Geophysical Prospecting (200?) **XX,** 000–000

Full Waveform Inversion and the truncated Newton method: quantitative imaging of complex subsurface structures

L. Métivier[1,2], F. Bretaudeau[2], R. Brossier[2], S. Operto[3], and J. Virieux[2]

[1] *LJK, CNRS, Université de Grenoble I, BP 53, 38041 Grenoble cedex 09, France*

[2] *ISTerre, Université de Grenoble I, BP 53, 38041 Grenoble Cedex 09 France*

[3] *Géoazur, Université de Nice Sophia-Antipolis, CNRS, IRD, OCA, Villefranche-sur-mer, France*

Accepted 2008 ?? ??. Received 2008 June ??; in original form 2008 June ??

SUMMARY

Full Waveform Inversion (FWI) is a powerful tool for quantitative seismic imaging from wide-azimuth seismic data. The method is based on the minimization of the misfit between observed and simulated data. This amounts to the resolution of a large-scale nonlinear minimization problem. The inverse Hessian operator plays a crucial role in this reconstruction process. Accounting accurately for the effect of this operator within the minimization scheme should correct for illumination deficits, restore the amplitude of the subsurface parameters, and help to remove artifacts generated by energetic multiple reflections. Conventional preconditioned gradient-based minimization methods only roughly approximate the effect of this operator. We are interested in this study to another class of minimization methods, named as truncated Newton methods. These methods are based on the computation of the model update through a matrix-free conjugate gradient resolution of the Newton linear system. The aim of this study is to present a feasible implementation of this method for the FWI problem, based on a second-order adjoint state

formulation for the computation of Hessian-vector products. We compare this method with the nonlinear conjugate gradient and the l-BFGS method within the context of 2D acoustic frequency FWI for the reconstruction of P-wave velocity models. Two test cases are investigated. The first is the synthetic BP 2004 model, representative of the Gulf Of Mexico geology with high velocity contrasts associated with the presence of salt structures. The second is a 2D real data-set from the Valhall oil field in North sea. These tests emphasize the interesting properties of the truncated Newton method regarding conventional optimization methods within the context of FWI.

Key words: Full Waveform, Theory, Computing aspects, Numerical study, Imaging

1 INTRODUCTION

Full Waveform Inversion (FWI) is a powerful seismic imaging tool, dedicated to quantitative estimations of subsurface parameters such as P-wave and S-wave velocities, density, impedance, or anisotropy parameters. The method is based on the minimization of a misfit function that measures the distance between recorded seismic data and predicted data computed through the numerical simulation of wave propagation. An initial subsurface model is iteratively updated to produce the final estimation.

The formalism of the FWI method has been introduced by Lailly (1983) and Tarantola (1984), based on a time domain discretization of the wave equation. Its first application to 2D synthetic data in the acoustic approximation was performed by Gauthier *et al.* (1986). Later on, a hierarchical frequency domain approach has been introduced by Pratt for cross-hole tomography (Pratt and Worthington 1990; Pratt 1990). During the past ten years, the simultaneous advances in acquisition systems (development of wide-azimuth seismic surveys for instance) and high performance computing facilities have made possible the successful application of FWI to surface data, both in the 2D acoustic or 2D elastic approximation to reconstruct one or several parameters (Operto *et al.* 2004, 2005; Ravaut *et al.* 2004; Gao *et al.* 2006; Brossier *et al.* 2009; Prieux *et al.* 2011; Plessix *et al.* 2012; Prieux *et al.* 2013a,b). Applications of FWI to real surface data in the 3D acoustic approximation have also been

performed (Sirgue *et al.* 2008; Vigh *et al.* 2010; Plessix and Perkins 2010; Plessix *et al.* 2012). For an overview on the FWI methodology and its applications to synthetic and real case studies, the reader is referred to the survey proposed by Virieux and Operto (2009).

The simplest optimization methods used in the context of FWI are gradient-based algorithms, such as the steepest descent or the nonlinear conjugate gradient algorithms. From a given initial model, the sequence of updates yielding the final model is defined by the gradient of the misfit function. However, it is well known that these methods have poor convergence properties.

Conversely, Newton-based methods possess better convergence properties (superlinear to quadratic convergence rate). These methods are based on a model update given by the multiplication of the gradient by the inverse Hessian operator*. The importance of this operator in the context of FWI has been emphasized by Pratt *et al.* (1998). The inverse Hessian operator acts as a deconvolution operator that accounts for the limited bandwidth of the seismic data and corrects for the loss of amplitude of poorly illuminated subsurface parameters. In addition, it helps to remove artifacts that the second order reflected waves may generate on the model update.

However, because of the large-scale aspect of the FWI problem, which easily involves millions of discrete unknowns in 2D up to billions discrete unknowns in 3D, explicit computation of the inverse Hessian operator is beyond current computational capabilities. As a consequence, research efforts have been mainly directed toward direct approximation of this operator.

A first possibility consists in approximating the diagonal of the Hessian. For instance, Operto *et al.* (2006) compute the diagonal terms of the Gauss-Newton approximation of the Hessian, which requires some extra-computation. A cheaper strategy based on the so-called pseudo-Hessian operator is also proposed by Shin *et al.* (2001).

A second possibility consists in approximating the inverse Hessian operator using previous values of the gradient of the misfit function. Among this class of methods, known as

*The Hessian is the matrix of the second-order derivatives of the misfit function.

quasi-Newton methods, the l-BFGS method is quite popular (Nocedal and Wright 2006; Byrd *et al.* 1995). Instead of approximating only the diagonal elements, a positive definite approximation of the full inverse Hessian is computed.

The approximation of the diagonal elements of the Hessian and the l-BFGS strategy can be combined to produce a more accurate approximation of the inverse Hessian operator. Indeed, the accuracy of the l-BFGS approximation of the inverse Hessian operator can be improved when based on a first estimation of the inverse Hessian operator. This method has been applied in the framework of 2D elastic FWI: Brossier *et al.* (2009) implemented a l-BFGS optimization using a diagonal pseudo-Hessian as initial guess. This method shows good convergence properties compared to preconditioned non-linear conjugate-gradient. The diagonal estimation can even be updated along the iterations (Nocedal and Wright 2006).

The truncated Newton represents an alternative to these already described optimization methods. At each iteration, the model update is computed as an approximate solution of the Newton equations through a linear iterative solver (namely a conjugate gradient solver) (Nash 2000). Implemented in a "matrix-free" fashion, this iterative solver only requires to compute Hessian-vector products. It is not necessary to form the Hessian operator explicitly.

Although this class of methods is well known in the numerical optimization community, the application of truncated Newton method in the FWI context has still not been fully investigated. Given the importance of the inverse Hessian operator in the FWI reconstruction scheme, we believe that this method could benefit from a better approximation of the inverse Hessian effect, and provides more accurate subsurface parameter estimations than standard optimization schemes. Therefore, the ambition of this study is to focus on the two following points:

- Designing a feasible implementation of the truncated Newton method for FWI in terms of computational time.
- Compare the performance of this method with conventional methods (nonlinear conjugate gradient, l-BFGS) on two realistic test cases.

In Section 2, we describe in more details the principle of the truncated Newton method

compared to preconditioned gradient-based methods. In Section 3, we present how the method can be implemented efficiently in the FWI context. In Section 4, we present two application cases. The first case is based on the synthetic BP 2004 model, partly inspired from the deep water Gulf of Mexico geology. The presence of salt structures in a marine environment is responsible for high velocity contrasts which make the seismic imaging task difficult. The second test case concerns a 2D line of a real Ocean Bottom Cable (OBC) data-set, acquired in a shallow-water environment at the Valhall oil field, in the North Sea. This test case is investigated to emphasize how the truncated Newton method behaves when the data is noise-contaminated. Conclusion and perspectives are given in Section 5.

2 THE TRUNCATED NEWTON SCHEME

For the sake of clarity, the mathematical results presented in this section are formulated for a number of sources equal to 1. The extension to a multi-source context is straightforward, as only explicit summations over the sources are required.

2.1 Problem settings

We consider the frequency-domain forward problem

$$A(m)u = s \tag{1}$$

where

- $m \in \mathcal{M}$ denotes the subsurface model;
- $u \in \mathcal{W}$ is the complex-valued seismic wavefield;
- s is a source term;
- $A(m)$ is a discretized partial differential operator related to the wave equation (from the acoustic dynamics to the visco-elastic anisotropic dynamics).

The FWI problem is defined as the minimization over the parameter space of a distance between the data predicted by the forward problem and the recorded data.

$$\min_{m \in \mathcal{M}} f(m) = \frac{1}{2} \| Ru(m) - d \|^2, \tag{2}$$

where

- $u(m)$ is the solution of the forward problem (1) for the source term and the subsurface parameter s and m;
- R is a mapping of the wavefield to the receivers locations;
- d is the data set associated to the source s;
- $\|.\|$ is a norm in the data space $\mathcal{D}$.

For practical reasons, the use of the L^2 norm is common. However, more general L^p norm could be also selected (Tarantola 2005). The L^1 norm is, for instance, a good choice when high-amplitude noise (outliers) corrupts the data (Brossier *et al.* 2010). More complex measurements of the distance between data sets can also be proposed to mitigate the sensitivity of FWI to the initial model. This is, however, beyond the scope of the work presented here.

2.2 Preconditioned gradient based-methods

From a numerical point of view, FWI is a large-scale nonlinear minimization problem. The high number of discrete parameters prevents from using global or semi-global optimization techniques to solve this problem. Therefore, we focus on local optimization methods, which are based on the following recurrence: from an initial guess m_0, a sequence m_k is computed such that

$$m_{k+1} = m_k + \gamma_k \Delta m_k, \tag{3}$$

where Δm_k is the model update and γ_k is a scalar parameter computed through a linesearch or a trust-region procedure (Bonnans *et al.* 2006; Nocedal and Wright 2006).

Within the framework of Newton algorithms, the increment Δm_k is defined as

$$H(m_k) \Delta m_k = -\nabla f(m_k). \tag{4}$$

where $H(m)$ denotes the Hessian operator and $\nabla f(m)$ is the gradient of the misfit function with respect to the model parameter m.

The exact resolution of this linear system at each iteration is beyond current computational capacities in the context of FWI. This is why many large-scale optimization schemes rely on an approximation of the inverse Hessian operator. These schemes are based on the computation of Δm_k as

$$\Delta m_k = -P_k \nabla f(m_k), \tag{5}$$

where P_k is an approximation of the inverse Hessian operator $H(m_k)^{-1}$. We shall refer to these schemes as preconditioned gradient-based methods in what follows.

Preconditioned steepest-descent, nonlinear conjugate gradient and l-BFGS methods fall into the scope of these minimization algorithms. These methods only differ in the way the matrix P_k is computed. The preconditioned steepest-descent uses a "direct" approximation of the inverse Hessian (for instance an approximation of its diagonal (Operto *et al.* 2006; Shin *et al.* 2001)). The nonlinear conjugate gradient method and the l-BFGS method (Byrd *et al.* 1995) use former values of the gradient to estimate the inverse Hessian operator[†]. As mentioned in the introduction, nonlinear conjugate gradient and l-BFGS techniques can be used in combination of a prior estimation of the diagonal to compute the matrix P_k (Nocedal and Wright 2006).

2.3 Truncated Newton algorithm

Instead of building an approximation P_k, the linear system (4) can be solved using a matrix-free version of the conjugate gradient algorithm (Saad 2003). This only requires the capability of computing Hessian-vector products $H(m_k)v$ where v is an arbitrary vector in the model space $\mathcal{M}$. The truncated Newton method thus results in a two nested loops algorithm:

[†]Only the value of the gradient at the previous iteration is used for the nonlinear conjugate gradient, while l previous values are used for the l-BFGS algorithm.

• The external loop consists in the iterative update of the current subsurface parameter estimation, following equation (3).

