Joint minimization of the mean and information entropy of the matching filter distribution for a robust misfit function in full-waveform inversion

Bingbing Sun, Tariq Alkhalifah

Abstract—Full-Waveform Inversion (FWI) is a highly nonlinear inversion methodology. FWI tends to converge to a local minimum rather than a global one. We refer to this phenomena as "cycle skipping" in FWI. A cost-effective solution for resolving this issue is to design a more convex misfit function for the optimization problem. A global comparison based on using a matching filter admits more robust misfit functions. In this case, we would compute a matching filter first by deconvolving of the predicted data from the measured ones. When the velocity model is accurate, the predicted data resemble the measured ones, and the resulting matching filter would be an approximated Dirac delta function. If the velocity produces data that are different from the observed ones, a misfit function can be formulated by penalizing the energy away from the zero-lag time (the center). Here, we develop a general mechanism for an evolution of the matching filter to our objective in FWI. Specifically from the statistics point of view, rather than using a penalty, we propose a novel misfit by minimizing the mean and information entropy of the matching filter distribution. We show that the resulting misfit function can mitigate the "cycle skipping" as well as reduce the mean and variance of the resulting matching filter distribution. We use a modified Marmousi example to demonstrate the features of the proposed misfit. We also evaluate the robustness of the proposed method using inaccurate (rotation in phase) source wavelets and measured data with different levels of Gaussian random noise.

Index Terms—full-waveform inversion, matching filter, information entropy, nonlinear inversion

I. INTRODUCTION

FULL waveform inversion is a high-resolution model-building tool often used to generate a subsurface model, but has applications in concrete ultrasound investigations and medical imaging among other fields [1]. Unlike conventional ray-based tomography methods, which utilize travel time information only, FWI utilizes all the information (both amplitude and phase) embedded in the seismic data, e.g., reflections, refractions, diving waves, P- and S- waves. Theoretically, FWI can invert for all the physical properties, which effect the wave propagation phenomena, such as P-and S- wave velocities, density, attenuation, and anisotropic parameters. Recently, due to advances in seismic data acquisition and high-performance computing (HPC), application of FWI to real datasets is becoming common and due to its high-resolution potential (up to half of the wavelength of the highest frequency), numerous success stories of FWI applications has been reported [2], [3], [4], [5], [6], [7], [8].

Mathematically, FWI is a highly nonlinear inverse problem. Given a current model (possibly from other velocity model building approaches), we computed what we refer to as predicted data using a forward modeling engine often approximating the physics of the true subsurface. Then, we measure the discrepancy between the predicted data and the measured ones using a misfit function such as the $l_2$ norm. Based on the adjoint state method, we can back-propagate the residual (the difference) computed from the misfit function to update the model. As with any nonlinear inversion problem, FWI faces the issue of being trapped in a local minimum rather than converging to a global one. Specifically, in FWI, if the starting frequency is not low enough or the initial velocity is far from the true one, the inversion will diverge. We refer to this phenomenon as "cycle-skipping," and it usually occurs when we use the "$l_2$ norm as the misfit function. Inversion from a relatively low frequency can possibly resolve the "cycle-skipping" issue. However the acquisition of low-frequency signals is difficult and costly.

Thus, the potential solution to the cycle-skipping issue in FWI is to use global updating (optimization) schemes, such as the Monte Carlo method [9], [10], genetic algorithms [11], [12], [13], or simulating annealing [14], [15]. However, those global optimization schemes typically require the computation of the misfit function for tens of thousands of models. This is often impractical, especially for real 3D applications. Thus, a cost-effective solution is to design a more robust misfit function, which gradually pulls the model towards a global optimum using a gradient-based updating approach.

The $l_2$ norm misfit, though its high resolution measure (sample-by-sample), is prone to severe "cycle-skipping" problems. The reason is that it forms a local comparison between the predicted and measured data. If the shift between the predicted and measured signal is larger than a half of a cycle, the $l_2$ norm would try to match different events, and thus, becomes "cycle-skipped." A remedy to this is to seek global rather than local comparison methods, such as matching filter based approaches [16], [17], [18], [19], [20] or optimal transport misfit functions [21], [22], [23], [19].

Matching filters have been widely used in signal processing [24]. In waveform inversion, a simplification of it computed based on crosscorrelation was adopted by [25]. The formal suggestion of a matching filter in waveform inversion can be attributed to [16]. The basic idea related to matching filter is
that instead of comparing the predicted data with the measured ones directly, we compute a matching filter first by deconvolving the predicted data from the measured ones. If the current model is accurate, the corresponding predicted data should be the same as the measured ones, and the resulting matching filter would reduce to a band-limited Dirac delta function focusing at zero time-lag. Otherwise, we formulate an optimization problem, such as using a penalty function to penalize the energy away from the zero-lag region of the matching filter, and update the model accordingly. Unlike the $l_2$ norm, the resulting matching filter is a global measurement of the similarity between the predicted and measured data. Thus, the resulting misfit function shows better convexity with regard to time-shifted signals, and as a result, can reduce the "cycle-skipping" issue considerably. The adaptive waveform inversion approach suggested by [17] added a normalization term to improve the convergence of the matching filter approach. This normalization term has its roots in the instantaneous traveltime objective function as demonstrated by [26]. Based on the optimal transport theory, [18], [20], [27] proposed a more general framework for focusing the resulting matching filter to be a Dirac delta function. In their method, rather than applying a penalty or a normalization term, they formulate the misfit function by measuring the Wasserstein $W$ distance between the resulting matching filter and the Dirac delta function. Based on their framework, the critical normalization term in AWI can be explained as a natural requirement for probability distributions. The mechanism of focusing the matching filter is given by reducing its distribution mean and variance. Their analysis was the inspiration for the work presented in this paper.

