Tutorial on the continuous and discrete adjoint state method and basic implementation

Matthew J. Yedlin*, Daryl G. Van Vorst*

ABSTRACT

This tutorial endeavours to lay out the basic scheme using the adjoint state method for setting up the basic equations for inversion of acoustic, elastic or electromagnetic data. The basic scheme is simple and relies on three basic ingredients consisting of a forward modelling package, the definition of a misfit norm and a gradient descent method. The adjoint state method will be presented in its continuous and discrete forms. The continuous adjoint state method will be developed for the full elastic equations in the time domain, while the discrete adjoint state method will be developed in the frequency domain.

INTRODUCTION

There is an abundance of inversion schemes archived in the literature of many different fields including exploration geophysics, aerodynamics, weather prediction and financial derivatives. These fields are so diverse, and yet, there is a common thread running through the methods applied to map acquired data to model parameters. At the core of all these inversion schemes is the application of the adjoint state method.

In this tutorial, we will briefly describe the adjoint state method and then in detail present its application in both the discrete and continuous settings. The continuous adjoint state method will be presented through the examination of the application by Liu and Tromp (Liu and Tromp, 2006) and the continuous discrete adjoint state method will be developed through the analysis of the classic paper of Pratt (Pratt, 1999). An excellent review by Plessix (Plessix, 2006) of the adjoint state method is presented in a more abstract setting. As we shall see in what follows, though the idea is elementary in concept, is is not simple and the devil is in the details.

Adjoint State Method

The adjoint state method has a very long history dating back really to the work of Lagrange (1760), who first presented the famous Lagrange identity to define the adjoint operator. We will see this identity in a more general form in the next section, when we discuss the continuous application. Curiously, successive theoretical developments in this area of application (adjoint methods, calculus of variations, and functional analysis) were also pioneered by French and Russian mathematicians, both pure and applied. We are all familiar with the Fréchet derivative (Maurice René Fréchet,1 878-1973) but how many of us know about Gâteaux (René Gâteaux,1889-1914), Jacques-Louis Lions (Jacques-Louis Lions, 1928-2001) or Gury Ivanovich Marchuk (1925-). These mathematicians have had significant impact on the development of the adjoint state method.

*Department of Electrical and Computer Engineering, University of British Columbia

While we are familiar, for example of the excellent work on inversion by the late Albert Tarantola, whose first paper in Geophysics can be viewed as the launch of formal inversion theory in exploration geophysics (Tarantola, 1984), the mathematical development post Lagrange, has a longer history. It was Marchuk (Marchuk, 1962) who developed the adjoint state technique in nuclear reactor design, which focused on obtaining the parameters for the critical regime of reactor operation. Further theoretical work was done by Lions (Lions, 1971) who laid the foundation for the introduction of optimal control methodology to systems governed by partial differential equations. We intuit, then, that the term adjoint state method originates from optimal control.

In optimal control, we begin with a description of the physical system via a coupled set, in general, of partial differential equations with associated boundary and initial conditions. The variables in this system comprise the state of the system. In addition to these state variables, there are parameters or control variables that are used to drive the system. For example, in seismology we view the state variables as the vector of displacements and the control variables as the material properties (elastic constants and density) and the source properties.

With the system now presented in this general setting, we need to define a performance measure under the application of the particular set of control variables. In exploration seismology, we usually employ the integrated squared error as the system performance measure, with the error defined as the difference between the measurements obtained and the simulation of those measurements. Lions found a clever way to find the best control variables such that the performance measure would be maximized, or in the case of an error measure, minimized. He introduced Lagrange multipliers as a means of incorporating the state equations that describe the evolution of the state variables in time. The Lagrange multipliers that he introduced are called the adjoint state variables and we shall see exactly how they function to give us very well-known results from prestack migration.

