A priori estimates of attraction basins for nonlinear least squares, with application to Helmholtz seismic inverse problem

Hélène Barucq, Guy Chavent, Florian Faucher

A priori estimates of attraction basins for nonlinear least squares, with application to Helmholtz seismic inverse problem

Hélène Barucq[1], **Guy Chavent**[2] **and Florian Faucher**[1]

[1] Inria Project-Team Magique-3D, E2S UPPA, CNRS, Pau, France.

[2] Inria Project-Team Serena, Paris, France.

E-mail: florian.faucher@inria.fr

Abstract. In this paper, we provide an *a priori optimizability* analysis of nonlinear least squares problems that are solved by *local* optimization algorithms. We define *attraction* (*convergence*) *basins* where the misfit functional is guaranteed to have only one local - and hence global - stationary point, provided the data error is below some *tolerable error level*. We use geometry in the data space (strictly quasiconvex sets) in order to compute the size of the attraction basin (in the parameter space) and the associated tolerable error level (in the data space). These estimates are defined *a priori*, i.e., they do not involve any least squares minimization problem, and only depend on the forward map. The methodology is applied to the comparison of the optimizability properties of two methods for the seismic inverse problem for a time-harmonic wave equation: the Full Waveform Inversion (FWI) and its Migration Based Travel Time (MBTT) reformulation. Computation of the size of attraction basins for the two approaches allows to quantify the benefits of the latter, which can alleviate the requirement of low-frequency data for the reconstruction of the background velocity model.

Keywords: Time-harmonic waves, Convergence analysis, Helmholtz inverse problem, A priori estimates, Seismic, Full Waveform Inversion, Migration Based Travel Time, Quantitative reconstruction. Submitted to: *Inverse Problems*

1. Introduction

When it comes to the *inverse problem* of determining a parameter m from data $\boldsymbol{d}$, a natural and widely used approach consists in trying to minimize the data misfit functional

$$\min_{m\in\mathcal{M}} \mathcal{J}(m) = \frac{1}{2}\|\mathcal{F}(m) - \boldsymbol{d}\|^2_{\mathcal{D}}. \tag{1}$$

The data relates to m by some *forward map* $\mathcal{F}$, and $\mathcal{M}$ is an admissible parameter set which encodes the a priori knowledge on m. In a large number of situations, $\mathcal{M}$ is convex, but the map $\mathcal{F}$ is not linear, and (1) is a *nonlinear least squares problem*.

The resolution of (1) is not easy, in particular because the nonlinearity of $\mathcal{F}$ can result in several local minima in the misfit functional, which local optimization algorithm cannot avoid. The reconstruction depends on the initial guess m_{init} for the minimization algorithm, which will converge to the first stationary point it encounters - and not necessarily to the global minimum. One can figure out approximately the "attraction basin" of the global minimum in the parameter space by solving (1) with different m_{init} and synthetic data $\boldsymbol{d}$. Such an optimizability study can give only partial answers, as one cannot cover all possible combinations m_{init} and $\boldsymbol{d}$.

In opposition, we perform in this paper an *a priori optimizability study* of the least squares problem (1): we *quantify* the size of *attraction basins* and *tolerable level errors* which ensure that a local algorithm with an initial guess inside the basin will converge to the global minimum, provided the error on the data is less than the tolerable error. With this definition, $\mathcal{M}$ is an attraction basin if and only if the attainable set $\mathcal{F}(\mathcal{M})$ is strictly quasiconvex (s.q.c.), as defined in [13] where sufficient conditions and a characterization are given in terms of *deflection* and *global radius of curvature* along curves of the data space. By construction, these quantities depend solely on the parameter-to-synthetic forward map $\mathcal{F}$, but not on the data, and can be computed without solving any optimization problem. However, their numerical determinations can require a large number of evaluation of $\mathcal{F}$ and its derivatives, and becomes intractable as soon as there are more than a few parameters. This is why we shall consider only *directional attraction basins* along lines of the parameter space.

We demonstrate the interest of this optimizability approach for the analysis of the Helmholtz inverse problem in a seismic context. The problem consists in recovering the subsurface Earth properties m from wave measurements $\boldsymbol{d}$ at the surface, using the Helmholtz acoustic equation for $\mathcal{F}$. In this context, the minimization of the least squares formulation (1) is referred to as the Full Waveform Inversion (FWI). The method was introduced for time-domain acoustic problem in [3, 4] for one dimension, followed by the work of [23, 38]. The time-frequency domain formulation was then developed by Pratt et al. [32, 30, 31]. The FWI approach to seismic imaging has become more and

more popular with the increase of computational power, and it has been investigated with respect to several aspects such as the choice of misfit function, using logarithmic function [39, 34, 35] the signal envelop [8], or optimal transport distance [26, 33, 45]. Convergence of the scheme is studied in [17, 19].

However, some difficulties inherent to FWI remain: because of the long distance traveled by the signal from the source to the deep reflectors and back to the surface receivers, a small change in the low spatial frequencies of the velocity (the "background velocity") will cause phase shifts of more than one cycle in the computed wavefield, and hence create local minima in the data misfit $\mathcal{J}$ - which motivates the use of this problem to test our a priori optimizability analysis. These local minima hamper the FWI approach when it comes to the determination of the background velocity model, as local algorithms will stop at the nearest local minimum, unless the initial background velocity is already accurate, or the data contain unrealistically low frequencies [10, 36]. One way out of this dilemma could be random optimization, e.g., [20], which has the ability to find the global minimum even in presence of many parasitic local minima. But it requires a very large number of misfit evaluations for the determination of a small number of parameters, which is not well adapted to seismic inversion, when the number of parameter is very large (it is of several thousands in our applications).

Quite early, FWI has been reformulated to overcome the local minima problem, at the price of an increased computational complexity. The Differential Semblance Optimization (DSO), [37], extends the depth reflectivity model to account for the various illuminations in the data, and defines a semblance objective function to retrieve the background model. With the same objective, the Migration Based Travel Time (MBTT) reformulation of FWI has been introduced in [15, 18, 5], where the Earth model m is parameterized by a background velocity $\mathfrak{p}$ and data-space reflectivity $\mathfrak{s}$.

So we apply in this paper our a priori analysis to the determination of directional attraction basins for both the original FWI formulation and its MBTT data-space reflectivity reformulation, which allows to quantify the effectiveness of the reformulation.

The paper is organized as follows. Section 2 defines the geometrical tools, based on [13], needed to define and analyze the *optimizability* of problem (1): sufficient and necessary optimizability conditions are given, and estimates of the size of the *attraction basin* in the parameter space, and of the *tolerable error level* in the data space are derived. Section 3 presents the time-harmonic inverse problem associated with the Helmholtz equation, and the associated nonlinear iterative minimization problem for the reconstruction of the velocity model. The strategy is based upon a global model parametrization (standard FWI) or using the MBTT reformulation (background/data space decomposition). Numerical estimates of the optimizability are provided in Section 4, following the formulas of Section 2, and highlight quantitatively

the size increase of the attraction basin provided by the MBTT approach. Appendix A reviews the gradient computation in the frequency domain, emphasizing the specificity of complex valued fields for the adjoint state method. In Appendix B, we provide some experiments of reconstruction to highlight the influence of the background velocity. details on the MBTT decomposition are given in Appendix C. Note that the research report [5] contains several additional experiments with this methodology.

2. Optimizability of least squares minimization problems

In this section, we define precisely the optimizability of the general nonlinear least squares problem (1). It refers to the possibility for a local (deterministic) optimization algorithm to converge to a global minimum, without stopping prematurely in a local minimum or stationary point. This analysis follows the work of [13], whose main results are given in Subsection 2.2. In Subsections 2.3, 2.4 and 2.5, we provide the methodology to compute *numerical estimates* that evaluate (quantitatively) the optimizability of least squares problems. These estimates are *a priori* and only depend on the forward problem. We provide *local* estimates, which are a first approximation and are computationally inexpensive, and *exact* estimates which require more computations, but are more accurate.

2.1. Problem statement and definition of optimizability

We consider the (possibly nonlinear) least squares minimization problem (1) where the *forward map* (*operator*) is $\mathcal{F}$, and $\boldsymbol{d}$ denotes the *data* (observations). We shall refer to $\mathcal{F}(\mathcal{M})$ as the *attainable set* of the least squares problem.

Assumption 1. *The following set of hypotheses is required for optimizability, cf. [13].*

- *The* model space *(or* admissible parameter set*)* $\mathcal{M}$ *is a closed convex and bounded subset of the finite dimensional* parameter space E *equipped with the norm* $\|\cdot\|_E$.
- *The* data space $\mathcal{D}$ *is a finite dimensional Hilbert space, equipped with the norm* $\|\cdot\|_{\mathcal{D}}$.
- *The* forward map $\mathcal{F} : \mathcal{M} \to \mathcal{D}$ *is continuous and twice differentiable along segments of* $\mathcal{M}$.
- *There exists* $\mathcal{C} \geq 0$ *such that* $\forall m_1, m_2 \in \mathcal{M}$, $\forall t \in [0,1]$,

$$\|D_t \mathcal{F}((1-t)m_1 + tm_2)\|_{\mathcal{D}} \leq \mathcal{C}\|m_2 - m_1\|_E,$$

where D_t *stands for the derivative with respect to* t.

The parameter and data space have been taken finite dimensional for convenience only, in order to avoid technical difficulties, but the theory can be put to work in an infinite dimensional setting.

Definition 1 (Path)**.** *A curve P drawn on $\mathcal{F}(\mathcal{M}) \subset \mathcal{D}$ is a* path *of $\mathcal{F}(\mathcal{M})$ if it is of the form:*

$$P \ : \ t \in [0,1] \to \mathcal{F}((1-t)m_1 + tm_2) \quad \text{where } m_1, m_2 \text{ are two parameters of } \mathcal{M}. \tag{2}$$

Definition 2 (Velocity and acceleration)**.** *P is twice differentiable and we denote by V and A (velocity and acceleration along P) its two first derivatives:*

$$V(t) = P'(t), \qquad A(t) = P''(t). \tag{3}$$

For simplicity, we shall consider only paths for which $V(t) \neq 0$ for all t, so we can define the unit tangent velocity v, and the normal acceleration a by

$$v(t) = \frac{V(t)}{\|V(t)\|_{\mathcal{D}}}, \qquad a(t) = \frac{A(t) - \langle A(t), v(t)\rangle_{\mathcal{D}}\ v(t)}{\|V(t)\|^2_{\mathcal{D}}}, \tag{4}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ is the inner product in $\mathcal{D}$.

Due to the limited accuracy of the recording devices, model error and noise, the observed data $\boldsymbol{d}$ do not belong in general to the attainable set $\mathcal{F}(\mathcal{M})$. Therefore, it is important that the least squares misfit function does not have parasitic local minima for data $\boldsymbol{d}$ which are "not too far" from the attainable set. This property is made precise by the following definition.

Definition 3 (Optimizability/Attraction Basin)**.** *The least squares problem (1) is* optimizable *on $\mathcal{M}$, or equivalently the* parameter set *$\mathcal{M}$ is an* attraction basin *for (1), if there exists a neighborhood $\mathcal{V}$ of $\mathcal{F}(\mathcal{M})$ such that*

- *uniqueness: all data $\boldsymbol{d} \in \mathcal{V}$ have a unique projection $\boldsymbol{d}_\dagger$ on $\mathcal{F}(\mathcal{M})$,*
- *unimodality: for any $\boldsymbol{d} \in \mathcal{V}$, the distance to $\boldsymbol{d}$ has no parasitic stationary point over $\mathcal{F}(\mathcal{M})$,*
- *convergence: if $\boldsymbol{d} \in \mathcal{V}$, any minimizing sequence $d_n \in \mathcal{F}(\mathcal{M})$ of the distance to $\boldsymbol{d}$ is a Cauchy sequence for both the norm $\|\cdot\|_{\mathcal{D}}$ and the arc length distance $\ell(P)$ along the path P defined by (2). Hence d_n converges in F to the unique projection $\boldsymbol{d}_\dagger$ of $\boldsymbol{d}$ onto $\mathcal{F}(\mathcal{M})$.*

Therefore, the absence of local minimum (unimodality) and uniqueness of the projection guarantee that the resolution of an *optimizable* (Definition 3) least squares problem by a local gradient algorithm will converge to a global (but not necessarily unique) minimizer, whatever the initial guess in its *basin of attraction* $\mathcal{M}$.

Remark 1. *The size of attraction basins depends only on the forward map to be inverted, but not on the optimization algorithm used (e.g. Newton method, gradient descent, etc). The choice of the method naturally affects the rate of convergence and the speed at which the final solution is eventually reached, but it has no influence on the presence or absence of local minimum, and none of the deterministic local algorithms is robust with respect to local minimum. On the other hand, for an* optimizable *problem as given in Definition 3, local algorithms would be able to find the solution because any local minimum in the attraction basin is a global minimum in such cases.*

2.2. Global Radius of Curvature and Deflection

Following [13, pp. 167–172 and 300-308], we define the *global radius of curvature* and the *deflection* along a path P, and further give in Propositions 1 and 2 a characterization and a sufficient condition of optimizability.

Definition 4 (Radius of curvature)**.** *The (possibly infinite) radius of curvature $R(t)$ of a path P at t is given by:*

$$\frac{1}{R(t)} = \|a(t)\|_{\mathcal{D}} = \frac{\|A(t)\|_{\mathcal{D}}}{\|V(t)\|_{\mathcal{D}}^2} \sin(A(t), V(t)). \tag{5}$$

The radius of curvature of the whole path P is then defined as:

$$\frac{1}{R(P)} \stackrel{\text{def}}{=} \sup_{t\in[0,1]} \frac{1}{R(t)}. \tag{6}$$

It is straightforward to see that

$$\frac{1}{R(t)} \leq \frac{\|A(t)\|_{\mathcal{D}}}{\|V(t)\|_{\mathcal{D}}^2}, \qquad \text{for a.e. } t \in [0,1]. \tag{7}$$

Definition 5 (Global radius of curvature)**.** *The (possibly infinite)* global radius of curvature $R_G(t,t')$ *of a path P at t seen from t', with $t \neq t'$, is given by:*

$$R_G(t,t') = \begin{cases} N^+/D & \text{if } \langle v(t), v(t')\rangle_{\mathcal{D}} \geq 0, \\ N^+ & \text{if } \langle v(t), v(t')\rangle_{\mathcal{D}} \leq 0, \end{cases} \tag{8}$$

where $v(t), v(t')$ are the normalized velocities along P defined by (4), and

$$\begin{cases} N^+ &= max(N,0) \quad \text{where } N = \text{sign}(t'-t)\langle P(t') - P(t), v(t')\rangle_{\mathcal{D}}, \\ D &= \left(1 - \langle v(t), v(t')\rangle_{\mathcal{D}}^2\right)^{1/2}. \end{cases} \tag{9}$$

The global radius of curvature of the path P is then defined by

$$R_G(P) \stackrel{\text{def}}{=} \inf_{t,t'\in[0,1]} R_G(t,t') \geq 0. \tag{10}$$

The interest of global radius of curvature comes from the following proposition.

