Geophys. J. Int. (2006) **167,** 495–503 doi: 10.1111/j.1365-246X.2006.02978.x

A review of the adjoint-state method for computing the gradient of a functional with geophysical applications

R.-E. Plessix

Shell International E&P, PO Box 60, 2280 *AB Rijswijk, The Netherlands. E-mail: reneedouard.plessix@shell.com*

Accepted 2006 February 28. Received 2006 February 28; in original form 2005 July 4

SUMMARY

Estimating the model parameters from measured data generally consists of minimizing an error functional. A classic technique to solve a minimization problem is to successively determine the minimum of a series of linearized problems. This formulation requires the Fréchet derivatives (the Jacobian matrix), which can be expensive to compute. If the minimization is viewed as a non-linear optimization problem, only the gradient of the error functional is needed. This gradient can be computed without the Fréchet derivatives. In the 1970s, the adjoint-state method was developed to efficiently compute the gradient. It is now a well-known method in the numerical community for computing the gradient of a functional with respect to the model parameters when this functional depends on those model parameters through state variables, which are solutions of the forward problem. However, this method is less well understood in the geophysical community. The goal of this paper is to review the adjoint-state method. The idea is to define some adjoint-state variables that are solutions of a linear system. The adjoint-state variables are independent of the model parameter perturbations and in a way gather the perturbations with respect to the state variables. The adjoint-state method is efficient because only one extra linear system needs to be solved.

Several applications are presented. When applied to the computation of the derivatives of the ray trajectories, the link with the propagator of the perturbed ray equation is established.

Key words: adjoint state, gradient, migration, tomography.

1 INTRODUCTION

One of the important tasks in data processing consists of determining model parameters from observed data. These tasks can be formulated as inverse problems, namely as the minimization of a functional, for instance the least-squares misfit between synthetic and observed data. In geophysics, this includes tomography, migration/inversion and automatic velocity analysis. When a local (descent) optimization technique, such as the conjugate gradient method, is used, the gradient of the functional is required, Gauthier *et al.* (1986), Liu & Bleistein (2001) and Mora (1989). The efficiency of the method greatly depends on the accuracy and efficiency of the computation of this gradient. With physical problems, the functional depends on so-called state variables. These state variables are the variables computed from the state equations, namely the equations that define the problem, sometimes called forward equations. For example, in the tomography problem, the state equations can be the ray equations, and the state variables the spatial coordinates and the slowness vectors describing the ray trajectories. The definition of the state variables depends on the mathematical formulation of the physical problem. The state equations depend on the model parameters. For the tomography problem this can be the background velocity (or slowness). The functional depends on those model parameters mainly through the dependency on the state variables.

The gradient of the functional, which depends on a set of state variables solutions of the forward equations, can be obtained with (a set of) the Fréchet derivatives of the state variables. The Fréchet derivatives are the derivatives of the state variables with respect to the model parameter. For instance, for the tomography problem described earlier, these derivatives are the derivatives of the spatial coordinates and the slowness vectors with respect to the slowness background. This gives the so-called Jacobian or sensitivity matrix. The Jacobian matrix can be used to linearize the functional, and the minimization problem can be solved by successively solving linearized problems using linear optimization techniques. However, the computation of the Fréchet derivatives can be expensive.

If non-linear optimization techniques, such as the non-linear conjugate method, are used, only the gradient of the functional may be needed, Gill *et al.* (1981). In the 1970s, a method based on the adjoint state has been introduced in the theory of inverse problems by Chavent (1974) to efficiently compute the gradient of a functional without the Fréchet derivatives. This approach originated from control theory, Lions (1972). Several authors in geophysics have applied this method, for instance, Lailly (1983), Bécache (1992), Chavent & Jacewitz (1995), Plessix *et al.* (1999) and Shen *et al.* (2003).

The goal of this note is to review this method that is well known in the numerical community, to give a recipe for applying it based on an augmented functional also called associated Lagrangian, and to describe several examples demonstrating its practical use.

The adjoint-state method is a general method to compute the gradient of a functional that depends on a set of state variables, which are solutions of forward equations. The adjoint-state variables are the solutions of an adjoint linear system and can be seen as variables which gather a global measure of the perturbation of the problem with respect to the state variables. Numerically this approach is attractive because only one extra linear system needs to be solved and often the computation of the gradient with respect to the model parameters is equivalent to one or two evaluations of the forward modelling. This cost is often almost independent of the number of model parameters, which is not always the case when the Fréchet derivatives are computed. However the adjoint-state method does not provide the sensitivity of the solution to errors. For that, the Fréchet derivatives are needed or a Monte Carlo type of methods with a large number of forward computations is required.

The outline of the paper is the following. In a first section the adjoint-state variables are introduced from the perturbation theory. Then based on an augmented functional a recipe to systematically define the adjoint-state equations is described. In the three next sections, examples are given. The first one is the least-squares migration, Tarantola (1987). This is almost a school example. The functional is the least-squares misfit between the synthetics and the measured reflection seismic data. The adjoint states correspond to the backpropagated field and the gradient of this least-squares misfit is a migration operator, Lailly (1983) and Tarantola (1984). The second example is the computation of the gradient of the differential semblance optimization (DSO) functional, Symes & Carazzone (1991) and Shen *et al.* (2003). This example shows the power of the adjoint-state technique. The third example is the stereotomography, Billette & Lambaré (1998) and Lambaré *et al.* (2004). The link between the adjoint-state variables and the propagator of the differential equation defining the ray trajectory perturbations, Cervený (2001) and Farra & Madariaga (1987), is established.

2 METHOD

The goal of this section is to explain the adjoint-state method for computing the gradient of a functional, $J(m)$, when J depends on $u(m)$. J is defined with the functional, h, by

$$J(m) = h(u(m), m). \tag{1}$$

The state variables, u, satisfy the state equations defined with the mapping, F,

$$F(u(m), m) = 0. \tag{2}$$

F is also called the forward problem or forward equation. m is the model parameter and belongs to the model parameter space $\mathbf{M}$. $\mathbf{M}$ is a real space in this article. u belongs to the state variable space, $\mathbf{U}$. $\mathbf{U}$ is a real or complex space. A state variable, u, is a physical realization if $F(u, m) = 0$. F is a mapping from $\mathbf{U} \times \mathbf{M}$ to $\mathbf{U}$. In order to distinguish between a physical realization and any element of $\mathbf{U}$, the elements of $\mathbf{U}$ are denoted by $\tilde{u}$. h is a functional from $\mathbf{U} \times \mathbf{M}$ to $\mathbf{R}$, the real space, whereas J is a functional from $\mathbf{M}$ to $\mathbf{R}$. Synthetic data are generally a subset of the state variables. It is assumed that h, F, and J are at least continuously differentiable and $u(m)$ is uniquely defined and continuously differentiable.

A simple example is the linear case $F(\tilde{u}, m) = \tilde{u} - A\,m$ and $h(\tilde{u}, m) = \frac{1}{2}||\tilde{u} - d||^2$, with d the observed data. This corresponds to the simple least-squares misfit with a linear forward problem. The physical realization is defined by $u(m) = A\,m$, with A a linear operator, in the discrete case a rectangular matrix.

2.1 The adjoint-state method from the perturbation theory

A perturbation, δm, of the model parameter, m, induces a perturbation, δu, of the physical realization, u, and a perturbation δJ of the error functional, J. $u + \delta u$ should be a physical realization with the model parameter $m + \delta m$. Therefore, to the first order:

$$\begin{aligned} 0 &= F(u + \delta u, m + \delta m) \\ &= F(u, m) + \frac{\partial\, F(u, m)}{\partial\, \tilde{u}}\delta u + \frac{\partial\, F(u, m)}{\partial\, m}\delta m. \end{aligned} \tag{3}$$

Since $F(u, m) = 0$, the first order development gives:

$$\frac{\partial\, F(u, m)}{\partial\, \tilde{u}}\delta u = -\frac{\partial\, F(u, m)}{\partial\, m}\delta m. \tag{4}$$

For the linear case this gives $\delta u = A\,\delta m$.

The first order development of J gives:

$$\delta J = \left\langle \frac{\partial\, h(u, m)}{\partial\, \tilde{u}}, \delta u \right\rangle_{\mathbf{U}} + \frac{\partial\, h(u, m)}{\partial\, m}\,\delta m, \tag{5}$$

where $\langle , \rangle_U$ is the scalar product in $\mathbf{U}$.

For the simple least-squares misfit, $\delta J = \langle u - d, \delta u \rangle_{\mathbf{U}}$.

