Review of "Full waveform inversion by model extension: theory, design and optimization", GEO22-0350

September 7, 2022

1 Summary

This paper and its companion (22-0382) make a major contribution to the literature on full waveform inversion (FWI) and avoidance of cycle-skipping. The exposition has a number of easily-corrected deficiencies, and raises several questions implicitly which should be explicitly recognised.

2 Overview

FWI is often (though not always) formulated as a least squares problem, the minimization of

$$\mathbf{\Phi}^{\mathrm{FWI}}(\mathbf{m}) = \|\mathbf{f}(\mathbf{m}) - \mathbf{d}\|^2. \tag{1}$$

In the definition 1, $\mathbf{f}$ is a modeling operator, computed by solving an initial-boundary value problem for a system of wave equations, $\mathbf{m}$ is a model vector consisting of (functions of) the coefficients of these equations, and $\mathbf{d}$ is a field of data traces. The definition is often augmented by constraints and regularization terms designed to enforce physically sensible qualities of optimal models. Starting with the pioneering contributions of Tarantola (1984), Lailly (1983) [these and other citations appear in the manuscript bibliography, unless otherwise noted], and others in the 1980s, minimization of $\mathbf{\Phi}^{\mathrm{FWI}}$ and relatives has been based continuous optimization, that is, on Newton's method and its relatives, which are in turn based on on linearization of the modeling operator: in the notation of this paper,

$$\mathbf{f}(\mathbf{m} + \mathbf{p}) \approx \mathbf{f}(\mathbf{m}) + \mathbf{B}(\mathbf{m})\mathbf{p} \tag{2}$$

in which $\mathbf{B}$ is the linearization or derivative of the forward modeling operator $\mathbf{f}$ (known in this business as the Born modeling operator, hence the use of "B"), and $\mathbf{p}$ is a model perturbation. As nicely illustrated in Figure 4 and the discussion on pp. 15-16, the least-squares solution arising from this approximation is unable to fit data with any choice of model perturbation $\mathbf{p}$ if the slowness $\mathbf{m}$ is substantiallly incorrect in its prediction of data

kinematics. Therefore any attempt to update $\mathbf{m}$ through reducing the linearized (Born) fit error may fail through inability to access kinematic information in the data, unless the initial estimate of slowness is already near-correct in the kinematic sense. This failure mode has come to be (somewhat loosely) termed "cycle-skipping", because of the typical appearance of predicted waveforms (in $\mathbf{f}(\mathbf{m})$) shifted by more than a half-cycle from observed waveforms.

The work discussed in this paper is in the *extended modeling* genre. That is, it provides additional parameters to the model and modeling operator so that local improvement in data fit becomes possible. Using this manuscript's convention for extended quantities, denote the extended model parameter field by $\tilde{\mathbf{m}}$, the extended modeling operator by $\tilde{\mathbf{f}}$, and its derivative by $\tilde{\mathbf{B}}$. The extensions used in the present work amounts to permitting some of the components of $\tilde{\mathbf{m}}$, which are coefficients in the wave equation, to become *operators*, rather than merely multiplication by functions. This extension is physically equivalent to relaxing the no-action-at-a-distance axiom of continuum mechanics. [For example, the square slowness is proportional to the compliance (reciprocal bulk modulus) for constant density acoustics, the physical setting of this paper. The bulk modulus is the ratio of stress and strain (pressure and divergence of displacement) in acoustics, so permitting it to become a (non-local) operator represents action-at-a-distance.]

The first order approximation to the extended modeling operator is

$$\tilde{\mathbf{f}}(\tilde{\mathbf{m}}+\tilde{\mathbf{p}}) \approx \tilde{\mathbf{f}}(\tilde{\mathbf{m}}) + \tilde{\mathbf{B}}(\tilde{\mathbf{m}})\tilde{\mathbf{p}}. \tag{3}$$

As has been verified in some of the cited references, and illustrated in Figure 4, for suitable choices of model extension, any reference extended model $\tilde{\mathbf{m}}$ and data field $\mathbf{d}$, an extended model perturbation $\tilde{\mathbf{p}}$ can be found so that the right hand side of equation 3 closely approximates $\mathbf{d}$. Thus various approximation to Newton's method can at least achieve local improvement in data fit.

