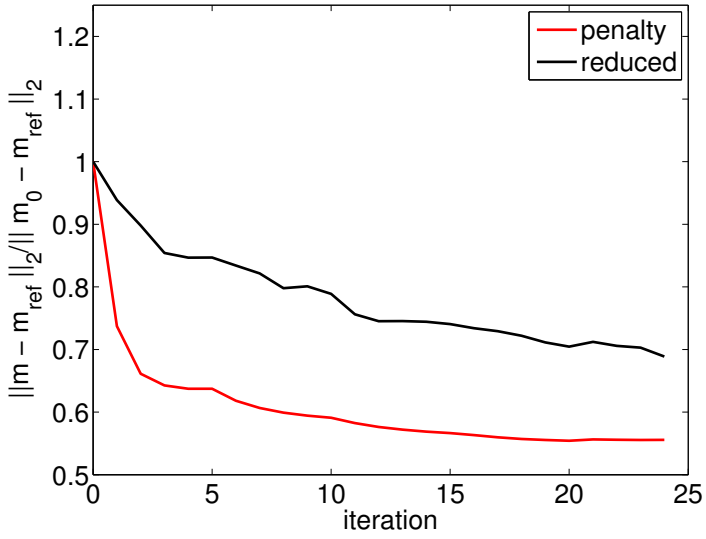
Figure 3. Marmousi example. (a) True model, (b) initial model.



(a)

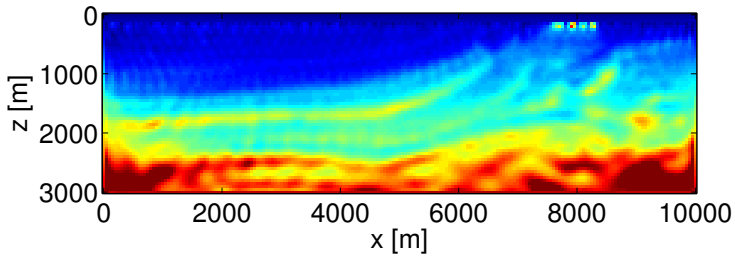


(b)

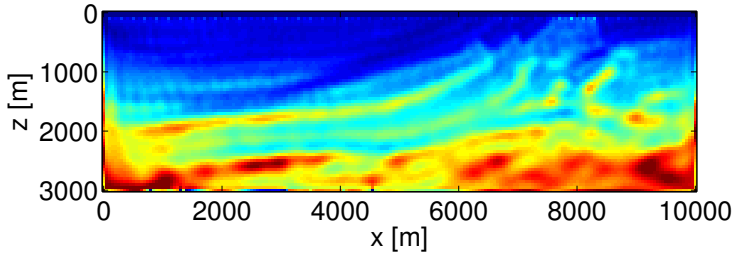


(c)

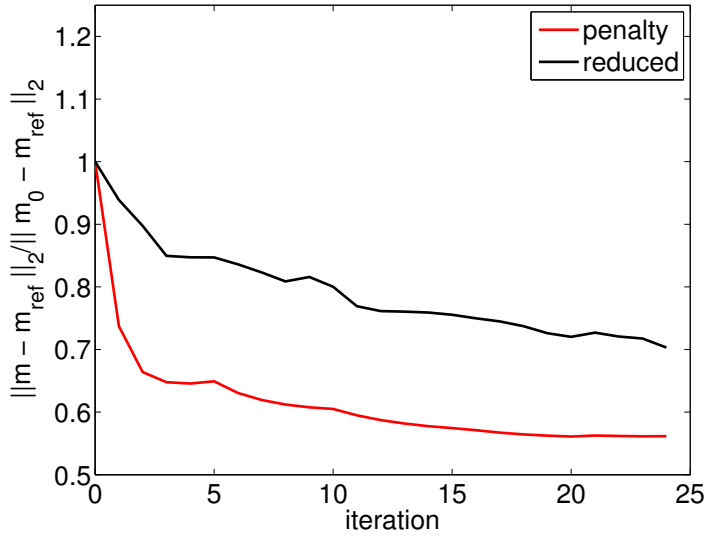
Figure 4. Marmousi example on data without noise. The reconstructed models after inverting frequencies 1,2,3,4,5 Hz. with 5 iterations each with (a) the penalty method and (b) the conventional approach are shown on the same colorscale as the true model. The relative model error at each iteration is shown in (c).



(a)

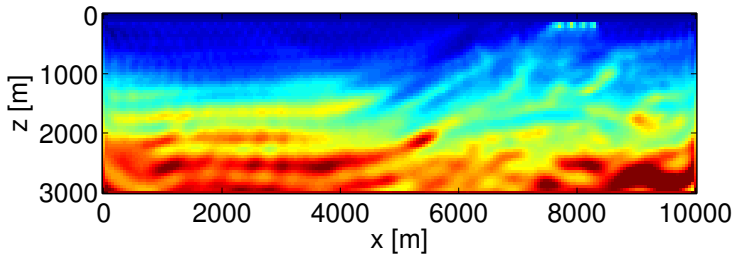


(b)

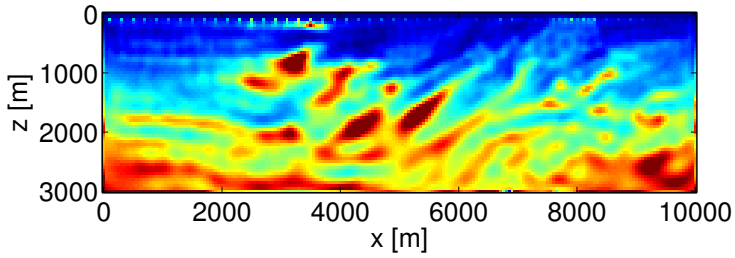


(c)

