Lecture 19: Waveform tomography

David Al-Attar, Michaelmas term 2023

Outline and motivation

In the previous lecture we discussed delay time tomography, this being the first and simplest tomographic method developed. While this approach is still used, it has been superseded by **waveform tomography**. Delay time tomography is based on the measurement of a small number of delay times from each seismogram, with these data modelled using linearised ray theory. By contrast, waveform tomography uses more of the data contained the seismograms, fully accounts for the complexities of elastic wave propagation in the Earth, and does not linearise the inverse problem. This lecture will outline the ideas behind waveform tomography, focusing on the solution of the inverse problem using gradient-based optimisation.

Numerical wavefield simulations

Waveform tomography depends crucially on the ability to solve the full elastic wave equation numerically within realistic earth models. This has been possible routinely on a global scale for around 15 years, with this process facilitated by both algorithmic and hardware advances. These lecture are not about numerical methods. But it will be seen that within this subject computational costs must be carefully considered. As a number to bear in mind, the simulation of elastic wave propagation on a global scale within a realistic 3D earth model takes of order one hour running in parallel on a modest number of cores (say 100). Within a tomographic study, it would be typical to use data from around 10^2-10^3 earthquakes, and a separate numerical simulation is required for each one. By running these different calculations separately on a very large cluster, the time taken for the full forward problem could be reduced to about an hour.

Waveform measurements

Let $u_i^{\text{obs}}(t)$ denote the displacement vector recorded by a seismograph at the surface location, $\mathbf{x}_r$, following an earthquake. In reality, seismographs do not record the displacement vector directly, with the signal instead being modified due to the mechanical and electronic properties of the instrument along with additional processing done by the seismologist. For simplicity, however, we neglect such effects, but they can be incorporated at the cost of only more complicated notations.

Suppose that we have models for both the Earth and earthquake, and so can numerically solve the elastic wave equation to obtain a **synthetic seismogram** $u_i(\mathbf{x}_r, t)$ at this location (see Fig.1). To quantify the misfit between the observed and synthetic seismograms we can define the least squares misfit

$$J = \frac{1}{2}\int_0^T \|\mathbf{u}(\mathbf{x}_r, t) - \mathbf{u}^{\text{obs}}(t)\|^2 \,\mathrm{d}t, \tag{1}$$

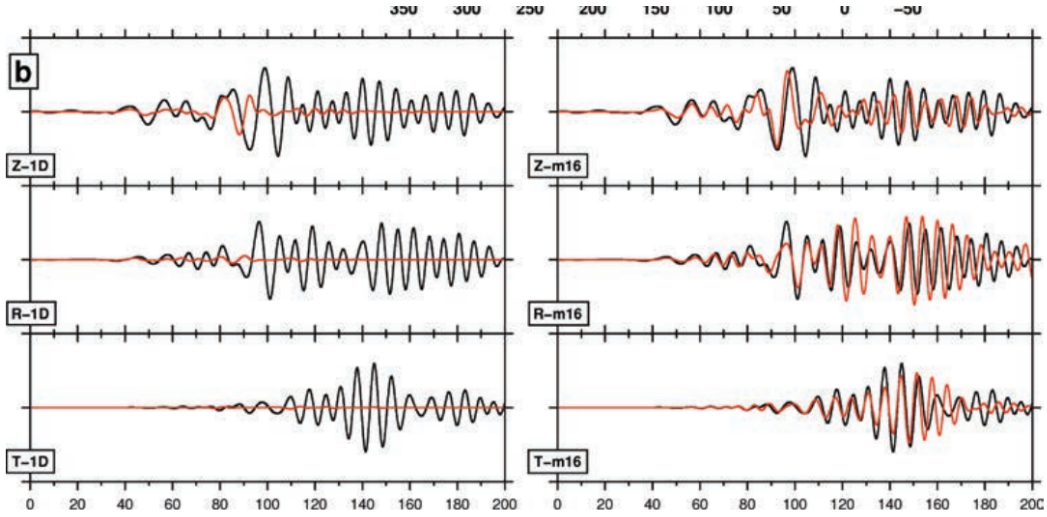

Fig. 1: An example of observed and synthetic three-component seismograms from a waveform tomographic study of Southern California. Data is shown in black, and synthetics in red. On the left the comparison is made prior to inversion, with the synthetics calculated in a laterally homogeneous model. The right hand side then shows the data fit following inversion involving 16 iterative model updates. Taken from Tape *et al.* (2010).

with $[0, T]$ is the time interval over which measurements have been made. Recorded seismograms are, of course, discretely sampled, and so really we should write a sum and not an integral over time. But in practice sampling rates are sufficiently high that this distinction can be ignored.

The value of J depends on the synthetic seismogram, u_i, and hence ultimately on the underlying parameters for the earth model and the earthquake. We could, therefore, try to minimise J in the hope that the model parameters are then closer their true values. In practice, of course, we would not actually use J as defined above, but sum such terms over many different locations and earthquakes, while regularisation terms might also be added. In fact, least-squares misfit between seismograms is rarely used in practice, with other more complicated measures of misfit being preferred. We will see one such example at the end of this lecture.

Functional derivatives and sensitivity kernels

The waveform misfit, J, defined above depends on the synthetic wavefield, $\mathbf{u}$, and hence we can write it as $J(\mathbf{u})$. But the value of $\mathbf{u}$ is determined through solution of the equations of motion from the underlying model parameters. To simplify notations, let us write $\mathbf{m}$ for these model parameters. Here $\mathbf{m}$ could be all of ρ, A_{ijkl} and $\bar{S}_{ij}$, but in practice it is often some subset of these parameters. For example, within a tomographic problem it might be assumed that the source parameters are known. For a given value of $\mathbf{m}$ we can compute the corresponding synthetic wavefield, $\mathbf{u}$, and from this the value of the misfit, $J(\mathbf{u})$. To indicate that the misfit can be viewed as an implicit function of $\mathbf{m}$, we will write it as $J(\mathbf{m})$. This is a slight abuse of notation, but it should always be clear from context which variables the misfit depends on.