Since the (action-at-a-distance) extension violates basic continuum mechanics, it must be suppressed at the solution somehow. One approach to ensuring that the ultimate solution is physical (that is, not extended) uses an operator ("annihilator") $\mathbf{A}$ that yields a vanishing result when applied to a physical model field $\mathbf{m}$: $\mathbf{Am} = 0$, so that any action at non-zero distance is penalized. A simple way to employ such a penalty is in the form of a "penalty function":

$$\mathbf{\Phi}_\epsilon^{\mathrm{NL}}(\tilde{\mathbf{m}}) = \|\tilde{\mathbf{f}}(\tilde{\mathbf{m}}) - \mathbf{d}\|^2 + \epsilon^2 \|\mathbf{A}\tilde{\mathbf{m}}\|^2. \tag{4}$$

in which the vertical bars denote the root mean square, or L^2 norm (Symes, 2008). If $\mathbf{A}\tilde{\mathbf{m}}$ is small at a minimizer of $\mathbf{\Phi}^{\mathrm{NL}}$, then $\tilde{\mathbf{m}}$ is close to physical, and approximates a (non-extended) minimizer of $\mathbf{\Phi}^{\mathrm{FWI}}$.

A major disadvantage to this approach is computational: the extended model field $\tilde{\mathbf{m}}$ acts as an operator on the wavefield and its space and time derivatives, so is represented after discretization by a full matrix, which must be applied at every time step in a typical time domain wave solver. A physical model field $\mathbf{m}$, in contrast, is represented by a diagonal matrix, much cheaper to apply.

This computational disadvantage can be ameliorated to some extent by *replacing* the extended modeling operator $\tilde{\mathbf{f}}$ with its extended Born approximation 3 at a physical model

$\tilde{\mathbf{m}} = \mathbf{m}$, with extended perturbation $\tilde{\mathbf{p}}$. Since physical models are (special cases of) extended models, the first order perturbation approximation 3 with $\tilde{\mathbf{m}} = \mathbf{m}$ still approximates essentially arbitrary data with suitable choice of extended perturbation $\tilde{\mathbf{p}}$. At the solution, the model estimate $\mathbf{m} + \tilde{\mathbf{p}}$ should be physical. Since the annihilator is linear (in most version of this approach), this condition is equivalent to $\mathbf{A}\tilde{\mathbf{p}} = 0$. A penalty function capturing this idea reads

$$\mathbf{\Phi}_\epsilon^{\mathrm{MVA}}(\mathbf{m}, \tilde{\mathbf{p}}) = \|\mathbf{f}(\mathbf{m}) + \tilde{\mathbf{B}}(\mathbf{m})\tilde{\mathbf{p}} - \mathbf{d}\|^2 + \epsilon^2\|\mathbf{A}\tilde{\mathbf{p}}\|^2. \tag{5}$$

Minimization of $\mathbf{\Phi}_\epsilon^{\mathrm{MVA}}$ is a version of *Wave Equation Migration Velocity Analysis*, or WEMVA. Because the extended modeling operator $\tilde{\mathbf{f}}$ has been replaced by its linearization about physical models, even a noise-free solution at which the objective value = 0 is not an FWI solution, but rather a Born-FWI solution, that is, $\mathbf{d} \approx \mathbf{f}(\mathbf{m}) + \mathbf{B}(\mathbf{m})\mathbf{p}$. This approximation is satisfactory exactly when the Born approximation is satisfactory, meaning that the data consists essentially of primary reflections away from critical angle. This status can sometimes be achieved by preprocessing, but field data often contain significant amounts of refracted, diving wave, and non-primary reflection energy, all carrying information about the subsurface but treated as noise by WEMVA. This is one possible reason that WEMVA has not migrated into common industrial practice. Another is that the computation of the extended Born operator $\tilde{\mathbf{B}}$ involves full matrix multiplies in each time step of a simulation loop, just as does the extended modeling operator $\tilde{\mathbf{f}}$. This intrinsic expense dominates the cost of WEMVA.

