Inner product preconditioned trust-region methods for frequency-domain full waveform inversion

Xavier Adriaens, Ludovic Métivier, Christophe Geuzaine

Inner product preconditioned trust-region methods for frequency-domain full waveform inversion

Xavier Adriaens[a,1,*], Ludovic Métivier[b], Christophe Geuzaine[a]

[a] *Department of Electrical Engineering and Computer Science, Montefiore Institute B28, Université de Liège, Belgium*

[b] *Université Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France*

Abstract

Full waveform inversion is a seismic imaging method which requires solving a large-scale minimization problem, typically through local optimization techniques. Most local optimization methods can basically be built up from two choices: the update direction and the strategy to control its length. In the context of full waveform inversion, this strategy is very often a line search. We here propose to use instead a trust-region method, in combination with non-standard inner products which act as preconditioners. More specifically, a line search and several trust-region variants of the steepest descent, the limited memory BFGS algorithm and the inexact Newton method are presented and compared. A strong emphasis is given to the inner product choice. For example, its link with preconditioning the update direction and its implication in the trust-region constraint are highlighted. A first numerical test is performed on a 2D synthetic model then a second configuration, containing two close reflectors, is studied. The latter configuration is known to be challenging because of multiple reflections. Based on these two case studies, the importance of an appropriate inner product choice is highlighted and the best trust-region method is selected and compared to the line search method. In particular we were able to demonstrate that using an appropriate inner product greatly improves the convergence of all the presented methods and that inexact Newton methods should be combined with trust-region methods to increase their convergence speed.

Keywords: numerical optimization, large-scale inverse problems, trust-regions methods, operator preconditioning, seismic imaging, full waveform inversion.

Highlights

- Unified presentation and comparison of line search and trust-region globalization methods
- Innovative introduction of preconditioning through the inner product
- Comprehensive comparison of the steepest, the l-BFGS and the Newton descent directions
- First extensive comparison of their combinations for full waveform inversion based on two case studies

1. Introduction

Full waveform inversion is a high-resolution seismic imaging technique formulated as a data fitting problem, whose aim is to recover some model parameters by minimizing the discrepancy between recorded data and data simulated by solving wave propagation problems [43, 53, 55]. By nature these data are oscillatory and consequently the misfit quantifying the discrepancy features local minima [4, 30]. Global optimization techniques should ideally be used but the typically very high dimensions of the search space prohibits their use

*Corresponding author

Email addresses: xavier.adriaens@uliege.be (Xavier Adriaens), ludovic.metivier@univ-grenoble-alpes.fr (Ludovic Métivier), cgeuzaine@uliege.be (Christophe Geuzaine)

[1] X. Adriaens is a research fellow funded by the F.R.S-FNRS.

and only local optimization tools can practically be employed, with care [10]. A straightforward direction to iteratively update the model properties is of course the gradient, *i.e.* the direction of steepest decrease. However it is well-known that the inverse Hessian plays a crucial role in the reconstruction in addition to offering the possibility to account for coupling effects between parameter classes for multi-parameter inversion. A theoretically simple way to incorporate these second-order derivatives is to minimize the misfit using Newton methods. Preliminary studies on small synthetic test cases have shown the benefits of inverting the exact Hessian operator, for both mono-parameter [43] and multi-parameter [38, 33] inversions, but they have also highlighted the computational cost of such inversions. In practice, the pure Newton method is too computationally intensive to implement, specifically because it requires inverting the Hessian operator. In addition, the misfit in full waveform inversion is not quadratic, thus the exact Newton direction is not necessarily appropriate. Consequently, it is natural to turn to inexact Newton methods, where the search direction is constructed iteratively to approximate the pure Newton direction, or to quasi-Newton methods. State-of-the-art methods rely on the quasi-Newton l-BFGS algorithm, which implicitly builds an approximation of the inverse Hessian operator from l previously saved gradients and model parameters [8, 32, 50]. However it has been illustrated that on some specific cases involving multiple reflections, such quasi-Newton methods fail to provide an accurate reconstruction where inexact Newton methods do succeed [28]. The latter compute the descent direction through a few iterations of a linear system involving the Hessian operator (the Newton system). One of the first implementation of such methods in the context of full waveform inversion is due to [12], which solved the Gauss-Newton system, a positive-definite approximation of the Newton system, with the conjugate gradient algorithm. Since then, the approach has been extended to solve the indefinite Newton system and extensively compared to quasi-Newton methods, first in the frequency domain [28], then in the time domain [57]. The benefits of Newton approaches were finally confirmed through their application for multiparameter inversions on real data sets in 2D [25, 26]. One advantage over l-BFGS is the locality of the quadratic approximation: such methods do not rely on the convergence history of the algorithm, which might yield inaccurate inverse Hessian approximation for non quadratic misfit functions. The bottleneck of these methods lies in the compromise to find between a direction built in few iterations, but which hardly takes the Hessian into account and a nearly exact direction which is very expensive to compute. A complementary strategy to reduce this number of inner iteration is to apply a preconditioner to both sides of the Newton system. Most widely used preconditioners are approximations of the inverse Hessian operator, and more specifically approximations of the inverse Gauss-Newton Hessian operator: firstly because it is positive definite and secondly because it can be expressed analytically in terms of receiver-side and emitter-side Green's functions - emitter-side Green's function appear because the shot which generates the wavefield is often modeled by a point source, while receiver-side Green's functions appear in the context of the adjoint state method [1, 17, 40], which expresses the misfit derivatives in terms of an artificial wavefield back-propagated from the receivers, which are also considered as point sources. The simplest preconditioners are obtained when using analytical formulas for the Green's functions and keeping only the diagonal of the resulting operator [6, 41]. Computing instead the exact Green's function yields more accurate preconditioners, but at some extra cost, as the receiver-side Green's functions are typically not computed during the inversion [3]. For that specific reason, pseudo Hessian operators are also often used. Pseudo Hessian operators are constructed by slightly modifying the analytical expression of the exact Hessian operator, in order to reduce its computational burden [37, 41, 46]. These operators have also been used for multi-parameter inversions through diagonal [7, 34] or block diagonal preconditioners [21, 26, 54]. Alternatively, to avoid computing the receiver-side Green's functions, phase encoding methods can also be used [35]. Less conventional strategies have also been explored, such as a band diagonal Hessian approximation, which thus required solving a band diagonal system [20], or exploiting the preconditioning properties of a change of variable, *i.e.* a model reparametrization [2]. Finally, in the context of Newton methods, the l-BFGS inverse operator itself can be used as the preconditioner [36, 37].

To implement any of the three above mentioned schemes, *i.e.* the steepest descent, the l-BFGS method or a Newton method, one can rely either on line search algorithms, or on trust-region methods. In the former case, once a direction is chosen, the outer iteration is completed by finding the optimal length of the step that should be performed along that direction. Among the non linear optimization community, it is sometimes argued however that line search is not well suited with Newton directions, especially when the Hessian is nearly singular. Indeed when the Hessian is nearly singular, the Newton direction becomes excessively long such that the local quadratic approximation implicitly made when computing it ceases to hold. Much computational effort must then be made by the line search procedure to reduce the step size [32]. Stopping the iterative solution of the Newton system earlier appears as a solution to this problem. For example, its convergence requirements could be relaxed such that they reflect the accuracy of the local quadratic approximation [11, 25].

Alternatively, a trust-region method could be used instead [1, 24, 56, 58, 59]. The latter limits the length of the update direction depending on the accuracy of the local quadratic approximation. The length of a direction is given by its norm, itself induced by the inner product chosen for the model parameters space [50].

The choice of this inner product is thus pivotal in the implementation of a trust-region method [8, 22]. Moreover changing the inner product modifies both the gradient and the Hessian and is equivalent to applying a preconditioner [31]. Consequently it also has a major impact on line search based local optimization methods [9, 16, 23, 60]. It is important to highlight here that even though the gradient and the Hessian are modified by the inner product, the misfit to be minimized remains the same. Modifying the inner product or the misfit function are two distinct, complementary strategies to define non-standard update directions. Only the former is considered in this work.

In this paper, we tackle the three following important questions:

- Which descent direction to compute: the gradient, the l-BFGS direction or an inexact Newton direction?
- Which globalization method to select: a line search method or a trust-region method?
- Which preconditioning strategy to apply? How to enforce it?

Answering these three questions and determining the good combinations (good practices) between them is crucial for effective full waveform inversion. From our study, it appears that preconditioning is essential and that enforcing preconditioning through the inner product is elegant and, more interestingly, implies no modification to the practical implementation of the optimization algorithms. The l-BFGS method is found to be the most efficient method for the considered single-parameter inversions. It is also found to be insensitive to the globalization choice. Inexact Newton methods should not be discarded though, as considering the exact Hessian might lead to better model parameter decoupling in the case multi-parameter inversions. When using inexact Newton methods, our case studies show that using a trust region globalization consistently improves convergence.

The paper is organized as follows. In the first part, full waveform inversion is stated very generally. The optimization problem and its solution procedures using either a line search or a trust-region are introduced. The Newton system, which is pivotal in local minimization theory, is also derived. A particular emphasis is given to the inner product choice. More specifically, its link with preconditioning the Newton system is established. Local minimization methods commonly used in the context of full waveform inversion are then recalled. In the second part, the application to acoustic imaging is detailed. The (adjoint) procedure to compute gradients and Hessian vector products is given and its computational cost is explained. The overall computational cost of each optimization method is then deduced. Finally, convergence results on the acoustic Marmousi case study are analyzed to determine the best inner product and the best parameters for a trust-region method. This best candidate is then compared to line search methods on both the Marmousi model and on a case study involving strong reflectors.

2. Local optimization methods

Full wave inversion is an imaging method based on the minimization of a misfit functional J, which exclusively depends on some model parameters m. The recovered model parameters m^* are defined as the minimizer of this misfit, *i.e.* $m^* = \arg\min J(m)$. Local optimization techniques are based on a local quadratic expansion of the misfit J around the current model estimate

$$J(m+\delta m) \approx J(m) + \{D_m J\}(\delta m) + \frac{1}{2}\{D^2_{mm} J\}(\delta m, \delta m). \tag{1}$$

This expansion can also be written in terms of the gradient j' and the Hessian operator H once an inner product $\langle \cdot, \cdot \rangle_M$ is chosen for the model space M

$$J(m+\delta m) \approx J(m) + \langle j', \delta m \rangle_M + \frac{1}{2} \langle H\delta m, \delta m \rangle_M . \tag{2}$$

The pure Newton direction p_N is then defined as the minimizer of this local quadratic expansion, which is also the solution of a linear system

$$p_N = \underset{p \in M}{\arg\min}\, J(m) + \langle j', p \rangle_M + \frac{1}{2} \langle Hp, p \rangle_M \quad \text{or} \quad Hp_N = -j'. \tag{3}$$

The large-scale nature of this linear system requires either the use of approximate Hessian operators that are straightforward to invert, or the use of Hessian-free iterative methods. Both approaches are usually referred to as quasi-Newton methods and inexact Newton methods. In the latter case, the conjugate gradient method is the ideal candidate for the iterative solver because the Hessian operator is symmetric. The conjugate gradient method is however designed for positive definite operators while the full Hessian can be indefinite, especially far from the global minimum [28, 43]. As a consequence, either an additional safeguard is added to exit prematurely when directions of negative curvature are encountered or the exact Hessian is modified such that it becomes positive definite, *e.g.* using the Gauss-Newton approximation [37]. Loops appearing inside the computation of this descent direction, *e.g.* the conjugate gradient algorithm, are referred to as inner loops, in contrast to the optimization loop that updates the model for a given direction at each iteration, which is referred to as the outer loop. In the sequel, quantities computed at the n^{th} outer iteration will be denoted with the subscript 'n', while the subscript 'k' is rather used for inner iterations.

2.1. Globalization methods

As mentioned in the introduction, the misfit is not quadratic and thus the pure Newton direction or its approximations are not always the best directions. For that reason the length of the search direction is often tweaked using a line search or a trust-region method, which ensures convergence towards the nearest stationary point [8, 11, 15, 14, 32]. However, the meaning of 'nearest' depends again on the metric and could thus differ depending on the inner product choice [8, 50].

2.1.1. Line search

When using a line search procedure, a direction p must first be identified. An appropriate length γ is then given to this direction p, ideally the global minimum along the line $m+\gamma p$. In practice however less stringent satisfactory conditions are used instead to spare expensive wave problem resolutions. The most widely used examples are strong Wolfe conditions

$$J(m+\gamma p) \leq J(m) + c_1\gamma\{D_m J(m)\}(p) \tag{4}$$

$$|\{D_m J(m+\gamma p)\}(p)| \leq c_2\,|\{D_m J(m)\}(p)| \tag{5}$$

for some constant c_1 and c_2 such that $0 < c_1 < c_2 < 1$. The first condition is called the sufficient decrease condition. It ensures that updating the model in the direction γp produces a decrease smaller than a fraction c_1 of what is expected from a local linear approximation of the misfit. The second condition, called the curvature condition, ensures that the updated model $m+\gamma p$ is sufficiently close to a local minimum along the line, where the directional derivative $\{D_m J(m+\gamma p)\}(p)$ would be zero. When this derivative is very smaller (resp. larger) than zero, then a larger (resp. smaller) step could produce a significantly bigger decrease. We choose here a line search algorithm that satisfies strong Wolfe conditions and accepts steps easily (Algorithm 3.2 from [32] with $c_1 = 10^{-4}$, $c_2 = 0.9$). The outer loop is finally obtained by repeating these two steps iteratively until convergence.

