LARGE SCALE RANDOMIZED LEARNING GUIDED BY PHYSICAL LAWS WITH APPLICATIONS IN FULL WAVEFORM INVERSION

Rui Xie[†]*, Fangyu Li*[⋆]*, Zengyan Wang*[‡]*, and WenZhan Song*[⋆]

[†] Department of Statistics, University of Georgia
[⋆]Center for Cyber-Physical Systems, University of Georgia
[‡] Department of Computer Science, University of Georgia

ABSTRACT

The rapid convergence rate, high fidelity learning outcome and low computational cost are key targets in solving the learning problem of the complex physical system. Guided by physical laws of wave propagation, in full waveform inversion (FWI), we learn the subsurface images through optimizing the media velocity model in a large scale non-linear problem. In this paper, we combine randomized subsampling techniques with a second-order optimization algorithm to propose the Sub-Sampled Newton (SSN) method for learning velocity model of FWI. By incorporating the curvature information, SSN preserves comparable convergence rate to Newtons method and significantly reduces the iteration cost by approximating the Hessian matrix through a non-uniform subsampling scheme. The numerical experiments demonstrate that the proposed SSN method has a faster convergence rate, and achieves a more accurate velocity model in terms of mean squared error than commonly used methods.

Index Terms— randomized learning, full waveform inversion, sub-sampling, big data

1. INTRODUCTION

The full waveform inversion (FWI) is a state-of-the-art method in subsurface imaging for providing the high-resolution estimate of the complex subsurface structure [1, 2]. FWI uses measured wave-field data to learn the media velocity model that wave propagated through to invert the subsurface image. The learning procedure is through a nonlinear minimization of a penalized error function, which is the normed discrepancy between observed wave-field data and calculated data with the constraint of wave equation [3, 4]. The calculated data is formulated by the most appropriate physical model for the system of wave propagation, which is mathematically described by a certain type of partial differential equation (PDE), such as wave equation, with unknown model coefficients, such as the velocity model in FWI.

Our research is partially supported by NSF-CNS1066391, NSF-CNS-0914371, NSF-CPS-1135814, NSF-CDI-1125165, NSF-DMS-1222718, NIH-R01-GM113242, NIH-R01-R01GM122080 and Southern Company.

Due to the complexity and ultra large data size of tracking the wave propagation, the iterative optimization method is used to achieve the minimization of the error function so that we can find the optimal velocity model for subsurface imaging [5]. The iterative optimization can be roughly divided into three part, forward modeling (solving PDE given the velocity model), error calculation (including gradient and Hessian) and velocity model update (first or second order iterative optimization) [3, 2]. Even though the relatively early introduction of the FWI technique, the intensive computation cost of these three components retard the development of FWI until the recent advances in high performance computing facilities along with the acquisition systems. As an example, for frequency domain FWI, the discretization of the Helmholtz equation, solving the forward modeling, calculating the error, and effectively updating the velocity, all involve the huge amount of data calculation and storage. The high computational burden calls for an efficient algorithm to provide a fast and accurate solution to the velocity model learning of FWI.

With the current development of FWI, the multiple-frequency or hierarchical data acquisition design promote FWI to more and more multi-scale and various field real data applications [6, 7]. The development of data quality also requires a more realistic forward modeling and a more accurate velocity model learning method. Among the traditional methods for FWI, the first order methods such as gradient based methods [8] and quasi-Newton methods such as the l-BFGS method [9, 10] preserve the stable numerical performance and low computational costs, but their convergence rate are not satisfied [11]. On the other hand, the family of second order method, the Newton-type methods [12, 13], has the advantages of fast convergence, exact update to second order Taylor approximation system, less geometric spreading and scattering artifacts [14]. However, in term of explicit calculation, the naive Newton-type methods of a large scale problem is impractical due to huge memory and storage cost and heavy computational burden. In recent years, developing fast and nearly scalable second order optimization methods is drawing more and more attentions on the optimizing and learning of complex optimization system [15, 16, 17].

To conquer the high iteration computational cost of Newton-type method, extract the useful information from the highly correlated data and preserve the fast convergence rate of the second order method, we propose a novel Sub-Sampled Newton (SSN) method with non-uniform/important sampling scheme to deliver the fast and accurate subsurface imaging through FWI.