• The internal loop consists in the iterative resolution of the linear system (4), in order to compute the model update Δm_k. An approximate solution of this system is computed, as described in the following

The advantage of using a truncated Newton method is two-fold:

• The approximation of the inverse Hessian operator is local, while the approximation computed in l-BFGS or nonlinear conjugate gradient method depends on previous iterations. For minimizing strongly nonlinear misfit functions, this can be advantageous, since the information on the local curvature carried out by previous iterate can be erroneous. Setting the memory parameter l is thus known to be a difficult issue as it is strongly problem dependent (Nocedal and Wright 2006).

• The truncation strategy in the inner loop consists in accounting only for the higher eigenvalues of the inverse Hessian operator. This has an intrinsic regularization effect on the computation of the model update.

In addition, the truncated Newton method offers the possibility of using the approximations of the inverse Hessian operator which have been developed in the context of preconditioned gradient-based methods as preconditioners of the inner linear systems, yielding the resolution of the preconditioned linear system

$$P_k H(m_k) \Delta m_k = -P_k \nabla f(m_k), \tag{6}$$

at each outer iteration. Hence, this method could be seen as a complementary evolution of previously developed minimization strategies.

However, compared to the preconditioned gradient-based methods, the truncated Newton method involves an additional computational expense associated with the iterative resolution of the linear system (4). This should be balanced by an improvement of the convergence speed of the external loop. Therefore, as mentioned by Nash (2000), an efficient implementation

of the truncated Newton method relies on the reduction of this additional cost. This can be achieved in the context of FWI by:

- Defining second-order adjoint formulas for the efficient computation of Hessian vector products $H(m_k)v$ for any $v \in \mathcal{M}$.
- Defining an adapted stopping criterion for the approximate resolution of the linear system (4) to limit as much as possible the number of iterations of the conjugate gradient algorithm at each step of the external loop.
- Using an appropriate preconditioner to accelerate the convergence of the resolution of the linear system (4).

The truncated Newton strategy can also be implemented using the Gauss-Newton approximation $B(m)$ of the Hessian operator $H(m)$, which consists in neglecting the second-order terms of the Hessian operator. The Gauss-Newton approximation is

$$B(m) = J(m)^\dagger R^\dagger R J(m), \tag{7}$$

where $J(m)$ is the Jacobian matrix:

$$J(m) = \partial_m u(m), \tag{8}$$

and † denotes the transpose conjugate operator.

The Gauss-Newton method is appealing for the following reasons:

- $B(m)$ is positive definite by construction, therefore the conjugate gradient algorithm is well adapted for the resolution of the linear system (4);
- Close to the solution, the residuals are small, therefore the matrix $B(m)$ should be a good approximation of the full Hessian matrix $H(m)$, since the second-order part which is neglected in the Gauss-Newton approximation is proportional to the residuals.

From the implementation point of view, the truncated Gauss-Newton procedure differs only in the computation of matrix-vector products $B(m)v$ instead of $H(m)v$ from the truncated Newton method.

In the next section, we investigate how the truncated Newton and Gauss-Newton methods can be efficiently implemented in the context of FWI.

3 IMPLEMENTATION OF THE TRUNCATED NEWTON METHOD

3.1 Computation of the model update

The resolution of the linear systems (4) first requires to compute the right-hand side $-\nabla f(m_k)$. The computation of $\nabla f(m_k)$ is efficiently achieved through the first-order adjoint-state method, introduced by Lions (1968). For the sake of generality, we will use here the notations introduced by Plessix (2006) in his review of the adjoint-state technique for seismic imaging. In this framework, the forward problem should be rewritten as

$$F(m,u) = 0, \tag{9}$$

where

$$F(m,u) = A(m)u - s. \tag{10}$$

3.1.1 Computation of the gradient and first-order adjoint method

In the context of FWI, the first-order adjoint method amounts to compute $\nabla f(m_k)$ as the zero-lag cross-correlation of the incident wavefield and the adjoint wavefield (Chavent 1974). The adjoint wavefield is defined as the solution of

$$\frac{\partial F(m,u)^\dagger}{\partial u}\lambda = R^\dagger(d - Ru(m)), \tag{11}$$

where $u(m)$ is the solution of (1). Based on the definition of the adjoint state λ, the gradient can be expressed as

$$\nabla f(m) = \mathbf{Re}\left(\frac{\partial F(m,u)^\dagger}{\partial m}\lambda(m)\right), \tag{12}$$

where **Re** denotes the real part operator.

Using this method, the computation cost of the gradient amounts to the resolution of two wave propagation problems per shot: one forward problem (1) and one adjoint problem (11).

3.1.2 Computation of $H(m)v$ through second-order adjoint method

The computation of Hessian-vector products through second-order adjoint methods is a topic that has already been investigated in the field of data assimilation and weather forecasting (Wang *et al.* 1992). However, the control variable in data assimilation is an initial condition for the system, whereas in seismic imaging, the control variable is a coefficient of the partial differential equation that describes the system. A formula for the computation of Hessian-vector products have been given by Pratt *et al.* (1998) in the seismic imaging context, for the Gauss-Newton approximation in the discrete frequency domain. Fichtner and Trampert (2011) also propose more general formulas for the computation of Hessian kernels. Epanomeritakis *et al.* (2008) give the formulas corresponding to the elastic case.

We propose here a general framework in the frequency domain for deriving these formulas, with no assumption on the discretization and the kind of partial differential equations that are used for the wave propagation description. The method can be straightforwardly adapted to the time-domain formulation by adding proper initial and final conditions, and boundary conditions. In addition, no prior assumption on the linearity of the forward problem is required, as it was the case in Pratt *et al.* (1998) and a previous work (Métivier *et al.* 2013).

We first define the functional $h_v(m)$ as

$$h_v(m) = (\nabla f(m), v)_{\mathcal{M}}, \tag{13}$$

where $(.,.)_{\mathcal{M}}$ denotes the scalar product on the parameter space $\mathcal{M}$. By definition, the gradient of the functional $h_v(m)$ is

$$\nabla h_v(m) = H(m)v. \tag{14}$$

We use the Lagrangian formalism to compute $\nabla h_v(m)$. We introduce the Lagrangian oper-

ator $L(m,u,\lambda,g,\mu_1,\mu_2,\mu_3)$ such that

$$\begin{aligned}
&L_v(m,u,\lambda,g,\mu_1,\mu_2,\mu_3)_{\mathcal{W}} = \\
&(g,v)_{\mathcal{M}} + \\
&\mathbf{Re}\left(g - \tfrac{\partial F(m,u)^\dagger}{\partial m}\lambda, \mu_1\right)_{\mathcal{W}} + \\
&\mathbf{Re}\left(\tfrac{\partial F(m,u)^\dagger}{\partial u}\lambda - R^\dagger\left(d - Ru\right), \mu_2\right)_{\mathcal{W}} + \\
&\mathbf{Re}\left(F(m,u), \mu_3\right)_{\mathcal{W}}.
\end{aligned} \tag{15}$$

In this expression, $(.,.)_{\mathcal{W}}$ denotes the scalar product in the wavefield space $\mathcal{W}$. In addition, u and λ play the role of the incident and adjoint wavefields respectively, while g plays the role of the gradient. Let $\overline{u}(m), \overline{\lambda}(m), \overline{g}(m)$ satisfying the constraints

$$F(m,\overline{u}) = 0 \quad \frac{\partial F(m,\overline{u})^\dagger}{\partial u}\overline{\lambda} = R^\dagger\left(d - R\overline{u}\right) \quad \overline{g} = \frac{\partial F(m,\overline{u})^\dagger}{\partial m}\overline{\lambda}. \tag{16}$$

We have

$$L_v(m,\overline{u},\overline{\lambda},\overline{g},\mu_1,\mu_2,\mu_3) = h_v(m), \tag{17}$$

and

$$\frac{\partial L_v}{\partial m}(m,\overline{u},\overline{\lambda},\overline{g},\mu_1,\mu_2,\mu_3) = \nabla h_v(m). \tag{18}$$

In addition

$$\begin{aligned}
&\frac{\partial L_v}{\partial m}(m,\overline{u},\overline{\lambda},\overline{g},\mu_1,\mu_2,\mu_3) = \\
&\mathbf{Re}\left(\left(\frac{\partial^2 F(m,u)^\dagger}{\partial m^2}\lambda\right)^\dagger \mu_3\right) + \\
&\mathbf{Re}\left(\left(\frac{\partial^2 F(m,u)^\dagger}{\partial m \partial u}\lambda\right)^\dagger \mu_2\right) + \\
&\mathbf{Re}\left(\frac{\partial F(m,u)^\dagger}{\partial m}\mu_1\right) + \\
&\frac{\partial L_v}{\partial g}(\overline{m},\overline{u},\overline{g},\mu_1,\mu_2,\mu_3)\frac{\partial \overline{g}(m)}{\partial m} + \\
&\frac{\partial L_v}{\partial \lambda}(\overline{m},\overline{u},\overline{g},\mu_1,\mu_2,\mu_3)\frac{\partial \overline{\lambda}(m)}{\partial m} + \\
&\frac{\partial L_v}{\partial u}(\overline{m},\overline{u},\overline{g},\mu_1,\mu_2,\mu_3)\frac{\partial \overline{u}(m)}{\partial m}
\end{aligned} \tag{19}$$

We can choose $\mu_i, i = 1, 3$ such that for any perturbations $dg, d\lambda, du$ we have

$$\left\{\begin{array}{rcl} \dfrac{\partial L_v}{\partial g}(\overline{m},\overline{u},\overline{g},\mu_1,\mu_2,\mu_3).dg &=& 0 \\ \dfrac{\partial L_v}{\partial \lambda}(\overline{m},\overline{u},\overline{g},\mu_1,\mu_2,\mu_3).d\lambda &=& 0 \\ \dfrac{\partial L_v}{\partial u}(\overline{m},\overline{u},\overline{g},\mu_1,\mu_2,\mu_3).du &=& 0. \end{array}\right. \tag{20}$$

This yields

$$\left\{\begin{array}{rcl} \mu_1 &=& -v \\ \dfrac{\partial F(m,u)}{\partial u}\mu_2 &=& \dfrac{\partial F(m,u)}{\partial_m}\mu_1 \\ \dfrac{\partial F(m,u)}{\partial u}^{\dagger}\mu_3 &=& \left(\dfrac{\partial^2 F(m,u)}{\partial u^2}^{\dagger}\lambda\right)^{\dagger}\mu_2 - \\ & & \left(\dfrac{\partial^2 F(m,u)}{\partial u \partial m}^{\dagger}\lambda\right)^{\dagger}\mu_1 - R^{\dagger}R\mu_2. \end{array}\right. \tag{21}$$

Using the relation (10) which defines our forward problem, we see that

$$\frac{\partial F(m,u)}{\partial u} = A(m), \quad \frac{\partial^2 F(m,u)}{\partial u^2} = 0. \tag{22}$$

This yields

$$\left\{\begin{array}{rcl} A(m)\mu_2 &=& -\dfrac{\partial F(m,u)}{\partial m}v \\ A(m)^{\dagger}\mu_3 &=& -\left(\dfrac{\partial^2 F(m,u)}{\partial u \partial m}^{\dagger}\lambda\right)^{\dagger}\mu_1 - R^{\dagger}R\mu_2. \end{array}\right. \tag{23}$$

We thus obtain the three-terms Hessian-vector product formula

$$\begin{array}{rcl} H(m)v &=& \mathbf{Re}\left(\left(\dfrac{\partial^2 F(m,u)^{\dagger}}{\partial m^2}\lambda\right)^{\dagger}\mu_1\right) + \\ & & \mathbf{Re}\left(\left(\dfrac{\partial^2 F(m,u)^{\dagger}}{\partial m \partial u}\lambda\right)^{\dagger}\mu_2\right) + \\ & & \mathbf{Re}\left(\dfrac{\partial F(m,u)^{\dagger}}{\partial m}\mu_3\right). \end{array} \tag{24}$$

where

$$\frac{\partial^2 F(m,u)}{\partial m \partial u} \tag{25}$$

is the radiation matrix.