Within waveform tomography, the goal is to find model parameters that explain the observed data, and this is done by minimising $\mathbf{m} \mapsto J(\mathbf{m})$. For a small model perturba-

tion, $\delta\mathbf{m}$, we can write to first-order accuracy

$$J(\mathbf{m}+\delta\mathbf{m}) = J(\mathbf{m}) + \langle\, DJ(\mathbf{m}) \,|\delta\mathbf{m}\,\rangle + O(\|\delta\mathbf{m}\|^2), \tag{2}$$

where $DJ(\mathbf{m})$ denotes the functional derivative of J with respect to $\mathbf{m}$ and $\langle\, \cdot\,|\cdot\,\rangle$ is an appropriate inner product. Note that the term **gradient** can be used interchangeably with functional derivative within this context. As an example, supposing that the shear velocity, β, is the only model parameter we can concretely write

$$\langle\, DJ(\mathbf{m}) \,|\delta\mathbf{m}\,\rangle = \int_M K_\beta\, \delta\beta\, \mathrm{d}^3\mathbf{x}, \tag{3}$$

where the function K_β is known as a **sensitivity kernel** for J with respect to β. We discuss later how these functional derivatives can be obtained practically, but for the moment we just assume that this can be done.

Gradient based optimisation

If $\bar{\mathbf{m}}$ is a minimum of J, then for any sufficiently small $\delta\mathbf{m}$ we should have

$$J(\bar{\mathbf{m}}+\delta\mathbf{m}) \geq J(\bar{\mathbf{m}}), \tag{4}$$

which is true if and only if

$$DJ(\bar{\mathbf{m}}) = \mathbf{0}. \tag{5}$$

This condition is, of course, familiar from calculus in finite-dimensional spaces.

There are a number of approaches to solve this optimisation problem. One is a systematic "grid search" of the model space. Doing this would require the model space to be discretised using a finite-dimensional basis as in Lecture 18. The issue with this method is the computational cost. Suppose, for example, that the discretised model space had n-dimensions, and that along each dimension experience had shown that m points needed to be tested. The number of forward calculations required would then be m^n. In realistic problems, n is of order thousands or millions, while we have already noted the high cost of solving the forward problem. The result is that grid search methods are not viable. Random sampling methods such as **Markov Chain Monte Carlo** (MCMC) simulations perform better within high-dimensional spaces (the number of forward calculations scales with something like n^2), but the costs are still far to high to be used in practically in waveform tomography. As a result, tomographers must rely on local optimisation methods that seek a point such that eq.(5) holds. An obvious downside to this approach is that the model obtained may only be a local and not a global minimum of the misfit. But this is what is currently possible, the results obtained do seem to have value, and there are some approximate but useful methods available for uncertainty quantification.

Suppose we start from an initial model $\mathbf{m}_0$. At this point we can calculate the value of the misfit, $J(\mathbf{m}_0)$, along with its gradient, $DJ(\mathbf{m})$. The gradient points in the direction along which the misfit most rapidly increases, and the negative gradient gives the direction of most rapid decrease. This suggests that we can look for a new model of the form

$$\mathbf{m}_1 = \mathbf{m}_0 - \lambda\, DJ(\mathbf{m}_0), \tag{6}$$

where $\lambda > 0$ is some step-length to be determined. To find a suitable value for λ, we need only consider a one-dimensional minimisation problem for the function

$$f(\lambda) = J[\mathbf{m}_0 - \lambda\, DJ(\mathbf{m}_0)], \tag{7}$$

which can be solved quite easily. Having now arrived at a new model, $\mathbf{m}_1$, with a lower misfit we can simply repeat the whole process until the gradient vanishes and a local minimum has been found[1]. The above method is the simplest example of **gradient-based optimisation** known as **steepest descents**. More sophisticated methods for choosing the **descent direction** tend to be used in practice which have superior convergence properties. In all cases, however, this direction is built from linear combinations of the current and past gradients of the misfit.

How not to calculate the gradient of of J

Gradient based optimisation depends on being able to both solve the forward problem, and to calculate the gradient of the misfit. As noted, the first task can be done using numerical wavefield simulations at a high but manageable cost. Schematically, such a code provides a means for performing the mapping $\mathbf{m} \mapsto J(\mathbf{m})$. With this at hand, one way to get the required gradient is to make use of the finite-difference approximation

$$\langle\, DJ(\mathbf{m}) \,|\delta\mathbf{m}\,\rangle \approx J(\mathbf{m}+\delta\mathbf{m}) - J(\mathbf{m}). \tag{8}$$

Using this formula, we can get the derivative of the misfit along any direction at the cost of two forward calculations. Suppose we introduce an n-dimensional parametrisation for the model space. Using $n+1$ forward calculations we could then approximate the gradient relative to this basis. The advantage of this approach is that it requires no additional theory or coding once a numerical wave propagation code has been developed. The problem is, as noted before, the number of model parameters, n, is typically large and so the overall cost is far too high.

