A tutorial on the adjoint-state method

Kris Innanen

ABSTRACT

A brief tutorial of the ideas behind the adjoint-state method for gradient calculations in full waveform inversion. In particular, the geometry of the problem is focused on, which allows us to distinguish between the use of Lagrange multipliers in this setting and those used for linear constrained optimization problems.

INTRODUCTION

If you are like me, you appreciated interpretable, visualizable ways of thinking when complex calculations are necessary. When discussin FWI, my own feeling over the years has always been that the adjoint-state method for computing the gradient lacked the "visualizability" of other approaches. Also, for those of us who were familiar with linear constrained optimization problems prior to learning the adjoint-state method, the apparently very different roles played by the Lagrange multipliers in those settings was a source of confusion. So, I present here a slightly scattershot tutorial – or perhaps better yet set of excursive remarks – on those two problems and their relations.

PRIMER: LINEAR CONSTRAINED OPTIMIZATION PROBLEMS

A standard underdetermined linear constrained optimization problem might involve solving for the model vector $\mathbf{m} \in \mathbb{R}^M$ which satisfies

$$\min_{\mathbf{m}} \phi, \ \ \text{subject to} \ \ G_i = 0, \ i = 1, ..., N, \tag{1}$$

where $N < M$,

$$\phi(\mathbf{m}) = \frac{1}{2}\big(\mathbf{m} - \mathbf{m}_0\big)^T \mathbf{W}\big(\mathbf{m} - \mathbf{m}_0\big), \tag{2}$$

and

$$G_i = d_i - \mathbf{g}_i^T \mathbf{m}. \tag{3}$$

Let us take this apart a little, using the $M = 3$, $N = 2$ case illustrated in Figure 1. The job is to select an element of model space, which here is the vector space $\mathbb{R}^{M=3}$. We make this selection by balancing two requirements.

First, we ask that the particular element we choose be that which minimizes the objective function ϕ. The objective function, which is set out in (2), is a scalar function of $\mathbf{m}$. It is quadratic in $\mathbf{m}$, which means it can be described with contours that are elliptical surfaces in $\mathbb{R}^{M=3}$ (the light blue shapes in Figure 1a). It comes equipped with a weighting matrix $\mathbf{W}$, which rotates and scales the axes of the ellipse, and it is translated in $\mathbb{R}^{M=3}$ such that it is centred around a non-zero reference point $\mathbf{m}_0$. This alone does not make for a compelling inverse problem, since ϕ is immediately minimized simply by selecting $\mathbf{m} = \mathbf{m}_0$. But, making it small is only one of the requirements.

Second, we also must ensure that the $\mathbf{m}$ we settle on satisfies N constraint equations, i.e., $G_i = d_i - \mathbf{g}_i^T \mathbf{m} = 0$. In this case, the constraints are that the data d_i must be satisfied, and, since, we are focusing on *linear* optimization, that these equations enforce a linear relationship between the d_i and the elements of $\mathbf{m}$. In our simple problem, we assume that we have two data, d_1 and d_2, and those data can be computed via a linear combination of the elements of $\mathbf{m} = [m_1, m_2, m_3]^T$, involving a known sets of weights $\mathbf{g}_1 = [g_{11}, g_{12}, g_{13}]^T$ and $\mathbf{g}_2 = [g_{21}, g_{22}, g_{23}]^T$:

$$d_1 = g_{11}m_1 + g_{12}m_2 + g_{13}m_3, \; d_2 = g_{21}m_1 + g_{22}m_2 + g_{23}m_3. \tag{4}$$

You can quickly confirm that asking that these hold is the same as requiring that $G_1 = 0$ and $G_2 = 0$. In the same way that we can visualize ϕ geometrically, we can visualize the constraints. A linear equation on $\mathbb{R}^M$ defines a subspace of $\mathbb{R}^M$ with a dimension one less than M, i.e., $\mathbb{R}^{M-1}$. So, if $M = 3$, $\mathbb{R}^{M=3}$ is the volume of all space, and the subspace defined by a linear equation is a plane, which is a $\mathbb{R}^2$ subspace of $\mathbb{R}^3$. Similarly, if we had been working on a problem in which the plane was our full space, i.e., $\mathbb{R}^{M=2}$, a single linear equation would have instead defined a line, which is a $\mathbb{R}^{2-1} = \mathbb{R}^1$ subspace of the plane. And so forth. So in our case, each of the two equations in (4) defines a plane in $\mathbb{R}^3$. These are illustrated as red and green polygons in Figure 1a. Since both constraint equations are in force, meaning the model we choose must lie on both planes, we are actually restricting ourselves to models which lie on the line at which the two planes intersect (AB in Figure 1a).

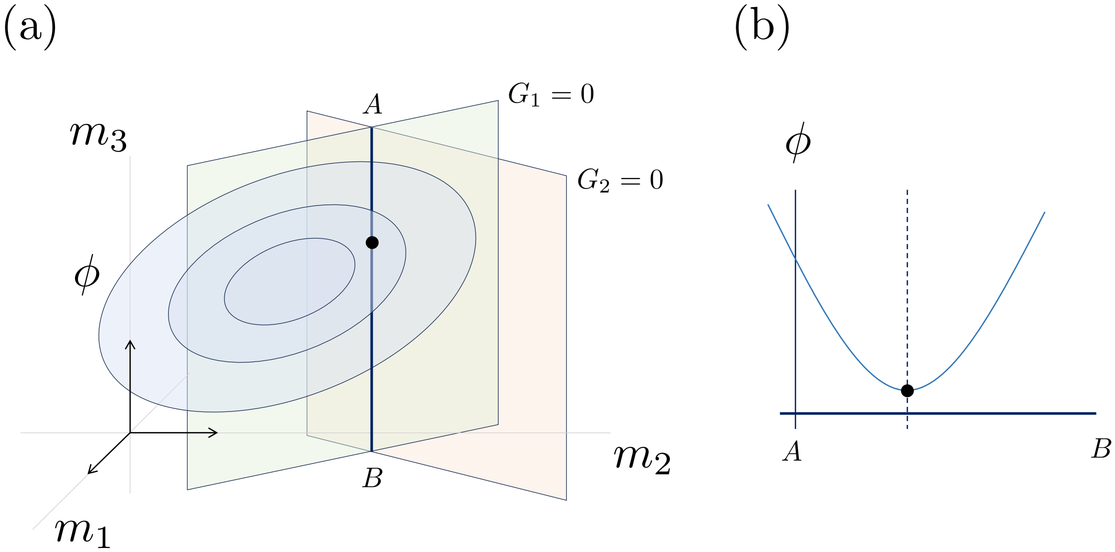

FIG. 1. The geometry of constrained optimization.

So, all told, our job is to explore the sub-region of model space allowed by the two constraints (in this case, this means moving up and down along the line AB), and to find, on that sub-region, the model $\mathbf{m}$ which minimizes ϕ. This is in general a lower dimensional problem (in Figure 1b, by extracting the values of ϕ lying along the line AB, we see it in fact to be a 1D problem), and it seems easier. But, how do we actually solve it? In unconstrained linear inversion, jumping to a minimum is a relatively simple calculation of the point at which $\frac{\partial \phi}{\partial \mathbf{m}} = 0$. How to we jump to a minimum while satisfying these constraints?

The method of Lagrange multipliers gives us an approach which is elegant, transforming the constrained optimization problem into an equivalent unconstrained optimization problem (which is then easy to solve). Furthermore, it allows us to continue to make geometric pictures of the steps we take (as realized in low-dimensional versions of the problems, of course). We need to establish a few more facts to proceed.

First, we need to start thinking not so much about the constraints and their geometric interpretation (i.e., planes in the $\mathbb{R}^{M=3}$ case), but about the normals to these planes. If G_i gives us a plane, then the normal to that plane is

$$\mathbf{n}_i = \frac{\partial G_i}{\partial \mathbf{m}} = \mathbf{g}_i, \tag{5}$$

where on the right-hand side we have recognized that in taking derivatives we have recovered the weights in the data-model mapping (we will continue to write these in the left-hand form, since we are heading towards a problem where this feature of linear problems will not hold. If you need convincing of the fact that these are normals, try the simple case $G_1 = m_3 - 2 = 0$, which represents a horizontal plane parallel to the m_1-m_2 coordinate plane, raised to a height of 2. The normal to this, according to the rule in (5), after carrying out the trivial differentiation, is $\mathbf{n}_1 = [0, 0, 1]^T$, which of course is the expected normal vector, pointing in the m_3 direction.

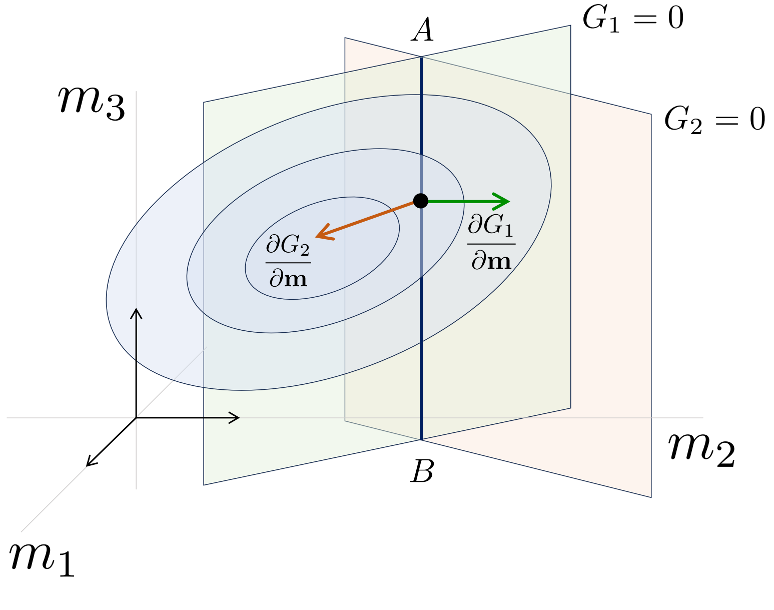
A
$G_1 = 0$
$G_2 = 0$
m_3
$\frac{\partial G_2}{\partial \mathbf{m}}$
$\frac{\partial G_1}{\partial \mathbf{m}}$
B
m_2
m_1

FIG. 2. Normals to the constraint equations.