2. ALGORITHM DESIGN

An accurate and fast optimization tool that can solve the complex subsurface imaging problem through FWI is crucially needed. Heavy computational burden has always been the bottleneck of the development in subsurface imaging, especially imaging through FWI. To deliver a high-resolution learning outcome of velocity model, we have to rely on the whole wave field information and accurate realization of the physical properties of wave propagation. The process that making efforts to get an accurate subsurface imaging comes along with huge quantity of data for processing and storing. The efficient use of these information will accelerate the learning procedure of FWI. Relative to first-order methods, second-order methods enjoy plenty of advantages in the nonlinear optimization problem, such as the superior convergence, robustness to ill-conditioned problem, and global convergence guarantee under mild assumptions [18].

Subsurface imaging through FWI, from computation perspective, is a constraint nonlinear learning problem. We consider the wave propagation as a multi-dimensional dynamic process $u(\boldsymbol{z},\omega)$, where $\boldsymbol{z} = (z_1,\ldots,z_p)^T \in \mathbb{R}^p$ is a multi-dimensional covariate and ω is the additional covariate, if exists, that is associated with the varying model coefficients. We assume that the dynamic process $u(\boldsymbol{z},\omega)$ follows an varying coefficients PDE,

$$\mathcal{F}\Big\{\boldsymbol{z}, \frac{\partial u(\boldsymbol{z},\omega)}{\partial z_1},\ldots,\frac{\partial u(\boldsymbol{z},\omega)}{\partial z_p}, \frac{\partial^2 u(\boldsymbol{z},\omega)}{\partial z_1^2}, \frac{\partial^2 u(\boldsymbol{z},\omega)}{\partial z_1 \partial z_2},\ldots, \boldsymbol{\theta}(\boldsymbol{z},\omega), f(\boldsymbol{z})\Big\} = 0, \tag{1}$$

where $\boldsymbol{\theta}(\boldsymbol{x},\omega) = (\theta_1(\boldsymbol{x},\omega),\ldots,\theta_m(\boldsymbol{x},\omega))^T$ is the varying coefficient vector depending on $\boldsymbol{x}$ and ω, and $f(\boldsymbol{x})$ is a known "forcing term" or "source". In the application of FWI, the varying coefficients PDE is specified as the wave equation in the time domain, or equivalently, the Helmholtz equation in the frequency domain. The PDE in (1) then becomes the Helmholtz equation,

$$\left(\boldsymbol{m}(\boldsymbol{x},\omega) + \nabla^2\right) u_\omega(\boldsymbol{x}) = -f_\omega(\boldsymbol{x}_s), \tag{2}$$

where $\boldsymbol{m}(\boldsymbol{x},\omega) = \frac{\omega^2}{v^2(\boldsymbol{x})}$, $u_\omega(\boldsymbol{x}) = u(\boldsymbol{x},t)e^{i\omega t}$, and $f_\omega(\boldsymbol{x}_s) = f(\boldsymbol{x}_s,t)e^{i\omega t}$ correspondingly with $\omega \in \Omega$ and $\boldsymbol{x},\boldsymbol{x}_s \in \mathcal{X} \subset \mathbb{R}^d$ given the frequency domain Ω and spatial domain $\mathcal{X}$. In the frequency domain Helmholtz equation, the varying coefficient $\boldsymbol{m}(\boldsymbol{x},\omega)$ depends not only on the spatial covariate $\boldsymbol{x}$, but also an additional covariate ω that is free from the derivatives in PDE.

In practice, we do not observe the dynamic process $u_\omega(\boldsymbol{x})$ and source term $f_\omega(\boldsymbol{x})$ on the whole domain. For the frequency domain FWI, instead we observe a surrogate $Y(\boldsymbol{x},\omega)$ at the locations where the sources and sensors are placed (usually on the surface of the ground) and within a certain frequency range. We assume that sources are located at source positions $\boldsymbol{x}_s$ with $s = 1,\ldots,n_s$, sensors are located at receiver positions $\boldsymbol{x}_r$ with $r = 1,\ldots,n_r$ and the frequency is observed at ω_w with $w = 1,\ldots,n_w$. The wave-field data we observed are denoted as

$$d_{obs,srw} = u_\omega(\boldsymbol{x}_r,\boldsymbol{x}_s,\omega_w,\boldsymbol{m}) + \epsilon_{srw}, \tag{3}$$

where ϵ_{srw}'s are assumed to be independent and identically distributed errors with zero mean and finite variance. Our goals are to estimate the varying model coefficient surface from the observed noisy data with the constraint of the PDE and to establish the statistical inference of the estimates [3].