In this paper, rather than using a penalty, we design a robust misfit function by joint minimization of the mean and entropy of the resulting matching filter (JMME). The concepts of "information entropy" were given formal mathematical definitions in the context of data communications by Shannon in 1948 [28]. Here, we introduce entropy as a regularization term to constrain the variance of the resulting matching filter. Together with reducing the mean value, the proposed misfit will physically force the resulting matching filter to evolve to zero-time lag and at the same time become more compact. Mathematically, it achieves this by minimizing a weighted sum of the mean and logarithm of the variance of the resulting matching filter as a probability distribution. The contribution of this paper can be summarized as follows:

- We demonstrate that the resulting matching filter should be interpreted as a probability distribution rather than a normal signal. The critical information in the matching filter is given by the degree in which its coefficients are spread over the time-lag axis. Specific events in the matching filter do not have physical meanings as that in the predicted or measured data. (Except for when only single event present in the data, in this case, the event in the matching filter denotes the time-shift lag.)
- The focusing mechanism of the resulting matching filter lies in reducing its mean and variance. As long as this is satisfied, any matching-filter based misfit function including AWI and JMME (proposed in this paper), can lead to a robust misfit function for mitigating cycle-skipping.
- Since reducing the mean and variance of the resulting matching filter is a global updating approach, the corresponding misfit function shows improved robustness, e.g., reduced requirement for an accurate source wavelet.

In the following, we will first review the FWI method and the conventional matching filter approach. Later we analyze the mechanism of focusing for the matching filter during inversion. After we discuss the proposed misfit and an analysis of its properties, we verify our method using a modified Marmousi example.

II. FULL WAVEFORM INVERSION AND MATCHING FILTER

A. Conventional full-waveform inversion using $l_2$ norm

For an isotropic acoustic FWI, we assume, for simplicity that, the propagation of seismic waves in the subsurface satisfies the following wave equation:

$$\frac{1}{v^2(x)} \frac{\partial^2}{\partial t^2} u(t, x_s, x) = \Delta u(t, x_s, x) + f(t) \delta(x - x_s),$$  \hspace{1cm}(1)$$

where $v(x)$ denotes the P-wave velocity at position $x$, $u(t, x_s, x)$ is the wavefield, $\Delta$ denotes the Laplacian operator. $x_s$ is the source position and $f(t)$ is the source signature. For a given acquisition, we sample the wavefield $u(t, x_s, x)$ with a sampling matrix $S_{s,r}$ to obtain the predicted data $p(t) = p(t, x_s, x) = S_{s,r} u(t, x_s, x)$ (We denote it as $p(t)$ and omit the notation of $x_s$ and $x$ in this paper for simplicity).

We can formulate a misfit function, which characterizes the discrepancy between the predicted data $p(t)$ and the measured ones $d(t)$, for example, using the $l_2$ norm:

$$J = \frac{1}{2} \int (p(t) - d(t))^2 dt.$$  \hspace{1cm}(2)$$

The adjoint source, i.e., the derivative of the misfit $J$ with respect to the predicted data $p$, is given by:

$$\delta s = p(t) - d(t).$$  \hspace{1cm}(3)$$

Gradient based optimization techniques can be utilized to solve FWI by back propagating of adjoint source in equation 3. As the seismic data signal is oscillatory, the resulting misfit function from equation 2 given by $l_2$ norm shows strong nonlinearity with regards to the time shift, and thus, is prone to "cycle-skipping". A global comparison rather than a local subtraction can reduce the nonlinearity and as a result reduce the "cycle-skipping" issue.

B. Full-waveform inversion using matching filter

In order to formulate a global comparison, we compute a matching filter $w(t)$ first by deconvolving of the predicted data $p(t)$ with the measured data $d(t)$:

$$d(t) * w(t) = p(t),$$  \hspace{1cm}(4)$$
where * denotes the convolution operation. Equation 4 is a linear equation, and the matching filter can be computed in the frequency domain as:

$$w(t) = \mathcal{F}^{-1}\left[ \mathcal{F}[p(t)]\mathcal{F}[d(t)] / \mathcal{F}[d(t)]\mathcal{F}[d(t)] + \epsilon \right],$$

(5)

where \(\mathcal{F}\) and \(\mathcal{F}^{-1}\) denote the Fourier transform and its inverse, respectively. Overline denotes the complex conjugation operation, and \(\epsilon\) is a small positive number to avoid dividing over zero. In our applications, we compute the maximum value of \(\mathcal{F}[d(t)]\mathcal{F}[d(t)]\) and set \(\epsilon\) to be 10% of this maximum value to achieve a stable inversion of the examples shown in this paper.

