

## A deep-learning inverse Hessian preconditioning for iterative least-squares reverse time migration

### Introduction

Least-squares reverse time migration (LSRTM) provides true amplitude and high-resolution reflectivity images from pre-stack seismic data (Nemeth et al., 1999). Conventional applications formulate LSRTM as a linearized waveform inversion problem where the Hessian operator, related to the seismic resolution and subsurface illumination, is implicitly inverted through iterations. However, due to limitations in the forward modelling operator, noise in the data and a non-trivial null space, iterative LSRTM presents slow convergence, translating into high computational costs.

Multiple formulations have been proposed to alleviate the computational aspect of LSRTM in both data and image domains (e.g. Tang, 2009; Liu et al., 2013). This study presents a novel deep-learning-based preconditioning strategy for LSRTM consisting of two distinct building blocks. The first component estimates the effect of the inverse Hessian by training a convolutional neural network (CNN) from pairs of migrated images. For this, we introduce a convolutional autoencoder (CAE) with a 1D lower-rank representation to benefit from its dimensionality reduction capabilities later. The technique only requires paired training samples obtained from the available seismic data and the action of the physical operators, circumventing the need for a representative dataset of ground-truth reflectivity labels. After training the CAE that approximates the inverse Hessian, the second component solves LSRTM directly on a lower-dimensional space by integrating the decoder in the optimization problem through a (non-linear) change of variables. Since the decoder learns to synthesize model realizations from low-dimensional representations of the discrepancy between high-fidelity and low-fidelity images, it has a preconditioning effect that can be used to enhance model features related to images with reduced artifacts and more illumination balance. Combining the deep-learning inverse Hessian with LSRTM potentially overcomes the resolution limitations of single-iteration matching filtering. Also, given the reduced number of inferred parameters and the fast generation of enhanced model realizations, the deep-learning parameterization significantly improves the LSRTM inversion performance.

### Method

Defining  $\mathbf{L}$  and  $\mathbf{L}^T$  as the demigration and migration operators, we can reproduce the effect of the Hessian by the sequence

$$\mathbf{L}^T \mathbf{L} \mathbf{m}_1 = \mathbf{m}_2, \quad (1)$$

with  $\mathbf{m}_1 = \mathbf{L}^T \mathbf{d}_{\text{obs}}$  denoting the migration image, and  $\mathbf{d}_{\text{obs}}$  the observed data. As  $\mathbf{m}_1$  and  $\mathbf{m}_2$  are known, a crude approximation of the inverse Hessian can be found by defining a non-stationary convolutional operator  $\mathbf{P} \approx \mathbf{H}^{-1}$  and minimizing the cost function  $E(\mathbf{P}) = \|\mathbf{m}_1 - \mathbf{P}\mathbf{m}_2\|_2^2 + \lambda R(\mathbf{P})$ , where  $R(\mathbf{P})$  is a regularization term and  $\mu$  is the trade-off parameter. After solving for  $\mathbf{P}$ , we obtain an improved image through a single-step filtering,  $\hat{\mathbf{m}} = \mathbf{P}\mathbf{m}_1$ , or by incorporating  $\mathbf{P}$  into an iterative scheme to precondition the data misfit gradient at each iteration.

Alternatively, we propose parameterizing the sought preconditioning operator by the weights of a deep CAE. A CAE comprises the deterministic pair  $(E_\phi, D_\theta)$  of CNNs, where  $E_\phi : \mathbb{R}^n \rightarrow \mathbb{R}^h$  denotes the encoder network parameterized by weights  $\phi$ ,  $D_\theta : \mathbb{R}^h \rightarrow \mathbb{R}^n$  corresponds to the decoder network parameterized by weights  $\theta$ , and  $h$  is the latent space dimension. For  $n > h$ ,  $E_\phi$  is trained to encode samples  $\mathbf{x} \in \mathbb{R}^n$  in the lower-dimensional space  $\mathbb{R}^h$  such that  $D_\theta$  can reconstruct an estimated sample  $\hat{\mathbf{x}}$  from its latent representation  $\mathbf{z} \in \mathbb{R}^h$  through a reverse mapping. In this case, the latent representation vector of fixed dimensions  $\mathbf{z} = E_\phi(\mathbf{x})$  is an informational bottleneck, which induces the CAE to capture the most important features of the input sample. Generally, CAEs are trained with unsupervised algorithms so that the output samples approximate the inputs based on the latent representation. The inputs are selected from a representative training dataset, which might not be available for computationally consuming tasks such as LSRTM, where the true reflectivity is unknown. We partially overcome this issue by relying on the migration and re-migration approach. Based on equation 1, we re-write the inverse Hessian approximation problem as  $\mathbf{m}_1 = D_\theta(E_\phi(\mathbf{m}_2))$ , and establish a supervised training strategy given