CONTINUOUS ADJOINT STATE METHOD – ANALYSIS FOR THE FULL ELASTODYNAMIC EQUATIONS IN THE TIME DOMAIN

In this section, we will present a detailed analysis of the adjoint state method for the elastodynamic system with the state variables denoted as the vector displacements, $\mathbf{s}(\mathbf{x}, t)$, and the control variables: the density, ρ, the elastic constant tensor, $\mathbf{c}$, and the the source, $\mathbf{f}$. We will follow explicitly the development of Liu and Tromp (Liu and Tromp, 2006) We will, for the most part, use exactly the same notation and equation order, and argument development, but we will fill in some missing steps and hopefully give some intuition regarding the final results. We wish to make it abundantly clear that what follows is their original work and we are simply filling in some pieces that not everyone may have in their mathematical handbag.

First, as described earlier, we begin with the performance measure, which is the usual least squares wave-form misfit:

$$\chi = \frac{1}{2}\sum_r \int_0^T \|\mathbf{s}(\mathbf{x_r}, t) - \mathbf{d}(\mathbf{x_r}, t)\|^2 \mathrm{d}t \tag{1}$$

This misfit is the sum over all receivers of the energy difference between the predicted displacements, $\mathbf{s}(\mathbf{x_r}, t)$, and the recorded data, $\mathbf{d}(\mathbf{x_r}, t)$.

We now need our description of the state via the partial differential equations, the state and control variables and initial and boundary conditions, and source description:

State equations and state variables:

The state equation is the elastodynamic equation given by

$$\rho\partial_t^2\mathbf{s} - \nabla\cdot\mathbf{T} - \mathbf{f} = \mathbf{0}, \tag{2}$$

with ρ the density distribution, $\mathbf{f}$ the source function, and $\mathbf{T}$ the stress tensor. The stress tensor is given by Hooke's Law, which is written in compressed form as

$$\mathbf{T} = \mathbf{c}\colon \nabla\mathbf{s}. \tag{3}$$

Equation (3) is a compact form of writing the index representation of the stress tensor, which is normally given in component form by

$$T_{ij} = cijkl\frac{\partial s_k}{\partial x_l}. \tag{4}$$

The colon notation in (3) represents a sum over the two dummy indices, k and l, in (4) and makes for a compact representation.

Boundary and initial conditions:

The boundary conditions correspond to the vanishing of the traction vector on the earth's surface, which will be denoted by $\partial\Omega$. Thus,

$$\hat{\mathbf{n}}\cdot\mathbf{T} = \mathbf{0} \text{ on } \partial\Omega. \tag{5}$$

Implicit, of course, in the foregoing is that often we have to impose artificial absorbing boundary conditions on our computational domain, as we do not often have the complete surface of the earth in our computation. In addition, we must impose the initial conditions for our displacement

$$\begin{aligned} \mathbf{s}(\mathbf{x}, t=0) &= \mathbf{0} \\ \partial_t\mathbf{s}(\mathbf{x}, t=0) &= \mathbf{0}. \end{aligned} \tag{6}$$

Source Description

The source in general is a very complex function but will be modelled in space as a point vector dipole pre-multiplied by the moment tensor $\mathbf{M}$,(contracted over one index), which represents radiation from the source, and post-multiplied by the source-time function, $S(t)$:

$$\mathbf{f} = -\mathbf{M}\cdot\nabla\delta\left(\mathbf{x} - \mathbf{x_{source}}\right)S(t) \tag{7}$$

The point vector dipole is simply the gradient of the location of the source.

With all now in place, the key point is to write down a modified misfit function, which we will continue to denote as χ. To the misfit function in (1), we add zero in a special way. We take the elastic equation (2) and multiply it by a Lagrange multiplier and add it to the misfit. This is the key point in the whole paper and the development of the adjoint method. We note first of all, that the mapping that takes displacements to forces via the elastic equation, is one that maps the vector space in which the displacement lies, to its dual space [Since an appropriate scalar product of force and displacement is a scalar, the energy]. Thus the Lagrange multiplier, $\boldsymbol{\lambda}$, which in this case is really a Lagrange field that depends on spatial and temporal co-ordinates, must be a displacement variable. In fact, the time-reverse Lagrange field is exactly the adjoint field. This result will be demonstrated at the end of this section.

