Seismic Wavefield Processing with Deep Preconditioners

Matteo Ravasi, King Abdullah University of Science and Technology (KAUST)

SUMMARY

In the last decade, seismic wavefield processing has begun to rely more heavily on the solution of wave-equation-based inverse problems. Especially when dealing with unfavourable data acquisition conditions (e.g., poor, regular or irregular sampling of sources and/or receivers), the underlying inverse problem is generally very ill-posed; sparsity promoting inversion coupled with fixed-basis sparsifying transforms has become the de-facto approach for many processing algorithms. Motivated by the ability of deep neural networks to identify compact representations of N -dimensional vector spaces, we propose to learn a mapping between the input seismic data and a latent manifold by means of an Autoencoder. The trained decoder is subsequently used as a nonlinear preconditioner for the inverse problem we wish to solve. Using joint deghosting and data reconstruction as an example, we show that nonlinear learned transforms outperform fixed-basis transforms and enable faster convergence to the sought solution (i.e., fewer applications of the forward and adjoint operators are required).

INTRODUCTION

Geophysical inverse problems are notoriously ill-posed and require regularisation techniques to guide their solutions with available prior knowledge. The quintessential example in seismic processing is the problem of *seismic data interpolation*. Interpolation methods can be divided into three main categories: wave-equation (Fomel, 2003), rank-reduction (Yang et al., 2003), and domain transform (Trad et al., 2002). The latter family of algorithms exploits the fact that seismic data can be represented by few non-zero coefficients in a suitable transformed domain (e.g., FK, Radon, Curvelet), whilst acquisition gaps introduce noise in such a domain. Reconstructing missing traces becomes a denoising problem, which is usually tackled by means of sparsity promoting inversion (Hennenfent and Herrmann, 2002). Transform-based algorithms lend nicely to the introduction of additional physics that can further mitigate the ill-posed nature of the inverse problem: for example, the recorded data can be separated into its up- and down-going components whilst being reconstructed using either single-sensor data (Grion, 2017) or multi-sensor data (Ozbek et al., 2010), which is likely to reduce the propagation of errors from one step of processing to the next.

With the recent success of deep learning in various scientific fields, a new family of methods for seismic interpolation has started to emerge: Siahkoobi et al. (2018) and Mandelli et al. (2019) approach data reconstruction as an end-to-end learning problem using Generative Adversarial Networks (GANs) and a U-net inspired Encoder-Decoder network, respectively. More recently, Kuijpers and Vasconcelos (2021) leverage Recurrent Inference Machines, which are able to include knowledge of the forward operator in the learning process. Such an approach

is shown to outperform end-to-end learning in various applications including seismic data reconstruction. Finally, Kong et al. (2020) propose to solve the seismic reconstruction problem in an unsupervised manner using a deep prior preconditioner. Whilst this approach relieves the need for any training data, it is currently hindered by very slow convergence.

Motivated by the ability of deep neural network to identify compact representations of N -dimensional vector spaces, in this work we investigate a two-step approach to the solution of ill-posed inverse problems: first, an Autoencoder (AE - Kramer (1991)) is trained to learn a latent representation of the input seismic data. Subsequently its decoder is used as a nonlinear preconditioner to the inverse problem we wish to solve. By reducing the dimensionality of the search space in the inversion, our approach shows higher accuracy and faster convergence than fixed-basis transform based algorithms. We consider the problem of joint deghosting and data reconstruction and show that a representative latent manifold can be identified using data that are not exactly in the same form of the model we wish to invert for, meaning that the AE can be trained on the available ghosted data whilst its latent representation is later on used to invert for a finely sampled, deghosted data.

THEORY

In this work we are concerned with finding a stable solution to:

$$\mathbf{y} = \mathbf{G}\mathbf{x} \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ are the model and data vectors, respectively, and $\mathbf{G} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ represents a linear mapping between these two vectors. When the problem is well-posed the solution of eq. 1 can be obtained by simply inverting the modelling operator: $\hat{\mathbf{x}} = \mathbf{G}^{-1}\mathbf{y}$. Unfortunately, most geophysical problems are ill-posed and require some form of prior information to find a satisfactory solution; this can be done by approximating eq. 1 with a neighbouring well-posed problem that enforces a stable solution.

One such option is regularized least-squares inversion:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{G}\mathbf{x}\|_2 + \epsilon_R \|\mathbf{R}\mathbf{x}\|_p \quad (2)$$

where $\mathbf{R}$ is the regularization operator and we define the L_p norm as $\|\mathbf{x}\|_p = \sum_i |x_i|^p$. Examples of regularization with $p = 2$ are the so-called Tikhonov regularization ($\mathbf{R} = \mathbf{I}$ – the identity operator) that enforces the norm of the solution to be small, or roughening terms that favour smoothness in the recovered model. Similarly Total Variation regularization ($p = 1$, $\mathbf{R} = \nabla$) is a popular choice in the case where the model is expected to be piece-wise constant.