The recorded data are acquired from an array of seismic receivers and denoted as $\mathbf{d}_{obs} = \{d_{obs,srw}\}_{srw}$. We track the wave propagation through PDE (2) given the velocity $\mathbf{m}$ and solve it numerically using finite-difference method [19], where the wave fields are projected at the receiver positions $\boldsymbol{x}_r$. The velocity model $\mathbf{m} \triangleq [m(\mathbf{x}_1),...,m(\mathbf{x}_{N_zN_x})]$, $m(\mathbf{x}_i)$ indicates the squared-slowness value at the 2D coordinate $\mathbf{x}_i, i = 1,...,N_zN_x$, where N_z and N_x are the vertical and lateral grid number, respectively.

In velocity model learning, we estimate the varying model coefficients $\mathbf{m}$ by minimizing the penalized squared errors of the estimated PDE solution $\mathbf{d}_{cal}$ to the noisy data in (3),

$$\begin{aligned} E(\mathbf{m}) &= \frac{1}{2}||\mathbf{d}_{obs} - \mathbf{d}_{cal}||_2^2 + \lambda\mathcal{J}(\mathbf{m}) \\ &= \frac{1}{2}||\mathbf{d}_{obs} - u_\omega(\boldsymbol{x},\mathbf{m})||_2^2 + \lambda\int \mathcal{F}\Big\{\boldsymbol{x}, \frac{\partial u(\boldsymbol{x},\omega)}{\partial x_1},\ldots, \\ &\quad \frac{\partial^2 u(\boldsymbol{x},\omega)}{\partial x_1^2},\ldots,\frac{\partial^2 u(\boldsymbol{x},\omega)}{\partial x_1\partial x_2},\ldots,\boldsymbol{m}(\boldsymbol{x},\omega), f(\boldsymbol{x})\Big\}d\boldsymbol{x}, \end{aligned} \tag{4}$$

where the first term measures the goodness-of-fit of (3), and the second term measures the fidelity of (3) to the physical system, i.e. PDE model, dened in (1). The tuning parameter λ controls the trade-off between fitting to the observed data and fidelity to the physical system.

Method for Learning Velocity Model

We learn the varying coecient vector $\boldsymbol{\theta}(\boldsymbol{x},\omega)$ in PDE model (1) in an alternating way with two nested levels of optimization. On one hand, we solve the PDE (1) given the current estimation of coefficient vector $\boldsymbol{\theta}(\boldsymbol{x},\omega)$ through finite-difference method to generate the calculated data $\mathbf{d}_{cal}$, which is usually called forward modeling. On the other hand, we update the

estimation of the coefficient vector $\boldsymbol{\theta}(\boldsymbol{x}, \omega)$ by minimizing the penalized squared errors of (4). This learning procedure is usually called inverse problem.

More specifically, in the application of FWI, to minimize $E(\mathbf{m})$, we search in an iterative manner, $\mathbf{m}_{k+1} = \mathbf{m}_k + \delta\mathbf{m}_k$, with $\delta\mathbf{m}_k$ being the optimal model perturbation at the k-th iteration step that minimizes $E(\mathbf{m})$. The optimal model perturbation comes from the expansion of $E(\mathbf{m}_k)$ in a small vicinity $\delta\mathbf{m}$ of $\mathbf{m}_k$ with a Taylor polynomial of degree two:

$$E(\mathbf{m}) = E(\mathbf{m}_k) + \delta\mathbf{m}^T\mathbf{g}_k + \frac{1}{2}\delta\mathbf{m}^T\mathbf{H}_k\delta\mathbf{m} + o(\|\delta\mathbf{m}\|^3),$$

where $\mathbf{g}_k \triangleq \partial E(\mathbf{m}_k)/\partial\mathbf{m}$ is the gradient of the error function $E(\mathbf{m})$ at $\mathbf{m}_k$ and $\mathbf{H}_k \triangleq \partial^2 E(\mathbf{m}_k)/\partial\mathbf{m}^2$ denotes the corresponding Hessian matrix. The Newton-type method is used to get the optimal model perturbation $\delta\mathbf{m}_k$ through the normal equation:

$$\mathbf{H}_k\delta\boldsymbol{m}_k = -\mathbf{g}_k, \tag{5}$$

and then update the velocity model according to

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha_k[\mathbf{H}(\mathbf{m}_k)]^{-1}\mathbf{g}_k, \tag{6}$$

where α_k is the learning rate.