The rest of the analysis is very algebraic, so we will highlight the main points, without going into every detail. The augmented misfit can now be explicitly written as

$$\chi = \frac{1}{2}\sum_r \int_0^T \|\mathbf{s}(\mathbf{x_r},t) - \mathbf{d}(\mathbf{x_r},t)\|^2 \mathrm{d}t - \int_0^T \int_\Omega \boldsymbol{\lambda}\cdot\left[\rho\partial_t^2\mathbf{s} - \nabla\cdot\mathbf{T} - \mathbf{f}\right]\mathrm{d}^3\mathbf{x}\mathrm{d}t. \tag{8}$$

The next step is to take the variation or differential of (8) with respect to functions that can be varied. This includes all state variables and control variables. To write everything in terms of double integrals, we include the delta function to turn on the misfit at the receiver locations $\mathbf{x} = \mathbf{x_r}$. We use the common notation for variation as the delta operator, which takes differentials in function space and obeys the usual product and chain rules for derivatives. The variation operator δ commutes with integration if everything is nice and smooth from a physical perspective. These results have been formally proven in the functional analysis community and we will accept them as given here. Taking the variation of (8) results in

$$\begin{aligned}
\delta\chi \;=\; & \int_0^T\int_\Omega \sum_r \left[\mathbf{s}(\mathbf{x_r},t) - \mathbf{d}(\mathbf{x_r},t)\right]\delta(\mathbf{x}-\mathbf{x_r})\cdot\delta\mathbf{s}\;\mathrm{d}^3\mathbf{x}\mathrm{d}t \\
& -\int_0^T\int_\Omega \boldsymbol{\lambda}\cdot\left[\delta\rho\partial_t^2\mathbf{s} - \boxed{\nabla\cdot(\delta\mathbf{c}\colon\nabla\mathbf{s})} - \delta\mathbf{f}\right]\mathrm{d}^3\mathbf{x}\mathrm{d}t \\
& -\int_0^T\int_\Omega \boldsymbol{\lambda}\cdot\left[\rho\partial_t^2\delta\mathbf{s} - \boxed{\nabla\cdot(\mathbf{c}\colon\nabla\delta\mathbf{s})}\right]\mathrm{d}^3\mathbf{x}\mathrm{d}t,
\end{aligned} \tag{9}$$

where we have dropped the explicit dependency of the displacement field, the Lagrangian field and control variables on the spatial co-ordinates and time. Note that the variation of the Lagrangian field vanishes, since its coefficient simply yields the equation of state (2). Now if we were at a minimum of the augmented misfit, then of course $\delta\chi$ would equal zero. This is not the case, so we must proceed further. We will want to push all derivatives from the displacement field, $\mathbf{s}(\mathbf{x},t)$, and its variation, $\delta\mathbf{s}(\mathbf{x},t)$, onto $\boldsymbol{\lambda}(\mathbf{x},t)$. How do we do this?

We now embark on the big technical step which is the integration by parts of (9). The integration by parts twice in time is direct and the initial conditions for the displacement field (6) are used. To integrate the boxed terms in the previous equation, we must use multi-dimensional integration by parts (an analog of Guass' Theorem). We need to use the

following result

$$\int_\Omega [\nabla \mathbf{u} \colon \mathbf{T}] \, \mathrm{d}^3\mathbf{x} = \int_{\partial\Omega} [\mathbf{u} \cdot (\hat{\mathbf{n}} \cdot \mathbf{T})] \, \mathrm{d}^2\mathbf{x} - \int_\Omega [\mathbf{u} \cdot (\nabla \cdot \mathbf{T})] \, \mathrm{d}^3\mathbf{x} \tag{10}$$

three times with $\mathbf{u} = \boldsymbol{\lambda}$ and $\mathbf{T} = \mathbf{c} \colon \nabla\delta\mathbf{s}$, $\mathbf{u} = \delta\mathbf{s}$ and $\mathbf{T} = \mathbf{c} \colon \nabla\boldsymbol{\lambda}$, and $\mathbf{u} = \boldsymbol{\lambda}$ and $\mathbf{T} = \delta\mathbf{c} \colon \nabla\delta\mathbf{s}$. Employing the above integrations by parts, the initial conditions (6), and setting the variation also of the boundary conditions to zero,