[16] proposed a matching filter (MF) based misfit function that penalizes energy beyond the zero lag of the resulting matching filter \(w(t)\):

$$J_{MF} = \int A^2(t)w^2(t)dt,$$

(6)

where \(A(t)\) is a penalty function, e.g., \(A(t) = |t|\). The penalty-based matching filter approach of equation 6 shows improved convexity property with respect to a time shift between the predicted and observed data and thus, reduces the “cycle-skipping” issue. It has another issue of being sensitive to the amplitude of the matching filter (as well as the predicted data due to relationship in equation 5). As demonstrated in Figure 1, if the matching filter \(w(t)\) (Figure 1a) evolves to \(\lambda w(t)\) (Figure 1b), where \(\lambda < 1\), the misfit value using equation 6 will drop although the focusing property of \(\lambda w(t)\) remains the same compared to the original matching filter \(w(t)\). Considering equation 4, which relates the matching filter to the predicted data, the inversion based on equation 6 may try to decrease the amplitude of the predicted data and potentially lead to no reflections in the predicted data. In order to guarantee that the resulting matching filter in Figure 1a evolves to the ideally focused one in Figure 1c, we need to consider the process of forcing the resulting matching filter to be a Dirac delta function as an evolution of probability distributions, and this would help in mitigating the amplitude sensitivity issue.

For a probability density function, it must obey 1) the nonnegative and 2) the mass conservation rules. In order to fulfill this requirement, we choose to square and normalize the resulting matching filter \(w(t)\) as follows:

$$\tilde{w}(t) = \frac{w^2(t)}{\int w^2(t)dt} = \frac{w^2}{||w||^2}.\quad(7)$$

If we apply the penalty to the resulting matching filter distribution \(\tilde{w}(t)\), we actually obtain a misfit similar to Adaptive Waveform Inversion (AWI) [17]:

$$J_{AWI} = \int t^2\tilde{w}(t)dt.\quad(8)$$

Likewise, considering the example shown in Figure 1, the resulting AWI misfit of equation 8 would be the same for the original matching filter \(w(t)\) and its scaled one \(\lambda w(t)\), and this will force the matching filter to focus rather than reduce in amplitude during inversion. The normalization term is emphasized by the authors of AWI [17]: “Without some form of normalization, the misfit function could be minimized simply by decreasing all the coefficients of the matching filter \(w\). This would be equivalent to decreasing all the amplitudes within the predicted data trace \(p\), for example, by decreasing all reflection coefficients in the model. This is not a behavior that will move toward the true earth model, and so normalization of the objective function is always necessary in some form to obtain a practical AWI algorithm.” However, we introduced the normalization from the perspective of the probability distribution, and this would help us in understanding the mechanism of the focusing of the matching filter.

C. the mechanism of the focusing of the matching filter

The resulting AWI misfit of equation 8 is the second order moment of the probability distribution \(\tilde{w}(t)\). It can be easily verified that, it equals to the sum of the square of the mean and variance, i.e.,

$$J_{AWI} = \int t^2\tilde{w}(t)dt = \mu^2 + \sigma^2,$$

(9)

here \(\mu\) is the mean of the matching filter distribution \(\tilde{w}(t)\)

$$\mu = \text{Mean}(\tilde{w}(t)) = \int t\tilde{w}(t)dt,$$

(10)

and \(\sigma^2\) is the variance

$$\sigma^2 = \text{Var}(\tilde{w}(t)) = \int (t - \mu)^2\tilde{w}(t)dt.\quad(11)$$

As demonstrated in Figure 2, the purpose of focusing on the resulting matching filter distribution is actually trying to reduce the mean value and the variance of the resulting matching filter. Suppose the resulting matching filter is approximated by a Gaussian distribution such as the one represented by the black curve in Figure 2. The mean value \(\mu\) would control the position of the matching filter distribution in the time-lag axis, while the variance \(\sigma^2\) would control its compactness (width). Focusing of the matching filter aims at making the resulting matching filter distribution move toward the position of zero time-lag while becoming more compact (with limited support).

III. METHODOLOGY

In the previous section, we have described that due to the limitation of the conventional \(l_2\) norm for “cycle-skipping”, the matching filter approach is introduced as a global comparison method for reducing the nonlinearity. The conventional penalty-based matching filter approach has a drawback of being sensitive to the amplitude, and thus, a normalization term should be incorporated. The evolution of the resulting matching filter should be considered as the evolution of a probability distribution, i.e., reducing the mean and variance values.