Gradient and Hessian

The conventional computation of gradient $\mathbf{g}_k$ and Hessian $\mathbf{H}_k$ can be specified as

$$\begin{aligned}\mathbf{g}_k = \frac{\partial E(\mathbf{m}_k)}{\partial\mathbf{m}} &= -\Re\left\{\left[\frac{\partial\mathcal{F}(\mathbf{m}_k)}{\partial\mathbf{m}}\right]^\dagger(\mathbf{d}_{obs} - \mathcal{F}(\mathbf{m}_k))\right\} \\ &= \Re\left\{\mathbf{J}_k^\dagger\delta\mathbf{d}_k\right\},\end{aligned} \tag{7}$$

where $\delta\mathbf{d}_k = \mathbf{d}_{obs} - \mathcal{F}(\mathbf{m}_k)$, $(\cdot)^\dagger$ denotes the adjoint of operator $(\cdot)$, $\Re\{z\}$ denotes the real part of complex number z, and $\mathbf{J}_k = -\partial\mathcal{F}(\mathbf{m}_k)/\partial\mathbf{m}$ is the Jacobian of $-\mathcal{F}(\cdot)$. By taking the derivative of $\mathbf{g}_k$, we get Hessian of $E(\cdot)$

$$\mathbf{H}_k = \frac{\partial^2 E(\mathbf{m}_k)}{\partial\mathbf{m}^2} = \Re\left\{\mathbf{J}_k^\dagger\mathbf{J}_k + \frac{\partial\mathbf{J}_k^\dagger}{\partial\boldsymbol{m}^T}[\delta\mathbf{d}_1^* \ \cdots \ \delta\mathbf{d}_k^*]\right\}, \tag{8}$$

where $(\cdot)^*$ denotes the conjugate of a complex number. Note that the first part of the Hessian matrix, $\Re\left\{\mathbf{J}_k^\dagger\mathbf{J}_k\right\}$, contains the most useful information of the curvature, and the second-order part of the Hessian matrix is easy to be contaminated be the noise since it usually presents numerous small or negative eigenvalues during the evaluation [20].

Sub-sampled Newton (SSN) Method

To update the velocity model, we need to solve the normal equation (5). The computational bottleneck for solving equation (5) is the inverse of the Hessian matrix $\mathbf{H}_k$, which takes $O((N_zN_x)^3)$ flops and $O((N_zN_x)^2)$ memories. To preserve the quadratic convergence rate of the second order method, extract the useful information from the highly correlated data, and conquer such high per-iteration computational cost in forming and inverting the Hessian matrix, along with the line of [15, 16, 17], we propose a randomized second order learning method called Sub-Sampled Newton (SSN) for FWI inverse problem.

We note that on realization of the Hessian of $E(\cdot)$ at certain receivers and frequencies,

$$\mathbf{H}(\mathbf{m}_k) = \sum_{s=1}^{N}\mathbf{A}_s^T(\mathbf{m}_k)\mathbf{A}_s(\mathbf{m}_k) + \mathbf{Q}(\mathbf{m}_k), \tag{9}$$

where $\sum_{s=1}^{N}\mathbf{A}_s^T(\mathbf{m}_k)\mathbf{A}_s(\mathbf{m}_k) = \Re\left\{\mathbf{J}_k^\dagger\mathbf{J}_k\right\}$, $\mathbf{Q}(\mathbf{m}_k) = \Re\left\{\frac{\partial\mathbf{J}_k^\dagger}{\partial\mathbf{m}^T}[\delta\mathbf{d}_1^* \ \cdots \ \delta\mathbf{d}_k^*]\right\}$, and $N = N_sN_rN_\omega$ with N_s, N_r, and N_ω are the source, receiver, and frequency numbers respectively.

