The SEISCOPE Optimization Toolbox: A large-scale nonlinear optimization library based on reverse communication

Ludovic Métivier[*,†], **Romain Brossier**[†]

[*] *LJK, Univ. Grenoble Alpes, CNRS, France. E-mail: ludovic.metivier@ujf-grenoble.fr.*

[†] *ISTerre , Univ. Grenoble Alpes, France. E-mail: romain.brossier@ujf-grenoble.fr*

Peer-reviewed code related to this article can be found at http://software.seg.org/2016/0001

Code number: 2016/0001

(December 2, 2015)

Running head: **The SEISCOPE Optimization Toolbox**

ABSTRACT

The SEISCOPE optimization toolbox is a set of `FORTRAN 90` routines which implement first-order methods (steepest-descent, nonlinear conjugate gradient) and second-order methods (l-BFGS, truncated Newton), for the solution of large-scale nonlinear optimization problems. An efficient linesearch strategy ensures the robustness of these implementations. The routines are proposed as black boxes easy to interface with any computational code where such large-scale minimization problems have to be solved. Travel-time tomography, least-squares migration, or full waveform inversion are examples of such problems in the context of geophysics. Integrating the toolbox for solving this class of problems presents two advantages. First, it helps to separate the routines depending on the physics of the problem from the ones related to the minimization itself, thanks to the reverse communication protocol. This enhances flexibility in code development and maintenance. Second, it allows

us to switch easily between different optimization algorithms. In particular, it reduces the complexity related to the implementation of second-order methods. As the latter benefit from faster convergence rates compared to first-order methods, significant improvements in terms of computational efforts can be expected.

INTRODUCTION

Nonlinear optimization problems are ubiquitous in geophysics. Methods such as travel-time tomography (Nolet, 1987), least-squares migration (Lambaré et al., 1992; Nemeth et al., 1999), or full waveform inversion (FWI) (Virieux and Operto, 2009), are based on the minimization of an objective function which measures the misfit between simulated and observed data from seismic. The simulated data (travel-time, partial measurement of the pressure and/or displacement velocity wavefields) depends on subsurface parameters (P- or S-wave velocities, density, anisotropy parameters, reflectivity among others). The minimization of the objective function consists in computing the subsurface parameters which best explain the observed data. In the general case, the simulated data depends nonlinearly on the subsurface parameters, yielding a nonlinear optimization problem.

Two general classes of methods exist to solve this type of problem: global and semi-global methods (Sen and Stoffa, 1995; Spall, 2003) or local descent algorithms (Nocedal and Wright, 2006; Bonnans et al., 2006). The methods belonging to the first class have the capability to find the global minimum of an objective function from any given starting point. They are said to be globally convergent. These methods proceed in a guided-random sampling of the parameter space to find the global minimum of the objective function. However, as soon as the number of discrete parameters exceeds a few tenths to hundreds, the number of evaluations of the objective function required to find its global minimum becomes excessively important for these methods to be applied in a reasonable time, even using modern high performance computing (HPC) platforms. In geophysics, realistic applications yield optimization problems for which the number of unknowns parameters is easily several orders of magnitude higher (billions may be involved for 3D FWI applications), rendering

global approaches intractable.

Local descent algorithms only guarantee local convergence: from a given starting point, the nearest local minimum of the objective function is found. Instead of proceeding to a random sampling, these methods are based on a guided exploration of the model space following the descent directions of the objective function. From an initial guess, an estimation of the local minimum is found by successive updates following these directions. The convergence and the monotonic decrease of the objective function is guaranteed by a proper linesearch or trust-region strategy (Nocedal and Wright, 2006). This consists in giving an appropriate scale to the update brought to the current model estimate. This guided exploration strategy is drastically cheaper than a random sampling of the model space: the number of iterations required to converge to a minimum is far lower than the number of evaluations of the objective function which would be required to obtain an acceptable sampling of the model space when its dimension is large. This is the reason why local descent methods are used for solving large scale optimization problems.

Among possible local descent methods, the simplest ones use only the first-order derivatives of the objective function (the gradient) to define the descent direction. These methods are referred to as first-order methods. The steepest-descent and nonlinear conjugate gradient algorithms are examples of these methods. More elaborated descent schemes account for the curvature of the objective function through quasi-Newton algorithms. These algorithms are based on approximations of the inverse Hessian operator (the matrix of the second-order derivatives of the objective function) following three different (possibly complementary) strategies.

The first strategy is referred to as the preconditioning strategy. In some situations,

an approximation of the inverse Hessian operator can be computed from analytic or semi-analytic formulas. Diagonal approximations can be used. An example in the context of FWI is the pseudo-Hessian preconditioning strategy proposed by Shin et al. (2001). This strategy can be used to derive an approximate inverse of the Hessian operator, which is integrated within first-order algorithms by multiplying the gradient by this approximation. The second strategy is based on the computation of an approximate inverse Hessian operator, from finite-differences of the gradient, such as in the l-BFGS algorithm (Nocedal, 1980). Instead of using an approximate inverse of the Hessian operator, the third strategy, which is referred to as the truncated Newton method, solves incompletely the linear system associated with the Newton equations, namely, the linear system relating the gradient to the descent direction, through a conjugate gradient iterative solver (Nash, 2000).

In geophysics, the solution of minimization problems is often computed using basic type methods such as the steepest descent or the nonlinear conjugate gradient. In addition, linesearch algorithms are not always implemented, and as a consequence the convergence and the monotonic decrease of the objective function is not guaranteed. Moreover, it is not unusual to see computer codes in which the minimization procedure is embedded into the routines related to the physical problem at stake. This type of implementation is not flexible, as it implies that the optimization process has to be re-developed for each application. Improving the optimization process from first-order to second-order methods may also require significant modifications of the code.

The SEISCOPE optimization toolbox has been designed to propose a solution to these issues. The first objective of the toolbox is to propose minimization routines which can be easily interfaced with any computational code through a reverse communication protocol (Dongarra et al., 1995). The principle is the following: the computation of the solution

of the optimization problem is performed in a specific routine of the code. This routine is organized as a minimization loop. At each iteration of the loop, the minimization routine from the toolbox chosen by the user is called. This routine communicates what is the quantity required at the current iteration: objective function, gradient, or Hessian-vector product. These quantities are computed by the user in specific routines, external to the minimization loop. The process continues until the convergence is reached. This implementation paradigm yields a complete separation between the code related to the physics of the problem and the code related to the solution of the minimization problem. This ensures a greater versatility as one can easily modify one of these parts while keeping the other unchanged.

The second objective of the toolbox is to propose robust and efficient methods for large-scale nonlinear optimization problems. Four different optimization schemes are implemented: steepest-descent, nonlinear conjugate gradient, l-BFGS and truncated Newton method. All these methods share the same linesearch strategy, which assures the convergence and the monotonic decrease of the objective function, and allows to compare the efficiency of the methods on a given application. In addition, the four methods can be combined with preconditioning strategies: any information regarding the curvature of the objective function can be easily incorporated to improve the convergence rate of the algorithms. Finally, the toolbox offers the possibility to use second-order methods as well as first-order methods. The complexity associated with the implementation of the l-BFGS approximation is hidden to the user. The use of truncated Newton methods only requires to implement Hessian-vector products.

It should be mentioned that a similar development in the `C++` language has been provided by members of the Rice project. The developments are based on RVL, the Rice Vector

library. The available methods are l-BFGS and the truncated Newton method implemented in the framework of the trust-region globalization strategy, also known as the Steihaug-Toint algorithm (Steihaug, 1983; Gould et al., 1999). However, to the best of our knowledge, the implementation is not based on a reverse communication protocol.

We devote this article to the presentation of the SEISCOPE optimization toolbox. First, we give an overview of the optimization methods which are implemented. No details on convergence proofs and computation of convergence rates are given. Those can be found in reference textbooks (Nocedal and Wright, 2006; Bonnans et al., 2006). We focus on the general principles of these methods. We present a code sample in which the toolbox is interfaced to illustrate the reverse communication protocol in a practical example. We provide details on the implementation of the routines. In the numerical study section, we investigate the use of the toolbox on two different applications. We minimize the bi-dimensional Rosenbrock function using the different methods proposed in the toolbox. We compare the convergence of these algorithms and the optimization path they follow. We give a second illustration on a large scale nonlinear minimization problem related to a 2D acoustic frequency-domain FWI case study. We consider synthetic data acquired on the Marmousi 2 model. We compare the convergence of the different minimization algorithms. These two examples emphasize the superiority of second-order optimization methods over first-order methods. We give a conclusion and perspectives in the final section.

THEORY

Generalities

The routines implemented in the SEISCOPE optimization toolbox are designed to solve unconstrained and bound constrained nonlinear minimization problems, under the general form

$$\min_{x\in\Omega} f(x), \tag{1}$$

where

$$\Omega = \prod_{i=1}^{n} [a_i, b_i] \subset \mathbb{R}^n, \;\; n \in \mathbb{N}, \tag{2}$$

and $f(x)$ is a sufficiently smooth function (at least twice differentiable) depending nonlinearly on the variable x.

All the routines considered in the toolbox belong to the class of local descent algorithms. From an initial guess $x_0 \in \Omega$, a sequence of iterates is built following the recurrence

$$x_{k+1} = x_k + \alpha_k \Delta x_k, \tag{3}$$

where $\alpha_k \in \mathbb{R}_+^*$ is a steplength and $\Delta x_k \in \mathbb{R}^n$ is a descent direction. The recurrence (3) is repeated until a certain stopping convergence criterion is reached.

The steplength α_k is computed through a linesearch process to satisfy the standard Wolfe conditions*. Satisfying the Wolfe conditions ensures the convergence toward a local minimum, provided $f(x)$ is bounded and sufficiently smooth (Nocedal and Wright, 2006). The satisfaction of the bound constraints is ensured through the projection of each iterate within the feasible domain Ω in the linesearch process.

*Standard is understood as opposed to the strong Wolfe conditions, which are more restrictive. This issue is important regarding the nonlinear conjugate gradient method implementation (see the nonlinear conjugate gradient subsection). The Wolfe conditions are described later in equations (13) and (14).

The different nonlinear minimization methods differ by the computation of the descent direction Δx_k. First, a description of the computation of these quantities depending on the minimization routines is given. More details on the linesearch and the bound constraints enforcement algorithm are given after.

