APPROXIMATE MESSAGE PASSING MEETS EXPLORATION SEISMOLOGY

Felix J. Herrmann

The University of British Columbia
Department of Earth and Ocean Sciences
Vancouver, Canada

ABSTRACT

Data collection, data processing, and imaging in exploration seismology increasingly hinge on large-scale sparsity promoting solvers to remove artifacts caused by efforts to reduce costs. We show how the inclusion of a "message term" in the calculation of the residuals improves the convergence of these iterative solvers by breaking correlations that develop between the model iterate and the linear system that needs to be inverted. We compare this message-passing scheme to state-of-the-art solvers for problems in missing-trace interpolation and in dimensionality-reduced imaging with phase encoding.

Index Terms— Exploration seismology, compressive sensing, curvelets, sparsity promotion, seismic imaging

1. INTRODUCTION

Modern-day exploration seismology increasingly relies on the solution of large-scale optimization problems that require multiple passes through the data. This leads to major challenges because seismic imaging increasingly depends on long-offset and full-azimuth sampling, which leads to exponential growing costs for seismic data collection, storage, and processing. We discuss how recent insights from compressive sensing, convex optimization, and statistical physics can be used to reduce these processing costs by minimizing the number of required passes through the data. More specifically, we look at the method of approximate message passing (AMP, see e.g. [1]), which can lead to significant improvements in the convergence of large-scale sparsity-promoting solvers for specific sparsity promoting problems that involve properly normalized matrices. The outline is as follows. First, we briefly introduce ideas behind randomized dimensionality reduction that include compressive sensing for seismic acquisition, processing, and imaging. Second, we discuss conventional solution strategies for large-scale sparsity-promoting optimization problems. Third, we introduce approximate message passing including a discussion of its limitations and possible remedies. We conclude by applying this specialized framework to missing-trace interpolation and imaging, which both represent significant excursions from the original framework.

2. RANDOMIZED DIMENSIONALITY REDUCTION

By working on smaller randomized subproblems, while exploiting structure within signals, challenges related to the so-called 'curse of dimensionality' can be addressed in the field of exploration seismology [2, 3]. This curse leads to exponential growth in data-collection and processing costs as the survey area and desired resolution increase. Recent developments in randomized acquisition and efficient imaging with source encoding are examples that address these challenges. However, in both examples cost reductions by randomized subsampling create artifacts, such as aliasing and source crosstalk. Therefore, the challenge has been to devise acquisition techniques that render these interferences incoherent so they can be removed from the data by some sort of processing. In exploration seismology, this processing typically involves a combination of matched and median filtering (see e.g. [4, 5] for processing of simultaneous marine data).

3. COMPRESSIVE SENSING

Unfortunately, matched filtering techniques are of limited value for high degrees of subsampling that lead to high levels of subsampling related interferences. By recognizing recently introduced randomized seismic acquisitions [6, 5], such as acquisition with missing or simultaneous sources, as instances of compressive sensing (or compressive sampling)—championed by Candès, Romberg, and Tao [7] and Donoho [8]—transform-domain sparsity promotion can be used to remove these artifacts. CS has emerged as a novel paradigm for sensing such signals more efficiently as compared to the classical approach based on Shannon-Nyquist sampling theory. Signals that admit sparse approximations can be acquired from significantly fewer measurements than their ambient dimension using nonlinear recovery algorithms, e.g., ℓ_1 minimization.

During sparsity promotion incoherent artifacts, such as spectral leakage or source crosstalk, are mapped back to co-

herent signal. This is done via a iterative procedure during which we transform the data to separate the significant transform coefficients from incoherent interference "noise". Mathematically, this approach corresponds to solving the convex sparsity-promoting program (Basis Pursuit):

$$\text{BP}: \quad \underset{\mathbf{x}}{\text{minimize}} \ \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{Ax} = \mathbf{b}. \quad (1)$$

During recent years, a tremendous body of work has been developed in support of this convex-optimization program where sparse vectors $\mathbf{x} \in \mathbb{C}^N$ are recovered from incomplete measurements $\mathbf{b} = \mathbf{Ax}_0$ with $\mathbf{x}_0$ sparse and $\mathbf{b} \in \mathbb{C}^n$ with $n \ll N$. So far, focus has been on (i) recovery guarantees of k-sparse vectors with sparsity levels $\rho = k/n$ from measurements by sampling matrices $\mathbf{A}$ with aspect ratios $\delta = n/N$; (ii) designing sampling schemes that favour recovery, e.g., the design of randomized acquisition or phase encoding; (iii) implementing fast ($\mathcal{O}(N \log N)$) sparsifying transforms that exploit structure, e.g. curvelets in seismic exploration; (iv) developing algorithms that are frugal in the number of matrix-vector multiplies they require to solve BP.