Proposition 1 ($R_G > 0 \iff$ optimizability)**.**
The least squares problem (1) is optimizable *- or equivalently $\mathcal{M}$ is an* attraction basin *for (1) - if and only if there exists $R_G > 0$ such that $R_G(P) \geq R_G > 0$ for all path P of $\mathcal{F}(\mathcal{M})$. The associated neighborhood $\mathcal{V}$ is defined by:*

$$\mathcal{V} = \{\boldsymbol{d} \in F \quad | \quad \mathrm{dist}(\boldsymbol{d}, \mathcal{F}(\mathcal{M})) < R_G\}. \tag{11}$$

The proofs can be found in [13]. The global radius of curvature can be computed numerically using (8) and (9), as will be done in Sections 4. It can also be estimated via the usual radius of curvature depending on the value of the *deflection*, which we define now, and which is illustrated Figure 1(a).

Definition 6 (Deflection)**.** *The deflection between two points t and t' of the curve P is the angle between the two velocities $V(t)$ and $V(t')$ (see Figure 1(a)). It is given by:*

$$\Theta(t, t') = \arccos\Big(\frac{\langle\, V(t), V(t')\,\rangle_{\mathcal{D}}}{\|V(t)\|_{\mathcal{D}}\|V(t')\|_{\mathcal{D}}}\Big) \in [0, \pi[. \tag{12}$$

The deflection $\Theta(P)$ of the curve P is defined as the largest angle $\Theta(t, t') \in [0, \pi]$ between any two tangent vectors $V(t)$ and $V(t')$ for any two points t and t' of $[0, 1]$. An infinitesimal variation of the deflection $\mathrm{d}\Theta$ satisfies

$$\mathrm{d}\Theta \leq \frac{\|A(t)\|_{\mathcal{D}}}{\|V(t)\|_{\mathcal{D}}}\,\mathrm{d}t. \tag{13}$$

Denoting t_1 and t_2 the values of t for which the deflection is maximum, the deflection $\Theta(P)$ along the curve P satisfies

$$\Theta(P) = \int_{t_1}^{t_2} \mathrm{d}\Theta \leq \int_0^1 \frac{\|A(t)\|_{\mathcal{D}}}{\|V(t)\|_{\mathcal{D}}}\,\mathrm{d}t. \tag{14}$$

This upper bound is sharp, but it is very conservative: equality holds only when P is an arc of circle with constant velocity $\|V(t)\|$, i.e. *when the path P turns always in the same direction with a constant radius.*

The relation between global and local radii of curvature is then given by the following proposition.

Proposition 2 (Local and Global Radii of curvature)**.** *For any path P of $\mathcal{F}(\mathcal{M})$ one has*

$$R(P) \geq R_G(P) \geq 0 \quad \textit{and} \quad R(P) = R_G(P) \quad \textit{as soon as} \quad \Theta(P) \leq \pi/2, \tag{15}$$

Definition 7 (Finite Curvature/Limited deflection (FC/LD) problem)**.** *The minimization Problem (1) is a FC/LD least squares problem if:*

$$\begin{aligned} &\textit{there exists } R > 0 \textit{ such that:} \quad \|A(t)\|_{\mathcal{D}} \leq \frac{1}{R}\|V(t)\|_{\mathcal{D}}^2 \\ &\textit{for a.e. } t \in [0, 1] \textit{ and all paths } P, \end{aligned} \tag{16}$$

$$\Theta(P) \leq \frac{\pi}{2} \quad \textit{for all paths } P. \tag{17}$$

From Definition 7 and using (15), a FC/LD problem verifies that

$$R_G(P) = R(P) \geq R > 0 \quad \text{for all paths } P, \tag{18}$$

which shows that FC/LD problems (also referred to as *weakly nonlinear* inverse problem in [17]) are necessarily optimizable.

Notice that Proposition 1 gives a characterization of optimizable problems, whereas Definition 7 provides only a sufficient condition.

2.3. Directional Attraction Basins

Numerical application of previous section to evaluate whether or not a given least squares problem is optimizable becomes quickly intractable when the number of parameters increases, as it is the case in seismic inversion. So we limit ourselves to *directional* (or *one-dimensional*) *parameter sets* of the form:

$$\mathcal{M}(m_0, \mathfrak{u}, \Delta) = [m_0 - \Delta\mathfrak{u}, m_0 + \Delta\mathfrak{u}] \quad \text{with} \quad \|\mathfrak{u}\|_E = 1. \tag{19}$$

Here, m_0 is a nominal model, $\mathfrak{u}$ a normalized perturbation direction, and Δ gives the size of the domain of investigation. The associated attainable set is the image of the path P defined by:

$$P \ : \quad t \in [0,1] \rightsquigarrow \mathcal{F}(m_0 + (2t-1)\Delta\mathfrak{u}). \tag{20}$$

We refer to directional optimizability when the problem is optimizable for an interval such as (19), and this interval is a directional attraction basin. Directional optimizability is only a necessary condition for optimizability, but it will allow to analyze the behavior of seismic inverse problems and to compare formulations: the size Δ of a directional attraction basin in a descent direction tells us how far one can move away in this direction without being stopped by parasitic local minima. Our objective now is to determine (see illustration Figure 1):

(i) the *size* $\Delta^{\mathfrak{u}}_{m_0}$ of the directional attraction basin centered at m_0. The larger $\Delta^{\mathfrak{u}}_{m_0}$, the better the least squares problem is amenable to minimization by local algorithm, because we allow a larger area for investigation. In our numerical experiments, we shall scale the estimate with the norm of the nominal model, $\|m_0\|_E$, to provide relative (to the model) quantity.

(ii) the associated *tolerable error level* $R^{\mathfrak{u}}_{G,m_0}$. It is the largest tolerable error on the data $\boldsymbol{d}$ which ensures the absence of parasitic local minima for the least squares objective function

$$t \in [0,1] \rightsquigarrow \frac{1}{2}\|\mathcal{F}(m_0 + (2t-1)\Delta^{\mathfrak{u}}_{m_0}\mathfrak{u}) - \boldsymbol{d}\|^2_{\mathcal{D}} \tag{21}$$

over $[0,1]$. The larger $R^{\mathfrak{u}}_{G,m_0}$ is, the better is the robustness of the minimization procedure to noise in the data. In our numerical experiments, we divide the

estimates with the norm of the synthetic data $\boldsymbol{d}_0 = \mathcal{F}(m_0)$ to provide relative (to the data) quantity.

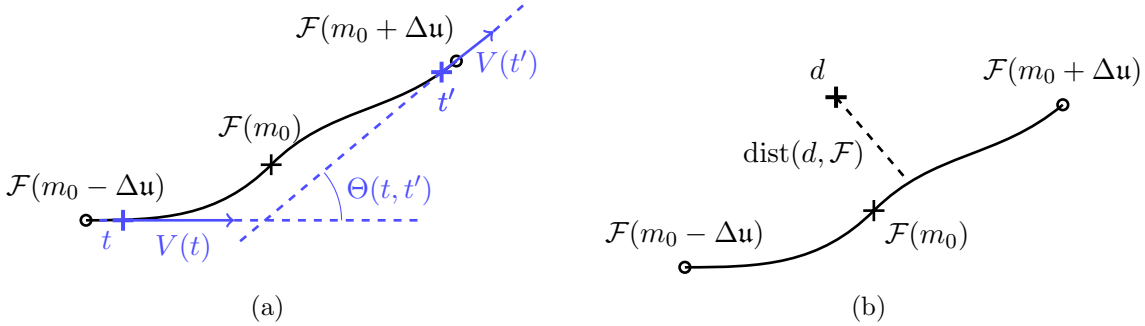


Figure 1. A one-dimensional setup for least squares problems. The figure lives in the data space, the attainable set is the path P image of the interval $\mathcal{M}(m_0, \mathfrak{u}, \Delta)$ of the model space. (a) Illustration of the computation of the deflection $\Theta(t, t')$ between two arbitrary points t and t'. (b) The path has a finite curvature and the deflection is smaller than $\pi/2$, so the FC/LD Property 7 is satisfied, and $R_G = R > 0$ by Proposition 2. Hence the "distance to $\boldsymbol{d}$" function cannot have local minimum over P provided the data $\boldsymbol{d}$ is at a distance of the attainable set $P = \mathcal{F}(\mathcal{M}(m_0, \mathfrak{u}, \Delta))$ smaller than R.

We shall use two types of estimate:

- Θ-estimates of $\Delta^{\mathfrak{u}}_{m_0}$, where the optimizability over $\mathcal{M}$ is obtained by satisfying the sufficient condition $\Theta(P) \leq \pi/2$ of Definition 7. In this case, $R_G(P) = R(P)$, so the tolerable error level is given by the minimum over the $[0, 1]$ interval of $R(t)$ given by (5).
- R_G-estimates of $\Delta^{\mathfrak{u}}_{m_0}$, where optimizability is obtained by satisfying the $R_G > 0$ characterization of optimizability of Proposition 1. In this case, the associated tolerable error level R_G has to be computed by evaluating numerically the infimum in (10) using (8) and (9).

2.4. Local Θ-estimate for $\Delta^{\mathfrak{u}}_{m_0}$ and associated tolerable error level $R^{\mathfrak{u}}_{m_0}$

We provide here a *local* Θ*-estimate* Δ of the attraction basin, in the sense that it is *based only on* the velocity V and acceleration A *at* m_0 *in the direction* $\mathfrak{u}$. In order to ensure that the deflection of the path P defined by (20) is smaller than $\pi/2$, we use the upper bound (14) of $\Theta(P)$, which, according to the optimizability condition of Definition 7, ensures in turn that $\mathcal{M}(m_0, \mathfrak{u}, \Delta)$ is an attraction basin.

With the notations of [11] for the directional derivative (indicated between

parenthesis), the chain rule differentiation gives:

$$\begin{cases} V(t) = \dfrac{\partial \mathcal{F}}{\partial m}\dfrac{\partial m}{\partial t} = 2\Delta D\mathcal{F}(m)(\mathfrak{u}), \\ A(t) = \dfrac{\partial^2 \mathcal{F}}{\partial m^2}\Big(\dfrac{\partial m}{\partial t}\Big)^2 + \dfrac{\partial \mathcal{F}}{\partial m}\dfrac{\partial^2 m}{\partial t^2} = 4\Delta^2 D^2\mathcal{F}(m)(\mathfrak{u},\mathfrak{u}), \end{cases} \tag{22}$$

where $\mathfrak{u}$ acts as the direction of derivation. Then we use a rectangle approximation in (14), which gives the approximate *upper bound* $\Theta^{\mathfrak{u}}_{m_0}$ to the deflection $\Theta(P)$:

$$\Theta(P) \leq \int_0^1 \frac{\|A(t)\|_{\mathcal{D}}}{\|V(t)\|_{\mathcal{D}}}\,\mathrm{d}t \sim \frac{\|A(1/2)\|_{\mathcal{D}}}{\|V(1/2)\|_{\mathcal{D}}} = 2\Delta\frac{\|D^2\mathcal{F}(m_0)(\mathfrak{u},\mathfrak{u})\|_{\mathcal{D}}}{\|D\mathcal{F}(m_0)(\mathfrak{u})\|_{\mathcal{D}}} \stackrel{\text{def}}{=} \Theta^{\mathfrak{u}}_{m_0}. \tag{23}$$

This gives immediately a *local* Θ*-estimate* of the size Δ of an attraction basin at m_0 in the direction $\mathfrak{u}$:

$$\Delta^{\mathfrak{u}}_{m_0} = \frac{\pi}{4}\frac{\|D\mathcal{F}(m_0)(\mathfrak{u})\|_{\mathcal{D}}}{\|D^2\mathcal{F}(m_0)(\mathfrak{u},\mathfrak{u})\|_{\mathcal{D}}}. \tag{24}$$

This estimate is an *approximate* (because of the rectangle approximation of the integral) *lower bound* (because it is based on the upper bound (14)) to the size of the largest attraction basins at m_0 in the direction $\mathfrak{u}$ but it is computationally cheap.

The associated *tolerable error level* $R^{\mathfrak{u}}_{m_0}$ is then the minimum of the radius of curvature along $P = \mathcal{F}(\mathcal{M}(m_0,\mathfrak{u},\Delta))$, which is approximated by its value at m_0, that is $R(t=1/2)$ given by (5):

$$R^{\mathfrak{u}}_{m_0} = \Big(\frac{\|V(t)\|^2_{\mathcal{D}}}{\|A(t)\|_{\mathcal{D}}|\sin\left(A(t),V(t)\right)|}\Big)|_{t=\frac{1}{2}}. \tag{25}$$

where $V(t)$ and $A(t)$ have been defined in (22).

2.5. Exact Θ- and R_G-estimates of $\Delta^{\mathfrak{u}}_{m_0}$ and associated tolerable error levels

The determination of *the exact* Θ*- and* R_G*-estimates* of the attraction basin centered at m_0 in a direction $\mathfrak{u}$ inside an interval $\mathcal{M}(m_0,\mathfrak{u},\Delta)$ of given size Δ requires the numerical computation of the deflection $\Theta(t,t')$ and the global radius of curvature $R_G(t,t')$ between any two points

$$\mathcal{F}(m_0+t\mathfrak{u}),\ \ -\Delta \leq t \leq \Delta \quad \text{and} \quad \mathcal{F}(m_0+t'\mathfrak{u}),\ \ -\Delta \leq t' \leq \Delta \tag{26}$$

of the path P, which is the image by $\mathcal{F}$ of the investigated interval $\mathcal{M}(m_0,\mathfrak{u},\Delta)$.

For this purpose, *deflection maps* and *global radius maps* are computed, which display the values of $\Theta(t,t')$ (Definition 6) and of $R_G(t,t')$ (Definition 5) between the points of $\mathcal{M}(m_0,\mathfrak{u},\Delta)$. On the diagonal of the maps, where $t=t'$, R_G is not defined by (8) (9), and we indicate instead the values of $R(t)$ given by (4) (5), which represent the limits of $R_G(t,t')$ when $t' \to t$. One can then read on these maps (cf. Section 4):

– the exact Θ-estimate of the attraction basin size $\Delta^{\mathfrak{u}}_{m_0}$, given by the largest square centered at $(0,0)$ where $\Theta(t,t') \leq \pi/2$ for all t, t';
– the exact R_G-estimate of the attraction basin size $\Delta^{\mathfrak{u}}_{m_0}$, given by the largest square centered at $(0,0)$ where $R_G(t,t') > 0$ for all t, t'.