by

$$E(\phi, \theta) = \frac{1}{N} \sum_{i=1}^N \|\mathbf{m}_1^i - D_\theta(E_\phi(\mathbf{m}_2^i))\|_2^2 + \beta \|\theta\|_2^2 + \lambda \|\phi\|_2^2, \quad (2)$$

with  $\mu, \lambda > 0$ . Like the linear matching filtering procedure (Guitton, 2004), this training strategy enforces the CAE to learn a filtering task as a function of the migrated and re-migrated paired samples. To have access to multiple realizations of  $\mathbf{m}_1$  and  $\mathbf{m}_2$  within the same acquisition setup, we build our training dataset with random crops of overlapping patches from the volumes of unstacked images  $\tilde{\mathbf{m}}_1$  and re-migrated images  $\tilde{\mathbf{m}}_2$ , generated by a source extended imaging condition (Huang et al., 2016).

After the training stage, we use the decoder as a non-linear synthesis operator to solve data-domain LSRTM in the latent space. This amounts to formulating the LSRTM problem as:

$$J(\mathbf{z}) = \frac{1}{2} \|\mathbf{L}D_\theta(\mathbf{z}) - \mathbf{d}_{\text{obs}}\|_2^2 \quad (3)$$

$$\hat{\mathbf{z}} = \operatorname{argmin}_{\mathbf{z}} J(\mathbf{z}) \quad (4)$$

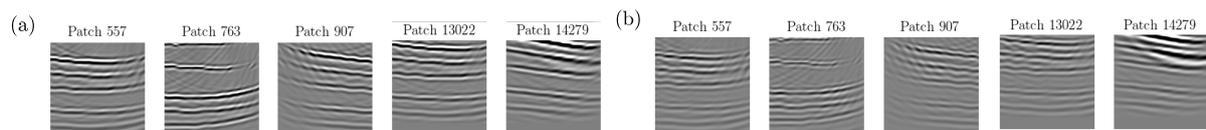
$$\hat{\mathbf{m}} = D_\theta(\hat{\mathbf{z}}), \quad (5)$$

where,  $\hat{\mathbf{m}}$  is the final inverted reflectivity. Because we split the sought reflectivity model in  $N_p$  overlapping patches of pre-defined size, we re-write  $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}_i\}_{i=1}^{N_p}$ . For simplicity, before performing the forward modelling on the full image, we attach an unpatching operator as the last layer of the decoder that assembles individually decoded patches  $\{D_\theta(\mathbf{z}_i)\}_{i=1}^{N_p}$  back together using weighing functions (Claerbout and Fomel, 2008). We employ the L-BFGS solver (Nocedal, 1980), which provides fast convergence and additional information about the local curvature of the new cost function to steer its search direction along the variable space. We rely on the decoder as a preconditioning operator to produce high-fidelity images relatively similar to those from the target distribution. To avoid convergence to a useless local minimum, we set  $\mathbf{z}_0 = E_\phi(\mathbf{m}_1)$ , which produces stable results. Furthermore,  $\mathbf{m}_1$  is available from the training stage (through a stacking operator acting on  $\tilde{\mathbf{m}}_1$ ), so initializing the inversion with the encoded adjoint image does not incur in additional computational costs.

## Numerical experiment

We aim to find the reflectivity model of a layered medium with slightly dipping reflectors (Figure 2d). The experiment simulates a fixed spread acquisition of 65 sources and 128 receivers. The source and receiver spacings are 24 m and 12 m, respectively. The first source and receiver are at  $x = 0$  m, and all sources and receivers are at  $z = 0$  m depth. The spatial grid interval is 12 m in both horizontal and depth dimensions. The seismic source is a Ricker wavelet of 20 Hz (dominant frequency), and the recording time is 1.8 seconds sampled at 1 ms. In this example, the observed data  $\mathbf{d}_{\text{obs}}$  is free of noise, and our migration velocity is accurate.