3.1. Solutions by cooling

Even though the framework of compressive sensing has let to major breakthroughs in many research areas including MRI, the application of its principles to exploration seismology has been challenging because of the large scale of problems that easily involve vectors with billions of unknowns. In addition, BP corresponds to solving the limiting case ($\lambda \downarrow 0$) of the quadratic problem:

$$\text{BP}_\lambda: \quad \underset{\mathbf{x}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{b} - \mathbf{Ax}\|_2 + \lambda\|x\|_1, \quad (2)$$

which is known to converge slowly as a function of the number of matvec's as the trade-off parameter $\lambda \downarrow 0$. To overcome this problem, 'optimizers' use continuation methods that employ properties of the Pareto tradeoff curve that traces the one-norm of the solution against the two-norm of the residual. This leads to approaches were series of subproblems are solved that allow components to enter the solution contollably by slowly increasing the one-norm of the solution. Each subproblem involves gradient updates, $\mathbf{x}^{t+1} = \mathbf{x}^t + \mathbf{A}^H(\mathbf{b} - \mathbf{Ax}^t)$ with t the iteration number, in combination with a nonlinear projection promoting sparsity. Not withstanding the success of these continuation methods, which undergird state-of-the-art versatile solvers such as SPGℓ_1 [9], convergence for large systems remains challenging in particular when given memory limitations and a small budget ($\mathcal{O}(50)$) of matvecs. However, this type of solvers provably solve BP, and the related problems BPDN and LASSO [9].

3.2. Solutions by approximate message passing

To understand the performance of solvers for BP, consider the second row of Fig. 1 that includes the first two model
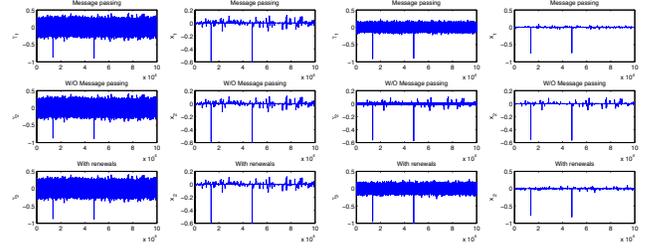
Fig. 1. Model iterates after one (left) and and two (right) iterations before and after soft thresholding. Notice the spurious spiky artifacts that develop after the first iteration in the second row. These correlations are removed by the message term (top row) or by drawing a new sampling matrix and data for each iteration (bottom row).

iterates before and after soft thresholding for $\mathbf{A}$ a Gaussian 10×10^5 matrix and $\mathbf{x}$ a vector with two ones at random locations. (Remember that soft thresholding is an elementwise operation that jointly minimizes the two-norm on the residual and the absolute value of the output and is given by $\eta(x, \tau) = \text{sgn}(x)\max(|x| - \tau, 0)$, with $\tau > 0$ the threshold value.) We use Gaussian matrices because they lead to Gaussian interferences, i.e., $\mathbf{A}^H \mathbf{Ax}_0 = \mathbf{x}_0 + \mathbf{w}$ with $\mathbf{w}$ zero-mean Gaussian noise with variance $n^{-1}\|\mathbf{x}_0\|_2^2$ [10]. This property, which is common amongst compressive-sensing matrices that favor recovery by BP, reduces the recovery problem to a simple "denoising" problem, which can be solved by soft thresholding with the proper threshold level. Unfortunately, this property no longer holds after the first iteration because dependencies emerge between the model iterate $\mathbf{x}^t$ and the matrix $\mathbf{A}$ for iterations $t > 1$. These dependencies cause spurious artifacts and this leads to slow convergence and the algorithm spends many iterations to remove these wrongly identified entries.