$$\hat{\mathbf{n}} \cdot (\delta\mathbf{c} \colon \nabla\mathbf{s} + \mathbf{c} \colon \nabla\delta\mathbf{s}) = 0 \text{ on } \partial\Omega, \tag{11}$$

we finally obtain:

$$\begin{aligned}
\delta\chi \;=\; & \int_0^T \int_\Omega \sum_r [\mathbf{s}(\mathbf{x_r}, t) - \mathbf{d}(\mathbf{x_r}, t)] \, \delta(\mathbf{x} - \mathbf{x_r}) \cdot \delta\mathbf{s} \, \mathrm{d}^3\mathbf{x}\mathrm{d}t \\
& - \int_0^T \int_\Omega \left(\delta\rho\boldsymbol{\lambda} \cdot \partial_t^2\mathbf{s} + \nabla\boldsymbol{\lambda} \colon \delta\mathbf{c} \colon \nabla\mathbf{s} - \boldsymbol{\lambda} \cdot \delta\mathbf{f} \right) \mathrm{d}^3\mathbf{x}\mathrm{d}t \\
& - \int_0^T \int_\Omega \left[\rho\partial_t^2\boldsymbol{\lambda} - \nabla \cdot (\mathbf{c} \colon \nabla\boldsymbol{\lambda}) \right] \cdot \delta\mathbf{s} \, \mathrm{d}^3\mathbf{x}\mathrm{d}t \\
& - \int_\Omega \left[\rho(\boldsymbol{\lambda} \cdot \partial_t\delta\mathbf{s} - \partial_t\boldsymbol{\lambda} \cdot \delta\mathbf{s}) \right]_T \mathrm{d}^3\mathbf{x} \\
& - \int_0^T \int_{\partial\Omega} \hat{\mathbf{n}} \cdot (\mathbf{c} \colon \nabla\boldsymbol{\lambda}) \cdot \delta\mathbf{s} \, \mathrm{d}^2\mathbf{x}\mathrm{d}t
\end{aligned} \tag{12}$$

We note that $\delta\chi$ must vanish when we find the true model, such that without perturbing the model, $\delta\rho$, $\delta\mathbf{c}$ and $\delta\mathbf{f}$ all vanish. For this to happen, we obtain the following equations for the Lagrangian field,$\boldsymbol{\lambda}$:

$$\rho\partial_t^2\boldsymbol{\lambda} - \nabla \cdot (\mathbf{c} \colon \nabla\boldsymbol{\lambda}) - \sum_r [\mathbf{s}(\mathbf{x_r}, t) - \mathbf{d}(\mathbf{x_r}, t)] \, \delta(\mathbf{x} - \mathbf{x_r}) = \mathbf{0}, \tag{13}$$

with free surface boundary conditions,

$$\hat{\mathbf{n}} \cdot (\mathbf{c} \colon \nabla\boldsymbol{\lambda}) = \mathbf{0} \text{ on } \partial\Omega, \tag{14}$$

and **end** conditions

$$\begin{aligned}
\boldsymbol{\lambda}(\mathbf{x}, t = T) &= \mathbf{0} \\
\partial_t\boldsymbol{\lambda}(\mathbf{x}, t = T) &= \mathbf{0}.
\end{aligned} \tag{15}$$

If the Lagrangian field is chosen to satisfy (13) to (15), then in the presence of perturbations, the variation of the action (12) becomes

$$\delta\chi = -\int_0^T \int_\Omega (\delta\rho\boldsymbol{\lambda} \cdot \partial_t^2\mathbf{s} + \nabla\boldsymbol{\lambda} \colon \delta\mathbf{c} \colon \nabla\mathbf{s} - \boldsymbol{\lambda} \cdot \delta\mathbf{f}) \, \mathrm{d}^3\mathbf{x}\mathrm{d}t \tag{16}$$

