Efficient least-squares migration with sparsity promotion

Felix J. Herrmann and Xiang Li

January 13, 2011

Abstract

Seismic imaging relies on the collection of multi-experimental data volumes in combination with a sophisticated back-end to create high-fidelity inversion results. While significant improvements have been made in linearized inversion, the current trend of incessantly pushing for higher quality models in increasingly complicated regions reveals fundamental shortcomings in handling increasing problem sizes numerically. The so-called "curse of dimensionality" is the main culprit because it leads to an exponential growth in the number of sources and the corresponding number of wavefield simulations required by 'wave-equation' migration. We address this issue by reducing the number of sources by a randomized dimensionality reduction technique that combines recent developments in stochastic optimization and compressive sensing. As a result, we replace the current formulations of imaging that rely on all data by a sequence of smaller imaging problems that use the output of the previous inversion as input for the next. Empirically, we find speedups of at least one order-of-magnitude when each reduced experiment is considered theoretically as a separate compressive-sensing experiment.

Introduction

Modern-day imaging relies increasingly on computationally and data-intensive wave-equation based least-squares migration to meet society's continued demand for hydrocarbons. This is problematic because in 3D direct solvers are unavailable and this leads exponentially to increasing costs as the dimensionality of imaging increases.

Motivated by early work of Morton and Ober (1998); Romero et al. (2000), we confront this challenge of the "curse of dimensionality" with a randomized dimensionality-reduction technique that decreases the number of source experiments and, as a result, the computational costs significantly. Conceptually, our contribution is a natural extension of earlier work, where costs of wavefield simulations were reduced by adapting compressive sensing (Herrmann et al., 2009; Neelamani et al., 2010).

However, contrary to wavefield simulations, migration entails the inversion of a large, but in the absence of noise, consistent 'overdetermined' system of equations. At the same time, finite-aperture effects and shadow zones make this system ill conditioned (Symes, 2008), which makes least-squares imaging challenging because of high costs related to the matrix-vector multiplies. By combining stochastic optimization (Bertsekas and Tsitsiklis, 1996; Nemirovski et al., 2009; Haber et al., 2010) with compressive sensing (CS—in short throughout the paper Candès et al., 2006; Donoho, 2006), we cast the original imaging problem into a series of smaller problems that works on different subsets of random source-encoded supershots (Li and Herrmann, 2010). As a result, we are able to drastically reduce the computation time at little or no cost in quality.

Randomized dimensionality reduction

Wave-equation migration involves the solution of the following separable least-squares problem:

$$\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2K} \sum_{i=1}^{K} \|\mathbf{b}_i - \mathbf{A}_i \mathbf{x}\|_2^2, \tag{1}$$

with $K = n_f \times n_s$ the batch size, given by the total number of monochromatic sources with n_f, n_s the number of frequencies and sources, respectively. In this expression, the vectors $\mathbf{b}_i$ represent the vector-

ized monochromatic shot records and the matrix $\mathbf{A}_i$ represents the corresponding scattering matrix for the i^{th} source. Evaluation of the action of this matrix, or its adjoint $\mathbf{A}_i^H$, requires the solution of two discretized partial-differential equations (PDEs). Thus, the majority of the inversion costs grow linearly with the number of monochromatic experiments, multiplied by the number of matrix-vector multiplies required by the solver of the least-squares migration problem.

Stochastic optimization: In machine learning, separable optimization problems (cf. Eq. 1) are solved efficiently by either the stochastic-average approximation (SAA) or by the stochastic approximation (SA). With SSA, Eq. 1 is approximated by selecting random subsets of frequencies and shots (possibly simultaneous), yielding $K' = n'_f \times n'_s \ll K$ for the reduced batch size. Seismologically, this corresponds to carrying out least-squares migration with randomly selected subsets of frequencies and sources. Unfortunately, the error of this approximation decays only slowly $\mathcal{O}(K')$ because this is essentially a Monte-Carlo sampling method. However, the advantage of SAA is its simplicity that allows us to use generic solvers such as LSQR (Paige and Saunders, 1982). This explains SAA's recent use in least-squares migration with phase-encoded sources (see e.g. Fei et al., 2010).