How you should calculate the gradient of of J

Our aim is to calculate the gradient of $J(\mathbf{m})$. This is equivalent to finding the gradient of $J(\mathbf{u})$ subject to the *constraint* that $\mathbf{u}$ is related to $\mathbf{m}$ through solution of the forward problem. To proceed, we make use of the method of **Lagrange multipliers** which most of you will have seen previously, though likely in simpler contexts. Within this course it is necessary to know how to *apply* this method, but not to explain why it works. For those interested a **non-examinable** outline is provided in the appendix to this lecture.

We begin by defining a Lagrangian for the problem which takes the following form

$$\begin{aligned} L(\mathbf{u},\mathbf{u}',\mathbf{w}',\mathbf{m}) = J(\mathbf{u}) &- \int_0^T \int_M \left[\rho \frac{\partial^2 u_i}{\partial t^2} - \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial u_k}{\partial x_l} - \bar{S}_{ij}\right)\right] u_i' \,\mathrm{d}^3\mathbf{x}\,\mathrm{d}t \\ &- \int_0^T \int_{\partial M} \left(A_{ijkl}\frac{\partial u_k}{\partial x_l} - \bar{S}_{ij}\right) \hat{n}_j w_i' \,\mathrm{d}S\,\mathrm{d}t. \end{aligned} \tag{9}$$

Here we see first on the right hand side the unconstrained misfit, $J(\mathbf{u})$. The second term is associated with the constraint that $\mathbf{u}$ should satisfy the equations of motion and includes a Lagrange multiplier field $\mathbf{u}'$. The final term does the same for the boundary conditions, and features a second Lagrangian multiplier field $\mathbf{w}'$. The **Lagrange multiplier theorem** tells us that the equality

$$DJ(\mathbf{m}) = D_{\mathbf{m}} L(\mathbf{u},\mathbf{u}',\mathbf{w}',\mathbf{m}), \tag{10}$$

[1] In practice, the process might be stopped when the misfit has decreased sufficiently much from a statistical perspective.

holds so long as the following conditions are met

$$D_{\mathbf{u}} L(\mathbf{u}, \mathbf{u}', \mathbf{w}', \mathbf{m}) = \mathbf{0}, \tag{11}$$
$$D_{\mathbf{u}'} L(\mathbf{u}, \mathbf{u}', \mathbf{w}', \mathbf{m}) = \mathbf{0}, \tag{12}$$
$$D_{\mathbf{w}'} L(\mathbf{u}, \mathbf{u}', \mathbf{w}', \mathbf{m}) = \mathbf{0}. \tag{13}$$

Here subscripts are used for partial functional derivatives in an obvious manner.

To understand this result, we begin with eq.(12). The dependence of the Lagrangian on the field $\mathbf{u}'$ is very simple, and so we readily find that

$$\langle\, D_{\mathbf{u}'} L(\mathbf{u}, \mathbf{u}', \mathbf{w}', \mathbf{m}) \,|\delta \mathbf{u}'\,\rangle = -\int_0^T \int_M \left[\rho \frac{\partial^2 u_i}{\partial t^2} - \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial u_k}{\partial x_l} - \bar{S}_{ij}\right)\right] \delta u_i' \,\mathrm{d}^3\mathbf{x}\,\mathrm{d}t. \tag{14}$$

We need this expression to vanish for all possible choices of $\delta\mathbf{u}'$, and hence we conclude that $\mathbf{u}$ must satisfy the expected equations of motion

$$\rho \frac{\partial^2 u_i}{\partial t^2} - \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial u_k}{\partial x_l} - \bar{S}_{ij}\right) = 0. \tag{15}$$

Similarly, the final condition in eq.(13) gives the boundary conditions for $\mathbf{u}$.

Turning now to the first condition in eq.(11), we find that for any $\delta\mathbf{u}$

$$\begin{aligned}\langle\, D_{\mathbf{u}} L(\mathbf{u}, \mathbf{u}', \mathbf{w}', \mathbf{m}) \,|\delta \mathbf{u}\,\rangle =& \langle\, DJ(\mathbf{u}) \,|\delta \mathbf{u}\,\rangle - \int_0^T \int_M \left[\rho \frac{\partial^2 \delta u_i}{\partial t^2} - \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right)\right] u_i' \,\mathrm{d}^3\mathbf{x}\,\mathrm{d}t \\ &- \int_0^T \int_{\partial M} \left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right) \hat{n}_j w_i' \,\mathrm{d}S\,\mathrm{d}t,\end{aligned} \tag{16}$$

must vanish. This is a more complicated expression, and it will take some work to see what it means. But we know that $\mathbf{u}$ is fixed uniquely the previous constraints, and hence the above expression must somehow determine the Lagrange multiplier fields.

As a first step, we can use eq.(1) to explicitly write the first term on the right hand side as

$$\langle\, DJ(\mathbf{u}) \,|\delta \mathbf{u}\,\rangle = \int_0^T [u_i(\mathbf{x}_r, t) - u_i^{\mathrm{obs}}(t)] \delta u_i(\mathbf{x}_r, t)\,\mathrm{d}t. \tag{17}$$

Making use of the Dirac delta function, this can be conveniently written

$$\langle\, DJ(\mathbf{u}) \,|\delta \mathbf{u}\,\rangle = \int_0^T \int_{\partial M} h_i' \delta u_i \,\mathrm{d}S\,\mathrm{d}t, \tag{18}$$

where we have set

$$h_i'(\mathbf{x}, t) = [u_i(\mathbf{x}_r, t) - u_i^{\mathrm{obs}}(t)]\delta(\mathbf{x} - \mathbf{x}_r). \tag{19}$$