Each of the constraint equations contributes a normal vector (see Figure 2). If the problem is well-posed, the planes are not parallel to one another, which in turn means these normals are independent. Therefore, the normals span an N-dimensional subspace of $\mathbb{R}^M$.

In the $M = 3$, $N = 2$ case we have been working with, we have two independent constraint normals, as illustrated in Figure 2, so the normal vectors span a $\mathbb{R}^2$ subspace, or plane, in $\mathbb{R}^{M=3}$. Any vector that lies in this plane can be constructed through a linear combination of the two normals:

$$\alpha \frac{\partial G_1}{\partial \mathbf{m}} + \beta \frac{\partial G_2}{\partial \mathbf{m}}. \tag{6}$$

This is relevant to our current problem because of a fact about the gradient of the original objective function ϕ, i.e., the quantity

$$\frac{\partial \phi}{\partial \mathbf{m}}, \tag{7}$$

which lives in the same space $\mathbb{R}^M$.

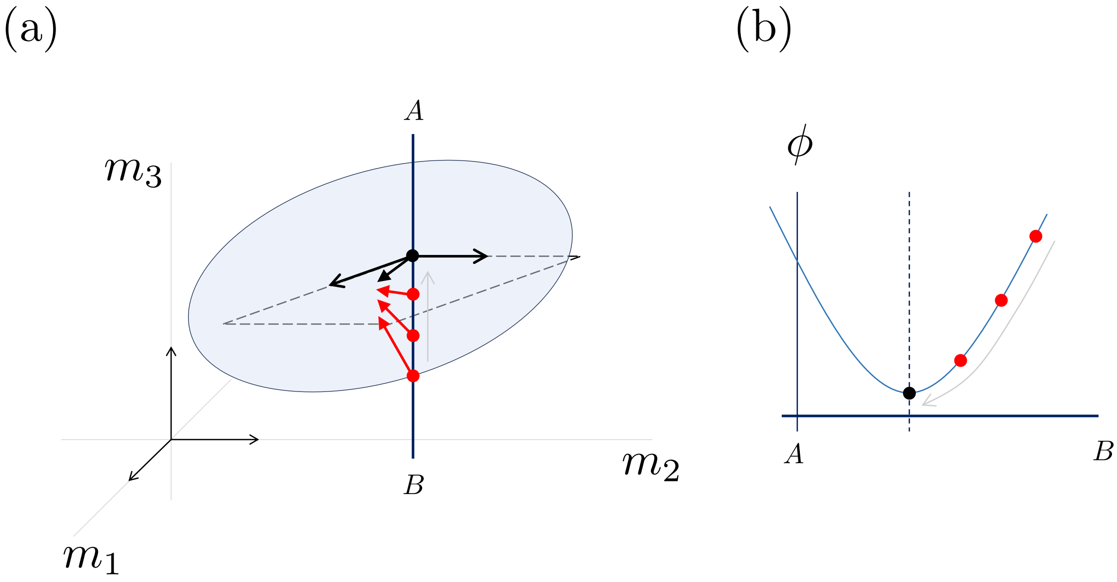

FIG. 3. A special condition in place at the constrained minimum.

To develop this fact, let us strip away much of the detail from our illustrative case, leaving only one elliptical surface of ϕ for reference, and the two normal vectors as well as a sketch of the plane they span (see Figure 3a). Now, let us travel along the line AB, i.e., the subregion of model space we have constrained ourselves to lie on, from the direction of B upward, following the red dots in the figure. At each intervening position, plot the gradient $\partial\phi/\partial\mathbf{m}$ (red arrows). We observe that, as we move upwards towards the plane spanned by the normal vectors, the gradient vector becomes increasingly parallel with this plane. Eventually, we hit a point along the line AB at which the gradient sits precisely in the plane. If we kept going (we don't in the figure, which would become messy otherwise), the gradient would again tilt out of the plane. Is there anything special about this point at which $\partial\phi/\partial\mathbf{m}$ sits precisely in the plane formed by $\partial G_1/\partial\mathbf{m}$ and $\partial G_2\ \partial\mathbf{m}$? There is indeed. If we repeat the exercise, but this time track the red dotted points on the extracted ϕ profile (see Figure 3b), a rather striking fact is revealed: *The gradient vector $\partial\phi/\partial\mathbf{m}$ lands in the plane formed by $(\partial G_1/\partial\mathbf{m})$ and $(\partial G_2\ \partial\mathbf{m})$ at exactly the point where ϕ is minimized along AB.*

Of course, this is the very point we are looking for. This fact can be proved mathematically, and is completely general as far as the dimension of the problem goes. You can find the proof in the text of Bob Parker, and no doubt in many other places. However, it is not particularly enlightening, in the sense of adding more insight (Parker, for instance, finds it necessary to prove it by contradiction, rather than by developing ideas that we can easily visualize). Because this is a tutorial about visualization, we will not reproduce it here, but we can attempt to make it more palatable by lowering the dimensions of the problem even further, such that we are working with the plane $\mathbb{R}^{M=2}$, and one equation of constraint, which forces us to seek a minimum along a line. See Figure 4. Here, the gradient $\partial\phi/\partial\mathbf{m}$ is easy to sketch, being perpendicular to the contours of ϕ in the plane, which we do at various points along the line, in red. There is only one constraint, and only one normal, which we plot at the same points in black. To say that a vector lies in the space spanned by the normals, in this almost trivial case, is just to say that it is parallel to the one normal vector. If we track the vector pairs, we see quickly that the point at which the two are parallel is naturally going to fall on the point at which the constraint line grazes the lowest-valued contour it is going to hit, which is, by definition, the minimum being sought.

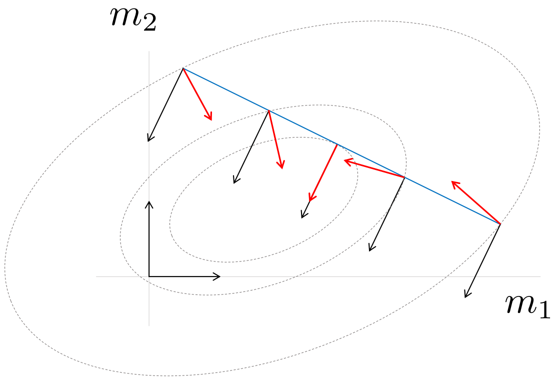

FIG. 4. Gradient direction versus normal to the constraint and their relations.

To continue, what we have discovered is the fact that, at the constrained minimum we are seeking, and nowhere else, we can write

$$\frac{\partial\phi}{\partial\mathbf{m}} = \lambda_1\frac{\partial G_1}{\partial\mathbf{m}} + \lambda_2\frac{\partial G_2}{\partial\mathbf{m}}. \tag{8}$$

The "stroke of genius" we associate with Lagrange is to package this fact into a new problem, which takes the form of a larger, but completely unconstrained, optimization. We form a new objective function $\mathcal{L}$:

$$\mathcal{L}(\mathbf{m},\boldsymbol{\lambda}) = \phi(\mathbf{m}) - \sum_{i=1}^{N}\lambda_i G_i = \phi(\mathbf{m}) - \boldsymbol{\lambda}^T\mathbf{G} = \phi(\mathbf{m}) - \langle\boldsymbol{\lambda},\mathbf{G}\rangle. \tag{9}$$

We have written it three times. The leftmost version directly uses the terms we have been developing so far: the original objective function ϕ is present, and a sum has been added in, involving the constraint equations and the coefficients λ_i from the assemblage in (8). The next two forms are just conveniences. The sum over i implies an inner product, provided the λ_i and the G_i are arranged into vectors, which is what we do in the middle version.

The rightmost version is here so that later, when we discuss the adjoint-state solution, the terminology one runs into in papers is more familiar. The use of angle-brackets to denote this inner product acts as an important reminder that this is *not* an inner product for vectors living in $\mathbb{R}^M$, but rather for vectors living in $\mathbb{R}^N$. As we shall see, in the adjoint state calculation, keeping these issues straight in our minds is more difficult, and more important.

Staying with the first version, let us now treat this objective function with techniques we would employ in any unconstrained optimization problem: let us take the gradient of $\mathcal{L}$, and seek the particular arrangement of unknowns for which this gradient is zero. The only extra thing to remember is that, in this expanded problem, there are more than the original number of unknowns: we have the elements of the vector $\mathbf{m}$, but now we also have the unknown λ_i, for $i = 1, ..., N$. So, we set

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}} = \frac{\partial \phi}{\partial \mathbf{m}} - \sum_{i=1}^{N} \lambda_i \frac{\partial G_i}{\partial \mathbf{m}} = 0, \tag{10}$$

and

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = G_i = 0, \;\; i = 1, ..., N. \tag{11}$$

What we want to establish is that by doing this we get what we want from the method; in a moment we will see that the approach also "spits out" an algorithm or workflow as well. First, take (10). Let us re-write it and set $M = 3$, $N = 2$ to recover our test case. Setting $\partial\mathcal{L}/\partial\mathbf{m} = 0$ evidently implies

$$\frac{\partial \phi}{\partial \mathbf{m}} = \lambda_1 \frac{\partial G_1}{\partial \mathbf{m}} + \lambda_2 \frac{\partial G_2}{\partial \mathbf{m}}, \tag{12}$$

meaning, comparing this to (8), that at the minimum of this new objective function $\mathcal{L}$ with respect to the $\mathbf{m}$ elements, we have found the minimum of ϕ along the constraint line[*]. Next, take (11). Actually there is no more to say here: at the minimum of $\mathcal{L}$ with respect to the λ_i, we correctly recover the constraints, i.e., the statements that $G_i = 0$.