As opposed to SAA, SA changes the solver by randomly selecting a different monochromatic source—i.e., a different $\mathbf{A}_i$—for each gradient update of steepest descents. For linear problems, this approach corresponds to randomized 'block' Kaczmarz (Strohmer and Vershynin, 2009), which was used by Natterer (2001) in tomography. Like SAA, SA extends to nonlinear inversion problems (see e.g. Nemirovski et al., 2009), and was recently introduced by Haber et al. (2010) in the context of parameter-estimation problems with PDE's. Even though SA provides insights into observed lack of convergence and instabilities with respect to noise (van Leeuwen et al., 2010), the theory does not readily extend to more sophisticated solves such as nonlinear least squares or l-BFGS proposed by Krebs et al. (2009); Ben-Hadj-Ali et al. (2009).

Compressive sensing: Randomized-dimensionality reduction also underlies recent advances in sampling theory for signals that exhibit structure that translates into transform-domain sparsity. As opposed to stochastic optimization, where randomization is used to reduce variance and stochasticity amongst different gradient updates, CS turns coherent subsampling-related interferences—such as aliasing and shot 'cross talk'—into relatively harmless Gaussian noise. According to CS, the energy of this noise depends on the degree of subsampling and transform-domain sparsity. Consequently, sampling is no longer dominated by Nyquist, but by transform-domain sparsity (see e.g. Herrmann, 2010, for an overview of the application of CS in exploration seismology). This insight represents a new paradigm of which we took advantage in speeding up wavefield simulations with the time-harmonic Helmholtz solver in Herrmann et al. (2009). Independently, Neelamani et al. (2010) made a similar observation in the context of simulations using time-domain finite difference. In both cases, fully-sampled wavefields are recovered with curvelet-domain sparsity promotion from small subsets of (monochromatic) simultaneous-source experiments. This procedure is efficient because the overhead of the recovery is small compared to computational gain of the subsampling.

The seismic mini batch—a collection of supershots: As in earlier work (Herrmann et al., 2009), we base our algorithm on forming compressive seismic experiments—or to use the language of online machine learning *mini batches*—that consist of collections of small numbers of supershots. These supershots are made of randomized superpositions of sequential sources. To maximize the randomization, each supershot is composed of a different small set of randomly selected frequencies. Mathematically, imaging experiments for mini batches with $n'_s \ll n_s$ supershots, require the solution of a reduced system

$$\mathbb{P}_{\ell_2}(\mathbf{RM}): \qquad \widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{RM}\,(\mathbf{b}-\mathbf{Ax})\,\|_2^2, \tag{2}$$

where $\mathbf{b}$ and $\mathbf{A}$ are now defined by the full set of data. Each simultaneous-source experiment in the collection of supershots is given by a different restriction, which the frequencies and selects a different random superposition of shots. Hence, the i^{th} block of the restiction matrix $\mathbf{R}$ is defined by the Kronecker product $\mathbf{R}_i := \mathbf{R}_i^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{R}_i^{\Omega}$ for $i = 1 \cdots n'_s$ with $\mathbf{R}_i^{\Sigma}$ a randomly selected row from the identity basis and $\mathbf{R}_i^{\Omega}$ the frequency restriction. The measurement matrix $\mathbf{M}$ is given by the Kronecker product $\mathbf{M} := \mathbf{M}^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{I}$. As in Lin and Herrmann (2007), we use Romberg (2009)'s method for the phase encoding.

Using the linearity of randomized subsampling by the operator **RM**, in combination with linearity and separability of monochromatic sources, it is a relatively simple exercise to show that the number of PDE solves required per iteration of Eq. 2 is slimmed down by a factor of K'/K (see also Herrmann et al., 2009, for details). However, remember this speed up goes at the expense of leaking coherent energy away from imaged reflectors to incoherent artifacts. Hence, the key question is to find a solver that mitigates these artifacts and restores the amplitudes at a cost small compared to this speed up.