As second order methods have been demonstrated to be effective in finding high precision minimizer, we propose a randomized second order learning method, Sub-Sample Newton, that exploit *non-uniform* sub-sampling of $\Re\{\mathbf{J}_k^\dagger\mathbf{J}_k\}$ to reduce the computational cost and achieve comparable convergence rate to Newtons method. We construct the non-uniform sub-sampling π_i over $\mathbf{A}_i(\boldsymbol{m}_k)$ with $i = 1, \ldots, N$ according to $\pi_i = \frac{\|\mathbf{A}_i\|_F^2}{\|\mathbf{A}\|_F^2}$ [21, 22], or the block partial leverage score [16] and take s sub-sample terms proportional to π_i. The details of the proposed method is summarized in Algo-

Algorithm 1: Sub-sampled Newton method with Non-uniform Sampling

Input: Initial velocity $\boldsymbol{m}_0$, frequency ω, number of iteration K, sampling scheme $\mathcal{S}$ and solver $\mathcal{A}$.
Output: $\mathbf{m}_K$
for $k = 0, \ldots, K-1$ **do**
- Construct the sampling distribution $\{\pi_i\}_{i=1}^N$ according to sampling scheme $\mathcal{S}$: leveraging score of $\mathbf{A}(\mathbf{m}) = [\mathbf{A}_1^T \cdots \mathbf{A}_N^T]^T$, say, $\pi_i = \frac{\|\mathbf{A}_i\|_F^2}{\|\mathbf{A}\|_F^2}$ [21], or the block partial leverage score [16].
- Draw the sample, i.e. $S_\mathbf{A}^T\mathbf{A}$, according to an importance sampling distribution $\{\pi_i\}_{i=1}^N$, where $S_\mathbf{A}^T$ is a random sampling matrix.
- Calculate randomized Hessian sketch $\widetilde{\mathbf{H}}(\mathbf{m}_k) = \sum_{l\in\mathcal{I}}\mathbf{A}_l^T(\mathbf{m}_k)\mathbf{A}_l(\mathbf{m}_k)/\pi_l + \mathbf{Q}(\boldsymbol{m}_k)$, where $\mathcal{I}$ is the subsampled indices set with size s.
- Calculate $\mathbf{g}(\boldsymbol{m}_k)$, and learning rate α_k using line search.
- Update velocity $\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha_k[\widetilde{\mathbf{H}}(\mathbf{m}_k)]^{-1}\mathbf{g}_k$, using solver $\mathcal{A}$ to inverse Hessian $\widetilde{\mathbf{H}}(\mathbf{m}_k)$.

end
return $\mathbf{m}_K$

rithm 1.

When the Hessian matrix dimension is large, the inexact solver, e.g. matrix free optimization [23] or conjugate gradient method [24], can be used to update the velocity model using a few iterations to produce a high-quality approximated solution to the normal equation.

3. NUMERICAL EXPERIMENTS

A 2D SEG/EAGE overthrust model (Fig. 1a) is used to test the proposed algorithm. The initial model is a smoothed version of the true model (Fig. 1b). The original model consists of 801×187 grid cells in a 2-D section with 25 m horizontal and vertical grid intervals. There are 100 sources and 100 receivers laid on the surface, which are spread out with 25 m spatial interval. A multi-scale inversion approach is adopted in our numerical experiments in frequency bands $0.5 - 4$ Hz in every 0.5 Hz.

Fig. 1c – Fig. 1e demonstrate the learning results of (c) gradient decent, (d) l-BFGS and (e) Sub-Sampled Newton based on the data set of the lowest frequency band ($0.5-4$Hz). Fig. 2 provides the convergence rate comparison among the three methods, which shows the overall performances of them given the same number of forward modelling evaluations.

From both Fig. 1 and Fig. 2, we see that the proposed SSN recover the best velocity model among three methods. In Fig. 2, given the same numbers, 10 per 0.5 Hz, of forward modelling evaluations, SSN method converge faster than gradient decent and l-BFGS by achieving smaller mean squared error (MSE) across almost all frequencies. The gradient descent method has a large MSE, which implies that the numbers of forward modelling evaluations may not be sufficient for efficiently recover the velocity model or even obtaining a correct search direction. l-BFGS shows a better convergence performances as the frequency raise, but still significantly slower than SSN. Considering the expense of evaluating forward modeling, SSN outperforms the first order method, gradient decent and l-BFGS, and save the high per-iteration cost of vanilla Newton type method so that achieves the fast and accuracy learning results.