Fine (I can hear you saying) but this seems to be nothing more than an interesting statement. What do we actually do here in order to solve for the $\mathbf{m}$ in question? Actually, for linear problems, the ballgame is already over. It is just a little hard to see it with the mathematics in its current form. Take (12), and substitute in (2), recalling rules for vector differentiation:

$$\frac{\partial \phi}{\partial \mathbf{m}} = \mathbf{W}\mathbf{m} = \sum_{i=1}^{N} \lambda_i \mathbf{g}_i. \tag{13}$$

This can be solved for the minimizer $\mathbf{m}^*$, if we have access to the λ_i:

$$\mathbf{m}^* = \sum_{i=1}^{N} \lambda_i \big(\mathbf{W}^{-1}\mathbf{g}_i\big). \tag{14}$$

[*] Another way to say it, rather than saying we "found" something, might be to say that in these equations we are making statements that *can only be true at the minimum* of ϕ along the constraint line.

The λ_i are, as their name implies, so far undetermined, but now that we have this form, note that we can return to our linear data-model relations, and substitute in this special case of the model, i.e., $\mathbf{g}_k^T\mathbf{m}^* = d_k$, whereby:

$$\mathbf{g}_k^T\left[\sum_{i=1}^{N}\lambda_i\big(\mathbf{W}^{-1}\mathbf{g}_i\big)\right] = d_k \rightarrow \sum_{i=1}^{N}\big(\mathbf{g}_k^T\mathbf{W}^{-1}\mathbf{g}_i\big)\lambda_i = d_k. \tag{15}$$

This can be compactly expressed as a matrix equation $\boldsymbol{\Gamma\lambda} = \mathbf{d}$ in $\mathbb{R}^N$, where $\boldsymbol{\Gamma}$ is an $M \times N$ [CHECK THIS!!!] matrix with elements

$$(\boldsymbol{\Gamma})_{ki} = \mathbf{g}_k^T\mathbf{W}^{-1}\mathbf{g}_i, \tag{16}$$

and solved for:

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}^{-1}\mathbf{d}. \tag{17}$$

The workflow is then to use these results in reverse order: the Lagrange multipliers λ_i are first solved for via (17), after which they can be used in (14) to produce $\mathbf{m}^*$. In the context of our $M = 3$, $N = 2$ problem, this workflow is as follows. First, we construct $\boldsymbol{\Gamma}$

FULL WAVEFORM INVERSION

FWI quantities

Let us first set up the elements of the FWI problem, after which we can develop it as a nonlinear constrained optimization problem which is amenable to an adjoint-state approach. We introduce vectors into the problem which are elements of three different vector spaces, which we will label $\mathbb{R}^M$, $\mathbb{R}^U$, and $\mathbb{R}^D$ to remind ourselves that their dimensions are in general different.[†] The vectors are

$$\begin{aligned}&\mathbf{m} \in \mathbb{R}^M, \ \text{“model space”},\\ &\mathbf{u} \in \mathbb{R}^U, \ \text{“wavefield space”},\\ &\mathbf{d} \in \mathbb{R}^D, \ \text{“data space”}.\end{aligned} \tag{18}$$

The interrelations between these vectors are as follows. First, the wavefield vector $\mathbf{u}$ satisfies the wave equation, which in the frequency domain is the matrix equation

$$\mathbf{S}(\mathbf{m},\omega)\mathbf{u} = \mathbf{f}(\omega, X), \tag{19}$$

where $\mathbf{S}$ is the $U \times U$ impedance matrix, $\mathbf{f} = \mathbf{f}(\omega, X)$ is the source vector, and ω is the temporal frequency. We have made $\mathbf{f}$ a function of the frequency, which it typically is, and we have also temporarily given it a dependence on X, which stands for all of the

[†] Of course, in frequency-domain FWI, which is what we are setting out, several of the spaces should actually be labelled $\mathbb{C}^U$ (etc.), since the vectors are complex-valued. However, since we are illustrating these methods with small examples that allow pictures to be sketched, we will pretend that the spaces are real. This does not introduce any incorrect insights or results.

differences the source vector undergoes when we adjust the location or character of the seismic source. Elements of the model vector $\mathbf{m}$, which in FWI is the vector of unknown medium properties, appear in various combinations in the elements of $\mathbf{S}$, hence we write $\mathbf{S} = \mathbf{S}(\mathbf{m}, \omega)$. Since solving the wave equation amounts to determining $\mathbf{u} = \mathbf{S}^{-1}\mathbf{f}$, and $\mathbf{S}$ is a function of $\mathbf{m}$, so to is $\mathbf{u}$ implicitly a function of $\mathbf{m}$. However, $\mathbf{u}$ also is treated as an independent variable, and functions of $\mathbf{u}$ for all values of $\mathbf{u}$, not just ones which satisfy (19), are regularly considered. One of the tricks in understanding FWI is getting used to this way of looking at $\mathbf{u}$. We need to be able to see it both ways.

The forward problem

Solving the *forward problem* involves:

1. Choosing a model $\mathbf{m}$ and a source term $\mathbf{f}$;
2. Embedding elements of $\mathbf{m}$ into the impedance matrix $\mathbf{S}$;
3. Solving for the associated $\mathbf{u}$ via $\mathbf{S}^{-1}\mathbf{f}$;
4. Extracting elements of the wavefield to be compared to measured seismic data.

The extraction process is realized through a sampling matrix $\mathbf{R}$:

$$\mathbf{d} = \mathbf{R}\mathbf{u}. \tag{20}$$

So, the model $\mathbf{m}$ determines in some generally complicated way the wavefield, some of whose values are extracted to form the data vector; inverting this process so that we can infer $\mathbf{m}$ via $\mathbf{d}$ is of course the goal.

Two low-dimensional illustrative examples

Let us again produce small, explicit, examples to sit alongside the more general developments, so that we can sketch the quantities and keep the geometry of the situation in the forefront of our minds. The first will be the simplest, with all spaces of dimension 2. This will allow very clear plots and illustrations to be set up; it will also mean, however, that there will be no significant difference between the data and the wavefield. To make sure we see some of those differences appear in the analysis, we will move, in the second case, to dimension 3 for both model and wavefield spaces, maintaining a dimension 2 data space.

Case 1: $M = 2$, $U = 2$, $D = 2$

Here we set $M = U = D = 2$, in which case model vectors have the form

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \tag{21}$$

wavefield vectors are likewise

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \tag{22}$$

and $\mathbf{R} = \mathbf{I}$, i.e., $[d_1, d_2]^T = [u_1, u_2]^T$. The system $\mathbf{Su} = \mathbf{f}$ is

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}. \tag{23}$$

In this case, which is the smallest one we will study, it will be useful to form some plots of these quantities "in action". For this, it is helpful if we force a relationship between the elements s_{ij} and the model elements m_i, in particular one in which the nonlinearity arising in FWI is present. Because there is no way to form a meaningful wave problem on a 2×2 template, we will have to "fake" the form for **S**, but we will find that even a fake form is insightful. We choose:

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} 1 + am_1^2 & bm_1m_2 \\ bm_2m_1 & 1 + cm_2^2 \end{bmatrix}, \tag{24}$$

letting a, b, and c be constants that affect the degree and kind of nonlinearity we examine. To make concrete examples, let us in fact choose values for the constants and stick with them: $a = 3$, $b = 0.5$, and $c = 2$[‡]. To solve this system we also need to supply a source vector $\mathbf{f}$, so again, in order to be able to make concrete examples, let us also choose a specific set of numbers to use: $\mathbf{f} = [f_1, f_2]^T = [0.1, 0.2]^T$.

Case 2: $M = 3$, $U = 3$, $D = 2$

Let us set $M = 3$ again, in which case our model vectors occupy something similar to physical space, and have 3 elements:

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}. \tag{25}$$

If $M = 3$, in an FWI problem (this is ridiculous of course) the implication is that there are 3 grid cells to be characterized, and m_1, m_2, and m_3 are the seismic velocities of those grid cells. If so, it is reasonable to assume that $U = 3$ also, since the general problem in (19) is to solve for the wavefield at each grid cell. If so, it would not be unreasonable to assign to $\mathbf{u}$ 3 elements also:

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}. \tag{26}$$

[‡] It should be emphasized that these are not constants of the medium, or unknowns we determine from the inversion or optimization problem. These are true constants of the operator **S**, playing a similar role to the 1 and 2 weights we would tend to find in a finite difference stencil containing second spatial derivatives.