The adjoint wavefield $\overline{\mu}_1$ is equal to v. The computation of one Hessian-vector products thus requires to compute the the radiation wavefield $\overline{u}(m)$ the adjoint wavefield $\overline{\lambda}(m)$, as

well as two additional wavefield $\overline{\mu}_2(m)$ and $\overline{\mu}_3(m)$. The two latter are computed through the resolution of respectively one forward and one adjoint problem with new source terms.

3.1.3 Gauss-Newton approximation

From the definition of the Gauss-Newton operator $B(m)$ (7), we see that only first-order derivatives of the wavefield $\overline{u}(m)$ with respect to the model parameter m (also called Jacobian matrix or Fréchet derivatives) are taken into account. The computation of $B(m)v$ can thus be derived from equations (23) and (24) by neglecting the contribution of all the second-order terms. This yields the following simplifications:

$$\left\{ \begin{array}{lcl} A(m)\mu_2 & = & -\dfrac{\partial F(m,u)}{\partial m} v \\ A(m)^\dagger \mu_3 & = & -R^\dagger R \mu_2, \end{array} \right. \tag{26}$$

and the following formula for the Gauss-Newton approximation

$$B(m)v = \mathbf{Re}\left(\frac{\partial F(m,u)^\dagger}{\partial m} \mu_3 \right). \tag{27}$$

This formula is consistent with the one derived in Pratt *et al.* (1998) or Métivier *et al.* (2013). Note however that in these two articles, the method used to derive this formula relies on the assumption of a linear forward problem, which is not the case here.

3.1.4 Computation cost

Consider the resolution of the linear system (4) for the computation of the model update Δm_k. The right hand side in (4) is the opposite gradient $-\nabla f(m)$. Using the first-order adjoint state method, it can be computed at the expense of one forward problem for $\overline{u}(m)$ and one adjoint problem for $\overline{\lambda}(m)$.

Provided these wavefields can be stored, the computation of $H(m)v$ of $B(m)v$ only requires to solve two additional problems: one forward problem for the computation of $\overline{\mu_2}(m)$, and one adjoint problem for the computation of $\overline{\mu_3}(m)$.

Note that in practice, the computation of $B(m)v$ does not make use of the adjoint wavefield $\overline{\lambda}(m)$. However, its computation is imposed by the computation of the gradient,

which is the right-hand side of the linear system (4). The computation cost of the action of the Hessian operator or its Gauss-Newton approximation on an arbitrary vector is thus *the same* in terms of number of wave equations to be solved. Nonetheless, only the incident wavefield $\overline{u}(m)$ have to be stored in the Gauss-Newton approximation.

The overall computation cost of the truncated Newton method in terms of wave propagation simulation is thus given by

$$C = N_{ext}(2 + 2 \times N_{int,k}), \tag{28}$$

where N_{ext} is the total number of iterations of the external loop and $N_{int,k}$ is the number of conjugate gradient iterations performed at the k^{th} iteration of the external loop. The choice of an appropriate stopping criterion for the conjugate gradient helps to reduce the quantities $N_{int,k}$. Note that this computational cost depends on the ability of storing $\overline{u}(m)$ and $\overline{\lambda}(m)$. This is a reasonable assumption for 2D applications. This is a more complex issue in 3D: more sophisticated memory and I/O management methods should be required. In particular, the increase of the number of sources is critical, as one incident and one adjoint wavefield $\overline{u}(m)$ and $\overline{\lambda}(m)$ have to be stored per sources. The use of source encoding techniques should therefore be crucial in this context.

3.2 Definition of an adapted stopping criterion

The Newton method is an iterative minimization of local quadratic expansions of the misfit function. Indeed, the resolution of the system (4) amounts to the minimization of the quadratic form

$$q_k(\Delta m) = f(m_k) + (\nabla f(m_k), \Delta m) + \frac{1}{2}(H(m_k)\Delta m, \Delta m). \tag{29}$$

The definition of the stopping criterion for the truncated Newton method is related to the accuracy of these quadratic expansions. The idea exploited by Eisenstat and Walker (1994) is the following. Consider a stopping criterion for the CG iterations of the form

$$\|H(m_k)\Delta m_k + \nabla f(m_k)\| \leq \eta_k \|\nabla f(m_k)\|. \tag{30}$$

where η_k is the forcing term. The role devoted to this forcing term is to account for the accuracy of the local quadratic approximation. When this accuracy increases, η_k should decrease, so as to require to solve the linear system (4) more accurately. Conversely, when the accuracy decreases, η_k should increase, so as to allow a less precise resolution of the linear system (4). This is achieved by defining η_k as the measure of the distance between the first order Taylor expansion of the gradient at the iteration $k-1$ and the gradient at iteration k:

$$\eta_k = \frac{\|\nabla f(m_k) - \nabla f(m_{k-1}) - \gamma_{k-1} H(m_{k-1}) \Delta m_{k-1}\|}{\|\nabla f(m_{k-1})\|}. \tag{31}$$

The definition of the stopping criterion is complemented with an appropriate strategy to deal with the detection of negative eigenvalues of the Hessian operator. The conjugate gradient algorithm is designed for the resolution of symmetric definite positive systems. However, far from the solution, the full Hessian operator $H(m_k)$ may be indefinite. Therefore, the iterative construction of the solution of (4) in the Krylov space

$$\left\{ r_0, H(m_k) r_0, H(m_k)^2 r_0, \dots \right\}, \tag{32}$$

where r_0 is the initial residual

$$H(m_k) \Delta m_k^0 + \nabla f(m_k), \tag{33}$$

may use a ascent direction associated with a negative eigenvalue of the operator $H(m_k)$. In this case, the linear iterations are stopped and the last value of the model update Δm_k which is computed is returned. If this ascent direction is met at the very first linear iteration, the steepest-descent direction is returned. This strategy, proposed by Eisenstat and Walker (1994), ensures superlinear convergence properties far from the solution, and quadratic convergence when entering the attraction basin of the minimum.

3.3 Preconditioning

In order to speed-up the convergence of the resolution of the linear system (4), it is natural to introduce a preconditioning matrix. In this study we focus on the special preconditioner related to the FWI problem proposed by Shin *et al.* (2001). The diagonal elements of the Gauss-

Newton part of the Hessian $B(m)$ are approximated using the pseudo-Hessian approach, which appears to be relevant for surface seismic survey. Let us denote $\alpha_j(m) = \partial_{m_j} u(m)$ the column j of the Jacobian matrix. Deriving the forward problem with respect to the jth component of the model parameter m_j yields

$$A(m)\frac{\partial u}{\partial_{m_j}} + A(m)\frac{\partial A}{\partial_{m_j}} u = 0. \tag{34}$$

Thus, $\alpha_j(m)$ is the solution of the forward problem

$$A(m)\alpha_j = -\partial_{m_j} A(m)u. \tag{35}$$

Using this formula, an exact computation of the entire Jacobian matrix $J(m)$ would thus require to solve m forward problems, which is intractable from a computational cost point of view. Nonetheless, a cheap approximation can be built by approximating the forward problem operator $A(m)$ as the identity matrix I in the left-hand side of equation (35). This leads to the definition of the pseudo-Hessian matrix entries:

$$\widetilde{H}_{ij}(m) = \left(\left[\partial_{m_i} A(m)u(m)\right]^T \left[\partial_{m_j} A(m)u(m)\right] \right), \quad i,j = 1,\ldots,M. \tag{36}$$

The preconditioner used by Shin *et al.* (2001) is defined by

$$P_k = diag\left(\frac{1}{\widetilde{H}_{ii}(m_k)} \right) \quad i = 1,\ldots,M. \tag{37}$$

However, because of the fast decrease of the wavefield with depth, very small values appear on the diagonal entries of $\widetilde{H}(m)$ corresponding to deep subsurface parameters. Therefore, using directly P_k as a preconditioner may yield numerical instabilities. We thus introduce a threshold parameter $\theta \in \mathbb{R}$, the constant $C_k \in \mathbb{R}$ such that

$$C_k = \max_j \widetilde{H}_{jj}(m_k), \tag{38}$$

and we define the matrix P_k^θ such that

$$P_k^\theta = diag\left(\frac{1}{\widetilde{H}_{ii}(m_k) + \theta C_k} \right), \quad i = 1,\ldots,M. \tag{39}$$

Finally the norm of the misfit gradient $\nabla f(m)$ should be preserved by the preconditioner*.

*Since the stopping criterion for the linear system (4) is based on the reduction of the linear residuals with respect to the norm of the gradient, it appears natural that the preconditioner conserves the gradient norm.

Therefore, we introduce

$$P_k^{\nu,\theta} = \nu P_k^{\theta}, \tag{40}$$

such that

$$\nu = \frac{\|\nabla f(m_k)\|}{\|P_k^{\theta} \nabla f(m_k)\|}. \tag{41}$$

We use $P_k^{\nu,\theta}$ as a preconditioner of the linear system (4) at each iteration of the external loop. The preconditioner $P_k^{\nu,\theta}$ can also be used as a preconditioner of the nonlinear conjugate gradient method and the l-BFGS method.

4 CASE STUDIES

4.1 Numerical framework

4.1.1 Forward problem

The numerical tests we present are performed in the 2D frequency domain. In the first case study, we use an isotropic acoustic approximation of the wave propagation. In the second case study, we consider the propagation of acoustic anisotropic waves in a VTI media . In both cases an optimized second-order finite differences scheme with a compact stencil is used (Hustedt *et al.* 2004; Operto *et al.* 2009). Perfectly Matched Layers (PML) (Berenger 1994; Métivier 2011) are introduced to avoid fictitious reflections on the boundaries of the computation domain, excepted on top, where a free surface condition is implemented.

The numerical resolution of the forward problem amounts to the resolution of a sparse linear system. This is performed through a parallel LU factorization using the MUMPS algorithm (Amestoy *et al.* 2000). The LU factorization of the impedance matrix associated with the discretization of the forward problem is reused to solve the adjoint problems, as MUMPS offers the possibility of solving the adjoint system once the factorization has been performed. This is especially important when the number of sources is large: the same LU factorization is used to solve the forward and adjoint problems associated with each source. This interesting feature is one of the reason for working in the frequency domain: provided the LU factorization of the impedance matrix can be stored, this approach largely reduces

the computational costs, compared to the time domain approach. In the particular case of the truncated Newton method, this also makes possible to use the factorization of the impedance matrix for the computation of the Hessian-vector products. The resolution of the extra forward and adjoint problems induced by the truncated Newton method can thus be performed without a new factorization of the impedance matrix.

4.1.2 Minimization scheme settings

In the following two tests an estimation of the P-wave velocity model is computed using a FWI scheme. For each of these two tests, we compare the performances of four minimization strategies:

- nonlinear conjugate gradient method;
- l-BFGS method;
- truncated Newton method using the full Hessian operator;
- truncated Newton method using the Gauss-Newton approximation;

In our implementation of the truncated Newton algorithm, we complement the Eisenstat stopping criterion by setting to 30 the maximum number of inner iterations that can be performed in the first case study, and only 3 in the second case study. These rather small values is chosen to enhance the smoothing effect related to the truncation strategy, which is appropriate for the inversion of noisy data.

The stopping criterion we use is the following: the iterations end as soon as

$$f(m_k)/f(p_0) < \epsilon. \tag{42}$$

The quantity ϵ is set to 10^{-2} for the first experiment. Such an accuracy can be reached since in this case we work with synthetic data without noise. In the second experiment, we use real noisy data, and ϵ is set to 6×10^{-1}, which is adapted to the expected decrease of the misfit function.

This stopping criterion is complemented with a maximum number of nonlinear iteration, which is set to 100 for the BP 2004 test case and 20 for the Valhall test case. In addition, if

no acceptable step length is found after 20 linesearch iterations, the minimization is stopped and an error flag is returned.