Assuming that for any model parameter m of $\mathbf{M}$ there exists a unique solution u of $\mathbf{U}$, $u + \delta u$ is the unique solution of $F(u + \delta u, m + \delta m) = 0$. Therefore, at the first order, δu is the unique solution of eq. (4) and can be written with the inverse of $\frac{\partial\, F(u,m)}{\partial\, \tilde{u}}$. This gives:

$$\begin{aligned} \delta J = {} & \frac{\partial\, h(u, m)}{\partial\, m}\,\delta m - \\ & \left\langle \frac{\partial\, h(u, m)}{\partial\, \tilde{u}}, \left(\frac{\partial\, F(u, m)}{\partial\, \tilde{u}}\right)^{-1} \frac{\partial\, F(u, m)}{\partial\, m}\delta m \right\rangle_{\mathbf{U}} \\ \delta J = {} & \frac{\partial\, h(u, m)}{\partial\, m}\,\delta m - \\ & \left\langle \left(\left(\frac{\partial\, F(u, m)}{\partial\, \tilde{u}}\right)^{-1}\right)^{*} \frac{\partial\, h(u, m)}{\partial\, \tilde{u}}, \frac{\partial\, F(u, m)}{\partial\, m}\delta m \right\rangle_{\mathbf{U}}. \end{aligned} \tag{6}$$

(* denotes the adjoint). For the linear example with the least-squares misfit, this gives $\delta J = \langle u - d, A\,\delta m \rangle$ since $\frac{\partial\, F(u,m)}{\partial\, \tilde{u}} = I$ in this case; I is the identity operator.

In the second line of eq. (6), the terms that do not depend on the perturbation δm have been gathered, the idea is to avoid the computation of the Fréchet derivatives, $\frac{\partial\, u}{\partial\, m}$, because this can be expensive. Let us now define λ by

$$\left(\frac{\partial\, F(u, m)}{\partial\, \tilde{u}}\right)^{*} \lambda = \frac{\partial\, h(u, m)}{\partial\, \tilde{u}}. \tag{7}$$

The perturbation δJ now reads:

$$\delta J = \left(-\left\langle \lambda, \frac{\partial\, F(u, m)}{\partial\, m} \right\rangle_{\mathbf{U}} + \frac{\partial\, h(u, m)}{\partial\, m} \right) \delta m. \tag{8}$$

For the linear example with the least-squares misfit this simply gives $\lambda = u - d$ and $\delta J = \langle u - d, A\,\delta m \rangle_U$.

λ belongs to the dual space of $\mathbf{U}$. It is called the adjoint-state variable and eq. (7) is the adjoint-state equation. This is a system of linear equations. The linear operator is the adjoint of the operator formed by the derivatives of the state equations (the mapping F) with respect to the state variables. The right-hand side consists of the derivatives of the functional, h, with respect to the state variables.

In a sense the adjoint states gather the information on the perturbations of the state variables, viewed as independent variables. The computation of $\frac{\partial J}{\partial m}$ with eqs (7) and (8) is called the adjoint-state method.

The gradient of J can be computed either via the Fréchet derivatives of u with eqs (4) and (5) or via the adjoint-state method with eqs (7) and (8). The important differences between the two approaches are related to eqs (4) and (7). Indeed, on one hand, eq. (7) is independent of δm and needs to be solved only once. On the other hand, the right-hand side of eq. (4) depends on δm and this equation needs to be solved for each perturbation to obtain $\frac{\partial u}{\partial m}$, namely M times if M is the number of elements in m. The computational time of the adjoint-state method is often almost independent of M, because the time to compute eq. (8) is often negligible compared with the time to solve eq. (7). This makes this approach very efficient. eq. (7) depends on the adjoint of $\frac{\partial F}{\partial \tilde{u}}$ evaluated at (u, m), this means that u should be completely known before solving it.

The adjoint-state equations can also be obtained with the use of an augmented functional, also called associated Lagrangian.

2.2 A recipe with the augmented functional

Let us define the augmented functional, $\mathcal{L}$, from $\mathbf{U} \times \mathbf{U}^* \times \mathbf{M}$ to $\mathbf{R}$ ($\mathbf{U}^*$ is the dual of $\mathbf{U}$) by:

$$\mathcal{L}(\tilde{u}, \tilde{\lambda}, m) = h(\tilde{u}, m) - \langle \tilde{\lambda}, F(\tilde{u}, m) \rangle_{\mathbf{U}}, \tag{9}$$

where $\tilde{\lambda}$ is any element of $\mathbf{U}^*$ and, therefore, does not depend on m, as $\tilde{u}$ is any element of $\mathbf{U}$.

u is a physical realization, therefore, $F(u, m) = 0$, and for any $\tilde{\lambda}$

$$\mathcal{L}(u, \tilde{\lambda}, m) = h(u, m) = J(m), \tag{10}$$

and since $\tilde{\lambda}$ is independent of m,

$$\frac{\partial \mathcal{L}(u, \tilde{\lambda}, m)}{\partial \tilde{u}} \frac{\partial u}{\partial m} + \frac{\partial \mathcal{L}(u, \tilde{\lambda}, m)}{\partial m} = \frac{\partial J}{\partial m}. \tag{11}$$

We can then choose λ in $\mathbf{U}^*$ such that:

$$\frac{\partial \mathcal{L}(u, \lambda, m)}{\partial \tilde{u}} = \frac{\partial h(u, m)}{\partial \tilde{u}} - \left(\frac{\partial F(u, m)}{\partial \tilde{u}} \right)^* \lambda = 0. \tag{12}$$

This equation is identical to eq. (7) and is the adjoint-state equation. With this choice we retrieve the result of eq. (8).

$$\begin{aligned} \frac{\partial J}{\partial m} &= \frac{\partial \mathcal{L}(u, \lambda, m)}{\partial m} \\ &= \frac{\partial h(u, m)}{\partial m} - \left\langle \lambda, \frac{\partial F(u, m)}{\partial m} \right\rangle_{\mathbf{U}}. \end{aligned} \tag{13}$$

$\mathcal{L}$ can be also viewed as the Lagrangian associated with the minimization problem: find the minimum u of $h(\tilde{u}, m)$ under the constraint $F(u, m) = 0$. The theory of optimization with equality constraints, Ciarlet (1989), tells us that u is the minimum, if (u, λ) is a saddle point of $\mathcal{L}$. λ are called the Lagrange multipliers. At the saddle point the derivatives of $\mathcal{L}$ are equal to 0. The derivatives of $\mathcal{L}$ with respect to $\tilde{u}$ and $\tilde{\lambda}$ are:

$$\begin{cases} \frac{\partial \mathcal{L}(\tilde{u}, \tilde{\lambda}, m)}{\partial \tilde{\lambda}} = -F(\tilde{u}, m); \\ \frac{\partial \mathcal{L}(\tilde{u}, \tilde{\lambda}, m)}{\partial \tilde{u}} = \frac{\partial h(\tilde{u}, m)}{\partial \tilde{u}} - \left(\frac{\partial F(\tilde{u}, m)}{\partial \tilde{u}} \right)^* \tilde{\lambda}. \end{cases} \tag{14}$$

Therefore, $\frac{\partial \mathcal{L}(u, \lambda, m)}{\partial \tilde{\lambda}} = 0$ gives the state equations and $\frac{\partial \mathcal{L}(u, \lambda, m)}{\partial \tilde{u}} = 0$ gives the adjoint-state equations. And $\frac{\partial \mathcal{L}(u, \lambda, m)}{\partial m} = \frac{\partial J}{\partial m}$ as seen previously. Notice again that when deriving $\mathcal{L}$ with respect to m, $\tilde{u}$ and $\tilde{\lambda}$ are independent of m.

This link with the optimization theory is not needed to apply the adjoint-state method. For those familiar with this theory, it helps to recall the method. As in the optimization theory with equality constraints where one scalar Lagrange multiplier is associated with each scalar equation defining the constraints, one scalar adjoint state is associated with each scalar equation defining the mapping F in the augmented functional.

The computation of the gradient with the adjoint states can be summarized in the following recipe when u has been found from $F(u, m) = 0$:

(i) Build the augmented functional (associated Lagrangian) $\mathcal{L}$. $\mathcal{L}$, a functional of independent variables $\tilde{u}$, $\tilde{\lambda}$, and m is defined by

$$\mathcal{L}(\tilde{u}, \tilde{\lambda}, m) = h(\tilde{u}, m) - \langle \tilde{\lambda}, F(\tilde{u}, m) \rangle_{\mathbf{U}}. \tag{15}$$

If $F(\tilde{u}, m)$ is composed of N scalar equations, $F_i(\tilde{u}, m)$, $\tilde{\lambda}$ is a vector with N components, since at each scalar equation of F an adjoint state is associated, and $\mathcal{L}$ is defined by:

$$\mathcal{L}(\tilde{u}, \tilde{\lambda}, m) = h(\tilde{u}, m) - \sum_{i=1}^{N} \langle \tilde{\lambda}_i, F_i(\tilde{u}, m) \rangle. \tag{16}$$

For the linear case with the least-squares misfit $\mathcal{L}(\tilde{u}, \tilde{\lambda}, m) = \frac{1}{2} ||\tilde{u} - d||^2 - \langle \tilde{\lambda}, \tilde{u} - A\, m \rangle_{\mathbf{U}}$.