With this done, we can write the required condition as

$$\begin{aligned}\int_0^T \int_{\partial M} h_i' \delta u_i \,\mathrm{d}S\,\mathrm{d}t - \int_0^T \int_M \left[\rho \frac{\partial^2 \delta u_i}{\partial t^2} - \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right)\right] u_i' \,\mathrm{d}^3\mathbf{x}\,\mathrm{d}t \\ - \int_0^T \int_{\partial M} \left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right) \hat{n}_j w_i' \,\mathrm{d}S\,\mathrm{d}t = 0,\end{aligned} \tag{20}$$

for all $\delta\mathbf{u}$. To proceed, we need to successively integrate by parts to shift derivatives from the perturbed displacement field to the Lagrange multipliers. Starting with the time-derivatives, we have

$$\int_0^T \frac{\partial^2 \delta u_i}{\partial t^2} u_i' \,\mathrm{d}t = -\int_0^T \frac{\partial \delta u_i}{\partial t}\frac{\partial u_i'}{\partial t}\,\mathrm{d}t + \left[\frac{\partial \delta u_i}{\partial t} u_i'\right]_{t=T}, \tag{21}$$

where we have used the vanishing initial conditions on u_i (and hence δu_i) to eliminate one of the boundary terms. Integrating by parts once more gives

$$\int_0^T \frac{\partial^2 \delta u_i}{\partial t^2} u_i' \,\mathrm{d}t = \int_0^T \delta u_i \frac{\partial^2 u_i'}{\partial t^2}\,\mathrm{d}t + \left[\frac{\partial \delta u_i}{\partial t} u_i'\right]_{t=T} - \left[\delta u_i \frac{\partial u_i'}{\partial t}\right]_{t=T}. \tag{22}$$

Looking now at the spatial derivatives, we first obtain

$$\begin{aligned}
&\int_M \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right) u_i' \,\mathrm{d}^3\mathbf{x} - \int_{\partial M}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right)\hat{n}_j w_i' \,\mathrm{d}S \\
&= -\int_M A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\frac{\partial u_i'}{\partial x_j}\,\mathrm{d}^3\mathbf{x} - \int_{\partial M}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right)\hat{n}_j (w_i' - u_i')\,\mathrm{d}S.
\end{aligned} \tag{23}$$

Making use of the hyperelastic symmetry, $A_{ijkl} = A_{klij}$, we can then integrate by parts once more to arrive at

$$\begin{aligned}
&\int_M \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right) u_i' \,\mathrm{d}^3\mathbf{x} - \int_{\partial M}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right)\hat{n}_j w_i' \,\mathrm{d}S \\
&= \int_M \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial u_k'}{\partial x_l}\right)\delta u_i \,\mathrm{d}^3\mathbf{x} - \int_{\partial M}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right)\hat{n}_j (w_i' - u_i')\,\mathrm{d}S \\
&\qquad - \int_{\partial M}\left(A_{ijkl}\frac{\partial u_k'}{\partial x_l}\right)\hat{n}_j \delta u_i \,\mathrm{d}S.
\end{aligned} \tag{24}$$

Putting these results together, we have shown that eq.(20) is equivalent to

$$\begin{aligned}
&-\int_0^T\int_M \left[\rho\frac{\partial^2 u_i'}{\partial t^2} - \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial u_k'}{\partial x_l}\right)\right]\delta u_i \,\mathrm{d}^3\mathbf{x}\,\mathrm{d}t \\
&-\int_0^T\int_{\partial M}\left(A_{ijkl}\frac{\partial \delta u_k}{\partial x_l}\right)\hat{n}_j (w_i' - u_i')\,\mathrm{d}S\,\mathrm{d}t \\
&+\int_0^T\int_{\partial M}\left[h_i' - \left(A_{ijkl}\frac{\partial u_k'}{\partial x_l}\right)\hat{n}_j\right]\delta u_i \,\mathrm{d}S\,\mathrm{d}t \\
&-\int_M \rho\left\{\left[\frac{\partial \delta u_i}{\partial t}u_i'\right]_{t=T} - \left[\delta u_i\frac{\partial u_i'}{\partial t}\right]_{t=T}\right\}\mathrm{d}^3\mathbf{x} = 0,
\end{aligned} \tag{25}$$

for all $\delta\mathbf{u}$. From the first term in this equation, we conclude that the Lagrange multiplier field $\mathbf{u}'$ must satisfy the elastic wave equation

$$\rho\frac{\partial^2 u_i'}{\partial t^2} - \frac{\partial}{\partial x_j}\left(A_{ijkl}\frac{\partial u_k'}{\partial x_l}\right) = 0. \tag{26}$$

In dealing with the boundary terms, we recall that for an arbitrary function its boundary value and that of its normal derivative can be chosen separately. As a result, the second term implies that $\mathbf{w}'$ is the restriction of $\mathbf{u}'$ to ∂M, the third term implies

$$\left(A_{ijkl}\frac{\partial u_k'}{\partial x_l}\right)\hat{n}_j = h_i', \tag{27}$$

and finally we have the **terminal** conditions

$$u_i'(\mathbf{x}, T) = \frac{\partial u_i'}{\partial t}(\mathbf{x}, T) = 0. \tag{28}$$

It has taken some effort, but we have now shown that eq.(11) requires the Lagrange multiplier field, $\mathbf{u}$, to satisfy an elastic wave equation, with this problem driven by a traction applied at the observation point and being equal to the difference between the predicted and observed seismograms. It has also shown that the second Lagrange multiplier field, $\mathbf{w}$, is just the restriction of $\mathbf{u}$ to the boundary, and hence no additional work is necessary for its calculation.