On the other hand, we would also likely set $D < U$, reflecting the idea that the data are values of the wavefield, sensed over some incomplete set of all possible grid cells. In fact, in reality D will be considerably less[§] than U. However, in order that the data space is large enough to contain somewhat interesting elements, we will set $D = 2$ in this case. The sampling matrix $\mathbf{R}$ might then be

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \tag{27}$$

in which case $\mathbf{d} = \mathbf{R}\mathbf{u}$ is

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{28}$$

$\mathbf{R}$ has no inverse, but we shall need to nonetheless use its transpose to map quantities in $\mathbb{R}^D$ into $\mathbb{R}^U$. This produces zeros in elements of the wavefield vector *not* associated with a sensor:

$$\mathbf{R}^T\mathbf{d} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ 0 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ 0 \end{bmatrix}. \tag{29}$$

The system $\mathbf{S}\mathbf{u} = \mathbf{f}$ in this case is

$$\begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}. \tag{30}$$

THE ADJOINT-STATE METHOD

FWI as a nonlinear constrained optimization problem

FWI is a *nonlinear constrained optimization* problem. As in the linear case, it is based on a quadratic objective function ϕ, but this time, we frame it in terms of the difference between data we observe, $\mathbf{d}^o$, and data we have predicted through the forward problem, $\mathbf{d}^p$:

$$\phi' = \frac{1}{2}\sum_{\omega}\sum_{X} ||\mathbf{d}^p(X,\omega) - \mathbf{d}^o(X,\omega)||_2^2. \tag{31}$$

The sums reflect the fact that we are seeking a solution which minimizes the squared difference of the predicted and the observed data for a dataset which includes both many different source locations/characteristics, and frequencies. The result of including these two sums is that they appear at the beginning of all subsequent formulas. While they certainly belong in those formulas, for an exposition, their main effect is to clutter up the expressions while

[§] For instance, on a square $N \times N$ grid, U would tend to be close to N^2, whereas, if we imagined arraying sensors along the top edge of the grid, we would have $D \approx N$.

not adding much insight. So, we will remove them, which amounts to pretending that we are inverting with a single frequency and single source; returning them requires change in the derivations. This leaves us with

$$\phi = \frac{1}{2}||\mathbf{d}^p - \mathbf{d}^o||_2^2 = \frac{1}{2}||\mathbf{R}\mathbf{u} - \mathbf{d}^o||_2^2, \tag{32}$$

or

$$\phi = \frac{1}{2}\big(\mathbf{R}\mathbf{u} - \mathbf{d}^o\big)^T\big(\mathbf{R}\mathbf{u} - \mathbf{d}^o\big). \tag{33}$$

We can make things even more compact with two additional adjustments. First, since we will from now on express the predicted data $\mathbf{d}^p$ as $\mathbf{R}\mathbf{u}$, the superscripts have become irrelevant, so let us now let $\mathbf{d}$ refer to the observed data. Second, if we expand the form in (33), the rightmost term goes as $\mathbf{d}^T\mathbf{d}$, which is a constant with respect to the model and the wavefield, and thus has no impact on the relative size of ϕ. So, again to maximize simplicity, we will omit it. With these two changes in place, we have the objective function in a convenient form:

$$\phi = \frac{1}{2}\mathbf{u}^T\mathbf{R}^T\mathbf{R}\mathbf{u} - \mathbf{u}^T\mathbf{R}^T\mathbf{d}. \tag{34}$$

There are a number of ways we could proceed from here to set up FWI. The approach taking us to the adjoint state method is to solve this by minimizing ϕ, subject to the additional constraint that $\mathbf{u}$ must satisfy (19). Expressed in full, we seek:

$$\min_{\mathbf{m}} \phi, \ \text{ subject to } \mathbf{S}\mathbf{u} = \mathbf{f}. \tag{35}$$

It is instructive to compare this constrained optimization problem to the linear problem we set out in the primer (equation 1). On the surface, they appear very similar: we are minimizing a quadratic functional ϕ, subject to a set of linear constraint equations $S_i = 0$. In our small examples, for instance, we have for $\mathbf{S}\mathbf{u} = \mathbf{f}$ in Case 1

$$\begin{aligned} S_1 &= s_{11}u_1 + s_{12}u_2 - f_1 = 0, \\ S_2 &= s_{21}u_1 + s_{22}u_2 - f_1 = 0, \end{aligned} \tag{36}$$

and for Case 2

$$\begin{aligned} S_1 &= s_{11}u_1 + s_{12}u_2 + s_{13}u_3 - f_1 = 0, \\ S_2 &= s_{21}u_1 + s_{22}u_2 + s_{23}u_3 - f_1 = 0, \\ S_3 &= s_{31}u_1 + s_{32}u_2 + s_{33}u_3 - f_1 = 0, \end{aligned} \tag{37}$$

i.e., a set of equations that describe lines, planes (or hyperplanes) with normals and a fulsome geometric character.

However, on closer inspection, some issues appear. Although ϕ is quadratic, it is not quadratic in $\mathbf{m}$, the unknown model vector we are interested in determining. It is quadratic in $\mathbf{u}$. This is where it becomes important to distinguish between (i) $\mathbf{u}$ as an independent variable, i.e., one which varies over all possible vectors in $\mathbb{R}^U$, and (ii) $\mathbf{u}$ as the solution of $\mathbf{S}(\mathbf{m})\mathbf{u} = \mathbf{f}$, in which case there is only one unique $\mathbf{u}$ for any $\mathbf{m}$, so $\mathbf{u}$ is not an independent variable. In (i), $\phi = \phi(\mathbf{u})$ is quadratic, as a function of $\mathbf{u}$, but very few actual vectors $\mathbf{u}$ entering this construction are useful, since we are only interested in $\mathbf{u}$ vectors which are possible wavefields. To fix this, we will on occasion need to discuss $\phi = \phi(\mathbf{u})$ evaluated at $\mathbf{u} = \mathbf{u}'$, where $\mathbf{u}'$ *does* satisfy $\mathbf{S}\mathbf{u}' = \mathbf{f}$. In contrast, in (ii), $\phi = \phi(\mathbf{u}) = \phi(\mathbf{u}(\mathbf{m}))$ is not quadratic, as a function of $\mathbf{m}$, but every $\mathbf{u}$ vector arising in this problem is a legitimate solution of the wave equation.

Ultimately, the fact that ϕ is determined through the complex, nonlinear relationship $\mathbf{u}$ has with $\mathbf{m}$, means that we will not be able to create a simple set of normal equations for the minimum, as we did in equations (14) and (17). Instead, we will have to settle for determining the gradient of ϕ with respect to $\mathbf{m}$, and using this gradient to drive iterative updating towards the minimum. It also means that the geometrical / pictorial insight we built up in the linear case will not transfer directly to the FWI problem. Does any remain? Fortunately, yes – but we need to do quite a bit more work to find it.

The geometrical constructions underlying the adjoint state method

So, our goal has changed: we are now bent on determining $\partial\phi/\partial\mathbf{m}$, as opposed to finding normal equations for the minimizer $\mathbf{m}^*$. The adjoint state method leads to an algorithm to this end. Finding the equations that underlie the algorithm is quick. However, to learn to think about the adjoint state geometrically, we will come at it slowly, through a process of "accidental discovery" (or inspired guessing, or something).

Let us start by focusing at first entirely on the character of ϕ and the constraints in the space of $\mathbf{u}$ vectors, where we think of $\mathbf{u}$ as an independent variable that ranges over all of the space $\mathbb{R}^U$. It is

$$\phi(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathbf{R}^T\mathbf{R}\mathbf{u} - \mathbf{u}^T\mathbf{R}^T\mathbf{d}. \tag{38}$$

To have a concrete example to look at, let us take Case 1, and assume that the data we have measured are $\mathbf{d} = [0.5, 0.5]^T$, for the moment not worrying about the model vector that led to these data. In Case 1, remembering that $\mathbf{R} = \mathbf{I}$, ϕ has the form

$$\phi(\mathbf{u}) = \frac{1}{2}[u_1, u_2]\left[\begin{array}{c} u_1 \\ u_2 \end{array}\right] - [u_1, u_2]\left[\begin{array}{c} d_1 \\ d_2 \end{array}\right]. \tag{39}$$

In the $\mathbf{u}$ plane ϕ is very simple, essentially a circularly-symmetric scalar function, centred around the minimum $\mathbf{d} = [0.5, 0.5]^T$, which can be plotted as a set of contours (Figure 5a). Let us now start allowing elements of model space to produce objects in this space. We select a point in $\mathbb{R}^M$, and allow it to determine via forward modelling its partner point in $\mathbb{R}^U$. The point is not special in either domain; we could consider it to be some intermediate iterate in model space we find ourselves at, or some initial model we have selected. Again to be completely concrete, let this point be $\mathbf{m} = [m_1, m_2]^T = [-0.56, 0.7]^T$. This is entered into our forward modeling system $\mathbf{d} = \mathbf{I}\mathbf{u} = \mathbf{S}^{-1}\mathbf{f}$, from which we obtain a wavefield (and, in Case 1, data) vector:

$$\left[\begin{array}{c} d_1 \\ d_2 \end{array}\right] = \left[\begin{array}{cc} 1 + am_1^2 & bm_1m_2 \\ bm_2m_1 & 1 + cm_2^2 \end{array}\right]^{-1}\left[\begin{array}{c} f_1 \\ f_2 \end{array}\right] = \left[\begin{array}{c} 0.06 \\ 0.11 \end{array}\right], \tag{40}$$