Suppose now that we select a new matrix $\mathbf{A}$ and corresponding measurement vector $\mathbf{b}$ at each iteration as suggested by [10]. In that case, correlations can no longer develop between the model iterate and the matrix rendering soft thresholding effective for each iteration (juxtapose the second and third row of Fig. 1). While this idea has been used successfully in situations where data is abundant, e.g., in seismic imaging where different independent randomized subsets of fully-sampled data volumes are drawn to speed up convergence [2], this solution is unworkable in data-scarce situations, such as during the recovery from incomplete field data.

Using arguments from statistical physics, [1] addresses this issue by including a 'message-passing' term in iterative soft thresholding schemes that solve BP. We iterate the following approximate message-passing scheme (AMP):

$$\begin{aligned} \mathbf{x}^{t+1} &= \eta\left(\mathbf{A}^H \mathbf{r}^t + \mathbf{x}^t; \theta_t\right) \\ \mathbf{r}^t &= \mathbf{b} - \mathbf{Ax}^t + \frac{I_t}{n}\mathbf{r}^{t-1}, \end{aligned} \quad (3)$$

with $\mathbf{x}^0 = 0$, $\mathbf{r}^0 = \mathbf{b}$. Here, $\eta(\cdot; \theta_t)$ is a iteration-dependent soft thresholding, and I_t is the number of entries that survived

the threshold of the previous iteration. As we can see from the first row in Fig. 1, inclusion of this extra term in Eq. 3 cancels the correlations and corresponds to adding the 'residual' of the previous iteration scaled by the ratio of the number of elements that survived the threshold and the number of measurements. This extra term clearly breaks the spurious interferences and guarantees that each iteration becomes a simple denoising problem ideally suited for soft thresholding. Clearly, the implication of including this extra term is profound because it leads to much faster convergence of the algorithm as illustrated in Fig. 1. These improvements prove to be important if we want to solve large scale (order of billions unknowns) sparse recovery problems.

3.3. Large dimensionality: a blessing in disguise

The theory explaining the improved performance by adding the message term in Eq. 3 is involved but can be summarized as follows. First, the message term leads to asymptotic cancellation of the damaging correlations and is valid for system sizes going to infinity ($N \to \infty$) for recovery problems with Gaussian matrices. In that limit, [10] argues that the linear system of equations decouples such that the recovery for each entry of the model iterate boils down to estimating x_i^t using the property $\left(\mathbf{x}^t + \mathbf{A}^H \mathbf{r}^t \right)_i = (\mathbf{x} + \tilde{\mathbf{w}})_i$ for $i = 1 \cdots N$. This can be done effectively by soft thresholding using the fact that the $\{\tilde{w}_i\}_{i=1 \cdots N}$ are asymptotically Gaussian in the large-scale limit. Second, and this is somewhat of a hand waiving argument, large matrices tend to behave as Gaussian matrices even if their entries are not Gaussian or if they are obtained by some other randomization such as random phase encoding or random restriction. For example, the randomly restricted Fourier matrix behaves approximately as a Gaussian matrix. Third, the iterative procedure with the message-passing term solves BP and therefore recovers k-sparse vectors with high probability if certain conditions are met on the aspect ratio of $\mathbf{A}$ and the sparsity level of the vector $\mathbf{x}_0$ (see e.g., [1] for a detailed overview on phase diagrams that predict the transition from recoverable to non-recoverable combinations of (ρ, δ)). Now, if these large-scale limit asymptotic arguments indeed hold then the message-passing algorithm is truly remarkable because we may be able to improve the performance of highly sophisticated one-norm solvers for exceedingly large problem sizes. Evidently, this promise comes with the caveat that message-passing algorithms are specifically designed to solve sparse-recovery problems for Gaussian matrices while methods such as SPGℓ_1 are versatile and solve BP for any $\mathbf{A}$ and $\mathbf{b}$ as long one is willing to spend a sufficient number of iterations to bring the residual down.