During this process, when the size of the investigated square increases from 0 to the exact Θ-estimate, the associated exact tolerable error $R^{\mathfrak{u}}_{m_0} = \inf_{-\Delta \leq t \leq \Delta} R(t)$ decreases from its value R_0 at m_0 to the tolerable error $R^{\mathfrak{u}}_{m_0}$ of the Θ-attraction basin. When the size of the square increases further to the exact R_G-estimate, the tolerable error is $R_G = \inf_{-\Delta \leq t,t' \leq \Delta} R_G(t,t')$, which continues to decrease, until it reaches the value 0 of the R_G-attraction basin. Naturally, the *exact* estimates are computationally more demanding than the *local* estimates, as they require evaluation of R_G and Θ for many couples (t,t').

3. The Helmholtz inverse problem for seismic

To illustrate the optimizability study of Section 2, we describe now two formulations of a seismic inverse problem associated to the Helmholtz equation: the objective is here to reconstruct the sound velocity in the Earth (the parameter) given partial surface measurements of reflected (backscattered) energy (the data), obtained from one side illumination (the surface). Of course, the methodology developed in Section 2 is not restricted to inverse wave problem or geophysical setup, and can be applied in any context involving (nonlinear) least squares minimization schemes.

3.1. Time-harmonic wave equations

We consider a bounded domain Ω of $\mathbb{R}^2$ with boundary $\partial\Omega$, which represents the region of interest (the analysis holds similarly in three dimensions). We consider the Helmholtz equation where the pressure field p is solution to,

$$\begin{cases} -(\omega^2 c^{-2}(\mathbf{x}) - \Delta)p(\mathbf{x}) = g(\mathbf{x}), & \text{in } \Omega, \\ p(\mathbf{x}) = 0, & \text{on } \partial\Omega. \end{cases} \tag{27}$$

The angular frequency is ω, the velocity (wavespeed) of the medium is denoted by $c(\mathbf{x})$ and the (interior) source of the phenomenon is g. The domain boundary is divided into $\partial\Omega = \Gamma_1 \cup \Gamma_2$, where we distinguish the upper free surface (physical interface, Γ_1) from the rest of the boundary (artificial boundary, Γ_2), see Figure 2. Due to the numerical truncation of the real domain (the Earth), appropriate conditions are imposed on Γ_2 to ensure that waves that reach Γ_2 are not reflected back to the domain. Here, we consider

Perfectly Matched Layers (PML, see [6]), which rewrite the derivative formula in the *layers* (sides and bottom of the domain here, see Figure 2):

$$\partial_x \to (1 + \mathrm{i}\frac{\sigma(x)}{\omega})^{-1}\partial_x, \qquad \text{in } \Omega_{\Gamma_x} \text{ (Perfectly Matched Layer)}, \tag{28}$$

and analogously for the other direction, in Ω_{Γ_z}. In our implementation, the damping function σ is defined following the work of [41, 44].

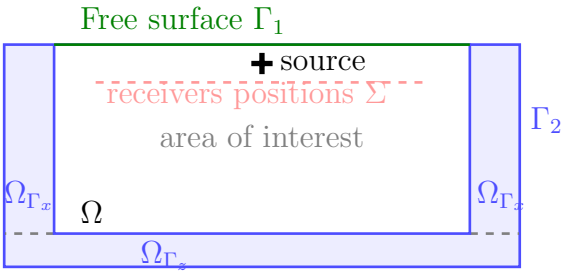


Figure 2. Illustration of the two-dimensional computational domain using Perfectly Matched Layers (PML) at the lateral and bottom boundaries. The sources and receivers that generate the data are located in the upper part, creating partial, backscattered data, according to a seismic configuration.

3.2. Inversion via classical FWI: global model representation

The FWI formulation is the most natural one: the parameter is the squared slowness $m = c^{-2}$, discretized at the n cells or nodes of a grid covering Ω: $m \in E = \mathbb{R}^n$, the data $\boldsymbol{d} \in \mathcal{D} = \mathbb{C}^q$ consist in $q = n_{rcv} \times n_{src} \times n_{freq}$ complex measurements of p at the receivers locations Σ for a finite number of sources g and frequencies ω. The spaces E and $\mathcal{D}$ are equipped with the norms:

$$\begin{aligned} &\|m\|_E = \Big(\sum\nolimits_{i=1}^{n} m_i^2\Big)^{1/2}, \qquad \|\boldsymbol{d}\|_{\mathcal{D}} = \Big(\sum\nolimits_{i=1}^{q} d_i\overline{d_i}\Big)^{1/2}, \\ &\langle d, d'\rangle_{\mathcal{D}} = Re\Big(\sum\nolimits_{i=1}^{q} d_i\bar{d}'_i\Big), \end{aligned} \tag{29}$$

where m_i and d_i are the i^{th} component of m and $\boldsymbol{d}$ respectively, and $\bar{}$ denotes the complex conjugate. Note that the representation of m with piecewise constant function over a partition of Ω is also used to estimate the stability of the inverse problem, see [1, 7].

The essence of FWI ([38, 32, 43]) is to reconstruct the subsurface properties by minimizing a misfit functional defined as the difference between the observed and simulated signals, starting from an initial model. The information on the deep Earth structure is brought by backscattered energy only, so one has to suppress the energy that has traveled directly from the source to the receivers (direct arrivals) from the observed

and simulated data. Let us denote by $p_\omega^{(g)}$ and $p_{s,\omega}^{(g)}$ the solutions of (27) for m and m_s, for the source g at frequency ω. Here, m_s is a 'smooth' version of m, close enough to m near the surface to generate the same direct arrivals, and smooth enough so that it scatters back negligible energy. So we define the forward operator $\mathcal{F} : E \rightsquigarrow \mathcal{D}$ by:

$$\mathcal{F}(m) = \left\{ \mathcal{F}_\omega^{g,\mathbf{x}}(m) = p_\omega^g(\mathbf{x}) - p_{s,\omega}^g(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \Sigma, g, \omega \right\} \in \mathcal{D}. \tag{30}$$

When computing derivatives of $\mathcal{F}$, it will be necessary to remember that m_s depends also on m. Then, FWI amounts to solve the nonlinear least squares minimization problem (1), which we recall here for convenience:

$$\min_{m \in \mathcal{M}} \mathcal{J}(m) = \frac{1}{2} \|\mathcal{F}(m) - \boldsymbol{d}\|_{\mathcal{D}}^2. \tag{31}$$

The minimization is usually performed by a Quasi-Newton algorithm, which requires only the gradient of the cost function:

$$\nabla \mathcal{J}(m) = D\mathcal{F}(m)^* \left(\mathcal{F}(m) - \boldsymbol{d}\right), \tag{32}$$

where $D\mathcal{F}$ stands for the Fréchet derivative of $\mathcal{F}$ and * is the adjoint. This gradient can be efficiently computed by the adjoint method, which does not require the formation of the Jacobian matrix. Appendix A describes a careful adaptation of this method to the case of complex variables (contrarily to the time-domain formulation, the data and the wavefields are, in the harmonic formulation, complex).

As recalled in the introduction, the determination of the background (low spatial frequencies) of m by (31) is hampered by the many local minima of $\mathcal{J}$, caused by phase shifts in the synthetics (see Figure 6). This can be overcome only if the data contain very low frequencies. These difficulties are a motivation for alternative techniques such as the MBTT reformulation of FWI below, and for the optimizability study developed in this paper.

*3.3. Inversion via MBTT–FWI (*background/data-space-reflectivity *decomposition)*

In the MBTT (Migration-Based Traveltime) approach, see [18], the model m is parameterized by a *smooth background* $\mathfrak{p} \in E$ and a *data-space reflectivity* $\mathfrak{s} \in \mathcal{D}$ using a *migration* operator:

$$m = m(\mathfrak{p}, \mathfrak{s}) = \mathfrak{p} + \mathfrak{r} = \mathfrak{p} + \mathcal{W}\, D\mathcal{F}(\mathfrak{p})^*\, \mathfrak{s} = \mathfrak{p} + \sum_\omega \mathcal{W}(\omega) D\mathcal{F}_\omega(\mathfrak{p})^* \mathfrak{s}(\omega), \tag{33}$$

where $\mathfrak{r}$ is the depth reflectivity associated to $\mathfrak{s}$ and $\mathfrak{p}$; $\mathcal{W}$ is a scaling operator (which possibly depends on the frequency) and * denotes the adjoint. The weight $\mathcal{W}$ is meant to compensate for the lower amplitude of deep migrated events, see [28]. In our experiments we use a simple scaling proportional to the square root of the depth. We refer to Appendix C for more details regarding the computational aspects of the

decomposition. Note that the solution of the linearized version (Born approximation) of the FWI problem (31) is of the form (33), in which case parameterization (33) is *not underparameterizing*. Hence, when the full model (27) is used, the parameterization by the data space reflectivity $\mathfrak{s}$ will be able to generate all primary events of the data, but maybe not all *multiple* events (i.e., it will miss the events associated to multiple reflections involving at least one reflector which generates no primary reflection), cf. [14].

When this decomposition is employed, the natural choice for the smooth version m_s of m in (30) (in order to suppress direct arrivals in the forward map $\mathcal{F}$) is simply $m_s = \mathfrak{p}$. With this change of parameter, the forward map given by (30) rewrites

$$\mathsf{F}(\mathfrak{p}, \mathfrak{s}) \stackrel{\text{def}}{=} \mathcal{F}(m) \qquad \text{with } \mathcal{F} \text{ given by (30) and } m \text{ by (33).} \tag{34}$$

By construction, $\mathcal{F}$ does not contain the direct arrivals, which implies that, for a background $\mathfrak{p}$ smooth enough, F satisfies:

$$\mathsf{F}(\mathfrak{p}, 0) = \mathcal{F}(\mathfrak{p}) \approx 0. \tag{35}$$

The motivation for this parameterization is to eliminate phase shifts induced in the synthetics by changes in the background $\mathfrak{p}$: the events in the synthetics are obtained from the *data-space* reflectivity $\mathfrak{s}$ by migration followed by simulation with *the same kinematic*, and hence are expected to have the same phase as those of $\mathfrak{s}$, as illustrated in Figure 6. Besides controlling the phase, this migration-demigration process has the additional property that the stack involved in $D\mathcal{F}_\omega(\mathfrak{p})^*$ turns, for a fixed data space reflectivity $\mathfrak{s}$, the data misfit into a coherency measure for the current background $\mathfrak{p}$, [14]. The price to pay is that the computational times are multiplied by three, because the evaluation of $D\mathcal{F}(\mathfrak{p})^*$ in (33) requires the resolution of two Helmholtz problems (see Appendix C) and the evaluation of $\mathcal{F}(m)$ in (34) requires the resolution of one additional Helmholtz equation (i.e. total of three forward problems instead of one).

Then, the MBTT–FWI minimization problem is

$$\min_{\mathfrak{p}\in\mathcal{M}_s, \mathfrak{s}\in\mathcal{D}} \mathsf{J}(\mathfrak{p}, \mathfrak{s}) = \frac{1}{2}\|\mathsf{F}(\mathfrak{p}, \mathfrak{s}) - \boldsymbol{d}\|^2_{\mathcal{D}}, \tag{36}$$

where $\mathcal{M}_s \subset \mathcal{M}$ is the *set of admissible smooth backgrounds*, $\mathcal{D}$ is the data space.

This approach has been shown successful in [18, 16] for the inversion of synthetic data, in particular when low frequencies are missing. Hence another motivation for the optimizability study of this paper is to quantify how far the MBTT reformulation of FWI succeeds in overcoming the local minima problem inherent to classical FWI.

4. Comparison of optimizability for FWI and MBTT

In this section, we analyze numerically the directional optimizability of the two least squares minimization problem of Section 3, using the computational estimates

obtained in Section 2. Namely, optimizability of the original FWI problem (31) (where the unknown model is the squared slowness m) is compared to that of the MBTT formulation (36) of FWI (where the unknown model is parametrized by a smooth background $\mathfrak{p}$ and a data space reflectivity $\mathfrak{s}$). Our objective is twofold:

- compare *local* estimates of attraction basins and tolerable error, which are numerically inexpensive, with the, more expensive, *exact* ones.
- Quantify the gain with respect to optimizability - if any - of the MBTT formulation over the classical FWI.

Remark 2. *The research report [5] associated with this paper contains several additional experiments where the same methodology is applied to analyze the convergence properties of least squares minimization. In particular, [5, Section 4] investigates the optimizability properties of global model reconstruction in FWI with respect to the search direction geometry, and the use of sequential or multiple (possibly complex) frequency data; the experiments are extended for elasticity and alternative boundary conditions problems in [5, Section 6].*

4.1. Choice of a nominal model

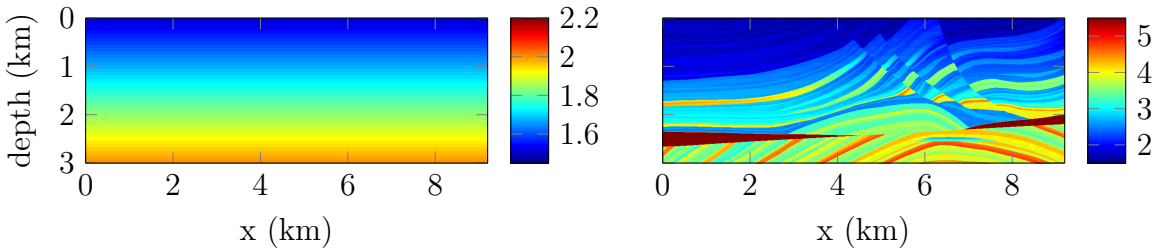
For the numerical estimates, we consider a two-dimensional geophysical setup for the Helmholtz equation (27), with a domain of size 9.2×3 km. The domain follows the *Marmousi* model, which is a geophysical subsurface wavespeed profile designed by the Institut Français du Pétrole (IFP) in the late eighties, [42], see Figure 3(b). We consider $n_{src} = 19$ sources and $n_{rcv} = 183$ receivers associated with each source (the receivers remain in the same position for all sources). Both are located near the surface, according to Figure 2. Therefore, we work with reflection data obtained from a one side (the surface) illumination. For a given frequency ω, the forward map $\mathcal{F}_\omega$ associates a vector $\mathcal{F}_\omega(m)$ of $\mathbb{C}^{n_{rcv} \times n_{src}}$ to any squared slowness model m.