(ii) Define the adjoint-state equations. The adjoint-state equations are simply defined by $\frac{\partial \mathcal{L}(u, \lambda, m)}{\partial \tilde{u}} = 0$, where the derivatives are evaluated at the point (u, λ). This gives

$$\left(\frac{\partial F(u, m)}{\partial \tilde{u}} \right)^* \lambda = \frac{\partial h(u, m)}{\partial \tilde{u}}, \tag{17}$$

or

$$\frac{\partial h(u, m)}{\partial \tilde{u}_j} - \sum_{i=1}^{N} \left(\frac{\partial F_i(u, m)}{\partial \tilde{u}_j} \right)^* \lambda_i = 0. \tag{18}$$

The solution of this system determines the adjoint state, λ. For the linear case with the least-squares misfit $\lambda = u - d$.

(iii) Computation of the gradient of J. The gradient of J consists of the derivatives of $\mathcal{L}$ with respect to m:

$$\frac{\partial J}{\partial m} = \frac{\partial h(u, m)}{\partial m} - \left\langle \lambda, \frac{\partial F(u, m)}{\partial m} \right\rangle_{\mathbf{U}}, \tag{19}$$

or

$$\frac{\partial J}{\partial m} = \frac{\partial h(u, m)}{\partial m} - \sum_{i=1}^{N} \left\langle \lambda_i, \frac{\partial F_i(u, m)}{\partial m} \right\rangle. \tag{20}$$

To compute the derivative of the augmented functional, we recall that $\tilde{u}$ and $\tilde{\lambda}$ are independent of m.

For the linear case with the least-squares misfit $\frac{\partial J}{\partial m} = \langle \lambda, \frac{\partial A m}{\partial m} \rangle_{\mathbf{U}} = A^* \lambda$.

If u has complex values, since $J(m)$ is a real, the real part should be taken in the right-hand side term of eqs (19) and (20).

The linear example with the least-squares misfit is a trivial example. In the next sections more complicated examples are described.

3 LEAST-SQUARES MIGRATION

In this section, we formulate the migration as an inverse problem. The problem consists of minimizing with respect to the square of the slowness, the least-squares misfit between the synthetics, obtained by solving the wave equation, and the recorded (observed) reflection seismic data. The minimization of J should give the exact slowness. Unfortunately, in practice, J has many local minima, and a gradient

optimization will only provide the best perturbation of the initial model inside a certain basin of attraction. This basin is generally not the basin of the global minimum, Gauthier *et al.* (1986). This is the reason why this problem is called the least-squares migration problem in this paper.

This example is a good example to understand how the adjoint-state method can be applied. It also allows us to redemonstrate that the gradient of J is a migration. This was discovered in the 1980s, Lailly (1983) and Tarantola (1984).

We will develop the computation for multiple sources and multiple receivers, first in the frequency domain because it is simple, then in the time domain.

3.1 Frequency domain

In frequency domain, the wave equation operator reads: $L = -\omega^2 \sigma^2 - \Delta$, with σ the slowness. Note that the dependency on the spatial coordinates, $\mathbf{x}$, is not written. The finite-difference discretization of $Lu_s = f_s$ with given boundary conditions leads to a complex linear system Marfurt (1984):

$$\mathbf{A}(\omega, \mathbf{m})\mathbf{u}_s(\omega, \mathbf{m}) = \mathbf{f}_s(\omega). \tag{21}$$

$\mathbf{A}$ is a complex matrix of size n by n, where n is the total number of discretization points of the earth model. $\mathbf{f}_s$, a complex vector of n elements, represents the source function at the source point s. $\mathbf{u}_s$, a complex vector of n elements, corresponds to the pressure field due to the shot at s. The model parameter, $\mathbf{m}$, is a vector of M elements, and represents the values of the squared slowness at the discretization points.

The least-squares functional is

$$J(\mathbf{m}) = \frac{1}{2} \sum_{\omega} \sum_{s,r} ||\mathbf{S}_{s,r}\mathbf{u}_s(\omega, \mathbf{m}) - \mathbf{d}_{s,r}(\omega)||^2. \tag{22}$$

$\mathbf{d}_{s,r}$ are the data recorded at the receiver position r due to the source f_s. $\mathbf{S}_{s,r}$ is the restriction matrix onto the receiver r of the shot s.

The augmented functional reads, with $\tilde{\mathbf{u}} = (\tilde{\mathbf{u}}_s(\omega))$ and $\tilde{\lambda} = (\tilde{\lambda}_s(\omega))$ (the dependence on $\mathbf{x}$ is not written, but $\tilde{\mathbf{u}}_s$ and $\tilde{\lambda}_s$ depend on the space variables $\mathbf{x}$):

$$\mathcal{L}(\tilde{\mathbf{u}}, \tilde{\lambda}, \mathbf{m}) = Re\left[\frac{1}{2} \sum_{\omega} \sum_{s,r} ||\mathbf{S}_{s,r}\tilde{\mathbf{u}}_s(\omega) - d_{s,r}(\omega)||^2 - \sum_{\omega} \sum_{s} \langle \tilde{\lambda}_s(\omega), \mathbf{A}(\omega, \mathbf{m})\tilde{\mathbf{u}}_s(\omega) - \mathbf{f}_s(\omega)\rangle_{\mathbf{x}}\right], \tag{23}$$

where $\langle , \rangle_{\mathbf{x}}$ is the scalar product in $\mathbf{C}^n$. As $\tilde{\mathbf{u}}_s(\omega)$, $\tilde{\lambda}_s(\omega)$ are complex vectors of n elements, since the forward system, eq. (21), contains n scalar equations.

The derivative of $\mathcal{L}$ with respect to $\tilde{\mathbf{u}}_s$ evaluated at $(\mathbf{u}, \tilde{\lambda}, \mathbf{m})$ gives:

$$\frac{\partial\, \mathcal{L}(\mathbf{u}, \tilde{\lambda}, \mathbf{m})}{\partial\, \tilde{\mathbf{u}}_s(\omega)} = \sum_{r} \mathbf{S}^*_{s,r}(\mathbf{S}_{s,r}\mathbf{u}_s(\omega) - d_{s,r}(\omega)) - \mathbf{A}^*(\omega, \mathbf{m})\tilde{\lambda}_s(\omega). \tag{24}$$

The adjoint state is defined by $\frac{\partial\, \mathcal{L}(\mathbf{u}, \lambda, \mathbf{m})}{\partial\, \tilde{\mathbf{u}}_s} = 0$:

$$\mathbf{A}^*(\omega, \mathbf{m})\lambda_s(\omega) = \sum_{r} \mathbf{S}^*_{s,r}(\mathbf{S}_{s,r}\mathbf{u}_s(\omega) - d_{s,r}(\omega)). \tag{25}$$

There is one adjoint system per shot and per angular frequency.

The matrix $\mathbf{A}$ propagates the shot into the earth and $\mathbf{u}_s$ is the incident field originating at s. The adjoint of $\mathbf{A}$ propagates backward its source term, Lailly (1983). The source term, the right-hand side of eq. (25), is the sum over the receivers of the shot s of the residual between the synthetics and data. λ_s is then the backpropagation of the residual field.

The gradient of J is:

$$\frac{\partial\, J}{\partial\, \mathbf{m}} = -Re\left(\sum_{\omega} \sum_{s} \left\langle \lambda_s, \frac{\partial\, \mathbf{A}}{\partial\, \mathbf{m}}\mathbf{u}_s \right\rangle_{\mathbf{x}}\right). \tag{26}$$

The gradient of J is a vector of M elements. If we impose that $\mathbf{m}$ is discretized on the same grid as $\mathbf{u}_s$, $M = n$. Outside the boundary points $\frac{\partial\, \mathbf{A}}{\partial\, \mathbf{m}}$ is equal to $-\omega^2$. At the discretization point $\mathbf{x}$, we obtain

$$\frac{\partial\, J}{\partial\, \mathbf{m}}(\mathbf{x}) = Re\left(\sum_{\omega} \sum_{s} \omega^2\, \lambda^*_s(\mathbf{x}, \omega)\, \mathbf{u}_s(\mathbf{x}, \omega)\right). \tag{27}$$

Up to a multiplication factor, the gradient is similar to a migrated image and the formula is kinematically similar to the imaging principle, Clearbout (1985). A demonstration of this result without the adjoint-state method can be found in Plessix & Mulder (2004).

3.2 Time domain

We here develop the same approach but in time domain. The application of the adjoint-state method is slightly more complicated because of the initial boundary conditions.

The wave operator is $L = \sigma^2 \frac{\partial\, ^2}{\partial\, t^2} - \Delta$. With the initial boundary conditions, the pressure field u_s due to the source f_s satisfies:

$$\begin{cases} u_s(0) = 0; \\ \frac{\partial\, u_s(0)}{\partial\, t} = 0; \\ L\, u_s = f_s. \end{cases} \tag{28}$$

u_s and f_s depend on the time and on the spatial coordinates.

The least-squares functional reads

$$J(m) = \frac{1}{2} \sum_{s,r} \int_0^T (S_{s,r}u_s(t) - d_{s,r}(t))^2\, dt. \tag{29}$$

T is the recording time. $S_{s,r}$ is the restriction operator onto the receiver position, it depends on the spatial coordinates. The model parameter is the squared slowness, $m = \sigma^2$.