We can now use eq.(10) to obtain sensitivity kernels with respect to the different model parameters. For arbitrary variations in these parameters we have

$$\begin{aligned}\langle\, DJ(\mathbf{m}) \,|\delta\mathbf{m}\,\rangle = &- \int_0^T \int_M \left[\delta\rho \frac{\partial^2 u_i}{\partial t^2} - \frac{\partial}{\partial x_j}\left(\delta A_{ijkl}\frac{\partial u_k}{\partial x_l} - \delta\bar{S}_{ij}\right)\right] u_i' \,\mathrm{d}^3\mathbf{x}\,\mathrm{d}t \\ &- \int_0^T \int_{\partial M} \left(\delta A_{ijkl}\frac{\partial u_k}{\partial x_l} - \delta\bar{S}_{ij}\right)\hat{n}_j u_i' \,\mathrm{d}S\,\mathrm{d}t. \end{aligned} \tag{29}$$

Integrating by parts once, this expression can be simplified

$$\langle\, DJ(\mathbf{m}) \,|\delta\mathbf{m}\,\rangle = \int_0^T \int_M \left[\delta\rho \frac{\partial u_i}{\partial t}\frac{\partial u_i'}{\partial t} - \delta A_{ijkl}\frac{\partial u_i'}{\partial x_j}\frac{\partial u_k}{\partial x_l} + \delta\bar{S}_{ij}\frac{\partial u_i'}{\partial x_j}\right] \mathrm{d}^3\mathbf{x}\,\mathrm{d}t, \tag{30}$$

from which we can read off the various sensitivity kernels

$$K_\rho = \int_0^T \frac{\partial u_i}{\partial t}\frac{\partial u_i'}{\partial t}\,\mathrm{d}t, \tag{31}$$

$$K_{A_{ijkl}} = -\int_0^T \frac{\partial u_i'}{\partial x_j}\frac{\partial u_k}{\partial x_l}\,\mathrm{d}t, \tag{32}$$

$$K_{\bar{S}_{ij}} = \frac{\partial u_i'}{\partial x_j}. \tag{33}$$

The approach we have described is known as the **adjoint method**, and in this context u_i' is known as the **adjoint wavefield** and h_i' the **adjoint traction**[2]. The steps needs to calculate the gradient of the misfit can be summarised as follows:

1. Solve the forward equations for u_i ;
2. From this solution determine J and the associated adjoint traction h_i';
3. Solve the adjoint equations for u_i';
4. Construct the sensitivity kernels by combining u_i and u_i' appropriately.

Noting that the adjoint problem is identical in form to the original, the cost of this method is just *two* numerical solutions of the elastic wave equation. In fact, within practical applications it is usual to perform three solutions, this being done so the forward field u_i need not be stored at all times when assembling the sensitivity kernels. Even so, the

[2] In detail, it is conventional to define the adjoint wavefield by time-reversing u_i' such that the terminal conditions obtained are transformed into initial conditions.

saving relative to the finite-difference approach is by a factor of $\frac{3}{n+1}$ with n the number of model parameters. Given that n in practice is of order 10^4-10^6 this is an enormous improvement and justifies a few pages of algebra. It is the adjoint method that allows waveform inversion to be a practical tool within seismology and also in a range of other fields. Indeed, it is this method that underlies numerical weather prediction.