In the following, the gradient of the objective function is denoted by $\nabla f(x)$, and the Hessian operator by $H(x)$. A preconditioner for the Hessian operator $H(x)$ is denoted by $P(x)$, such that

$$P(x) \simeq H^{-1}(x). \tag{4}$$

Steepest descent

The simplest optimization routine implemented is the steepest descent algorithm. This method uses the following descent direction

$$\Delta x_k = -P(x_k)\nabla f(x_k). \tag{5}$$

The convergence rate of this method is linear. When no preconditioner is available (P is the identity matrix), the descent direction is simply the opposite of the gradient. In this case, examples show that the number of iterations required to reach the convergence can be extremely high.

Nonlinear conjugate gradient

The conjugate gradient algorithm is an iterative linear solver for symmetric positive definite systems. This method can be interpreted as a minimization algorithm for quadratic functions. The nonlinear conjugate gradient method is conceived as an extension of this algorithm to the minimization of general nonlinear functions. The model update is computed

as a linear combination of the opposite of the gradient and the descent direction computed at the previous iteration

$$\begin{cases} \Delta x_0 = -P(x_0)\nabla f(x_0), \\ \Delta x_k = -P(x_k)\nabla f(x_k) + \beta_k \Delta x_{k-1}, \quad k \geq 1, \end{cases} \tag{6}$$

where $\beta_k \in \mathbb{R}$ is a scalar parameter. Different formulations can be used to compute β_k, giving as many versions of the nonlinear conjugate gradient algorithm. Standard implementations use formulas from Fletcher and Reeves (1964), Hestenes and Stiefel (1952) or Polak and Ribière (1969). In the SEISCOPE optimization toolbox, an alternative formula, proposed by Dai and Yuan (1999), is used. In this particular case, the scalar β_k is computed as

$$\beta_k = \frac{\|\nabla f(x_k)\|^2}{\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)^T \Delta x_{k-1}}. \tag{7}$$

This formulation ensures the convergence toward the nearest local minimum as soon as the steplength satisfies the Wolfe conditions. This is not the case for the standard formulations, which require satisfying the strong Wolfe conditions. This is the reason why the formulation from Dai and Yuan (1999) is preferred. This allows us to use the same linesearch strategy for all the optimization routines proposed in the toolbox, including the nonlinear conjugate gradient. The convergence of the nonlinear conjugate gradient is linear, as well as for the steepest descent algorithm. In some cases, the reduction of the number of iterations can be significant, however this potential acceleration is case dependent and therefore unpredictable.

Quasi-Newton *l*-BFGS method

Among the class of quasi-Newton methods, the *l*-BFGS algorithm has become increasingly popular because of its simplicity and efficiency (Nocedal, 1980). The *l*-BFGS approximation

relies on an approximate inverse Hessian operator Q_k computed through finite-differences of l previous values of the gradient. The resulting descent direction is computed as

$$\Delta x_k = -Q_k \nabla f(x_k). \tag{8}$$

The equation (8) is noteworthy: it is similar in form as the equation (5), where Q_k plays the role of the preconditioner $P(x_k)$. The difference comes from the fact that the preconditioning matrix Q_k is computed in a systematical way, only from the objective function gradient values, and updated at each iteration following the l-BFGS formula. In preconditioned steepest descent algorithm, the user is in charge to provide the preconditioning matrix $P(x_k)$. The l-BFGS method achieves a superlinear convergence rate, and is most of the time faster than nonlinear conjugate gradient and steepest descent algorithms.

In terms of practical implementation, the matrix Q_k is never explicitly built. The product $Q_k \nabla f(x_k)$ is directly computed instead, following a two-loop recursion algorithm (Nocedal and Wright, 2006). Interestingly, it is possible to incorporate an estimation of the inverse Hessian operator through the preconditioner $P(x_k)$ into this two-loop recursion. This means that the information from the preconditioner $P(x_k)$ and the l-BFGS formula can be combined to approximate as accurately as possible the action of the inverse Hessian operator. Surprisingly, this possibility is rarely mentioned in the literature, while it seems to be one of the most efficient strategy for large-scale nonlinear optimization problems. In the context of FWI, the combination of a diagonal preconditioner and the l-BFGS approximation can be shown to yield the best computational efficiency among the different optimization strategies available from the toolbox (see the numerical results on the Marmousi 2 case study). The details of this approach are given in the Appendix.

Truncated Newton method

Instead of using an approximate inverse of the Hessian operator, the truncated Newton method computes an inexact (truncated) solution of the linear system associated with the Newton equation (Nash, 2000)

$$H(x_k)\Delta x_k = -\nabla f(x_k). \tag{9}$$

This approximate solution of the linear system is computed through a matrix-free conjugate gradient algorithm: only the action of the Hessian operator on a given vector is required. In the context of FWI, the gradient is usually computed through first-order adjoint methods. Second-order adjoint strategies can be used to compute the action of the Hessian on a given vector (Métivier et al., 2012). In both cases, the computational complexity is reduced to the solution of the wave equation for different source terms. This allows us to use truncated Newton methods for solving FWI problems at a reasonable computational cost. This strategy has been investigated in Métivier et al. (2013, 2014); Castellanos et al. (2015)

The stopping criterion for the inner linear system is

$$\|H(x_k)\Delta x_k + \nabla f(x_k)\| \leq \eta_k \|\nabla f(x_k)\|, \tag{10}$$

where η_k is a forcing term. The equation (10) reveals that η_k controls the accuracy of the relative residuals of the linear system (9). For large-scale nonlinear applications, particular choices for η_k are recommended, as it is detailed by Eisenstat and Walker (1994). Since the targeted applications for this toolbox are precisely large-scale nonlinear minimization problems, the default implementation of the computation of η_k is based on this work: η_k should decreases as soon as the function which is minimized is locally quadratic. The measure of the local quadratic approximation of the misfit function is based on the gradient

current and previous values. For FWI application, this strategy reveals to be efficient (Métivier et al., 2013, 2014; Castellanos et al., 2015). However, for simpler problems, a constant and small value for η_k might produce better results. In numerical results section, for the Rosenbrock experiment, η_k is set to 10^{-5} and remains constant throughout the iterations.

In the context of the truncated Newton method, the preconditioner $P(x)$ is used naturally as a preconditioner for the linear system (9). This implies that the symmetry of (9) has to be preserved through the preconditioning operation. Therefore, $P(x)$ cannot be used simply as a left or right preconditioner. The preconditioner $P(x)$ has to be symmetric definite positive. In this case, it can be factorized as

$$P(x) = C(x)^T C(x), \tag{11}$$

and the following symmetric preconditioning strategy can be used

$$C(x_k)^T H(x_k) C(x_k) \Delta y_k = -C(x_k)^T \nabla f(x_k), \quad \Delta x_k = C(x_k) \Delta y_k. \tag{12}$$

The symmetry of (9) is preserved as $C(x_k)^T H(x_k) C(x_k)$ is symmetric by construction.

In terms of implementation, the formulation (12) may imply that a factorization of the preconditioner $P(x)$ under the form (11) is required. However, a particular implementation of the preconditioned conjugate gradient algorithm, such as the one proposed in Nocedal and Wright (2006), allows us to solve (12) through the conjugate gradient method using only matrix-vector products from $H(x)$ and $P(x)$. In practice, this preconditioned conjugate gradient implementation differs from the non-preconditioned conjugate gradient implementation only by the multiplication of the residuals by the matrix $P(x)$. Thus, from the user point of view, only the computation of the action of $P(x)$ on a given vector is required.

The truncated Newton method has been shown to converge even faster than the l-BFGS method, achieving a nearly quadratic convergence rate close to the solution. However, the computation cost of each iteration is increased by the (inexact) solution of the Newton equation. In practice, l-BFGS and truncated Newton methods have been shown to be competitive on a collection of benchmark problems from the nonlinear optimization community (Nash and Nocedal, 1991). The same conclusions have been drawn for mono-parameter FWI applications (Métivier et al., 2013, 2014; Castellanos et al., 2015), and this is what can be observed in the numerical experiments proposed in this study.

Linesearch and bound constraints

Using a linesearch strategy is necessary to ensure the robustness of the optimization routines, that is guaranteeing a monotonic decrease of the objective function and the convergence toward a local minimum. The linesearch algorithm enforces the Wolfe conditions. These are the sufficient decrease condition

$$f(x_k + \alpha \Delta x_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T \Delta x_k, \tag{13}$$

and the curvature condition

$$\nabla f\left(x_k + \alpha \Delta x_k\right)^T \Delta x_k \geq c_2 \nabla f(x_k)^T \Delta x_k, \tag{14}$$

where c_1 and c_2 are coefficients such that

$$0 < c_1 < c_2 \leq 1. \tag{15}$$

In practice, the values

$$c_1 = 10^{-4}, \quad c_2 = 0.9, \tag{16}$$

are used, following Nocedal and Wright (2006). The first condition indicates that the steplength α should be computed to ensure that the update in the direction Δx_k will generate a sufficient decrease of the objective function. The second condition is used to rule out too small values for α itself, which would lead to undesirable small updates of the estimation x_k.

The bound constraints satisfaction is enforced through a projection method, similarly to the strategy proposed by Byrd et al. (1995). At each iteration of the linesearch, before testing the two Wolfe conditions, the updated model $x_k^* = x_k + \alpha \Delta x_k$ is projected into the feasible domain Ω, through the operator $Proj(x)$ defined as

$$Proj(x)_i = \begin{cases} x_i \ \text{ if } \ a_i \leq x_i \leq b_i \\ a_i \ \text{ if } \ x_i < a_i \\ b_i \ \text{ if } \ x_i > b_i, \end{cases} \tag{17}$$

where the subscript i denotes the ith component of the vectors of $\mathbb{R}^n$. This procedure ensures that the estimated model always remains in the feasible domain Ω. Although there is no proof of convergence for optimization algorithms using this bound constraints enforcement, the method has yielded satisfactory results in practice. In some sense, it can be viewed as a simplification of the method of projection on convex sets (POCS) proposed in the context of seismic inversion by Baumstein (2013).

IMPLEMENTATION

The reverse communication protocol on which are based all the routines of the SEISCOPE optimization toolbox requires a specific implementation of the different algorithms. The principle of reverse communication from the user point of view is first introduced. A specific code example is used, where the minimization of a function is performed with the

preconditioned l-BFGS routine from the toolbox. Next, details on the implementation of the algorithms of the toolbox are given. The general algorithmic structure of the different optimization methods in the toolbox is presented, as well as the structure of the linesearch algorithm. Note that in the current version, all the toolbox routines are implemented in single precision, with the purpose to save computational time on large-scale applications.