For a fair comparison, the two inversion approaches (FWI and MBTT) have to be applied to *the same nominal model*, so we construct a model $\mathfrak{m}_0$ whose MBTT decomposition, $\mathfrak{p}_0$, $\mathfrak{s}_0$, is known exactly, i.e. which satisfies

$$\mathfrak{m}_0 = m(\mathfrak{p}_0, \mathfrak{s}_0) \text{ according to (33), it implies that } \mathcal{F}(\mathfrak{m}_0) = \mathsf{F}(\mathfrak{p}_0, \mathfrak{s}_0). \quad (37)$$

We first choose the smooth background $\mathfrak{p}_0$ as the one-dimensional ramp pictured in Figure 3(a). Note that our figures plot the wavespeed (in $\mathrm{km\,s^{-2}}$) per consistency with the geophysical settings but we remind that we have chosen the squared slowness as unknown parameter i.e. $\mathfrak{p}_0 = c_0^{-2}$.

Then we choose for nominal $\mathfrak{s}_0$ the first guess approximation of the data space reflectivity of the Marmousi model of Figure 3(b), given by the Marmousi synthetic



(a) Smooth background model c_0.

(b) Marmousi velocity model c_m.

Figure 3. Wavespeed models of size 9.2×3 km used for the numerical estimates, the values are indicated in $\mathrm{km\,s^{-1}}$. For the computation, the models are $\mathfrak{p}_0 = c_0^{-2}$, $m_m = c_m^{-2}$.

section deprived from its direct arrivals:

$$\mathfrak{s}_0(\omega) = \mathcal{F}_\omega(m_m) \qquad \text{from (30), using } m_s = \mathfrak{p}_0, \tag{38}$$

where m_m and $\mathfrak{p}_0$ are shown in Figure 3. Hence, we remove the direct arrivals given by $\mathfrak{p}_0$ from pressure fields simulated with m_m.

Finally, the nominal model $\mathfrak{m}_0$ is simply *defined* by (37), that is

$$\mathfrak{m}_0 = \mathfrak{p}_0 + \mathfrak{r}_0, \quad \mathfrak{r}_0 = \sum_\omega \mathfrak{r}_0(\omega), \quad \mathfrak{r}_0(\omega) = \mathcal{W}(\omega) D\mathcal{F}_\omega^*(\mathfrak{p}_0)\mathfrak{s}_0(\omega) \quad \forall \omega. \tag{39}$$

where $\mathfrak{r}_0$ (respectively $\mathfrak{r}_0(\omega)$) is the depth reflectivity associated to the sum of all frequencies (respectively to frequency ω‡).

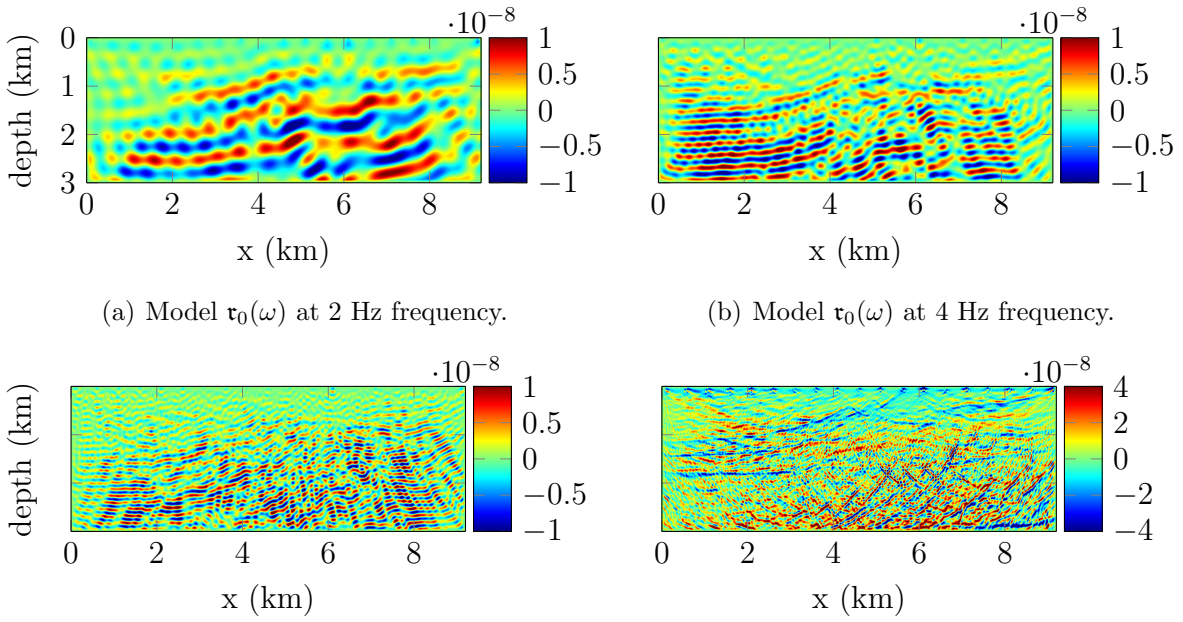
We choose the weight $\mathcal{W}$ proportional to the square root of depth, as proposed in Section 3, and adjust its amplitude by (arbitrarily) imposing a model reflectivity level of 1%,

$$\|\mathfrak{r}_0(\omega)\|/\|\mathfrak{p}_0\| = 10^{-2}, \qquad \forall\ \omega. \tag{40}$$

In Figure 4, we illustrate the resulting models $\mathfrak{r}_0(\omega)$ for three frequencies: 2, 4 and 7 Hz. We also show the model $\mathfrak{r}_0$ where the frequency sum contains frequencies between 0.5 to 15 Hz, with 0.5 Hz increment. We observe that the reflectivity, defined from the difference between observations and simulations using a smooth background, provides structures of size consistent with the selected frequency. For the global model, shown in Figure 4(d), we see the contributions of all wavelengths, and we can distinguish some structures of the Marmousi medium given in Figure 3(b).

For simplicity, in the following, we restrict ourselves by studying only *single frequency* nominal models, which means that we only work with models resulting from

‡ Note that with $\mathfrak{s}_0$ given by (38), the resulting $\mathfrak{r}_0(\omega)$ is the gradient of the misfit function (1) at frequency ω at $m = \mathfrak{p}_0$, see Appendix A.



(a) Model $\mathfrak{r}_0(\omega)$ at 2 Hz frequency.

(b) Model $\mathfrak{r}_0(\omega)$ at 4 Hz frequency.

(c) Model $\mathfrak{r}_0(\omega)$ at 7 Hz frequency.

(d) Model $\mathfrak{r}_0$ using frequencies from 0.5 to 15 Hz with 0.5 Hz increment.

Figure 4. Reflectivity models $\mathfrak{r}_0$ obtained from the MBTT representation defined by (39). The model $\mathfrak{s}$ is defined from (38) as the difference between the data obtained from the Marmousi model Figure 3(b) and the smooth background Figure 3(a). The figures correspond with squared slowness and the values are given in $(\mathrm{m\,s}^{-1})^{-2}$.

a single, fixed ω:

$$\mathfrak{m}_0(\omega) = \mathfrak{p}_0 + \mathfrak{r}_0(\omega) = \mathfrak{p}_0 + \mathcal{W}(\omega) D\mathcal{F}^*_\omega(\mathfrak{p}_0)\mathfrak{s}_0(\omega). \tag{41}$$

It allows us to study the behaviour of both approaches (FWI and MBTT) with individual frequency.

4.2. Choice of perturbation directions

We define now the unit norm directions $\mathfrak{u}$ to be used for the determination of the directional attraction basins introduced in Section 2.

Background perturbation The direction for the background perturbation, $\mathfrak{u}$, is selected as the one-dimensional ramp of Figure 5. This perturbation is either applied onto the global model m (FWI), or onto the background unknown $\mathfrak{p}$ (MBTT). We first illustrate the effect of the background perturbation onto the forward map in Figure 6. It shows the unperturbed and perturbed synthetic data for the center source at frequency 4 Hz. It corresponds to the solution of the Helmholtz equation (27) recorded at the receivers

location. Note that, from (30) the direct arrivals are removed from the forward operator. One sees on this figure that, when the perturbation in the direction $\mathfrak{u}$ is applied to the global model $\mathfrak{m}$ (FWI), both phase and amplitude of the signal are modified. On the contrary, when it is applied to the background part $\mathfrak{p}$ (MBTT), the phase of the original signal is preserved, and only the amplitudes are modified.

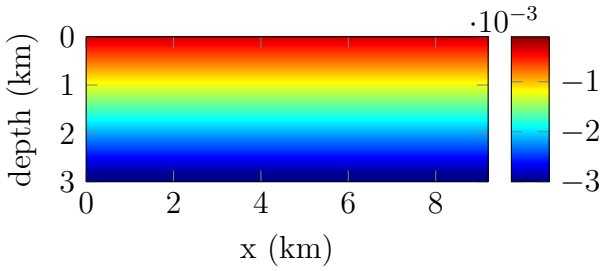


Figure 5. Perturbation $\mathfrak{u}$ used for the background model $\mathfrak{p}$. The amplitude is determined such that $\|\mathfrak{u}\| = 1$ and the values are given in $(\mathrm{m\,s^{-1}})^{-2} = \mathrm{s^2\,m^{-2}}$.

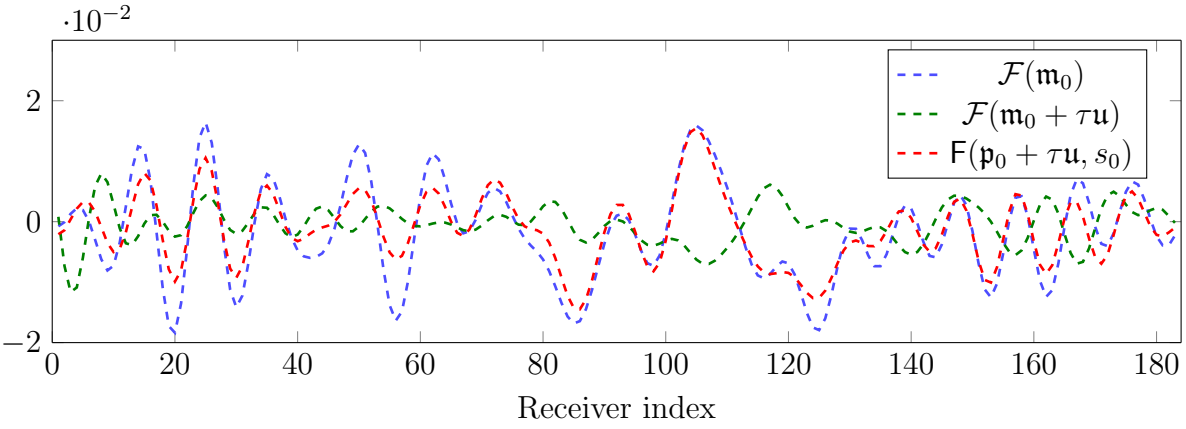


Figure 6. Comparison of the synthetic data associated with the center shot at 4 Hz using a model perturbed by the direction $\mathfrak{u}$ of Figure 5 applied onto the global model m or on the part $\mathfrak{p}$ using the MBTT model decomposition. The step for the perturbation is $\tau = 5 \times 10^{-5}$.

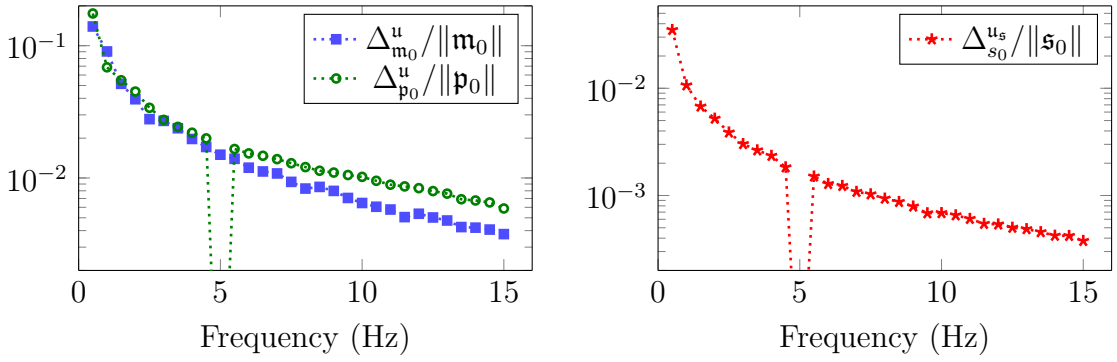
Reflectivity perturbation The FWI objective function is known to be nearly quadratical with respect to reflectivity, i.e. to the high spatial frequency part of m, and the same property holds by construction for the dependance of the MBTT objective function with respect to the data space reflectivity $\mathfrak{s}$. Hence one expects large basins of attraction with respect to $\mathfrak{s}$ in the MBTT formulation. There is no clear strategy to select the direction $\mathfrak{u}_\mathfrak{s}$ for $\mathfrak{s}$, hence, we choose for a random vector of the data space.

4.3. Comparison of local Θ-estimates

The formulas for the local estimate of the size $\Delta^{\mathfrak{u}}_{\mathfrak{m}_0}$ of the Θ-attraction basin have already been derived in Section 2.4 for the classical FWI. Application of the same formulas (24) (25) to $\mathsf{F}(\mathfrak{p},\mathfrak{s})$ instead of $\mathcal{F}(m)$ gives immediately for the MBTT formulation the local estimates of the sizes $\Delta^{\mathfrak{u}}_{\mathfrak{p}_0}$ and $\Delta^{\mathfrak{u}_\mathfrak{s}}_{\mathfrak{s}_0}$ of the Θ-attraction basins with respect to $\mathfrak{p}$ and $\mathfrak{s}$ in directions $\mathfrak{u}$ and $\mathfrak{u}_\mathfrak{s}$ at $\mathfrak{p}_0, \mathfrak{s}_0$. Figure 7 shows the evolution of the local Θ-estimates with frequency. We observe that:

- the size of the corresponding attraction basins decreases with frequency, when the perturbation is applied on m, $\mathfrak{p}$ and $\mathfrak{s}$. It is the expected behaviour as one knows that high frequencies are more prone to local minima (with the decrease of the wavelength).
- Figure 7(a) shows a slightly larger attraction basin in the direction of the background perturbation $\mathfrak{u}$ when it is applied to the propagator part $\mathfrak{p}_0$ of the MBTT parameterization rather than when applied directly to $\mathfrak{m}_0$. But it is not up to the expectations raised by the claim that the MBTT parameterization allows to overcome the phase shift problem [14, 15].
- Regarding $\mathfrak{s}$, the estimated size appears surprisingly small compared to the large attraction basin expected.

Yet, one has to remember that these are *local* Θ-estimate, which can be very pessimistic, as explained in Section 2.4, and we postpone more definitive comments to the end of Subsection 4.4, where exact Θ-estimates are calculated.



(a) Perturbation of the background model $\mathfrak{p}$.

(b) Perturbation of the reflectivity $\mathfrak{s}$.