In the time domain u_s is real. For simplicity, we don't mention the spatial boundary conditions.

We associate the adjoint states $\tilde{\mu}^0_s$ and $\tilde{\mu}^1_s$ with the initial boundary conditions, and $\tilde{\lambda}_s$ with the wave equation. The augmented functional is defined by:

$$\begin{aligned} &\mathcal{L}((\tilde{u}_s), (\tilde{\lambda}_s), (\tilde{\mu}^0_s), (\tilde{\mu}^1_s), m) = \\ &\frac{1}{2} \sum_{s,r} \int_0^T (S_{s,r}u_s(t) - d_{s,r}(t))^2\, dt \\ &- \sum_{s} \int_0^T \left\langle \tilde{\lambda}_s(t), m\frac{\partial\, ^2\tilde{u}_s(t)}{\partial\, t^2} - \Delta\tilde{u}_s(t) - f_s(t) \right\rangle_{\mathbf{x}} dt \\ &- \sum_{s} \langle \tilde{\mu}^0_s, \tilde{u}_s(0)\rangle_{\mathbf{x}} - \sum_{s} \left\langle \tilde{\mu}^1_s, \frac{\partial\, \tilde{u}_s(0)}{\partial\, t} \right\rangle_{\mathbf{x}}, \end{aligned} \tag{30}$$

with $\langle \tilde{\lambda}_s, \tilde{u}_s\rangle_{\mathbf{x}} = \int_{\mathbf{X}} \tilde{\lambda}_s(\mathbf{x})\, \tilde{u}_s(\mathbf{x})\, d\mathbf{x}$ the real scalar product in the coordinate space.

After two integrations by part:

$$\int_0^T \left\langle \tilde{\lambda}_s, m \frac{\partial^2 \tilde{u}_s}{\partial t^2} \right\rangle_{\mathbf{x}} dt = \int_0^T \left\langle m \frac{\partial^2 \tilde{\lambda}_s}{\partial t^2}, \tilde{u}_s \right\rangle_{\mathbf{x}} dt + \left\langle \tilde{\lambda}_s(T), m \frac{\partial \tilde{u}_s(T)}{\partial t} \right\rangle_{\mathbf{x}} - \left\langle \tilde{\lambda}_s(0), m \frac{\partial \tilde{u}_s(0)}{\partial t} \right\rangle_{\mathbf{x}} - \left\langle m \frac{\partial \tilde{\lambda}_s(T)}{\partial t}, \tilde{u}_s(T) \right\rangle_{\mathbf{x}} + \left\langle m \frac{\partial \tilde{\lambda}_s(0)}{\partial t}, \tilde{u}_s(0) \right\rangle_{\mathbf{x}}. \quad (31)$$

With eqs (30) and (31) we can now compute the derivatives with respect to $\tilde{u}_s$ and evaluate them at (u, λ) to obtain the adjoint-state equations:

$$\begin{cases} \lambda_s(T) = 0; \\ \frac{\partial \lambda_s(T)}{\partial t} = 0; \\ m \frac{\partial^2 \lambda_s}{\partial t^2} - \Delta \lambda_s = \sum_r S^T_{s,r}(S_{s,r} u_s - d_{s,r}); \\ \mu^0_s = m \frac{\partial \lambda_s(0)}{\partial t}; \\ \mu^1_s = m \lambda_s(0). \end{cases} \quad (32)$$

(^{T} denotes the transpose.)

The gradient of J at the point $\mathbf{x}$ is

$$\frac{\partial J}{\partial m}(\mathbf{x}) = -\sum_s \int_0^T \lambda_s(\mathbf{x}, t) \frac{\partial^2 u_s(\mathbf{x}, t)}{\partial t^2} dt. \quad (33)$$

The adjoint states, μ^0_s and μ^1_s do not play a role in the gradient of J. We can ignore them.

The system (32) has final boundary conditions. To solve it the computation is done backwards from T to 0. To give a physical sense to the adjoint state and to interpret the integral, eq. (33), a new adjoint state, q_s, is defined by a change of variables in the time axis:

$$q_s(t) = \lambda_s(T - t). \quad (34)$$

The new adjoint-state system reads

$$\begin{cases} q_s(0) = 0; \\ \frac{\partial q_s(0)}{\partial t} = 0; \\ m \frac{\partial^2 q_s(t)}{\partial t^2} - \Delta q(t) = \\ \sum_r S^T_{s,r}(S_{s,r} u_s(T - t) - d_{s,r}(T - t)). \end{cases} \quad (35)$$

q_s satisfies the same wave equation that u_s but with a different source term. $S_{s,r} u_s(T - t) - d_{s,r}(T - t)$ is the residual, the difference between the synthetics and the recorded data, in reverse time. eq. (35) propagates the residual into the earth starting from the final time. q_s is called the backpropagated field of the residual. The gradient of J now reads:

$$\frac{\partial J}{\partial m}(\mathbf{x}) = -\sum_s \int_0^T q_s(\mathbf{x}, T - t) \frac{\partial^2 u_s(\mathbf{x}, t)}{\partial t^2} dt. \quad (36)$$

This result has been demonstrated by Lailly (1983).

4 SHOT-BASED DIFFERENTIAL SEMBLANCE OPTIMIZATION

As explained in the introduction of the previous section, the least-squares formulation is not satisfactory to retrieve the long wavelength components of the velocity model (background) from reflection seismic data, because the least-squares misfit as a function of the background has many local minima. To reformulate the problem and obtain a larger basin of attraction for the global minimum, an idea is to exploit the fact that in the reflection seismic data the earth is seen through different angles of incident. If the background velocity is correct, the pre-stack migration of the data should give the same images, Al Yahya (1989). If the pre-stack migration gives different earth structures, this means that the background slowness used in the migration is erroneous. Several mathematical formulations of this idea have been proposed in the last 20 years, among them Chavent & Jacewitz (1995), Clément *et al.* (2001), Plessix *et al.* (2000) and Symes & Carazzone (1991). In order to compute the gradient of the reformulated cost functions, the authors generally use the adjoint-state formulation because it is the most systematic method, without forgetting that they are mainly mathematicians.

As an example, I will describe the gradient computation of the DSO functional introduced in Symes & Carazzone (1991) for a common shot-based approach. The principle is to migrate each shot individually and then to differentiate the pre-stack migrated result with respect to the shot position for fixed points in the migrated images. If the derivative with respect to the shot position of the pre-stack migrated data is zero, it means that the migrated images are independent of the shot position, that is, of the angle of incident and that the background is correct. To obtain a global formulation, the DSO functional is used as a regularization of the least-squares functional.

Using a finite-difference scheme in frequency-domain, eq. (21), the incident wavefield, $\mathbf{u}_i$, due to the source function $\mathbf{f}_i$ located at the shot position i satisfies:

$$\mathbf{A}(\omega, \mathbf{m})\mathbf{u}_i(\omega, \mathbf{m}) = \mathbf{f}_i(\omega). \quad (37)$$

The dependency on the spatial coordinates is not explicitly written to simplify the notation. The synthetics at the angular frequency, ω, are $\mathbf{S}_{i,j}\,\mathbf{u}_i$, with $\mathbf{S}_{i,j}$ the restriction operator onto the receiver, j, of the shot, i.

To compute the migration, we introduce the backpropagated field, $\mathbf{v}_i$, defined by:

$$\mathbf{A}^*(\omega, \mathbf{m})\mathbf{v}_i(\omega, \mathbf{m}) = \sum_j \mathbf{d}_{i,j}(\omega), \quad (38)$$

with $\mathbf{d}_{i,j}$ the measured seismic data due to the shot i recorded at the receiver j.

The shot migrated image, $\mathbf{r}_i$, is then defined by:

$$\mathbf{r}_i = -\sum_\omega \omega^2 \mathbf{v}^*_i(\omega)\mathbf{u}_i(\omega). \quad (39)$$

Here we abuse the notation and eq. (39) means $\mathbf{r}_i(\mathbf{x}) = \sum_\omega \mathbf{v}^*_i(\mathbf{x})\mathbf{u}_i(\mathbf{x})$, where $\mathbf{x}$ is a discretization point.

The functional J reads

$$J(\mathbf{m}) = \frac{\alpha_1}{2} \sum_{i=1}^{n_s} \sum_j \sum_\omega ||\mathbf{S}_{i,j}\mathbf{u}_i(\omega) - \mathbf{d}_{i,j}||^2 + \frac{\alpha_2}{2} \sum_{i=1}^{n_s - 1} \| Re(\mathbf{r}_{i+1} - \mathbf{r}_i)\|^2. \quad (40)$$

n_s is the number of shots. α_1 and α_2 are the weights of the least-squares functional (the first term) and the DSO functional (the second term). The real part of $\mathbf{r}_{i+1} - \mathbf{r}_i$ is taken in the DSO functional because $\mathbf{r}_i$ is a complex number and only the real part corresponds to the migrated image.

We recall that $\mathbf{m}$ is the squared slowness at the discretization point and all the state variables are differentiable functions with respect

to $\mathbf{m}$. The goal is to compute the gradient of J with respect to the squared slowness.