We note that the end conditions for the Lagrangian field can be converted to initial conditions by a simple time reversal replacing t by $T - t$. The new time-reversed field obeys the

same equations as given by (13) to (15), the variable t replaced by $T-t$. So end conditions become initial conditions. This time-reversed Lagrangian field only differs from the initial state equations as described in (2) to (7) only in the nature of the source term, in which the adjoint field is driven by the error between the predicted and measured fields. If we define the adjoint field as the time-reversed Lagrange field, with

$$\mathbf{s}^{adjoint}(\mathbf{x},t) = \boldsymbol{\lambda}(\mathbf{x},T-t), \tag{17}$$

it clearly follows that

$$\boldsymbol{\lambda}(\mathbf{x},t) = \mathbf{s}^{adjoint}(\mathbf{x},T-t). \tag{18}$$

If we substitute (18) into (12) and keep all the variable dependencies, something quite remarkable happens. The variation in χ is now given by

$$\begin{aligned}
\delta\chi &= -\int_0^T\int_\Omega \left[\delta\rho(\mathbf{x})\mathbf{s}^{adjoint}(\mathbf{x},T-t)\cdot\partial_t^2\,\mathbf{s}(\mathbf{x},t)\right]\mathrm{d}^3\mathbf{x}\mathrm{d}t \\
&- \int_0^T\int_\Omega \left[\nabla\mathbf{s}^{adjoint}(\mathbf{x},T-t)\colon\delta\mathbf{c}(\mathbf{x})\colon\nabla\mathbf{s}(\mathbf{x},t)\right]\,\mathrm{d}^3\mathbf{x}\mathrm{d}t \\
&+ \int_0^T\int_\Omega \left[\mathbf{s}^{adjoint}(\mathbf{x},T-t)\cdot\delta\mathbf{f}\right]\,\mathrm{d}^3\mathbf{x}\mathrm{d}t
\end{aligned} \tag{19}$$

In its present form, (19) can be used in a gradient descent method to update the material parameters. This can be seen by interchanging the order of integration and factoring out the spatial perturbations of the density and elastic tensor. Then we obtain

$$\begin{aligned}
\delta\chi &= \int_\Omega \delta\rho(\mathbf{x})\underbrace{\left[-\int_0^T \mathbf{s}^{adjoint}(\mathbf{x},T-t)\cdot\partial_t^2\,\mathbf{s}(\mathbf{x},t)\,\mathrm{d}t\right]}_{\text{Fréchet kernel for density}}\,\mathrm{d}^3\mathbf{x} \\
&+ \int_\Omega \delta\mathbf{c}(\mathbf{x})::\underbrace{\left[-\int_0^T \nabla\mathbf{s}^{adjoint}(\mathbf{x},T-t)\,\nabla\mathbf{s}(\mathbf{x},t)\,\mathrm{d}t\right]}_{\text{Fourth order Fréchet kernel tensor}}\,\mathrm{d}^3\mathbf{x} \\
&+ \int_0^T\int_\Omega \left[\mathbf{s}^{adjoint}(\mathbf{x},T-t)\cdot\delta\mathbf{f}\right]\mathrm{d}^3\mathbf{x}\mathrm{d}t.
\end{aligned} \tag{20}$$

The Fréchet kernels in (20) can be used directly in a gradient descent method to update the density and the components of the elastic tensor. Note that to compute these kernels, we only need access to two fields and therefore, in principle, we only need to solve two forward problems. In practice we may need to solve a third forward problem if we are using conjugate gradient descent methods. We will derive a similar expression again in an entirely different way in the next section.

If we look at the Fréchet kernels for both the density and the elastic tensor, we see that at each point we compute the zero-lag cross-correlation of a filtered forward propagated field and a filtered time-reversed adjoint field. This is exactly a different form of the Claerbout reflection mapping principle of reflectors defined to be the locations at which there is

the time-coincidence of downgoing and upgoing waves (Claerbout, 1971). What is truly amazing is that Claerbout expressed the foregoing ideas some 27 years before this analytical tour-de-force of Liu and Tromp. We shall now re-examine the above in the simpler context of discrete systems for the adjoint equation, derived directly in the frequency domain, with the final result in frequency again an independent verification of the Claerbout reflection mapping principle.