Figure 7. Evolution with frequency of the local Θ-estimates of the size of the attraction basins given by (24), in the context of FWI and MBTT. Here $\mathfrak{p}_0$ is the smooth velocity background of Figure 3(a), the direction $\mathfrak{u}$ for $\mathfrak{p}$ is given Figure 5, and the direction $\mathfrak{u}_\mathfrak{s}$ for $\mathfrak{s}$ is a random vector. In the MBTT representation, the reflectivity uses only the selected frequency.

4.4. Comparison of exact Θ- and R_G-estimates

We apply the method described in Section 2.5 for the case of classical FWI, which translates immediately to the case of MBTT by replacing the FWI forward map $m \rightsquigarrow \mathcal{F}(m)$ by the MBTT forward map $\mathfrak{p}, \mathfrak{s} \rightsquigarrow \mathsf{F}(\mathfrak{p}, \mathfrak{s})$. This leads to the computation of *deflection* and *global radius of curvature maps* between the following points:

$$\begin{array}{lllll} \text{FWI} & \text{(attraction basin for } \mathfrak{m}) & : & \mathcal{F}(\mathfrak{m}_0 + t\mathfrak{u}) \quad \text{and} \quad \mathcal{F}(\mathfrak{m}_0 + t'\mathfrak{u}); \\ \text{MBTT} & \text{(attraction basin for } \mathfrak{p}) & : & \mathsf{F}(\mathfrak{p}_0 + t\mathfrak{u}, \mathfrak{s}_0) \quad \text{and} \quad \mathsf{F}(\mathfrak{p}_0 + t'\mathfrak{u}, \mathfrak{s}_0); \\ \text{MBTT} & \text{(attraction basin for } \mathfrak{s}) & : & \mathsf{F}(\mathfrak{p}_0, \mathfrak{s}_0 + t\mathfrak{u}_\mathfrak{s}) \quad \text{and} \quad \mathsf{F}(\mathfrak{p}_0, \mathfrak{s}_0 + t'\mathfrak{u}_\mathfrak{s}); \end{array}$$

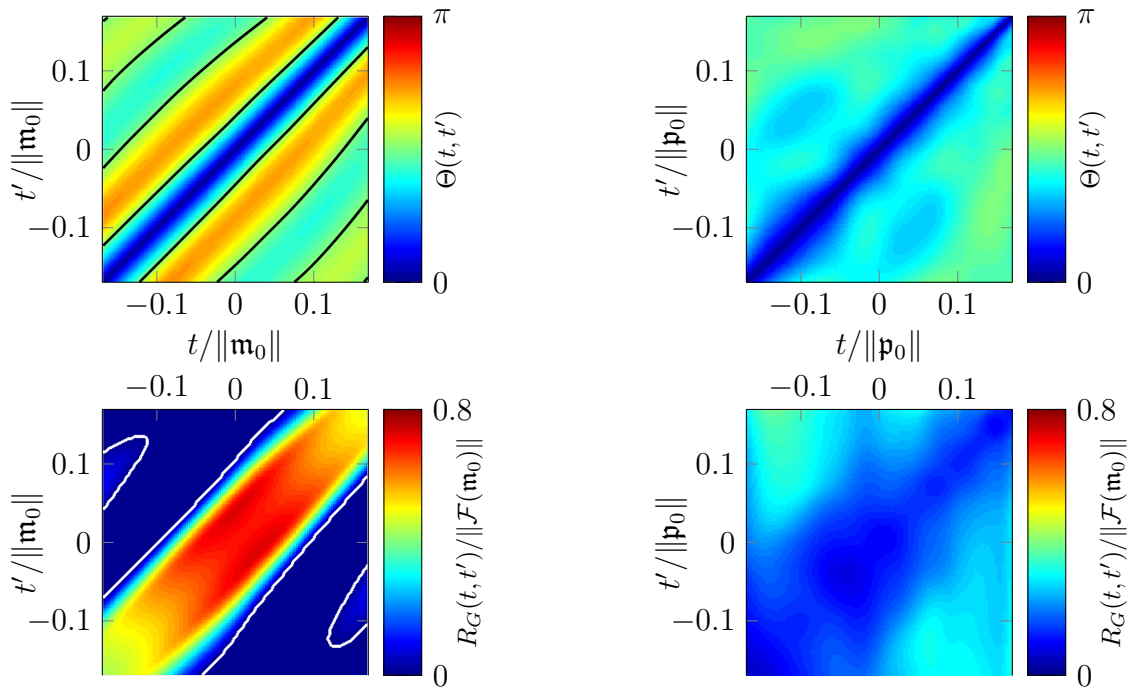
for all $-\Delta \leq t, t' \leq \Delta$.

We first compute the deflection and global radius of curvature maps for $\mathfrak{m}$ and $\mathfrak{p}$, using values of t and t' in an interval $[-\Delta, \Delta]$ which is chosen to represent in each case about $\pm 20\%$ of the norm of $\mathfrak{m}_0$ or $\mathfrak{p}_0$ defined in (39). Figures 8 and 9 show these maps at two selected frequencies: 4 and 7 Hz, and Table 1 summarizes the extracted exact estimates of $\Delta^{\mathfrak{u}}_{\mathfrak{m}_0}$, $\Delta^{\mathfrak{u}}_{\mathfrak{p}_0}$, together with the local estimates extracted from Figure 7(a).

Table 1. Size Δ of attraction basins centered at $\mathfrak{m}_0$ and corresponding maximal tolerable error R_G for the different estimations, at 4 and 7 Hz. By construction, the R_G-estimates correspond to the limit case of a zero tolerable error, the other values are extracted from Figures 7, 8, 9 and 10.

	model m		model $\mathfrak{p}$		model $\mathfrak{s}$	
	$\frac{\Delta^{\mathfrak{u}}_{\mathfrak{m}_0}}{\|\mathfrak{m}_0\|}$	$\frac{R^{\mathfrak{u}}_{G,\mathfrak{m}_0}}{\|\mathcal{F}(\mathfrak{m}_0)\|}$	$\frac{\Delta^{\mathfrak{u}}_{\mathfrak{p}_0}}{\|\mathfrak{p}_0\|}$	$\frac{R^{\mathfrak{u}}_{G,\mathfrak{p}_0}}{\|\mathcal{F}(\mathfrak{m}_0)\|}$	$\frac{\Delta^{\mathfrak{u}}_{\mathfrak{s}_0}}{\|\mathfrak{s}_0\|}$	$\frac{R^{\mathfrak{u}}_{G,\mathfrak{s}_0}}{\|\mathcal{F}(\mathfrak{m}_0)\|}$
4 Hz						
Local Θ-estimates	0.02	1.6	0.022	0.8	2×10^{-3}	6.5
Exact Θ-estimates	0.02	0.6	0.2	0.05	54	6.5
Exact R_G-estimates	0.05	0.0	0.23	0	60	0
7 Hz						
Local Θ-estimates	0.01	1.6	0.014	0.7	1×10^{-3}	4.9
Exact Θ-estimates	0.01	0.6	0.11	0.06	23	4.9
Exact R_G-estimates	0.025	0.0	0.20	0	>35	0

– The first observation is that lower values of deflection are achieved when the background perturbation $\mathfrak{u}$ is applied to $\mathfrak{p}$ (MBTT) rather than to $\mathfrak{m}$ (FWI): at 4 Hz, Figures 8(b), it never reaches $\pi/2$, and at 7 Hz, Figure 9(b), only a few portions attain this value. On the contrary, for FWI, Figures 8(a) and 9(a), the deflection rapidly reaches $\pi/2$ at both frequencies. This indicates that the MBTT formulation produces larger Θ-attraction basins than the standard FWI formulation, roughly



(a) FWI: perturbation of the global model $\mathfrak{m}$.

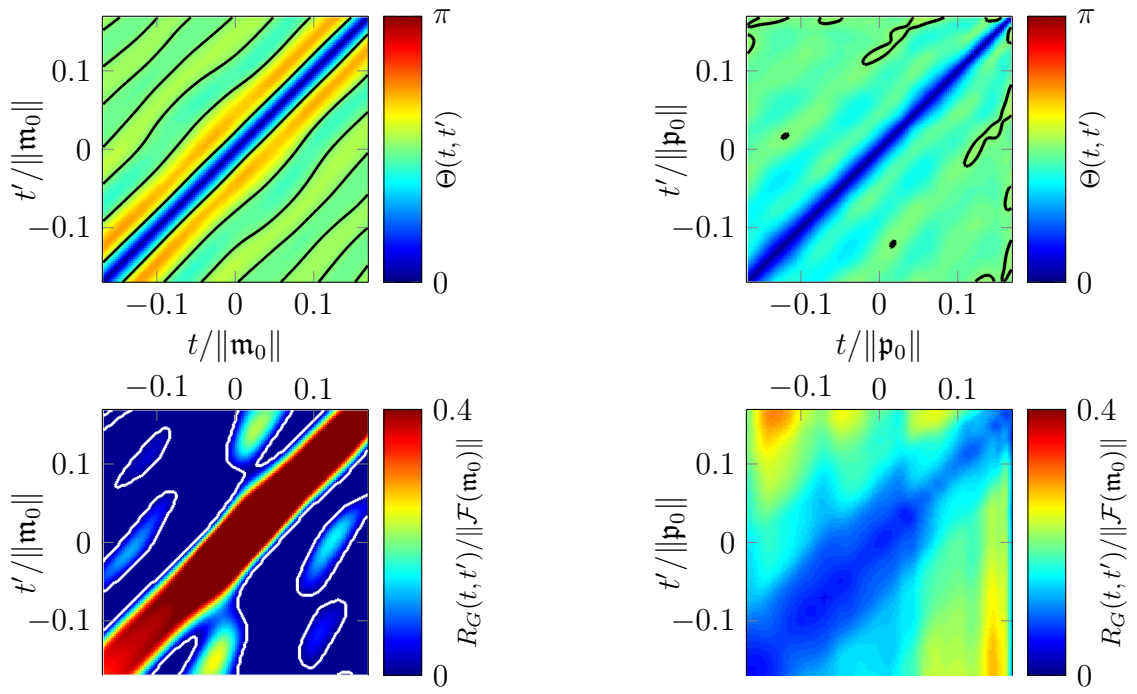
(b) MBTT: perturbation of the background $\mathfrak{p}$.

Figure 8. Maps of the deflection (12) (top) and global radius (8) (bottom) between two perturbed velocity or background models at frequency 4 Hz. The perturbation direction is the ramp of Figure 5, it is either applied to the global model $\mathfrak{m}$ (left) or to the background parameter $\mathfrak{p}$ (right). The black lines indicate when the deflection becomes higher than $\pi/2$, the white lines indicate when the global radius becomes 0.

by a factor ten (Table 1). Notice that the size of the Θ-attraction basin is divided by two for both FWI and MBTT when the frequency increases from 4 to 7 Hz.

– The second observation concerns the *strict positivity* of the global radius of curvature R_G (bottom of Figures 8 and 9), which determines the R_G-attraction basin characterized by a zero tolerable error (see Section 2.5). For the MBTT formulation, Figures 8(b) and 9(b), R_G remains strictly positive all over the map, which shows that the R_G-basin is larger than the investigated interval. Its size is of approximately 20% at both 4 and 7 Hz (Table 1). On the contrary, for the usual FWI formulation, Figures 8(a) and 9(a), R_G decreases very rapidly to zero when one moves away of the diagonal, producing smaller R_G-attraction basins, with size of 5% at 4 Hz and 2.5% at 7 Hz, smaller by a factor four to eight to the corresponding MBTT attraction basins.

– Concerning the *magnitude* of R_G, whose minimum over the attraction basin gives



(a) FWI: perturbation of the global model m.

(b) MBTT: perturbation of the background $\mathfrak{p}$.

Figure 9. Maps of the deflection (12) (top) and global radius (8) (bottom) between two perturbed velocity or background models at frequency 7 Hz. The perturbation direction is the ramp of Figure 5, it is either applied to the global model $\mathfrak{m}$ (left) or to the background parameter $\mathfrak{p}$ (right). The black lines indicate when the deflection becomes higher than $\pi/2$, the white lines indicate when the global radius becomes 0.

the tolerable error level, one sees that it takes *larger values* for FWI near the main diagonal (i.e. for small attraction basins) than for MBTT over the whole map (i.e. for larger attraction basins). It is confirmed by the values of R_G in Table 1 which gives the tolerable error level associated with the Θ-attraction basins (this level is zero by definition for the R_G attraction basins).

To summarize, MBTT extends significantly the size of attraction basins with respect to background perturbations, at the price of a reduction in the admissible error level. This explains the success of MBTT's alternate minimization algorithm, as reported in [18, 16]. We further illustrate in Appendix B.

We compare now the above exact Θ-estimates with the local Θ-estimates of Subsection 4.3.

– For the FWI approach, Figures 7(a), 8(a) and 9(a) and Table 1, it shows that both *local* and *exact* Θ-estimates of $\Delta^{\mathfrak{u}}_{\mathfrak{m}}$ are of the same size. In sight of the upper bound

estimate (14) on which the local Θ-estimate is based, one can think that the FWI formulation corresponds to the worst situation, where the image of a segment in the background space is a curve close to an arc of circle in the data space.

– The situation is completely different for the MBTT formulation: Figures 7(a), 8(b) and 9(b) and Table 1, we see that the exact Θ-estimate $\Delta^{\mathfrak{u}}_{\mathfrak{p}}$ is about ten times larger than its *local* Θ-estimate.

We can also determine the exact attraction basins in the MBTT formulation for the data space reflectivity $\mathfrak{s}$ at $\mathfrak{s}_0$ in the direction $\mathfrak{u}_{\mathfrak{s}}$, which we expect to be large because the forward map F is nearly linear with respect to $\mathfrak{s}$. Figure 10 shows the corresponding deflection and global radius of curvature maps for values of t and t' in an interval $[-\Delta, \Delta]$ which is chosen to represent in each case about ± 35 times the norm of $\mathfrak{s}_0$ defined in (39). As expected, the exact Θ-attraction basin is large (23 to 54 times the norm of $\mathfrak{s}_0$ depending on frequency), and is 10^5 times larger than its local estimate, which, together with the previous results on the estimation of $\Delta^{\mathfrak{u}}_{\mathfrak{p}}$, confirms the necessity of exact estimates for accuracy in MBTT.