2.1.2. Trust region

At the opposite when using a trust-region method, first a maximum length Δ is chosen. Then the best approximate solution, meaning the direction that minimizes a local prediction of the misfit but smaller than this length, is used

$$p = \underset{p\in M, \|p\|_M \leq \Delta}{\arg\min} \left[J^{\text{pred}}(m;p) := J(m) + \langle j'(m), p\rangle_M + 0.5 \left\langle \tilde{H}(m)p, p \right\rangle_M \right]. \tag{6}$$

This local misfit prediction J^{pred} is typically constructed based on the local quadratic approximation (2) through a particular choice of some approximate Hessian operator $\tilde{H}$. Of course the approximate Newton direction $\tilde{H}p = -j'$ is the solution of this problem if it lies inside the trust region. There are several possibilities to choose this length Δ and our particular choice is detailed later. More importantly, as we pointed out in the introduction, the length constraint is formulated in terms of the norm induced by the inner product $\|p\|_M^2 = \langle p, p\rangle_M \leq \Delta^2$. Modifying this inner product therefore changes the shape of the trust region and it is then desirable to choose it carefully [8]. The size of the trust region is actually controlled by the outer iterations. The decision of modifying the trust region is based on the accuracy of the local prediction of the

misfit. When the prediction is accurate but the updates are limited by the length constraint, then the trust region radius is increased. At the opposite, when the updates are out of the range of validity of the prediction, then the trust region radius is decreased. The decrease (resp. increase) rate of the radius is controlled by some parameter $c_0 < 1$ (resp. $c_1 > 1$). The quality of the prediction is quantified by the ratio between the actual decrease $\delta J_{\mathrm{a}} := J(m_n) - J(m_{n+1})$ and the decrease predicted by the local prediction of the misfit. There are two ways to compute this predicted decrease [14]. On the one hand the expansion can be written in terms of the gradient and the Hessian operator at the previous model estimate

$$J(m_{n+1}) = J(m_n + p_n) \tag{7}$$

$$\approx J(m_n) + \langle j'(m_n), p_n \rangle_M + 0.5 \left\langle \tilde{H}(m_n) p_n, p_n \right\rangle_M = J^{\mathrm{pred}}(m_n; p_n) \tag{8}$$

which defines the prospective predicted decrease

$$\delta J_{\mathrm{p,p}} := J(m_n) - J^{\mathrm{pred}}(m_n; p_n) \tag{9}$$

$$= - \langle j'(m_n), p_n \rangle_M - 0.5 \left\langle \tilde{H}(m_n) p_n, p_n \right\rangle_M . \tag{10}$$

On the other hand, it can also be written in terms of the gradient and the Hessian operator at the next model estimate

$$J(m_n) = J(m_{n+1} - p_n)$$

$$\approx J(m_{n+1}) - \langle j'(m_{n+1}), p_n \rangle_M + 0.5 \left\langle \tilde{H}(m_{n+1}) p_n, p_n \right\rangle_M = J^{\mathrm{pred}}(m_{n+1}; -p_n)$$

which defines the retrospective predicted decrease

$$\delta J_{\mathrm{p,r}} := J^{\mathrm{pred}}(m_{n+1}; -p_n) - J(m_{n+1}) \tag{11}$$

$$= - \langle j'(m_{n+1}), p_n \rangle_M + 0.5 \left\langle \tilde{H}(m_{n+1}) p_n, p_n \right\rangle_M . \tag{12}$$

These ratios between the actual decrease and one of both the predicted decreases $\rho_{\mathrm{p}} := \delta J_{\mathrm{a}} / \delta J_{\mathrm{p,p}}$ and $\rho_{\mathrm{r}} := \delta J_{\mathrm{a}} / \delta J_{\mathrm{p,r}}$ are actually both equal to one when the approximate Hessian in the update direction and the second order expansion (2) are exact. When the misfit is not quadratic or the Hessian approximation is not accurate, then these ratios can go away from one. Using anything else than the full Newton method can degrade these ratios, even if the misfit is quadratic. In particular for a pure quadratic misfit, neglecting the negative definite part of the Hessian makes the prospective ratio bigger than one ($\delta J_{\mathrm{p,p}}$ is underestimated) and the retrospective ratio smaller than one ($\delta J_{\mathrm{p,r}}$ is overestimated). Standard trust-region methods directly control the radius Δ. However it is an absolute quantity, in the sense that it is compared to $\|p\|_M$, which depends on the inner product. Thus, it seems more natural to control this radius relatively to the gradient norm ($\Delta = \mu \|j'\|_M$), which provides a length reference for the (approximate) Newton system. In this way, even when the (approximate) Newton system changes scale from one iteration to another, the trust region remains relevant. This particular variant (Algorithm 1) has been first introduced in [15]. According to this algorithm, a direction p_n is rejected when the prospective misfit prediction J_n^{pred} used to compute it is not accurate, in the sense that the prospective ratio is smaller than some threshold ρ_0. If not rejected, then the trust region size is updated according to either the prospective or the retrospective ratio, based on a comparison with a second threshold ρ_1. Because the updated radius Δ_{n+1} constrains the direction search around the next model estimate m_{n+1}, it makes sense to use the retrospective ratio which also involves the next model estimate m_{n+1} and not the prospective ratio which involves the current model estimate m_n. Using the retrospective ratio is however slightly more expensive because the next (approximate) Hessian operator in the current direction must be computed in addition. Moreover the accuracy of the retrospective prediction might be good in the direction $-p_n$ while still being bad in the direction p_{n+1} and inversely. There is also no safeguards for large value of the ratios, which means that when the model is not accurate but the predicted decrease underestimates the true decrease, the radius can still be increased.

Three sets of values for the threshold ρ_1 and the rates c_0/c_1 have been tested. The acceptance threshold ρ_0 is always tiny such that steps are often accepted, similarly to the line search algorithm.

$$\text{(A)} \;\; \rho_0 = 10^{-4},\, \rho_1 = 0.25 \quad \text{and} \quad c_0 = 0.20,\, c_1 = 5.$$

$$\text{(B)}\ \ \rho_0 = 10^{-4},\ \rho_1 = 0.75 \quad \text{and} \quad c_0 = 0.25,\ c_1 = 2.$$
$$\text{(C)}\ \ \rho_0 = 10^{-4},\ \rho_1 = 0.90 \quad \text{and} \quad c_0 = 0.50,\ c_1 = 2.$$

The first one (A) is very similar to what was originally proposed in [14]. The other two (B,C) are more cautious because they modify the radius more rarely and when they do, it increases by a smaller factor. Note that the second one (B) is also close to what is proposed in [32].

Algorithm 1 Fan trust-region algorithm

Require: retrospective or prospective, $0 \leq \rho_0 < \rho_1 < 1$ and $0 < c_0 < 1 < c_1$
  $\mu_0 = 1$
  loop
    $\Delta_n = \mu_n \left\| j'(m_n) \right\|_M$
    $p_n = \begin{cases} -\mu_n j'_n \\ (28) \text{ with } \Delta = \Delta_n \\ \text{Algorithm 5 with } \Delta = \Delta_n \end{cases}$
    $\delta J_{\mathrm{a}} = J(m_n) - J(m_n + p_n)$ and $\delta J_{\mathrm{p,p}} = J(m_n) - J^{\mathrm{pred}}(m_n; p_n)$
    $\rho_{\mathrm{p}} = \delta J_{\mathrm{a}} / \delta J_{\mathrm{p,p}}$
    if $\rho_{\mathrm{p}} \geq \rho_0$ **then** $m_{n+1} = m_n + p_n$ **else** $m_{n+1} = m_n$
    if prospective **or** $\rho_{\mathrm{p}} < \rho_0$ **then**
      $\rho = \rho_{\mathrm{p}}$
    else if retrospective **then**
      $\delta J_{\mathrm{p,r}} = J^{\mathrm{pred}}(m_{n+1}; -p_n) - J(m_{n+1})$
      $\rho = \rho_{\mathrm{r}} = \delta J_{\mathrm{a}} / \delta J_{\mathrm{p,r}}$
    end if
    if $\rho < \rho_1$ **then** $\mu_{n+1} = c_0 \mu_n$
    else if $\rho \geq \rho_1$ **and** $\left\| p_n \right\|_M > 0.5 \Delta_n$ **then** $\mu_{n+1} = c_1 \mu_n$
    else **then** $\mu_{n+1} = \mu_n$
  end loop

2.2. Inner product

The choice of the inner product plays a central role in the inversion because it defines through the norm how directions length are measured but also because it defines both gradients and Hessian operators. Indeed the equivalence between both expansions (1) and (2) is granted by the defining property of the gradient and the Hessian operator in terms of directional derivatives

$$\langle j', \delta m_1 \rangle_M := \{D_m J\}(\delta m_1) \qquad \forall \delta m_1, \tag{13}$$
$$\langle H \delta m_2, \delta m_1 \rangle_M := \left\{D^2_{mm} J\right\}(\delta m_1, \delta m_2) \qquad \forall \delta m_1,\ \delta m_2. \tag{14}$$

This link between directional derivatives and kernels is actually a straightforward application of the Fréchet-Riesz representation theorem [19].

The model parameter space is a function space defined on some region Ω and conventionally, the inner product is chosen as the $L_2(\Omega)$ inner product

$$\langle m_2, m_1 \rangle_M = \langle m_2, m_1 \rangle := \int_\Omega m_1(\boldsymbol{x}) m_2(\boldsymbol{x})\ d\Omega. \tag{15}$$

This straightforward choice leads to the conventional gradient j'_{L_2} and the conventional Hessian operator H_{L_2}, that can both be computed efficiently using the adjoint state method [1, 17, 40]. As an illustration, a conventional gradient is represented in Fig. 1b. It is actually the first gradient computed during the acoustic imaging of the Marmousi model. This case study is described in detail in subsection 3.1. As can be seen, shallow contributions have much greater amplitudes than deeper parts [29, 35, 41, 42, 43, 46]. This actually reflects the bad scaling properties of this inner product and motivates the use of a spatially weighted inner product

$$\langle m_2, m_1 \rangle_M := \left\langle m_2 \sqrt{w}, \sqrt{w}\ m_1 \right\rangle, \tag{16}$$

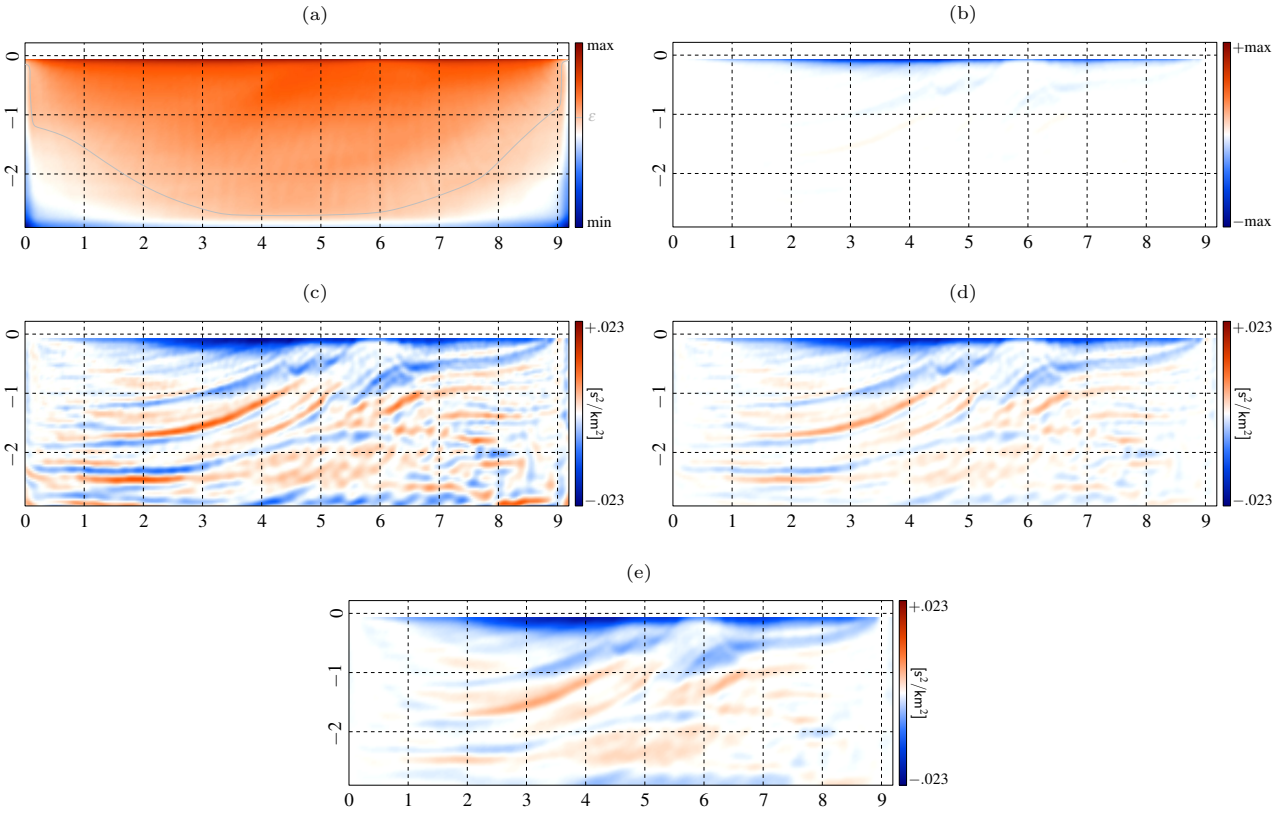

Figure 1: Diagonal part of the Gauss-Newton Hessian (a). Conventional gradient (b). Weighted gradient (c). Weighted and thresholded gradient (d). Weighted and smoothed gradient ($2\pi l_c = 0.250$ [km])(e). The level curve whose value is the stabilization parameter ϵ is given graphically in the top figure (▬). All quantities are computed for the initial Marmousi model Fig. 2b. The case study is described in detail in subsection 3.1.

with an appropriate spatially dependant weight w. Insights on how to design w can be gained by relating the conventional and the weighted gradients. Indeed, both are defined by (13) then by transitivity of the equality

$$\left\langle j'_{L_2}, \delta m_1 \right\rangle = \left\langle j' \sqrt{w}, \sqrt{w}\ \delta m_1 \right\rangle \quad \forall \delta m_1 \quad \text{such that} \quad j' = w^{-1} j'_{L_2}. \tag{17}$$

The same reasoning can be applied to both Hessian operators ($H = w^{-1} H_{L_2}$). Choosing this weight close to the Hessian operator then makes the gradient closer to the pure Newton direction and the Hessian operator closer to the identity. In other words, the Newton system (3) is better conditioned and iterative solvers are therefore expected to converge faster. We choose here to take this weight as the diagonal part of the Gauss-Newton Hessian ($w = \mathrm{diag}\,(H_{\mathrm{GN}})$) because it can be computed semi-analytically for a given model at no extra computational cost under certain circumstances [37]. A weight that has the same units as the Hessian also has the advantage that the corresponding weighted gradient has the same units than the model parameters. Model parameters, weighted gradients and weighted Hessian vector products therefore all have the units of model parameters and the coefficients between them, for example the length γ and μ involved respectively in line search and trust region techniques, are then always dimensionless and thus easier to interpret. The weights and the corresponding weighted gradient are given in Fig. 1a and 1c respectively. As expected, the weighted inner product compensates for the geometrical spreading and restores balance between shallow and deep contributions. It is however dangerous to use this weight alone because it can be very close to zero in poorly illuminated zones as for example in the corners of the model. In these regions, the weighted inner product is insensitive and consequently the preconditioner is unstable. The simplest stabilization strategy consists in the introduction of a threshold ϵ in the weights

$$\langle m_2, m_1 \rangle_M := \left\langle m_2\ \sqrt{w}, \sqrt{w}\ m_1 \right\rangle + \epsilon \left\langle m_2, m_1 \right\rangle. \tag{18}$$

The corresponding preconditioning effect is to keep silent some regions, where the weight is much smaller than

the threshold. Another strategy is to use an inner product with the following stabilization term

$$\langle m_2, m_1 \rangle_M := \langle m_2 \sqrt{w}, \sqrt{w}\, m_1 \rangle + \epsilon\, l_c^2\, \langle \boldsymbol{\nabla} m_2, \boldsymbol{\nabla} m_1 \rangle \tag{19}$$

where l_c is a characteristic length. This second term, related to spatial derivation, increases the norm of directions that are rapidly varying and also prevents the inner product from being insensitive in regions where the diagonal Hessian is close to zero. In regions where the diagonal Hessian is close to the threshold, then directions with details smaller than the characteristic length l_c are penalized with respect to smoother directions. This inner product is actually very similar to the one introduced in [60], except that the Gauss-Newton diagonal Hessian weight is used in addition. As far as preconditioning is concerned, this inner product can be reformulated through an integration by parts as

$$\langle m_2, m_1 \rangle_M := \langle w\ m_2, m_1 \rangle - \epsilon\, l_c^2\, \langle \Delta m_2, m_1 \rangle\,. \tag{20}$$

Then as previously, conventional and preconditioned gradients are linked

$$\langle j', \delta m_1 \rangle_M = \langle j'_{L_2}, \delta m_1 \rangle \quad \forall \delta m_1 \tag{21}$$

$$\left\langle \left(w - \epsilon\, l_c^2 \Delta\right) j', \delta m_1 \right\rangle = \langle j'_{L_2}, \delta m_1 \rangle \quad \forall \delta m_1 \qquad \Leftrightarrow \quad j' = \left(w - \epsilon\, l_c^2 \Delta\right)^{-1} j'_{L_2}. \tag{22}$$

From the point of view of preconditioning, this inner product generates a rescaling thanks to the Gauss-Newton diagonal Hessian weight and a Laplacian filtering, whose smoothing length equals $2\pi l_c$ where the diagonal Hessian equals the threshold. The effect of these inner products is illustrated in Fig. 1d and 1e. In addition of stabilizing the weights, [60] have shown that a filtering inner product can help the convergence of full waveform inversion by mitigating its non linearity.