We suggest to use information entropy to replace the variance to control the compactness of the resulting matching filter:

$$\text{Entropy}(\tilde{w}(t)) = - \int \tilde{w}(t)\log(\tilde{w}(t))dt.\quad(12)$$

The information entropy would attain its maxima when \(\tilde{w}(t)\) is uniformly distributed, and achieve its minima when it is
compact. In the Appendix, we prove that for a Gaussian distribution with mean $\mu$ and variance $\sigma^2$, the Entropy is:

$$\text{Entropy}(\mu, \sigma^2) = \frac{1}{2} \left[ \ln(2\pi\sigma^2) + 1 \right]$$

(13)

Thus, we can see that the entropy performs the function of measuring the variance of the resulting matching filter, and thus, we can suggest the following misfit function by joint minimizing of the mean and entropy (JMME) of the resulting matching filter distribution:

$$J_{\text{JMME}} = \left[ \text{Mean}(\hat{w}(t)) \right]^2 + \lambda \text{Entropy}(\hat{w}(t))$$

$$= \left[ \int t\hat{w}(t)dt \right]^2 - \lambda \int \hat{w}(t) \log(\hat{w}(t)) dt,$$

(14)

where $\lambda$ is the weighting parameter, which balances the minimization between the mean and entropy terms. Compared to the misfit function of AWI in equation 9, both AWI and JMME misfits apply to the matching filter distribution $\hat{w}(t)$ rather than the matching filter itself $w(t)$, and both have terms related to the mean value of the distribution. However, the compactness constraint is different in affecting the variance.

The adjoint source for $J_{\text{JMME}}$ is given by:

$$\delta s = \left[ \frac{\partial \hat{w}}{\partial \phi} \right]^T \left[ \frac{\partial \hat{w}}{\partial \omega} \right]^T \frac{\partial J_{\text{JMME}}}{\partial \hat{w}},$$

(15)

where we interpret the derivatives as Jacobi matrices and $T$ denotes the matrix transpose operator. $\partial J_{\text{JMME}}/\partial \hat{w}$ is the derivative of the misfit with respect to the matching filter distribution, and can be computed from equation 14:

$$\frac{\partial J_{\text{JMME}}}{\partial \hat{w}} = 2t \int t\hat{w}(t)dt - \lambda [1 + \log(\hat{w}(t))].$$

(16)

The term $\partial \hat{w}/\partial w$ is the derivative of the matching filter distribution with respect to the original matching filter, and it can be computed from equation 7:

$$\left[ \frac{\partial \hat{w}}{\partial w} \right]^T = 2 \frac{\text{diag}(w(t) - w(t)\hat{w}(t))^T}{w(t)^T(w(t))}.$$  \hspace{1cm} (17)

The term $\partial \hat{w}/\partial p$ is the derivative of the matching filter with respect to the predicted data, and it can be computed using the adjoint analysis of equation 5:

$$\left[ \frac{\partial \hat{w}}{\partial p} \right]^T = F^* \text{diag} \left[ \frac{\mathcal{F}[d(t)]}{\mathcal{F}[d(t)] + \epsilon} \right] \mathcal{F}.$$  \hspace{1cm} (18)

For a detailed derivation of those derivatives for the adjoint source, we refer you to Appendix B.

IV. EXPERIMENTAL RESULTS

A. dataset description

We use a modified Marmousi example to verify the effectiveness of the proposed JMME misfit. The model is shown in Figure 3a extending 2 km in depth and 8 km horizontally. We modeled data using 80 shots with a spacing of 100 m, and 400 receivers placed at the surface with interval 20 m. We use an absorbing boundary condition at the surface, and the source is a Ricker wavelet with a peak frequency of 10 Hz. Figure 3b is a $v(z)$ model which is relatively far away from the true model shown in Figure 3a and used as one of the initial models for evaluating the "cycle skipping". This initial model is similar to that in [17], i.e., "it increases approximately linearly from 1550 m/s at the seabed to 2550 m/s at 1000 m depth, then less slowly to reach a velocity of 3250 m/s at the base of the model". In Figure 4, we show the profiles at different positions. We can see that in the left part of the model, the initial model is relatively far from the true one and this will lead to "cycle-skipping" at the current frequency band as we show later. Another initial velocity is shown in Figure 14. It is produced by Gaussian smoothing of the true velocity model in Figure 3a. The standard deviation of the Gaussian smoothing windows is selected to be 200 m for both the vertical and horizontal directions. As it is a smoothed version of the true model and thus relatively close to the true one, we use it as the initial model for evaluating the robustness of the schemes with respect to inaccurate sources and noise. Figure 5 shows the record at distance 2 km using the true model and the $v(z)$ initial model (Figure 3b), respectively.

B. Cycle-skipping experiments

In the inversion, to verify the ability of the misfits in overcoming "cycle-skipping," we do not perform frequency continuation, i.e., from low to high frequencies. We band-pass filter the data in the frequency range 3 to 10 Hz and perform a full band inversion. We use nonlinear conjugate gradient schemes for the optimization and a quadratic line-search scheme. We apply smoothing to the gradient and do not include any regularization such as TV [29] in this inversion process.

Figure 6 shows the inverted result by the $l_2$ norm, AWI and the proposed JMME misfits, respectively, after 200 iterations. The JMME result presented in Figure 6c corresponds to a weighting $\lambda = 0.01$ which provided a good result. It is obvious that the result from the $l_2$ norm is highly cycle-skipped and geologically meaningless in general. On the other hand, both AWI and JMME misfits lead to a meaningful result and compared to the true model, both the background low wavenumber and fine-scale high wavenumber components are well recovered. It is reasonable to have degenerated results at the boundary due to limited illumination. Considering the challenging situation we set up for the experiments (raw $v(z)$ initial model and relatively high peak frequency of 10 Hz), the converging results by AWI and JMME are courtesy of using the matching filter and the application of controls on its mean and variance over iterations. We can also notice that the result of JMME in Figure 6c is less noisy than the result in Figure 6b by AWI. To further demonstrate the accuracy of the inverted models, we show the modeled record at source location 2 km in Figure 7 for the inverted models and their difference with the true record in Figure 8. We can easily observe that both AWI and JMME records resemble the true record well while the $l_2$ norm record hardly matches the true one due to cycle-skipping. Figure 9 shows the $l_2$ residuals over iterations for different misfit functions. We can see that the $l_2$ norm misfit drops fast initially then it gets stuck in a local minimum leaving it with relatively large residuals. While the
AWI and JMME methods drops gradually leading to relative small residuals. We observe that JMME is slightly better than AWI in reducing the residuals over iterations.