Stochastic-average approximations with warm restarts: In practice, we found that combining SAA and SA with restarts gives superior results (Haber et al., 2010). For linearized inversion, this approach corresponds to drawing a mini batch, followed by solving Eq. 2, and using this solution as a warm start for a new linearized inversion with a freshly drawn mini batch. This process is repeated until there is no longer progress towards the solution. For more details, see Algorithm 1, which outlines this solver for a general optimization scheme $\mathbb{P}(\mathbf{RM};\mathbf{x}_0)$, with warm start $\mathbf{x}_0$.

We compare the performance of this algorithm for least-squares (LSQR, Paige and Saunders, 1982, denoted by $\mathbb{P}_{\ell_2}(\mathbf{RM})$), where we solve each subproblem for 10 iterations, and (SPGℓ_1, Berg and Friedlander, 2008, denoted by $\mathbb{P}_{\ell_1}(\mathbf{RM})$). For the second solver, we leverage properties of the Pareto curve (Hennenfent et al., 2008) by selecting new mini batches each time the Pareto curve is reached. At that point, the solution is sparse, and as long as the batch size is not too small, we argue that the Pareto curves will be similar for each **RM**. This means we can continue to use the root finding of SPGℓ_1.

Example: To illustrate the performance of our dimensionality-reduced algorithm, we conduct a series of synthetic imaging experiments based on the smoothed background model and the slowness perturbation plotted in Fig.1 (left, middle). We model data with the linearized Born scattering operator for 384 co-located shot and receiver positions with a spacing of 24 m, yielding a maximal offset of 9.2km. We used 3 frequencies, weighted by a Ricker wavelet with central frequency of 12Hz, from a time record of 3s. We use this dataset as input to our two inversion algorithms. For reference, we also include the least-squares image for the total batch size of $K = 36 \times 10^3$ in Fig.1 (right). Fig.2 summarizes our recovery results from a minibatch defined by only eight simultaneous shots and three frequencies, yielding a reduction of the batch size by a factor of $K'/K = 7 \times 10^{-4}$. Juxtaposition of the results for $\mathbb{P}_{\ell_2}(\mathbf{RM})$ for 10 restarts and $\mathbb{P}_{\ell_1}(\mathbf{RM})$ in Fig.2(left, middle) clearly shows improved recovery with ℓ_1, even though the total number of PDE solves are approximately the same. Comparison of Fig.'s 2(middle, right) also shows the importance of drawing new minibatches after each subproblem is solved. We observe this for both $\mathbb{P}_{\ell_2}(\mathbf{RM})$ and $\mathbb{P}_{\ell_1}(\mathbf{RM})$. Because we needed approximately 200 iterations, the overall speedup of our algorithm is roughly a factor of 25 compared to solving the least-squares problem for all data. To get an idea on the influence of the batch size, we also ran an example with The signal-to-noise ratios with respect the true perturbation for this experiment, and an experiment where we increased the number of simultaneous shots and frequencies to 16 and 5, are (with the SNR for the second example in brackets) 2.9 dB for the reference; 1.86 (3.7) dB for $\mathbb{P}_{\ell_2}(\mathbf{RM})$; 2.7 (3.87) dB for $\mathbb{P}_{\ell_2}(\mathbf{RM})$; and 1.5 dB for ℓ_1 without restarts. The improvements are consistent typically seen in CS. Because the system is square for the larger batch size, we do not expect as much of an uplift from ℓ_1.

Algorithm 1: Stochastic-average approximation with warm restarts

$\mathbf{x}_0 \longleftarrow \mathbf{0}; \mathbf{k} \longleftarrow \mathbf{0}$; // initialize
while $\|\mathbf{x}_0 - \widetilde{\mathbf{x}}\|_2 \geq \varepsilon$ **do**
  $k \longleftarrow k+1$; // increase counter
  $\widetilde{\mathbf{x}} \longleftarrow \mathbf{x}_0$; // update warm start
  $\mathbf{RM} \longleftarrow \text{Draw}(\mathbf{RM})$; // draw new subsampler
  $\mathbf{x}_0 \longleftarrow \text{Solve}(\mathbb{P}(\mathbf{RM});\widetilde{\mathbf{x}})$; // solve the subproblem
end