In general, any inner product that can be related to the conventional inner product (15) through some preconditioner P yields a preconditioned gradient and a preconditioned Hessian operator

$$\langle m_2, m_1 \rangle_M = \langle P m_2, m_1 \rangle \quad \Rightarrow \quad j' = P^{-1} j'_{L_2} \text{ and } H = P^{-1} H_{L_2}. \tag{23}$$

Changing the inner product is formally equivalent to preconditioning both the gradient and the Hessian operator. We choose to introduce preconditioning through a change in the inner product rather than through the application of an operator because it appears more elegant and rigorous to us. Moreover, this approach has the pedagogical advantage to include preconditioning inside the inner product choice and thus it does not need to appear explicitly in the description of the optimization algorithms. In terms of practical implementation, it implies that the optimization routines need not be rewritten, only the subroutine which computes the inner product has to be modified, hence providing a lot of flexibility. Basically, a different choice for the inner product does not modify the pure Newton direction because the same preconditioner is applied to both sides of the Newton system (3), but does modify the subspace constructed by the conjugate gradient method and does modify norms which are involved in any stopping criterion. A good choice can thus lead to better approximate directions and better truncation rules.

2.3. Steepest descent

The steepest descent is actually the simplest local optimization algorithm. It consists in taking the search direction as the opposite gradient. This is the best direction at first order ($\tilde{H} = 0$) but it can also be seen as a quasi-Newton step where the approximate Hessian operator is the identity operator ($\tilde{H} = I$). In practice however, this approximation is very crude because the Hessian operator is far from the identity operator, even after preconditioning. The downside of this simple method is its linear convergence rate. This slow convergence speed is one of the main motivation for the investigation of higher order algorithms.

2.3.1. Line search globalization

No length information can be captured from the approximate Hessian operator in this case, because it is simply the identity operator ($\tilde{H} = I$). The first trial step length is then chosen based on the history of the outer iterations to save as many step length trials as possible *e.g.* $\gamma = 2(J(m_n) - J(m_{n-1}))/\{D_m J\}(-j')$ [32].

2.3.2. Trust region globalization

Trust-region methods are barely used with steepest descent. Mostly because the linear misfit prediction

$$J^{\text{pred}}(m;p) := J(m) + \langle j'(m), p\rangle_M \tag{24}$$

is not accurate enough. Moreover the solution to the trust-region sub-problem (6) is trivially $p = -\mu j'$ and is always on the boundary, because of the absence of a second order term. An upper bound on the relative size of the trust region (μ) is then added to compensate the fact that the trust-region algorithm will never keep it constant. This bound is set to $\mu_{\max} = 4,\ 4,\ 5$ for parameter sets A, B, C respectively.

2.4. Limited memory BFGS method

Quasi-Newton methods are expected to provide a huge improvement over the steepest descent and an attractive alternative to Newton methods because they do not involve any expensive Hessian vector product. In place of the exact Hessian, an approximation $\tilde{H} = B$ is used instead. This approximation is built only with the successive gradients and model parameters of each iteration. Moreover, since expensive Hessian vector product are avoided, quasi-Newton methods are sometimes more efficient than Newton methods. The Broyden-Fletcher-Goldfarb-Shanno algorithm, abbreviated BFGS, is maybe the most widely used quasi-Newton method. This method constructs a symmetric and positive definite approximation of the Hessian operator based on all the previous gradients and model parameters. This approximation B_{n+1} is chosen such that it verifies the secant equation

$$B_{n+1}\Delta m_n = \Delta j'_n \quad \text{with } \Delta m_n = m_{n+1} - m_n \text{ and } \Delta j'_n = j'_{n+1} - j'_n \tag{25}$$

while being close to the previous approximation B_n and positive definite. Note that imposing the positive definiteness of this approximation also imposes that the update direction must satisfy the (BFGS) curvature condition $\langle \Delta m_n, \Delta j'_n\rangle_M > 0$. One of the biggest advantage of the BFGS algorithm is that it is possible to directly build the approximate inverse Hessian operator B_n^{-1} from the memorized gradients and model parameters. However building explicitly this inverse operator in the context of large-scale optimization is still prohibitively expensive, as well as storing in memory all the previous gradients and model parameters. For these reasons, a limited memory version of the algorithm has been derived. Instead of memorizing all the previous iterates, it only requires the l last iterates and above all, it comes with a two-loop recursive procedure to compute the application of the inverse operator on any direction. The approximate Newton direction associated with the l-BFGS operator is therefore straightforward to compute. This two-loop recursive l-BFGS algorithm is given in Algorithm 2. A constant initial Hessian approximation, *i.e.* $B_n^0 = \left\langle \Delta m_{n-1}, \Delta j'_{n-1}\right\rangle_M / \left\langle \Delta j'_{n-1}, \Delta j'_{n-1}\right\rangle_M$, is here chosen [8, 32].

Algorithm 2
Inverse l-BFGS operator application

Require: $q,\ \Delta m_k,\ \Delta j'_k, \quad \forall k \in [n-l, n-1]$
  for $k = n-1$ **down to** $k = n-l$ **do**
    $\alpha_k = \langle \Delta m_k, q\rangle_M / \langle \Delta j'_k, \Delta m_k\rangle_M$
    $q = q - \alpha_k \Delta j'_k$
  end for
  $r = B_n^0\, q$
  for $k = n-l$ **up to** $k = n-1$ **do**
    $\beta_k = \langle \Delta j'_k, r\rangle_M / \langle \Delta j'_k, \Delta m_k\rangle_M$
    $r = r + (\alpha_k - \beta_k)\Delta m_k$
  end for
  return $r\ \left(= B_n^{-1} q\right)$

Algorithm 3
Direct l-BFGS operator application

Require: $q,\ \Delta m_k,\ \Delta j'_k, \quad \forall k \in [n-l, n-1]$
  for $k = n-l$ **up to** $k = n-1$ **do**
    $b_k = \Delta j'_k / \sqrt{\langle \Delta j'_k, \Delta m_k\rangle_M}$
    $a_k = B_n^0 \Delta m_k$
    for $i = n-l$ **up to** $i = k-1$ **do**
      $a_k = a_k + \langle b_i, \Delta m_k\rangle\, b_i - \langle a_i, \Delta m_k\rangle\, a_i$
    end for
    $a_k = a_k / \sqrt{\langle \Delta m_k, a_k\rangle_M}$
  end for
  $r = B_n^0\, q$
  for $k = n-l$ **up to** $k = n-1$ **do**
    $r = r + b_k \langle b_k, q\rangle_M - a_k \langle a_k, q\rangle_M$
  end for
  return $r\ (= B_n\, q)$

It is important to highlight here that this method also benefits from the modification of the inner product. Indeed the building blocks of this approximate Hessian operator are the successive gradients, which are preconditioned through the inner product. By measuring gradient variations, this method constructs a representation

of the misfit which is good enough to produce super-linear convergence, a great improvement over the steepest descent, at no extra cost. This approximation is however positive definite while the exact Hessian might be indefinite, especially during the early iteration of the inversion. In such cases, this quasi-Newton method may fail to provide an accurate reconstruction while Newton methods may succeed [28].

2.4.1. Line search globalization

The unit step length $\gamma = 1$ is always tried first because the length information should be captured by the inverse approximate Hessian. Importantly, it can be shown that the (BFGS) curvature condition is always satisfied if the strong Wolfe conditions (4) and (5) are enforced [32]. Therefore the l-BFGS algorithm combined with a line search will always construct a positive definite approximate Hessian operator B.

2.4.2. Trust region globalization

Finding the exact solution to the trust-region sub-problem (6) with the l-BFGS predicted misfit

$$J^{\text{pred}}(m;p) := J(m) + \langle j'(m), p\rangle_M + 0.5\,\langle Bp, p\rangle_M \tag{26}$$

is difficult for a general trust region radius. However when this radius is large enough, in particular larger than the unconstrained solution $p^{\text{u}} := -B^{-1}j'$, then it is actually also the exact constrained solution. On the other hand, when the radius is small enough, the quadratic term in the misfit prediction is negligible and the sub-problem is equivalent to the steepest descent, which indicates following the gradient up to the boundary. Based on these solutions for the extreme value of the radius, the exact solution to the sub-problem (6) might be substituted by an interpolation between these two solutions. Namely, the gradient is followed each time the minimum of the misfit prediction along the gradient, *i.e.* the Cauchy point $p^{\text{c}} = -\alpha j'$ (with $\alpha = \langle j', j'\rangle_M \,/\, \langle Bj', j'\rangle_M$), is outside the radius. Then for intermediate radii, which contains this Cauchy point but not the unconstrained solution, an interpolation between both is done

$$p(\Delta) = p^{\text{c}} + \tau^* \left(p^{\text{u}} - p^{\text{c}}\right) \text{ with } 0 < \tau^* < 1 \text{ such that } \|p\|_M = \Delta. \tag{27}$$

Finally for large radii, the unconstrained solution is accepted. In summary

$$p(\Delta) = \begin{cases} p^{\text{u}} & \text{when } \|p^{\text{u}}\|_M \leq \Delta, \\ -\mu j' & \text{when } \|p^{\text{c}}\|_M \geq \Delta, \\ p^{\text{c}} + \tau^*(p^{\text{u}} - p^{\text{c}}) & \text{when } \|p^{\text{c}}\|_M \leq \Delta \leq \|p^{\text{u}}\|_M . \end{cases} \tag{28}$$

The approximate solution (28) to the trust-region sub-problem (6) is called the dogleg method [8, 32].

A huge difference with the line search implementation of the l-BFGS algorithm is that now the direct application of the approximate Hessian operator B on some directions must be computed. Unfortunately there is no equivalent to Algorithm 2 for the direct l-BFGS operator and its application must then be computed from its recursive definition at iteration n

$$B_i q = B_n^0 q + \sum_{k=n-l}^{i-1} b_k \langle b_k, q\rangle_M - a_k \langle a_k, q\rangle_M \tag{29}$$

$$\text{with} \quad a_k = \frac{B_k \Delta m_k}{\sqrt{\langle B_k \Delta m_k, \Delta m_k\rangle_M}} \quad \text{and} \quad b_k = \frac{\Delta j'_k}{\sqrt{\langle \Delta j'_k, \Delta s_k\rangle_M}}. \tag{30}$$

It is important to highlight that the sequence of directions a_k could not be memorized because at each iterations the oldest information is discarded, which modifies the whole a_k sequence. A complete procedure to compute the application of the direct l-BFGS operator is given in Algorithm 3. Faster but more sophisticated procedure do exist [32]. However, manipulations in the model parameter space are computationally negligible with respect to wave propagation problems hence the speedup would also be negligible. Thanks to this procedure the prospective and retrospective predicted decrease (10) and (12) can be evaluated. Interestingly, the prospective decrease is evaluated with the current Hessian approximation B_n while the retrospective decrease is evaluated with the next Hessian approximation B_{n+1}. The retrospective ratio is therefore expected to be more often close to one because this next Hessian approximation B_{n+1} is specifically constructed from the update direction $p_n = \Delta m_n = m_{n+1} - m_n$.

2.5. Newton methods

In contrast to quasi-Newton methods, Newton methods use the Hessian operator explicitly, as they try to solve the Newton system (3). The interest of these method lies in their independence on the convergence history and in their quadratic convergence rate in the vicinity of the minimum. Far from this minimum, the Hessian operator might however be indefinite, which complicates the solution procedure for the Newton system. For that reason, it is frequent to make the Gauss-Newton approximation ($\tilde{H} = H_{\mathrm{GN}}$), which consist in keeping only the positive definite part of the Hessian operator. The downside of this approximation is then that the second order representation (2) of the misfit is less accurate, especially if the negative definite part of the Hessian is not negligible, which might prevent the method from reaching an accurate reconstruction. In this section, we present inexact Newton methods based on a line search procedure or a trust region method. Both are valid for the full Hessian and for its Gauss-Newton approximation.

2.5.1. Line search globalization

Newton methods can be combined with a line search procedure. In this case a direction p is first found by solving the Newton system approximately with the conventional conjugate gradient method (Algorithm 4) [32]. This algorithm constructs iteratively the solution of a linear system without requiring the explicit expression of the Hessian matrix but only its action in particular directions. The iterative procedure is stopped when the residuals have decreased more than some threshold, called the forcing sequence η, which is typically close to zero

$$(\|r_k\|_M :=) \|Hp_k + j'\|_M < \eta \|j'\|_M (= \eta \|r_0\|_M) . \tag{31}$$

Over-solving is here avoided through this forcing term η, which is not systematically close to zero but which is instead chosen to reflect the accuracy of the second-order expansion. Three possible choices for this sequence have been described and studied by [11]. These three choices were then compared in the context of acoustic imaging in [28], who advise to use the forcing sequence

$$\eta_n = \frac{\|j'(m_n) - j'(m_{n-1}) - \gamma_{n-1} H(m_{n-1}) p_{n-1}\|_M}{\|j'(m_{n-1})\|_M}. \tag{32}$$

If the accuracy of the local quadratic approximation is good then this forcing term is close to zero and the Newton system is solved accurately. If not, then iterations are truncated sooner. This forcing sequence plays a similar role than the prospective ratio for trust-region method. It is however based on a (prospective) expansion of the gradient while the prospective ratio is based on an expansion of the misfit. Additional safeguards are also added to prevent this forcing term to decrease too fast or to increase above $\eta_0 = 0.9$. Interestingly, directions of negative curvatures are never investigated, except if it is the gradient. As previously an appropriate length γ is then given to this direction p through a line search. The unit step length $\gamma = 1$ is again tried first because it is the best choice if the misfit were quadratic.