4. CONCLUSIONS

We present an efficient Sub-Sampled Newton (SSN) method to solve complex non-linear system guided by physical laws with application to FWI problem. SSN significantly reduces the computational complexity while preserving a fast convergence property, by using the non-uniform subsampling techniques. SSN captures the important information in the second order term thus having a rapid rate of convergence. In numerical experiments of the Overthrust velocity model, we demonstrate that SSN significantly outperformed gradient descent and l-BFGS, resulting in high-quality inverted velocity model.

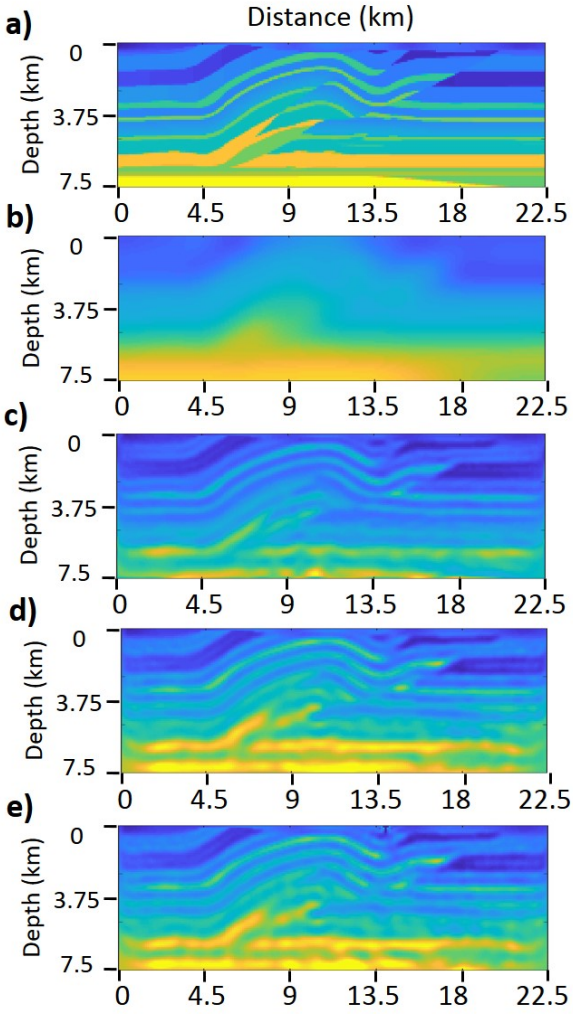

Fig. 1. (a) Overthrust model, (b) Initial velocity model, (c-e) The learning results of (c) gradient decent, (d) l-BFGS and (e) Sub-Sampled Newton using the data set of the lowest frequency band ($0.5 - 4$Hz).

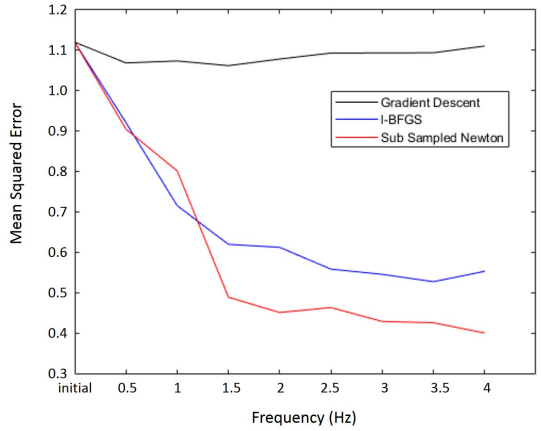

Fig. 2. Convergence comparison of different methods. The mean squared error (MSE) of velocity model is plotted every 0.5 Hz with 10 forward modelling are evaluated at each of the 0.5 Hz frequencies.

5. REFERENCES

[1] Stéphane Operto, Yaser Gholami, V Prieux, Alessandra Ribodetti, R Brossier, L Metivier, and Jean Virieux, "A guided tour of multiparameter full-waveform inversion with multicomponent data: From theory to practice," *The Leading Edge*, vol. 32, no. 9, pp. 1040–1054, 2013.

[2] Jean Virieux and Stéphane Operto, "An overview of full-waveform inversion in exploration geophysics," *Geophysics*, vol. 74, no. 6, pp. WCC1–WCC26, 2009.

[3] Albert Tarantola, "Inversion of seismic reflection data in the acoustic approximation," *Geophysics*, vol. 49, no. 8, pp. 1259–1266, 1984.