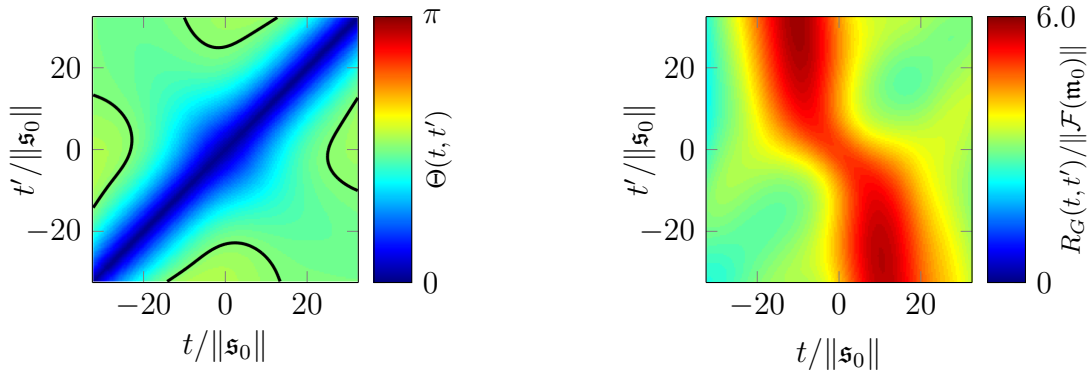


Figure 10. Maps of the deflection (12) (left) and global radius (8) (right) at frequency 7 Hz between two perturbed data-space reflectivities, for the perturbation direction $\mathfrak{u}_{\mathfrak{s}}$ chosen in Section 4.2. The black lines indicate when the deflection becomes higher than $\pi/2$, the white lines indicate when the global radius becomes 0.

4.5. Parameter tuning

Numerical experimentation in [5] with the smoothness of the background $\mathfrak{p}$ and the reflectivity level (40) have shown that the effectiveness of the MBTT reformulation of FWI decreases when the reflectivity level is too small (the energy backscattered by the reflectors becomes of the same order of magnitude as the "negligible" energy backscattered by the background), or too large (the energy of multiples, whose phase

is not controlled by MBTT, becomes comparable to that of the primary reflections). A priori computation of attraction basins allows to fine tune the parameters of the inverse problem to produce the largest attraction basin and hence to determine the precision required for the initial guess to ensure convergence of local algorithms to the global minimum.

5. Conclusion

We have presented theoretical and numerical tools for the a priori analysis of the optimizability of nonlinear least squares minimization problems by local algorithms. They consist in the definition of *attraction basins* around a nominal parameter and the associated *tolerable errors* such that *for any data below tolerable error*, one is sure that the data misfit has *a unique local - and hence global - minimum* over the basin. The computation of these quantities can be intensive, but it only depends on directional derivatives of the map to be inverted, and it provides *a priori* information on the model space size where there is no local minimum, without having to experiment with the misfit function for different data.

These optimizability tools have been applied to seismic inversion in the time-frequency domain, where the misfit function exhibits local minima in the directions associated with low spatial frequencies perturbation of the background velocity. Computation of directional background attraction basins for the FWI approach and its MBTT reformulation has confirmed and quantified the benefits associated to the reformulation, in terms of optimizability. This provides a strong incentive for the use of the MBTT decomposition in order to alleviate the low frequency requirement of FWI, despite its larger computational burden. It is the subject of our future work (implementation and analysis of the choice of tuning parameters).

More important, the computation of attraction basins has been shown to be a useful quantitative tool for tuning the parameters of the inverse problem, in order to ensure an *as-large-as-possible* attraction basin. It also tells how precise the initial parameter guess has to be for a local algorithm to converge to the global minimum. Note that the methodology is applicable for other least squares minimization problems (see [5]).

Acknowledgments

The authors would like to thank the anonymous referees that have provided valuable comments to improve the quality of the paper. The research of F. Faucher is supported by the Inria–TOTAL strategic action DIP.

Appendix A. Adjoint-state for complex variables, directional derivatives

The quantitative reconstruction method follows an iterative minimization of the cost function defined as the difference between simulation and observations. We follow the standard least squares formulation of (31), and consider the Helmholtz equation (27) to write

$$\mathcal{J}(m) = \frac{1}{2}\sum_{\omega}\sum_{g} \|\mathcal{F}_{\omega}^{g}(m) - d_{\omega}^{g}\|_{\mathcal{D}}^{2} = \frac{1}{2}\sum_{\omega}\sum_{g} \|\mathcal{R}p_{\omega}^{g} - d_{\omega}^{g}\|_{\mathcal{D}}^{2}, \tag{A.1}$$

where the forward problem is written with the restriction operator to receiver location $\mathcal{R}$, and we use the index g for the sources. For the minimization, one needs to obtain the gradient of the cost function, which is usually obtained using adjoint state method, see [29] for a review of the method in geophysical application. In this appendix, we specify the computations associated with complex-valued fields. For the sake of clarity, we omit the source and frequency sums, and consider

$$\mathcal{J}(m) = \frac{1}{2}\|\mathcal{F}(m) - \boldsymbol{d}\|_{\mathcal{D}}^{2} = \frac{1}{2}\|\mathcal{R}p - \boldsymbol{d}\|_{\mathcal{D}}^{2}. \tag{A.2}$$

In the frequency domain, the pressure field p is complex, which requires some precaution for the application of the adjoint state method. In particular, note that the functional

$$\mathsf{J}(p) = \frac{1}{2}\|\mathcal{R}p(m) - \boldsymbol{d}\|_{\mathcal{D}}^{2} = \frac{1}{2}(\mathcal{R}p(m) - \boldsymbol{d})\overline{(\mathcal{R}p(m) - \boldsymbol{d})} \tag{A.3}$$

is not analytic (holomorphic) with respect to the field p. A workaround is relatively standard, see for example [9, 24, 22], with elements of complex calculus based on Wirtinger calculus. We believe it is important to mention this aspect which is too often disregarded in seismic applications and hereby present the steps involved.

Appendix A.1. Complex derivation

The derivation of complex functional is conducted by taking independently the complex variable and its conjugate, respectively z and $\overline{z}$, for a complex parameter $z = x + \mathrm{i}y$, with $\mathrm{i}^2 = -1$.

Theorem 1. *[9, Theorem 1] Let $g : \mathbb{C}\times\mathbb{C} \to \mathbb{C}$ be a function of a complex number z and its conjugate $\overline{z}$ and let g be analytic with respect to each variable (z and $\overline{z}$) independently. Let $h : \mathbb{R}\times\mathbb{R} \to \mathbb{C}$ be the function of the real variables x and y such that $g(z,\overline{z}) = h(x,y)$ where $z = x + \mathrm{i}y$. Then the partial derivative $\partial_z g$ (treating $\overline{z}$ as a constant) gives the same result as $(\partial_x h - \mathrm{i}\partial_y h)/2$. Similarly, $\partial_{\overline{z}} g$ is equivalent to $(\partial_x h + \mathrm{i}\partial_y h)/2$.*

Corollary 1. *Following the statement of Theorem 1, we have*

$$\overline{\frac{\partial g}{\partial \overline{z}}} = \frac{\partial g}{\partial z}. \tag{A.4}$$

Proof. By direct application of Theorem 1,

$$\overline{\frac{\partial g}{\partial \overline{z}}} = \frac{1}{2}\Big(\overline{\frac{\partial f}{\partial x} + \mathrm{i}\frac{\partial f}{\partial y}}\Big) = \frac{1}{2}\Big(\frac{\partial f}{\partial x} - \mathrm{i}\frac{\partial f}{\partial y}\Big) = \frac{\partial g}{\partial z}. \tag{A.5}$$

□

We straightforwardly apply the theorem to the misfit function where we identify $p := z = x + \mathrm{i}y$.

$$\mathsf{J} \;:\; (x, y) \quad \to \quad \frac{1}{2}\|\mathcal{R}(x + \mathrm{i}y) - \boldsymbol{d}\|^2_{\mathcal{D}}, \tag{A.6}$$

where x, y and $\boldsymbol{d}$ can be assimilated with vectors in the discrete setting. Then by deriving independently with respect to x and y we obtain

$$\begin{cases} \dfrac{\partial \mathsf{J}}{\partial x} = \dfrac{1}{2}[\mathcal{R}^*(\mathcal{R}(p) - \boldsymbol{d}\,)]^T + \dfrac{1}{2}(\mathcal{R}(p) - \boldsymbol{d}\,)^*\mathcal{R}, \\ \dfrac{\partial \mathsf{J}}{\partial y} = -\dfrac{\mathrm{i}}{2}[\mathcal{R}^*(\mathcal{R}(p) - \boldsymbol{d}\,)]^T + \dfrac{\mathrm{i}}{2}(\mathcal{R}(p) - \boldsymbol{d}\,)^*\mathcal{R}. \end{cases} \tag{A.7}$$

We can further deduce the derivative of J with respect to p and $\overline{p}$, where they are considered independent such that $\mathsf{J} = \mathsf{J}(p, \overline{p})$, with Theorem 1,

$$\begin{cases} \dfrac{\partial \mathsf{J}}{\partial p} = \dfrac{1}{2}[\mathcal{R}^*(\mathcal{R}(p) - \boldsymbol{d}\,)]^T = \dfrac{1}{2}(\mathcal{R}(p) - \boldsymbol{d}\,)^T\mathcal{R}, \\ \dfrac{\partial \mathsf{J}}{\partial \overline{p}} = \dfrac{1}{2}(\mathcal{R}(p) - \boldsymbol{d}\,)^*\mathcal{R} \quad = \dfrac{1}{2}(\mathcal{R}(\overline{p}) - \overline{\boldsymbol{d}}\,)^T\mathcal{R}. \end{cases} \tag{A.8}$$

The following theorems give the framework of what can be seen as the chain rule for complex derivation.

Theorem 2. *Consider the complex-valued function f of a real parameter m and the real-valued functions g_1 and g_2 such that $f(m) = g_1(z(m), \overline{z}(m)) + \mathrm{i}g_2(z(m), \overline{z}(m))$. The derivative with respect to the real parameter m is defined by*

$$\frac{\partial f}{\partial m} = \frac{\partial g}{\partial z}\frac{\partial z}{\partial m} + \frac{\partial g}{\partial \overline{z}}\frac{\partial \overline{z}}{\partial m}. \tag{A.9}$$

Proof. From the definition of f we have

$$\begin{cases} \dfrac{\partial f}{\partial m} = \dfrac{\partial g_1(z(m),\overline{z}(m))}{\partial m} + \mathrm{i}\dfrac{\partial g_2(z(m),\overline{z}(m))}{\partial m} \\ \quad = \dfrac{\partial g_1}{\partial z}\dfrac{\partial z}{\partial m} + \dfrac{\partial g_1}{\partial \overline{z}}\dfrac{\partial \overline{z}}{\partial m} + \mathrm{i}\dfrac{\partial g_2}{\partial z}\dfrac{\partial z}{\partial m} + \mathrm{i}\dfrac{\partial g_2}{\partial \overline{z}}\dfrac{\partial \overline{z}}{\partial m} \\ \quad = \dfrac{\partial (g_1 + \mathrm{i}g_2)}{\partial z}\dfrac{\partial z}{\partial m} + \dfrac{\partial (g_1 + \mathrm{i}g_2)}{\partial \overline{z}}\dfrac{\partial \overline{z}}{\partial m} \\ \quad = \dfrac{\partial g}{\partial z}\dfrac{\partial z}{\partial m} + \dfrac{\partial g}{\partial \overline{z}}\dfrac{\partial \overline{z}}{\partial m} \end{cases} \tag{A.10}$$

□

Theorem 3. *Consider the real-valued functions f and g defined by $f(m) = g(z(m),\overline{z}(m))$,*

$$\frac{\partial f}{\partial m} = 2\mathrm{Re}\Big(\frac{\partial g}{\partial z}\frac{\partial z}{\partial m}\Big) = 2\mathrm{Re}\Big(\frac{\partial g}{\partial \overline{z}}\frac{\partial \overline{z}}{\partial m}\Big). \tag{A.11}$$

Proof. Direct application of Theorem 2 gives

$$\frac{\partial f}{\partial m} = \mathrm{Re}\Big(\frac{\partial g}{\partial z}\frac{\partial z}{\partial m} + \frac{\partial g}{\partial \overline{z}}\frac{\partial \overline{z}}{\partial m}\Big). \tag{A.12}$$

We use Theorem 1 and Corollary 1, and take $z(m) = x(m) + \mathrm{i}y(m)$ to have

$$\mathrm{Re}\Big(\frac{\partial g}{\partial \overline{z}}\frac{\partial \overline{z}}{\partial m}\Big) = \mathrm{Re}\Big(\overline{\frac{\partial g}{\partial z}}\,\overline{\frac{\partial \overline{z}}{\partial m}}\Big) = \mathrm{Re}\overline{\Big(\frac{\partial g}{\partial z}\overline{\frac{\partial \overline{z}}{\partial m}}\Big)} = \mathrm{Re}\Big(\frac{\partial g}{\partial z}\overline{\frac{\partial \overline{z}}{\partial m}}\Big), \tag{A.13}$$

where

$$\overline{\frac{\partial \overline{z}}{\partial m}} = \overline{\frac{\partial (x - \mathrm{i}y)}{\partial m}} = \overline{\Big(\frac{\partial x}{\partial m} - \mathrm{i}\frac{\partial y}{\partial m}\Big)} = \frac{\partial x}{\partial m} + \mathrm{i}\frac{\partial y}{\partial m} = \frac{\partial z}{\partial m}. \tag{A.14}$$

We inject in (A.12) to obtain

$$\frac{\partial f}{\partial m} = \mathrm{Re}\left(\frac{\partial g}{\partial z}\frac{\partial z}{\partial m}\right) + \mathrm{Re}\left(\frac{\partial g}{\partial z}\frac{\partial z}{\partial m}\right) = 2\mathrm{Re}\left(\frac{\partial g}{\partial z}\frac{\partial z}{\partial m}\right). \tag{A.15}$$

The alternative expression is obtained similarly but by replacing $\partial_z g$ in (A.12), instead of $\partial_{\overline{z}} g$. □

Application of Theorem 3 gives the gradient of the cost function with respect to m,

$$\begin{cases} \nabla_m \mathsf{J} = \dfrac{\partial}{\partial m}\Big(\mathsf{J}(m,p)\Big)^T = 2\mathrm{Re}\Big(\dfrac{\partial \mathsf{J}}{\partial p}\dfrac{\partial p}{\partial m}\Big)^T = \mathrm{Re}\Big((\mathcal{R}(p) - \boldsymbol{d}\,)^*\mathcal{R}\dfrac{\partial p}{\partial m}\Big)^T \\ \qquad = \mathrm{Re}\left(\Big(\dfrac{\partial p}{\partial m}\Big)^*\mathcal{R}^*(\mathcal{R}(p) - \boldsymbol{d}\,)\right), \end{cases} \tag{A.16}$$

where ^{T} stands for the transposed.