2.5.2. Trust region globalization

When the Newton method is associated with a trust-region technique, the direction is found by minimizing the local quadratic expansion of the misfit

$$J^{\mathrm{pred}}(p) := J(m) + \langle j', p \rangle_M + 0.5 \langle Hp, p \rangle_M \tag{33}$$

inside a sphere of radius Δ. The constraint $\|p\|_M \leq \Delta$ limits the size of the direction and aims at preventing over-solving. This trust-region sub-problem can be solved approximately with the Steihaug conjugate gradient method (Algorithm 5) [8, 49]. This method actually exploits two properties of the conjugate gradient algorithm: successive approximate solutions always grow in norm ($\|p_k\|_M < \|p_{k+1}\|_M$) while the misfit prediction always decrease ($J^{\mathrm{pred}}(p_k) > J^{\mathrm{pred}}(p_{k+1})$). The underlying idea of the method is then to minimize the second order expansion of the misfit iteratively using the conventional conjugate gradient algorithm until either convergence is achieve, either the boundary is reached. Basically there are only two modifications compared to Algorithm 4. First, the inner iterations are cropped to the trust region radius Δ when the unconstrained solution increases beyond it. Second, when a direction of negative curvature is encountered, it is followed up to the boundary of

the trust region and the algorithm is stopped. Interestingly these directions were never investigated in the conventional version. The convergence criterion is unchanged but here the forcing term is kept constant ($\eta = 0.5$).

Algorithm 4
Conventional conjugate gradient algorithm

Require: $0 < \eta \leq 1$
$p_0 = 0$, $r_0 = j'$, $q_0 = -j'$
if $\langle Hj', j' \rangle_M \leq 0$ **then return** $-j'$
loop
if $\langle Hq_k, q_k \rangle_M \leq 0$ **then return** p_k
$\alpha_k = \langle r_k, r_k \rangle_M \,/\, \langle Hq_k, q_k \rangle_M$
$p_{k+1} = p_k + \alpha_k q_k$ and $r_{k+1} = r_k + \alpha_k H q_k$
if $\|r_{k+1}\|_M < \eta \|j'\|_M$ **then return** p_{k+1}
$\beta_{k+1} = \|r_{k+1}\|_M^2 \,/\, \|r_k\|_M^2$
$q_{k+1} = -r_{k+1} + \beta_{k+1} q_k$
end loop

Algorithm 5
Steihaug conjugate gradient algorithm

Require: $0 < \eta \leq 1$ and $\Delta > 0$
$p_0 = 0$, $r_0 = j'$, $q_0 = -j'$
loop
if $\langle Hq_k, q_k \rangle_M \leq 0$ **then**
$\tau^* = \tau > 0 \mid \|p_k + \tau q_k\|_M = \Delta$
return $p_k + \tau^* q_k$
end if
$\alpha_k = \langle r_k, r_k \rangle_M \,/\, \langle Hq_k, q_k \rangle_M$
if $\|p_k + \alpha_k q_k\|_M \geq \Delta$ **then**
$\tau^* = \tau > 0 \mid \|p_k + \tau q_k\|_M = \Delta$
return $p_k + \tau^* q_k$
end if
$p_{k+1} = p_k + \alpha_k q_k$ and $r_{k+1} = r_k + \alpha_k H q_k$
if $\|r_{k+1}\|_M < \eta \|j'\|_M$ **then return** p_{k+1}
$\beta_{k+1} = \|r_{k+1}\|_M^2 \,/\, \|r_k\|_M^2$
$q_{k+1} = -r_{k+1} + \beta_{k+1} q_k$
end loop

3. Numerical investigations

Numerical studies are performed in the context of subsurface acoustic imaging in the frequency domain [43, 47]. In that particular case, the misfit is conventionally chosen as the least-squares distance between some acoustic pressure measurements $d_{\omega er}$ (at some receiver r, for several excitation sources e and for different frequencies ω) and the corresponding computed acoustic pressures $p_{\omega e}(\boldsymbol{x}_r)$, obtained by solving the Helmholtz equation

$$J(s^2) = 0.5 \sum_{\omega,e,r} \left| p_{\omega e}(\boldsymbol{x}_r; s^2) - d_{\omega er} \right|^2 \quad \text{with} \quad \Delta p + \omega^2 s^2 p = \delta(\boldsymbol{x} - \boldsymbol{x}_e). \tag{34}$$

It is here chosen that the subsurface model parameter is the slowness squared distribution s^2 [$\mathrm{s}^2/\mathrm{km}^2$] (also called the sloth), as could be guessed from the expression of the Helmholtz operator $A_\omega(s^2) := \Delta + \omega^2 s^2$. The slowness squared s^2 is actually the squared inverse of the velocity v. Several other parametrizations are also possible but it has been shown that the slowness squared can yield a fast convergence and accurate results [2, 5, 18, 39]. Implementation of any of the above described local optimization algorithms requires an efficient procedure to compute the misfit and the gradient for a given slowness squared distribution s^2 and the action of the Hessian operator for any given slowness squared perturbation δs^2. The well-known adjoint state method has been developed for that specific purpose. It is summarized here below and detailed in [1, 17, 40]. The two terms in gray should be removed under the Gauss-Newton approximation. Also note that the preconditioning operator P in steps 3 and 6 need not to be constructed explicitly: the preconditioned gradient and Hessian are actually computed from their defining property (13) and (14), which express them in terms of inner products.

1. Find the forward fields $p_{\omega e}$ such that

$$A_\omega(s^2) p_{\omega e} = \delta(\boldsymbol{x} - \boldsymbol{x}_e). \tag{35}$$

2. Find the adjoint fields $p^\dagger_{\omega e}$ such that

$$A_\omega(s^2) p^\dagger_{\omega e} = \sum_r (\overline{p}_{\omega e}(\boldsymbol{x}_r) - \overline{d}_{\omega er}) \delta(\boldsymbol{x} - \boldsymbol{x}_r). \tag{36}$$

3. Find the preconditioned gradient j' such that

$$Pj' = -\sum_{\omega} \omega^2 \sum_{e} p^{\dagger}_{\omega e} \overline{p}_{\omega e}. \tag{37}$$

4. Find the perturbed forward fields $\delta p_{\omega e}$ such that

$$A_{\omega}(s^2)\delta p_{\omega e} = -\omega^2 \delta s^2 p_{\omega e}. \tag{38}$$

5. Find the perturbed adjoint fields $\delta p^{\dagger}_{\omega e}$ such that

$$A_{\omega,e}(s^2)\delta p^{\dagger}_{\omega e} = \sum_{r} \delta p_{\omega e}(\boldsymbol{x}_r)\delta(\boldsymbol{x} - \boldsymbol{x}_r) - \omega^2 \delta s^2 p^{\dagger}_{\omega e}. \tag{39}$$

6. Find the preconditioned Hessian operator $H\delta s^2$ in the direction δs^2 such that

$$PH\delta s^2 = -\sum_{\omega} \omega^2 \sum_{e} (\delta p^{\dagger}_{\omega e} \overline{p}_{\omega e} + p^{\dagger}_{\omega e} \delta \overline{p}_{\omega e}). \tag{40}$$

Independently of any practical solver for these wave propagation problems, a misfit evaluation only requires performing step 1 and thus only requires solving a single wave propagation problem. A gradient evaluation requires steps 1 to 3, thus a single supplementary wave propagation problem must be solved if the misfit has already been computed. Similarly, steps 1 to 6 are necessary for the application of the (Gauss-)Newton Hessian operator in a particular direction, thus again two supplementary wave propagation problems if the gradient has already been computed for the same model parameters.

Consequently the steepest descent and the l-BFGS directions require solving two wave problems while any Newton-based direction has an initial cost of four wave propagation problems and each supplementary conjugate gradient iteration requires two more wave problems. To the price of the directions must be added the cost of the line search or the trust-region methods. Line search typically accepts a step length if it verifies sufficient conditions (4) and (5) which involves the misfit and its gradient. Thus it requires one or two additional wave problems each time a trial step length is rejected. Prospective trust-region has no additional cost because the evaluation of the trust region only depends on quantities already computed. At the opposite, retrospective (Gauss-)Newton trust-region requires the application of the Hessian operator on the preceding direction and thus needs to solve two additional wave propagation problems. When a trust-region iteration fails, the model parameter is not updated and both the misfit and the gradient can be re-used for the following iteration, without solving any new wave propagation problem. Hessian-vector products could also be remembered to spare computational cost after a failed trust-region (Gauss-)Newton iteration, but the number of fields to memorize is larger, as it is proportional to the number of inner iterations. Hence, we choose not to store the Hessian vector products from the previous iteration. Because failed trust-region (Gauss)-Newton iterations barely appears, the speedup would be small anyway. Table 1 summarizes this accounting.

	Base	CG	LS	TR
SD	2	-	$2N_{\text{LS}}$	0
l-BFGS	2	-	$2N_{\text{LS}}$	0
LS-NCG	2	$2N_{\text{CG}}$	$2N_{\text{LS}}$	-
TR-NCG (P)	2	$2N_{\text{CG}}$	-	0
TR-NCG (R)	2	$2N_{\text{CG}}$	-	2

Table 1: Wave propagation problem solution count for a single outer iteration of the steepest descent (SD), the l-BFGS (LB) or the (Gauss-)Newton (NCG) methods combined with a line search (LS) or a trust region (TR) with a prospective (P) or retrospective (R) radius update. N_{CG} is the number of inner iteration of the conjugte gradient algorithm. N_{LS} is the number of rejected values of γ during the line search.

It is interesting to highlight here that the first inner iteration of any conjugate gradient Newton method is simply the steepest descent but it is twice more expensive because the curvature is computed. Subsequent inner iterations must therefore provide large decrease of the misfit to compensate this high entry cost. This phenomenon is even worse with the retrospective trust region algorithm because there is a systematical additional cost to update the trust region radius.

In this work, solutions to partial differential equations (35) to (40) are obtained numerically with the finite element method. In what follows, we specify the exact numerical procedure in that context. Note however that the analysis would nearly be identical with finite differences. As far as wave equations in the frequency domain are concerned, direct or hybrid solvers are typically used, because they outperform standard iterative methods, for which no robust preconditioner currently exists for the high-frequency regimes encountered in full waveform inversion [13, 48]. In that context, finite element discretization assembles operators into matrices and source terms into vectors. Wave propagation problems (35), (36), (38) and (39) therefore transform into a linear system whose left-hand-side matrix A is always the same for a given frequency while the right-hand-side source b is different for any field type, frequency and excitation index. The solution of this system is obtained by first computing its lower-upper factorization then by performing an upward-backward substitution for each right-hand-side source

$$\mathrm{Ap} = \mathrm{b} \quad \Leftrightarrow \quad \mathrm{A} = \mathrm{LU}, \quad \mathrm{Lq} = \mathrm{b} \text{ and } \mathrm{Up} = \mathrm{q}. \tag{41}$$

Huge computational reduction is therefore obtained because only one matrix per frequency is assembled and factorized. The computation of any wave field then requires the assembly and the upward-backward substitution of a vector per excitation source, but no more matrix factorization. The numerical equivalence of the preceding six steps procedure is given here below.

1.	• Factorize wave propagation operators	(n_ω)
	• Substitute forward sources	$(n_\omega \times n_e)$
2.	• Substitute adjoint sources	$(n_\omega \times n_e)$
3.	• Factorize the preconditioner	(1)
	• Substitute the conventional gradient	(1)
4.	• Substitute perturbed forward sources	$(n_\omega \times n_e)$
5.	• Substitute perturbed adjoint sources	$(n_\omega \times n_e)$
6.	• Substitute the conventional Hessian	(1)

It is interesting to highlight that model problems (steps 3 and 6) are negligible with respect to wave problems. Indeed while wave problems involve a matrix per frequency and a vector per excitation source, model problems only involve a single matrix (*i.e* the preconditioner) and a single source vector (*i.e* the conventional gradient or Hessian). Moreover the model discretization is usually coarser than the wave field discretization. Consequently not considering these model problems when quantifying the computational complexity is not dramatic. It should however be highlighted that forward problems are more expensive than the corresponding adjoint problem, because the matrix factorization is reused. Moreover the perturbed forward problem and the perturbed adjoint problem are slightly heavier than the adjoint problem, because both their sources are dense, at the opposite of forward and adjoint sources, which are sparse. Nevertheless we weight equally all of these four problems when quantifying the computational complexity.

In the next two sections, two synthetic numerical case studies are investigated. The first one is based on the widely used Marmousi benchmark [52] while the second one, replicated from [28], is inspired from a near-surface imaging of close concrete structures and features important multiple scattering. Multiple scattering is responsible for the indefiniteness of the Hessian operator, which, as mentioned in the previous part, is challenging for optimization algorithms [17, 28, 27, 43]. This second example is thus chosen to emphasize which optimization methods are able to overcome such difficulties. For both case studies, the influence of the inner product choice on the convergence speed and the quality of the inverted model is studied first. Once the inner product is chosen, prospective and retrospective trust-region methods with different parameter sets are compared and the best option is selected. Advantages and drawbacks of trust-region methods in the context of full waveform inversion are then finally discussed based on a comparison with the corresponding line search methods. In the remainder of this section, data misfit are normalized such that the misfit corresponding to the initial model is one and computational complexity is measured in numbers of forward problems solved, as explained above.

3.1. Case study 1

Numerical inversions are performed on the 2D Marmousi model (Fig. 2a) [52] in the frequency domain. Three frequencies (4, 6 and 8 [Hz]) are inverted simultaneously. The surface acquisition system is composed of 122 equally spaced (72 [m]) excitation sources and 243 equally spaced (36 [m]) receivers. Outer iterations are stopped when satisfying the convergence criterion $J(s^2)/J(s^2_{\text{init}}) < 10^{-3}$. A smoothed version of the exact Marmousi model is used as an initial guess (Fig. 2b). This initial model is computed with a Laplacian filter $s^2_{\text{init}} = (1-l_c^2\Delta)^{-1}s^2_{\text{exact}}$ with $2\pi l_c = 2$ [km]. Slowness squared fields and pressure fields at the three frequencies are discretized on a square grid (36 [m]) by hierarchical finite elements, respectively of order 1 and of order 2, 3, 4. A water layer (216 [m]) is also added at the top of the model but it is kept constant during the inversion. The model is spatially truncated by Sommerfeld boundary conditions [45] on all four sides. The top boundary is thus not a free surface. Recorded data are generated synthetically using the same hierarchical finite elements setting than for the inversion. An inversion result, *i.e.* an estimated squared slowness, is shown in (Fig. 2c). From a relatively low resolution initial guess, full waveform inversion indeed provides a high-resolution estimation of the exact model. Images obtained with the other methods do not differ significantly.