Reverse communication protocol

The principle of reverse communication is illustrated from the code sample presented in Figure 1. In this example, the user minimizes an objective function $f(x)$ which corresponds to the Rosenbrock function. In this example, x is a two-dimensional real vector such that $x \in \mathbb{R}^2$. In general, x is a n-dimensional vector. It is implemented in an array structure. The minimization is performed from an initial value x_0 which initializes x. At the end of the minimization loop, x contains the solution of the minimization problem. In this example, the minimization is performed with the preconditioned version of the l-BFGS algorithm proposed in the toolbox by the routine `PLBFGS`.

The code is organized in a `do while` loop. The principle is the following: while the convergence has not been reached, further iterations of the minimization process are needed. Each iteration of this loop starts by a call to the optimization routine. The communication with the optimization routine is achieved through the four characters chain `FLAG`. This character chain is an input/output variable of the optimization routine. On return of each call, it is updated to communicate with the user and to require the computation of particular quantities. On first call, this variable is initialized by the user to the chain of characters ‘‘`INIT`’’.

Following this principle, after each call to the optimization routine, the user tests the possible values for the communicator and performs the subsequent actions. If `FLAG` has been set to ‘‘`GRAD`’’, the computation of the objective function and its gradient at the current value of `x` is required. In this case, the code calls the subroutine `Rosenbrock` with the current `x` as an input. On return, the subroutine `Rosenbrock` computes the objective function as well as its gradient at the current `x`, and store these quantities in the variables `fcost` and `grad` respectively. The code reaches the end of the `if/end if` loop and returns to the minimization function `PLBFGS` with the required values for the variables `fcost` and `grad`.

If `FLAG` is set to ‘‘`PREC`’’, the optimization loop requires the preconditioner to be applied. In this case, the user extracts the vector `y` from the variable `optim`. The data type of `optim` is proper to the optimization toolbox and is described in the file `optim.h`. The field of interest is `optim%q_plb`. The user then applies a preconditioner to `y` through the call to the routine `apply_Preco`. On return of this routine, the quantity `Py` is transferred into the data structure `optim` in the field `optim%q_plb`. The code reaches the end of the `if/end if` loop and returns to the minimization function `PLBFGS` with the required values for the variable `optim%q_plb`. A more efficient strategy in terms of high performance computing would consist in performing the two copies involving the vector `optim%q_plb` directly in the routine `apply_Preco` to preserve the locality of the data. However, the purpose here is to illustrate the use of the toolbox rather than proposing an optimized implementation.

When `FLAG` is set to ‘‘`NSTE`’’, this is an indication that the linesearch process just terminated and has yielded a new iterate x_{k+1} of the minimization sequence. This new iterate is stored in the variable `x`. This information can be useful to track the history of the computed approximation of the solution. In this example, the user calls a routine

`print_information` to this purpose.

If `FLAG` is `''CONV''`, the default convergence criterion implemented in the optimization routine has been reached. This convergence criterion is based on the reduction of the relative objective function. The convergence is declared when the condition

$$\frac{f(x_k)}{f(x_0)} \leq \varepsilon, \tag{18}$$

is satisfied, where ε is a threshold parameter initialized by the user.

If `FLAG` is `''FAIL''`, a failure in the linesearch process has been detected. The user is in charge of defining the minimization loop and is responsible for defining the actual stopping criterion. This stopping criterion can be based on the information provided by the optimization routine. For instance, in the example provided, the minimization loop stops as soon as `FLAG` is set to `''CONV''` or to `''FAIL''`. However, it is also possible to ignore this information, or to complete it with additional tests to define the actual stopping criterion. On exit of the `do while` loop, `x` contains the solution reached when the convergence criterion is satisfied.

The code sample presented in this example can be used as a template for the minimization of any objective function $f(x)$ for which it is possible to compute the gradient. The optimization algorithm used here is the preconditioned l-BFGS method, however, this code sample can be straightforwardly extended to the use of the other methods from the toolbox. To this purpose, the table 2 presents the different values that can be taken by the variable `FLAG` and the corresponding actions required depending on the choice of the minimization routine. For practical use, more details are given in the toolbox documentation.

In the following subsections, details on the implementation of the toolbox are given. Readers only interested in the use of the toolbox may skip the two following subsections

and jump directly to the numerical examples section.

General algorithmic structure of the optimization routines

The diagram in Figure 2 presents the general structure of the toolbox optimization routines. The workflow starts with a call of the routine from the user. The value of the communicator `FLAG` is first tested.

If it is equal to "`INIT`", this means that this is a first call from the user. In this case, subsequent initializations have to be performed. An initialization routine (here `PSTD_init`) is thus called. This routine is not described in details here; this is where memory allocation and parameters settings (such as the scalar c_1 and c_2 for the linesearch) are performed. The memory allocation consists in allocating different fields in the data structure `optim`. Depending on the minimization routines, certain fields may be allocated or not. The l-BFGS algorithm requires specific fields related to the l-BFGS approximation to be allocated for instance. After this initialization has been performed, the communicator `FLAG` is set to "`GRAD`" and the routine returns to the user.

If the communicator `FLAG` has any value different from "`INIT`", the linesearch algorithm is called. Returning from the linesearch, the algorithm asks if the linesearch has terminated or not.

If the linesearch has terminated, this means that a new iterate has been computed. This new iterate is stored in the variable `x`. In this situation, the algorithm tests if the stopping criterion (18) is satisfied. If this is the case, the communicator `FLAG` is set to "`CONV`" and the routine returns to the user. If not, this means that a new step has been taken in the minimization process, while the convergence has not been reached yet. In this case, a new

descent direction is computed, and the communicator `FLAG` is set to "`NSTE`". This value of the communicator is useful for intermediate printing of the solution for instance. The routine returns to the user.

If the linesearch process is not terminated, this means that a new value of the objective function and its gradient are required to continue the linesearch process. In this case, the steplength is updated, the communicator `FLAG` is set to "`GRAD`", and the routine returns to the user.

The steepest descent, nonlinear conjugate gradient and l-BFGS algorithms share exactly this algorithmic structure in the toolbox implementation. The only difference is related to the routine called for the initialization and the computation of the descent direction.

The structure has to be modified as soon as the computation of the descent direction also requires a direct intervention of the user. This is the case for instance for the preconditioned l-BFGS routine. In this situation, the computation of the descent direction is split into two parts, corresponding to the two-loop recursion (see Appendix for more details). Between these two loops, a call to the user is required for applying the preconditioner.

The same is true for truncated Newton methods (with or without preconditioning). The computation of the descent direction requires additional communications with the user for applying the Hessian operator and the preconditioner to particular vectors, in order to compute the incomplete solution of the system (12).

From this overview of the data structure, one can see that the linesearch algorithm plays a central role in the optimization routines implementation. More details on its implementation are given in the next subsection.

Linesearch implementation

A diagram summarizing the linesearch algorithmic structure is presented in Figure 3. The structure is also based on a reverse communication paradigm. The calling program is in charge of the computation of the quantities required by the linesearch algorithm: objective function and gradient computed at particular points.

The first call to the linesearch subroutine is indicated by an internal linesearch flag. In this case, a first candidate for the next model update x_{k+1} is computed. This candidate is denoted by x_k^*, computed as

$$x_k^* = Proj(x_k + \alpha_k \Delta x_k). \tag{19}$$

The formula (19) means that the candidate is computed as the current value x_k updated in the descent direction Δx_k with a step-length α_k and projected into the set Ω to satisfy the bound constraints which could be potentially imposed.

At this stage, the Wolfe conditions have to be tested on x_k^*. However, this requires the computation of the quantities $f(x_k^*)$ and $\nabla f(x_k^*)$. The linesearch routine thus goes back to the calling program with a request for computing these quantities. In particular, it indicates that the linesearch did not terminate.

When these quantities have been computed, the calling program goes back to the linesearch routine. This time the internal linesearch flag indicates it is not a first call to the linesearch. As a consequence, the two Wolfe conditions are tested in a sequential order. If they are satisfied, the linesearch algorithm terminates successfully with a new model update x_{k+1} which is set to the candidate value x_k^*. The internal linesearch flag is set back to "`first_call`".

If one of the two Wolfe conditions is not satisfied, the steplength α_k has to be adjusted. The modification of α_k is performed following the bracketing strategy proposed in Bonnans et al. (2006). Once α_k has been modified, a new candidate x_k^* is computed, and the linesearch algorithm goes back to the calling program with a request for the computation of $f(x_k^*)$ and $\nabla f(x_k^*)$.

The optimization routines implemented in the toolbox are thus based on a two-level reverse communication process. The first-level is between the user and the optimization routine. The second level is embedded in the optimization routine itself and corresponds to the linesearch implementation. The potential complexity of such an implementation is transparent for the user. When asked to compute the gradient of the objective function, the user does not have to know if this computation is requested for the linesearch process or the evaluation of a new descent direction.

In terms of practical implementation, the linesearch algorithm requires to define an initial step-length α_0. In the current implementation, this step-length is simply set to 1. Besides, at each iteration k of the minimization, the new step-length α_k is initialized to the previous value α_{k-1}. From our experience, this approach results in the following behavior. At the first iteration of the minimization process, the linesearch algorithm may require several internal iterations to adjust the step-length. However, for further iterations, the step-length computed at the previous iteration often directly yields a candidate which satisfies the Wolfe conditions. The additional computational cost related to the linesearch is thus limited to the first iteration of the minimization. Note that this cost can be reduced by the user by using an appropriate scaling of the objective function. Indeed, at the first iteration, for most optimization methods, the descent direction is equal to the gradient. The step-length adjustment from the linesearch thus corresponds to an automatic scaling of

the misfit function. This process can be accelerated by a proper scaling performed directly by the user.

Summary of the routines implemented in the toolbox

Table 1 summarizes the name of the minimization routines implemented in the toolbox together with their basic properties. Although four main minimization methods are presented, the toolbox proposes a total of six subroutines. This is due to the implementation of preconditioning. While it is straightforward for steepest-descent and nonlinear conjugate gradient, the implementation of preconditioning for l-BFGS and truncated Newton method requires substantial modifications. Our choice is thus to implement a single version for steepest-descent and nonlinear conjugate gradient, where the user only has to apply the identity operator if no preconditioner is available, while to distinct versions (with and without preconditioning) are implemented for l-BFGS and truncated Newton method.

In theory, first-order methods have a linear convergence rate, while second-order methods benefits from superlinear convergence. While the Newton method converges with a quadratic rate close from the solution, l-BFGS and the truncated Newton method only benefit from a superlinear convergence rate as only an approximation of the inverse Hessian operator is used. The six routines require the computation of the gradient. Only the truncated Newton method (with and without preconditioning) requires the computation of Hessian-vector products.