Alternatively the inverse problem can be preconditioned, meaning that the solution is found in a projected space:

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{G}\mathbf{P}\mathbf{z}\|_2 + \epsilon_P \|\mathbf{z}\|_p \quad (3)$$

Seismic Wavefield Processing with Deep Preconditioners

Figure 1: a) Sinusoid reconstruction for the different inversion approaches. b) Residual norm and c) error norm as function of iterations.

where $\mathbf{P} \in \mathbb{R}^k \rightarrow \mathbb{R}^m$ is the preconditioner and $\mathbf{z} \in \mathbb{R}^k$ is the projected variable. The model is finally obtained as $\hat{\mathbf{x}} = \mathbf{P}\hat{\mathbf{z}}$. Smoothness in the solution can be accomplished by using a smoothing operator, possibly together with a L_2 penalizing term on the projected variable. Another very popular choice of preconditioner is a transform that projects the model into a possibly overcomplete space ($k \geq m$) where the model can be explained by a small number of non-zero coefficients. This concept lies at the basis of so-called sparsity promoting inversion (Candès et al., 2006).

For both of the approaches discussed so far, the choice of the regularizer and preconditioner is driven by experience and it can take a great deal of human effort and time to find a suitable transform for a specific problem. Dictionary learning can partially overcome this limitation (Zhu et al., 2017), nevertheless linear bases struggle to accurately approximate complex vector spaces such as those spanned by seismic data. This is however not the case for nonlinear dimensionality reduction techniques that leverage deep neural networks. In mathematical terms, the inverse problem we wish to solve becomes:

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{G}D_{\theta}(\mathbf{z})\|_2 + \varepsilon_Z \|\mathbf{z}\|_p \quad (4)$$

where $D_{\theta} : \mathbb{R}^k \rightarrow \mathbb{R}^m$ is the decoder part of an AE. Here, since the functional in eq. 4 is nonlinear, second-order nonlinear solvers such as L-BFGS (Nocedal, 1980) will be used to estimate the latent variable $\mathbf{z}$. The parameters θ of the decoder are learned alongside those of an encoder $E_{\theta} : \mathbb{R}^m \rightarrow \mathbb{R}^k$ prior to solving the inverse problem in eq. 4. The learning process is performed by using a training dataset of n_s samples, $U = \{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{n_s}\}$, that is representative of the model vector $\mathbf{x}$ we wish to invert. This is accomplished by minimizing the following cost function:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n_s} \sum_{i=1}^{n_s} \|\mathbf{u}_i - D_{\theta}(E_{\theta}(\mathbf{u}_i))\|_p + \varepsilon_{\theta} \|\theta\|_2 \quad (5)$$

TOY EXAMPLE

We first apply the proposed methodology to a problem of 1D signal reconstruction. This examples provides an intuitive understanding of the value of finding a suitable nonlinear latent representation of the model vector to solve eq. 4. As a comparison we also solve i) eq. 3 with a preconditioner based on linear dimensionality reduction (i.e., Principal Component Analysis – PCA) ii) eq. 2 with a regularizer that penalizes the second-order derivative of the model (i.e., enforces smoothness).

The forward problem is defined by a restriction operator that extracts values from the finely sampled signal $\mathbf{x} \in \mathbb{R}^{500}$ (black line in Fig1a) at irregular locations to form the data vector $\mathbf{y} \in \mathbb{R}^{100}$ (black dots). To train both dimensionality reduction techniques, we assume that our signal originates from a parametric family of curves: $x = \sum_{i=1}^N a_i \sin(2\pi f_i t + \phi_i)$ where N , a , f and ϕ are sampled from uniform distributions. We sample 30000 curves and split them as, 70% for training and 30% validation. The dimensionality of the latent space is chosen to be equal to $k = 40$. Both the encoder and decoder are fully connected neural networks composed of a single hidden layer with 80 elements. Training is performed using the Adam optimizer (Kingma and Ba, 2014) with learning rate $l_r = 10^{-3}$, weight decay regularization $\varepsilon_{\theta} = 10^{-5}$, and $p = 2$ in eq. 5. After 15 epochs, the reconstruction error for both the train and validation set is virtually zero. The trained decoder is used to solve eq. 4 with 30 iterations of L-BFGS (green line in Fig1a). This solution is compared to that of the regularized problem after 30 (red line) and 200 (magenta line) iterations of LSQR, respectively, and to the solution of the PCA preconditioned problem after 30 iterations of LSQR (blue line). Faster convergence is observed in terms of the residual norm for both preconditioned solutions compared to the regularized solution (Fig1b). More importantly, the error norm of the regularized solution decays very slowly compared to their preconditioned counterparts. A major difference is also observed between the PCA and AE error norms: the former plateaus at around 2.5 after a few iterations, whilst the latter goes to zero after about 20 iterations.