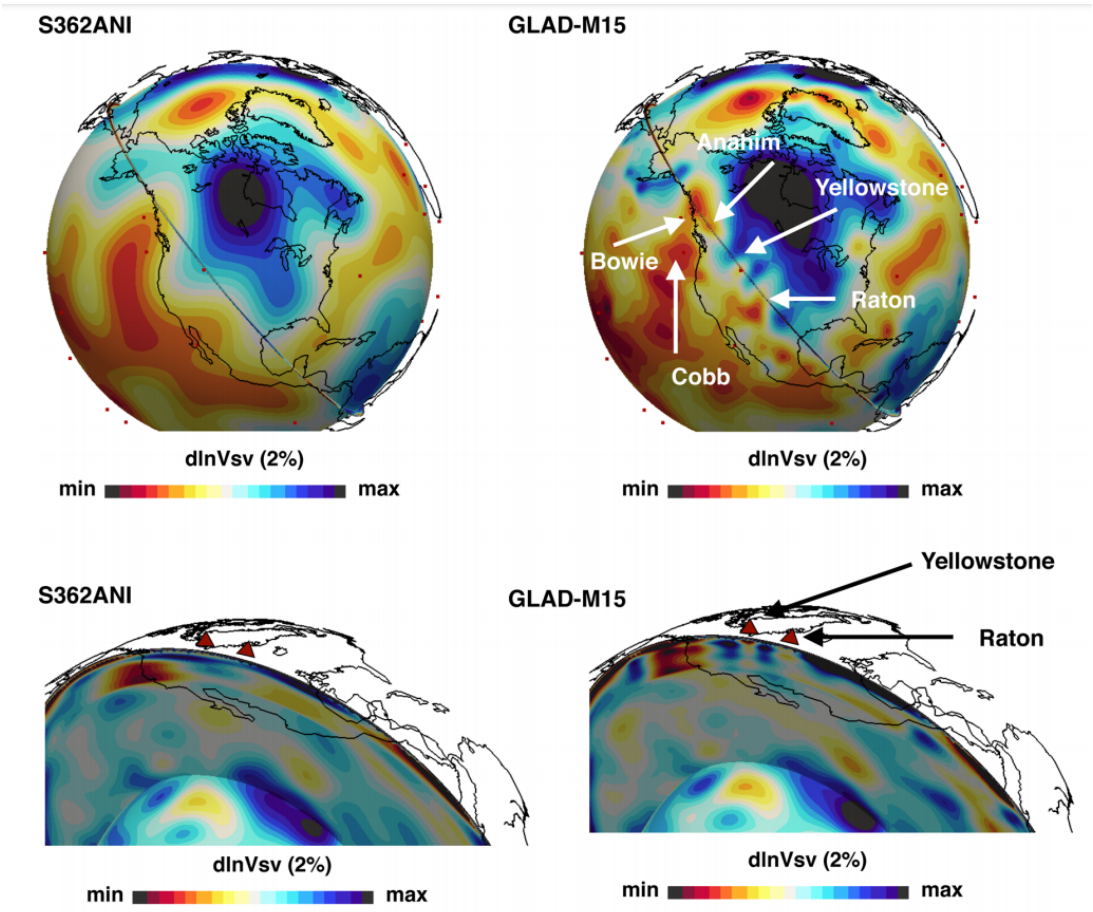

Fig. 2: Images from the first global waveform tomographic model produced by Bozdag *et al.* (2016). On the left is shown the starting model S362ANI which was obtained using ray theoretic methods, and on the right can be seen the waveform model following 15 iterations. While the difference in the models might not seem huge, they are still significant. There is considerably more detail within the waveform model, including structures associated with a range of hot spots including that under Yellowstone. Moreover, as methods continue to improve the difference between ray theoretic and waveform tomographic models will only grow further.

It is worth commenting that though we have obtained the adjoint equations for a specific choice of misfit, the form of J only enters into the problem through the adjoint traction h_i'. This means that the theory can be readily applied to other measures of misfit without having to go through the lengthy derivation each time.

Banana Doughnut kernels

Let $u_i^{\text{obs}}(t)$ again denote the displacement vector recorded by a seismograph at the surface location, $\mathbf{x}_s$, following an earthquake. Suppose we define

$$s^{\text{obs}}(t) = w(t)\hat{\nu}_i u_i^{\text{obs}}(t), \tag{34}$$

with $\hat{\nu}_i$ a unit vector that selects a component of the displacement of interest, and $w(t)$ a windowing function that zeros everything out except for a particular seismic phase of interest. Exactly the same thing can be done with a synthetic seismogram

$$s(\mathbf{x}_s, t) = w(t)\hat{\nu}_i u_i(\mathbf{x}_s, t), \tag{35}$$

with u_i calculated in some reference model. The difference between $s^{\text{obs}}(t)$ and $s(\mathbf{x}_s, t)$ could be quantified using a least-squares misfit as before, but this is not the only choice. Suppose instead we consider the **cross-correlation** of these two time series as defined by

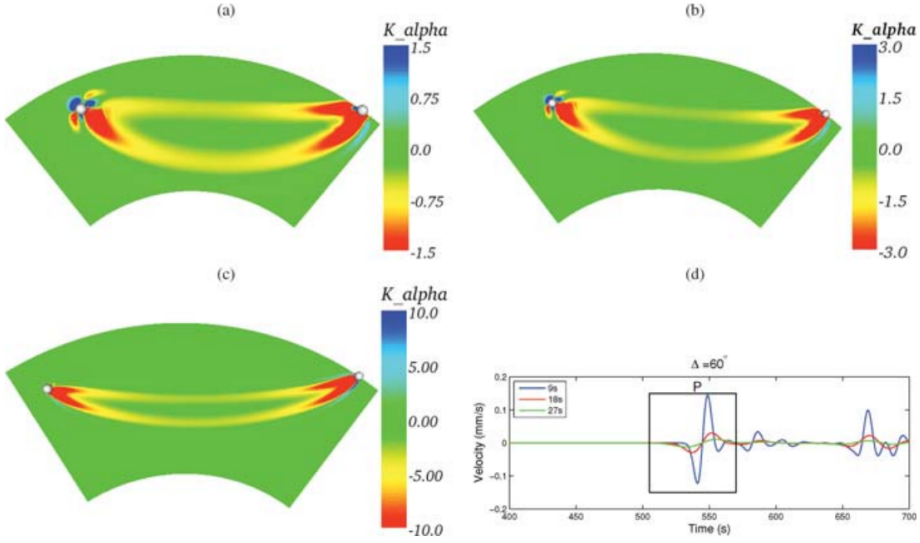

Fig. 3: Cross sections through the functional derivative of a cross-correlation delay time for a P-wave phase with respect to α as calculated by the adjoint method. Here we see that the delay time of the same phase can be made at different frequencies by filtering the data and synthetic waveforms. The dominant sensitivity lies within the first Fresnel zone, but it curiously vanishes along the ray path. Figure taken from Liu & Tromp (2008).

$$C(\tau) = \int_{-\infty}^{\infty} s^{\text{obs}}(t-\tau)s(\mathbf{x}_s, t)\,\mathrm{d}t. \tag{36}$$

This quantity will be largest when s^{obs} and s are maximally aligned, and hence the associated value of τ gives a measurement of the delay time between the observed and synthetic waveforms. Indeed, this is typically what is done in practical to automate delay time measurements. The condition for $\bar{\tau}$ to maximise the cross-correlation of the waveforms is given by

$$\frac{\mathrm{d}C}{\mathrm{d}\tau}(\bar{\tau}) = 0, \tag{37}$$

and this acts to define $\bar{\tau}$ implicitly as a function of the synthetic wavefield u_i, and hence of the underlying model parameters. The adjoint method can be applied to determine the functional derivatives of $\bar{\tau}$, and here we need only determine the appropriate adjoint source (details in the second problem set). The result is that we can write

$$\delta\bar{\tau} = \int_M \left(K_\alpha\,\delta\alpha + K_\beta\,\delta\beta\right)\,\mathrm{d}^3\mathbf{x}, \tag{38}$$

for the the first-order perturbation in this **cross-correlation delay time** with respect to the P- and S-wave speeds, holding other model parameters fixed. This expression can be compared to the ray theoretic result

$$\delta T = -\int_{\text{ray}} \frac{\delta c}{c^2}\,\mathrm{d}s, \tag{39}$$

obtained using Fermat's principle, with c the appropriate phase speed. An example of the resulting functional derivatives are shown in Fig.3 for a P-wave phase, with the delay time measured at a range of different frequencies.