Finally, note that the memory parameter l for the two l-BFGS methods, which corresponds to the number of gradient stored to compute the approximation of the inverse Hessian, is set to $l = 20$ in the first case and $l = 5$ in the second case.

4.1.3 Linesearch algorithm and nonlinear conjugate gradient implementation

The linesearch algorithm we use is designed so that the step γ_k computed at each nonlinear iteration satisfies the Wolfe criterion (see equation (4)). Because we want the four minimization methods to use the same linesearch algorithm, we implement a particular form of the nonlinear conjugate gradient method. Indeed, as mentioned by Nocedal and Wright (2006), standard implementations of this algorithm (such as Fletcher-Reeves or Polak-Ribière implementations) require to use a linesearch algorithm satisfying the *strong* Wolfe conditions to guarantee global convergence toward local minima. We select instead the nonlinear conjugate gradient algorithm proposed by Dai and Yuan (1999), which is compatible with a linesearch process that only enforces the *standard* Wolfe conditions.

4.1.4 Preconditioner

The four minimization methods use the same preconditioner $P_k^{\nu,\theta}$. For the two case studies, a trial-and-error approach has been used to determine a common usable values for θ. Pragmatical values range from 10^{-1} to 10^{-5}. In practice, the value $\theta = 10^{-2}$ has been used along all the experiments, for all the methods.

4.2 The 2004 BP model

4.2.1 Presentation

The 2004 BP model has been originally designed as a benchmark model for testing sub-salt seismic imaging methods (Billette and Brandsberg-Dahl 2004). We perform a decimation of the original model taking one parameter value each ten grid points, and we choose a 25 m

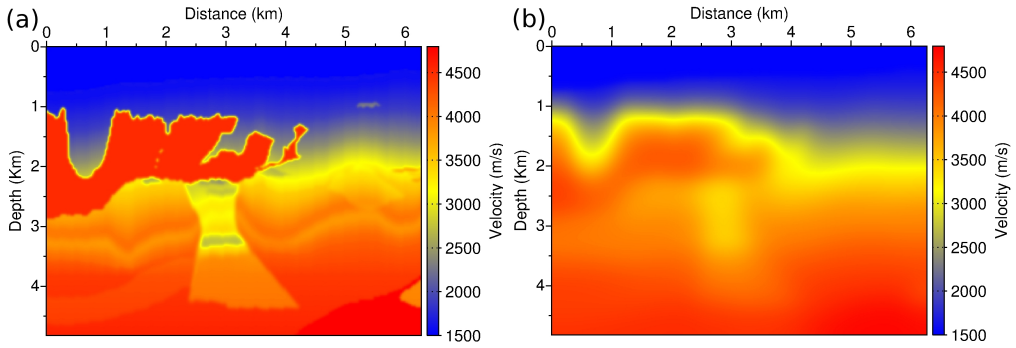

Figure 1. BP 2004 synthetic exact model (a), initial model (b).

discretization grid. We end up with a 6.2 km wide and 4.2 km deep reduced model, described by approximately 5×10^4 discrete parameters.

The resulting model is presented in figure 1. It presents a complex rugose salt body, and sub-salt slow velocity anomalies that represent over-pressured zones. This intends to mimic the geology that can be found in the Gulf of Mexico. The main challenges in this area are related to the definition of a precise delineation of the salt and recovering information on the sub-salt velocity variations. The P-wave velocity in the salt reaches 4790 m.s^{-1}, while it is equal to 1486 m.s^{-1} in the water. The discrepancy between these two values is responsible for high amplitude reflections. These energetic reflected waves are reflected back at the top of the water layer through the free surface condition. The proximity of these two reflectors generates multiple scattering.

We use a surface acquisition configuration with 62 sources and 248 receivers, from $x = 50$ m to $x = 6225$ m at 25 m below the sea-level. The spatial sampling of the receivers and the sources is set up to 25 m and 100 m, respectively. We use a free-surface condition at the top of the model, to account for the surface multiples. The water layer is kept constant throughout the iterations, so as to stabilize the problem. The bathymetry of the sea-bottom is respected. No regularization strategy is used here, as we consider synthetic data without additional noise.

The cycle of FWI starts with an initial model computed as a smooth version of the exact model using again a Gaussian smoothing. Compared to previous work (Métivier *et al.*

Group 1	2 Hz	2.25 Hz	2.5 Hz	2.75 Hz	
Group 2	2.5 Hz	3 Hz	3.5 Hz	4 Hz	
Group 3	4 Hz	4.5 Hz	5 Hz	5.5 Hz	
Group 4	5.5 Hz	6 Hz	6.5 Hz	7 Hz	
Group 5	7 Hz	7.5 Hz	8 Hz	8.5 Hz	
Group 6	8.5 Hz	9.5 Hz	10.5 Hz	11.5 Hz	
Group 7	11.5 Hz	12.5 Hz	13.5 Hz	14.5 Hz	15.5 Hz
Group 8	15.5 Hz	16.5 Hz	17.5 Hz	18.5 Hz	19.5 Hz

Table 1. Frequency group strategy for the BP 2004 case study

2012), we use here a smoother initial model. The characteristic length used for this smoothing is set to 500 m instead of 375 m for this former study. The resulting initial model is presented in figure 1. The use of this smoother initial model requires a careful hierarchical frequency strategy. We generate 27 data sets, from 2 Hz frequency to 19.5 Hz gathered into 8 overlapping subgroups, as presented in table 1.

4.2.2 Estimated models

The models estimated by the four minimization methods are presented in figure 2. From km 1 to km 6 in the horizontal direction, the top salt-structure is correctly delineated in the four estimations. The reconstruction of the basin between $x = 0$ km and $x = 1$ km seems more difficult. This basin is responsible for high amplitude multi-scattered waves difficult to interpret, and is located at one extremity of the acquisition. The nonlinear conjugate gradient method and the l-BFGS method seem to be the most affected by this particular configuration. The geometry of the basin is not recovered and is filled with high amplitude velocities. These perturbations are also responsible for obscuring the sub-salt targets, and creating erroneous slow velocity anomalies.

Conversely, the results provided by the truncated Gauss-Newton or the truncated Newton method seem more reliable. The best estimation is provided by the truncated Newton method. The geometry of the basin is better recovered, and the sub-salt slow velocity anomalies better reconstructed. Note also that only these two methods are able to provide details on

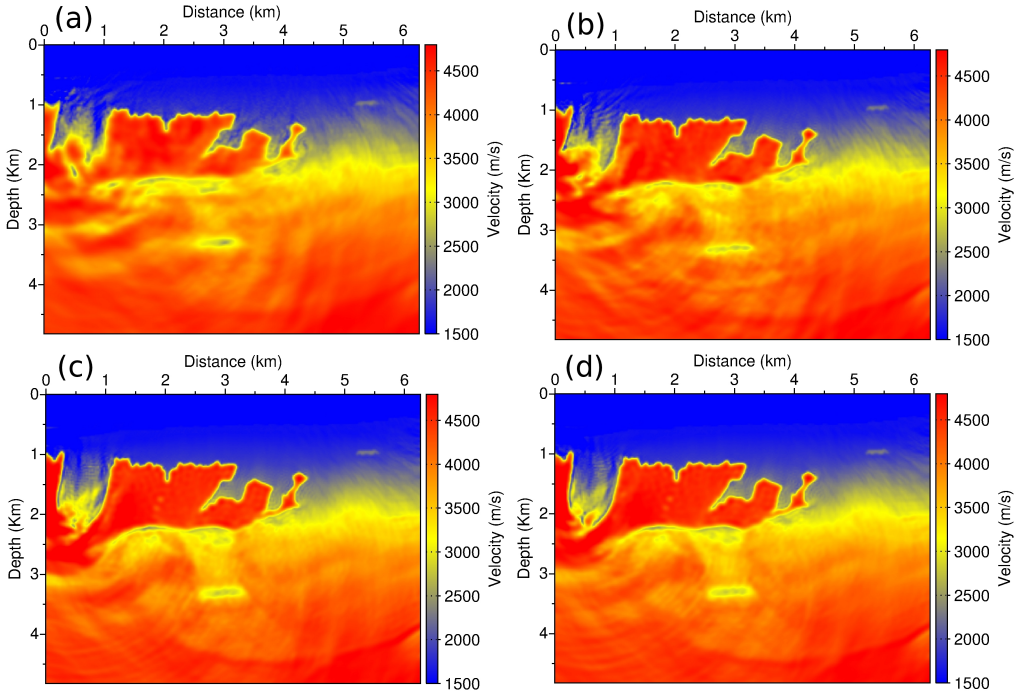

Figure 2. Estimated models for the BP case study. Nonlinear conjugate gradient (a), l-BFGS method (b), truncated Gauss-Newton method (c), truncated Newton method (d).

the salt structure itself, namely the very small heterogeneities located at $x = 2km, z = 2km$. The possible enhancement of the inverse Hessian approximation yielded by the truncated Newton method may explain this improvement in the resolution and the stability of the inversion.

4.2.3 Convergence profiles

The convergence profiles of the four methods are presented for the frequency group 1, 4 and 8 in figure 3. Excepted for the frequency group 1 for which the l-BFGS method converges slightly faster, the nonlinear conjugate gradient and the l-BFGS method converge slower than the two truncated Newton methods in terms of nonlinear iterations. This is satisfactory since each nonlinear iteration is more expensive for the truncated Newton method. In addition, only the two truncated Newton method satisfies the convergence criterion before reaching the maximum number of nonlinear iteration, set to 100.

In terms of number of forward problems resolution, the slowest method is the truncated Newton method. This indicates that the speed-up in terms of nonlinear iteration does not

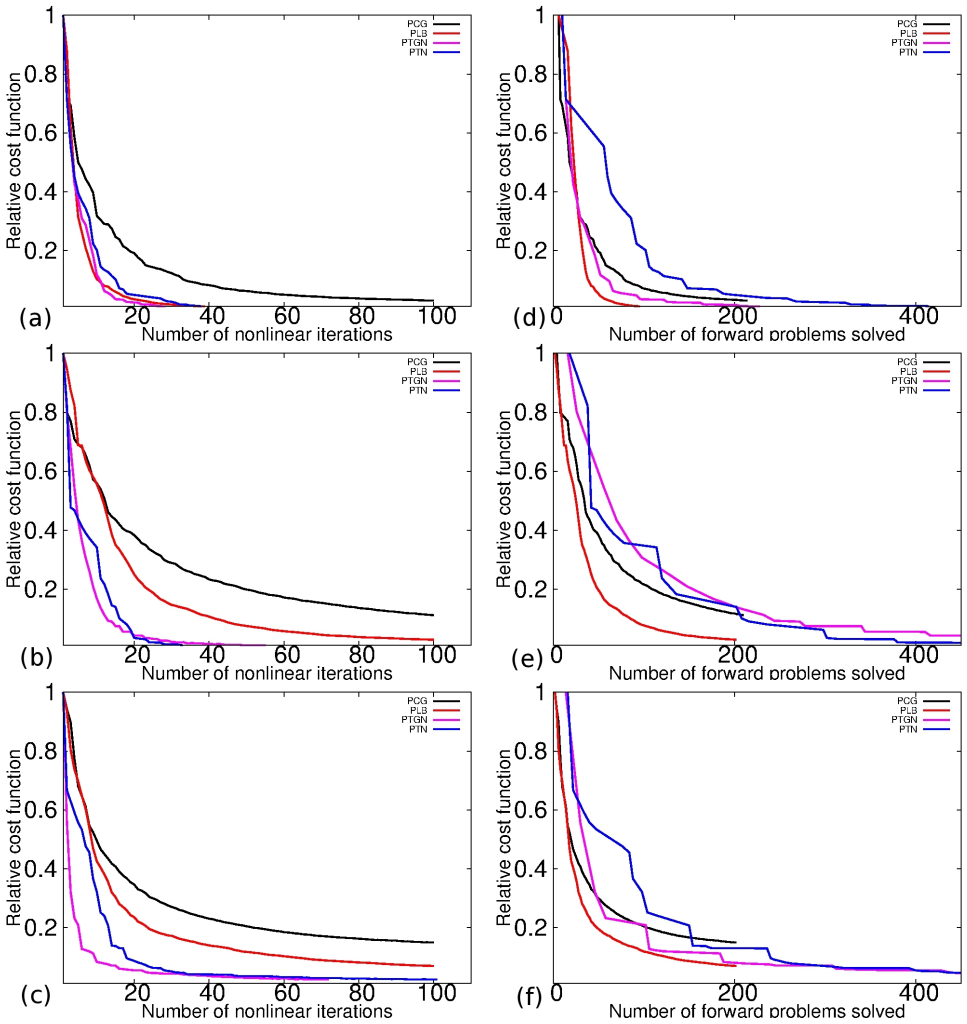

Figure 3. Misfit function decrease for the BP 2004 case study. Decrease with respect to the the number of nonlinear iterations: first frequency group (a), second frequency group (b), third frequency group (c). Decrease with respect to the number of wave equation problem solved: first frequency group (d), second frequency group (e), third frequency group (f). PCG: nonlinear conjugate gradient, PLB: l-BFGS, PTN: truncated Newton, PTGN: truncated Gauss-Newton method.