The forward equations, eqs (37), (38) and (39), depend on the state variables $\mathbf{u}_i$, $\mathbf{v}_i$ and $\mathbf{r}_i$. $\mathbf{u}_i$, $\mathbf{v}_i$, $\mathbf{r}_i$ are discretized on the same grid, therefore, $\mathbf{u}_i, \mathbf{v}_i, \mathbf{r}_i$ belong to $\mathbf{C}^n$, with n the number of discretization points. To define the augmented functional, we associate, for each shot i, $\tilde{\lambda}_i^u$ with eq. (37), $\tilde{\lambda}_i^v$ with eq. (38) and $\tilde{\lambda}_i^r$ with eq. (39). $\tilde{\lambda}_i^u$, $\tilde{\lambda}_i^v$ and $\tilde{\lambda}_i^r$ belong to $\mathbf{C}^n$ because eqs (37) or (38) or (39) define n scalar equations. These quantities depend on ω and $\mathbf{x}$. The augmented functional reads with the state variables, $\tilde{\mathbf{u}} = (\tilde{\mathbf{u}}_i)$, $\tilde{\mathbf{v}} = (\tilde{\mathbf{v}}_i)$, $\tilde{\mathbf{r}} = (\tilde{\mathbf{r}}_i)$, and the adjoint-state variables, $\tilde{\lambda}^u = (\tilde{\lambda}_i^u)$, $\tilde{\lambda}^v = (\tilde{\lambda}_i^v)$ and $\tilde{\lambda}^r = (\tilde{\lambda}_i^v)$:

$$
\begin{aligned}
\mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}, \tilde{\mathbf{r}}, \tilde{\lambda}^u, \tilde{\lambda}^v, \tilde{\lambda}^r, \mathbf{m}) = Re\Bigg[& \\
& \frac{\alpha_1}{2} \sum_{i=1}^{n_s} \sum_{\omega} \sum_{j} ||\mathbf{S}_{i,j}\tilde{\mathbf{u}}_i(\omega) - \mathbf{d}_{i,j}(\omega)||^2 \; + \\
& \frac{\alpha_2}{2} \sum_{i=1}^{n_s-1} ||Re(\tilde{\mathbf{r}}_{i+1} - \tilde{\mathbf{r}}_i)||^2 - \\
& \sum_{i=1}^{n_s} \sum_{\omega} \left\langle \tilde{\lambda}_i^u(\omega), \mathbf{A}(\omega, \mathbf{m})\tilde{\mathbf{u}}_i(\omega) - \mathbf{f}_i(\omega) \right\rangle_{\mathbf{x}} - \\
& \sum_{i=1}^{n_s} \sum_{\omega} \left\langle \tilde{\lambda}_i^v(\omega), \mathbf{A}^*(\omega, \mathbf{m})\tilde{\mathbf{v}}_i(\omega) - \sum_j \mathbf{d}_{i,j}(\omega) \right\rangle_{\mathbf{x}} - \\
& \sum_{i=1}^{n_s} \left\langle \tilde{\lambda}_i^r, \tilde{\mathbf{r}}_i + \sum_{\omega} \omega^2 \tilde{\mathbf{v}}_i^*(\omega)\tilde{\mathbf{u}}_i(\omega) \right\rangle_{\mathbf{x}} \Bigg].
\end{aligned} \tag{41}
$$

The adjoint states λ^u, λ^v and λ^r are obtained by taking the derivatives of $\mathcal{L}$ with respect to $\tilde{\mathbf{u}}$, $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{r}}$ equal to zero at the point ($\mathbf{u}$, $\mathbf{v}$, $\mathbf{r}$, λ^u, λ^v, λ^r, $\mathbf{m}$):

$$
\begin{cases}
\lambda_i^r = \alpha_2 \left(Re(\mathbf{r}_i - \mathbf{r}_{i-1}) - Re(\mathbf{r}_{i+1} - \mathbf{r}_i) \right); \\
\mathbf{A}(\omega, \mathbf{m})\lambda_i^v(\omega) = -\omega^2 \lambda_i^{r*} \mathbf{u}_i(\omega); \\
\mathbf{A}^*(\omega, \mathbf{m})\lambda_i^u(\omega) = \\
-\omega^2 \lambda_i^r \mathbf{v}_i(\omega) \; + \alpha_1 \sum_j \mathbf{S}_{i,j}^* (\mathbf{S}_{i,j}\mathbf{u}_i - \mathbf{d}_{i,j}).
\end{cases} \tag{42}
$$

The gradient of J is obtained by

$$
\frac{\partial J}{\partial \mathbf{m}} = \omega^2 Re\left(\sum_i \sum_{\omega} \left\langle \lambda_i^u(\omega), \mathbf{u}_i(\omega) \right\rangle_{\mathbf{x}} + \left\langle \lambda_i^v(\omega), \mathbf{v}_i(\omega) \right\rangle_{\mathbf{x}} \right), \tag{43}
$$

since $\frac{\partial A}{\partial \mathbf{m}} = \frac{\partial A^*}{\partial \mathbf{m}} = -\omega^2$ outside the boundary points. Notice that $\langle \lambda_i^u(\omega), \mathbf{u}_i(\omega) \rangle_{\mathbf{x}}$ is a vector, the scalar product in $\mathbf{x}$ is taken per component i, since we should have written $\lambda_i^u(\omega, \mathbf{x})$ and $\mathbf{u}_i(\omega, \mathbf{x})$.

This application shows the numerical interest of the adjoint-state method with the augmented functional. Indeed, a systematic use of the method automatically produces the result. Using the perturbation as described in the first section can lead to the same result, but its application is a bit more difficult and cumbersome, as shown in the Appendix. The physical interpretation of the adjoint states is difficult to find. We can notice that λ^r is just the perturbation with respect to $\mathbf{r}$ of the DSO functional, λ^u satisfies the adjoint wave equation and λ^v the wave equation.

5 STEREOTOMOGRAPHY

The last example describes an application based on the ray equations. The functional depends on the traveltimes and on the other ray-based parameters. The derivatives of the traveltimes with respect to the velocity are efficiently computed by integrating the slowness perturbations along the ray. There is no real gain to introduce the adjoint states when only the derivatives of the traveltimes with respect to the velocity parameters are required, because the adjoint-state method does not give a more efficient algorithm. The derivatives of the ray trajectories can be evaluated from the paraxial ray equations and the propagator associated with this linear differential system, Cervený (2001) and Farra & Madariaga (1987). When the functional depends on the ray trajectories, the adjoint-state method provides a faster approach to the computation of the gradient of the functional. This case is illustrated with the stereotomography functional.

The purpose of the stereotomography, as described in Billette & Lambaré (1998), is to retrieve the velocity background not only from traveltimes picked on the seismic data but also from slopes of the locally coherent events in the common source and common receiver gathers. This approach differs from the classic traveltime tomography because the picks are interpreted independently from each other without any association to a given interface and they can represent either reflection or refraction events. The (observed) data are a set of source positions, $\mathbf{x}_s$, receiver positions, $\mathbf{x}_r$, two-way traveltimes, T_{sr}^m, the slopes, $\mathbf{p}_s$, at the source locations, and the slopes, $\mathbf{p}_r$, at the receiver locations.

Following Billette & Lambaré (1998), the model parameters are $\mathbf{x}_d$ the subsurface reflection points, θ_s and θ_r the take-off angles of the rays going towards the source, $\mathbf{x}_s$, and towards the receiver, $\mathbf{x}_r$, T_s and T_r the traveltimes along the rays from $\mathbf{x}_d$ towards the source and the receiver and the parameters $(\mathbf{v}_k)$ defining the continuous velocity field, v. The integration parameter along the ray is the time, t. The ray equations are:

$$
\begin{cases}
\frac{\partial \mathbf{x}}{\partial t} = v^2(\mathbf{x})\mathbf{p}; \\
\frac{\partial \mathbf{p}}{\partial t} = -\mathbf{p}^2 v(\mathbf{x}) \nabla v(\mathbf{x}).
\end{cases} \tag{44}
$$

With $\mathbf{y}(t) = \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{p}(t) \end{pmatrix}$ and $f(\mathbf{y}, v) = \begin{pmatrix} v^2(\mathbf{x})\mathbf{p} \\ -\mathbf{p}^2 v(\mathbf{x}) \nabla v(\mathbf{x}) \end{pmatrix}$ the ray equations becomes

$$
\begin{cases}
\mathbf{y}_a(0) = \mathbf{y}_0(\mathbf{x}_d, \theta_a, v); \\
\frac{\partial \mathbf{y}_a}{\partial t} = f(\mathbf{y}_a, v).
\end{cases} \tag{45}
$$

$\mathbf{y}_a$ is computed from $t = 0$ to $t = T_a$. The subscript a represents s or r. $\mathbf{y}_0$ is the function defining the initial conditions. The error functional is

$$
\begin{aligned}
J_{sr} = & \frac{1}{2} c_T (T_s + T_r - T_{sr}^m)^2 + \\
& \frac{1}{2} (\mathbf{y}_r(T_r) - \zeta_r)^T \mathbf{C}_r (\mathbf{y}_r(T_r) - \zeta_r) + \\
& \frac{1}{2} (\mathbf{y}_s(T_s) - \zeta_s)^T \mathbf{C}_s (\mathbf{y}_s(T_s) - \zeta_s),
\end{aligned} \tag{46}
$$

with $\zeta_r = \begin{pmatrix} \mathbf{x}_r \\ \mathbf{p}_r \end{pmatrix}$ and $\zeta_s = \begin{pmatrix} \mathbf{x}_s \\ \mathbf{p}_s \end{pmatrix}$. In eq. (46), c_T is a scalar coefficient and $\mathbf{C}_s$ and $\mathbf{C}_r$ are two diagonal matrices.