For comparison, in Figure 10, we show the inverted result with the weighting parameter $\lambda$ smaller than what we deem as optimal, used in the previous inversion, and in Figure 11 for larger weighting parameters. From the results, we conclude that the mean value of the resulting matching filter controls the low wavenumber background of the model (see Figure 10a, the weighting parameter is rather small, and in this case, the entropy can be set to zero throughout the inversion. Thus, a sole mean term can still recover the low wavenumber background information). The variance controlled by entropy will effect the resolution of the resulting model (From Figure 10a to Figure 11b, with increasing weighting parameters, the entropy terms become more dominant, and the resulting inverted models show higher resolution gradually). In FWI, the low wavenumber components are highly important as it mainly controls the travel times of the wave propagation. As reported by [30],[31], applying the matching filter with constraints only on the mean in the Radon transform, we can recover a reasonable good Marmousi model as well. However, here we compute the matching filter and implement the inversion in the time-space $t-x$ domain rather than Radon $\tau-p$ domain. Figure 12 shows the $l_2$ norm of the residual over iterations for the JMME misfit with different weighting parameters. Too large or too small a weighting parameter $\lambda$ hampers the inversion while a proper weighting between the mean and the entropy will lead to a robust FWI with minimum residuals.

C. Inaccurate source wavelet experiments

In previous experiments, we assumed the source signature (wavelet), i.e., $f(t)$ in equation 1, is known and definite. In real applications of FWI, the source signature must be inverted (or provided), and often its phase or amplitude are not accurate. Thus, in order to evaluate the robustness of proposed JMME misfit, we invert the Marmousi model with a rotated source wavelet as shown in Figure 13. We rotate the original 10 Hz Ricker wavelet (in balck) 45 degrees (in blue) and 90 degrees (in red), respectively. We use a relatively accurate initial model as shown in Figure 14 and the parameters for the other setup are the same as before. In general, the $l_2$ norm results in Figure 15 show low resolution compared to the results in Figure 16 using JMME misfit, and due to the local comparison nature, the $l_2$ norm is relatively more sensitive to the accuracy of the source wavelet. When a 90 degrees rotated wavelet is used, the inverted model shows rather low resolution demonstrating poor updating. Meanwhile the result from JMME in Figure 16 shows convergence to a high resolution model when the phase is rotated up to 90 degrees. We further setup the experiments in a more challenging situation, i.e., with the $v(z)$ initial model in Figure 3b for which the "cycle-skipping" is an issue using conventional $l_2$ norm. We only show the result after 200 iterations for the JMME misfit in Figure 17, and the result by $l_2$ norm misfit is cycle-skipped and similar to the inverted model shown in Figure 6a. We can see that, even in this challenging situation, JMME still performs well and can still invert for a geologically meaningful model similar to that in Figure 6c inverted using an accurate source. In order to further evaluate the effect of the inaccuracy of the source wavelet on the results, we plot records of the measured and predicted data as well as the matching filter distribution over iterations for the JMME method in Figures 18-20. Figures 18 and 19 correspond to using the correct source wavelet. We can see that the matching between the measured and predicted record is good and the resulting matching filter distribution evolves to a band-limited delta function over iterations gradually. Figures 20 and 21 corresponds to using the 45-degree rotated wavelet. We can see that though the source wavelet is wrong, the predicted data can still fits the kinematic part of the measured data. In Figures 22 and 23, the results using the 90-degree rotated wavelet show similar features. These demonstrate that due to the global updating features of the matching filter approach, JMME can focus on updating the kinematic parts of the model mainly and can reduce the requirement for an accurate source wavelet. In some sense, such a reduced requirement on the source wavelet suggests that JMME has less requirement for predicting the amplitude in the data, and thus, it focuses more on the phase of the signal of the data.

D. Noisy dataset experiments

The following examples aim at testing the behavior of the schemes when noise is introduced to the data. The frequency band used in these noisy dataset experiments is the same as before, i.e., 3 Hz to 10 Hz. We use an additive random Gaussian noise model, i.e, the record with noise $d_n(t)$ corresponds to the addition of the noise-free record $d(t)$ and the random noise $n(t)$:

$$d_n(t) = d(t) + n(t).$$

To mimic different noise levels, we set the Gaussian random noise $n(t)$ with zero mean and different variances. In Figure 24, we show the record in Figure 5a with different levels of noise, i.e., variance equals to 0.5, 2.5, 5. We first use the smooth initial model shown in Figure 14 to start the inversion using noisy dataset. After 50 iterations, we obtain the results in Figures 25 and 26, respectively. Due to the added noise, the result by the $l_2$ norm is not affected by the noise and shows strong robustness. While the JMME shows slight sensitivity, as the noise level increases, the results become slightly degenerated at the deeper part of the model. The reason is that JMME uses matching filters rather than the original seismic data to formulate the misfit. When adding noise to the dataset, the deeper reflections as well as the signals at far offsets are highly effected because of its relatively small amplitude. We further utilize the $v(z)$ initial model shown in Figure 3b and the noisy record to evaluate the JMME misfit. (Again, we do not show the $l_2$ norm result as it is cycle-skipped similar to the inverted model in Figure 6a). The results are as we expected, with increasing noise levels, the inverted model is less accurate in the deeper part. However, the inverted model still provides reasonable information at the shallow part and the inverted model is also geologically meaningful, no "cycle-skipping" artifacts similar to that in Figure 6a using the $l_2$ norm.
V. DISCUSSIONS

Both the AWI and JMME approaches use a deconvolution step to compute the matching filter. As the deconvolution is implemented in the Fourier domain for each trace. This process can be made efficient by parallel programming, such as using OPENMP. This makes the matching filter computation step almost negligible compared to the most time-consuming part of the inversion in solving the PDEs for the forward and backward propagating the wavefields. Compared to AWI, JMME has an extra parameter \( \lambda \) that we need to determine. According to our experiments, the parameter \( \lambda \) is easy to select and for the examples present in the paper, we only tried limited number of \( \lambda \) to obtain an optimal value for inversion. If we apply JMME to different models, we might use the optimal parameter \( \lambda \) for the Marmousi model directly or have a few trials nearby the optimal value. In the examples shown in the paper, we fixed the parameter \( \lambda \) during inversion. For testing, we tried to vary the value of \( \lambda \) during inversion. For example, we set \( \lambda \) to a small number initially and gradually increase it over iterations. Our tests demonstrated to us that a reduction in variance is also important at the initial stages of the inversion. The desired matching filter objective admits a global fitting between the predicted and measured data. If the predicted and measured data are simple, e.g., a single event, the misfit with only the mean term is enough. However, if the data are complex with multiple events (which is usually the case), cross-talks between those events will exist and this would increase the variance of the resulting matching filter (for example, the deconvolution between the data with two events will lead to four events, which will have a large variance). Thus, for complex media like the Marmousi model, we need to minimize the mean and variance jointly to push the resulting matching filter to a delta function. An inversion scheme that focuses only on the mean or variance will be incomplete, especially when we seek a high resolution model of the subsurface.

VI. CONCLUSION

We proposed a misfit function by minimizing of the mean and information entropy of the resulting matching filter computed by deconvolving of the predicted from the measured data. Similar to AWI, the proposed misfit function tries to reduce the mean and variance of the resulting matching filter distribution. Staring from a \( v(z) \) initial model and a dataset without frequencies below 3 Hz, the proposed misfit function can invert for a geologically meaningful model, with both the low-wavenumber background and the high-wavenumber fine-scale parts of the model well recovered. Due to the global updating nature of this approach, the resulting misfit function demonstrates strong robustness with regard to an erroneous source signature. Compared to the \( l_2 \) norm, though the proposed misfit function shows sensitivity to noise added to the data, it can still, however, update the model free of "cycle-skipping" even when moderate noise are present.

ACKNOWLEDGMENT

The authors would like to thank KAUST for funding this research and the members of SWAG for useful discussions.

APPENDIX A

ENTROPY OF A GAUSSIAN DISTRIBUTION

Below, we show that the entropy given by Equation 12 computes the logarithm of the variance of a Gaussian distribution. The Gaussian probability density function \( N(x; \mu, \sigma) \) is defined as:

\[
N(x; \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{20}
\]

All probability density functions must be normalized to unity, thus, we have:

\[
\int N(x; \mu, \sigma) dx = 1. \tag{21}
\]

The expectation of \( (x-\mu)^2, E(x-\mu)^2 \), i.e, the variance can be evaluated as:

\[
E(x-\mu)^2 = \int (x-\mu)^2 N(x; \mu, \sigma) dx = \sigma^2. \tag{22}
\]

Using Equation 12, the entropy of the Gaussian distribution can be computed as:

\[
\text{Entropy}(\mu, \sigma^2) = - \int \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\mu)^2/(2\sigma^2)} \ln \left[ \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\mu)^2/(2\sigma^2)} \right] dx
\]

\[
= \frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \int (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)} dx
\]

\[
+ \frac{1}{2\sigma^2} \int (2\pi\sigma^2)^{-1/2} (x-\mu)^2 e^{-(x-\mu)^2/(2\sigma^2)} dx. \tag{23}
\]

Using equations 21 and 22, the entropy of the Gaussian distribution can be written as:

\[
\text{Entropy}(\mu, \sigma^2) = \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sigma^2 = \frac{1}{2} [\ln(2\pi\sigma^2) + 1]. \tag{24}
\]

This form of the entropy reveals that the entropy of a Gaussian distribution equals to the logarithm of the variance plus a constant value.