The form of these functional derivatives – or **sensitivity kernels** as they are also known – is worth some comment. First, within ray theory all sensitivity lies along the reference ray path. Within the adjoint theory, however, there is a broad region of sensitivity around the ray path. This region can be shown to be the **first Fresnel zone** for the wave under consideration, i.e. the region in which a scattered wave arrives in phase with the main arrival. This is not too surprising, as ray theory ray neglects diffraction effects. What is perhaps more curious is that the adjoint kernels actually vanish along the reference ray path. When this was first observed from numerical calculations this result seemed counter-intuitive to some people: ray theory says that all the sensitivity is along the ray path, but the exact theory says none of it is! Subsequent work showed that there is no contradiction. The exact kernels vanish along the ray path because a scatterer on the ray path is associated with no phase delay relative to the reference wave. Moreover, it can be shown that as the dominant frequency of the wave increases, the width of the Fresnel zone shrinks, and in the limit as the frequency goes to infinity the two kernels agree in the sense of generalised functions, this meaning that their action on a smooth model perturbation produces the same delay time.

The final point to mention is that you might think that the kernels in Fig.3 look a little like a banana cut lengthwise, while in cross section they are hollow and can be said to resemble a doughnut. The first person to work on this theory, Tony Dahlen (1942-2007) from Princeton, certainly thought so, and as a result he jokingly named these functional derivatives **banana doughnut kernels**, and this remains the usual term used.

What you need to know and be able to do

(i) That waveform tomography depends on the ability to efficiently solve the elastic wave equation in realistic earth models.

(ii) How least-squares waveform misfits can be defined, and what is meant by their gradients. You might also be given other waveform misfits and asked to determine the corresponding adjoint traction.

(iii) Know in how gradient based optimisation works, why it is useful, and what its are limitations

(iv) How the adjoint method can be used calculate gradients, and what its advantages are. The derivation in this lecture too lengthy for you to reproduce in full. But you may be asked perform parts of the calculations with suitable guidance or to apply the method to simpler systems.

The Lagrange multiplier theorem – NON-EXAMINABLE

A sketch proof of the Lagrange multiplier theorem is provided for those interested. Let $\mathbf{u} \mapsto J(\mathbf{u})$ be a functional of the state vector, $\mathbf{u}$. Suppose that $\mathbf{u}$ is determined through solution of the following equation

$$\mathbf{a}(\mathbf{u}, \mathbf{m}) = \mathbf{0}, \tag{40}$$

with $\mathbf{m}$ the model vector. Here $\mathbf{a}$ is some potentially non-linear vector-valued function representing abstractly the partial differential equation satisfied by $\mathbf{u}$ and in which $\mathbf{m}$ act as parameters. We assume that for each $\mathbf{m}$ there is a unique solution to this equation, and hence it serves to define a mapping $\mathbf{m} \mapsto \hat{\mathbf{u}}(\mathbf{m})$ such that

$$\mathbf{a}[\hat{\mathbf{u}}(\mathbf{m}), \mathbf{m}] = \mathbf{0}. \tag{41}$$

Using this mapping, we can define a new functional

$$\hat{J}(\mathbf{m}) = J[\hat{\mathbf{u}}(\mathbf{m})]. \tag{42}$$

Within the main notes we do not distinguish notationally between $J(\mathbf{u})$ and $\hat{J}(\mathbf{m})$, but here it is useful to be more explicit. Our aim is to calculate the functional derivative, $D\hat{J}(\mathbf{m})$, this being the linear term within the first-order Taylor expansion

$$\hat{J}(\mathbf{m} + \delta\mathbf{m}) = \hat{J}(\mathbf{m}) + \langle\, D\hat{J}(\mathbf{m}) \,|\delta\mathbf{m}\,\rangle + O(\|\delta\mathbf{m}\|^2). \tag{43}$$

We will obtain this functional derivative first using a "brute-force" method via the chain rule, and then show that the Lagrange multiplier theorem leads to the same result. The advantage of the latter is the relative easy of its application in practice.

Using eq.(42), we can perturb $\mathbf{m}$ and expand to first-order to obtain

$$\hat{J}(\mathbf{m} + \delta\mathbf{m}) = J[\hat{\mathbf{u}}(\mathbf{m} + \delta\mathbf{m})] = J[\hat{\mathbf{u}}(\mathbf{m}) + \delta\mathbf{u} + \cdots] = \hat{J}(\mathbf{m}) + \langle\, DJ(\mathbf{u}) \,|\delta\mathbf{u}\,\rangle + \cdots, \tag{44}$$

where $\delta\mathbf{u}$ denotes the first-order perturbation to $\mathbf{u}$ due to the given change in $\mathbf{m}$. To determine the concrete form of $\delta\mathbf{u}$, we perturb eq.(41) to obtain

$$\mathbf{a}(\mathbf{u} + \delta\mathbf{u} + \cdots, \mathbf{m} + \delta\mathbf{m} + \cdots) = \mathbf{0}, \tag{45}$$

where it is understood that $\mathbf{u} = \hat{\mathbf{u}}(\mathbf{m})$. Expanding this out to first-order we find

$$D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})\delta\mathbf{u} + D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})\delta\mathbf{m} = \mathbf{m} = \mathbf{0}. \tag{46}$$