5.1 Gradient with the adjoint-state method

After the integration by part of the terms $\int_0^{T_a} (\tilde{\lambda}_a)^T (\frac{\partial \tilde{\mathbf{y}}_a}{\partial t} - f(\tilde{\mathbf{y}}_a, v))\, dt$, the augmented functional reads

$$\begin{aligned}\mathcal{L} = &\frac{1}{2}c_T(T_s + T_r - T_{sr}^m)^2 + \\ &\frac{1}{2}(\tilde{\mathbf{y}}_r(T_r) - \zeta_r)^T \mathbf{C}_r(\tilde{\mathbf{y}}_r(T_r) - \zeta_r) + \\ &\frac{1}{2}(\tilde{\mathbf{y}}_s(T_s) - \zeta_s)^T \mathbf{C}_s(\tilde{\mathbf{y}}_s(T_s) - \zeta_s) + \\ &\int_0^{T_s}\left[\left(\frac{\partial\,\tilde{\lambda}_s}{\partial\,t}\right)^T \tilde{\mathbf{y}}_s + (\tilde{\lambda}_s)^T f(\tilde{\mathbf{y}}_s, v)\right]dt - \\ &(\tilde{\lambda}_s(T_s))^T\tilde{\mathbf{y}}_s(T_s) + (\tilde{\lambda}_s(0))^T\tilde{\mathbf{y}}_s(0) + \\ &\int_0^{T_r}\left[\left(\frac{\partial\,\tilde{\lambda}_r}{\partial\,t}\right)^T \tilde{\mathbf{y}}_r + (\tilde{\lambda}_r)^T f(\tilde{\mathbf{y}}_r, v)\right]dt - \\ &\left(\tilde{\lambda}_r(T_r)\right)^T\tilde{\mathbf{y}}_r(T_r) + (\tilde{\lambda}_r(0))^T\tilde{\mathbf{y}}_r(0) - \\ &\left(\tilde{\lambda}_s^0\right)^T(\tilde{\mathbf{y}}_s(0) - \mathbf{y}_0(\mathbf{x}_d, \theta_s, v)) - \\ &\left(\tilde{\lambda}_r^0\right)^T(\tilde{\mathbf{y}}_r(0) - \mathbf{y}_0(\mathbf{x}_d, \theta_r, v)).\end{aligned} \tag{47}$$

$\mathcal{L}$ is a functional of the state variables, $\tilde{\mathbf{y}}_a$, the adjoint-state variables, $\tilde{\lambda}_a$, and the model parameters, T_s, T_r, θ_s, θ_r, $\mathbf{x}_d$ and $(\mathbf{v}_k)$. The derivatives with respect to the state variables, $\tilde{\mathbf{y}}_a$, give the adjoint-state equations:

$$\begin{cases}\lambda_a(T_a) = \mathbf{C}_a(\mathbf{y}_a(T_a) - \zeta_a);\\ \frac{\partial\,\lambda_a(t)}{\partial\,t} = -\mathbf{A}_a^T(t)\lambda_a(t);\\ \lambda_a^0 = \lambda_a(0),\end{cases} \tag{48}$$

with $\mathbf{A}_a(t) = \frac{\partial\, f(\mathbf{y}_a(t), v)}{\partial\,\tilde{\mathbf{y}}}$.

The derivatives of the augmented functional with respect to the model parameters correspond to the derivatives of J_{sr} with respect to the model parameters. This yields

$$\begin{aligned}\frac{\partial\,J_{sr}}{\partial\,\mathbf{v}_k} &= \int_0^{T_s}(\lambda_s(t))^T\,\frac{\partial\, f(\mathbf{y}_s(t), v)}{\partial\,\mathbf{v}_k}\,dt + \\ &\quad \left(\lambda_s^0\right)^T\frac{\partial\, y_0(\mathbf{x}_d, \theta_s, v)}{\partial\,\mathbf{v}_k} + \\ &\quad \int_0^{T_r}(\lambda_r(t))^T\,\frac{\partial\, f(\mathbf{y}_r(t), v)}{\partial\,\mathbf{v}_k}\,dt + \\ &\quad \left(\lambda_r^0\right)^T\frac{\partial\, y_0(\mathbf{x}_d, \theta_r, v)}{\partial\,\mathbf{v}_k};\\ \frac{\partial\,J_{sr}}{\partial\,\mathbf{x}_d} &= \left(\lambda_s^0\right)^T\frac{\partial\,\mathbf{y}_0(\mathbf{x}_d, \theta_s, v)}{\partial\,\mathbf{x}_d} + \\ &\quad \left(\lambda_r^0\right)^T\frac{\partial\,\mathbf{y}_0(\mathbf{x}_d, \theta_r, v)}{\partial\,\mathbf{x}_d};\\ \frac{\partial\,J_{sr}}{\partial\,\theta_a} &= \left(\lambda_a^0\right)^T\frac{\partial\,\mathbf{y}_0(\mathbf{x}_d, \theta_a, v)}{\partial\,\theta_a};\\ \frac{\partial\,J_{sr}}{\partial\,T_a} &= c_T(T_r + T_s - T_s r) + \\ &\quad (\mathbf{y}_a(T_a) - \zeta_a)^T\mathbf{C}_a f(\mathbf{y}_a(T_a), v).\end{aligned} \tag{49}$$

For the computation of the derivatives with respect to T_a, we have used the fact that the ray equations are satisfied at T_a.

5.2 Gradient with the Fréchet derivatives

A more traditional approach to compute the gradient is to first determine the Fréchet derivatives. In this case, this means the derivatives of T_s, T_r, $\mathbf{y}_s(T_s)$, and $\mathbf{y}_r(T_r)$ with respect to T_s, T_r, $\mathbf{x}_d$, θ_s, θ_r, and $(\mathbf{v}_k)$. A usual approach to compute those derivatives is to use the paraxial ray equations Cervený (2001) and Farra & Madariaga (1987). The perturbation δy_a of the rays $\mathbf{y}_a$ is obtained from the propagator $\mathbf{P}_a$ defined by

$$\begin{cases}\mathbf{P}_a(t_0, t_0) = \mathbf{I};\\ \frac{d\,\mathbf{P}_a(t, t_0)}{d\,t} = \mathbf{A}_a(t)\mathbf{P}_a(t, t_0).\end{cases} \tag{50}$$

This gives, Billette & Lambaré (1998)

$$\begin{cases}\delta y_a(0) = \frac{\partial\, y_0(\mathbf{x}_d, \theta_a, v)}{\partial\,\mathbf{x}_d}\delta\mathbf{x}_d + \\ \qquad\qquad \frac{\partial\, y_0(\mathbf{x}_d, \theta_a, v)}{\partial\,\theta_a}\delta\theta_a + \\ \qquad\qquad \frac{\partial\, y_0(\mathbf{x}_d, \theta_a, v)}{\partial\,v_k}\delta\mathbf{v}_k;\\ \delta y_a(t) = \mathbf{P}_a(t, 0)\delta y_a(0) + \\ \qquad\qquad \int_0^t \mathbf{P}_a(t, t')\frac{\partial\, f(y_a(t'), v)}{\partial\,\mathbf{v}_k}\delta\mathbf{v}_k dt' + \\ \qquad\qquad f(y_a(t), v)\delta t,\end{cases} \tag{51}$$

and the Fréchet derivatives are

$$\begin{cases}\frac{\partial\,\mathbf{y}_a(T_a)}{\partial\,T_a} = f(\mathbf{y}_a(T_a), v);\\ \frac{\partial\,\mathbf{y}_a(T_a)}{\partial\,\theta_a} = \mathbf{P}_a(T_a, 0)\frac{\partial\, y_0(\mathbf{x}_d, \theta_a, v)}{\partial\,\theta_a};\\ \frac{\partial\,\mathbf{y}_a(T_a)}{\partial\,\mathbf{x}_d} = \mathbf{P}_a(T_a, 0)\frac{\partial\, y_0(\mathbf{x}_d, \theta_a, v)}{\partial\,\mathbf{x}_d};\\ \frac{\partial\,\mathbf{y}_a(T_a)}{\partial\,\mathbf{v}_k} = \mathbf{P}_a(T_a, 0)\frac{\partial\, y_0(\mathbf{x}_d, \theta_a, v)}{\partial\,\mathbf{v}_k} + \\ \qquad\qquad \int_0^{T_a}\mathbf{P}_a(T_a, t')\frac{\partial\, f(\mathbf{y}_a(t'), v)}{\partial\,\mathbf{v}_k}\,dt';\\ \quad \frac{\partial\,T_s}{\partial\,T_s} = 1;\\ \quad \frac{\partial\,T_r}{\partial\,T_r} = 1.\end{cases} \tag{52}$$

The other Fréchet derivatives are equal to 0.