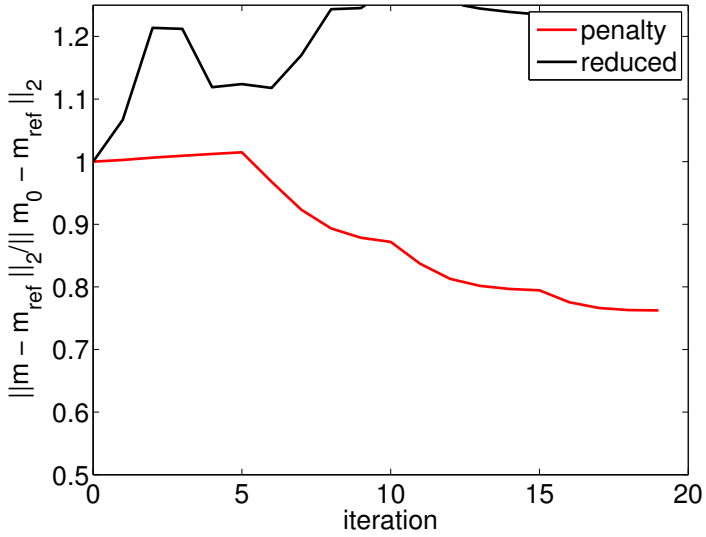
Figure 5. Marmousi example on data with 10% noise. The reconstructed models after inverting frequencies 1,2,3,4,5 Hz. with 5 iterations each with (a) the penalty method and (b) the conventional approach are shown on the same colorscale as the true model. The relative model error at each iteration is shown in (c).



(a)



(b)



(c)

Figure 6. Marmousi example with missing low frequencies (no noise). The reconstructed models after inverting frequencies 2,3,4,5 Hz. with 5 iterations each with (a) the penalty method and (b) the conventional approach are shown on the same colorscale as the true model. The relative model error at each iteration is shown in (c).

[1] George Biros and Omar Ghattas. Parallel Lagrange–Newton–Krylov–Schur Methods for PDE-Constrained Optimization. Part II: The Lagrange–Newton Solver and Its Application to Optimal Control of Steady Viscous Flows. *SIAM Journal on Scientific Computing*, 27(2):714–739, January 2005.

[2] L.T. Biegler. *Real-time PDE-constrained Optimization*. Computational science and engineering. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2007.

[3] Eldad Haber, Uri M. Ascher, and Douglas W. Oldenburg. Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. *Geophysics*, 69(5):1216, 2004.

[4] I Epanomeritakis, V Akçelik, O Ghattas, and J Bielak. A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion. *Inverse Problems*, 24(3):034015, June 2008.

[5] Gassan S Abdoulaev, Kui Ren, and Andreas H Hielscher. Optical tomography as a PDE-constrained optimization problem. *Inverse Problems*, 21(5):1507–1530, October 2005.

[6] Eldad Haber, Uri M Ascher, and Doug Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse Problems*, 16(5):1263–1280, October 2000.

[7] Aleksandr Y Aravkin and Tristan van Leeuwen. Estimating nuisance parameters in inverse problems. *Inverse Problems*, 28(11):115016, 2012.

[8] R.G. Richter. Numerical Identification of a Spatially Varying Diffusion Coefficient. *Mathematics of Computation*, 36(154):375–386, 1981.

[9] J. Eckstein. Augmented Lagrangian and Alternating Direction Methods for Convex Optimization: A Tutorial and Some Illustrative Computational Results. *Rutcor Research Report*, 32-2012, 2012.

[10] D.P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.

[11] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2nd edition, 2003.

[12] J Virieux and S Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

[13] G.R. Pratt, Changsoo Shin, and G.J. Hicks. Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophysical Journal International*, 133(2):341–362, May 1998.

[14] Carey Bunks. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457, September 1995.

[15] Å. Björck and T. Elfving. Accelerated projection methods for computing pseudoinverse solutions of systems of linear equations. *BIT*, 19(2):145–163, June 1979.

[16] Dan Gordon and Rachel Gordon. Robust and highly scalable parallel solution of the Helmholtz equation with large wave numbers. *Journal of Computational and Applied Mathematics*, 237(1):182–196, January 2013.

[17] Yair Censor, Paul P. B. Eggermont, and Dan Gordon. Strong underrelaxation in Kaczmarz's method for inconsistent systems. *Numerische Mathematik*, 41(1):83–92, February 1983.

[18] Eldad Haber, Matthias Chung, and Felix Herrmann. An Effective Method for Parameter Estimation with PDE Constraints with Multiple Right-Hand Sides. *SIAM Journal on Optimization*, 22(3):739–757, July 2012.

[19] Michael P. Friedlander and Mark Schmidt. Hybrid Deterministic-Stochastic Methods for Data Fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, May 2012.

[20] Tristan van Leeuwen and Felix J. Herrmann. Fast waveform inversion without source-encoding. *Geophysical Prospecting*, pages no–no, July 2012.

[21] X. Li, A.Y. Aravkin, T. van Leeuwen, and F.J. Herrmann. Fast randomized full-waveform inversion with compressive sensing. *Geophysics*, 77(3):A13, 2012.

[22] Jerome R. Krebs, John E. Anderson, David Hinkley, Anatoly Baumstein, Sunwoong Lee, Ramesh Neelamani, and Martin-Daniel Lacasse. Fast full wave seismic inversion using source encoding. *Geophysics*, 28(1):2273–2277, 2009.

[23] Eldad Haber, Matthias Chung, and Felix Herrmann. An Effective Method for Parameter Estimation with PDE Constraints with Multiple Right-Hand Sides. *SIAM Journal on Optimization*, 22(3):739–757, July 2012.