Here we see the partial derivatives of a with respect to $\mathbf{u}$ and $\mathbf{m}$, and note that these are linear operators that act on the perturbations to $\mathbf{u}$ and $\mathbf{m}$ to return an appropriate vector. If this notation is confusing, think of $\mathbf{u}$ as an n-dimensional vector and $\mathbf{m}$ as an m-dimensional one. Then the values of $\mathbf{a}(\mathbf{u}, \mathbf{m})$ are also n-dimensional vectors, and the linear operators $D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})$ and $D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})$ are matrices having dimensions $n \times n$ and

$n \times m$, respectively. Assuming that the linear operator, $D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})$ is invertible, we can solve the above equation to obtain

$$\delta\mathbf{u} = -[D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{-1} D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})\delta\mathbf{m}. \tag{47}$$

The assumption that this matrix is invertible is not actually an assumption at all. It can be shown via the implicit function theorem that this is a necessary (and locally sufficient) condition for eq.(41) to admit unique solutions. Substituting this expression for $\delta\mathbf{u}$ into eq.(48) we find

$$\hat{J}(\mathbf{m} + \delta\mathbf{m}) = \hat{J}(\mathbf{m}) - \langle\, DJ(\mathbf{u}) \,|[D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{-1} D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})\delta\mathbf{m} \,\rangle + \cdots . \tag{48}$$

Recalling the definition of the adjoint of a linear operator, we can transform the inner product term to obtain

$$\hat{J}(\mathbf{m} + \delta\mathbf{m}) = \hat{J}(\mathbf{m}) - \langle\, [D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{\dagger} [D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{-\dagger} DJ(\mathbf{u}) \,|\delta\mathbf{m} \,\rangle + \cdots , \tag{49}$$

where $(\cdot)^{-\dagger}$ is a shorthand for the inverse of the adjoint. By inspection, we have arrived at the desired result

$$D\hat{J}(\mathbf{m}) = -[D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{\dagger} [D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{-\dagger} DJ(\mathbf{u}). \tag{50}$$

We now apply the Lagrange multiplier theorem and show that it leads to an identical result. The starting point is defining the appropriate Lagrangian

$$L(\mathbf{u}, \mathbf{u}^{'}, \mathbf{m}) = J(\mathbf{u}) + \langle\, \mathbf{a}(\mathbf{u}, \mathbf{m}) \,|\mathbf{u}' \,\rangle, \tag{51}$$

with $\mathbf{u}'$ the Lagrange multiplier field used to impose eq.(41). Our claim is that

$$D\hat{J}(\mathbf{m}) = D_{\mathbf{m}}L(\mathbf{u}, \mathbf{u}', \mathbf{m}), \tag{52}$$

so long as the following conditions hold

$$D_{\mathbf{u}}L(\mathbf{u}, \mathbf{u}', \mathbf{m}) = \mathbf{0}, \quad D_{\mathbf{u}'}L(\mathbf{u}, \mathbf{u}', \mathbf{m}) = \mathbf{0}. \tag{53}$$

As in the main lecture, we readily see that the second condition implies eq.(41), and hence we have $\mathbf{u} = \hat{\mathbf{u}}(\mathbf{m})$. To examine the first condition, we note that

$$\langle\, D_{\mathbf{u}}L(\mathbf{u}, \mathbf{u}', \mathbf{m}) \,|\delta\mathbf{u} \,\rangle = \langle\, DJ(\mathbf{u}) \,|\delta\mathbf{u} \,\rangle + \langle\, D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})\delta\mathbf{u} \,|\mathbf{u}' \,\rangle = \mathbf{0}, \tag{54}$$

for any $\delta\mathbf{u}$. Again using the adjoint of a linear operator, this is equivalent to

$$\langle\, DJ(\mathbf{u}) + [D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{\dagger}\mathbf{u}' \,|\delta\mathbf{u} \,\rangle = \mathbf{0}, \tag{55}$$

and as $\delta\mathbf{u}$ is arbitrary, we see that the first condition in eq.(53) implies

$$DJ(\mathbf{u}) + [D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{\dagger}\mathbf{u}' = \mathbf{0}. \tag{56}$$

This is a linear equation for the Lagrange multiplier $\mathbf{u}'$, and it has solution

$$\mathbf{u}' = -[D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{-\dagger} DJ(\mathbf{u}). \tag{57}$$

This result can be substituted into eq.(52), but first we note that

$$\langle\, D_{\mathbf{m}}L(\mathbf{u}, \mathbf{u}', \mathbf{m}) \,|\delta\mathbf{m} \,\rangle = \langle\, D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})\delta\mathbf{m} \,|\mathbf{u}' \,\rangle = \langle\, \delta\mathbf{m} \,|[D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{\dagger}\mathbf{u}' \,\rangle, \tag{58}$$

for any $\delta\mathbf{m}$, and hence

$$D_{\mathbf{m}}L(\mathbf{u}, \mathbf{u}', \mathbf{m}) = [D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{\dagger}\mathbf{u}' = -[D_{\mathbf{m}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{\dagger} [D_{\mathbf{u}}\mathbf{a}(\mathbf{u}, \mathbf{m})]^{-\dagger} DJ(\mathbf{u}). \tag{59}$$

The final result is precisely what was obtained using the direct method, and hence we have established the validity of the Lagrange multiplier theorem. It is worth noting that eq.(56) for the Lagrange multiplier, $\mathbf{u}'$, involves the adjoint and the linearised forward operator. It is this fact that gives the adjoint method its name.