compensate the extra computation cost required. However, as suggested by the profile aspect, reducing the total number of authorized internal iterations could possibly improve the convergence in terms on forward problem resolutions. Indeed, we can detect plates which correspond to non-converging inner iterations, which reach the maximum allowed number of iterations, set to 30 for this experiment. In each case, the fastest method in terms of forward problem resolution is the l-BFGS method. The truncated Gauss-Newton method and the nonlinear conjugate gradient are in between.

The speed-up obtained in terms of convergence with respect to the number of nonlinear iterations does not compensate for the additional cost related to the resolution of the inner linear systems. However, compared to preconditioned gradient-based method, the truncated Newton method appears to provide better subsurface estimations.

4.3 2D Valhall case study

4.3.1 Description

The Valhall oil field is located in the North Sea, in production since the beginning of the 80s. In this shallow water environment field, the water depth reaches only 100 m. This particular configuration is adapted for the use of ocean bottom cables (OBC) seismic recorders. Four components sensors have thus been disposed at the sea bottom level. Several three-dimensional seismic surveys have been performed to follow the time-evolution of the oil field. The Valhall seismic data have been investigated by several authors in the context of 2D multi-parameter FWI (Prieux *et al.* 2011, 2013a,b) and 3D mono-parameter FWI(Sirgue *et al.* 2010; Etienne *et al.* 2012).

In this study, we are interested in the mono-parameter inversion of a 2D line from the Valhall OBC data-set. This data-set involves 320 sources located at 5 meters depth in the water layer, and 210 receivers located on the sea bottom, between 68 and 72 meter depth. Sources and receivers are equally spaced each 50 m. Only the hydrophone component is used. The data is interpreted down to 3.5 Hz. The signal-over-noise ratio (SNR) for lower frequencies is too weak for the data to be exploited in this frequency band.

Below the sea level, the presence of soft shale sediments is responsible for strong attenuation and anisotropy of the wave propagation. As demonstrated in the work of Prieux *et al.* (2011), it is necessary to account for this anisotropy in the modeling to correctly invert the Valhall data-set. For this case study, we thus change the forward problem from standard isotropic acoustic modeling to the VTI acoustic modeling proposed by Operto *et al.* (2009).

From 3.5 Hz to 7.45 Hz, we define 6 overlapping frequency groups to be inverted sequentially

Group 1	3.54 Hz	3.78 Hz	4.03 Hz
Group 2	4.03 Hz	4.28 Hz	4.64 Hz
Group 3	4.64 Hz	5.00 Hz	5.25 Hz
Group 4	5.25 Hz	5.62 Hz	5.99 Hz
Group 5	5.99 Hz	6.35 Hz	6.72 Hz

Table 2. Frequency group strategy for the Valhall case study

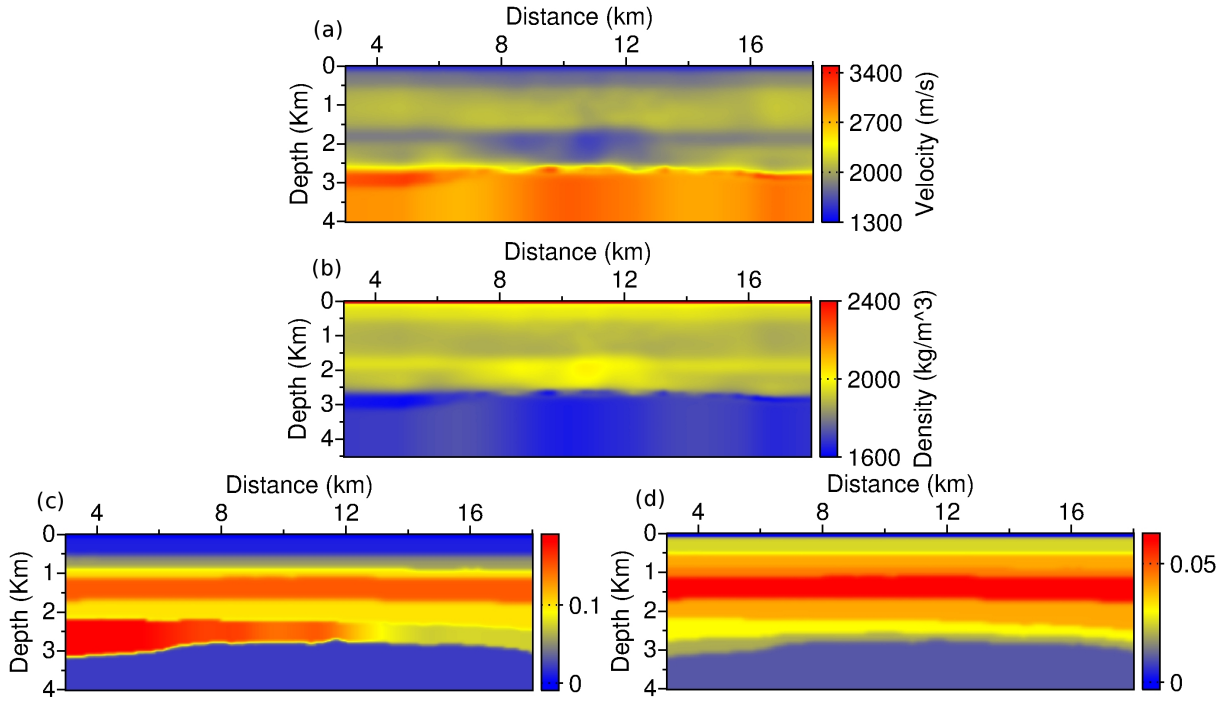

Figure 4. Initial models for P-wave velocity (a), density (b). Dimensionless Thomsen parameters ϵ (c), δ (d)

The VTI forward modeling requires to define initial models not only for P-wave velocity, density and attenuation, but also for the Thomsen anisotropy parameters ϵ and δ. The initial models for P-wave velocity and Thomsen parameters are determined by reflection travel-time tomography. The density initial model is derived from the P-wave velocity initial model through the Gardner law. The quality factor model is taken constant equal to 200. The corresponding models are presented in figure 4. In our experiments, the initial models for density, attenuation and Thomsen parameters remain the same, and we only focus on the reconstruction of the P-wave velocity model.

4.3.2 Regularization strategy

As the real seismic data is noise contaminated, an appropriate regularization strategy has to be designed. We choose to add a standard Tikhonov regularization term $T(m)$ to the misfit function $f(m)$ to define the regularized misfit function $f_T(m)$

$$f_T(m) = f(m) + \frac{\alpha}{2} T(m) \tag{43}$$

where

$$T(m) = \left(\alpha_x \|\partial_x m\|^2 + \alpha_z \|\partial_z m\|^2\right) \tag{44}$$

The parameter α accounts for the influence of the regularization term. The parameters α_x and α_z can be used to enforce a stronger regularization in the direction x or z. In the sequel, we use the settings

$$\alpha_x = 1, \quad \alpha_z = 0.5 \tag{45}$$

to enforce smoother variations in the horizontal direction.

In addition, the truncation strategy for the computation of the descent direction within the truncated Newton/Gauss-Newton methods is used as an additional regularization (see the work of Kaltenbacher *et al.* (2008) on this particular topic). The maximum number of inner conjugate gradient iteration is set to 3.

In order to investigate the sensitivity of the methods to the regularization parameter α, we have performed experiments for 2 different values of α, reflecting what we may call a "strong" α_1 and a "weak" α_2 regularization.

- For nonlinear conjugate gradient and l-BFGS, we use $\alpha_1 = 10^{-4}$ and $\alpha_2 = 5 \times 10^{-4}$.
- For the truncated Newton/Gauss-Newton method, we use $\alpha_1 = 10^{-6}$ and $\alpha_2 = 5 \times 10^{-6}$.

The combination of two types of regularization for the truncated Newton/Gauss-Newton strategies makes possible to decrease the value of the parameter α by two orders of magnitude for these methods. This reduces the smoothing effect yielded by the Tikhonov regularization term, and possibly the resolution loss associated with this smoothing.

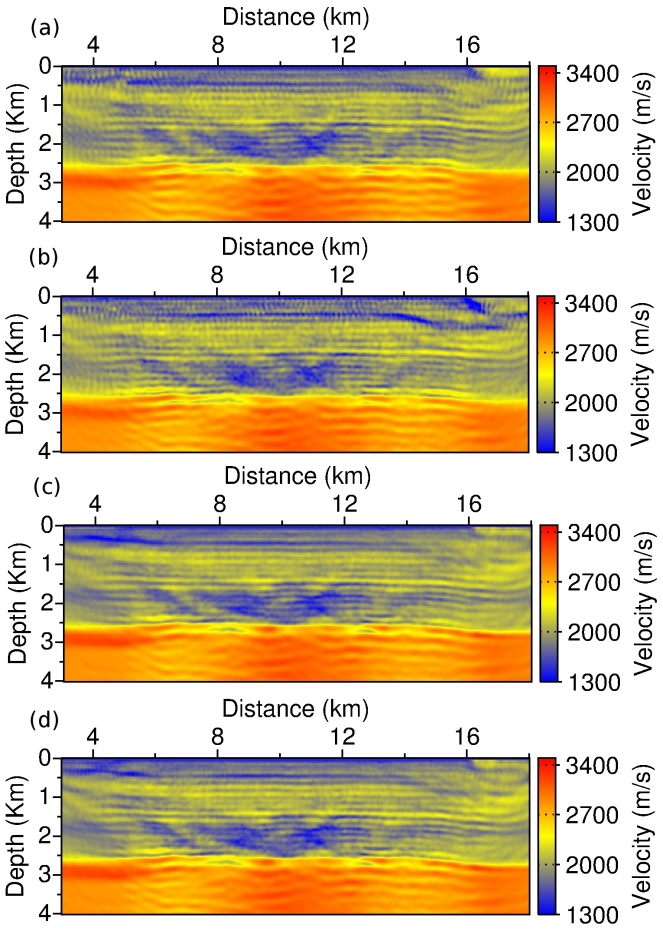

Figure 5. Weak regularization: computed P-wave velocity models using the nonlinear conjugate gradient (a), l-BFGS (b), truncated Gauss-Newton (c), truncated Newton (d).

4.3.3 FWI results

We presents the results obtained by our four optimization schemes in figures 5 and 6. Below shallow layers of shale sediments (from $z = 0$ to $z = 1.5$ km), we can see two low velocity zones, corresponding to the presence of gas layers (between $z = 1.5$ km and $z = 2.5$ km, $x = 8$ km and $x = 12$ km). Below these gas layers (between $z = 2.5$ km and $z = 4$ km), we can locate the cap rock of the reservoir and stronger reflectors.