Appendix A.2. Adjoint state method

In order to avoid explicit computation of $\partial_m p$ in (A.16), the gradient is computed with the first order adjoint state method. It has been introduced in the work of [25], and implemented by [12] for the computation of a functional gradient. The formulation for the elastic wave problem has been carried out by [39, 40]. It is a relatively standard techniques nowadays, e.g. [21], see [29] for a review in geophysical situations. Yet, the complex variable specification is less common in seismic literature. In order to compute the derivative $\nabla\mathcal{J}$, we formulate the constrained minimization problem (omitting the space dependency)

$$\min_{m\in\mathcal{M}} \mathcal{J}(m) = \mathsf{J}(p) \qquad \text{subject to } \mathcal{A}(m)p = g, \tag{A.17}$$

where we introduce the wave operator $\mathcal{A}$, which corresponds to the Helmholtz equation defined in (27). Note that we consider a single source for now, for clarity, and shall later reintroduce the source summation, by linearity, cf. (A.25). The problem (A.17) is recast into a formulation with Lagrangian such that

$$\mathcal{L}(m,\widehat{p},\widehat{\gamma}) = \mathsf{J}(m,\widehat{p}) + \langle \mathcal{A}\widehat{p} - g, \widehat{\gamma}\rangle, \tag{A.18}$$

where $\langle\cdot,\cdot\rangle$ stands for the complex inner product in L^2 such that $\langle v, w\rangle = v^* w$, with v^* the adjoint. By taking p solution of $\mathcal{A}p = g$, we have that $\nabla_m\mathcal{L}(m,p,\widehat{\gamma}) = \nabla_m\mathcal{J}(m)$. Furthermore, by application of complex derivation Theorem 2, we have

$$\frac{\partial}{\partial m}\Big(\mathcal{L}(m,p,\widehat{\gamma})\Big) = \mathrm{Re}\Big(\frac{\partial\mathcal{L}}{\partial m} + \frac{\partial\mathcal{L}}{\partial\overline{p}}\frac{\partial\overline{p}}{\partial m} + \frac{\partial\mathcal{L}}{\partial p}\frac{\partial p}{\partial m}\Big), \tag{A.19}$$

and with Corollary 1,

$$\frac{\partial}{\partial m}\Big(\mathcal{L}(m,p,\widehat{\gamma})\Big) = \mathrm{Re}\Big(\frac{\partial\mathcal{L}}{\partial m} + \Big(\frac{\partial\mathcal{L}}{\partial\overline{p}} + \frac{\partial\mathcal{L}}{\partial p}\Big)\frac{\partial p}{\partial m}\Big). \tag{A.20}$$

The adjoint state γ is now selected such that

$$\mathrm{Re}\Big(\frac{\partial\mathcal{L}}{\partial\overline{p}} + \frac{\partial\mathcal{L}}{\partial p}\Big) = 0, \tag{A.21}$$

which gives,

$$\mathrm{Re}\Big(\frac{\partial\mathsf{J}}{\partial p} + \frac{\partial\mathsf{J}}{\partial\overline{p}} + \mathcal{A}^*\gamma\Big) = 0. \tag{A.22}$$

We now incorporate (A.8), and the adjoint state γ solves the problem

$$\mathcal{A}^*\gamma = -\mathcal{R}^*(\mathcal{R}(p) - \boldsymbol{d}). \tag{A.23}$$

Using this formulation for γ, the gradient reduces to

$$\nabla_m\mathcal{J} = \mathrm{Re}\Big(\langle\partial_m\mathcal{A}p, \gamma\rangle\Big)^T. \tag{A.24}$$

We can eventually reintroduce the sum over the sources, which, by linearity, gives

$$\nabla_m \mathcal{J} = \sum_g \mathrm{Re}\Big(\langle (\partial_m \mathcal{A}) p^g, \gamma^g \rangle\Big)^T, \quad \text{where } \gamma^g \text{ solves } \mathcal{A}^* \gamma^g = -\mathcal{R}^*(\mathcal{R}(p^g) - d^g). \quad \text{(A.25)}$$

Using the adjoint-state approach, the gradient is derived from the resolution of additional (adjoint) forward problem, using the residuals for sources.

Appendix A.3. Directional derivative computation

For the computation of the directional Fréchet derivative, we consider the path

$$P(t) = \mathcal{F}(m_0 + t\mathfrak{u}), \quad \text{(A.26)}$$

associated with the pressure field p solution to the Helmholtz equation

$$(-\omega^2(m_0 + t\mathfrak{u}) - \Delta)p = g, \quad \text{(A.27)}$$

according to (27), where we omit the space dependency and boundary conditions. Deriving (A.27) with respect to t gives

$$(-\omega^2(m_0 + t\mathfrak{u}) - \Delta)\partial_t p = \omega^2 \mathfrak{u} p, \quad \text{(A.28)}$$

and we have

$$V(t) = \mathcal{R}(\partial_t p) = D\mathcal{F}(m_0)(\mathfrak{u}). \quad \text{(A.29)}$$

It is straightforward to reproduce the operation for the second order derivative:

$$(-\omega^2(m_0 + t\mathfrak{u}) - \Delta)\partial_t^2 p = 2\omega^2 \mathfrak{u} \partial_t p, \quad \text{(A.30)}$$

and we obtain,

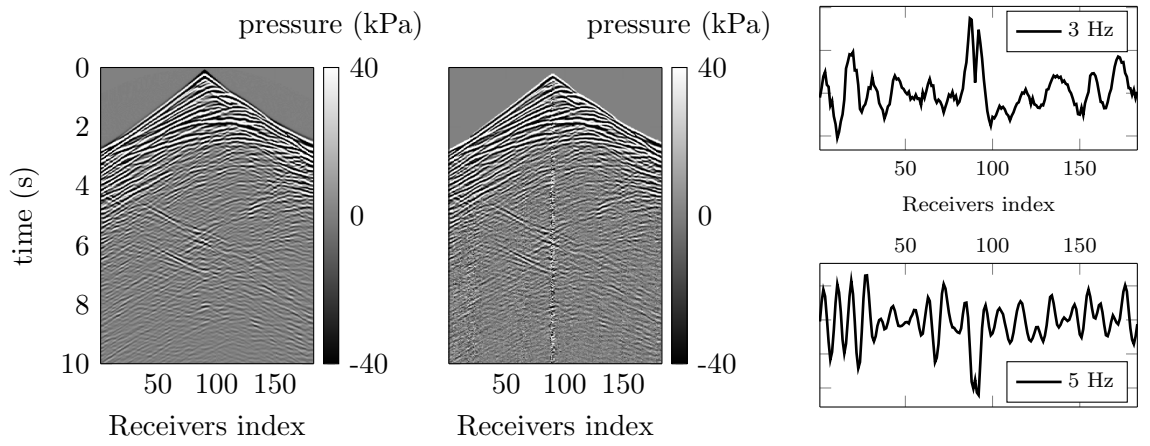
$$A(t) = \mathcal{R}(\partial_t^2 p) = D^2\mathcal{F}(m_0)(\mathfrak{u}, \mathfrak{u}). \quad \text{(A.31)}$$

Therefore, the directional derivative only required the resolution of Helmholtz equation with appropriate right-hand side. The technique can also be found in the context of elastic-fluid interaction in [2], where the derivation is conducted with respect to the Lamé parameters.

Appendix B. Influence of background wavespeed in FWI

In Section 4, the a priori estimates have shown that the MBTT-reformulation of FWI provides an increase of the size of the attraction basins, in particular with respect to the background velocity. In this appendix, we carry out numerical experiments of reconstruction to highlight the importance of this background velocity for the iterative reconstruction algorithm and how it impacts on the reconstructed models.

We use the FWI method for the identification of the wavespeed c in (27), and target the Marmousi model of Figure 3(b). We consider a seismic configuration, where the data consist in *time-domain* measurements of the pressure field (p in (27)). We take 91 sources equally distributed along the horizontal axis and located at a fixed depth of 10 m (i.e. near the surface, cf. Figure 2). We consider 183 receivers to acquire the data: they are positioned at a depth of 100 m. In order to mimic a realistic acquisition, the data $\boldsymbol{d}$ input to the time-harmonic FWI problem (31) is obtained by generating time-domain seismic traces, then adding noise with a signal-to-noise ratio of 15 dB, and finally applying a discrete Fourier transform. These steps are illustrated in Figure B1 for a source located in $x = 4500$ m.



(a) Noiseless time-domain trace.

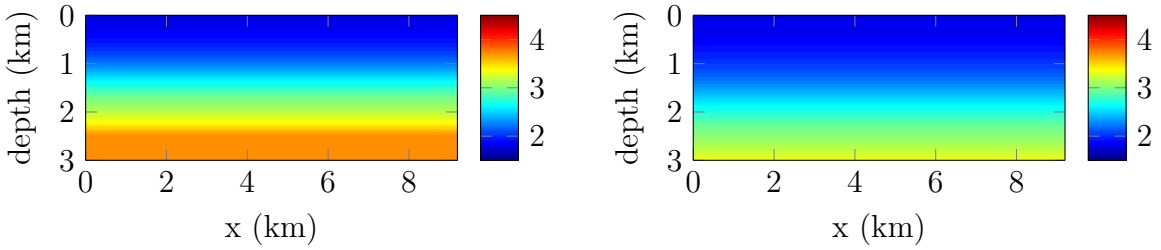
(b) Time-domain trace with 15 dB signal-to-noise ratio.

(c) Real part of the Fourier transform of the noisy trace.

Figure B1. The time-domain data used for the reconstruction of the Marmousi wavespeed encode noise and we compute a discrete Fourier transform at frequencies from 2 to 10 Hz for the reconstruction.

For the reconstruction, we follow a sequential frequency progression from 2 to 10 Hz, with a 1 Hz step (lower frequencies are not available because of the noise). We compare two choices of starting models, which are pictured in Figure B2. Both correspond to one-dimensional variations of the wavespeed (which only changes with the depth). They have similar values on the first 200 m in depth but below, they have different magnitude for the profile slope.

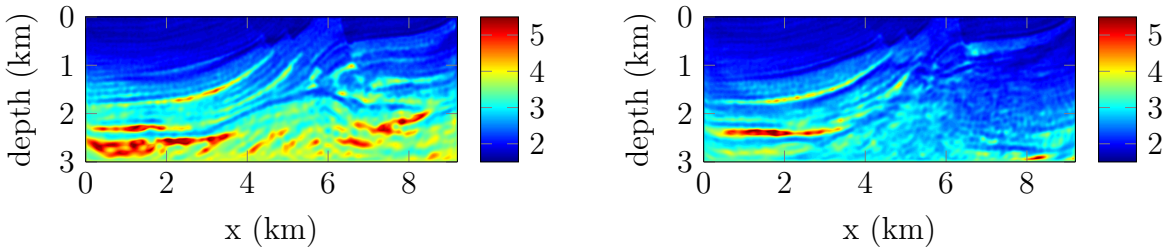
We perform the iterative reconstruction with FWI for these two initial guesses, i.e. we proceed with (31). The gradient is computed with the adjoint-state method (see Appendix A) and we use the nonlinear conjugate gradient method for the search direction, cf. [27]. We perform 20 iterations per frequency, for a total of 180 iterations. The final reconstructions (i.e. after 10 Hz frequency) are shown in Figure B3.



(a) Starting model A.

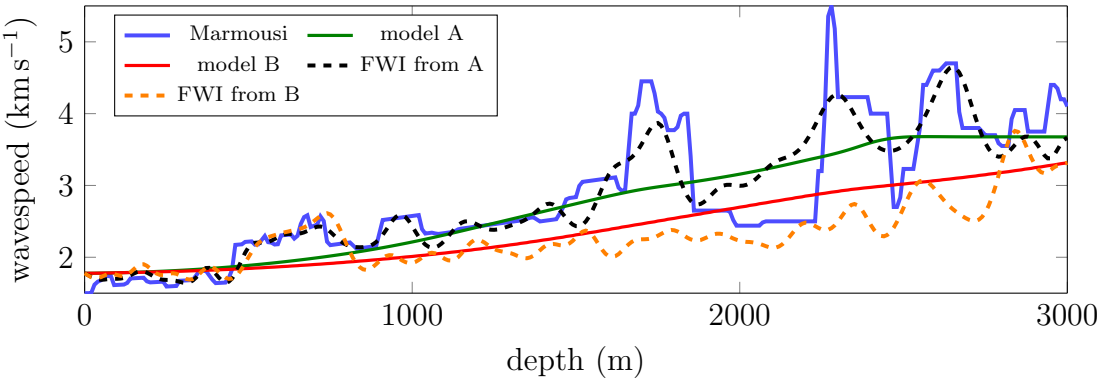
(b) Starting model B.

Figure B2. The starting wavespeeds for the reconstruction of the Marmousi model of Figure 3(b) consist in one-dimensional variation (with the depth only). On the left, the profile varies from 1.5 to 3.7 km s^{-1} while on the right from 1.5 to 3.3 km s^{-1}.



(a) Reconstruction starting from model A.

(b) Reconstruction starting from model B.



(c) Vertical profile at $x = 7$ km.

Figure B3. Wave speed reconstruction and vertical profile in $x = 7$ km of the Marmousi model Figure 3(b) using the initial guesses of Figure B2 with data of frequency between 2 and 10 Hz.

While the two initial models are quite close (see the vertical section Figure 3(c)), the final reconstructions are totally different.

– The reconstruction using starting model A (Figure 3(a)) is accurate and encodes the

appropriate velocity values and structures. Only the deepest parts are less accurate due to limited illumination.

- However, the reconstruction using starting model B (Figure 3(b)) only gives a low-valued wavespeed, where none of the actual structures appear.
- It is confirmed in the one-dimension section in $x = 7$ km of Figure 3(c), where we see that the reconstruction from initial model A follows the Marmousi structures, but the reconstruction from initial model B fails after about 1 km depth. For the latter, the reconstruction is actually sometimes worse (i.e. lower values) than its starting model B.

This experiment confirms the importance of the velocity background for the reconstruction algorithm and clearly, its absence of knowledge leads to the failure of the procedure. With FWI, this can only be overcome by accessing lower (unrealistically low) frequency content, see [10]. It is anticipated that the MBTT algorithm would not suffer from this issues, as it increases the attraction basins, cf. Section 4.

In a similar approach as what we did for Figure 6, we evaluate the misfit functional for a background variation applied either to the full model (i.e. FWI) or to $\mathfrak{p}$ in the MBTT-formulation, see Figure B4. It corresponds to 7 Hz frequency with the direction $\mathfrak{u}$ of Figure 5.