Table 1 also summarizes the comparisons in terms of convergence speed (reduction of the misfit function depending on the number of iteration) and computational efficiency (reduction of the misfit function depending on the number of gradient/Hessian-vector product

evaluations) observed on the two numerical examples presented in the following section.

NUMERICAL EXAMPLES

In this section, numerical examples of use of the SEISCOPE optimization toolbox are provided. First, the bi-dimensional Rosenbrock objective function is considered. This function serves as a benchmark in reference textbooks to calibrate optimization methods. Its particularity is to present a narrow valley of attraction where the function has a very flat minimum. It is therefore a challenging problem for nonlinear optimization. The interest of this example is to illustrate the properties of the different optimization methods proposed in the toolbox. This example can be reproduced using the test case implemented in the toolbox (see file `00README_for_GEOPHYSICS_submission`).The second example shows an application of 2D acoustic frequency-domain FWI on the Marmousi 2 model. The Marmousi 2 model is a standard benchmark for today FWI algorithms (Martin et al., 2006). The interest of this example is to illustrate the capability of the SEISCOPE optimization toolbox to handle realistic scale optimization problems and to perform comparisons between the different algorithms which are proposed. In this case the number of discrete unknowns reaches almost 100 000. This example is not directly proposed in the toolbox code attached to this submission. However, it could be reproduced using the open-source code from the SEISCOPE team `TOY2DAC` (dowloadable here *http://seiscope2.osug.fr/TOY2DAC,82?lang=en*, accessed on 09/23/2015). The Marmousi 2 model can be downloaded here *http://www.agl.uh.edu/downloads/downloads.html, accessed on 09/23/2015*

The Rosenbrock function

The analytic expression for the bi-dimensional Rosenbrock function is

$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2. \tag{20}$$

The gradient and the Hessian of the Rosenbrock function are, respectively,

$$\nabla f(x, y) = \begin{pmatrix} 2(x - 1) - 400x(y - x^2) \\ \\ 200(y - x^2) \end{pmatrix}, \tag{21}$$

$$H(x, y) = \begin{pmatrix} 1200x^2 - 400y + 2 & -400x \\ \\ -400x & 200 \end{pmatrix}. \tag{22}$$

The Rosenbrock function reaches its minimum value at the point $(1, 1)$. In Figure 4 the valley of attraction appears in blue. The steepness of the objective function in this valley is weak, which renders the convergence to the minimum $(1, 1)$ difficult for local descent algorithms.

As an illustration of the performance of the algorithms proposed in the toolbox, the different paths taken by steepest-descent, nonlinear conjugate gradient, l-BFGS and truncated Newton methods are presented in Figure 5. The starting point (initial guess) is taken as $(x_0, y_0) = (0.25, 0.25)$.

The memory parameter l, which controls the number of gradient stored to build the l-BFGS approximation of the inverse Hessian, is set to 20. Changes of this value do not show significant modification of the l-BFGS path. For realistic size case studies, 20 is already a "large" value for this parameter. For strongly nonlinear misfit functions (for instance in the presence of noisy data when the method is applied to nonlinear least-squares minimization), it may be useful to take lower values such that $l = 3$ or $l = 5$, as the information carried

out by older iterates may be not relevant to the local approximation of the inverse Hessian operator.

Starting from (x_0, y_0), all the methods head rapidly toward the valley of attraction. The path they take when they reach this area is however very different. The steepest-descent algorithm performs the worst. It generates very small steps and therefore numerous accumulation points appear on the path it takes. The convergence is reached only after several thousand iterations. The nonlinear conjugate gradient performs better, but presents a quite irregular convergence toward the minimum. In particular, the algorithm goes beyond the minimum before heading back to it. The l-BFGS algorithm path is more regular, however a zone of accumulation points is created near the minimum. Finally, the truncated Newton method proposes the most efficient path, taking long updates in the attraction valley before reaching the minimum around which smaller steps are taken.

This example is completed by the analysis of the convergence curves of the four algorithms in terms of number of iterations and computational effort in Figure 6. Not surprisingly, the steepest-descent algorithm convergence is very slow compared to the three other methods. Although this does not appear in Figure 6, the convergence is reached only after several thousand iterations. On the other hand, the three other methods converge in less than 60 iterations. Among these three, the methods follow the hierarchy which could be expected: the nonlinear conjugate gradient is the slowest (53 iterations), followed by the l-BFGS method (29 iterations); the most efficient one is the truncated Newton method (18 iterations).

However, this observation has to be mitigated by the analysis of the behavior of the methods in terms of computational effort. The latter is measured by the number of gra-

dient computation required to reach convergence. In the case of the truncated-Newton method, the cost of a Hessian-vector multiplication is equivalent to the cost of a gradient computation. In terms of computational effort, the hierarchy between l-BFGS and the truncated Newton method is inverted. This is understandable as even if the truncated Newton method converges in less iterations, each iteration of this algorithm has a significantly higher cost related to the additional Hessian-vector products it has to perform.

The Rosenbrock case study is a good illustration of what can be expected from the different descent methods implemented in the SEISCOPE optimization toolbox. For difficult problems, the information on the local curvature of the objective function helps to guarantee a faster convergence speed. This information is embedded in the second-order derivatives of the objective function. The two methods accounting at best for this information are the l-BFGS method and the truncated Newton method. The latter has the capability to integrate the information on the Hessian with a great accuracy, which decreases the number of iteration required to reach the minimum, at the expense of an increased computational complexity. The l-BFGS method proposes the better trade-off between computational efficiency and accuracy of the Hessian estimation.

FWI on the Marmousi case study

Full Waveform Inversion principle

FWI is a seismic imaging method based on the minimization between observed data and synthetic data computed through the numerical solution of a wave propagation problem. The method allows us to retrieve high resolution quantitative estimates of subsurface parameters such as P-wave and S-wave velocities, density, or anisotropy parameters. For this

simple illustration, a 2D acoustic frequency-domain application with constant density is used. In this context, the wave propagation is described by the Helmholtz equation

$$-\frac{\omega^2}{v_P^2}u - \Delta u = s, \tag{23}$$

where ω denotes the circular frequency, $v_P(x)$ the P-wave velocity, $u(x,\omega)$ the pressure wavefield, Δ the Laplacian operator, and $s(x,\omega)$ an explosive source term.

Given partial measurement of the pressure wavefield d_{obs} acquired at the surface, the FWI problem consists in the nonlinear least-squares minimization problem

$$\min_{v_P} \frac{1}{2}\|d_{cal}(v_P) - d_{obs}\|^2, \tag{24}$$

where $\|.\|$ is the L^2 norm and $d_{cal}(v_P)$ is computed as

$$d_{cal}\left(v_P\right) = Ru\left(x,\omega;v_P\right). \tag{25}$$

In equation (25), $u\left(x,\omega;v_P\right)$ is the solution of the Helmholtz equation (23) for the P-wave velocity model v_P, and R is a restriction operator mapping this wavefield to the receiver location where the measurements are performed.

In the following numerical examples, the `TOY2DAC` code from the SEISCOPE group is used. This code is based on the solution of the equation (23) through a fourth-order mixed staggered-grid finite differences discretization (Hustedt et al., 2004). The solution of the Helmholtz equation in an infinite domain is simulated with Perfectly Matched Layers (Bérenger, 1994). The associated linear system is solved by a LU factorization performed using the massively parallel solver MUMPS (Amestoy et al., 2000; MUMPS-team, 2011). The computation of the gradient and Hessian-vector products is performed following first-order and second-order adjoint methods. Using these techniques the computation of these quantities is reduced to the solution of wave propagation problems of type (23) with different

source terms (Plessix, 2006; Métivier et al., 2012, 2013). The `TOY2DAC` code is interfaced with the SEISCOPE optimization toolbox.

Comparison of the optimization routines on the Marmousi2 benchmark model

The Marmousi 2 benchmark model and the initial model which is used are presented in Figure 7. The initial model is obtained after applying a Gaussian smoothing to the exact model with a 500 m correlation length. A fixed-spread acquisition system with 336 sources and receivers positioned at 50 m depth, from $x = 0.15$ km to $x = 16.9$ km, equally spaced each 50 m, is used. The synthetic observed data-set is constructed using six frequencies from 3 Hz to 8 Hz, with 1 Hz sampling. No free-surface condition is implemented. The synthetic observed data thus do not contain any surface multiples. No multi-scale strategy is used: the data corresponding to the six frequencies is inverted simultaneously. After discretization, the P-wave velocity model is described by approximately 100 000 discrete parameters, which yields a large-scale nonlinear optimization problem.

For FWI, the computational efficiency of the different optimization methods can be measured in terms of the number of gradient required to reach a certain level of accuracy. Indeed, a complexity analysis of the method reveals that the average complexity of one gradient computation is in $O(N^4)$ for 2D experiments ($O(N^6)$ for 3D experiments), where N is the average number of discrete points in one direction. This approximation is obtained assuming that the number of sources is in $O(N)$ (resp. $O(N^2)$ for the 3D case). In comparison, the complexity if the operations performed within the different optimization methods is in $O^(N^2)$ (resp. $O^(N^3)$ for the 3D case). Therefore the computation associated with the optimization algorithm in itself is negligible compared to the computation cost

associated with gradient computation. It should be mentioned here that in the context of adjoint strategies for computing Hessian-vector product, the computation cost of this operation is the same as the one for the gradient, therefore this simplifies the comparison between first-order method and l-BFGS with the truncated Newton method.

The results obtained using the four different optimization methods available in the toolbox, steepest-descent, nonlinear conjugate gradient, l-BFGS, truncated Newton, are presented in Figure 8. All the methods are combined with a diagonal preconditioner. To compute this preconditioner, the diagonal elements of the Hessian operator are approximated through the pseudo-Hessian approach promoted by Shin et al. (2001). The preconditioner is computed as the inverse of the diagonal matrix formed with these elements. A similar comparison of optimization methods in the context of FWI has been performed by Castellanos et al. (2015) on different models.

The number of gradients stored for computing the l-BFGS approximation is set to $l = 10$. For the truncated Newton method, the maximum number of iterations for the incomplete solution of the Newton equations is set to 10. In practice, this bound is not reached: the stopping criterion from Eisenstat and Walker (1994) is satisfied at each nonlinear iteration in less than 10 iterations. The purpose of this experiment is to compare the different convergence rates of the methods, therefore the stopping criterion which is used is based on the maximum number of nonlinear iterations to be performed, which is set to 20. The convergence curves in function of the number of iterations and the number of gradient estimations are presented in Figure 9.

As can be observed in Figure 9, the nonlinear conjugate gradient does not yield any improvement with respect to the steepest-descent method. This is not really a surprise,

given the known erratic behavior of this method. In this situation, the method seems less efficient than the steepest-descent algorithm.