The manuscript under review explains a clever modification of WEMVA that frees it from the Born limitation, and delivers an approximate FWI solution instead. This modification has two components. The first component is the observation that if $\tilde{\mathbf{p}} \approx 0$ at the solution, then $\mathbf{m}$ is a minimizer of $\mathbf{\Phi}^{\mathrm{FWI}}$. To force this happy circumstance to occur, the annihilator $\mathbf{A}$ is augmented by a multiple of the identity operator:

$$\mathbf{D} = \begin{bmatrix} \alpha\mathbf{I} \\ \mathbf{A} \end{bmatrix} \tag{6}$$

with $\alpha > 0$ (see Figure 11). Thus $\mathbf{D}\tilde{\mathbf{p}} = 0$ implies not just that $\tilde{\mathbf{p}}$ is physical, but that $\tilde{\mathbf{p}} = 0$. The modified penalty term in the FWIME objective is then $\epsilon^2\|\mathbf{D}\tilde{\mathbf{p}}\|^2$, rather than $\epsilon^2\|\mathbf{A}\tilde{\mathbf{p}}\|^2$:

$$\mathbf{\Phi}_\epsilon^{\mathrm{IME}}(\mathbf{m}, \tilde{\mathbf{p}}) = \|\mathbf{f}(\mathbf{m}) + \tilde{\mathbf{B}}(\mathbf{m})\tilde{\mathbf{p}} - \mathbf{d}\|^2 + \epsilon^2\|\mathbf{D}\tilde{\mathbf{p}}\|^2. \tag{7}$$

Figures 14 - 18 show that this innovation may not be quite enough, in that the search direction (gradient) generated by the modified penalty function may have a tendency to overemphasize local oscillatory perturbations, similar to the the gradient of $\mathbf{\Phi}^{\mathrm{FWI}}$. It's not obvious that iteration would lead to stagnation, but the search direction displayed in Figure 18 (c) is certainly not efficient.

The second innovation is designed to produce a more efficient search direction, by application of a smoothing operator whose effective width decreases as the FWIME iteration proceeds, eventually permitting features of all scales to migrate into $\mathbf{m}$. Presumably the justification for model smoothing is the familiar observation that the longest spatial frequencies

in the model field have the strongest influence (for the most part) on travel times. Therefore to match kinematics, the smoothest parts of the model should be favored, at least initially. The authors accomplish this goal by representing $\mathbf{m}$ on a (possibly spatially non-uniform) spline space. The effect is to smooth the gradient of the FWIME objective function by the operator $\mathbf{SS}^T$, $\mathbf{S}$ being the spline representation operator (spline coefficients to fine grid values). Figures 21 (a)-(c) show the effect of spline smoothing on the gradient components displayed in Figures 18 (a)-(c), which is indeed constructive.

To achieve eventual equivalence to FWI, fine scale features must migrate into the model estimate. This migration is achieved by repeated refinement of the spline grid. Refinements occur when the model updates for a given spline grid stagnate. The current model estimate is then projected onto a finer spline grid and the iteration restarted. In the final block of iterations, the spline grid is identical to the fine simulation grid. The three major examples presented at the end of the paper illustrate the effectiveness of this strategy, the eventual vanishing of the extended perturbation $\tilde{\mathbf{p}}$, and the recovery of an effective FWI solution.

In this overview I have neglected several concepts emphasized in the exposition, such as Born and tomographic gradients. Also it should be mentioned that the variable projection method plays a critical role in the formulation and performance of the algorithm. However I believe that in some sense these are side issues, and the essential ideas are those I have described.