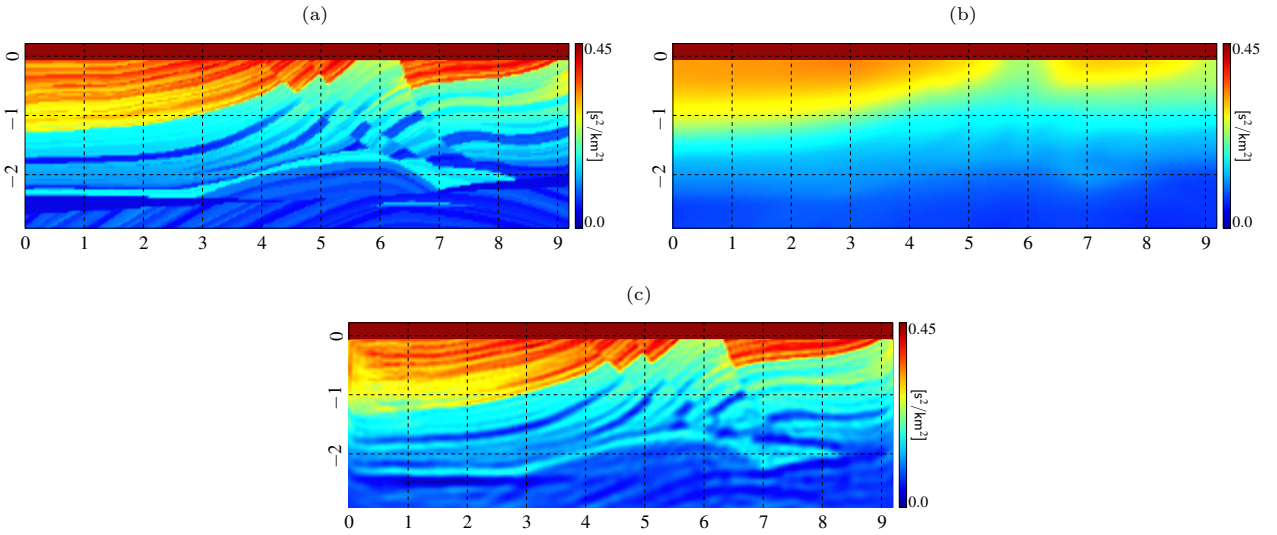

Figure 2: Marmousi model (a), initial guess (b) and inversion results using a line search l-BFGS algorithm with a weighted and thresholded inner product (c).

3.1.1. Inner product

As explained earlier, the inner product has an influence on both the gradient and the Hessian. Its choice is therefore expected to influence the convergence speed but also the particular minimizer that is reached [8]. To illustrate both these effects, the line search l-BFGS algorithm has been applied with the four different inner products introduced in this work, *i.e.* the conventional inner product (15), the weighted inner product (16) and its regularized variants (18) or (19). For the smoothing inner product (19), the characteristic length is set to the smallest propagated wavelength, *i.e.* $l_c = 250$ [m]. In the context of full waveform inversion, the expected resolution is a fraction of the wavelength. Enforcing a much smaller smoothing length therefore has hardly no effect. The smallest propagated wavelength actually provides an appropriate lower bound for the characteristic length. Corresponding convergence curves and error maps are given in Fig. 3 and 4 respectively. Both these figures are also summarized in Table 2. As can be seen from these figures, all weighted inner product increase the convergence speed with respect to the conventional, *i.e.* unweighted, one. However the minimizer obtained with the weighted inner product alone is further away from the exact solution, in particular in the right corner of the model. Avoiding such artifacts is precisely one of the reasons for the introduction of regularized inner products, as they dampen the contributions in these poorly illuminated regions. Both the thresholding and the smoothing strategy perform similarly in reducing the error back to the same level than the unweighted solution but the thresholding strategy converges faster. It is thus kept for the sequel of this case study. The advantages of the smoothing inner product will be highlighted during the second case study. In the next three subsections, the behaviour of the steepest descent method, the l-BFGS method, the full Newton and the Gauss-Newton methods is analysed. Convergence curves and interesting statistics for all these methods are given in Fig. 5 and Table 3 respectively.

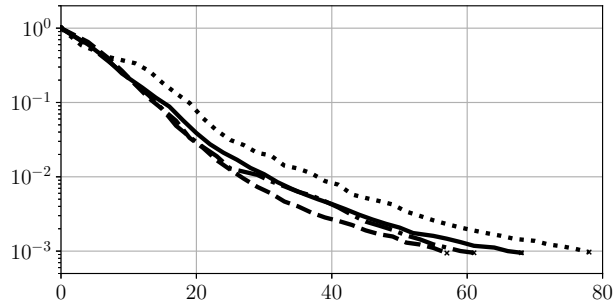

Figure 3: Data misfit as a function of the number of wave propagation problem solved for the line search l-BFGS algorithm with a conventional (••), only weighted (-•), weighted and thresholded (--) or weighted and smoothed (–) inner product.

		Wave sol. (tot)	Error rms ($[s^2/km^2]$)
Conventional		78	0.0174
Weighted	only	61	0.0202
	and thresholded	57	0.0174
	and smoothed	68	0.0173

Table 2: Number of wave propagation problem solved and root-mean squared error for the line search l-BFGS algorithm with different inner products.

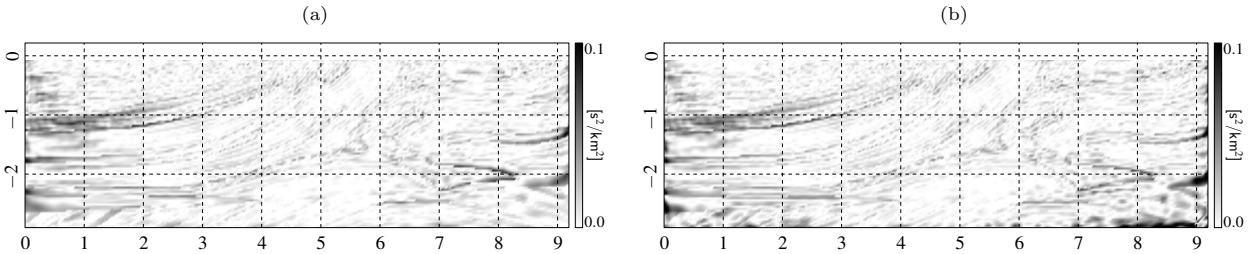

Figure 4: Final inversion error for the line search l-BFGS algorithm with a conventional (a) or a weighted (b) inner product. Inversion errors for both regularized inner products are not shown because these are very similar to those obtained with the conventional inner product.

3.1.2. Steepest descent

There is no dramatic improvement when using one or another direction scaling method, because actually the direction itself is bad. Nevertheless, it appears that methods which reject less frequently the proposed update direction are faster, *i.e* the prospective trust-region method with the more cautious parameters sets (B and C) and the line search method. Retrospective radius update does not speed up convergence. Actually we observed that the retrospective predicted decrease (12) sometimes largely underestimates the actual decrease, illustrating that the retrospective misfit prediction is very not accurate, but still producing an increase of the trust region radius. Finally, among the three best methods, the slope is slightly steeper for the two trust-region methods, probably because they systematically try to increase the length given to the gradient direction.

3.1.3. Limited memory BFGS method

There is hardly no difference between all the methods combined with the l-BFGS algorithm. We observed that the line search method only rejects the unit step length $\gamma = 1$ for the first two iterations. Similarly, we observed that the retrospective ratio is always very close to one, such that the trust region radius for retrospective methods quickly becomes large and thus the pure l-BFGS direction is always accepted after the first few iterations. An algorithm that unconditionally follows the pure l-BFGS direction would therefore already be very good and neither a line search nor a trust-region method can actually drastically improve it, as far as convergence speed is concerned. Nevertheless the more cautious prospective trust-region methods (B,C) also converge fast, which shows that, on the other hand, constraining the size of the update directions does not slow down the inversion.

		Wave sol. (tot)	Wave sys. (tot)	Outer it. (tot)	Inner it. (avg)	Rejected (%)	Constrained (%)	Negative curv. (%)
SD	LS	244	128	111	-	10	-	-
	TR-P (A)	317	198	198	-	40	100	-
	TR-P (B)	272	140	140	-	6	100	-
	TR-P (C)	258	132	132	-	5	100	-
	TR-R (A)	328	164	164	-	20	100	-
	TR-R (B)	354	177	177	-	20	100	-
	TR-R (C)	330	165	165	-	25	100	-
LB	LS	57	29	27	-	7	-	-
	TR-P (A)	57	29	29	-	3	10	-
	TR-P (B)	57	29	29	-	3	34	-
	TR-P (C)	60	32	32	-	13	50	-
	TR-R (A)	58	29	29	-	3	10	-
	TR-R (B)	56	28	28	-	0	11	-
	TR-R (C)	56	28	28	-	0	11	-
FN	LS	139	24	17	2.9	12	-	29
	TR-P (A)	171	22	22	3.0	32	77	0
	TR-P (B)	106	13	13	3.1	0	69	0
	TR-P (C)	106	16	16	2.3	0	75	0
	TR-R (A)	144	14	14	3.1	14	64	0
	TR-R (B)	128	14	14	2.6	0	79	0
	TR-R (C)	142	17	17	2.2	0	82	0
GN	LS	124	15	15	3.13	0	-	-
	TR-P (A)	130	11	11	4.9	0	10	-
	TR-P (B)	98	10	10	3.9	0	30	-
	TR-P (C)	98	10	10	3.9	0	30	-
	TR-R (A)	152	11	11	4.9	0	10	-
	TR-R (B)	132	14	14	2.7	0	79	-
	TR-R (C)	184	24	24	1.8	0	83	-

Table 3: Statistics related to the implementation of the steepest descent (SD), the l-BFGS (LB), the full Newton (FN) method and the Gauss-Newton (GN) methods combined with a line search (LS) or a trust region (TR) with a prospective (P) or retrospective (R) radius update with different parameter sets (A,B,C).

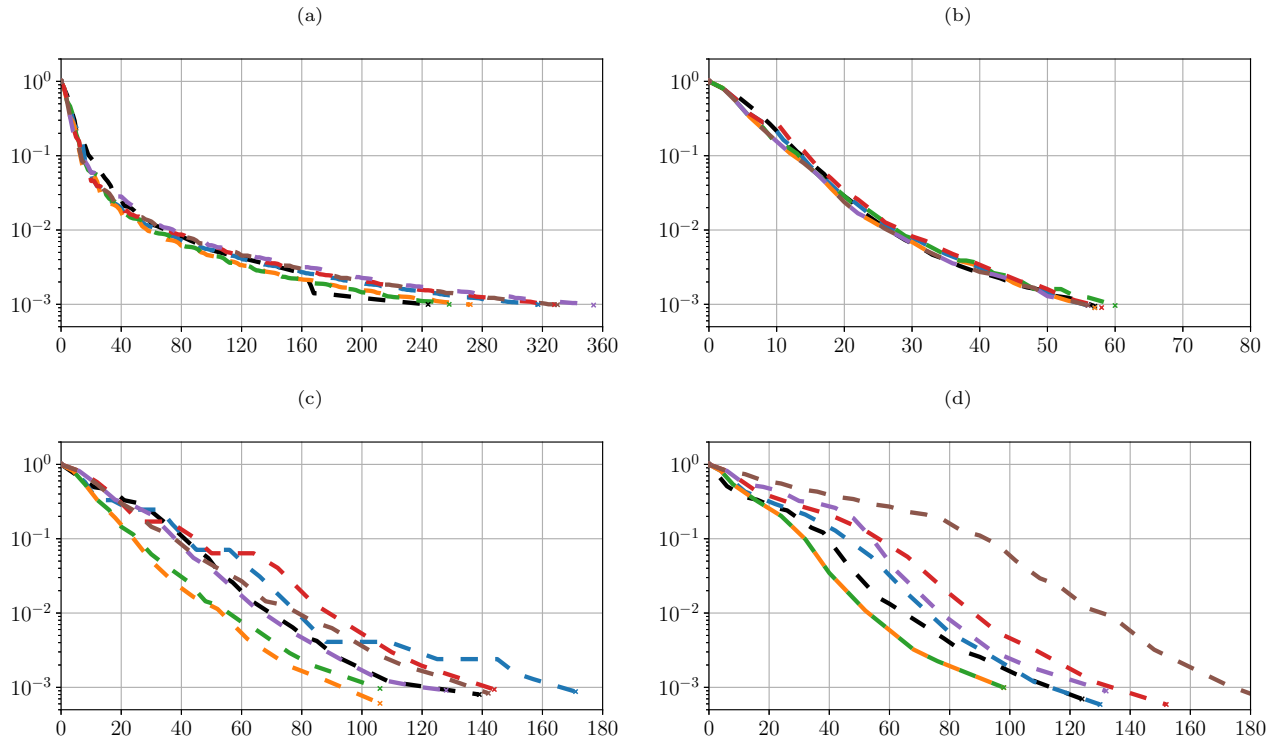

Figure 5: Data misfit as a function of the number of wave propagation problem solved for the steepest descent (a), the l-BFGS (b), the full Newton (c) and the Gauss-Newton (d) methods combined with either a line search (▬) or a prospective trust region (A (▬), B (▬), C (▬)) or a retrospective trust region (A (▬), B (▬), C (▬)).

3.1.4. Newton methods

As far as trust-region methods are concerned, it first clearly appears that the retrospective radius update is not worth its computational cost. Indeed it does not require less wave solutions than the best prospective ones, even if the computation cost of the retrospective predicted decrease is withdrawn (two wave solutions per outer iterations). Retrospective radius update has been introduced to anticipate and prevent failures. However the prospective Newton method combined with the more cautious parameters sets (B and C) does already not reject any direction and there is then no interest in computing the retrospective ratio.

Among the prospective methods, it appears that the two more cautious (B and C) yield the fastest convergence. Convergence speed decreases when using parameter set A with both the full Newton method and the Gauss-Newton method but for two different reasons. With parameter set A, the trust-region radius grows quickly and the full Newton method is thus allowed to explore large directions, beyond the validity of the exact second order expansion (2). Such directions produce a high rejection rate (32%) and thus a waste of computational effort. At the opposite, the Gauss-Newton method never rejects a direction and the explanation for its slower convergence can therefore not be the same. During the earliest iterations, far from the global minimum, the Gauss-Newton approximation is not valid (because data residuals are not small yet) and thus the Gauss-Newton Hessian is quiet different from the full Hessian. The misfit prediction under the Gauss-Newton approximation is thus cruder than the exact second order expansion (2) and the ratio ρ_{p} is even more likely to be away from one. This ratio ρ_{p} is given in Fig. 6c. As can be seen, during the first few iterations, this ratio is actually very larger than one, which indicates that the misfit prediction is indeed not accurate. Nevertheless, the trust region radius is still increased and the system is solved accurately while the Hessian and the misfit are not approximated accurately. This effect generates over-solving the system at the earliest iteration and slows down the Gauss-Newton method, as can be seen by comparing the initial slopes between variant A and B/C in Fig. 5d. This effect would be even more dominant if the convergence requirements, *i.e.* the forcing sequence η, was smaller. With the large value $\eta = 0.5$ chosen here, convergence of the conjugate gradient algorithm is attained relatively fast. Actually variant B and C perform better than variant A only because it takes more iterations for the trust region constraint to become inactive. Starting with a larger initial radius would result in the same convergence speed than variant A. Also, it is interesting to highlight that when using the retrospective

radius update with the Gauss-Newton approximation, the situation is reversed because the retrospective ratio is then smaller than one. Instead of over-solving, under-solving then appears. Therefore we believe that it is better to use trust-region methods with the full Newton Hessian, because it constructs the best possible misfit prediction while it does not introduce supplementary difficulties.