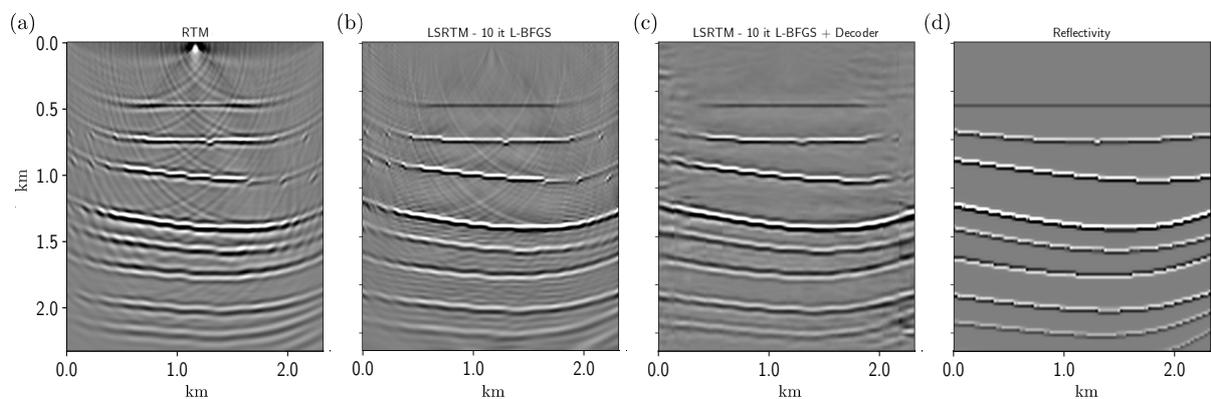
In the training stage, we define a fixed patch size of  $64 \times 64$  with a stride of  $8 \times 64$  grid points. The total number of patches is 2304. Figures 1a and 1b show several randomly chosen patches from the bank of labels  $\tilde{\mathbf{m}}_1$  and inputs  $\tilde{\mathbf{m}}_2$ , respectively. As expected, the patches of re-migrated images differ significantly from the  $\tilde{\mathbf{m}}_1$  patches, presenting much higher amplitude imbalance and more migration-related artifacts. For the deep autoencoder architecture, we employ the ResNet CAE model presented in Ravasi (2021), which includes multiple residual blocks as the backbone of the network, each composed



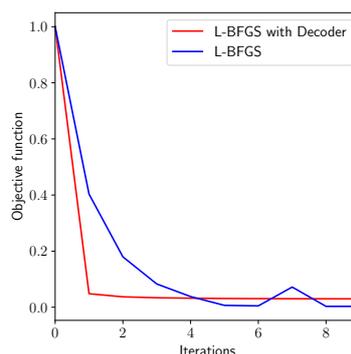
**Figure 1** Training dataset. Random selection of paired samples: (a) Labels from  $\tilde{\mathbf{m}}_1$  (high-fidelity samples). (b) Their corresponding input patches from  $\tilde{\mathbf{m}}_2$  (low-fidelity samples). All the images are plotted using the same amplitude range.

of two 2D convolutional layers (a sequence of multiple 2D convolutional filters, a batch normalization operator, and a leaky ReLU non-linear activation) and a skip connection over the two layers. To slightly improve performance, we concatenate two consecutive residual blocks on each level, producing a deeper architecture but retaining the same complexity level (He et al., 2016). We also change the number of 2D filter coefficients from  $3 \times 3$  to  $5 \times 5$  to expand the receptive field of the convolution kernel. At least for the following example, we obtain a moderately increased performance with these modifications regarding training stability and prediction quality. The input shape is hard coded to the dataset dimensionality (patch size), and the dimension of the latent space is fixed to  $h = 300$ . We use the Adam method (Kingma and Ba, 2014) to optimize the network parameters by minimizing equation 2 with 50 epochs, batch size of 256, a learning rate of  $1e^{-3}$ , and  $\beta, \lambda = 1e^{-5}$ .

Applying the patching technique with the same configuration as in the training stage, we split the sought reflectivity model into 18 patches. This corresponds to a dimensionality reduction factor of approximately 3 compared with the original model dimensions. First, we only inverted the central shot located at the central part of the model. The result is shown in Figure 2c. We also run the inversion without the decoder synthesizer for comparison (Figure 2b), showing that the imaging enhancement we obtain relies mostly on the action of the decoder. Most migration artifacts have been effectively reduced in the decoder-based inversion, and the image still presents resolution improvements compared to the adjoint image (Figure 2a). Figure 3 shows the normalized objective function versus the number of iterations. Although both inversions converge to a similar value, the decoded inversion achieves convergence in fewer iterations, translating into improved computational performance.