3 Comments

The approach described in this paper represents a significant advance in the literature on anti-cycle-skip methods. The paper itself however suffers from two defects: (1) not enough information, and (2) too much non-information. The purpose of a scientific paper is to present new results, and to describe the method used to produce these results in enough detail that a knowledgeable reader could (at least in principle) reproduce the results, *and* integrate the methodology into their own work. Note that availability of a computer program or programs with input data does not necessarily accomplish the second goal: if the reader can *only* reproduce the results of the paper, then the reader has learned nothing of any use. Too much information is left out of the current manuscript to allow even a knowledgeable reader to transfer the methodology to his/her own projects. On the other hand, repeated and repetitive claims for the virtues of the approach are distracting, and in some cases misleading. In the following couple of pages, I will try to give a representative list of missing information and unnecessary misinformation, hopefully in enough detail that the authors can easily fix both types of defect.

3.1 Model Specification

The physics used throughout is introduced on p. 7, in English: $\mathbf{f}$ is the "discretized acoustic isotropic constant-density forward modeling operator". Next up (skipping the WEMVA

subsection on p. 8, about which more later) is $\tilde{\mathbf{B}}$, the "extended Born modeling operator". Then $\mathbf{D}$, "an invertible modified form of the DSO operator".

It's likely true that most interested readers will understand what is meant by "discretized acoustic isotropic constant-density forward modeling operator". However, one cannot tell from the manuscript what discretization is used here, nor what source representation (isotropic point source, probably, but with or without correction for off-grid placement? And with what wavelets, in the various examples?).

Fewer will be familiar with the "extended Born modeling operator". For one thing, what extension? And for that matter, what is an extension? Only those marinated in the literature on anti-cycle-skip are likely to know.

Finally, only specialists are likely to recognize "an invertible modified form of the DSO operator", especially since the references cited on p 10 discuss extensions and DSO operators quite different from those used here. Furthermore the invertible modification is not described at all - what invertible modification?

The paper really needs to be modified to include explicitly all of this missing information. But this is easy: an appendix is a natural solution. Start with the acoustic constant density wave equation, introduce sampling with isotropic point sources and receivers. Take the slowness derivative to obtain the Born operator. Define the model extensions you actually use (subsurface horizontal offset and time shift). Describe briefly the finite difference approach you use to turn these into discrete problems, for both extended and non-extended contexts. Then define the operator $\mathbf{D}$ (I offer display 6 at no extra charge).

A reader who already knows all of this stuff, or doesn't care, can ignore the appendix, but one who really wants to be sure of following your reasoning will read it.

3.2 Hyperparameters, hyperparameters everywhere...

The authors assert that their method is "simple to use" (p. 1), its features are mathematically consistent "thereby reducing the number of optimization hyper-parameters to two" (p. 5), and result in a method that "can invert any type of seismic data using the same framework and without the need for intensive hyper-parameter tuning" (p. 6). Also, "...FWIME is formulated in a compact and mathematically consistent manner that only requires a simple tuning of one hyperparameter at the initial step" (p. 21). Many other statements throughout the paper echo this notion that the method is largely automated so that non-expert users can easily adopt it.

While this goal (automation via good default or computed choices of essential parameters) is very laudable, it's not yet achieved in this work. There are far more than one (or two) "hyper-parameters" that must be chosen as input to Algorithm 2. I address a number of them in this subsection, beginning with the two that are actually called out.

3.2.1 Penalty Parameter ϵ

This "trade-off" parameter is one of the two called out explicitly in the manuscript. It is to be selected by "a trial and error approach based on examining a subset of the CIGs extracted

from the optimal extended perturbation $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}$ computed at the initial step". Very well, but how? What aspect of the CIGs should be used to set this parameter? How many CIGs need to be examined? Note that any correction requires re-computation of $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}$, which (as later discussed) is the principal computational bottleneck of the method. So this step is manual *and* computationally expensive, but the authors do not describe how to carry it out!