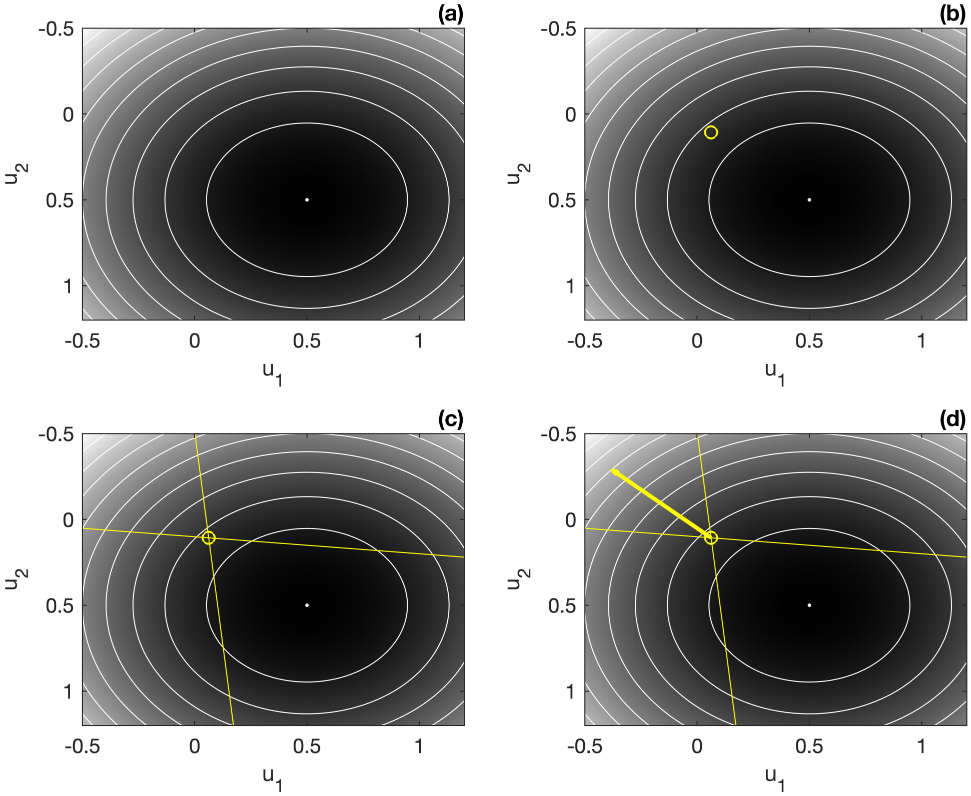

FIG. 5. Geometrical constructions within a low-dimensional adjoint-state example.

using the numerical values of a, b, c, f_1, and f_2 we set out earlier. We plot this as a yellow circle in Figure 5b. To start building up our geometrical interpretation, recall that we could also have broken the system solved in (40) up into a set of constraint equations, and examined them one at a time. In other words we could have considered

$$\begin{aligned} S_1 &= s_{11}u_1 + s_{12}u_2 - f_1 = \left(1 + 3m_1^2\right) u_1 + \left(\frac{1}{2}m_1 m_2\right) u_2 - 0.1 = 0, \\ S_2 &= s_{21}u_1 + s_{22}u_2 - f_2 = \left(\frac{1}{2}m_1 m_2\right) u_1 + \left(1 + 2m_2^2\right) u_2 - 0.2 = 0, \end{aligned} \tag{41}$$

individually. Each of these two equations defines a line, and the two lines intersect at the solution (see Figure 5c). We would like to use these lines to help describe the gradient of ϕ in $\mathbb{R}^U$, which we can obtain by directly differentiating (38):

$$\frac{\partial \phi}{\partial \mathbf{u}} = \frac{\partial}{\partial \mathbf{u}} \left(\frac{1}{2}\mathbf{u}^T\mathbf{u} - \mathbf{u}^T\mathbf{R}^T\mathbf{d} \right) = \mathbf{u} - \mathbf{d}, \tag{42}$$

which, in Case 1, is

$$\begin{bmatrix} \partial\phi/\partial u_1 \\ \partial\phi/\partial u_2 \end{bmatrix} = \begin{bmatrix} u_1 - d_1 \\ u_2 - d_2 \end{bmatrix}. \tag{43}$$

This result makes sense given the symmetry of this simple objective function – the gradient is a vector pointing directly away from the minimum $\mathbf{d} = [d_1, d_2]^T$. See Figure 5d.

Constraint equations describe lines (as in the case in equations 41), but also planes, or hyperplanes, depending on the dimensionality of the problem, so we are not ready to use them yet to help us characterize the gradient vector in (43). Let us address this now. In Figure 6a, we return to the plot of $\phi(\mathbf{u})$, with the special point mapped from model space, the two constraint lines, and the gradient $\partial\phi/\partial\mathbf{u}$. Let us now add to this the gradients of the constraints (which, you may recall, played an important role in the method of Lagrange multipliers in the case of linear optimization problems), which are vectors normal to the constraint lines. In Figure 6b we keep the objects from the first panel (but in a dimmer colour), and add in two dashed lines in yellow, passing through the special point with directions given by the gradients of the constraints:

$$\begin{aligned}\frac{\partial S_1}{\partial\mathbf{u}} &= \begin{bmatrix} {}^{\partial S_1}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{bmatrix} = \begin{bmatrix} s_{11} \\ s_{12} \end{bmatrix} = \begin{bmatrix} 1+3m_1^2 \\ ({}^1\!/\!_2)m_1m_2 \end{bmatrix}, \\ \frac{\partial S_2}{\partial\mathbf{u}} &= \begin{bmatrix} {}^{\partial S_2}\!/\!_{\partial u_1} \\ {}^{\partial S_2}\!/\!_{\partial u_2} \end{bmatrix} = \begin{bmatrix} s_{21} \\ s_{22} \end{bmatrix} = \begin{bmatrix} ({}^1\!/\!_2)m_1m_2 \\ 1+2m_2^2 \end{bmatrix}.\end{aligned} \tag{44}$$

Consider linear combinations of the two vectors ${}^{\partial S_1}\!/\!_{\partial\mathbf{u}}$ and ${}^{\partial S_2}\!/\!_{\partial\mathbf{u}}$ that produced them:

$$\alpha\frac{\partial S_1}{\partial\mathbf{u}} + \beta\frac{\partial S_2}{\partial\mathbf{u}} = \alpha\begin{bmatrix} {}^{\partial S_1}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{bmatrix} + \beta\begin{bmatrix} {}^{\partial S_2}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{bmatrix}. \tag{45}$$

Because these are vectors in $\mathbb{R}^U$, they can be substantively compared to the ϕ gradient vector in equation (43). In Figure 6c, we select (essentially at random) $\alpha = -0.1$ and $\beta = -0.2$, and plot three vectors, all in bold yellow, the first two being

$$\begin{aligned}\alpha\frac{\partial S_1}{\partial\mathbf{u}} &= \alpha\begin{bmatrix} {}^{\partial S_1}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{bmatrix}, \\ \beta\frac{\partial S_2}{\partial\mathbf{u}} &= \beta\begin{bmatrix} {}^{\partial S_2}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{bmatrix},\end{aligned} \tag{46}$$

and the third being their sum,

$$\alpha\frac{\partial S_1}{\partial\mathbf{u}} + \beta\frac{\partial S_2}{\partial\mathbf{u}} = \alpha\begin{bmatrix} {}^{\partial S_1}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{bmatrix} + \beta\begin{bmatrix} {}^{\partial S_2}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{bmatrix}. \tag{47}$$

By varying α and β, we can change the linear combination into a large suite of possible vectors (in fact, recalling linear algebra, since the two constraint normal vectors are not parallel, we know we can construct any vector in $\mathbb{R}^U$ with them).

We will want to land on a special pair α and β of this continuous range of possibilities. To help select this special pair, let us consider the model space $\mathbb{R}^M$, and the objective function as realized in this space, i.e.,

$$\phi(\mathbf{u}(\mathbf{m})) = \frac{1}{2}[u_1, u_2]\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} - [u_1, u_2]\begin{bmatrix} d_1 \\ d_2 \end{bmatrix}, \tag{48}$$

as before, but now with the wavefield vector elements $u_1 = u_1(m_1, m_2)$ and $u_2 = u_2(m_1, m_2)$ being functions of the new set of independent variables, m_1 and m_2, and being determined by solving the forward problem for each pair of $\mathbf{m}$ elements. We have made a very significant change by doing this, even though we are still studying the same objective function ϕ; now $\mathbf{m}$ is the independent variable, and it has a strongly nonlinear relationship with ϕ, through $\mathbf{u} = \mathbf{u}(\mathbf{m})$. In Case 1, this produces the much more complicated "topography" in Figure 6d. In a real FWI problem, the exact topography would be much different (and would be realized in a much higher dimensional space), but the warped and complex contours we see here serve very well to illustrate the situation.

The special point (yellow circle) we have been studying in the wavefield space came from a special point defined in model space, $[m_1, m_2]^T = [-0.56, 0.7]^T$, so we can focus on this counterpart in the model space (yellow circle, Figure 6d). We have set out as our final goal to compute the gradient $\partial\phi/\partial\mathbf{m}$. This is a much more involved problem than the gradient calculation was in the wavefield space. But, for a low dimensional problem like this, the calculation of the target gradient can be done by brute force, and this gives us "the right answer", something to which we can compare other constructions. By numerical differencing, we produce the vector plotted in Figure 6d in white (actually we plot its negative, to allow us to keep a close focus on the point and the minimum).