Conclusions

We introduced an efficient algorithm to solve the linearized imaging problem by working on subsets of data. Our method combines recent findings from stochastic optimization and compressive sensing and turns the originally 'overdetermined' seismic imaging problem into a series of underdetermined dimensionality-reduced experiments. By considering these subproblems as compressive-sampling experiments, we were able to create high-fidelity images at a fraction of the computational cost. We found that the inclusion of sparsity promotion in wave-related problems is essential. As in compressive sensing, the quality of the images depends on the dimensionality reduction and the transform-domain sparsity, and hence the complexity, of the model. This leads to the tentative conclusion that we arrived at a formulation where certain aspects of the 'curse of dimensionality' are removed. This removal is particularly important for full-waveform inversion, where each Gauss-Newton update corresponds to a linearized inversion problem. We will report on this topic in another contribution to these proceedings.

Acknowledgments: We would like to thank the sponsors of the SINBAD project. This work was partly supported by a NSERC Discovery Grant (22R81254) and CRD Grant DNOISE II (375142-08).

References

Ben-Hadj-Ali, H., Operto, S. and Vineux, J. [2009] Three-dimensional frequency-domain full waveform inversion with phase encoding. *SEG Technical Program Expanded Abstracts*, **28**(1), 2288–2292, doi:10.1190/1.3255317.

Berg, E.v. and Friedlander, M.P. [2008] Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, **31**(2), 890–912, doi:10.1137/080714488, to appear in *SIAM J. Sci. Comp.*

Bertsekas, D. and Tsitsiklis, J. [1996] *Neuro-Dynamic Programming (Belmont, MA: Athena Scientific)*.

Candès, E., Romberg, J. and Tao, T. [2006] Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, **59**(8), 1207–1223.

Donoho, D.L. [2006] Compressed sensing. *IEEE Trans. Inform. Theory*, **52**(4), 1289–1306.

Fei, T.W., Luo, Y., Aramco, S. and Schuster, G.T. [2010] De-blending reverse-time migration. SEG, vol. 29, 3130–3134, doi:10.1190/1.3513496.

Haber, E., Chung, M. and Herrmann, F.J. [2010] An effective method for parameter estimation with pde constraints with multiple right hand sides. Tech. Rep. TR-2010-4, UBC-Earth and Ocean Sciences Department.

Hennenfent, G., van den Berg, E., Friedlander, M.P. and Herrmann, F.J. [2008] New insights into one-norm solvers from the Pareto curve. *Geophysics*, **73**(4).

Herrmann, F.J., Erlangga, Y.A. and Lin, T. [2009] Compressive simultaneous full-waveform simulation. *Geophysics*, **74**, A35.

Herrmann, F.J. [2010] Randomized sampling and sparsity: Getting more information from fewer samples. *Geophysics*, **75**(6), WB173–WB187, doi:10.1190/1.3506147.

Krebs, J.R. et al. [2009] Fast full-wavefield seismic inversion using encoded sources. *Geophysics*, **74**(6), WCC177–WCC188, doi:10.1190/1.3230502.

Li, X. and Herrmann, F.J. [2010] Full-waveform inversion from compressively recovered model updates. SEG, vol. 29, 1029–1033, doi:10.1190/1.3513022.

Lin, T.T.Y. and Herrmann, F.J. [2007] Compressed wavefield extrapolation. *Geophysics*, **72**(5), SM77–SM93, doi:10.1190/1.2750716.

Morton, S.A. and Ober, C.C. [1998] Faster shot-record depth migrations using phase encoding. *SEG Technical Program Expanded Abstracts*, SEG, vol. 17, 1131–1134, doi:10.1190/1.1820088.

Natterer, F. [2001] *The mathematics of computerized tomography*. Society for Industrial Mathematics.

Neelamani, R.N., Krohn, C.E., Krebs, J.R., Romberg, J.K., Deffenbaugh, M. and Anderson, J.E. [2010] Efficient seismic forward modeling using simultaneous random sources and sparsity. *Geophysics*, **75**(6), WB15–WB27, doi:10.1190/1.3509470.

Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. [2009] Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, **19**(4), 1574–1609.

Paige, C.C. and Saunders, M.A. [1982] Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM TOMS*, **8**(1), 43–71.

Romberg, J. [2009] Compressive sensing by random convolution. *SIAM Journal on Imaging Sciences*, **2**(4), 1098–1128, doi:10.1137/08072975X.

Romero, L.A., Ghiglia, D.C., Ober, C.C. and Morton, S.A. [2000] Phase encoding of shot records in prestack migration. *Geophysics*, **65**(2), 426–436, doi:10.1190/1.1444737.

Strohmer, T. and Vershynin, R. [2009] A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, **15**(2), 262–278.

Symes, W.W. [2008] Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, **56**(765-790).

van Leeuwen, T., Aravkin, A. and Herrmann, F.J. [2010] Seismic waveform inversion by stochastic optimization. Tech. Rep. TR-2010-5, UBC-Earth and Ocean Sciences Department.

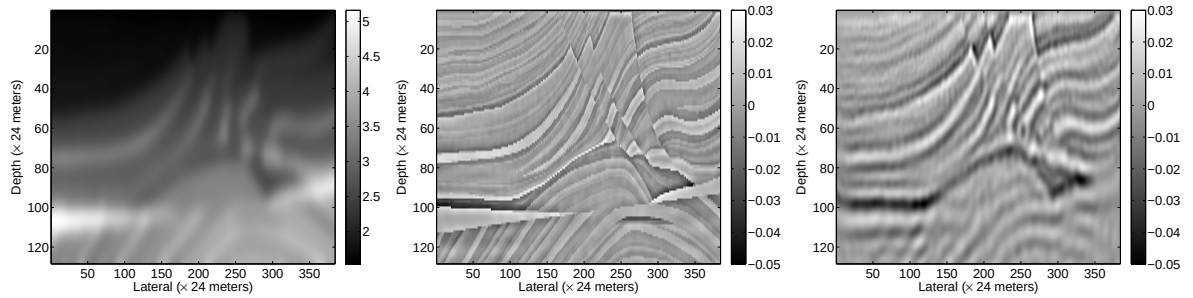

Figure 1 *Least-squares imaging experiment.* ***Left:*** *background velocity model and* ***middle:*** *slowness perturbation.* ***Right:*** *least-squares image obtained with 10 iterations of LSQR with all data, i.e., batch size* $K = 36 \times 10^3$. *The SNR of this image compared to the true perturbation* 2.9 dB.

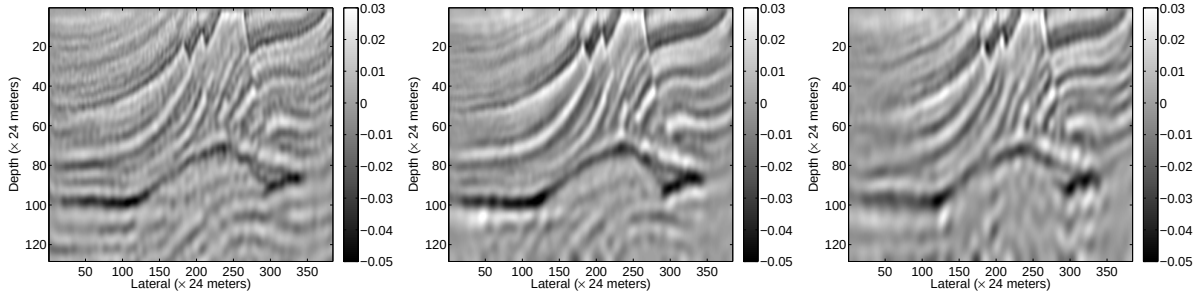

Figure 2 *Comparison between images obtained with dimensionality reduction.* ***Left:*** *Image obtained with LSQR for 10 restarts, yielding* 1.86 dB. ***Middle:*** *image obtained with SPG*ℓ_1 *for approximately the same number of PDE solves with* 2.7 dB. ***Right:*** *the same but without renewals, yielding SNR* 1.5 dB.