I suggest including an example that involves an initial incorrect (too large or too small) choice of ϵ, and its correction based on whatever the authors actually do by trial and error. The examples included in the current manuscript do not show this process - and it is absolutely critical! Too small, and not enough resolution of slowness; too large, and the method behaves like FWI. How do you tell where you are in this spectrum by looking at CIGs? The behaviour of the iteration may be relatively insensitive to ϵ, as noted on p. 21, however apparently it cannot be off by much more than an order of magnitude. Note that ϵ is a dimensional parameter, which makes its tuning even more difficult.

3.2.2 Extent of Extension

p. 29 bottom: "Figure 17 illustrates how $\tilde{\mathbf{p}}^{\text{opt}}$ contains crucial kinematic information and it shows the importance of using an extended perturbation with a large-enough extension (otherwise the information would be lost)."

The obvious question is, how does the user determine what is "large enough"? It appears that the answer is "trial and error", by spot-checking gathers. However even more important: what quality should the extended gather possess? In principle, this question has an easy answer: the extension axis (axes) should be large enough that the data is well-fit at the beginning. However, that is not the whole story. Should the test be that the data is well-fit with $\epsilon = 0$? Since $\epsilon > 0$ is in tension with data fit, perhaps that's the right answer, but note that this once again involves computing $\tilde{\mathbf{p}}^{\text{opt}}$, which is the single most expensive part of the algorithm.

Again, I think the simplest way to address this deficiency is to include a non-trivial example, in which an inadequate extension axis is adjusted. Since looking at individual data gathers is not going to reduce the expense, presumably the relative residual of the entire data set

$$\|\mathbf{f}(\mathbf{m}) + \tilde{\mathbf{B}}(\mathbf{m})\tilde{\mathbf{p}} - \mathbf{d}\|/\|\mathbf{d}\|$$

should be below some suitable tolerance. What tolerance? Or perhaps the authors use some other criterion. Please explain!

3.2.3 Type of Extension

The examples use the time-shift extension as in Biondi and Almomin (2014), which is fine for 2D examples. This choice gets little discussion. On p. 26, we find "...for 3D field applications, the time-lag extension requires a single additional axis (compared to two axes for space lags), thereby reducing the computational cost and memory footprint of the method.". However, the purpose of the extension is to permit the extended model to fit the data, thus transporting kinematic information from data space to model space. In 3D (full azimuth), the number of

data axes is 5, whereas with the time-shift extension the number of extended model axes is 4. This parameter count suggests that at least in some cases a time-shift extended model will not be adequate for 3D application.

3.2.4 Penalty parameter α

The operator $\mathbf{D}$ is implicitly defined in Figure 11, and depends on a parameter α. I have taken the liberty of creating an actual definition 6, which may describe the operator $\mathbf{D}$ used in this work. It depends on a parameter α. If $\alpha = 0$, then $\mathbf{D}$ is effectively a WEMVA penalty operator. As stated on p. 20, "...since the goal is not to obtain a well-cofused image (but to make $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}$ vanish), we modify the DSO operator by also penalizing energy located on the physical plane of $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}$."

So: how do we choose α? So far as I can tell, the choice of α used in the examples is nowhere given, nor is any principle suggested to make this choice. The authors need to include information about this. It's fine if they just say $\alpha = 1$ or some such arbitrary choice, even better if they give a reason for a specific choice, but they need to tell us what they chose, and why.

This choice raises another question: if the goal is to suppress all the energy in the extended part of the model $\tilde{\mathbf{p}}$, why not just dispense with the WEMVA part of $\mathbf{D}$ altogether, and choose $\mathbf{D} = \mathbf{I}$? The authors might try to answer this question as part of describing the effect of various choices of α.

3.2.5 Number of CG iterations, accuracy of $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}$

The number of CG iterations used to compute $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}$ is mentioned in each example, but not the reason for the choice (which varies from example to example). In fact, this is another number that the user much choose. How did the authors make this choice, and what advice do they have for a potential user?

This question is more subtle than it might at first appear.