The results we obtain demonstrate the difficulty of finding a suitable trade-off between regularization and resolution power. For nonlinear conjugate gradient and the l-BFGS method, changing from weak to strong regularization yields significant differences in the results. For weak regularization, the final results is contaminated by noise. For strong regular-

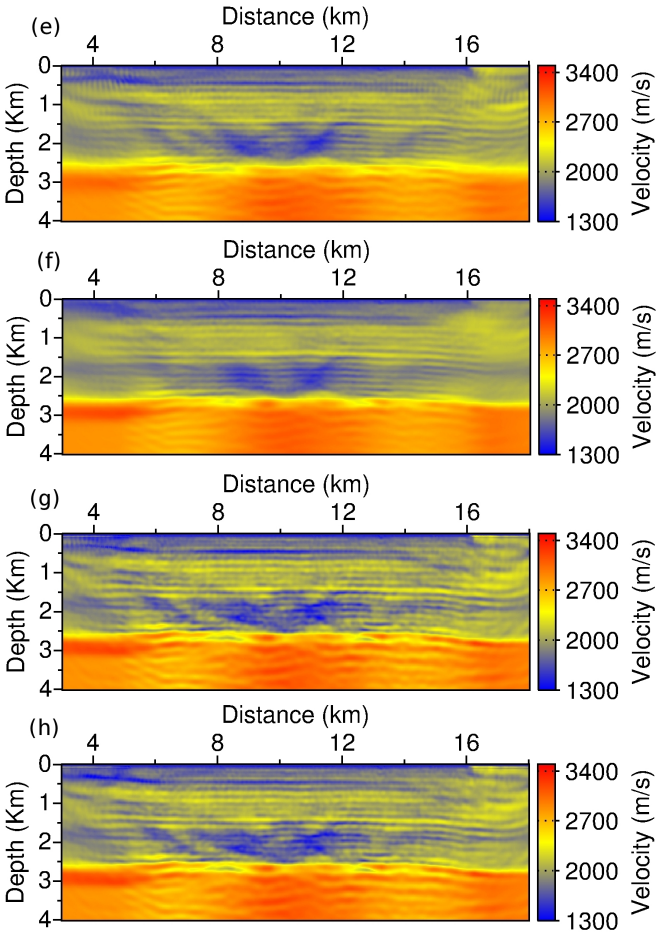

Figure 6. Strong regularization: computed P-wave velocity models using the nonlinear conjugate gradient (e), l-BFGS (f), truncated Gauss-Newton (g), truncated Newton (h).

ization, the results appears significantly smoother, to the cost of a potential loss of resolution in depth.

Conversely, the truncated Newton method seems to be more stable with respect to a modification of the regularization parameter. The two results obtained for weak and strong regularization do not show strong differences. In addition, even if the regularization parameter used is two order of magnitude smaller than the one used for the two latter methods, the final results is less affected by noise. This tends to confirm the regularization effect associated with the truncation strategy.

This analysis is confirmed by the well-logs presented in figures 7 and 8. The results from the truncated Newton methods provide a globally better fit to the well-log data with strong or weak regularization. We are particularly interested in the jump in depth at $z = 3.5$ km

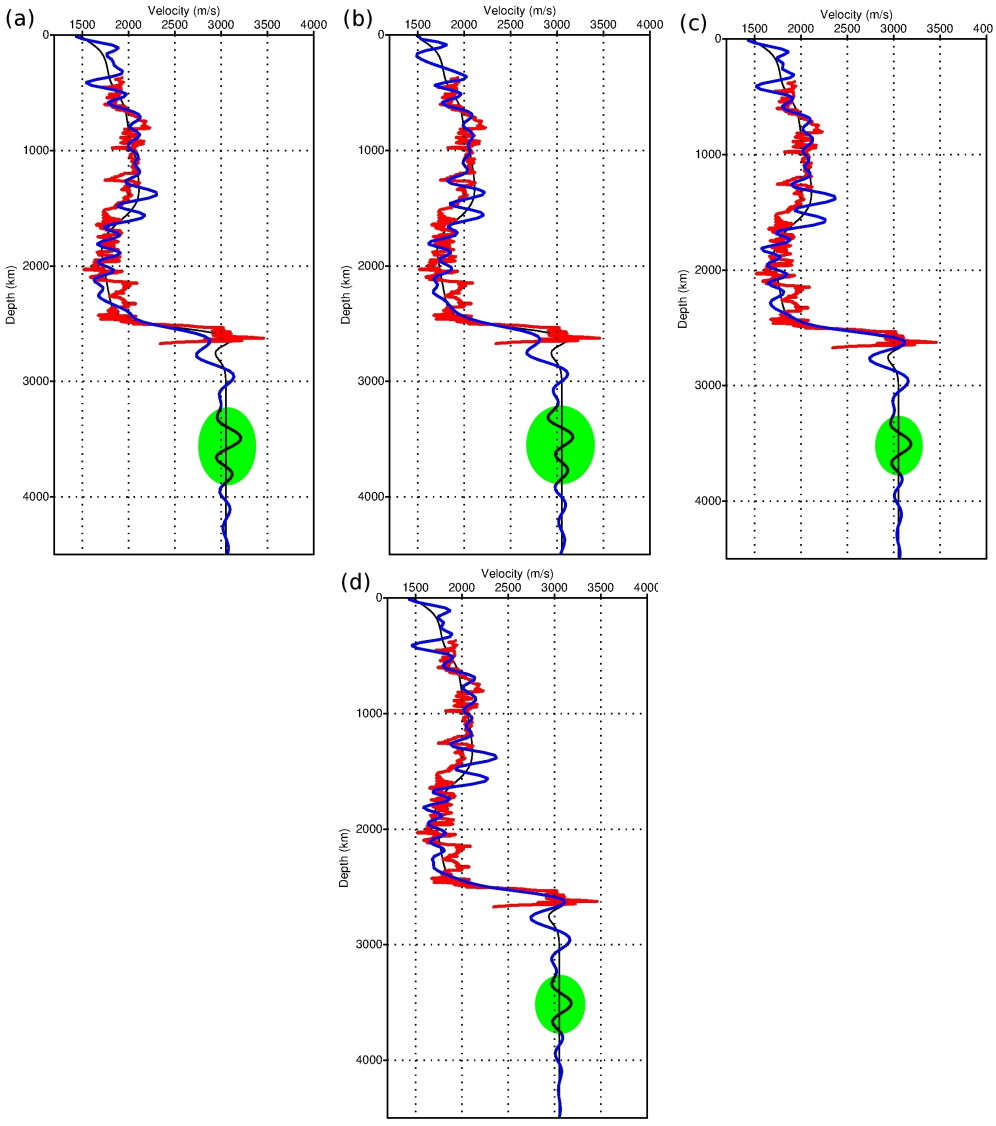

Figure 7. P-wave velocity logs compared to a reference well-log at $x = 9.5$ km,), with *weak* regularization. Results obtained using the nonlinear conjugate gradient (a), l-BFGS (b), truncated Gauss-Newton (c), truncated Newton (d)

(green circles), which can be quite clearly identified in the results provided by these two methods. This jump correspond to the presence of an actual reflector (see the migrated section using the initial velocity model 9). For the preconditioned gradient-based methods, the results obtained with a weak regularization present strong oscillations in depth which make difficult to identify this reflector. The results obtained with a strong regularization smooth out the results, and make equivalently difficult the identification of the reflector.

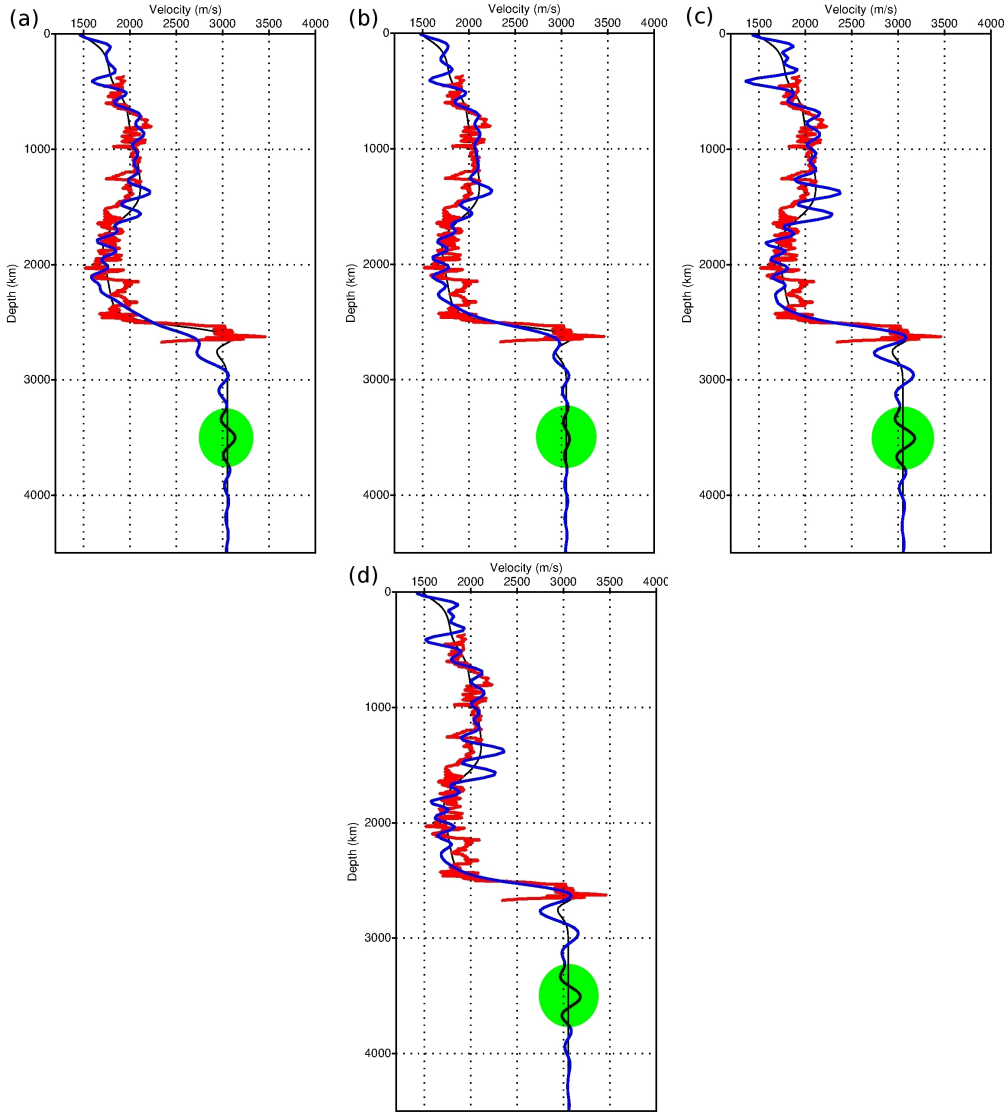

Figure 8. P-wave velocity logs compared to a reference well-log at $x = 9.5$ km,), with *strong* regularization. Results obtained using the nonlinear conjugate gradient (a), l-BFGS (b), truncated Gauss-Newton (c), truncated Newton (d)

4.3.4 Resolution analysis

The analysis we have led suggests that a better trade-off between regularization and resolution power can be obtained using the truncated Newton method. In order to investigate this particular point we compare the sensitivity of the preconditioned gradient-based methods and the truncated Newton method to a perturbation in the final estimations.