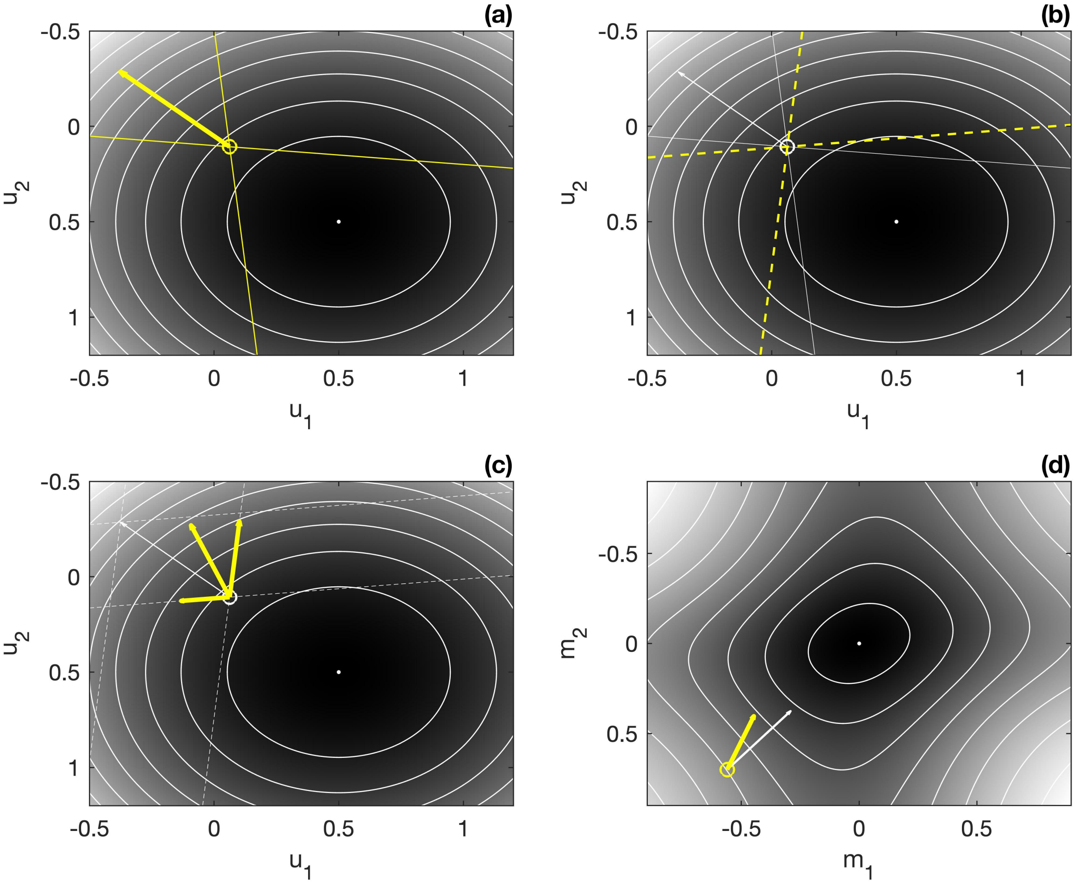

FIG. 6. Geometrical constructions within a low-dimensional adjoint-state example.

Now we make what will seem like a mysterious vector construction. Let us even pretend that we just dreamed it up, and allow what happens next to be a surprise discovery. Let the

U rows of the matrix $\mathbf{S}$ be column vectors $\mathbf{s}_i$, $i = 1, ..., U$. In Case 1 this means

$$\mathbf{s}_1 = \begin{bmatrix} s_{11} \\ s_{12} \end{bmatrix}, \; \mathbf{s}_2 = \begin{bmatrix} s_{21} \\ s_{22} \end{bmatrix}. \tag{49}$$

Take the transposes of these vectors, and form a matrix by taking the derivative of each element with respect to each element of the model vector:

$$\frac{\partial \mathbf{s}_1^T}{\partial \mathbf{m}} = \begin{bmatrix} \partial s_{11}/\partial m_1 & \partial s_{12}/\partial m_1 \\ \partial s_{11}/\partial m_2 & \partial s_{12}/\partial m_2 \end{bmatrix}, \; \frac{\partial \mathbf{s}_2^T}{\partial \mathbf{m}} = \begin{bmatrix} \partial s_{21}/\partial m_1 & \partial s_{22}/\partial m_1 \\ \partial s_{21}/\partial m_2 & \partial s_{22}/\partial m_2 \end{bmatrix}. \tag{50}$$

It will later be important to note that, though they look a little complicated, these matrices are actually very quick to fill in, knowing the elements of $\mathbf{S}$ and how they depend on the elements of $\mathbf{m}$. In Case 1,

$$\frac{\partial \mathbf{s}_1^T}{\partial \mathbf{m}} = \begin{bmatrix} 2am_1 & bm_2 \\ 0 & bm_1 \end{bmatrix}, \; \frac{\partial \mathbf{s}_2^T}{\partial \mathbf{m}} = \begin{bmatrix} bm_2 & 0 \\ bm_1 & 2cm_2 \end{bmatrix}. \tag{51}$$

We observe that the columns of these matrices are vectors in model space, but the rows are vectors in wavefield space, which means, we can take products of these matrices with our current special wavefield vector, $\mathbf{u} = [u_1, u_2]^T$. Call the results $\mathbf{v}_i$:

$$\mathbf{v}_1 = \frac{\partial \mathbf{s}_1^T}{\partial \mathbf{m}} \mathbf{u} = \begin{bmatrix} 2am_1 & bm_2 \\ 0 & bm_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \tag{52}$$

and

$$\mathbf{v}_2 = \frac{\partial \mathbf{s}_2^T}{\partial \mathbf{m}} \mathbf{u} = \begin{bmatrix} bm_2 & 0 \\ bm_1 & 2cm_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{53}$$

After the multiplication is complete, the resulting vectors are vectors in model space. Finally, assemble the $\mathbf{v}$ vectors in a linear combination, *using the same weights we used for the wavefield space construction in equation 45*, namely α and β:

$$\alpha \mathbf{v}_1 + \beta \mathbf{v}_2 = \alpha \begin{bmatrix} 2am_1 & bm_2 \\ 0 & bm_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \beta \begin{bmatrix} bm_2 & 0 \\ bm_1 & 2cm_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{54}$$

This vector is plotted as a yellow, bold arrow, alongside the white gradient arrow depicting $\partial\phi/\partial\mathbf{m}$, in Figure 6d.

So, we have two vectors, one in model space, in equation (54), and one in wavefield space, in equation (47), plotted side by side in Figures 6c and d. The two vectors are linked in that they are built using the same pair of coefficients, α and β. We want to watch how they behave in comparison to one another as we vary the coefficients. To befit what these coefficients actually represent, let us re-name them $\alpha = \lambda_1$ and $\beta = \lambda_2$, in which case the two vectors become

$$\lambda_1 \begin{bmatrix} 2am_1 & bm_2 \\ 0 & bm_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \lambda_2 \begin{bmatrix} bm_2 & 0 \\ bm_1 & 2cm_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \tag{55}$$

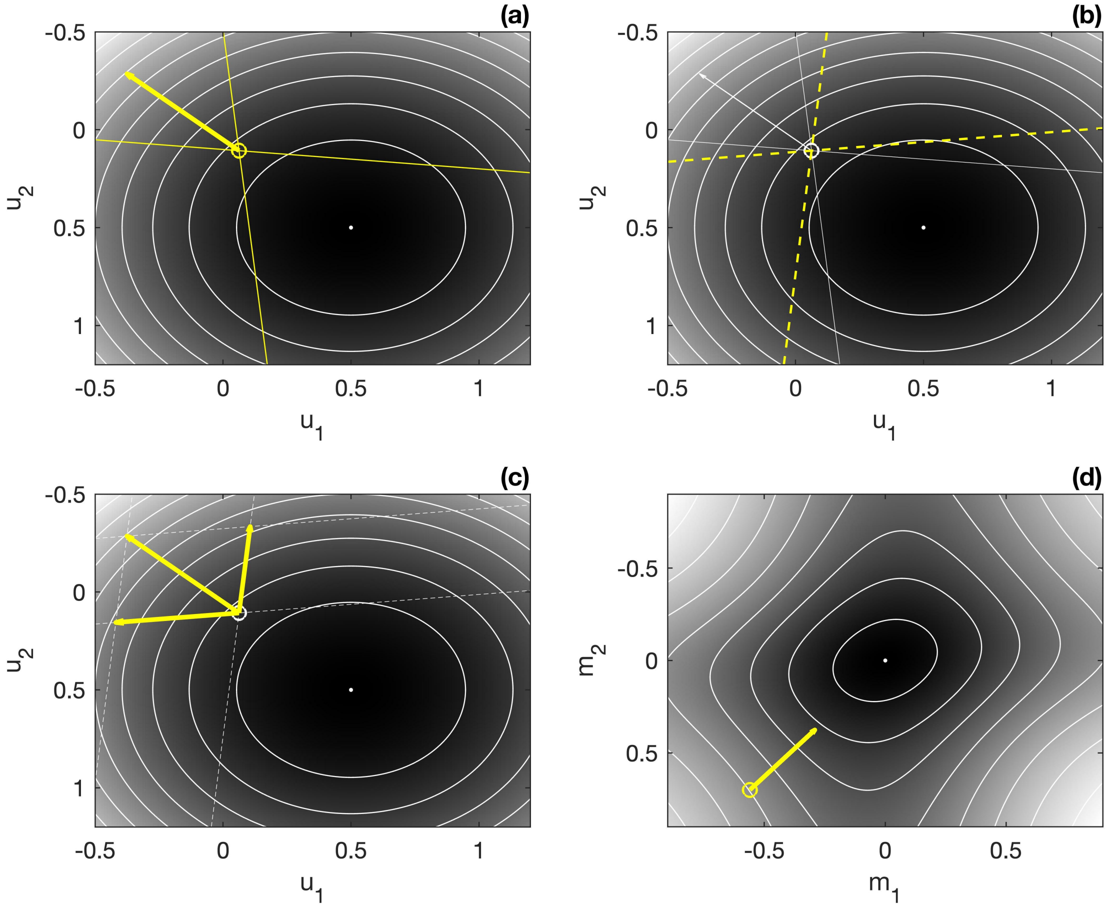

FIG. 7. Geometrical constructions within a low-dimensional adjoint-state example.