A natural criterion for truncation of the CG loop is that the gradient of $\Phi_{\epsilon,\mathbf{m}}(\tilde{\mathbf{p}})$ (equation 6, p. 10) decrease in length below a prescribed maximum (absolute tolerance) or below a prescribed factor of its length at iteration 0 (relative tolerance). The first iteration of CG is a steepest descent step along the gradient, which could also be viewed as an extended RTM image. Would that search direction be sufficient? If so, a great deal of the expense in carrying out FWIME could be eliminated. Actually, using RTM rather than a more accurate minimization of $\Phi_{\epsilon,\mathbf{m}}(\tilde{\mathbf{p}})$ almost surely won't deliver an accurate gradient of $\Phi_\epsilon(\mathbf{m})$, which is the main goal of computing $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}$.

This topic is ripe for further research, but there have been a couple of studies. Symes and Kern (1994) pointed out that RTM is insufficient to yield a useful gradient, but so is CG with any of the common truncation criteria. The reason is that the expression 16 (p. 22) of $\nabla\Phi_\epsilon(\mathbf{m})$ is incorrect. It is missing the term

$$(D\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m}))^*[\tilde{\mathbf{B}}(m)^*(\mathbf{f}(\mathbf{m}) + \tilde{\mathbf{B}}(\mathbf{m})\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m}) - d) + \epsilon^2\mathbf{D}^*\mathbf{D}\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m})],$$

which comes from applying the chain rule to the right-hand side of equation 6. Of course, the quantity in square brackets is precisely the gradient of $\Phi_{\epsilon,\mathbf{m}}(\tilde{\mathbf{p}})$ (with respect to $\tilde{\mathbf{p}}$. An iterative method such as CG eventually makes this gradient small (in the L^2 sense) but does not make it vanish, in any finite number of iterations. That would be fine, *except* that the operator $(D\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m}))^*$ involves the derivative of $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m})$ with respect to $\mathbf{m}$. This $\mathbf{m}$ derivative has a similar effect to spatial derivatives, that is, enhances high frequency content without limitation, and can therefore give a large result even when applied to an L^2-small $\nabla\Phi_{\epsilon,\mathbf{m}}(\tilde{\mathbf{p}})$. Thus the missing term in equation 16 can be quite large.

Symes and Kern (1994) give examples of this phenomenon showing that it can cause the computed gradient via the VPM formulae (of which equation 16 is an example), with the equivalent of $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m})$ computed via CG, to be uselessly inaccurate. They also describe a correction that compensates for the missing term using Chebyshev (rather than CG) iteration, and show examples of accurate gradient calculations using this alternative technique. Twenty years later, the group of Hervé Chauris revisited this topic, proposing several remedies. The *function* Φ_ϵ with CG-derived $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m})$ is differentiable, and the errors in its values are controlled by the L^2 error achieved by CG. Chuaris et al. applied automatic differentiation to calculate the adjoint derivative $(D\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m}))^*$, and thus computed the gradient of Φ_ϵ with CG-derived $\tilde{\mathbf{p}}_\epsilon^{\text{opt}}(\mathbf{m})$ to machine precision [Cocher et al., 2017,]. Alternatively, they showed that a stable gradient could be obtained by application of a suitable filter to the naive VPM gradient, without however giving a convergence proof [Cocher et al., 2018,].

Apart from these references, the issue is ignored altogether in the literature on WEMVA and related topics. I speculate that the neglect of gradient instability may be in some part responsible for the failure of extended model techniques to have a serious impact on geophysical practice.

3.2.6 Spline grids, refinement schedule

The authors are good about telling us what spline grids are used in the examples (though for the last example we learn only the first and the last (the finite difference grid) and that there is another grid in between). As for how a user might construct the spline grid schedule - how many grids, what degree of refinement from one to the next - I cannot find any advice. It would be appreciated. Too few grid levels could presumably result in failure to converge to an optimal model. Too many is just expensive. It may well be that the authors don't have a comprehensive theory for managing the grid schedule, but they clearly have some idea about how to do it. There are least two parameters here that need choosing - the coarsest grid level, and the refinement factor. That's for the simplest case, of similar grids.