DISCRETE ADJOINT STATE METHOD – ANALYSIS FOR THE ELASTIC WAVE EQUATION IN THE FREQUENCY DOMAIN

Pratt (Pratt, 1999) presented, in the frequency domain, a direct way to arrive at an expression for the gradient of the misfit function. He used again, the least squares error functional, in the frequency domain, but did not use the Lagrange dual variable approach. Instead, he directly took the gradient of the misfit, concatenated all the expressions into a single expression and then correctly extracted the adjoint state variables. We will follow his development, essentially equation by equation, and when useful, we will cross-reference the previous section. We will use the same variables described in the previous derivation of the continuous adjoint state method.

We start with the forward problem, which is now in discrete form and includes all boundary conditions. It is represented, in the frequency domain, after temporal Fourier transform of the original partial differential equation, as a linear system of equations given by

$$\mathbf{L}(\omega)\mathbf{s}(\omega) = \mathbf{f}(\omega) \tag{21}$$

The operator $\mathbf{L}(\omega)$ contains all the information about the linear operator which defines the original partial differential equation, including all boundary conditions. The state variables, $\mathbf{s}(\omega)$, are complex-valued functions of frequency and spatial co-ordinates, which are indicated implicitly by the index of the vector in (21). The state variables will depend on the material parameters, which will be denoted by the vector $\mathbf{p}$ of dimension m. The right-hand side of (21) represents all sources. For future clarity, we need to define the indices of the matrices and vectors appearing in (21). The vector $\mathbf{s}(\omega)$ has l elements and consists of the discrete values of the displacement vector at all points in the numerical grid. The size of the matrix $\mathbf{L}(\omega)$ is $l \times l$. We will define the data residual, $\delta\mathbf{d}(\omega)$, the difference between simulated data, $\mathbf{s}(\omega)$, and recorded data, $\mathbf{d}(\omega)$, as a vector, $\mathbf{s}(\omega) - \mathbf{d}(\omega)$, on a sub domain of n receiver positions.

The squared sum of the data residual vector is the norm of the complex data residual and is defined as

$$E(\mathbf{p}) = \frac{1}{2}\delta\mathbf{d}(\omega)^T\delta\mathbf{d}^*(\omega), \tag{22}$$

with the usual notation of superscript T for transpose and $*$ for complex conjugation. In (22) the factor of 1/2 is introduced for convenience.

The next step is to compute the gradient of $E(\mathbf{p})$ with respect to $\mathbf{p}$. This is completely analogous to the computation of the variation of χ in (9) in the analysis of the previous

section. Thus we have, via the chain rule, that

$$\nabla_{\mathbf{p}} E(\mathbf{p}) = \frac{\partial E(\mathbf{p})}{\partial \mathbf{p}} = \Re eal\{\mathbf{J}^t \delta \mathbf{d}^*(\omega)\}. \tag{23}$$

There is a big difference as to how we arrived at the gradient of $E(\mathbf{p})$ and the variation of χ. We did not introduce the Lagrange dual formulation, but computed the gradient directly. The problem is now the calculation of the matrix $\mathbf{J}$. What is this matrix, exactly?

We can see from the (23) that this matrix is the derivative of the displacement vector, $\mathbf{s}(\omega)$, at the n receiver locations, with respect to the m-vector $\mathbf{p}$. It is commonly known as the Fréchet derivative matrix or sensitivity matrix, whose elements are given by

$$J_{ij} = \frac{\partial s_i}{\partial p_j} \qquad i = 1 \ldots n \text{ and } j = 1 \ldots m \tag{24}$$

In practice, we would have to recompute the predicted measurements of the vector displacement m times, a very costly matter in three dimensions, in which we would have to solve, via LU decomposition, the matrix system of equations (21). We would like to avoid all these calculations by concatenating all the required operations in the gradient calculation, regrouping them and arriving at an adjoint formulation analogous to that presented in (20).