and

$$\lambda_1 \begin{bmatrix} {}^{\partial S_1}\!/_{\partial u_1} \\ {}^{\partial S_1}\!/_{\partial u_2} \end{bmatrix} + \lambda_2 \begin{bmatrix} {}^{\partial S_2}\!/_{\partial u_1} \\ {}^{\partial S_1}\!/_{\partial u_2} \end{bmatrix}. \tag{56}$$

In Figures 6c and d, the coefficients ($\alpha = \lambda_1$ and $\beta = \lambda_2$) were arbitrarily chosen. Let us now choose values for them such a specific goal is accomplished in the wavefield space, via equation (56). Let us choose λ_1 and λ_2 such that the construction in equation (56) exactly matches the gradient $\partial\phi/\partial\mathbf{u}$, which is plotted in white in Figure 6c. That is, we choose λ_1 and λ_2 to enforce

$$\lambda_1 \begin{bmatrix} 2am_1 & bm_2 \\ 0 & bm_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \lambda_2 \begin{bmatrix} bm_2 & 0 \\ bm_1 & 2cm_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} {}^{\partial\phi}\!/_{\partial u_1} \\ {}^{\partial\phi}\!/_{\partial u_2} \end{bmatrix}. \tag{57}$$

How we found the values for these two coefficients which accomplished this is not too important– we can imagine we just tried pairs of λ_1 and λ_2 values until we found a match.

OK, we have found a pair of λ_1 and λ_2 values for which the construction in equation (56) takes on special significance. What is the effect on the counterpart model space construction (equation 55) of making this particular choice? In Figures 7a-d, we reproduce Figure 6, having enforced this in the wavefield construction; in 7c, we can see that we have successfully constructed a vector equal to the gradient, which is now obscured by the yellow vector. Inspecting the corresponding model space construction, in Figure 7d we make our "discovery", which, it turns out, is the central fact of the adjoint state method. When (57) holds, the second construction matches precisely the gradient $\partial\phi/\partial\mathbf{m}$:

$$\lambda_1 \left[\begin{array}{c} {}^{\partial S_1}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{array} \right] + \lambda_2 \left[\begin{array}{c} {}^{\partial S_2}\!/\!_{\partial u_1} \\ {}^{\partial S_1}\!/\!_{\partial u_2} \end{array} \right] = \left[\begin{array}{c} {}^{\partial \phi}\!/\!_{\partial m_1} \\ {}^{\partial \phi}\!/\!_{\partial m_2} \end{array} \right]. \tag{58}$$

We can think of this geometric discovery as being the nonlinear constrained optimization version of the linear "discovery" depicted in Figure 3, in which the gradient was found to be an element of the subspace spanned by the constraint normal vectors at, and only at, the point at which the objective function was a (constrained) minimum. This is slightly more complicated, but then, we already knew we were not going to be able to solve directly for a minimum. Here, we find something at least as powerful, however, *given how complicated the problem of determining the gradient of ϕ with respect to $\mathbf{m}$ is*. The coefficients needed to construct the *simple* gradient $\partial\phi/\partial\mathbf{u}$, which are relatively easy to determine, are the same coefficients needed to build the *complex* gradient $\partial\phi/\partial\mathbf{m}$ out of a model space basis whose elements are *simple*. This is a mouthful, but hopefully the implication is clear: we have a basic path to follow wherein a complex construction is possible based only on simple ingredients. It is rare to find violations of the storied principle of conservation of aggravation, but this is an example, if there ever was one.

The adjoint state method

We still have some work to do. What, after all, does any of this have to do with Lagrange multipliers? Evidently by our choice of terminology we are hinting that the coefficients linking the two constructions are the multipliers, but they certainly don't seem to come from that kind of approach at the moment. Let us formulate answers to this question, this time staying fully general, but illustrating, as needed, with our Case 2 example. The Case 1 example was admirably simple, but it went so far towards simplicity that it erased the distinction between data space and wavefield space, and we should bring that distinction back as soon as we can.

The previous discussion, both the general aspects and the aspects special to Case 1, were geared towards analyzing (and almost, but not quite, solving) the problem set out in equation (35):

$$\min_{\mathbf{m}} \phi, \quad \text{subject to } \mathbf{S}\mathbf{u} = \mathbf{f}. \tag{59}$$

Recall that in our discussion of the method of Lagrange multipliers for linear problems, in the end we took our geometrical / pictorial insight and used it to formulate an unconstrained problem that (via the insight) we argued to produce the same answer that we seek for the constrained problem. Let us do something similar – let us guess at a new objective function, involving, like in the linear case, a set of undetermined multipliers.

Like we did above, let us break the $U \times U$ system $\mathbf{S}\mathbf{u} = \mathbf{f}$ up into U individual equations. Let $\mathbf{s}_i$ be a column vector containing the elements of the ith row of $\mathbf{S}$, and then let $G_i =$

$\mathbf{s}_i^T\mathbf{u} - f_i = 0$, where f_i is the ith element of $\mathbf{f}$, be the ith constraint equation implied by the problem in (35). The new inverse problem is

$$\mathbf{m}^* = \min_{\mathbf{m},\boldsymbol{\lambda}} \mathcal{L},\ \mathcal{L} = \phi - \sum_{i=1}^{U} \lambda_i G_i = \phi - \boldsymbol{\lambda}^T\mathbf{g} = \phi - \langle \boldsymbol{\lambda}, \mathbf{g} \rangle, \tag{60}$$

where $\mathbf{g} = [G_1, ..., G_U]^T$. We introduce again the angle bracket notation for inner products between vectors in $\mathbb{R}^U$, a terminology often used in the literature. But, we will stop short of finding $\mathbf{m}^*$ in the formalism, which, again, when ϕ is not a quadratic function of $\mathbf{m}$, will not be available through a simple set of normal equations. Instead, we will seek the gradient of ϕ with respect to $\mathbf{m}$, a quantity which will drive an iterative approach to estimating $\mathbf{m}^*$. There is a mathematical subtlety here, however, and now is as good a time as any to deal with it.

Aside: review of differentiation with indirect dependencies

The subtlety is that ϕ and $\mathcal{L}$ depend on $\mathbf{m}$ directly, since in general we see $\mathbf{m}$ appearing explicitly within them, but also indirectly, because they also depend on $\mathbf{u}$ and we know that $\mathbf{u} = \mathbf{u}(\mathbf{m})$. This situation is common in methods involving Lagrangians, and we will be borrowing the calculus used in these problems for the adjoint state method, so let us pursue that analogy for a moment. Consider a Lagrangian L built up to describe the motion of 2 particles in one dimension, with spatial coordinates $x_1(t)$ and $x_2(t)$. Depending on the problem, L itself may also be an explicit function of time, so

$$L = L(t, x_1, x_2) = L\big(t, x_1(t), x_2(t)\big). \tag{61}$$

What would happen to L over a short interval of time Δt in a problem like this? The rate of change of L with time, according to multivariate calculus rules, is

$$\frac{\partial L}{\partial t}, \tag{62}$$

which suggests the change in L would be, to leading order in Δt,

$$\Delta L \underset{?}{\approx} \frac{\partial L}{\partial t}\Delta t. \tag{63}$$

But, we know that, in the interval Δt, both x_1 and x_2 have also changed, which affects L, and so the total change in L would not be reflected in this partial derivative. Instead, we would use the chain rule:

$$\Delta L \approx \frac{\partial L}{\partial t}\Delta t + \frac{\partial L}{\partial x_1}\frac{\partial x_1}{\partial t}\Delta t + \frac{\partial L}{\partial x_2}\frac{\partial x_2}{\partial t}\Delta t, \tag{64}$$

to capture all the ways L can be affected. This implies a kind of derivative that really sees all the impacts on L of motion:

$$\frac{dL}{dt} = \lim_{\Delta t \to 0}\frac{\Delta L}{\Delta t} = \frac{\partial L}{\partial t} + \frac{\partial L}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial L}{\partial x_2}\frac{\partial x_2}{\partial t} = \frac{\partial L}{\partial t} + \left(\frac{\partial L}{\partial \mathbf{x}}\right)^T \frac{\partial \mathbf{x}}{\partial t}, \tag{65}$$

where $\mathbf{x} = [x_1, x_2]^T$. On the right side we have recognized that the sum over space variable partial derivative terms can be framed as a dot product. We are going to use this derivative in our development, and likely you can already see that the model vector is going to play a role similar to the time, and the wavefield vector a role similar to the particle coordinates. To make the adjoint state and Lagrangian dynamics derivatives look even more similar to one another, let us ask what would happen if the physical world had more than one time coordinate — weird, to be sure, but perfectly possible from the point of view of self-consistent mathematical theory! In that case, we would have something like

$$L = L(t_1, t_2, x_1, x_2) = L\big(t_1, t_2, x_1(t_1, t_2), x_2(t_1, t_2)\big). \tag{66}$$