The introduction of local refinements would introduce an entirely new set of choices. None of the FWIME examples use non-uniform grids, nor discuss how to include non-uniform refinement in FWIME. I think that even the minor amount of discussion of non-uniform grids, centering around Figures 19 and 20, is out of place here and should be reserved for another paper, where nonuniform grids are actually used.

A side-issue: the dismissal of a spatial filter as in Biondi and Almomin (2014) on p. 30 is misleading. Algorithm 2 precisely applies a spatial filter. The use of a spline grid to

construct it is a convenience, not an essential feature. Some mechanism to control the degree of smoothing, or width of the averaging kernel, is essential in this application, however there are many ways (other than the spline construction) to do that. The essential point is the coupling of the degree of smoothing to the optimization, reducing the spatial scale of $\mathbf{m}$ towards zero as the iteration proceeds.

3.2.7 Optimization Parameters

FInally, every optimization algorithm depends on parameter choices. For CG, it suffices to specify the number of iterations, as already discussed, though that may not be optimal, as also discussed. For LBFGS, the size of the approximate inverse Hessian and several line search parameters need specification. This is true for any application of these methods, and is not specific to FWIME.

3.2.8 Summary

I count at least 10 important user-selected parameters. Maybe 5 (mostly the optimization parameters, and maybe α) could be given usable defaults, leaving another 5 (at least, more in 3D) as user choices. It's not just one, or even two.

3.3 Forms of WEMVA

The objective $\Phi_\epsilon^{\text{MVA}}$ defined in equation 5, which I claimed to be a form of WEMVA, differs from the function Φ_{WEMVA} discussed on p. 8. However, the latter is a limiting case of the former: just as the FWIME function tends to the FWI objective as $\epsilon \to \infty$, so does $\epsilon^{-2}\Phi_\epsilon^{\text{MVA}}$ tend to Φ_{WEMVA} as $\epsilon \to 0$, provided that we make two identifications: (1) $\mathbf{E} = \mathbf{A}$, and (2) we solve for $\tilde{\mathbf{p}}$ using one CG iteration, so that $\tilde{\mathbf{p}}$ is the RTM image $\mathbf{I}$ as defined in equation (3). The earliest accounts of VPM in extended inversion, such as Symes and Kern (1994), used the penalty function form given in equation 5 of this review, along with much more accurate solution of the inner problem.

In any case, the relation between FWIME and WEMVA in the form of definition 5 in this review seems more evident than in the form of equation 2 in the manuscript, and the latter is actually a limiting special case of the former. The authors might consider revising this discussion accordinly.

3.4 Transmitted and refracted wave inversion

The first and third examples are excellent, and clearly differentiate FWIME from the standard viewpoint of MVA as a process applied to primary reflections. However the observation that extended inversion of the type considered here can accomodated transmitted and diving waves is not new. [Shen, 2013] and [Lameloise and Chauris, 2016] independently observed that subsurface offset extension provides a framework for model-domain analysis of velocity error from transmitted wavefields, including diving waves.

This earlier work did not produce algorithms that accommodate all wave types within one workflow. FWIME, through its two innovations (the disappearing extended perturbation, and integrated management of spatial scale) does exactly that.

3.5 FWIME vs. RFWI

How does FWIME actually compare to RFWI, in ability to converge to a good nonlinear least squares solution? RFWI is similar to WEMVA in that it splits the model into kinematic and dynamic components, though the decomposition changes during the iteration. It is reduced to a problem over the kinematic component by VPM, just as some WEMVAs and FWIME do, and ultimately returns an approximate FWI solution, as does FWIME. The remark at the end of p. 25 suggests that FWIME is "more robust". Can you cite a specific comparison? If not it would be better to remove this assertion.

3.6 Nonlinear inversion

Example 2 illustrates FWIME inversion of reflection data. Similar examples appear in accounts of various versions of WEMVA. However even construed as an inversion method, as I did earlier, WEMVA produces a solution of the Born inversion problem, with separate velocity and reflectivity models, usually differing in scale spectrum. Thus WEMVA does not actually solve the FWI problem, whereas FWIME does.