A second observation is the relatively good performance of the steepest-descent compared to the very slow convergence observed for the Rosenbrock test case for this method. Two reasons may be invoked to explain this observation. First, a preconditioner is used in the Marmousi 2 case, which improves significantly the convergence rate of this method. Second, the Rosenbrock case study is a pathological situation in which the steepest-descent has the highest difficulty to converge. For the large-scale Marmousi 2 application, this seems to be no longer true.

As in the Rosenbrock case study, second-order methods outperform first-order methods. The truncated Newton method provides the fastest convergence rate in terms of nonlinear iterations. In terms of computational efficiency, the l-BFGS method provides the best performance.

The analysis of the four P-wave estimations obtained after 20 iterations of each of the algorithm is in agreement with the objective function level attained by the four methods (Fig. 8). The steepest-descent and nonlinear conjugate gradient estimations are comparable. The shallow structures are well recovered. The low velocity anomalies accounting for the presence of gas located at ($z = 2$ km, $x = 6$ km) and ($z = 2$ km, $x = 21$ km) are efficiently reconstructed. However, the deeper structures below $z = 4.5$ km are not well delineated. Conversely, the l-BFGS and truncated Newton methods are able to better refocus the energy on this deeper part of the model. The deep reflectors are correctly reconstructed. This is particularly visible in the truncated Newton reconstruction. The inverse Hessian operator acts as an efficient deblurring filter in the context of FWI (Pratt et al., 1998).

Although the l-BFGS algorithm is more efficient than the truncated Newton method in terms of computational cost for this mono-parameter FWI case study, the situation may change in the context of multi-parameter FWI. In this context, non-negligible trade-offs between different classes of parameters are expected (Operto et al., 2013). Removing these trade-offs require to account accurately for the inverse Hessian operator at the first stages of the inversion (Métivier et al., 2014). As the l-BFGS algorithm builds a progressive estimation of this operator along the iterations of the minimization loop, the estimation in the first iterations depends strongly on the prior information injected by the user. If this information is poor, as it is often the case, it is expected that the trade-offs between parameter classes contaminate the solution at the early stages of the inversion. Removing these trade-offs in the next iterations is an extremely difficult task (see for instance the multi-parameter FWI for ground penetrating radar data performed by Lavoué et al. (2014)). In contrast, the truncated Newton method is based on an approximation of the inverse Hessian operator whose accuracy does not depend on the convergence history. Therefore, it is expected that the truncated Newton method performs better than the l-BFGS algorithm in a multi-parameter FWI context.

The comparison between the different minimization algorithms on the Marmousi case study should be ended up with the following comments. The settings of the experiments have been designed such that the initial model is close from the exact one. This could favor second-order algorithm over first-order algorithm because this is the configuration in which the superlinear convergence of second-order algorithms is more likely to be observed (in the vicinity of the minimum). In real case applications, the initial model can be poorer and the difference between first-order and second-order algorithm may be less obvious, at least in the early stages of the inversion. In any case, if the initial solution is very poor, all the four

minimization strategies will fail to produce a reliable subsurface model estimation. What can be observed is an acceleration of the convergence rate when the solution approaches the minimum of the misfit function. The use of Tikhonov regularization also has in impact on the differences between first-order and second-order methods. This regularization technique, mandatory in case of noisy data for instance, results in adding a multiple of the identity to the Hessian operator. To see the impact of this regarding optimization method, one can consider the limit case, for which the Hessian operator tends toward the identity operator. In this case, first-order methods, relying on this approximation, become equivalent to second-order methods. Therefore, using strong Tikhonov regularization weights will have the effect of reducing the differences between first-order and second-order methods regarding their convergence rate.

CONCLUSION

The SEISCOPE optimization toolbox is a set of `FORTRAN 90` routines for the solution of large-scale nonlinear minimization problems. This type of problems is ubiquitous in geophysics. The toolbox implements four different optimization schemes: the steepest-descent, the nonlinear conjugate gradient, the l-BFGS method and the truncated Newton methods. All these routines are implemented with a linesearch strategy ensuring the robustness of the methods: monotonic decrease of the objective function and guaranteed convergence toward the nearest local minimum. Bound constraints can also be activated.

The use of these routines within a computational code is completely uncoupled from the physical problem through a reverse communication protocol. Within this framework, the minimization of the objective function is brought back to the definition of a minimization loop controlled by the user in which the solver is called at each iteration. Depending on

the return values of a communication flag, the user performs the action required by the solver at a given point, namely: computation of the objective function and its gradient, application of a preconditioner, computation of a Hessian-vector product. This allows for a flexible implementation and an easier use of second-order optimization methods.

The general structure of the minimization algorithms of the toolbox is based on a two-level reverse communication principle. The first level consists in implementing the communication between the user code and the minimization code. The second level is embedded in the minimization code itself, and implements the communication between a linesearch algorithm and the minimization code. The minimization code transfers the requirement from the linesearch (computation of the objective function and its gradient) directly to the user. This imbrication of reverse communication levels is thus transparent for the user.

The example of use of the toolbox provided in the numerical results part show the benefit one can expect from second-order methods. The case of the Rosenbrock function is pathological, however it illustrates how powerful can be the introduction of the information carried out by the second-order derivatives into the optimization. The Marmousi 2 case for FWI also illustrates the benefit of introducing an appropriate estimation of the inverse Hessian operator within the minimization to speed-up the convergence of the algorithm and save considerable computation time.

Perspectives of development of the toolbox to adapt it to very-large scale HPC applications is currently considered. In its actual implementation, the method requires to collect one model and one gradient on a single node of a supercomputer. For very large-scale applications, such as the one involved in 3D FWI, these quantities are usually distributed and this should require additional global communications at each iteration of the minimization

loop. In some cases, the ability to store these quantities on a single node may even be questionable. Therefore, a completely distributed version of the toolbox may be derived to relax this requirement. One possible implementation is to externalize all the operations related to scalar products and vector additions involved in the minimization procedures through the reverse communication protocol. This could slightly complicate the interface of the toolbox with other computer codes, requiring more actions to be performed within the minimization loop. However, it would represent a very efficient way of performing the minimization on very large-scale problems for which all the vector quantities have to be distributed over the whole cluster.

On a longer term, the set of routines currently included the toolbox could be extended to routines dedicated to the solution of constrained optimization problems. Based on the same architecture, sequential quadratic programming solvers could be designed, for handling problems with not only bound constraints, but also linear and nonlinear equality and inequality constraints. This type of problems arise in geophysics as soon as regularization methods are considered, beyond simple additive techniques. Another longer term extension consists in integrating trust-region algorithms to propose an alternative to the linesearch technique currently implemented. In particular, the combination of the truncated Newton method and the trust-region method, known as a the Steihaug algorithm (Steihaug, 1983), is reputed to benefit from better convergence properties than the linesearch version of the truncated Newton method. These extensions may be the topic of future investigations.

ACKNOWLEDGEMENTS

This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by BP, CGG, CHEVRON, EXXON-MOBIL, JGI, PETROBRAS, SAUDI ARAMCO,

SCHLUMBERGER, SHELL, SINOPEC, STATOIL, TOTAL and WOODSIDE. This study was granted access to the HPC resources of the Froggy platform of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche, and the HPC resources of CINES/IDRIS under the allocation 046091 made by GENCI. The authors would like to thank all the actors of the SEISCOPE consortium research group for their active and constant support. Special thanks are due to Stéphane Operto for suggesting us this article and his (always) careful proof-reading. Special thanks also go to Joe Dellinger, Philippe Thierry, Esteban Diaz and two anonymous reviewers for their suggestions, which really helped us to improve our work.

APPENDIX

The l-BFGS algorithm is based on the computation of the descent direction Δx_k following the equation (8). In (8), Q_k is defined by

$$\begin{aligned} Q_k \;=\; & \left(V_{k-1}^T \dots V_{k-l}^T\right) Q_k^0 \left(V_{k-l} \dots V_{k-1}\right) \\ & +\rho_{k-l} \left(V_{k-1}^T \dots V_{k-l+1}^T\right) s_{k-l} s_{k-l}^T \left(V_{k-l+1} \dots V_{k-1}\right) \\ & +\rho_{k-l+1} \left(V_{k-1}^T \dots V_{k-l+2}^T\right) s_{k-l+1} s_{k-l+1}^T \left(V_{k-l+2} \dots V_{k-1}\right) \\ & +\dots \\ & +\rho_{k-1} s_{k-1} s_{k-1}^T, \end{aligned} \tag{26}$$

where the pairs s_k, y_k are

$$s_k = x_{k+1} - x_k, \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k), \tag{27}$$

the scalar ρ_k are

$$\rho_k = \frac{1}{y_k^T s_k}, \tag{28}$$

and the matrices V_k are defined by

$$V_k = I - \rho_k y_k s_k^T. \tag{29}$$

These formulas are directly extracted from Nocedal and Wright (2006). As is also explained in this reference textbook, the computation of the quantity $\Delta x_k = -Q_k \nabla f(x_k)$, can be performed through the double recursion algorithm presented below.

Data: $\rho_i, s_i, y_i, \;\; i = k-l, \ldots, k-1, H_k^0, \nabla f(x_k)$

Result: $\Delta x_k = -Q_k \nabla f(x_k)$

$q = -\nabla f(x_k)$;

for $i = k-1, \ldots, k-l$ **do**

$\alpha_i = \rho_i s_i^T \Delta x_k$;

$q = q - \alpha_i y_i$;

end

$\Delta x_k = H_k^0 q$;

for $i = k-l, \ldots, k-1$ **do**

$\beta = \rho_i y_i^T \Delta x_k$;

$\Delta x_k = \Delta x_k + (\alpha_i - \beta_i) s_i$;

end

Algorithm 1: Two-loop recursion algorithm

In Algorithm 1, the matrix H_k^0 is an estimation of the inverse Hessian matrix at the k^{th} iteration of the l-BFGS minimization. In most applications, this estimation is not updated throughout the l-BFGS iterations, and kept equal to H_0^0. However, nothing prevents from performing this update. In practice, in the FWI application presented in this study, this strategy reveals to be efficient. The inverse Hessian estimation which is used is a diagonal approximation based on the pseudo-Hessian strategy (Shin et al., 2001). The computation cost of this approximation is cheap, and therefore can be recalculated at each iteration. The implementation which is used in the toolbox allows for these updates to be perform. When the routine PLBFGS is used, the preconditioning operation requested by the communicator flag set to ``PREC'' corresponds to the multiplication $\Delta x_k = H_k^0 q$. The user may thus update the matrix H_k^0 at each iteration k before performing this multiplication. Surprisingly, this possibility is rarely applied (to the best of our knowledge) while it is explicitly specified

in Nocedal and Wright (2006).