We add a positive perturbation of 200 m.s^{-1} in the gas cloud localized around $x = 11$ km, $z = 2$ km. We compute synthetic data using the same surface acquisition geometry in

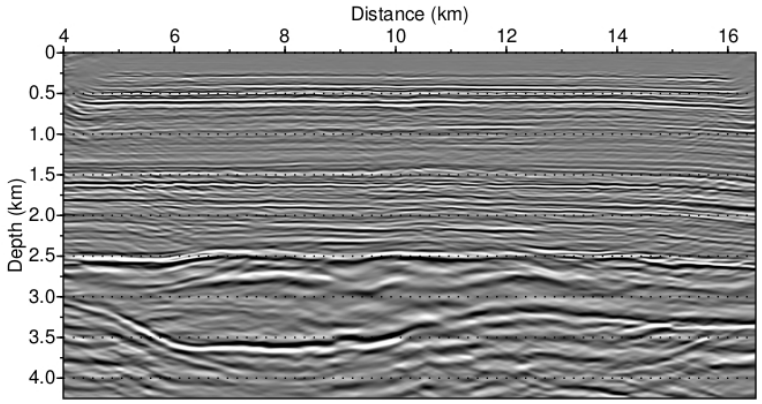

Figure 9. Reverse Time Migration result using the starting velocity model.

this perturbed model. We invert this synthetic data starting from the unperturbed model, and we perform only one nonlinear iteration. In this case, the nonlinear conjugate gradient and the l-BFGS method are equivalent: at the first iteration, the l-BFGS approximation is just given by the preconditioning matrix.

We compare the reconstruction of the perturbation for the nonlinear conjugate gradient model and for the truncated Newton model. For the truncated Newton experiment, we compute the perturbation using

- 3 inner linear iterations
- 7 inner linear iterations

The results are presented in figure 10. Not surprisingly, the focusing achieved by the nonlinear conjugate gradient method is poorer than the one obtained using the truncated Newton method. The amplitude of the perturbation reconstructed by the nonlinear conjugate gradient only reaches 10 m.s^{-1} while it reaches more than 30 m.s^{-1} for the truncated Newton method. More interestingly, we can see the refocusing effect associated with the inverse Hessian operator on the model update obtained after 7 inner linear iterations. The amplitude of the reconstructed perturbation is similar, but the artifacts around the perturbation tends to vanish.

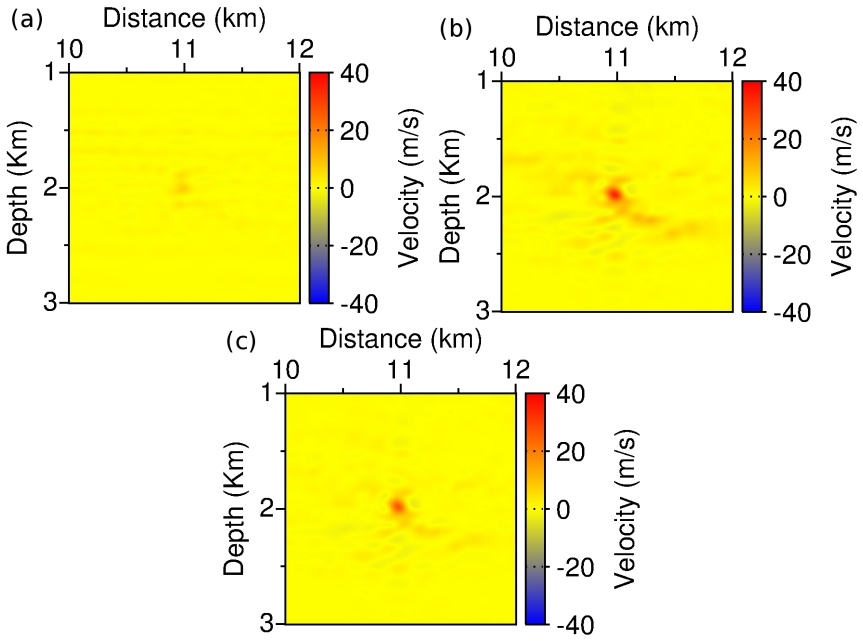

Figure 10. Reconstruction of a small amplitude perturbation in the final estimation. Zoom on the perturbation. Nonlinear conjugate gradient (a), truncated Newton method with 3 inner linear iterations (b), 7 inner linear iterations (c)

4.3.5 Convergence profiles

The convergence profiles for the Valhall case study are displayed in figure 11. They correspond to the inversion of the frequency groups 1, 3, and 5 in the *strong* regularization case.

Among the four methods, the l-BFGS method seems to be the less robust, as it fails to converge for the three frequency groups displayed. It means that the method terminates on a linesearch failure: the minimization of the misfit function does not reach the bound fixed at 0.6 or the maximum number of nonlinear iteration fixed at 20.

Compared to l-BFGS, the nonlinear conjugate gradient method seems to be more robust. For the frequency group 1 and 3, the maximum number of nonlinear iterations is reached and the method provides a better decrease of the misfit function.

The truncated Newton/Gauss-Newton methods systematically reach the maximum number of authorized nonlinear iterations and provide a better decrease of the misfit function

at each frequency group. However, the speed-up in terms of nonlinear iteration compared to the gradient-based methods is limited, excepted for the first frequency group.

In terms of number of resolution of wave propagation problems, the truncated Newton/Gauss-Newton methods appear not surprisingly as more expensive as their gradient-based counterpart. However, this extra computation cost is controlled by the maximum number of inner iteration, which is limited to 3. Note that for this case study, the results provided by the truncated Newton and the truncated Gauss-Newton are very similar (for the first frequency group, the convergence curves of these two methods are even superposed).

5 CONCLUSION AND PERSPECTIVES

5.1 Conclusion

In this study, we are interested in the implementation of the truncated Newton minimization scheme for FWI. The algorithm is based on a two nested loops architecture: the external loop consists in updating an initial model up to the final estimation. The inner loop consists in the computation of the model updates from an inexact resolution of the linear system associated with the computation of the Newton descent direction. This inexact resolution is performed with a matrix-free conjugate gradient solver. Only the action of the Hessian operator on an arbitrary vector has to be computed. This can be efficiently achieved using second-order adjoint state formulas: only two additional wave propagation problems have to be solved to compute this quantity. The use of an adaptive stopping criterion for the inner loop and a suitable preconditioner enhance the efficiency of the algorithm.

Compared to conventional preconditioned gradient-based methods operator such as the nonlinear conjugate gradient or the l-BFGS method, it seems that using the truncated Newton method could be advantageous. The synthetic BP 2004 case study emphasizes the difficulty of salt and sub-salt imaging. The high velocity contrasts caused by the presence of large salt structures is responsible for the presence of energetic multiple reflections in the data. This cause difficulties to standard minimization methods to provide correct images of the salt and sub-salt targets. In this case, the truncated Newton method seems to be

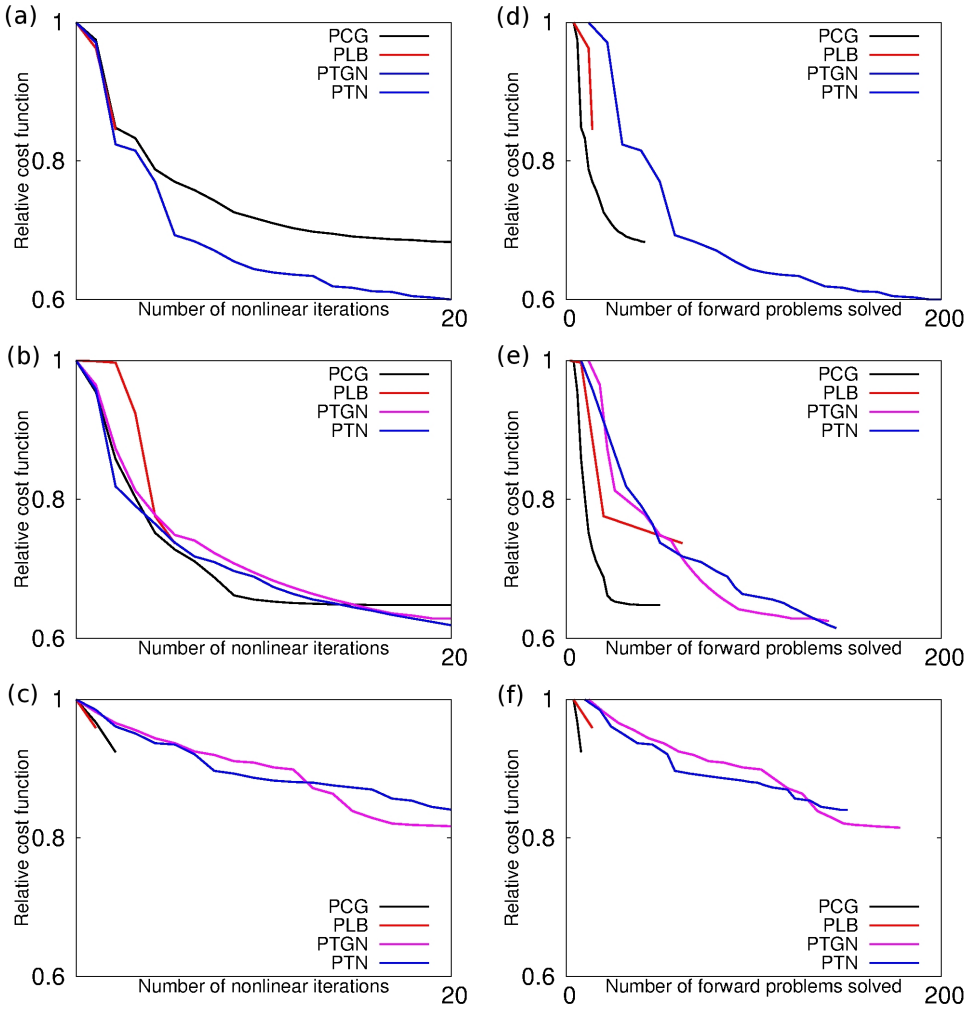

Figure 11. Misfit function decrease for the Valhall case study for the *strong* regularization case. Decrease with respect to the number of nonlinear iterations: frequency group 1 (a), frequency group 3 (b), and frequency group 5 (c). Decrease with respect to the number of wave equation problems solved: frequency group 1 (d), frequency group 3 (e), and frequency group 5 (f). PCG: nonlinear conjugate gradient, PLB: l-BFGS, PTGN: truncated Gauss-Newton method, PTN: truncated Newton.

more robust than the nonlinear conjugate gradient and the l-BFGS method. In particular, the truncated Newton method provide better results than the truncated Gauss-Newton method. A possible explanation is that the part of the Hessian operator which is neglected in the Gauss-Newton approximation is related to double scattered waves. In this context of multiple-scattering between the salt and the free surface, it could be crucial to account for the whole Hessian operator.

The Valhall real data case study also demonstrates the potential interest of using the truncated Newton method compared to the conventional methods. The noise-contamination of the data requires to use a suitable regularization. We use a standard Tikhonov regularization of the first order derivatives of the model. We identify the trade-off between regularization and the expected resolution of the final estimation. The nonlinear conjugate gradient and the l-BFGS method are highly sensitive to the regularization hyper-parameter which controls the amount of regularization. Conversely, the truncated Newton method possesses an inherent smoothing related to the truncation strategy for computing the model update (Kaltenbacher *et al.* 2008). This yields the possibility of choosing a regularization hyper-parameter 2 orders of magnitude lower than for the preconditioned gradient-based methods. It also reduces the sensitivity to this hyper-parameter. A better trade-off between resolution and regularization is thus achieved. A point-spread function test in the final estimated models clearly indicates that the information on the local curvature carried out by the inverse Hessian operator yields a better resolution.

5.2 Perspectives

The numerical performances of the different minimization schemes we have compared strongly depend on an accurate estimation of the inverse Hessian operator as a preconditioner. For now, we have only used the diagonal preconditioner from the pseudo-Hessian approach developed by Shin *et al.* (2001). The accuracy of this approximation is however limited and the use of this preconditioner is restricted to surface acquisition configurations. Therefore we are interested in the development of better approximation of the inverse Hessian. Several techniques could be used.