APPENDIX B

DERIVATION OF THE ADJOINT SOURCE FOR THE JMME MISFIT FUNCTION

Following the chain rule given by equation 15, we compute the adjoint source for the JMME misfit function. According to equation 15, there are three terms, i.e., the derivative of the misfit with respect to the matching filter distribution \( \partial J_{JMME} / \partial w \), the derivative of the matching filter distribution with respect to the matching filter \( (\partial \tilde{w} / \partial w) \) and the derivative of the matching filter with respect to the predicted data \( (\partial \tilde{w} / \partial \tilde{u}) \). The first and the second terms are relatively easy to obtain. For example, for the computation of \( \partial J_{JMME} / \partial \tilde{w} \), we refer to equation 14 and have:

\[
\frac{\partial J_{JMME}}{\partial \tilde{w}} = \frac{\partial}{\partial \tilde{w}} \left[ \left( \int \tilde{w}(t) dt \right)^2 - \lambda \int \tilde{w}(t) \log(\tilde{w}(t)) dt \right]
\]

\[
= 2 \left( \int \tilde{w}(t) dt \right) \frac{\partial}{\partial \tilde{w}} \left( \int \tilde{w}(t) dt \right) - \lambda \int \frac{\partial}{\partial \tilde{w}} [\tilde{w}(t) \log(\tilde{w}(t))] dt
\]

\[
= 2 \left( \int \tilde{w}(t) dt \right) - \lambda [1 + \log(\tilde{w}(t))] \tag{25}
\]
Using equation 5, we can easily obtain:
\[
\frac{\partial w}{\partial p} = F^* \text{diag} \left[ \frac{F[d(t)]}{F[d(t)]F[d(t)] + \epsilon} \right] F. \tag{26}
\]
Thus, the adjoint operator becomes:
\[
\left[ \frac{\partial w}{\partial p} \right]^T = \left[ F^* \text{diag} \left[ \frac{F[d(t)]}{F[d(t)]F[d(t)] + \epsilon} \right] F \right]^T
= F^T \text{diag} \left[ \frac{F[d(t)]}{F[d(t)]F[d(t)] + \epsilon} \right] F \left[ F - \epsilon \right]^T. \tag{27}
\]
Here, we utilize the property of Fourier transform that the adjoint operator of the Fourier transform is the same as the inverse operator, i.e., $F^T = F^*$. The derivative of the matching filter distribution $\hat{w}$ with respect to the matching filter $w$ is slightly complicated. In order to facilitate the derivation, we express equation (7) in discrete form:
\[
\hat{w}_i = \frac{w_i^2}{\sum_j w_j^2}. \tag{28}
\]
Thus, the corresponding discrete form of $\frac{\partial \hat{w}}{\partial w}$ can be expressed as:
\[
\frac{\partial \hat{w}}{\partial w} = \frac{2w_i \delta_{ij} - \sum_j w_j^2 2w_i}{\sum_j w_j^2}. \tag{29}
\]
We can further express it in matrix form as:
\[
\frac{\partial \hat{w}}{\partial w} = \frac{2}{w(t)^T w(t)} \left[ \text{diag}(w(t)) - w(t)w(t)^T \right]. \tag{30}
\]
Its adjoint, thus, becomes:
\[
\left[ \frac{\partial \hat{w}}{\partial w} \right]^T = \frac{2}{w(t)^T w(t)} \left[ \text{diag}(w(t)) - w(t)w(t)^T \right]^T
= \frac{2}{w(t)^T w(t)} \left[ \text{diag}(w(t)) - w(t)w(t)^T \right]^T
= \frac{2}{w(t)^T w(t)} \left[ \text{diag}(w(t)) - w(t)w(t)^T \right]^T.
\]
\[
\text{REFERENCES}
\]
[27] ——, “The application of an optimal transport to a preconditioned data matching function for robust waveform inversion,” *Geophysics*, vol. 84, no. 6, pp. no. 6, R923–R945, 2019.
Fig. 1: A cartoon demonstration of the evolution of the matching filter: The energy distribution for a) an example matching filter \( w(t) \); b) the scaled matching filter \( \lambda w(t) \) with \( \lambda < 1 \); c) the ideal matching filter with its energy focused at zero-lag. The conventional penalty-based matching-filter approach may not drive the matching filter shown in plot a) to evolve towards the ideal one in plot c), rather it may try to decrease its amplitude at all time-lags and lead to the solution in b), which definitely shows no improved focusing compared to a). We need to normalize the resulting matching filter and consider the evolution of its probability distribution to resolve this issue.
Fig. 2: A demonstration of the focusing mechanism of the matching filter. Suppose the matching filter is approximated by a Gaussian distribution. The ideal one and our target matching filter is given by the red curve, which is an approximated Dirac delta function with zero mean and relatively small variance. If the current velocity is inaccurate, the resulting matching filter can be a Gaussian distribution with a relatively large mean \( \mu = 1 \) and variance \( \sigma^2 = 0.5^2 \). The focusing of the matching filter given by equation 9 is actually trying to reduce the mean and variance of the resulting matching filter at the same time, i.e., evolve from the black curve to the red one (the mean value \( \mu \) changes from 1 to 0 and the variance \( \sigma^2 \) changes from 0.5^2 to 0.1^2).

Fig. 3: a) The true velocity and b) the initial velocity for the modified Marmousi model.