REFERENCES

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

Baumstein, A., 2013, POCS-based geophysical constraints in multi-parameter Full Wavefield Inversion: Presented at the 75th EAGE Conference & Exhibition incorporating SPE EUROPEC 2013.

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

Bonnans, J. F., J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, 2006, Numerical optimization, theoretical and practical aspects: Springer series, Universitext.

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

Castellanos, C., L. Métivier, S. Operto, R. Brossier, and J. Virieux, 2015, Fast full waveform inversion with source encoding and second-order optimization methods: Geophysical Journal International, **200(2)**, 720–744.

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

Dongarra, J., V. Eijkhout, and A. Kalhan, 1995, Reverse communication interface for linear algebra templates for iterative methods: Technical report, University of Tennessee.

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

Fletcher, R. and C. M. Reeves, 1964, Function minimization by conjugate gradient: Computer Journal, **7**, 149–154.

Gould, N. I. M., S. Lucidi, M. Roma, and P. Toint, 1999, Solving the trust-region subproblem using the lanczos method: Siam Journal on Optimization, **9**, 504–525.

Hestenes, M. R. and E. Stiefel, 1952, Methods of conjugate gradient for solving linear systems: Journal of Research of the National Bureau of Standards, **49**, no. 6, 409–436.

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

Lambaré, G., J. Virieux, R. Madariaga, and S. Jin, 1992, Iterative asymptotic inversion in the acoustic approximation: Geophysics, **57**, 1138–1154.

Lavoué, F., R. Brossier, L. Métivier, S. Garambois, and J. Virieux, 2014, Two-dimensional permittivity and conductivity imaging by full waveform inversion of multioffset GPR data: a frequency-domain quasi-Newton approach: Geophysical Journal International, **197**, 248–268.

Martin, G. S., R. Wiley, and K. J. Marfurt, 2006, Marmousi2: An elastic upgrade for Marmousi: The Leading Edge, **25**, 156–166.

Métivier, L., F. Bretaudeau, R. Brossier, S. Operto, and J. Virieux, 2014, Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures: Geophysical Prospecting, **62**, no. 6, 1353–1375.

Métivier, L., R. Brossier, J. Virieux, and S. Operto, 2012, Toward Gauss-Newton and exact Newton optimization for full waveform inversion: Presented at the 74^{tn} Annual International Meeting, EAGE.

———, 2013, Full Waveform Inversion and the truncated Newton method: SIAM Journal On Scientific Computing, **35(2)**, B401–B437.

MUMPS-team, 2011, MUMPS - MUltifrontal Massively Parallel Solver users' guide - version 4.10.0 (may 10, 2011). ENSEEIHT-ENS Lyon, http://www.enseeiht.fr/apo/MUMPS/ or http://graal.ens-lyon.fr/MUMPS.

Nash, S. and J. Nocedal, 1991, A Numerical Study of the Limited Memory BFGS Method and the Truncated-Newton Method for Large Scale Optimization: SIAM Journal on Optimization, **1**, 358–372.

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

Nemeth, T., C. Wu, and G. T. Schuster, 1999, Least-squares migration of incomplete reflection data:

Geophysics, **64**, 208–221.

Nocedal, J., 1980, Updating Quasi-Newton Matrices With Limited Storage: Mathematics of Computation, **35**, 773–782.

Nocedal, J. and S. J. Wright, 2006, Numerical optimization: Springer, 2nd edition.

Nolet, G., 1987, Seismic tomography with applications in global seismology and exploration geophysics: D. Reidel publishing Company.

Operto, S., R. Brossier, Y. Gholami, L. Métivier, V. Prieux, A. Ribodetti, and J. Virieux, 2013, A guided tour of multiparameter full waveform inversion for multicomponent data: from theory to practice: The Leading Edge, **Special section Full Waveform Inversion**, 1040–1054.

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**, 495–503.

Polak, E. and G. Ribière, 1969, Note sur la convergence de méthodes de directions conjuguées: Revue Française d'Informatique et de Recherche Opérationnelle, **16**, 35–43.

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

Sen, M. K. and P. L. Stoffa, 1995, Global optimization methods in geophysical inversion: Elsevier Science Publishing Co.

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

Spall, J. C., 2003, Introduction to Stochastic Search and Oprimization: Estimation, simulation and control: Wiley-Interscience Series in Discrete Mathematics and Optimization. 1st edition.

Steihaug, T., 1983, The conjugate gradient method and trust regions in large scale optimization: SIAM Journal on Numerical Analysis, **20**, 626–637.

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

TABLES

	First-order methods		Second-order methods			
Routine name	PSTD	PNLCG	LBFGS	PLBFGS	TRN	PTRN
Method	Steep. descent	Nonlin. CG	l-BFGS	l-BFGS	trunc. Newt.	trunc. Newt.
Preconditioning	Yes	Yes	No	Yes	No	Yes
Conv. rate	lin.	lin.	suplin.	suplin.	suplin.	suplin.
$\nabla f(m)$ required	Yes	Yes	Yes	Yes	Yes	Yes
$H(m)v$ required	No	No	No	No	Yes	Yes
Rank test 1	4,**4**	3,**3**	2,**1**	N/A	1,**2**	N/A
Rank test 2	3,**3**	4,**4**	N/A	2,**1**	N/A	1,**2**

Table 1: Summary of the minimization routines implemented in the toolbox. The ranks in the two last lines correspond to the performance on the numerical examples 1 (Rosenbrock function) and 2 (FWI on the Marmousi mode) in terms of convergence rate (normal font), and number of computed gradient (bold font).

FLAG values	Action required /Meaning
‘‘INIT’’	This flag is set only by the user, prior to the minimization loop. On reception of this flag, the solver performs the necessary initializations (parameter settings, memory allocations).
‘‘CONV’’	The convergence criterion $\frac{f(x_k)}{f(x_0)} < \varepsilon$ is satisfied. The variable x contains the solution x_k computed at this stage.
‘‘FAIL’’	The linesearch has terminated on a failure.
‘‘NSTE’’	The linesearch has terminated successfully. The variable x contains the new iterate x_{k+1}.
‘‘GRAD’’	Compute $f(x)$ and store it in fcost. Compute $\nabla f(x)$ and store it in grad. For PNLCG and PSTD, apply the preconditioner to $\nabla f(x)$ and store it in grad_preco.
‘‘PREC’’	Only for PLBGS and PTRN routines. If the routine called is PLBFGS, then apply the preconditioner to the variable optim%q_plb and store it in the same variable. If the routine called is PTRN, then apply the preconditioner to the variable optim%residual and store it in the variable optim%residual_preco
‘‘HESS’’	Only for TRN and PTRN routines. Apply the Hessian operator to the variable optim%d and store it in optim%Hd

Table 2: Summary of the different values taken by the communication variable FLAG and their meaning.

FIGURES CAPTION

- FIGURE 1: Example of a reverse communication loop
- FIGURE 2: General structure of the toolbox optimization routines
- FIGURE 3: Schematic view of the linesearch implementations
- FIGURE 4: Bi-dimensional Rosenbrock function map. The two white crosses highlight the starting point for the optimization process located at $(0.25, 0.25)$ and the minimum of the Rosenbrock function, at the point $(1, 1)$. The valley of attraction appears as a broad blue channel where the function has a very flat minimum
- FIGURE 5: Convergence paths taken by the four different optimization methods: steepest descent (a), nonlinear conjugate gradient (b), l-BFGS (c), truncated Newton (d). The starting point is $(0.25, 0.25)$. The convergence is reached at $(1, 1)$.
- FIGURE 6: Rosenbrock function case study. Convergence rate of the four optimization methods in terms of iterations (a) and computational cost in terms of the number of gradient evaluations (b). For the truncated Newton method, the computation cost of a Hessian-vector product is assimilated to the computation cost of a gradient.
- FIGURE 7: Marmousi 2 synthetic case study. Exact model (a). Initial model obtained after smoothing the exact model with a Gaussian filter (b). The correlation length for the Gaussian filter is set to 500 m. The number of discrete points is equal to 141×681, which yields a minimization problem involving approximately $96,000$ unknowns.
- FIGURE 8: Marmousi case study. Models reconstructed by the four different optimization methods. Steepest-descent (a), nonlinear conjugate gradient (b), l-BFGS (c), truncated Newton (d).
- FIGURE 9: Marmousi case study. Convergence curves of the four different optimization methods in terms of iterations (a), and computational cost in terms of the number of gradient evaluations (b). For the truncated Newton method, the computation cost of a Hessian-vector

product is assimilated to the computation cost of a gradient. This is consistent with the implementation of second-order adjoint formulas for the computation of these quantities.

FIGURES

```
do while ((FLAG.ne.'CONV').and.(FLAG.ne.'FAIL'))
    call PLBFGS(n,x,fcost,grad,grad_preco,optim,FLAG)
    if(FLAG.eq.'GRAD') then
        !compute objective function and its gradient at point x
        call Rosenbrock(x,fcost,grad)
    endif
    if(FLAG.eq.'PREC') then
        !apply preconditioning to optim%q_plb
        y(:)=optim%q_plb(:)
        call apply_preco(y,Py)
        optim%q_plb(:)=Py(:)
    endif
    if(FLAG.eq.'NSTE') then
        !a new step has been taken
        call print_information(x)
    endif
enddo
```

Figure 1: Example of a reverse communication loop.