The full Newton method and the Gauss-Newton method are slightly slower when combined with a line search method and also require more wave operator factorization. As far as the full Newton method is concerned, directions of very small curvature can produce large update directions, far beyond the validity of the expansion (2). In such cases the initial length $\gamma = 1$ is rejected and some computational cost must be involved to reduce it to satisfy Wolfe conditions. This effect has actually been observed twice using the full Newton method. Moreover during the first fifth outer iterations, the full Newton method using the line search globalization stops because a direction of negative curvature is encountered. At the opposite of its trust-region counterpart, the line search variant of the conjugate gradient algorithm discard any direction of negative curvature, thus wasting the associated computational cost. Of course within the Gauss-Newton approximation this second effect can not appear (and the first one was actually not observed). The line search globalization therefore seems more suited with the Gauss-Newton approximation. Nevertheless it is not much faster. In the context of line search globalization, the accuracy of the second order local expansion is expressed through the forcing sequence η, which is, as can be seen in Fig. 8, away from zero during the first few iterations. Consequently, the convergence of the conjugate gradient algorithm is quickly reached and only a few inner iterations are performed per outer iterations as can be seen from Fig. 7c. Fig. 7c actually show how hard it is to stop the Gauss-Newton inner iterations at the right time: the fastest method is the prospective trust region B/C and it performs less inner iterations then the variant A but more than the line search method. The difficulty to pick up an appropriate stopping criterion for the Gauss-Newton method is another motivation to use the full Newton method instead. Using the full Newton method, the line search variant actually suffers from directions of small or negative curvature while trust-region methods do not. Based on this case study, we would therefore recommend to use the full Newton method combined with a trust-region method and a prospective radius update.

Conclusions regarding the l-BFGS method and the full Newton method have been reinforced by a slightly modified example, in which noise is introduced into the data. The results and parameters of this additional case study are discussed in more detail in Appendix A.

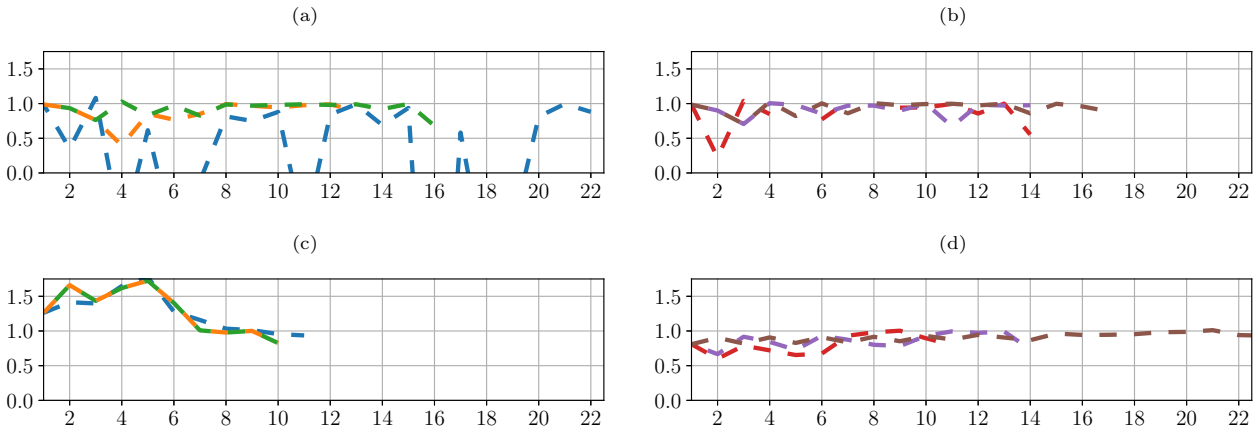

Figure 6: Prospective ratio ρ_{p} (a,c) or retrospective ratio ρ_{r} (b,d) during the outer iterations of the full Newton method (a,b) and the Gauss-Newton method (c,d) with different parameter sets using a prospective radius update (a,c) (A (▬), B (▬), C (▬)) or a retrospective radius update (b,d) (A (▬), B (▬), C (▬)).

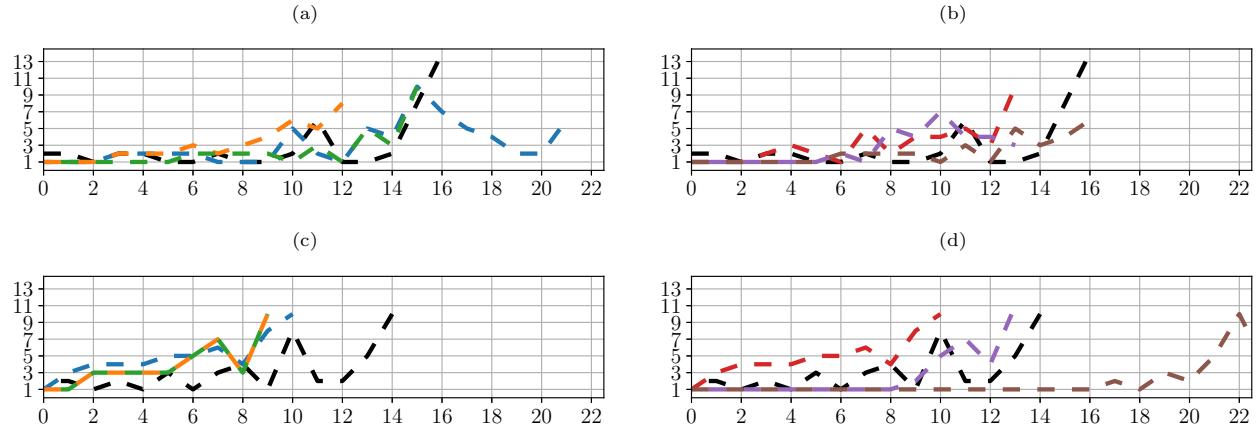

Figure 7: Inner iterations per outer iteration for the full Newton method (a,b) and the Gauss-Newton method (c,d) with different parameter sets using a prospective radius update (a,c) (A (▬), B (▬), C (▬)) or a retrospective radius update (b,d) (A (▬), B (▬), C (▬)).

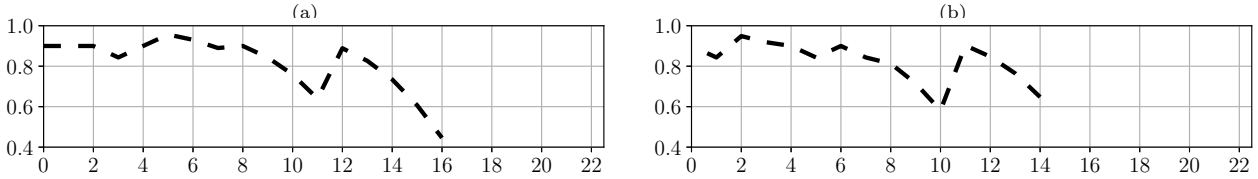

Figure 8: Forcing sequence η for the full Newton (a) and the Gauss-Newton (b) methods combined with a line search method (▬). The forcing sequence for methods combined with a trust-region is constant ($\eta = 0.5$).

3.2. Case study 2

The configuration of this second case study is replicated from [28]. The true velocity distribution is given in Fig. 10a. It presents two T-shaped concrete structures ($v_c = 4$ [km/s]) embedded in a homogeneous background ($v_b = 0.3$ [km/s]) with a horizontal layer reflector in the bottom ($v_r = 0.5$ [km/s]). The depth of investigation is limited to 3 [m] while the width is 30 [m]. The aspect ratio and the propagation scales are thus very different from the Marmousi model. These two concrete foundations, buried at few meters deep, generate high-amplitude reflections because of the very high velocity contrast with the background. Moreover, important multiple scattering appears between the two structures, as they are relatively close to each other. The acquisition system is divided into three segments: one on the surface and the two others inside boreholes on both lateral sides. Sources and receivers are equally spaced (15 [cm]) along these three segments. Note that the surface sources and receivers that would lie inside the two concrete structures are not considered in the modelling, leading to an actual number of sources and receivers totaling 227. Nine frequencies (100, 125, 150, 175, 200, 225, 250, 275, and 300 [Hz]) are inverted simultaneously from an initial model composed of the homogeneous background and the bottom reflector only. For this second case study, a logarithmic slowness squared parametrization is used $m := \ln s^2$. This parametrization has the advantage to be unable to produce negative values of the slowness squared. Inverting the slowness squared actually drives it into negative values, because of the two concrete structures whose slowness squared is really close to zero. Outer iterations are stopped when satisfying the convergence criterion $J(\ln s^2)/J(\ln s^2_{\text{init}}) < 10^{-2}$. Slowness squared fields and pressure fields at the nine frequencies are discretized on a square grid (15 [cm]) by hierarchical finite elements, respectively of order 1 and of order 2, 3, 4. At the light of the first case study, trust-region methods with parameter sets A and C will no longer be considered, as both were systematically outperformed by parameter set B.

3.2.1. Inner product

Illumination of the medium is nearly perfect and consequently, the diagonal part of the Gauss-Newton Hessian that we previously used as a weight can reasonably be approximated by a constant h_{GN}. However the part related to the change of variable is varying spatially $\delta s^2 = \frac{ds^2}{d\ln s^2}\,\delta \ln s^2 = s^2\,\delta \ln s^2$. Hence the weight for the inner product is chosen as $w = h_{\text{GN}}\ s^4$. As previously, the line search l-BFGS algorithm has been applied with the four different inner products introduced in this work. Convergence curves are given in Fig. 9 while inversion results are given in Fig. 10. For the weighted and smoothed variant, the threshold is set as $\epsilon = h_{\text{GN}}\ s_b^4$ while the characteristic length for the smoothing inner product is set to $l_c = 3$ [m]. It is important to highlight that this length is greater than the smallest wavelength in the background medium (1 [m]) while for the first case study, this length was actually close to the smallest wavelength. The weighted and thresholded variant has been tested for several values of the threshold, from $\epsilon = h_{\text{GN}}\ s_c^4$ to $\epsilon = h_{\text{GN}}\ s_b^4$ but none of them provided inversion results significantly different from the conventional or the weighted inner products. Only the smoothing inner product is able to reconstruct the model parameter accurately, provided the smoothing length is sufficiently large, *e.g.* $l_c = 3$ [m]. When the smoothing length is decreased below the wavelength (1 [m]), convergence issues appear again because the smoothing effect becomes negligible and the situation is then the same than with the conventional inner product. This smoothing inner product actually mitigates the non-linearity of the misfit, because spatial roughness is incorporated progressively in the model parameter [60]. During the inversion, the model parameter never explores extremely high velocity values, at the opposite of the other variants. It is thus able to converge to an accurate solution while more straightforward optimization is not, because of a numerical breakdown. Adding a regularization term to the misfit function, e.g. a Tikhonov penalization [51], eliminates this numerical breakdown but, depending on the value of the regularization parameter, the reconstruction then either features a lower contrast or the convergence is much slower compared to the smoothing inner product. Consequently, this smoothing inner product is used for the remainder of this study. The performance of the three optimization methods is described in the next three subsections. Convergence curves, inversion results and statistics are given in Fig. 12 and 11 and in Table 4 respectively.

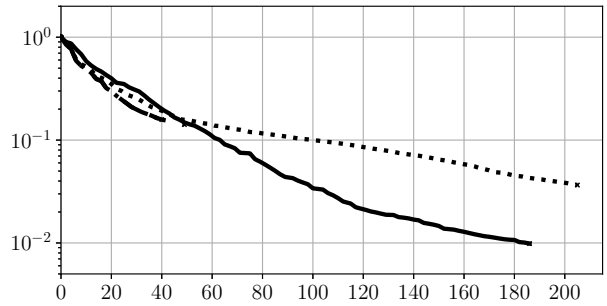

Figure 9: Data misfit as a function of the number of wave propagation problem solved for the line search l-BFGS algorithm with a conventional (••), only weighted (- •) or weighted and smoothed (–) inner product.

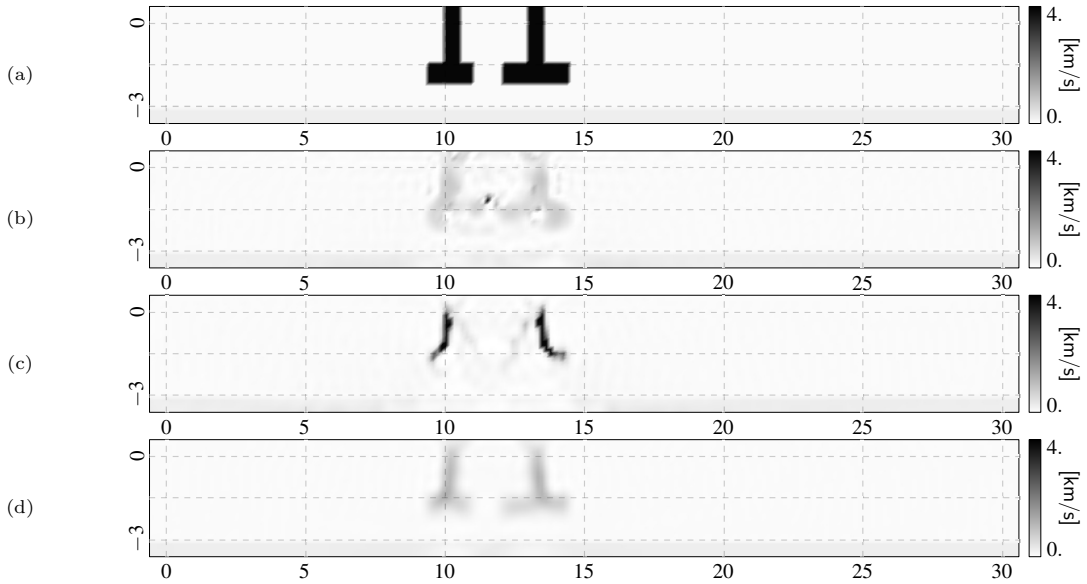

Figure 10: Near-surface concrete structures velocity model (a) and inversion results using a line search l-BFGS algorithm with a conventional (b), a weighted only (c) or a weighted and smoothed (d) inner product.