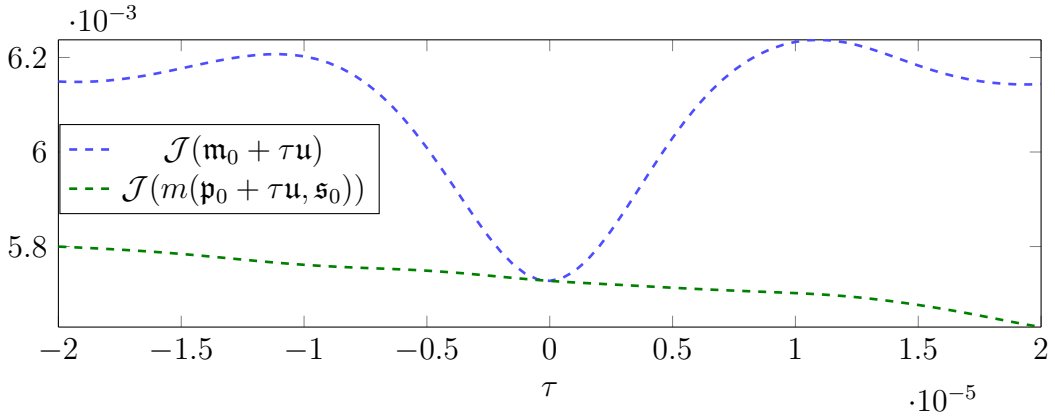


Figure B4. Comparison of the misfit functional associated with the Marmousi data (see Figure B1) at 7 Hz with a perturbation of the background wavespeed. The perturbation is either applied to the global model (FWI approach) or restricted to the background $\mathfrak{p}$ in the MBTT formulation.

As expected, we observe that the MBTT increases the size of the attraction basin with respect to background perturbation. Indeed, local minima appear on the right and left sides of Figure B4 for FWI (global model perturbation) while the cost function is monotone for MBTT (perturbation of $\mathfrak{p}$); see also the comparisons of [5]. Therefore,

it confirms that the MBTT would be able to overcome the lack of low frequency, as observed in the reconstruction experiments of [18, 16].

Our next step is now the implementation of the full MBTT framework for reconstruction, but the method remains a complicated task numerically speaking (e.g., choice of basis to ensure the smoothness of the background $\mathfrak{p}$) and it is part of ongoing investigations (cf. the conclusion section). The a priori estimates we have given Section 4 already provides a quantitative measure of the expected gain and advices for implementation (Subsection 4.5).

Appendix C. Details on the MBTT model decomposition

In this appendix, we provide additional details on the MBTT model decomposition, and the computational framework. In particular, we avoid the explicit computation of $D\mathcal{F}$ to obtain the reflectivity in (33), using the adjoint-state method. Then, we give the directional derivative computations.

Appendix C.1. Computation of reflectivity

The reflectivity part of the MBTT model representation is given by, cf. (33),

$$\mathfrak{r} = \mathcal{W} D\mathcal{F}_0^* \mathfrak{s}. \tag{C.1}$$

The adjoint state method of Appendix A allows to compute $\mathfrak{r}$ without explicitly forming $D\mathcal{F}_0$. Indeed, the adjoint state method provides, by identification from (A.16) and (A.25),

$$\begin{cases} \displaystyle\sum_g \Big(\langle (\partial_m \mathcal{A}) p^g, \gamma^g \rangle \Big)^T &= \displaystyle\sum_g \Big(\frac{\partial p^g}{\partial m} \Big)^* \mathcal{R}^* (\mathcal{R}(p^g) - d^g) \\ \Rightarrow \displaystyle\sum_g \Big((\partial_m \mathcal{A}) p^g)^* \gamma^g \Big)^T &= D\mathcal{F}^g(m)^* (\mathcal{R}(p^g) - d^g), \end{cases} \tag{C.2}$$

where $\mathcal{F}^g$ stands for the forward operator associated with source g. The fields p^g and γ^g solve respectively the forward and ajoint problems, see (27) and (A.25). Proceeding by analogy with the reasoning of Appendix A, it is straightforward to see that $\mathfrak{r}$ can be express as

$$\mathfrak{r} = \mathcal{W} \sum_g \Big(((\partial_{\mathfrak{p}} \mathcal{A}_0) p_0^g)^* \gamma_0^g \Big)^T, \tag{C.3}$$

where $\mathcal{A}_0$ is the Helmholtz operator with zero reflectivity (i.e., using $m = \mathfrak{p}$),

$$\mathcal{A}_0 := (\, -\omega^2 \mathfrak{p} - \Delta); \tag{C.4}$$

p_0^g solves the forward problem with $\mathcal{A}_0$ for the source g, and γ_0 solves for the source g,

$$\mathcal{A}_0^* \gamma_0^g = -\mathcal{R}^* \mathfrak{s}^g, \tag{C.5}$$

where $\mathfrak{s}$ (in the data space) writes as $\mathfrak{s} = \{\mathfrak{s}^{(1)} \dots \mathfrak{s}^{(n_{src})}\}$. The model representation (33) becomes

$$m(\mathfrak{p}, \mathfrak{s}) = \mathfrak{p} + \mathcal{W} \sum_g \Big(((\partial_{\mathfrak{p}} \mathcal{A}_0) p_0^g)^* \gamma_0^g \Big)^T. \tag{C.6}$$

Therefore, the model is expressed after the computation of direct and adjoint fields p_0 and γ_0 using the background $\mathfrak{p}$ only.

Appendix C.2. Directional derivative computation

For the estimation of the size of the basin of attraction and of the radius of curvature, we need the directional derivative of the forward operator. The method is given in Appendix A.3 and only necessitates the resolution of the forward problem, with an additional step for the MBTT decomposition. For clarity, we focus on the parameter $\mathfrak{p}$, the chain rule gives (where $(\mathfrak{u}_{\mathfrak{p}})$ indicates the directional derivative),

$$\frac{\partial \mathsf{F}}{\partial \mathfrak{p}}(\mathfrak{p}, \mathfrak{s})(\mathfrak{u}_{\mathfrak{p}}) = \frac{\partial \mathsf{F}}{\partial m} \frac{\partial m}{\partial \mathfrak{p}}(\mathfrak{u}_{\mathfrak{p}}). \tag{C.7}$$

We derive from (C.6),

$$\Big(\frac{\partial m}{\partial \mathfrak{p}}(\mathfrak{p}, \mathfrak{s})(\mathfrak{u}_{\mathfrak{p}}) \Big)^T = \mathfrak{u}_{\mathfrak{p}} + \mathcal{W} \sum_g ((\partial_{\mathfrak{p}} \mathcal{A}_0) \partial_{\mathfrak{p}} p_0^g + (\partial^2_{\mathfrak{p}^2} \mathcal{A}_0) p_0^g)^* (\mathfrak{u}_{\mathfrak{p}}) \gamma_0^g + ((\partial_{\mathfrak{p}} \mathcal{A}_0) p_0^g)^* \partial_{\mathfrak{p}} \gamma_0^g (\mathfrak{u}_{\mathfrak{p}}). \tag{C.8}$$

The workflow is as follows

(i) compute the directional derivative $(\partial_{\mathfrak{p}} p_0)(\mathfrak{u}_{\mathfrak{p}})$ and $(\partial_{\mathfrak{p}} \gamma_0)(\mathfrak{u}_{\mathfrak{p}})$ with the same method as presented in Appendix A.3 (thus, each requires the resolution of the wave equation with specific right-hand side).

(ii) Formulate $(\partial_{\mathfrak{p}} m)(\mathfrak{u}_{\mathfrak{p}})$ from (C.8).

(iii) Compute the directional derivative $\frac{\partial \mathsf{F}}{\partial m}(\mathfrak{u}_{\widehat{m}})$, where $\mathfrak{u}_{\widehat{m}} = (\partial_{\mathfrak{p}} m)(\mathfrak{u}_{\mathfrak{p}})$, using the same methodology as prescribed in Appendix A.3.

One can proceed similarly for $\mathfrak{s}$, adapting the chain rule. Regarding the second order derivatives, it is analogous with one degree more of derivation in the chain rule. Computationally speaking, it simply requires the resolution of additional forward problems.

References

[1] G. Alessandrini and S. Vessella, *Lipschitz stability for the inverse conductivity problem*, Adv. in Appl. Math., 35 (2005), pp. 207–241.

[2] I. Azpiroz, H. Barucq, R. Djellouli, and H. Pham, *Characterization of partial derivatives with respect to material parameters in a fluid–solid interaction problem*, Journal of Mathematical Analysis and Applications, 465 (2018), pp. 903–927.

[3] A. Bamberger, G. Chavent, and P. Lailly, *Une application de la théorie du contrôle à un problème inverse de sismique*, Annales de Géophysique, 33 (1977), pp. 183–200.

[4] ——, *About the stability of the inverse problem in the 1-d wave equation*, Journal of Applied Mathematics and Optimisation, 5 (1979), pp. 1–47.

[5] H. Barucq, H. Calandra, G. Chavent, and F. Faucher, *A priori estimates of attraction basins for velocity model reconstruction by time-harmonic Full Waveform Inversion and Data Space Reflectivity formulation*, Research Report RR-9253, Magique 3D ; Inria Bordeaux Sud-Ouest ; Université de Pau et des Pays de l'Adour, Feb. 2019.

[6] J.-P. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, Journal of Computational Physics, 114 (1994), pp. 185 – 200.

[7] E. Beretta, M. V. de Hoop, F. Faucher, and O. Scherzer, *Inverse boundary value problem for the helmholtz equation: quantitative conditional lipschitz stability estimates*, SIAM Journal on Mathematical Analysis, 48 (2016), pp. 3962–3983.

[8] E. Bozdağ, J. Trampert, and J. Tromp, *Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements*, Geophysical Journal International, 185 (2011), pp. 845–870.

[9] D. Brandwood, *A complex gradient operator and its application in adaptive array theory*, in IEE Proceedings F-Communications, Radar and Signal Processing, vol. 130, IET, 1983, pp. 11–16.

[10] C. Bunks, F. M. Saleck, S. Zaleski, and G. Chavent, *Multiscale seismic waveform inversion*, Geophysics, 60 (1995), pp. 1457–1473.

[11] H. Cartan, *Differential calculus*, vol. 1, Hermann, 1971.

[12] G. Chavent, *Identification of functional parameters in partial differential equations*, in Identification of Parameters in Distributed Systems, R. E. Goodson and M. Polis, eds., ASME, New York, 1974, pp. 31–48.

[13] G. Chavent, *Nonlinear least squares for inverse problems: theoretical foundations and step-by-step guide for applications*, Springer Science & Business Media, 2010.

[14] G. Chavent, *Data Space Reflectivity and the Migration based Travel Time approach to FWI*, in 79th EAGE Conference and Exhibition 2017-Workshops, 2017.

[15] G. Chavent and F. Clément, *Waveform inversion through MBTT formulation*, Inria, 1992.

[16] G. Chavent, K. Gadylshin, and V. Tcheverda, *Reflection fwi in mbtt formulation*, in 77th EAGE Conference and Exhibition 2015, 2015.

[17] G. Chavent and K. Kunisch, *On weakly nonlinear inverse problems*, SIAM Journal on Applied Mathematics, 56 (1996), pp. 542–572.

[18] F. Clément, G. Chavent, and S. Gómez, *Migration-based traveltime waveform inversion of 2-d simple structures: A synthetic example*, Geophysics, 66 (2001), pp. 845–860.

[19] M. V. de Hoop, L. Qiu, and O. Scherzer, *A convergence analysis of a multi-level projected steepest descent iteration for nonlinear inverse problems in banach spaces subject to stability constraints*, arXiv preprint arXiv:1206.3706, (2012).

[20] C. B. James Martin, Lucas C. Wilcox and O. Ghattas, *A stochastic newton mcmc method for large-scale statistical inverse problems with application to seismic inversion*, SIAM Journal

on Scientific Computing, 34 (2012), pp. A1460–A1487.

[21] M. Kern, *Numerical Methods for Inverse Problems*, John Wiley & Sons, 2016.

[22] K. Kreutz-Delgado, *The complex gradient operator and the cr-calculus*, arXiv preprint arXiv:0906.4835, (2009).

[23] P. Lailly, *The seismic inverse problem as a sequence of before stack migrations*, in Conference on Inverse Scattering: Theory and Application, J. B. Bednar, ed., Society for Industrial and Applied Mathematics, 1983, pp. 206–220.

[24] H. Li and T. Adali, *Optimization in the complex domain for nonlinear adaptive filtering*, in Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on, IEEE, 2006, pp. 263–267.

[25] J. L. Lions and S. K. Mitter, *Optimal control of systems governed by partial differential equations*, vol. 1200, Springer Berlin, 1971.

[26] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, *Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion*, Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society, 205 (2016), pp. 345–377.

[27] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Sries in Operations Research, 2 ed., 2006.

[28] R. Plessix, G. Chavent, and Y. De Roeck, *A quantitative kirchhoff migration to esimate the 2d velocity distribution*, in 3rd Internat. Conf. on Mathematical and Numerical Aspects of Wave Propagation, 1995, pp. 704–712.

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

[30] R. G. Pratt and N. R. Goulty, *Combining wave-equation imaging with traveltime tomography to form high-resolution images from crosshole data*, Geophysics, 56 (1991), pp. 208–224.

[31] R. G. Pratt, Z.-M. Song, P. Williamson, and M. Warner, *Two-dimensional velocity models from wide-angle seismic data by wavefield inversion*, Geophysical Journal International, 124 (1996), pp. 323–340.

[32] R. G. Pratt and M. H. Worthington, *Inverse theory applied to multi-source cross-hole tomography.*, Geophysical Prospecting, 38 (1990), pp. 287–310.

[33] L. Qiu, J. Ramos-Martínez, A. Valenciano, Y. Yang, and B. Engquist, *Full-waveform inversion with an exponentially encoded optimal-transport norm*, in SEG Technical Program Expanded Abstracts 2017, Society of Exploration Geophysicists, 2017, pp. 1286–1290.

[34] C. Shin and D.-J. Min, *Waveform inversion using a logarithmic wavefield*, Geophysics, 71 (2006), pp. R31–R42.

[35] C. Shin, S. Pyun, and J. B. Bednar, *Comparison of waveform inversion, part 1: conventional wavefield vs logarithmic wavefield*, Geophysical Prospecting, 55 (2007), pp. 449–464.

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

[37] W. Symes and J. J. Carazzone, *Velocity inversion by differential semblance optimization*, Geophysics, 56 (1991), pp. 654–663.

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

[39] A. Tarantola, *Inversion of travel times and seismic waveforms*, in Seismic tomography, Springer, 1987, pp. 135–157.

[40] A. Tarantola, *Theoretical background for the inversion of seismic waveforms including elasticity and attenuation*, Pure and Applied Geophysics, 128 (1988), pp. 365–399.

[41] E. Turkel and A. Yefet, *Absorbing pml boundary layers for wave-like equations*, Applied Numerical Mathematics, 27 (1998), pp. 533–557.

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

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

[44] S. Wang, M. V. de Hoop, and J. Xia, *On 3d modeling of seismic wave propagation via a structured parallel multifrontal direct helmholtz solver*, Geophysical Prospecting, 59 (2011), pp. 857–873.

[45] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt, *Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion*, Geophysics, 83 (2018), pp. R43–R62.