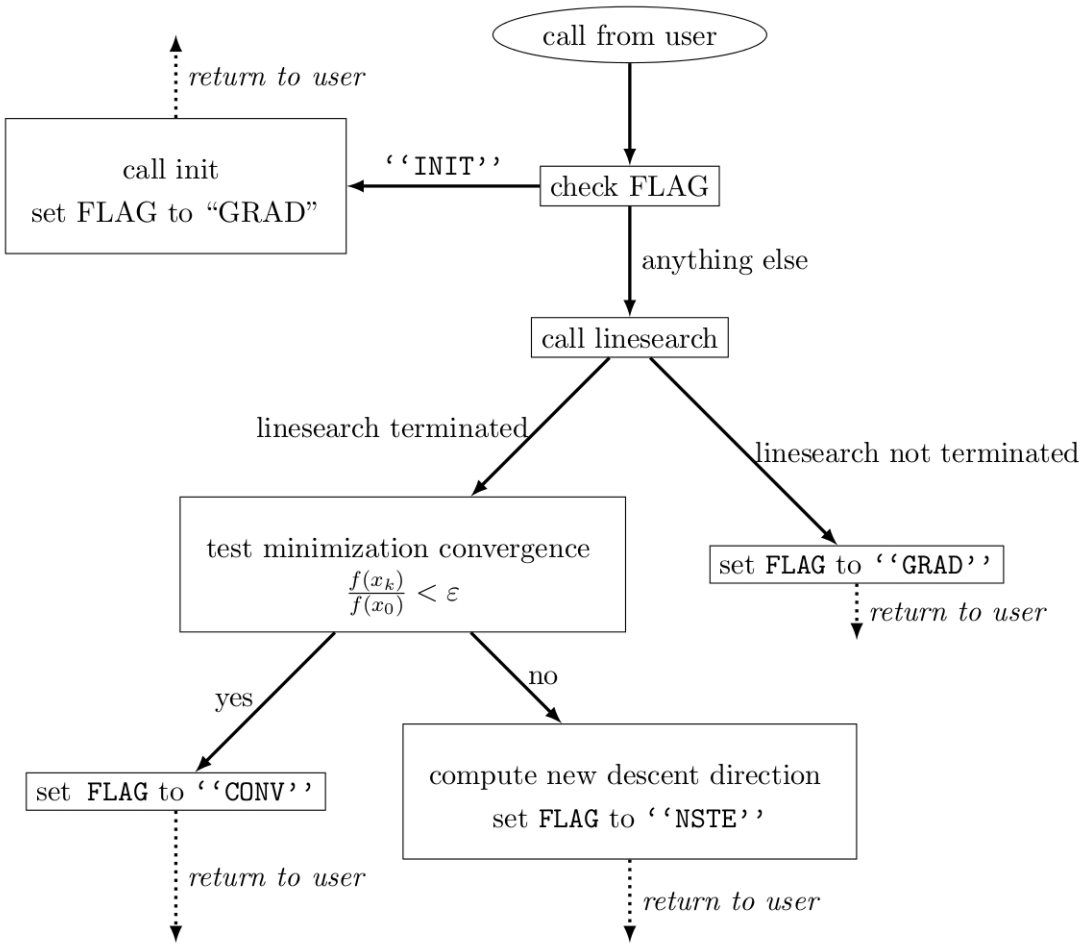

Figure 2: General structure of the toolbox optimization routines.

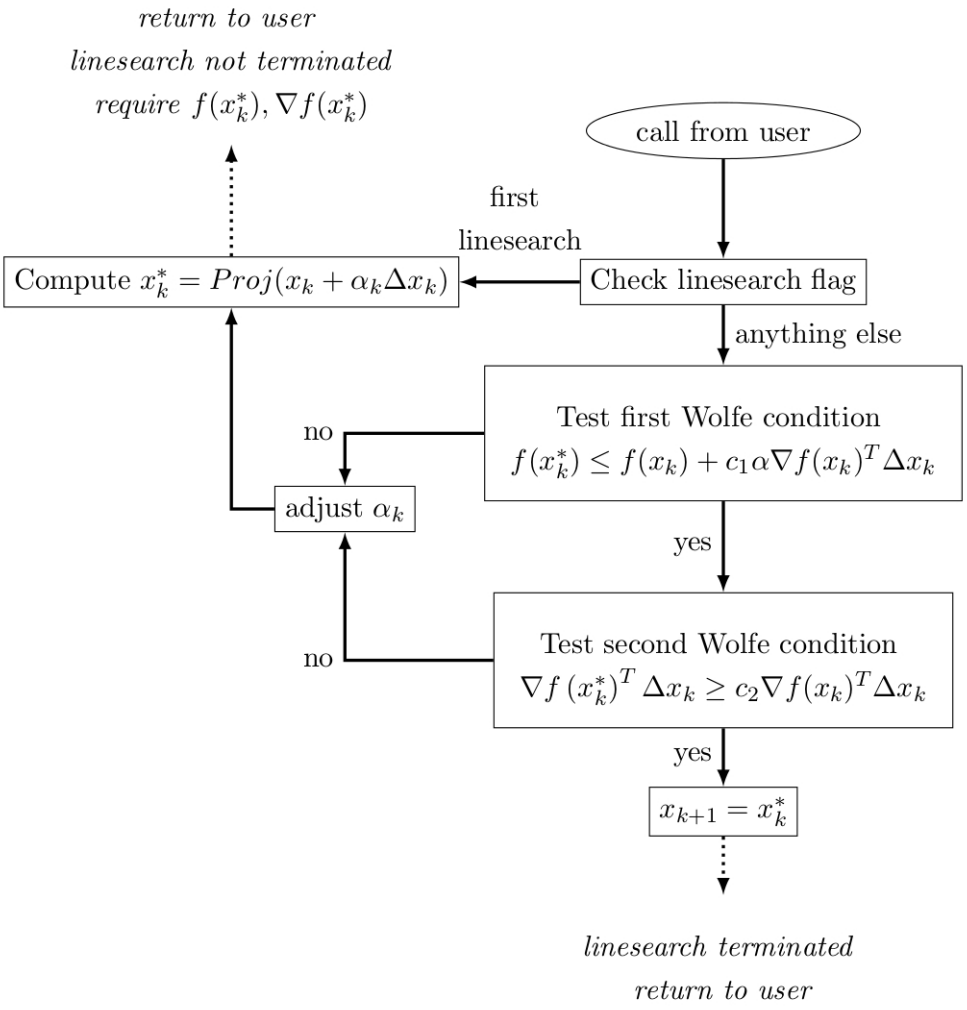

Figure 3: Schematic view of the linesearch implementation.

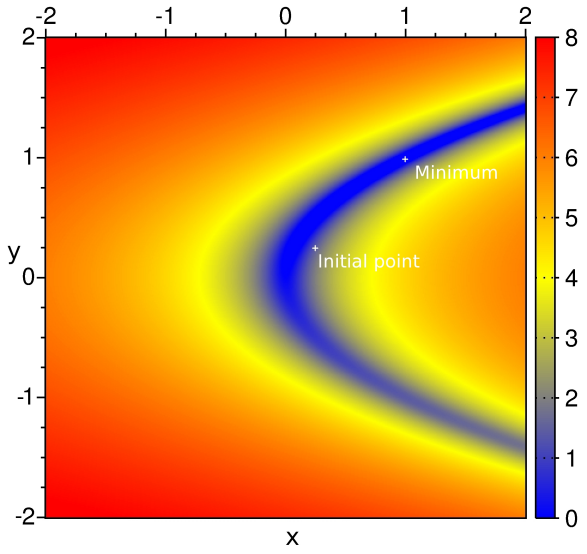

Figure 4: Bi-dimensional Rosenbrock function map. The two white crosses highlight the starting point for the optimization process located at $(0.25, 0.25)$ and the minimum of the Rosenbrock function, at the point $(1, 1)$. The valley of attraction appears as a broad blue channel where the function has a very flat minimum.

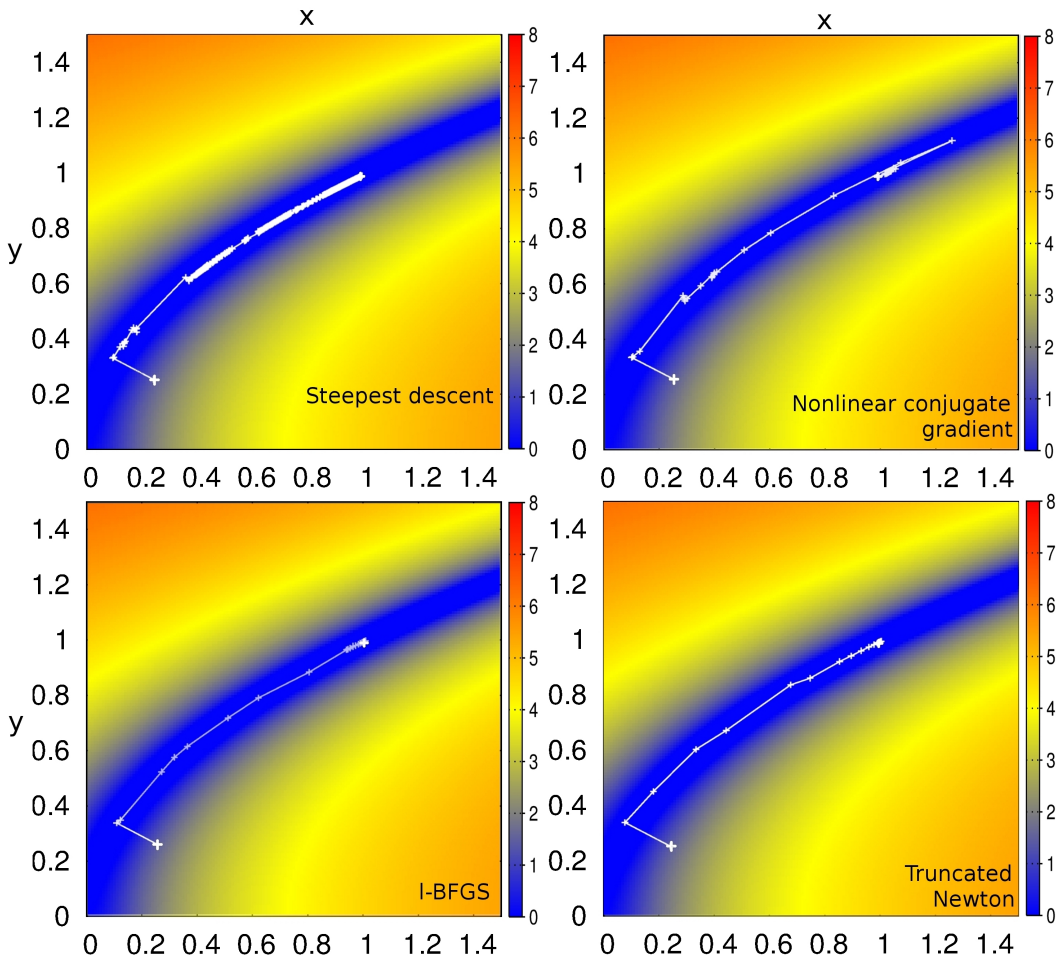

Figure 5: Convergence paths taken by the four different optimization methods: steepest descent (a), nonlinear conjugate gradient (b), l-BFGS (c), truncated Newton (d). The starting point is $(0.25, 0.25)$. The convergence is reached at $(1, 1)$.