Seismic Wavefield Processing with Deep Preconditioners

Figure 2: a) Autoencoder architecture. b) Schematic representation of the combined decoding and modelling operator.

JOINT DEGHOSTING AND RECONSTRUCTION

Joint receiver-side deghosting and data reconstruction is applied to a synthetic seismic dataset composed of 201 sources and receivers spaced every 15m at a depth of 10m and 50m below the free-surface, respectively. Data is modelled using a Ricker wavelet with $f_{dom} = 15Hz$. Receivers are further subsampled in two ways: i) irregularly by a factor of 40%, ii) regularly by keeping one receiver every 4 (25% available data). In order to train the AE, the training data is created by resorting the seismic data in the common receiver gather (CRG) domain. By taking advantage of the fact that sources are generally more densely sampled than receivers, CRGs are used to extract patches of size 64×64 , which are further normalized by their absolute maximum value and split into train (90%) and validation (10%) sets. The network architecture chosen for this example is shown in Fig2a. Both the contracting and expanding paths are composed of two MultiRes blocks (Kong et al., 2020). Two fully connected layers are also added at the end of the contracting path and at the beginning of the expanding path to reduce the dimensionality of the latent space to a chosen number of samples $k = 1000$. A final convolutional layer is also added to the expanding path to restore the number of channels to 1 as in the input data. The network is trained for 20 epochs using the Adam optimizer with learning rate $l_r = 10^{-3}$, weight decay of $\epsilon_\theta = 10^{-5}$, and $p = 1$ in eq. 5.

For the physical problem, the model is the up-going component of the recorded seismic data ($\mathbf{x} = \mathbf{p}^-$) whilst the data is the total pressure wavefield deprived of its direct wave ($\mathbf{y} = \mathbf{p}_-d$). The operator is defined as $\mathbf{G} = \mathbf{R}(\mathbf{I} + \Phi)\mathbf{P}\mathbf{W}$. Φ is a frequency-wavenumber phase shift operator, and $\mathbf{R}$ is a restriction operator. Moreover, since training is performed on patches of 64×64 samples, the projected model vector $\mathbf{z}$ is composed of a stack of multiple latent space vectors that are decoded by the decoder D_θ , re-scaled from the dynamical range of $(-1, 1)$ used in the network to the actual range of the seismic data via the operator $\mathbf{W}$, and finally assembled together by means of a patching operator $\mathbf{P}$ (Fig2b). Deghosting is initially applied to the fully sampled data for a source in the middle of the array. The FISTA solver (Beck and Teboulle, 2009) with a Curvelet

sparsifying transform is used to optimize the associated functional for a total number of 200 iterations. This ensures that we accurately deghost also small amplitude regions in the data such that we can use this estimate as our benchmark solution (Fig3b and 4b). The subsampled data in Fig3a and 4a are inverted with fixed-basis sparsifying transforms, namely the FK transform in overlapping time-space patches (Fig3c and 4c) and the Curvelet transform (Fig3d and 4d). Once again we use the FISTA solver for 80 iterations. Finally the trained decoder is used as preconditioner in eq. 4, which is minimized with 80 iterations of L-BFGS (Fig3e and 4e). Both visually and by means of $SNR = 10\log_{10}(\|\mathbf{x}\|_2^2 / \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2)$, we conclude that the AE-based inversion produces results of higher quality compared to other commonly used fixed-basis transforms.

DISCUSSION AND CONCLUSIONS

In this work, we have introduced a framework that leverages nonlinear based dimensionality reduction in the solution of physics-driven inverse problems. We have shown its applicability to the problem of joint deghosting and interpolation of seismic data for both irregular and regular coarse receiver geometries. An important finding of this contribution is that there is no need for any training data beyond the recorded seismic data itself. Moreover, despite training being performed on ghosted common-receiver gathers, the identified latent space is shown to transfer well to ghost-free common-source gathers which we invert for. At this stage, the choice of the network architecture and size of the latent space seem to be the determining factors to find a representative latent representation of our seismic data. Moreover, we observe that data normalization is crucial in ensuring a stable training process; this in turn requires the inclusion of a rescaling operator in the physical forward modeling operator to bring each decoded patch of the seismic data to its natural dynamical range. Finally, we argue that the proposed framework may have far wider applicability in seismic processing: acoustic and elastic wavefield separation, up/down deconvolution, and target-oriented redatuming will be subject of future studies.

Seismic Wavefield Processing with Deep Preconditioners

Figure 3: Joint deghosting and reconstruction from irregularly sampled data. a) Subsampled data, b) Deghosted data, c-d-e) Deghosted and reconstructed data using FK, Curvelet, and AE preconditioners, respectively. All data are shown in time-space domain in the top row and frequency-wavenumber domain in the bottom row.

Figure 4: Joint deghosting and reconstruction from regularly sampled data. Panels are in the same order as those in Fig3.