The derivatives of J_{sr} are then simply

$$\begin{aligned}\frac{\partial\,J_{sr}}{\partial\,\nu} &= (\mathbf{y}_s(T_s) - \zeta_s)^T\mathbf{C}_s\frac{\partial\,\mathbf{y}_s(T_s)}{\partial\,\nu} + \\ &\quad (\mathbf{y}_r(T_r) - \zeta_r)^T\mathbf{C}_r\frac{\partial\,\mathbf{y}_r(T_r)}{\partial\,\nu},\end{aligned} \tag{53}$$

with ν equals to $\mathbf{v}_k$, $\mathbf{x}_d$, θ_s, or θ_r and

$$\begin{aligned}\frac{\partial\,J_{sr}}{\partial\,T_a} &= c_T(T_s + T_r - T_s r) + \\ &\quad (\mathbf{y}_a(T_a) - \zeta_a)^T\mathbf{C}_a\frac{\partial\,\mathbf{y}_a(T_a)}{\partial\,T_a}.\end{aligned} \tag{54}$$

5.3 Relation between adjoint states and propagator

From eqs (49), (52) and (53), we deduce that

$$\lambda_a^T(t) = (\mathbf{y}_a(T_a) - \zeta_a)^T\mathbf{C}_a\mathbf{P}_a(T_a, t)\,. \tag{55}$$

In fact, the propagator $\mathbf{P}_a^T$ satisfies

$$\begin{cases}\frac{d\,\mathbf{P}_a^T(t, t')}{d\,t'} = -\mathbf{A}_a^T(t')\mathbf{P}_a^T(t, t');\\ \mathbf{P}_a^T(t, t) = \mathbf{I}.\end{cases} \tag{56}$$

$\mathbf{P}_a^T$ is the propagator of the adjoint-state differential equation with a final condition, eq. (48). The adjoint state, λ_a, is then equal to ($\mathbf{C}_a$ is a diagonal matrix):

$$\lambda_a(t) = \mathbf{P}_a^T(T_a, t)\lambda_a(T_a) = \mathbf{P}_a^T(T_a, t)\mathbf{C}_a(\mathbf{y}_a(T_a) - \zeta_a). \tag{57}$$

The main difference between the two approaches lies in the adjoint-state system (eq. 48) and the propagator system (eq. 50). Whereas the first one is a vectorial system, the second is a matrix system. This means that the adjoint-state system is d times smaller than the propagator system, with $d = 4$ in 2-D problems and $d = 6$ in 3-D problems. The adjoint-state method is then roughly d times faster. However, the adjoint-state method does not provide the Jacobian matrix, but

only the gradient of J and the Jacobian may be used to determine the sensitivity of the solution to errors. The minimization should rely on non-linear optimization techniques, such as non-linear conjugate or quasi-Newton methods, Gill *et al.* (1981). In Billette & Lambaré (1998), the authors solve the non-linear optimization problem, by successively solving the linear problems defined by the Jacobian matrices.

6 CONCLUSION

The adjoint-state method for the gradient computation of a functional has been reviewed. The technique applies when the functional depends on the model parameters through a set of state variables, solutions of forward equations. The method consists of the computation of one unique extra linear system. The linear operator is formed with the adjoint of the operator defined by the derivatives of the forward model with respect to the state variables and the second member consists of the derivatives of the functional with respect to the state variables. The adjoint-state variables are the solution of this linear system. In a sense, they gather the information of the perturbations with respect to the state variables, assuming that the state variables are independent variables. Since this linear system is independent of the derivatives with respect to the model parameters, the adjoint states have to be computed only once, making the method numerically very efficient. The gradient of the functional with respect to the model parameters is now simply the scalar product between the adjoint states and the derivatives of the forward model with respect to the model parameters.

To form this extra linear system (the adjoint-state system) a recipe based on an augmented functional has been reviewed. This provides a systematic approach. The main step in the definition of this augmented functional is to consider the state variables and the adjoint-state variables as independent variables.

Several examples have been described to show the power of this approach. For the complicated example with the DSO functional, the adjoint-state equations have been derived from perturbation theory. This shows that the use of the augmented functional is not strictly necessary but simplifies the approach.

When the forward equations are the ray equations, the transpose of the propagator of the perturbed ray equations is the propagator of the adjoint equations. The adjoint-state method is computationally more efficient because the adjoint-state system is a vectorial system whereas the system of the propagator is a matrix system. Nevertheless, the adjoint-state method only gives the gradient, and not the Fréchet derivatives. The optimization problem should be solved with a non-linear optimization method, such a quasi-Newton or non-linear conjugate gradient technique.

ACKNOWLEDGMENTS

I would like to thank Gilles Lambaré from Ecole des Mines de Paris and Colin Perkins from Shell for their comments on the manuscript.

REFERENCES

Al Yahya, K.M., 1989. Velocity analysis by iterative profile migration, *Geophysics,* **54,** 718–729.

Becache, E., 1992. Reflection tomography: how to cope with multiple arrivals? Part II: a new gradient computation method: continuous and discrete problem, in *PSI Consortium Annual Report*.

Billette, F. & Lambaré, G., 1998. Velocity macro-model estimation from seismic reflection data by stereotomography, *Geophys. J. Int.,* **135,** 671–690.

Cervený, V., 2001. *Seismic Ray Method*, Cambridge Univ. Press, New York.

Chavent, G., 1974. Identification of function parameters in partial differential equations, in *Identification of parameter distributed systems*, eds Goodson, R.E. & Polis, New-York, ASME 1974.

Chavent, G. & Jacewitz, C.A., 1995. Determination of background velocities by multiple migration fitting, *Geophysics,* **60,** 476–490.

Ciarlet, P.G., 1989. *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge University Press, New York.

Clearbout, J.F., 1985. *Imaging the Earth's Interior*, Blackwell Scientific Publications, London.

Clément, F., Chavent, G. & Gomez, S., 2001. Migration-based traveltime inversion of 2-D simple structures: a synthetic example, *Geophysics,* **66,** 845–860.

Farra, V. & Madariaga, R., 1987. Seismic waveform modeling in heterogeneous media by ray perturbation theory, *J. geophys. Res.,* **92,** 2697–2712.

Gauthier, O., Virieux, J. & Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: numerical results, *geophysics,* **51,** 1387–1403.

Gill, P.E., Murray, W. & Wright, M.H., 1981. *Practical Optimization*, Academic Press ltd, New York.

Lailly, P., 1983. The seismic inverse problem as a sequence of before stack migration, in *Proc. of Conf. on Inverse Scattering, Theory and Applications*, SIAM, Philadelphia, Pennsylvania.

Lambaré, G., Alerini, M., Baina, R. & Podvin, P., 2004. Stereotomography: a semi-automatic approach for velocity macromodel estimation, *Geophys. Prospect.,* **52,** 671–682.

Lions, J., 1972. *Nonhomogeneous boundary value problems and applications*, Springer Verlag, Berlin.

Liu, Z. & Bleistein, N., 2001. Migration velocity analysis: theory and iterative algorithm, *Geophysics,* **60,** 142–153.

Marfurt, K.J., 1984. Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations, *Geophysics,* **49,** 533–549.

Mora, P., 1989. Inversion = migration + tomography, *Geophysics,* **54,** 1575–1586.

Mulder, W.A. & ten Kroode, A.P.E., 2002. Automatic velocity analysis by differential semblance optimization, *Geophysics,* **67,** 1184–1191.

Plessix, R.-E. & Mulder, W.A., 2004. Frequency-domain finite-difference amplitude-preserving migration, *Geophys. J. Int.,* **157,** 975–987.

Plessix, R.-E., de Roech, Y.-H. & Guy Chavent, 1999. Waveform inversion of reflection seismic data for kinematic parameters by local optimization, *SIAM J. on Scientific Computing,* **20**(3), 1033–1052.

Plessix, R.-E., Mulder, W.A. & ten Kroode, A.P.E., 2000. Automatic cross-well tomography by semblance and differential semblance optimization: theory and gradient computation, *Geophys. Prospect.,* **48,** 913–935.

Shen, P., Symes, W.W. & Stolk, C., 2003. Differential semblance velocity analysis by wave-equation migration, in *Proc. of the 73th SEG int'l mtg*, Dallas, Expanded Abstract.

Symes, W.W. & Carazzone, J.J., 1991. Velocity inversion by differential semblance optimisation, *Geophysics,* **56,** 654–663.

Tarantola, A., 1984. Linearized inversion of seismic reflection data, *Geophys. Prospect.,* **32,** 998–1015.