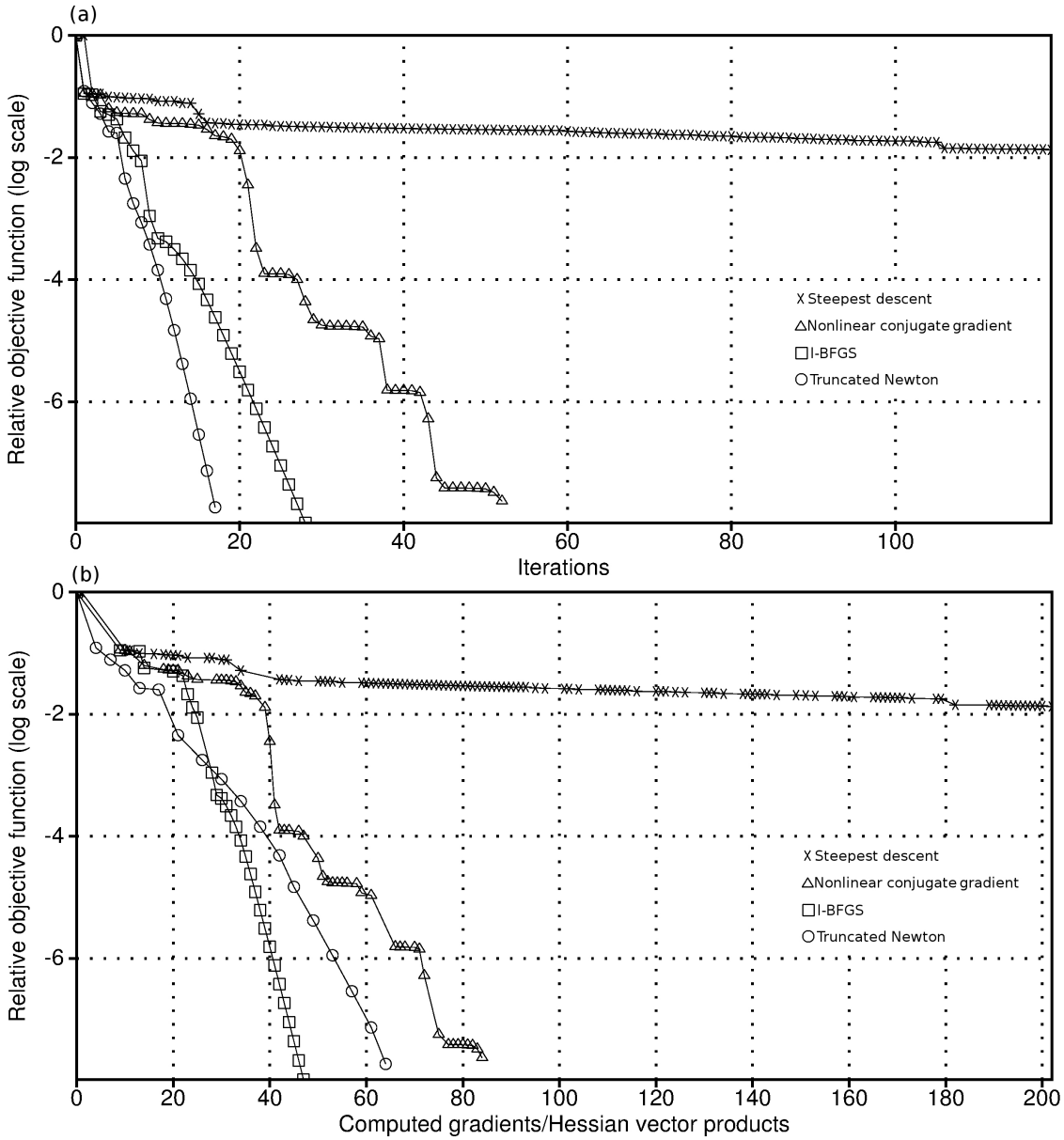

Figure 6: Rosenbrock function case study. Convergence rate of the four optimization methods in terms of iterations (a) and computational cost in terms of the number of gradient evaluations (b). For the truncated Newton method, the computation cost of a Hessian-vector product is assimilated to the computation cost of a gradient.

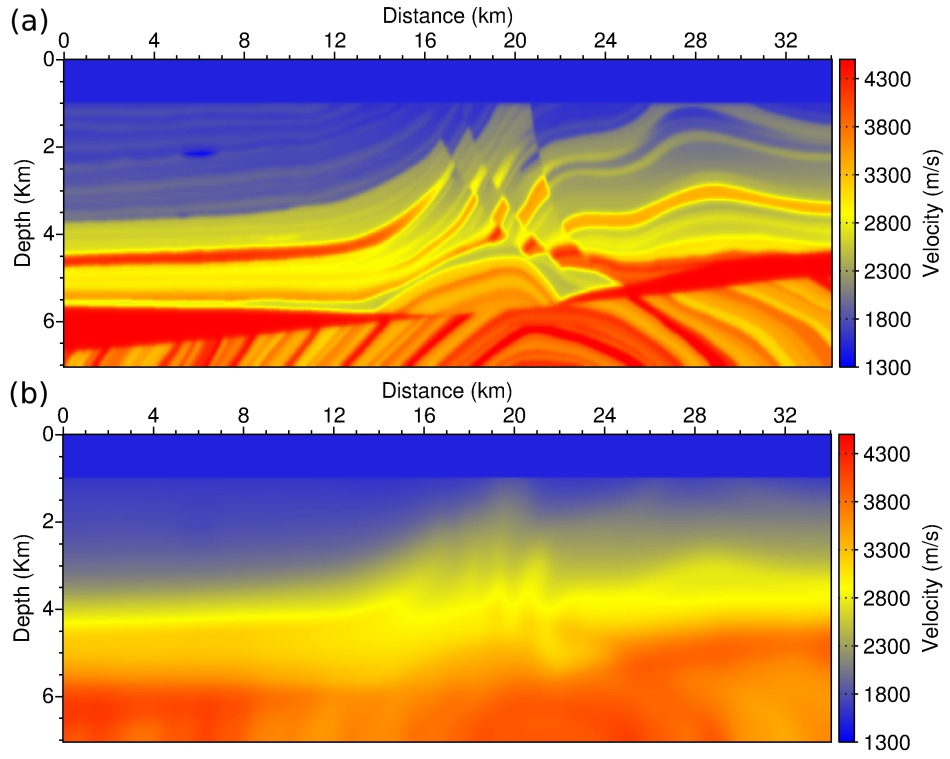

Figure 7: Marmousi 2 synthetic case study. Exact model (a). Initial model obtained after smoothing the exact model with a Gaussian filter (b). The correlation length for the Gaussian filter is set to 500 m. The number of discrete points is equal to 141×681, which yields a minimization problem involving approximately $96,000$ unknowns.

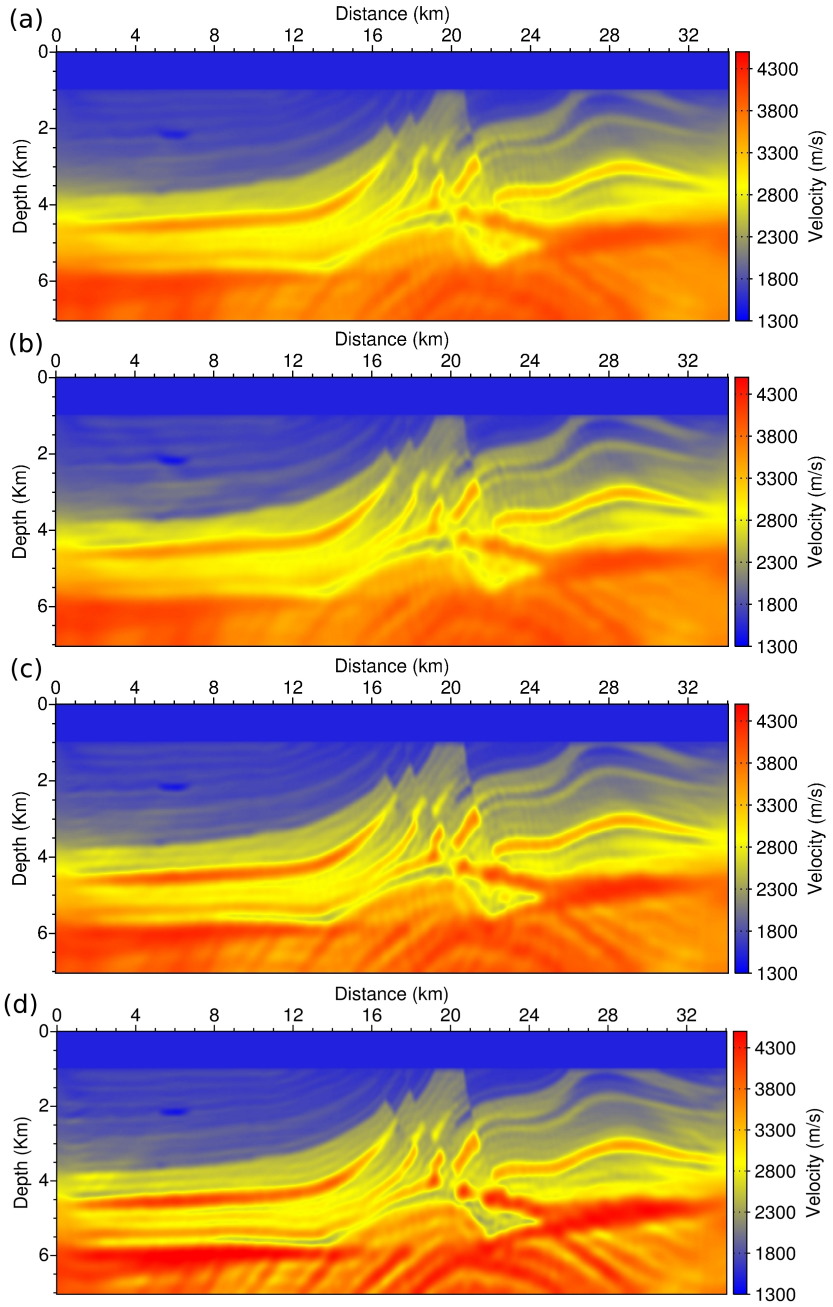

Figure 8: Marmousi case study. Models reconstructed by the four different optimization methods. Steepest-descent (a), nonlinear conjugate gradient (b), l-BFGS (c), truncated Newton (d).

Figure 9: Marmousi case study. Convergence curves of the four different optimization methods in terms of iterations (a), and computational cost in terms of the number of gradient evaluations (b). For the truncated Newton method, the computation cost of a Hessian-vector product is assimilated to the computation cost of a gradient. This is consistent with the implementation of second-order adjoint formulas for the computation of these quantities.