To facilitate this reduction of operations, which in fact will lead us to the adjoint state, we redefine our matrix $\mathbf{J}^T$ to be $m \times l$ and the vector $\delta \mathbf{d}^*(\omega)$ to be $l \times 1$ by padding with the appropriate number of zeros. The new matrix will be denoted by $\hat{\mathbf{J}}$, in which the partial derivatives are now computed for the entire grid while the augmented data residual vector will be denoted by $\delta \hat{\mathbf{d}}^*(\omega)$. The gradient of $E(\mathbf{p})$ is now given by

$$\nabla_{\mathbf{p}} E(\mathbf{p}) = \Re eal\{\hat{\mathbf{J}}^t \delta \hat{\mathbf{d}}^*(\omega)\}. \tag{25}$$

Pratt now looks at each column of the matrix $\hat{\mathbf{J}}$ and explicitly computes this column vector directly from the forward problem. He then assembles all the columns to create $\hat{\mathbf{J}}$. This new $\hat{\mathbf{J}}$ is substituted explicitly into (25) and the adjoint state is directly identified. The math steps are given below:

1. Compute the Columns of $\hat{\mathbf{J}}$

   (a) We differentiate both sides of (21), assuming that the sources are not dependent on $\mathbf{p}$ and obtain

   $$\frac{\partial}{\partial p_i} \left(\mathbf{L}(\omega)\mathbf{s}(\omega)\right) = 0. \tag{26}$$

   which becomes

   $$\mathbf{L}(\omega)\frac{\partial \mathbf{s}(\omega)}{\partial p_i} = -\frac{\partial \mathbf{L}(\omega)}{\partial p_i}\mathbf{s}(\omega) \tag{27}$$

(b) We solve (27) for the elements of $\hat{\mathbf{J}}$ in terms of virtual sources $\mathbf{f}^{(i)}_{virt}(\omega)$ to obtain

$$\frac{\partial \mathbf{s}(\omega)}{\partial p_i} = -\mathbf{L}^{-1}(\omega)\frac{\partial \mathbf{L}(\omega)}{\partial p_i}\mathbf{s}(\omega) = \mathbf{L}^{-1}(\omega)\mathbf{f}^{(i)}_{virt}(\omega). \tag{28}$$

2. Assemble the columns of $\hat{\mathbf{J}}$ and substitute into the gradient expression (25).

(a) We substitute the expression from (28) and assemble the matrix $\hat{\mathbf{J}}$ to obtain

$$\hat{\mathbf{J}} = \left[\frac{\partial \mathbf{s}(\omega)}{\partial p_1}\frac{\partial \mathbf{s}(\omega)}{\partial p_2}\cdots\frac{\partial \mathbf{s}(\omega)}{\partial p_m}\right] = \mathbf{L}^{-1}(\omega)\left[\mathbf{f}^{(1)}_{virt}(\omega)\mathbf{f}^{(2)}_{virt}(\omega)\cdots\mathbf{f}^{(m)}_{virt}(\omega)\right]. \tag{29}$$

This can be written in a compact form given by

$$\hat{\mathbf{J}} = \mathbf{L}^{-1}(\omega)\mathbf{F}. \tag{30}$$

(b) We put everything together to obtain a final expression for the gradient. We insert the compact form of the Fréchet derivative matrix (30) into the expression for the gradient (25) to arrive at

$$\nabla_{\mathbf{p}}E(\mathbf{p}) = \Re eal\{\hat{\mathbf{J}}^t\delta\hat{\mathbf{d}}^*(\omega)\} = \Re eal\{\mathbf{F}^T\left[\mathbf{L}^{-1}(\omega)\right]^T\delta\hat{\mathbf{d}}^*(\omega)\}. \tag{31}$$