The most obvious difference between acoustic data and its Born approximation is the (possible) presence in the former of multiple reflections. These are commonplace in field data, and are routinely suppressed in standard data processing. Such suppression of multiples is often imperfect, and renders the data unfaithful to the acoustic model. Since FWIME ultimately solves the FWI problem, fitting data via the acoustic model, it is irresistable to ask: does FWIME succeed in the presence of multiple reflections?

The data of example 2 may contain multiple reflections, though of small amplitude, since the reflection coefficients are all relatively small, and absorbing boundaries appear to have been used througout. A free surface would have produced considerably more multiple energy. Would FWIME have succeeded then?

While this question - can FWIME deal directly with multiples? - is fascinating, it opens up several aspects of inversion that this paper has not considered. In particular, source calibration is essential in interpreting multiply reflected energy. Since most sources in field use are anisotropic, this amounts to more than estimating a wavelet: radiation pattern must also be pinned down.

Still, even a "cheating" synthetic example, with a known isotropic point source and a nice free-surface multiple series, would represent a decisive advance beyond WEMVA. Can FWIME do it?

3.7 A final thought

The comments I have offered so far mostly concern clear expositional shortcomings of the current draft, and I would expect to see them addressed in a revision. My final comment is more stylistic in nature, and I offer it more as a statement of personal preference than as a requirement.

I find the material in the part of the paper between pages 12 and 30 to be redundant and mostly motivational. I suspect I am typical of the potential readers of this paper in that I don't need quite so much motivation, or "a high-level description of the main intuition" (p. 12): the ideas and eamples in this paper are sufficient motivation already.

This is a long and complicated piece of writing, so I will just pick out two examples. The section "FWIME inversion workflow" presents an fully described algorithm (Algorithm 1), then proceeds to show that it doesn't always work by a special example introduced only for this purpose. This takes five pages. The point of this discussion is the necessity of a spatial multi-scale approach, at least in some cases (but not example 1!). This point is made in the section "A model-space multi-scale approach for FWI", and the illustration could be folded in with example 2, which also does not work without multiple scales (or so the discussion suggests). My suggestion would be: get rid of Algorithm 1, which is not the algorithm ultimately used and is anyway a special case of Algorithm 2 with one spatial scale, and fold the illustration in with the example section.

Another example is equation 21 on p. 23 and surrounding discussion, which echoes the earlier subsection title "FWIME $\approx$ FWI + WEMVA". The text following the equation is an interesting description of (what the authors claim is) typical behaviour of the algorithm, in terms of the two components of the gradient. Part of this behaviour is illustrated in example 2, which shows the components of the initial FWIME search direction. Clearly the rest of the behaviour could be illustrated in by this example as well, instead of introducing a different but very similar example (Figure 12 etc.) to make the illustration. You could even get rabbit ears from Mora!

There are many other ways in which I feel that this mid-section of the paper is too repetitive and motivational, in a way that might be appropriate in a PhD thesis or a textbook, but which just gets in the way of the main message in a research paper. As stated earlier, I don't ask for specific revisions to address this issue, and am not even sure that every reader would react the same way. I do wish the authors to think about what really belongs in this exposition, and realize that unnecessary length will reduce the impact of their contribution.

References

[Cocher et al., 2017] Cocher, E., H. Chauris, and R.-E. Plessix, 2017, Seismic iterative migration velocity analysis: two strategies to update the velocity model: Computational Geosciences, **21**, 759–780.

[Cocher et al., 2018] ———, 2018, Towards a stable iterative migration velocity analysis scheme: Geophysics, **83**, R475–R495.

[Lameloise and Chauris, 2016] Lameloise, C.-A., and H. Chauris, 2016, Extension of migration velocity analysis to transmitted wavefields: Geophysical Journal International, **207**, 343–356.

[Shen, 2013] Shen, P., 2013, Subsurface focusing measurement of diving waves and its application to reflection tomography: Presented at the 75th Annual Conference, European Association of Geoscientists and Engineers.