For instance, Chiu and Demanet (2012) have proposed approximation of the inverse Hessian operator through matrix probing. However, this method only approximates the Gauss-Newton part of the operator, and relies on the acoustic approximation. Bekas *et al.* (2007) propose a general algebraic method to compute approximation of the diagonal of a given matrix in a matrix-free fashion. The explicit construction of the entire matrix is

not required. This makes this method adapted to FWI, as opposed with incomplete LU or Cholesky factorization (Benzi 2002).

Another way of improving the resolution of the inner linear systems is related to deflation strategies (Saad 2003). From the Krylov subspace created during the resolution of the inner linear system at iteration $k-1$, an estimation of the smallest eigenvectors of the Hessian operator can be computed. These eigenvectors can be inserted in the starting Krylov suspace at iteration k in order to improve the convergence. Related methods developed within the data assimilation community could also be investigated. Gratton *et al.* (2011) propose for instance to compute an approximate initial guess for the inner linear systems in a reduced space given by a spectral decomposition of the initial Hessian operator.

The importance of an accurate estimation of the inverse Hessian operator should be even more important in the context of multi-parameter FWI. The simultaneous reconstruction of parameters such as the P-wave velocity and the density for instance is affected by strong trade-offs. The perturbation in the P-wave models can be erroneously interpreted as perturbations in the density model, and *vice-versa*. A more accurate estimation of the inverse Hessian operator should help to mitigate these trade-off effects. In this context, the truncated Newton method could thus produce more reliable results. This topic will be investigated in future studies.

ACKNOWLEDGMENTS

This study was funded by the SEISCOPE II consortium ($http://seiscope2.osug.fr$), sponsored by BP, CGG-VERITAS, CHEVRON, EXXON-MOBIL, JGI, SAUDI ARAMCO, SHELL, STATOIL and TOTAL. This study was granted access to the HPC facilities of CIMENT (Université Joseph Fourier Grenoble), and of GENCI-CINES under Grant 2011-046091 of GENCI (Grand Equipement National de Calcul Intensif). The authors would like to thank V. Tcheverda for his invitation to publish this study. They also thank BP Norge AS and their Valhall partner Hess Norge AS, for allowing access to the Valhall dataset as well as the well-log velocities.

REFERENCES

Amestoy, P., Duff, I. S., and L'Excellent, J. Y., 2000. Multifrontal parallel distributed symmetric and unsymmetric solvers, *Computer Methods in Applied Mechanics and Engineering*, **184**(2-4), 501–520.

Bekas, C., Kokiopoulou, E., and Saad, Y., 2007. An estimator for the diagonal of a matrix, *Appl. Numer. Math.*, **57**(11-12), 1214–1229.

Benzi, M., 2002. Preconditioning techniques for large linear systems: A survey, *Journal of Computational Physics*, **182**, 418–477.

Berenger, J-P, 1994. A perfectly matched layer for absorption of electromagnetic waves, *Journal of Computational Physics*, **114**, 185–200.

Billette, F. J. and Brandsberg-Dahl, S., 2004. The 2004 BP velocity benchmark, in *Extended Abstracts, 67th Annual EAGE Conference & Exhibition, Madrid, Spain*, p. B035.

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, C. A., 2006. *Numerical Optimization, Theoretical and Practical Aspects*, Springer series, Universitext.

Brossier, R., Operto, S., and Virieux, J., 2009. Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion, *Geophysics*, **74**(6), WCC105–WCC118.

Brossier, Romain, Operto, Stéphane, and Virieux, Jean, 2010. Which data residual norm for robust elastic frequency-domain full waveform inversion?, *Geophysics*, **75**(3), R37–R46.

Byrd, R.H., Lu, P., and Nocedal, J., 1995. A limited memory algorithm for bound constrained optimization, *SIAM Journal on Scientific and Statistical Computing*, **16**, 1190–1208.

Chavent, G., 1974. Identification of parameter distributed systems, in *Identification of function parameters in partial differential equations*, edited by R. Goodson and M. Polis, pp. 31–48, American Society of Mechanical Engineers, New York.

Chiu, J. and Demanet, L., 2012. Matrix probing and its conditioning, *SIAM Journal on Numerical Analysis*, **50(1)**, 171–193.

Dai, Y. and Yuan, Y., 1999. A nonlinear conjugate gradient method with a strong global convergence property, *SIAM Journal on Optimization*, **10**, 177–182.

Eisenstat, S. C. and Walker, H. F., 1994. Choosing the forcing terms in an inexact Newton method, *SIAM Journal on Scientific Computing*, **17**, 16–32.

Epanomeritakis, I., Akçelik, V., Ghattas, O., and Bielak, J., 2008. A Newton-CG method for large-scale three-dimensional elastic full waveform seismic inversion, *Inverse Problems*, **24**, 1–26.

Etienne, V., Hu, G., Operto, S., Virieux, J., Barkved, O.I., and Kommedal, J., 2012. Three-dimensional acoustic full waveform inversion: algorithm and application to Valhall, in *Expanded Abstracts, 74th Annual EAGE Conference & Exhibition, Copenhagen*, EAGE.

Fichtner, A. and Trampert, J., 2011. Hessian kernels of seismic data functionals based upon adjoint techniques, *Geophysical Journal International*, **185**(2), 775–798.

Gao, F., Levander, A. R., Pratt, R. G., Zelt, C. A., and Fradelizio, G. L., 2006. Waveform tomography at a groundwater contamination site: surface reflection data, *Geophysics*, **72(5)**, G45–G55.

Gauthier, O., Virieux, J., and Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: numerical results, *Geophysics*, **51**(7), 1387–1403.

Gratton, S., Laloyaux, P., Sartenaer, A., and Tshimanga, J., 2011. A reduced and limited-memory preconditioned approach for the 4d-var data-assimilation problem, *Q.J.R. Meteorol. Soc.*, **137**, 452–466.

Hustedt, B., Operto, S., and Virieux, J., 2004. Mixed-grid and staggered-grid finite difference methods for frequency domain acoustic wave modelling, *Geophysical Journal International*, **157**, 1269–1296.

Kaltenbacher, B., Neubauer, A., and Scherzer, O., 2008. *Iterative Regularization Methods for Nonlinear Problems*, de Gruyter, Berlin, New York.

Lailly, P., 1983. The seismic inverse problem as a sequence of before stack migrations, in *Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia*, edited by R. Bednar and Weglein, pp. 206–220.

Lions, J. L., 1968. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris.

Métivier, L., 2011. Interlocked optimization and fast gradient algorithm for a seismic inverse problem, *Journal of Computational Physics*, **230**(19), 7502–7518.

Métivier, L., Brossier, R., Virieux, J., and Operto, S., 2012. Optimization schemes in FWI: the truncated Newton method, in *SEG*, p. This issue.

Métivier, L., Brossier, R., Virieux, J., and Operto, S., 2013. Full waveform inversion and the truncated newton method, *SIAM Journal On Scientific Computing*, **35(2)**, B401–B437.

Nash, S. G., 2000. A survey of truncated Newton methods, *Journal of Computational and Applied Mathematics*, **124**, 45–59.

Nocedal, J. and Wright, S. J., 2006. *Numerical Optimization*, Springer, 2nd edn.

Operto, S., Ravaut, C., Improta, L., Virieux, J., Herrero, A., and Dell'Aversana, P., 2004. Quantitative imaging of complex structures from dense wide-aperture seismic data by multiscale traveltime and waveform inversions: a case study, *Geophysical Prospecting*, **52**, 625–651.

Operto, S., Virieux, J., and Dessa, J. X., 2005. High-resolution crustal seismic imaging from OBS data by full-waveform inversion: application to the eastern-Nankai trough, in *EOS Trans. AGU*, vol. 86, American Geophysical Union.

Operto, S., Virieux, J., Dessa, J. X., and Pascal, G., 2006. Crustal imaging from multi-

fold ocean bottom seismometers data by frequency-domain full-waveform tomography: application to the eastern Nankai trough, *Journal of Geophysical Research*, **111**(B09306), doi:10.1029/2005JB003835.

Operto, S., Virieux, J., Ribodetti, A., and Anderson, J. E., 2009. Finite-difference frequency-domain modeling of visco-acoustic wave propagation in two-dimensional TTI media, *Geophysics*, **74 (5)**, T75–T95.

Plessix, R. E., 2006. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications, *Geophysical Journal International*, **167**(2), 495–503.

Plessix, R. E. and Perkins, C., 2010. Full waveform inversion of a deep water ocean bottom seismometer dataset, *First Break*, **28**, 71–78.

Plessix, R.-E., Baeten, G., de Maag, J. Willem, and ten Kroode, F., 2012. Full waveform inversion and distance separated simultaneous sweeping: a study with a land seismic data set, *Geophysical Prospecting*, **60**, 733 – 747.

Pratt, R. G., 1990. Inverse theory applied to multi-source cross-hole tomography. part II : elastic wave-equation method, *Geophysical Prospecting*, **38**, 311–330.

Pratt, R. G. and Worthington, M. H., 1990. Inverse theory applied to multi-source cross-hole tomography. Part I: acoustic wave-equation method, *Geophysical Prospecting*, **38**, 287–310.

Pratt, R. G., Shin, C., and Hicks, G. J., 1998. Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion, *Geophysical Journal International*, **133**, 341–362.

Prieux, V., Brossier, R., Gholami, Y., Operto, S., Virieux, J., Barkved, O.I., and Kommedal, J.H., 2011. On the footprint of anisotropy on isotropic full waveform inversion: the Valhall case study, *Geophysical Journal International*, **187**, 1495–1515.

Prieux, V., Brossier, R., Operto, S., and Virieux, J., 2013. Multiparameter full waveform inversion of multicomponent OBC data from valhall. Part 1: imaging compressional wavespeed, density and attenuation, *Geophysical Journal International*, **doi: 10.1093/gji/ggt177**.

Prieux, V., Brossier, R., Operto, S., and Virieux, J., 2013. Multiparameter full waveform inversion of multicomponent OBC data from valhall. Part 2: imaging compressional and shear-wave velocities, *Geophysical Journal International*, **doi: 10.1093/gji/ggt178**.

Ravaut, C., Operto, S., Improta, L., Virieux, J., Herrero, A., and dell'Aversana, P., 2004. Multi-scale imaging of complex structures from multi-fold wide-aperture seismic data by frequency-domain full-wavefield inversions: application to a thrust belt, *Geophysical Journal International*, **159**, 1032–1056.

Saad, Y., 2003. *Iterative methods for sparse linear systems*, SIAM, Philadelphia.

Shin, C., Jang, S., and Min, D. J., 2001. Improved amplitude preservation for prestack depth migration by inverse scattering theory, *Geophysical Prospecting*, **49**, 592–606.

Sirgue, L., Etgen, J. T., and Albertin, U., 2008. 3D Frequency Domain Waveform Inversion using Time Domain Finite Difference Methods, in *Proceedings 70th EAGE, Conference and Exhibition, Roma, Italy*, p. F022.

Sirgue, L., Barkved, O. I., Dellinger, J., Etgen, J., Albertin, U., and Kommedal, J. H., 2010. Full waveform inversion: the next leap forward in imaging at Valhall, *First Break*, **28**, 65–70.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics*, **49**(8), 1259–1266.

Tarantola, A., 2005. *Inverse Problem theory and methods for model parameter estimation*, Society for Industrial and Applied Mathematics, Philadelphia.

Vigh, Denes, Starr, Bill, Kapoor, Jerry, and Li, Hongyan, 2010. 3d full waveform inversion on a gulf of mexico waz data set, *SEG Technical Program Expanded Abstracts*, **29**(1), 957–961.

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

Wang, Z., Navon, I. M., Dimet, F.X., and Zou, X., 1992. The second order adjoint analysis: Theory and applications, *Meteorology and Atmospheric Physics*, **50**(1-3), 3–20.