Fig. 4: Profile for the true (Figure 3a) and the initial velocity (Figure 3b) at position a) 2 km; b) 4 km; and c) 6 km. Note the black solid lines denote the velocity for the true model while the red dashed line for the initial model.
Fig. 5: a) The measured (true) record and b) the initial record for the modified Marmousi model.

Fig. 6: The inverted modified Marmousi model using a) $l_2$ norm; b) AWI; c) JMME. For JMME, the regularization parameter $\lambda = 0.01$. 
Fig. 7: The record corresponding to the inverted models using a) $l_2$ norm; b) AWI; c) JMME, which are shown in Figure 5. For JMME, the regularization parameter $\lambda = 0.01$.

Fig. 8: The difference between the true record and those shown in Figure 7 corresponding to a) $l_2$ norm; b) AWI; c) JMME. For JMME, the regularization parameter $\lambda = 0.01$. The scale for plotting is set to be same for a fair comparison.
Fig. 9: The $l_2$ norm residual over iterations for different misfit functions.

Fig. 10: The inverted modified Marmousi model using JMME with the weighting values a) $\lambda = 1.0\times10^{-4}$; b) $\lambda = 1.0\times10^{-3}$.

Fig. 11: The inverted modified Marmousi model using JMME with the weighting values a) $\lambda = 1$; b) $\lambda = 10$.

Fig. 12: The $l_2$ norm residual over iterations for the JMME misfit with different weighting parameters $\lambda$. Note $\lambda = 0.01$ is the optimal parameter giving minimum residual.
Fig. 13: The original 10 Hz Ricker wavelet (black curve) and we phase rotate it 45 degrees (blue curve) and 90 degrees (red curve), respectively, to obtain an inaccurate source wavelet for testing the robustness of the misfits in following experiments.

Fig. 14: A smoothed Marmousi velocity as the initial model for the inversion. It is obtained by applying a Gaussian smoothing (variance equals 200 m) to the true model in Figure 3a.

Fig. 15: The inverted Marmousi model by the $l_2$ norm misfit using the initial model in Figure 14 and a) the correct, b) the 45 degrees and c) the 90 degrees phase rotated wavelet.
Fig. 16: The inverted Marmousi model by JMME misfit using the initial model in Figure 14 and a) the correct wavelet, b) the 45 degrees and c) the 90 degrees phase rotated wavelet.

Fig. 17: The inverted Marmousi model by JMME misfit using the $v(z)$ initial model in Figure 3b and a) the 45 degrees and b) the 90 degrees phase rotated wavelet. Compared to the result using the accurate source wavelet in Figure 6c, due to the global updating features of the matching filter, the results, by the JMME misfit function using an inaccurate wavelet, still maintain accuracy of the inversion.
Fig. 18: The record comparison of the measured data and predicted data. The final inverted model corresponds to Figure 6c. In this setup, the source wavelet is correct. We show the comparison at iterations a) 1; b) 10; c) 50; d) 199.

Fig. 19: The matching filter distribution (computed by equation 7) at different iterations. The final inverted model corresponds to Figure 6c. In this setup, the source wavelet is correct. We show the matching filter distribution at iterations a) 1; b) 10; c) 50; d) 199.
Fig. 20: The record comparison of the measured data and predicted data. The final inverted model corresponds to Figure 17a. In this setup, the source wavelet is rotated 45 degrees. We show the comparison at iterations a) 1; b) 10; c) 50; d) 199.

Fig. 21: The matching filter distribution (computed by equation 7) at different iterations. The final inverted model corresponds to Figure 17a. In this setup, the source wavelet is rotated 45 degrees. We show the matching filter distribution at iterations a) 1; b) 10; c) 50; d) 199.
Fig. 22: The record comparison of the measured data and predicted data. The final inverted model corresponds to Figure 17b. In this setup, the source wavelet is rotated 90 degrees. We show the comparison at iterations a) 1; b) 10; c) 50; d) 199.

Fig. 23: The matching filter distribution (computed by equation 7) at different iterations. The final inverted model corresponds to Figure 17b. In this setup, the source wavelet is rotated 90 degrees. We show the matching filter distribution at iterations a) 1; b) 10; c) 50; d) 199.
Fig. 24: The record added with Gaussian random noise. The noise has zero mean and variances equal to a) 0.5, b) 2.5 and c) 5.

Fig. 25: The inverted Marmousi model by the $l_2$ norm misfit using the initial model in Figure 14 and the dataset added with Gaussian random noise. The noise has zero mean and variances equal to a) 0.5 (Figure 24a), b) 2.5 (Figure 24b) and c) 5 (Figure 24c).
Fig. 26: The inverted Marmousi model by JMME norm misfit using the initial model in Figure 14 and the dataset with Gaussian random noise. The noise has zero mean and variances equal to a) 0.5 (Figure 24a), b) 2.5 (Figure 24b) and c) 5 (Figure 24c).

Fig. 27: The inverted Marmousi model by JMME misfit using the $v(z)$ initial model in Figure 3b and the dataset with added Gaussian random noise. The noise has zero mean and variances equal to a) 0.5 (Figure 24a), b) 2.5 (Figure 24b) and c) 5 (Figure 24c).