Tarantola, A., 1987. *Inverse Problem theory*, Elsevier, Amsterdam.

Vinje, V., Iversen, E. & Gjøystdal, H., 1993. Traveltime and amplitude estimation using wave front construction, *Geophysics,* **58,** 1157–1166.

APPENDIX : DSO GRADIENT WITH PERTURBATION APPROACH

In this appendix, we retrieve the gradient of the DSO function directly from the perturbation approach without the help of the augmented functional. If the model, $\mathbf{m}$, is perturbed by $\delta\mathbf{m}$, the wavefields, $\mathbf{u}_i$, are perturbed by $\delta\mathbf{u}_i$, the backpropagated wavefields, $\mathbf{v}_i$,

are perturbed by $\delta\mathbf{v}_i$, the migrated images, $\mathbf{r}_i$, are perturbed by $\delta\mathbf{r}_i$, and the functional J by δJ. From eq. (40) we obtain

$$\delta J = Re\Bigg[\alpha_1 \sum_{i=1}^{n_s} \sum_j \sum_\omega \left\langle \mathbf{S}_{ij}\mathbf{u}_i(\omega) - \mathbf{d}_{ij}(\omega),\, S_{ij}\delta\mathbf{u}_i(\omega) \right\rangle + \alpha_2 \sum_{i=1}^{n_s-1} \langle Re(\mathbf{r}_{i+1} - \mathbf{r}_i),\, \delta\mathbf{r}_{i+1} - \delta\mathbf{r}_i \rangle_{\mathbf{x}} \Bigg]. \quad \text{(A1)}$$

(n_s is the number of shots, i is the shot index and $\langle , \rangle$ is the scalar product in the data space, $\langle , \rangle_{\boldsymbol{x}}$ is the scalar product in spatial coordinate space.)

The perturbations of the state variable equations, eqs (37), (38) and (39), gives

$$\begin{cases} \frac{\partial\,\mathbf{A}}{\partial\,\mathbf{m}}\,\mathbf{u}_i\,\delta\mathbf{m} + \mathbf{A}\,\delta\mathbf{u}_i = 0; \\ \frac{\partial\,\mathbf{A}^*}{\partial\,\mathbf{m}}\,\mathbf{v}_i\,\delta\mathbf{m} + \mathbf{A}^*\,\delta\mathbf{v}_i = 0; \\ \delta\mathbf{r}_i = -\sum_\omega \omega^2(\mathbf{v}_i^*\,\delta\mathbf{u}_i + \delta\mathbf{v}_i^*\,\mathbf{u}_i). \end{cases} \quad \text{(A2)}$$

The dependency on the spatial coordinates and the angular frequency are not written.

The complicated part consists in defining the adjoint-state equations. However with some experience, this is possible. For this example, we first rewrite the second term of eq. (A1)

$$\alpha_2 \sum_{i=1}^{n_s-1} \langle Re(\mathbf{r}_{i+1} - \mathbf{r}_i),\, \delta\mathbf{r}_{i+1} - \delta\mathbf{r}_i \rangle_{\mathbf{x}} = \sum_{i=1}^{n_s} \alpha_2 \,\langle Re(\mathbf{r}_i - \mathbf{r}_{i-1}) - Re(\mathbf{r}_{i+1} - \mathbf{r}_i),\, \delta\mathbf{r}_i \rangle_{\mathbf{x}}\,, \quad \text{(A3)}$$

with $\mathbf{r}_0 = \mathbf{r}_1$ and $\mathbf{r}_{n_s+1} = \mathbf{r}_{n_s}$. And we define

$$\lambda_i^r = \alpha_2(Re(\mathbf{r}_i - \mathbf{r}_{i-1}) - Re(\mathbf{r}_{i+1} - \mathbf{r}_i)). \quad \text{(A4)}$$

Replacing $\delta\mathbf{r}_i$ by its value gives:

$$\delta J = Re\Bigg[\alpha_1 \sum_i \sum_\omega \sum_j \langle \mathbf{S}_{ij}\mathbf{u}_i - \mathbf{d}_{ij},\, \mathbf{S}_{ij}\delta\mathbf{u}_i \rangle + \sum_i \sum_\omega \omega^2 \left\langle \lambda_i^r,\, -(\mathbf{v}_i^*\delta\mathbf{u}_i + \delta\mathbf{v}_i^*\mathbf{u}_i) \right\rangle_{\mathbf{x}} \Bigg]. \quad \text{(A5)}$$

We then gather the terms depending on $\delta\mathbf{u}_i$ and the terms depending on $\delta\mathbf{v}_i$:

$$\delta J = Re\Bigg[\sum_i \sum_\omega \left\langle -\omega^2\mathbf{v}_i\lambda_i^r + \alpha_1 \sum_j \mathbf{e}_{ij},\, \delta\mathbf{u}_i \right\rangle_{\mathbf{x}} + \sum_i \sum_\omega \omega^2 \left\langle -\mathbf{u}_i\lambda_i^{r*},\, \delta\mathbf{v}_i \right\rangle_{\mathbf{x}} \Bigg] \quad \text{(A6)}$$

with $Re(\langle -\lambda_i^r\,\mathbf{u}_i^*,\, \delta\mathbf{v}_i^* \rangle_{\boldsymbol{x}}) = Re(\langle -\mathbf{u}_i\lambda_i^{r*},\, \delta\mathbf{v}_i \rangle_{\boldsymbol{x}})$ and $\mathbf{e}_{ij} = \mathbf{S}_{ij}^*\,(\mathbf{S}_{ij}\,\mathbf{u}_i - \mathbf{d}_{ij})$.

By replacing $\delta\mathbf{v}_i$ by its values in the second term of eq. (A6) we obtain

$$\omega^2 \left\langle -\lambda_i^{r*}\mathbf{u}_i,\, \delta\mathbf{v}_i \right\rangle_{\mathbf{x}} = \omega^2 \left\langle \lambda_i^{r*}\mathbf{u}_i,\, (\mathbf{A}^*)^{-1}\frac{\partial\,\mathbf{A}^*}{\partial\,\mathbf{m}}\mathbf{v}_i \right\rangle_{\mathbf{x}} \delta\mathbf{m} = \omega^2 \left\langle \mathbf{A}^{-1}\lambda_i^{r*}\mathbf{u}_i,\, \frac{\partial\,\mathbf{A}^*}{\partial\,\mathbf{m}}\mathbf{v}_i \right\rangle_{\mathbf{x}} \delta\mathbf{m}. \quad \text{(A7)}$$

By replacing $\delta\mathbf{u}_i$ by its value in the first term of eq. (A6) we obtain

$$\left\langle -\omega^2\mathbf{v}_i\lambda_i^r + \alpha_1 \sum_j \mathbf{e}_{ij},\, \delta\mathbf{u}_i) \right\rangle_{\mathbf{x}} = -\left\langle -\omega^2\lambda_i^r\mathbf{v}_i + \alpha_1 \sum_j \mathbf{e}_{ij},\, A^{-1}\frac{\partial\,A}{\partial\,m}\mathbf{u}_i \right\rangle_{\mathbf{x}} = -\left\langle (A^*)^{-1}\left(-\omega^2\lambda_i^r\mathbf{v}_i + \alpha_1 \sum_j \mathbf{e}_{ij} \right),\, \frac{\partial\,A}{\partial\,m}\mathbf{u}_i \right\rangle_{\mathbf{x}}. \quad \text{(A8)}$$

We can now define

$$\begin{cases} \mathbf{A}\lambda_i^v = -\omega^2\lambda_i^{r*}\mathbf{u}_i; \\ \mathbf{A}^*\lambda_i^u = -\omega^2\lambda_i^r\mathbf{v}_i + \alpha_1 \sum_j \mathbf{e}_{ij}. \end{cases} \quad \text{(A9)}$$

The equations eqs (A4) and (A9) are the adjoint-state equations. They are the same that the ones obtained with the use of the augmented functional in the section on the DSO function, however the derivation is less obvious and systematic in this appendix.

The perturbation δJ now reads:

$$\delta J = -Re\left(\sum_i \sum_\omega \left\langle \lambda_i^u,\, \frac{\partial\,\mathbf{A}}{\partial\,m}\mathbf{u}_i \right\rangle_{\mathbf{x}} + \left\langle \lambda_i^v,\, \frac{\partial\,\mathbf{A}^*}{\partial\,m}\mathbf{v}_i \right\rangle_{\mathbf{x}} \right) \delta m. \quad \text{(A10)}$$

And with $\frac{\partial\,\mathbf{A}^*}{\partial\,m} = \frac{\partial\,\mathbf{A}}{\partial\,m} = -\omega^2$, we obtain:

$$\delta J = \omega^2 Re\left(\sum_i \sum_\omega \left\langle \lambda_i^u(\omega),\, \mathbf{u}_i(\omega) \right\rangle_{\mathbf{x}} + \left\langle \lambda_i^v(\omega),\, \mathbf{v}_i(\omega) \right\rangle_{\mathbf{x}} \right) \delta m. \quad \text{(A11)}$$