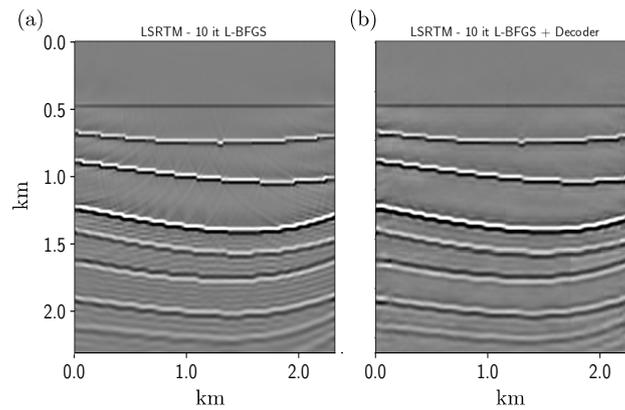


**Figure 2** Image space results for the inversion of the central shot. (a) RTM image. (b) L-BFGS result. (c) Proposed method. (d) True reflectivity.



**Figure 3** Convergence of the L-BFGS algorithm (with and without the decoder synthesizer) versus the number of iterations for the central shot inversion.

Finally, To show the method’s potential in dealing with sparse acquisitions, we limit the multiple shot inversion to only  $N_s = 10$  shots evenly spaced at the surface. Figure 4 shows the result of the stacked inversion. Naturally, stacking is effective in suppressing interferences observed in single-shot inversions. We still notice a substantial improvement in the decoded inversion over the traditional LSRTM result.



**Figure 4** Image space results for the inversion of 10 shots. (a) L-BFGS. (b) Proposed method.

## Conclusions

We developed a fast imaging framework that relies on deep learning and data domain inversion to recover an improved subsurface reflectivity model. A preliminary numerical example demonstrates the potential of the method. Compared to other supervised deep-learning techniques that need paired samples of ground-truth labels and initial reconstruction models, the proposed method does not use the former. The training stage does not need complicated pre-processing and requires minimal user interaction. It is also relatively cheap since we are simulating the effects of the Hessian operator at the cost of only one migration/re-migration sequence, equivalent to one iteration of conjugate gradients. Moreover, the deep synthesizer operator used in the inversion stage is similar in spirit to preconditioning schemes in linear inverse problems, in which the preconditioner promotes stable features in the model, improving the eigenvalue distribution of the forward operator and increasing the rate of convergence.

## Acknowledgements

The Alberta Innovates Graduate Student Scholarship has funded this research. We thank the sponsors of the Signal Analysis and Imaging Group (SAIG) at the University of Alberta for their financial support.

## References

- Claerbout, J.F. and Fomel, S. [2008] *Image estimation by example: geophysical soundings image construction: multidimensional autoregression*. Citeseer.
- Guittou, A. [2004] Amplitude and kinematic corrections of migrated images for nonunitary imaging operators. *Geophysics*, **69**(4), 1017–1024.
- He, K., Zhang, X., Ren, S. and Sun, J. [2016] Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- Huang, Y., Nammour, R. and Symes, W. [2016] Flexibly preconditioned extended least-squares migration in shot-record domain. *Geophysics*, **81**(5), S299–S315.
- Kingma, D.P. and Ba, J. [2014] Adam: A Method for Stochastic Optimization. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
- Liu, Y., Symes, W.W. and Li, Z. [2013] Multisource least-squares extended reverse time migration with preconditioning guided gradient method. In: *SEG Technical Program Expanded Abstracts 2013*, Society of Exploration Geophysicists, 3709–3715.
- Nemeth, T., Wu, C. and Schuster, G.T. [1999] Least-squares migration of incomplete reflection data. *Geophysics*, **64**(1), 208–221.
- Nocedal, J. [1980] Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, **35**(151), 773–782.
- Ravasi, M. [2021] Seismic wavefield processing with deep preconditioners. In: *First International Meeting for Applied Geoscience & Energy Expanded Abstracts*. 2859–2863.
- Tang, Y. [2009] Target-oriented wave-equation least-squares migration/inversion with phase-encoded Hessian. *Geophysics*, **74**(6), WCA95–WCA107.