3. Regroup the calculations and interpret the final form for the gradient as Claerbout's imaging principle.

(a) In (31) we can define a new variable

$$\mathbf{v} = \left[\mathbf{L}^{-1}(\omega)\right]^T\delta\hat{\mathbf{d}}^*(\omega). \tag{32}$$

The expression for the gradient of the error becomes

$$\nabla_{\mathbf{p}}E(\mathbf{p}) = \Re eal\{\mathbf{F}^T\mathbf{v}\}. \tag{33}$$

(b) We can interpret $\mathbf{v}$ as the field obtained by taking the conjugated residual vector defined at the receivers and applying the inverse of exactly the same forward modelling operator as used in (21) [The transpose in (32) disappears under the assumption of reciprocity]. Furthermore, we are in effect back-propagating the time-reversed residuals, since conjugation in frequency is equivalent to time-reversal. We can also interpret the gradient by explicitly writing out the expression for one element of the gradient. Thus we have

$$\frac{\partial E(\mathbf{p})}{\partial p_m} = \Re eal\{-\mathbf{s}(\omega)^T\frac{\partial \mathbf{L}(\omega)}{\partial p_m}^T\mathbf{v}\}. \tag{34}$$

If we neglect the wave-operator in the previous equation, and we consider the first step of our inversion iteration, we see that we have, as Pratt states on page 891 of his paper, the equivalent to prestack reverse-time migration. That is, we have at the reflector, the zero-lag cross-correlation in time. This is simply done by integrating (34) over all frequencies. The field, $\mathbf{v}$, in frequency, is the exact equivalent of

the adjoint field presented in the previous section. Clearly the derivation is much simpler than before. However, we are still left with using the gradient in our computation. Each iteration of our inversion involves solving the forward problem, solving the adjoint problem and solving a third forward problem for finding the size of the step-length as we descend down the gradient to achieve a minimum and our final model. We note again, that by using the adjoint method, we have considerably reduced the computational load by avoiding the solution of many forward problems for each iteration of our inversion algorithm.

OUTLINE OF BASIC IMPLEMENTATION

In the previous sections, we have presented, in tutorial fashion, the essence of the adjoint method, in all its different guises as applied to geophysical inverse problems. What remains is to program the foregoing algorithms. We now indicate at a high-level how this is done. We iterate the algorithm presented below until convergence or until the maximum number of iterations is reached:

- Solve the forward problem to find the residuals and $E(\mathbf{p})$. The forward problem is in itself not trivial, but can be implemented in either time or frequency and spatially discretized in many different ways.
- Solve the adjoint problem and in combination with the solution to the forward problem obtain $\nabla_{\mathbf{p}} E(\mathbf{p})$. That, for example in the temporal, continuous case, would involve calculation of the Fréchet kernels as described in (20). In the discrete frequency adjoint method, we would compute (31).
- Perform a line search to find α (one or more forward modelling steps) if a conjugate gradient-type method is used.
- Update the model: $\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \alpha \nabla_{\mathbf{p}} E(\mathbf{p})$

CONCLUSIONS

In this tutorial paper, we have presented in considerable detail, the continuous adjoint state method for the elastodynamic equation and showed how to derive the Fréchet kernels used in a gradient descent method. We followed this by deriving the adjoint state method for the discretized elastodynamic equations directly in the frequency domain. All the messy details were included (and hidden!) inside the forward modelling operator. In both cases, the final result for the gradient of the squared misfit function was related to Claerbout's reflector mapping principle.

REFERENCES

Claerbout, J. F., 1971, Toward a unified theory of reflector mapping: Geophysics, **36**, 467–481.

Lions, J. L., 1971, Optimal control of systems governed by partial differential equations: Springer Verlag.

Liu, Q., and Tromp, J., 2006, Finite-frequency kernels based on adjoint methods: Bulletin of the Seismological Society of America, **96**, No. 6, 2383–2397.

Marchuk, G., 1962, A survey of nuclear-reactor design methods: Atomic Energy, **11**, No. 4, 1000–1016.

Plessix, R. E., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, No. 2, 495–503.

Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model: Geophysics, **64**, No. 3, 888–901.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, No. 8, 1259–1266.