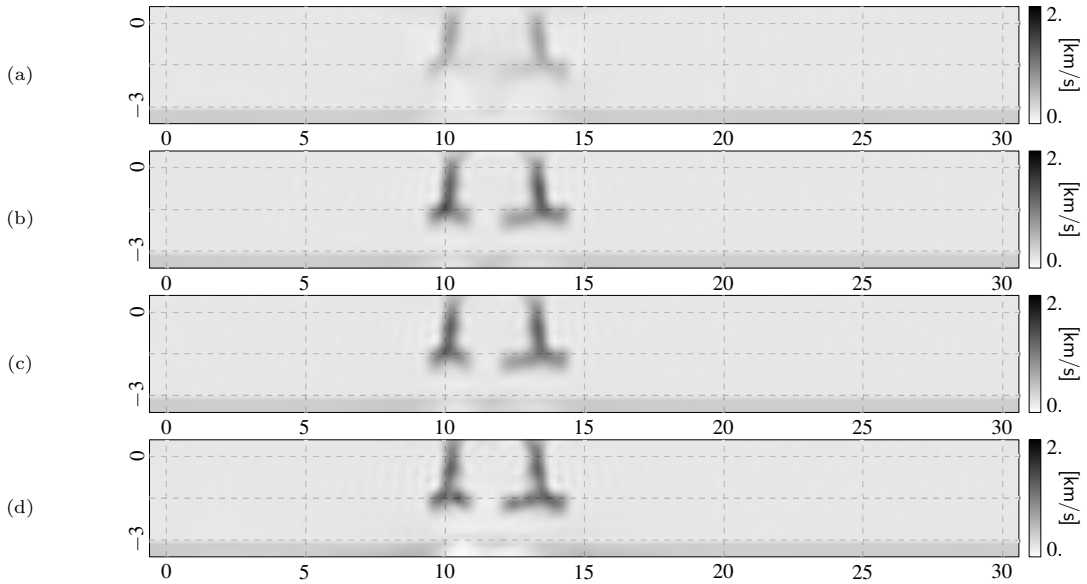

Figure 11: Inversion results for the steepest descent (a), the l-BFGS (b), the full Newton (c) and the Gauss-Newton (d) methods combined with trust-region method using a prospective radius update (B). Note the the upper color scale limit is only 2 [km/s].

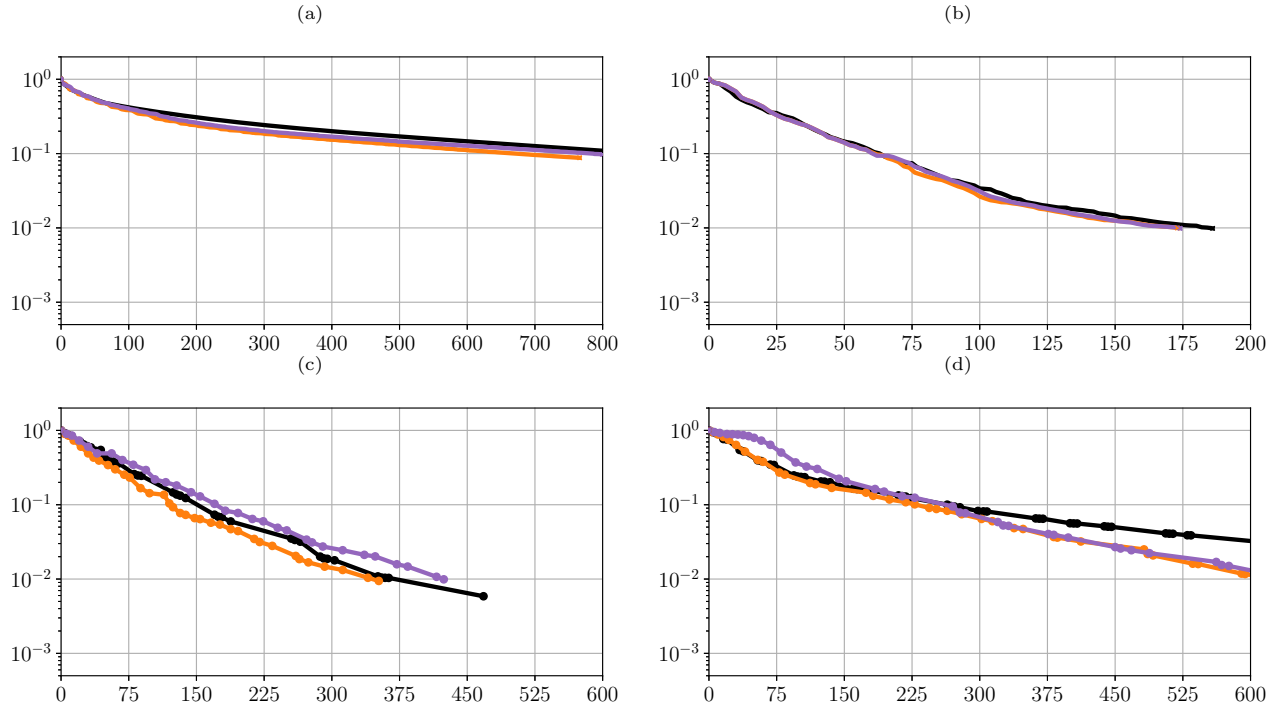

Figure 12: Data misfit as a function of the number of wave propagation problem solved for the steepest descent (a), the l-BFGS (b), the full Newton (c) and the Gauss-Newton (d) methods combined with either a line search (▬) or a trust-region with a prospective (B (▬)) or a retrospective (B (▬)) radius update. Dots on (Gauss)-Newton curves indicate outer iterations.

		Wave sol. (tot)	Wave sys. (tot)	Outer it. (tot)	Inner it. (avg)	Rejected (%)	Constrained (%)	Negative curv. (%)
SD	LS*	803	403	400	-	1	-	-
	TR-P (B)*	766	400	400	-	9	100	-
	TR-R (B)*	800	400	400	-	18	100	-
LB	LS	186	98	88	-	8	-	-
	TR-P (B)	173	87	87	-	1	23	-
	TR-R (B)	174	87	87	-	2	6	-
FN	LS	468	57	35	5.3	23	-	17
	TR-P (B)	352	34	34	4.2	0	56	0
	TR-R (B)	424	31	31	4.8	3	68	0
GN	LS*	923	65	60	6.6	8	-	-
	TR-P (B)	680	38	38	7.9	0	42	-
	TR-R (B)	672	39	39	6.6	0	41	-

Table 4: Statistics related to the implementation of the steepest descent (SD), the l-BFGS (LB), the full Newton (FN) and the Gauss-newton (GN) methods combined with a line search (LS) or a trust-region (TR) with a prospective (P) or retrospective (R) radius update with parameter set B. Star marker $*$ indicates methods that have been stopped before convergence.

3.2.2. Steepest descent

The steepest descent method is not able to reach convergence in a reasonable amount of computations. Progressively decreasing the smoothing length l_c during the inversion would accelerate the convergence [60], but it is not needed for more sophisticated methods and thus it is not done here neither. As for the first test case, the slope of trust-region methods is slightly steeper than the line search method. The prospective radius update rejects less often directions and hence converges faster than the retrospective radius update.

3.2.3. Limited memory BFGS method

Similarly to the first test case, the influence of the globalization method on the convergence speed is small. Trust-region methods actually spare a part of the line search cost, in terms of both the number of wave solutions and system factorizations, but it already represents only a tiny fraction (20 wave solutions) of the overall computational cost (186 wave solutions). Retrospective ratio is again always very close to one and the only difference between retrospective and prospective radius update is the frequency the size constraint is active, although it does not influence the convergence speed.

3.2.4. Newton methods

For this case study, the full Newton method clearly outperforms the Gauss-Newton method, independently of the globalization method used. On the one hand, the convergence speed is much higher and on the other hand the accuracy of the inversion results is superior. As demonstrated in [28], the missing negative definite part of the Hessian can prevent the Gauss-Newton method from reaching an accurate reconstruction. Here, thanks to the inner product preconditioning, every method is able to find a relevant minimum but the invalidity of the Gauss-Newton approximation impacts the convergence speed and the inversion results. Interestingly, for the Gauss-Newton method, the retrospective radius update succeeds to compensate its cost (2 wave solutions per outer iteration). Indeed, during the earliest outer iterations when the Gauss-Newton and the full Hessian are different, we observed that the retrospective ratio is smaller than one while the prospective ratio is bigger than one. Consequently the retrospective method performs less inner iterations per outer iterations than the prospective method (Fig. 13b), and thus avoids early over-solving. In the end both methods still converge at the same speed, but the retrospective method has spent less time in the computation of linear system solutions (680 versus $672 - 2 \times 39 = 594$ wave solutions). At the opposite, for the full Newton method, the retrospective method spent even more time in the computation of linear system solutions than the prospective method. The prospective method is actually already efficient because the prospective misfit prediction is accurate. The line search globalization also provides fast convergence in this case, despite the fact that directions of negative curvature are often encountered (12 wasted wave solutions) and that the unit step length is often rejected. However the flow of the method is very different from trust-region methods and particularly, requires much more wave operator factorizations. Indeed line search methods have a tendency to compute a single very accurate system solution, followed by several very inaccurate system solutions as can be seen from Fig. 13a and from the dots spacing in Fig. 12c while trust-region methods perform a nearly steadily increasing number of inner iterations per outer iteration. Whether a flow is better than the other has not been emphasized by our case studies, except that trust-region methods require less wave system factorization. In the case of noisy data, we however believe it could have an influence.

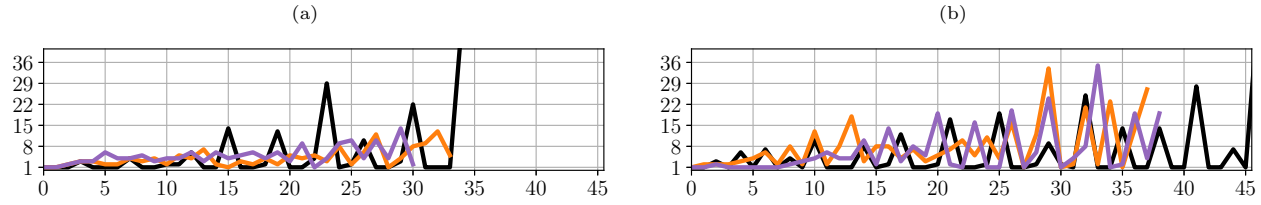

Figure 13: Inner iterations per outer iteration for the full Newton method (a) and the Gauss-Newton method (b) combined with either a line search (▬) or a trust-region with a prospective (B (▬)) or a retrospective (B (▬)) radius update.

4. Conclusion

In this work, we investigated the use of trust-region methods in the context of full waveform inversion in the frequency domain. At the heart of any trust-region method is the trust-region constraint, which is expressed in terms of the inner product chosen for the model parameter space. Consequently we began our analysis by investigating different inner product choices that could be implemented. We showed that changing the inner product does not only modify how lengths are measured but also acts as a preconditioner on both the gradient and the Hessian operator. Based on two numerical case studies, we showed that moving from a conventional inner product to a smoothed and/or weighted inner product can accelerate the convergence and mitigate the non-linearity of the misfit, for any optimization method independently of the globalization method (line search or trust region).

In parallel with this inner product choice, we also introduced line search and trust-region variants of the steepest descent, the l-BFGS and the (Gauss-)Newton methods. The number of wave propagation problems to be solved for each method was derived in order to compare them fairly. For each optimization method, the line search and the trust-region globalizations were then compared based on two different case studies. Thanks to the inner product preconditioning, every combination actually already yields very satisfying results. Nevertheless, we showed that trust-region methods outperform line search methods in numerous situations. In particular, we observed that the steepest descent converges slightly faster, because the trust-region methods always tried to increase the step length. As far as the l-BFGS method is concerned, very few differences were noted, but interestingly, constraining the size of the update direction did not decrease the convergence speed. The more dramatic differences appeared when using the full Newton method. Trust-region methods actually overcome the difficulties that appeared when using a line search method with the full Newton method. The Gauss-Newton approximation is not required with trust-region methods and actually degrades their performances, because this approximation also degrades the misfit prediction.

We believe that more sophisticated optimization methods, for example combining l-BFGS and Newton methods, could increase even more the convergence speed. Future works should also investigate the behaviour of inner product preconditioned trust-region methods in the presence of noise, possibly with new inner products that involve prior information on the model parameter space. We believe that the size constraint could act as a regularization method *per se*. Based on our study and these potential extensions, trust-region methods and inner product preconditioning seem to be two very useful tools for full waveform inversion.

5. Acknowledgements

The authors would like to thank Anthony Royer for his help on the finite element solver used in this work [44]. This research was funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) and by the ARC "WAVES" grant 15/19-03 from the Wallonia-Brussels Federation of Belgium. Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the F.R.S.-FNRS and by the Walloon Region.

Appendix A. Case Study 1 with noisy data

In this appendix, the Marmousi case study is performed on noisy data. Specifically, a band limited ($B = 15$ [Hz]) Gaussian noise (SNR $= 4$ [-]) is added to the noise-free data in the time domain, before frequency extraction. A Tikhonov penalization [51] is also added to the misfit for regularization. Only the l-BFGS method with a line-search or a retrospective/prospective trust-region and the full Newton method with a line search or a prospective trust-region are considered here, as these methods are the most efficient for the noise-free case. The iterative algorithm is here stopped when the total misfit drops below some absolute threshold.

The inversion result for the l-BFGS method combined with a line-search is given in Figure A.14. Inversion results for the other four methods are again very similar. The convergence curves are given in Figure A.15 while some statistics are given in Table A.5. These results show that there is no significant difference with noise-free situation. Indeed, the main conclusions remain valid: the l-BFGS method is not very sensitive to the globalization method while the full Newton method should better be combined with a trust-region method, as it is more efficient to prevent over-solving.

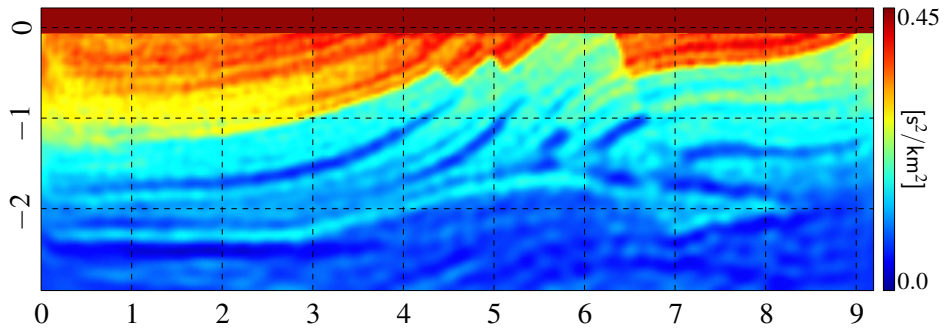

Figure A.14: Inversion results for the Marmousi model using a line search l-BFGS algorithm on noisy data, regularized with a Tikhonov additive misfit penalization.