[4] Ying Rao and Yanghua Wang, "Seismic waveform tomography with shot-encoding using a restarted l-bfgs algorithm," *Scientific Reports*, vol. 7, no. 1, pp. 8494, 2017.

[5] R Gerhard Pratt, Changsoo Shin, and GJ Hick, "Gauss–newton and full newton methods in frequency–space seismic waveform inversion," *Geophysical Journal International*, vol. 133, no. 2, pp. 341–362, 1998.

[6] René-Edouard Plessix, Guido Baeten, Jan Willem de Maag, Marinus Klaassen, Zhang Rujie, and Tao Zhifei, "Application of acoustic full waveform inversion to a low-frequency large-offset land data set," in *SEG Technical Program Expanded Abstracts 2010*, pp. 930–934. Society of Exploration Geophysicists, 2010.

[7] Jacques R Ernst, Alan G Green, Hansruedi Maurer, and Klaus Holliger, "Application of a new 2d time-domain full-waveform inversion scheme to crosshole radar data," *Geophysics*, vol. 72, no. 5, pp. J53–J64, 2007.

[8] R Gerhard Pratt, "Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model," *Geophysics*, vol. 64, no. 3, pp. 888–901, 1999.

[9] R-E Plessix and WA Mulder, "Frequency-domain finite-difference amplitude-preserving migration," *Geophysical Journal International*, vol. 157, no. 3, pp. 975–987, 2004.

[10] Wenyong Pan, Kristopher A Innanen, Gary F Margrave, and Danping Cao, "Efficient pseudo-gauss-newton full-waveform inversion in the τ-p domain," *Geophysics*, vol. 80, no. 5, pp. R225–R14, 2015.

[11] W Hu, A Abubakar, TM Habashy, and J Liu, "Preconditioned non-linear conjugate gradient method for frequency domain full-waveform seismic inversion," *Geophysical Prospecting*, vol. 59, no. 3, pp. 477–491, 2011.

[12] Ivar Stakgold and Michael J Holst, *Green's functions and boundary value problems*, vol. 99, John Wiley & Sons, 2011.

[13] L Métivier, F Bretaudeau, R Brossier, S Operto, and J Virieux, "Full waveform inversion and the truncated newton method: quantitative imaging of complex subsurface structures," *Geophysical Prospecting*, vol. 62, no. 6, pp. 1353–1375, 2014.

[14] Cai Liu, Fengxia Gao, Xuan Feng, Yang Liu, and Qianci Ren, "Memoryless quasi-newton (mlqn) method for 2d acoustic full waveform inversion," *Exploration Geophysics*, vol. 46, no. 2, pp. 168–177, 2014.

[15] Murat A Erdogdu and Andrea Montanari, "Convergence rates of sub-sampled newton methods," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*. MIT Press, 2015, pp. 3052–3060.

[16] Peng Xu, Jiyan Yang, and Farbod None Roosta-Khorasani, "Sub-sampled newton methods with non-uniform sampling," in *Advances In Neural Information Processing Systems*, 2016, pp. 2530–2538.

[17] Mert Pilanci and Martin J Wainwright, "Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 205–245, 2017.

[18] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright, *Optimization for machine learning*, Mit Press, 2012.

[19] Jean Virieux, "P-sv wave propagation in heterogeneous media: Velocity-stress finite-difference method," *Geophysics*, vol. 51, no. 4, pp. 889–901, 1986.

[20] Ludovic Métivier, Romain Brossier, Jean Virieux, and Stéphane Operto, "Full waveform inversion and the truncated newton method," *SIAM Journal on Scientific Computing*, vol. 35, no. 2, pp. B401–B437, 2013.

[21] Ping Ma, Michael W Mahoney, and Bin Yu, "A statistical perspective on algorithmic leveraging," *Journal of Machine Learning Research*, vol. 16, pp. 861–911, 2015.

[22] Xinlian Zhang, Rui Xie, and Ping Ma, *Statistical Leveraging Methods in Big Data*, pp. 51–74, Springer International Publishing, Cham, 2018.

[23] James Martens, "Deep learning via hessian-free optimization.," in *ICML*, 2010, vol. 27, pp. 735–742.

[24] Stephen Wright and Jorge Nocedal, "Numerical optimization," *Springer Science*, vol. 35, no. 67-68, pp. 7, 1999.