The variation ΔL, i.e., the change in L coming from evolution along both the t_1 and t_2 axes, would have to include explicit accommodation of changes over both t_1 and t_2, so we would have to allow for Δt_1 and Δt_2, and, the way each of x_1 and x_2 now vary with those two times:

$$\begin{aligned} \Delta L \approx & \frac{\partial L}{\partial t_1}\Delta t_1 + \frac{\partial L}{\partial t_2}\Delta t_2 + \frac{\partial L}{\partial x_1}\frac{\partial x_1}{\partial t_1}\Delta t_1 + \frac{\partial L}{\partial x_1}\frac{\partial x_1}{\partial t_2}\Delta t_2 \\ & + \frac{\partial L}{\partial x_2}\frac{\partial x_2}{\partial t_1}\Delta t_1 + \frac{\partial L}{\partial x_2}\frac{\partial x_2}{\partial t_2}\Delta t_2. \end{aligned} \tag{67}$$

The first two terms imply a dot product, and (after squinting at it for a while), the second four terms imply a bilinear form:

$$\Delta L \approx [\Delta t_1, \Delta t_2] \left[\begin{array}{c} \partial L/\partial t_1 \\ \partial L/\partial t_2 \end{array} \right] + [\Delta t_1, \Delta t_2] \left[\begin{array}{cc} \partial x_1/\partial t_1 & \partial x_2/\partial t_1 \\ \partial x_1/\partial t_2 & \partial x_2/\partial t_2 \end{array} \right] \left[\begin{array}{c} \partial L/\partial x_1 \\ \partial L/\partial x_2 \end{array} \right], \tag{68}$$

or

$$\Delta L \approx \Delta \mathbf{t}^T \left[\frac{\partial L}{\partial \mathbf{t}} + \frac{\partial \mathbf{x}}{\partial \mathbf{t}}\frac{\partial L}{\partial \mathbf{x}}\right] = \left[\frac{\partial L}{\partial \mathbf{t}} + \frac{\partial \mathbf{x}}{\partial \mathbf{t}}\frac{\partial L}{\partial \mathbf{x}}\right]^T \Delta \mathbf{t} \tag{69}$$

where $\Delta\mathbf{t} = [\Delta t_1, \Delta t_2]^T$, which is much more compact, as long as we remember what is intended by $^{\partial \mathbf{x}}/_{\partial \mathbf{t}}$, a matrix with a row for every time coordinate and a column for every space coordinate. This then implies a derivative which is a vector, with an element for each time:

$$\frac{dL}{d\mathbf{t}} = \frac{\partial L}{\partial \mathbf{t}} + \frac{\partial \mathbf{x}}{\partial \mathbf{t}}\frac{\partial L}{\partial \mathbf{x}}, \tag{70}$$

again, with the proviso that we remember what the matrix $^{\partial \mathbf{x}}/_{\partial \mathbf{t}}$ represents (i.e., refer to the explicit matrix in (68).

Back to the adjoint state method

We need to apply the derivative we developed above to the problem of change in our FWI problem. To accomplish this, we make the following assignments:

$$\begin{aligned} & L \to \mathcal{L}/\phi, \\ & \mathbf{t} \to \mathbf{m}, \\ & \mathbf{x} \to \mathbf{u}. \end{aligned} \tag{71}$$

The scalar functional is either of our objective functions ϕ, or $\mathcal{L}$, the fundamental independent variable, similar to time, is the model vector $\mathbf{m}$, and the sometimes dependent and sometimes independent varialbes, similar to the space coordinates, are the wavefield vectors $\mathbf{u}$. Let us now go back to the functional $\mathcal{L}$ in equation (60) and consider its derivatives. We had

$$\mathcal{L} = \phi - \sum_{i=1}^{U} \lambda_i G_i, \tag{72}$$

where

$$\phi = \frac{1}{2}\mathbf{u}^T\mathbf{R}^T\mathbf{R}\mathbf{u} - \mathbf{u}^T\mathbf{R}^T\mathbf{d}, \tag{73}$$

and

$$G_i = \mathbf{s}_i^T\mathbf{u} - f_i = 0, \tag{74}$$

where $\mathbf{s}_i$ is a column vector containing the elements of the ith row of the impedance matrix $\mathbf{S}$. In Case 2, the ingredients are

$$\phi = \frac{1}{2}[u_1, u_2, u_3]\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} - [u_1, u_2, u_3]\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} d_1 \\ d_2 \end{bmatrix}, \tag{75}$$

i.e.,

$$\phi = \frac{1}{2}[u_1, u_2, u_3]\left\{\begin{bmatrix} u_1 \\ u_2 \\ 0 \end{bmatrix} - \begin{bmatrix} d_1 \\ d_2 \\ 0 \end{bmatrix}\right\}, \tag{76}$$

and since

$$\begin{aligned} G_1 &= s_{11}u_1 + s_{12}u_2 + s_{13}u_3 - f_1 \\ G_2 &= s_{21}u_1 + s_{22}u_2 + s_{23}u_3 - f_2 \\ G_3 &= s_{31}u_1 + s_{32}u_2 + s_{33}u_3 - f_3, \end{aligned} \tag{77}$$

we have for $\mathcal{L}$

$$\begin{aligned} \mathcal{L}(\mathbf{u}) =& \frac{1}{2}[u_1, u_2, u_3]\left\{\begin{bmatrix} u_1 \\ u_2 \\ 0 \end{bmatrix} - \begin{bmatrix} d_1 \\ d_2 \\ 0 \end{bmatrix}\right\} \\ & - \lambda_1\big(s_{11}u_1 + s_{12}u_2 + s_{13}u_3 - f_1\big) \\ & - \lambda_2\big(s_{21}u_1 + s_{22}u_2 + s_{23}u_3 - f_2\big) \\ & - \lambda_3\big(s_{31}u_1 + s_{32}u_2 + s_{33}u_3 - f_3\big). \end{aligned} \tag{78}$$

Let us now set up derivatives of $\mathcal{L}$ with respect to all of the unknowns. The means

$$\frac{d\mathcal{L}}{d\mathbf{m}} = \frac{\partial\mathcal{L}}{\partial\mathbf{m}} + \frac{\partial\mathbf{u}}{\partial\mathbf{m}}\frac{\partial\mathcal{L}}{\partial\mathbf{u}}, \tag{79}$$

and

$$\frac{d\mathcal{L}}{d\boldsymbol{\lambda}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \mathbf{g}. \tag{80}$$

Now, the program from the linear problem was to set both of these quantities to zero, but we have already realized that this exact step will not be useful here, since we cannot determine in one step the minimum on account of the nonlinearity of the problem. We will do something similar here, but not exactly the same. The second set of equations will be set to zero as before:

$$\mathbf{g} = \mathbf{0}, \text{ or } G_i = 0, \ i = 1, ..., U. \tag{81}$$

This simply recovers the constraints, apparently, but it instructs us to do something pretty important. It instructs us not to permit $\mathbf{u}$ to be a general vector in $\mathbb{R}^U$, but only an element satisfying $\mathbf{Su} = \mathbf{f}$. In terms of our Case 2:

$$\begin{bmatrix} d\mathcal{L}/dm_1 \\ d\mathcal{L}/dm_2 \\ d\mathcal{L}/dm_2 \end{bmatrix} = \begin{bmatrix} \partial\mathcal{L}/\partial m_1 \\ \partial\mathcal{L}/\partial m_2 \\ \partial\mathcal{L}/\partial m_3 \end{bmatrix} + \begin{bmatrix} {}^{\partial u_1}\!/\!{}_{\partial m_1} & {}^{\partial u_2}\!/\!{}_{\partial m_1} & {}^{\partial u_3}\!/\!{}_{\partial m_1} \\ {}^{\partial u_1}\!/\!{}_{\partial m_2} & {}^{\partial u_2}\!/\!{}_{\partial m_2} & {}^{\partial u_3}\!/\!{}_{\partial m_2} \\ {}^{\partial u_1}\!/\!{}_{\partial m_3} & {}^{\partial u_2}\!/\!{}_{\partial m_3} & {}^{\partial u_3}\!/\!{}_{\partial m_3} \end{bmatrix} \begin{bmatrix} \partial\mathcal{L}/\partial u_1 \\ \partial\mathcal{L}/\partial u_2 \\ \partial\mathcal{L}/\partial u_3 \end{bmatrix}. \tag{82}$$

Examining the detailed form of $\partial\mathcal{L}/\partial\mathbf{u}$, we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = \frac{\partial \phi}{\partial \mathbf{u}} - \sum_{i=1}^{U} \lambda_i \frac{\partial G_i}{\partial \mathbf{u}}; \tag{83}$$

which clarifies what we have seen in the low-dimensional examples. When

$$\frac{\partial \phi}{\partial \mathbf{u}} = \sum_{i=1}^{U} \lambda_i \frac{\partial G_i}{\partial \mathbf{u}}, \tag{84}$$

then and only then is

$$\frac{d\mathcal{L}}{d\mathbf{m}} = \frac{\partial \mathcal{L}}{\partial \mathbf{m}}, \tag{85}$$

the right-hand side now being computable in terms of the λ_i values satisfying 84.

CONCLUSIONS

This set of excursive remarks connecting the adjoint-state method ideas with those of linear constrained optimization has been designed to bring a more geometrical viewpoint to the gradient calculation within FWI. There is no way the author, re-reading what he has written, can claim these remarks have simplified the problem. However, something of the main purpose, which was to force a reader who has made it this far to draw pictures in their minds when using the adjoint state method, seems to have been achieved.

ACKNOWLEDGEMENTS

The sponsors of CREWES are gratefully thanked for continued support. This work was funded by CREWES industrial sponsors, NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 543578-19, and in part by an NSERC-DG.