		Wave sol. (tot)	Wave sys. (tot)	Outer it. (tot)	Inner it. (avg)	Rejected (%)	Constrained (%)	Negative curv. (%)
LB	LS	53	28	25	-	12	-	-
	TR-P (B)	58	30	30	-	7	30	-
	TR-R (B)	68	34	34	-	15	32	-
FN	LS	197	41	11	6.5	55	-	55
	TR-P (B)	101	17	17	2.0	6	71	0

Table A.5: Statistics related to the implementation of the l-BFGS (LB) and the full Newton (FN) combined with a line search (LS) or a trust-region (TR) with a prospective (P) or retrospective (R) radius update with parameter set B, for the Marmousi model with noisy data.

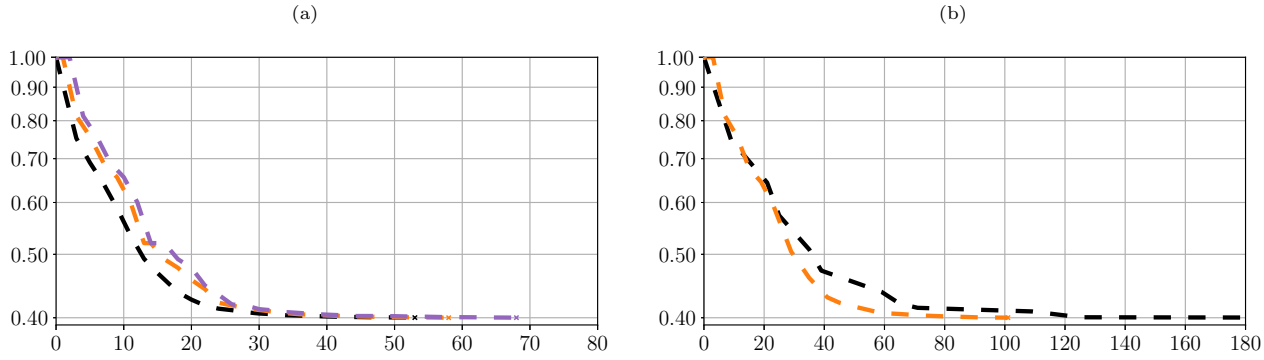

Figure A.15: Data misfit as a function of the number of wave propagation problem solved for the l-BFGS (a) and the full Newton (b) combined with either a line search (▬) or a trust-region with a prospective (B (▬)) or a retrospective (B (▬)) radius update, for the Marmousi model with noisy data.

References

[1] Adriaens, X., Métivier, L., and Geuzaine, C. A trust-region Newton method for frequency-domain full-waveform inversion. In *First International Meeting for Applied Geoscience & Energy Expanded Abstracts* (9 2021), vol. 2021-Septe, Society of Exploration Geophysicists, pp. 757–761.

[2] Anagaw, A. Y., and Sacchi, M. D. Model parametrization strategies for Newton-based acoustic full waveform inversion. *Journal of Applied Geophysics 157* (2018), 23–36.

[3] Brossier, R., Operto, S., and Virieux, J. Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion. *GEOPHYSICS 74*, 6 (11 2009), WCC105–WCC118.

[4] Brossier, R., Operto, S., and Virieux, J. Which data residual norm for robust elastic frequency-domain full waveform inversion? *GEOPHYSICS 75*, 3 (5 2010), R37–R46.

[5] Carneiro, M. d. S. R., Pereira-Dias, B., Soares Filho, D. M., and Landau, L. On the Scaling of the Update Direction for Multi-parameter Full Waveform Inversion: Applications to 2D Acoustic and Elastic Cases. *Pure and Applied Geophysics 175*, 1 (1 2018), 217–241.

[6] Causse, E., Mittet, R., and Ursin, B. Preconditioning of full-waveform inversion in viscoacoustic media. *GEOPHYSICS 64*, 1 (1 1999), 130–145.

[7] Choi, Y., Min, D.-J., and Shin, C. Frequency-Domain Elastic Full Waveform Inversion Using the New Pseudo-Hessian Matrix: Experience of Elastic Marmousi-2 Synthetic Data. *Bulletin of the Seismological Society of America 98*, 5 (10 2008), 2402–2415.

[8] Conn, A. R., Gould, N. I. M., and Toint, P. L. *Trust Region Methods*, vol. 3. Society for Industrial and Applied Mathematics, 2000.

[9] Consolvo, B., Zuberi, M., Pratt, R., and Cary, P. FWI with Scaled-Sobolev Preconditioning Applied to Short-offset Vibroseis Field Data. In *79th EAGE Conference and Exhibition 2017* (6 2017), pp. 10–15.

[10] Datta, D., and Sen, M. K. Estimating a starting model for full-waveform inversion using a global optimization method. *GEOPHYSICS 81*, 4 (7 2016), R211–R223.

[11] Eisenstat, S. C., and Walker, H. F. Choosing the forcing terms in an inexact Newton method. *SIAM Journal of Scientific Computing 17*, 1 (1996), 16–32.

[12] Epanomeritakis, I., Akçelik, V., Ghattas, O., and Bielak, J. A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion. *Inverse Problems 24*, 3 (6 2008), 034015.

[13] Ernst, O. G., and Gander, M. J. Why it is Difficult to Solve Helmholtz Problems with Classical Iterative Methods. *Numerical analysis of multiscale problems 83* (2012), 325–363.

[14] Fan, J., Pan, J., and Song, H. A Retrospective Trust Region Algorithm with Trust Region Converging to Zero. *Journal of Computational Mathematics 34*, 4 (6 2016), 421–436.

[15] Fan, J.-y., and Yuan, Y.-x. A new trust region algorithm with trust region radius converging to zero. *Proceedings of the 5th International Conference on Optimization: Techniques and Applications*, February (2001).

[16] Favennec, Y., Dubot, F., Le Hardy, D., Rousseau, B., and Rousse, D. R. Space-Dependent Sobolev Gradients as a Regularization for Inverse Radiative Transfer Problems. *Mathematical Problems in Engineering 2016* (2016), 1–18.

[17] Fichtner, A., and Trampert, J. Hessian kernels of seismic data functionals based upon adjoint techniques. *Geophysical Journal International 185*, 2 (5 2011), 775–798.

[18] Guitton, A., Ayeni, G., and Díaz, E. Constrained full-waveform inversion by model reparameterization. *GEOPHYSICS 77*, 2 (3 2012), R117–R127.

[19] Hinze, M., Pinnau, R., Ulbrich, M., and Ulbrich, S. *Optimization with PDE Constraints*, vol. 23 of *Mathematical Modelling: Theory and Applications*. Spinger, Dordrecht, 2009.

[20] Hu, W., Abubakar, A., Habashy, T. M., and Liu, J. Preconditioned non-linear conjugate gradient method for frequency domain full-waveform seismic inversion. *Geophysical Prospecting 59*, 3 (5 2011), 477–491.

[21] Innanen, K. A. Seismic AVO and the inverse Hessian in precritical reflection full waveform inversion. *Geophysical Journal International 199*, 2 (11 2014), 717–734.

[22] Kazemi, P., and Renka, R. Minimization of the Ginzburg–Landau energy functional by a Sobolev gradient trust-region method. *Applied Mathematics and Computation 219*, 11 (2 2013), 5936–5942.

[23] Li, D., Gurnis, M., and Stadler, G. Towards adjoint-based inversion of time-dependent mantle convection with non-linear viscosity. *Geophysical Journal International 209*, 1 (1 2017), ggw493.

[24] Lin, P., Peng, S., Lu, Y., and Du, W. The trust region method for time-domain full waveform inversion. In *Proceedings of the 7th International Conference on Environment and Engineering Geophysics & Summit Forum of Chinese Academy of Engineering on Engineering Science and Technology* (Paris, France, 2016), Atlantis Press, pp. 220–223.

[25] Métivier, L., Bretaudeau, F., Brossier, R., Operto, S., and Virieux, J. Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures. *Geophysical Prospecting 62*, 6 (11 2014), 1353–1375.

[26] Métivier, L., Brossier, R., Operto, S., and Virieux, J. Acoustic multi-parameter FWI for the reconstruction of P-wave velocity, density and attenuation: preconditioned truncated Newton approach. In *SEG Technical Program Expanded Abstracts 2015* (8 2015), vol. 34, Society of Exploration Geophysicists, pp. 1198–1203.

[27] Métivier, L., Brossier, R., Operto, S., and Virieux, J. Full Waveform Inversion and the Truncated Newton Method. *SIAM Review 59*, 1 (1 2017), 153–195.

[28] Métivier, L., Brossier, R., Virieux, J., and Operto, S. Full Waveform Inversion and the Truncated Newton Method. *SIAM Journal on Scientific Computing 35*, 2 (1 2013), B401–B437.

[29] Mora, P. Nonlinear two-dimensional elastic inversion of multioffset seismic data. *GEOPHYSICS 52*, 9 (9 1987), 1211–1228.

[30] Mulder, W., and Plessix, R.-E. Exploring some issues in acoustic full waveform inversion. *Geophysical Prospecting 56*, 6 (11 2008), 827–841.

[31] Neuberger, J. *Sobolev Gradients and Differential Equations*, vol. 1670 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[32] Nocedal, J., Wright, S. J., and Robinson, S. M. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.

[33] Operto, S., Gholami, Y., Prieux, V., Ribodetti, A., Brossier, R., Metivier, L., and Virieux, J. A guided tour of multiparameter full-waveform inversion with multicomponent data: From theory to practice. *The Leading Edge 32*, 9 (9 2013), 1040–1054.

[34] Operto, S., and Miniussi, A. On the role of density and attenuation in three-dimensional multiparameter viscoacoustic VTI frequency-domain FWI: an OBC case study from the North Sea. *Geophysical Journal International 213*, 3 (6 2018), 2037–2059.

[35] Pan, W., Innanen, K., and Margrave, G. A Comparison of Different Scaling Methods for Least-squares Migration/inversion. In *76th European Association of Geoscientists and Engineers Conference and Exhibition 2014: Experience the Energy - Incorporating SPE EUROPEC 2014* (2014), vol. 25, pp. 4325–4329.

[36] Pan, W., Innanen, K. A., and Liao, W. Accelerating Hessian-free Gauss-Newton full-waveform inversion via improved preconditioning strategies. In *SEG Technical Program Expanded Abstracts 2016* (9 2016), vol. 35, Society of Exploration Geophysicists, pp. 1455–1461.

[37] Pan, W., Innanen, K. A., and Liao, W. Accelerating Hessian-free Gauss-Newton full-waveform inversion via l-BFGS preconditioned conjugate-gradient algorithm. *GEOPHYSICS 82*, 2 (3 2017), R49–R64.

[38] Pan, W., Innanen, K. A., Margrave, G. F., Fehler, M. C., Fang, X., and Li, J. Estimation of elastic constants for HTI media using Gauss-Newton and full-Newton multiparameter full-waveform inversion. *GEOPHYSICS 81*, 5 (9 2016), R275–R291.

[39] Park, B., Ha, W., and Shin, C. A comparison of the preconditioning effects of different parameterization methods for monoparameter full waveform inversions in the Laplace domain. *Journal of Applied Geophysics 172* (2020), 103883.

[40] Plessix, R.-E. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International 167*, 2 (11 2006), 495–503.

[41] Plessix, R.-E., and Mulder, W. A. Frequency-domain finite-difference amplitude-preserving migration. *Geophysical Journal International 157*, 3 (6 2004), 975–987.

[42] Plessix, R.-E., and Mulder, W. A. Resistivity imaging with controlled-source electromagnetic data: depth and data weighting. *Inverse Problems 24*, 3 (6 2008), 1–22.

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

[44] Royer, A., Béchet, E., Geuzaine, C., Eric, B., and Geuzaine, C. GMSH-FEM : AN EFFICIENT FINITE ELEMENT LIBRARY BASED ON GMSH. *In preparation*, July (2020), 19–24.

[45] Schot, S. H. Eighty years of Sommerfeld's radiation condition. *Historia Mathematica 19*, 4 (1992), 385–401.

[46] Shin, C., Jang, S., and Min, D.-J. Improved amplitude preservation for prestack depth migration by inverse scattering theory. *Geophysical Prospecting 49*, 5 (9 2001), 592–606.

[47] Sirgue, L., and Pratt, R. G. Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies. *GEOPHYSICS 69*, 1 (1 2004), 231–248.

[48] Sourbier, F., Haidar, A., Giraud, L., Ben-Hadj-Ali, H., Operto, S., and Virieux, J. Three-dimensional parallel frequency-domain visco-acoustic wave modelling based on a hybrid direct/iterative solver. *Geophysical Prospecting 59*, 5 (9 2011), 834–856.

[49] Steihaug, T. The Conjugate Gradient Method and Trust Regions in Large Scale Optimization. *SIAM Journal on Numerical Analysis 20*, 3 (6 1983), 626–637.

[50] Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, 1 2005.

[51] Tikhonov, A. N., and Arsenin, V. Y. *Solutions of Ill-Posed Problems*. Wiley, New York, 1977.

[52] Versteeg, R. The Marmousi experience: Velocity model determination on a synthetic complex data set. *The Leading Edge 13*, 9 (9 1994), 927–936.

[53] Virieux, J., and Operto, S. An overview of full-waveform inversion in exploration geophysics. *GEOPHYSICS 74*, 6 (11 2009), WCC1–WCC26.

[54] Wang, Y., Dong, L., Liu, Y., and Yang, J. 2D frequency-domain elastic full-waveform inversion using the block-diagonal pseudo-Hessian approximation. *GEOPHYSICS 81*, 5 (2016), R247–R259.

[55] Warner, M., Ratcliffe, A., Nangoo, T., Morgan, J., Umpleby, A., Shah, N., Vinje, V., Štekl, I., Guasch, L., Win, C., Conroy, G., and Bertrand, A. Anisotropic 3D full-waveform inversion. *GEOPHYSICS 78*, 2 (2013).

[56] Yan, X., He, Q., and Wang, Y. Truncated trust region method for nonlinear inverse problems and application in full-waveform inversion. *Journal of Computational and Applied Mathematics 404* (4 2022), 113896.

[57] Yang, P., Brossier, R., Métivier, L., Virieux, J., and Zhou, W. A Time-Domain Preconditioned Truncated Newton Approach to Visco-acoustic Multiparameter Full Waveform Inversion. *SIAM Journal on Scientific Computing 40*, 4 (1 2018), B1101–B1130.

[58] Zhang, H., Li, X., Song, H., and Liu, S. An adaptive subspace trust-region method for frequency-domain seismic full waveform inversion. *Computers & Geosciences 78*, February (5 2015), 1–14.

[59] Zhang, W., and Li, Y. Elastic wave full-waveform inversion in the time domain by the trust region method. *Journal of Applied Geophysics 197*, May 2021 (2 2022), 104540.

[60] Zuberi, M. A., and Pratt, R. G. Mitigating nonlinearity in full waveform inversion using scaled-Sobolev pre-conditioning. *Geophysical Journal International 213*, 1 (4 2017), 706–725.