4. AMP IN EXPLORATION SEISMOLOGY

Unfortunately, physical constraints of seismic acquisition and imaging lead to matrices that do not meet assumptions of approximate message passing. Because these matrices are ex-

tremely large they can not be formed explicitly, which makes simple operations such as scaling of the columns challenging. We discuss two possibilities that allow us to still benefit from AMP even though we are outside the scope where the theory applies. Aside from requiring a Gaussian sensing matrix $\mathbf{A}$, AMP also requires the column norm of $\mathbf{A}$ to be approximately normalized to one, which is accomplished by generating the entries from $1/\sqrt{n} \cdot N(0, 1)$. Suppose now, that we are given $\mathbf{A} \to \mathbf{A} \Gamma^{1/2}$ with Γ a diagonal matrix with unknown positive entries. The questions we can ask ourselves now are: (i) how much can the entries of Γ deviate from unity, (ii) how accurate do we need to estimate these entries from the actions of $\mathbf{A}$, and (iii) are there situations in exploration seismology where the ideas from AMP can still be used but where corrections for the diagonal are infeasible.

4.1. Randomized estimation of the diagonal

We experimented with matrices $\mathbf{A}$ with a random Γ drawn from a uniform distribution with mean one. In this case, AMP fails to solve recovery problem for $\mathbf{A} \in \mathbb{R}^{200 \times 10^4}$ and $k = 20$ nonzeros. Following [11], we estimate the diagonal via

$$\widehat{\Gamma} = \left[\sum_{i=1}^K \mathbf{w}_i \otimes \mathbf{A}^H \mathbf{A} \mathbf{w}_i \right] \oslash \left[\sum_{i=1}^K \mathbf{w}_i \otimes \mathbf{w}_i \right], \quad (4)$$

with $\mathbf{w}_i$ random vectors and $\otimes$ and $\oslash$ elementwise multiplication and devision. With this estimate for Γ obtained with K trial vectors, we are able to solve for $\mathbf{x}$ by replacing the single-threshold soft thresholding operator in Equation 3 by varying thresholding $\eta(\cdot; \mathrm{diag}(\widehat{\Gamma})\theta_t)$ with the threshold multiplied by the estimate for Γ. While this procedure works in principle, the convergence of the estimator in Equation 4 is slow and deteriorates for fat matrices $\mathbf{A}$ ($n \ll N$) requiring a prohibitively large number of matvecs.

4.2. Curvelet-based sparse recovery

While AMP is strictly speaking designed for Gaussian matrices and strictly sparse vectors only, we examine its performance for the recovery of a seismic line from 50% randomly missing sources using roughly 100 matvec's. We use the restricted 3D curvelet transform for the recovery. Output shot records for SPGℓ_1 and AMP are plotted in Fig. 2 and show clear improvements for AMP despite violations of the underlying assumptions. (The restricted 3D curvelet transform matrix is not a Gaussian matrix and the vector is not strictly sparse but compressible.) The SNRs, 7.75dB for SPGℓ_1 and 9.75dB for AMP, confirm this observation. Aside from this remarkable improvement, the residue is significantly smaller after 50 iterations and the solution paths for AMP points to the ℓ_1-norm of the curvelet coefficient vector obtained by solving SPGℓ_1 for 300 matvecs on the complete data. This is remarkable and highly encouraging because this is a very large-scale problem ($N = 1.12 \times 10^9$) with a matrix that is remote from a Gaussian matrix.
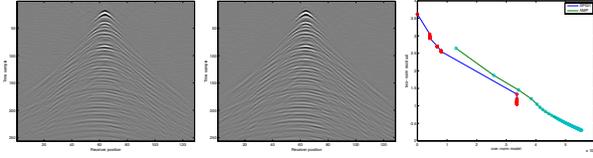
Fig. 2. Recovery from 50% missing shots with SPGℓ_1(left) and AMP (middle). Solution paths (right).

4.3. Efficient imaging

Aside from improvements in the recovery of missing data, message-passing ideas also provide an explanation for efficient sparsity-promoting imaging where new subsets of simultaneous shots are drawn each time one of the subproblems of SPGℓ_1 is solved [2]. Following Herrmann and Li [2], we replace the overdetermined least-squares imaging problem

$$\underset{\Delta\mathbf{x}}{\text{minimize}} \ \frac{1}{2K}\sum_{i=1}^{K}\|\Delta\mathbf{b}_i - \mathbf{A}_i\Delta\mathbf{x}\|_2^2 \qquad (5)$$

with K the number of sources, by the more efficient underdetermined sparsity-promoting program with $K' \ll K$ sources:

$$\underset{\Delta\mathbf{x}}{\text{minimize}} \ \|\mathbf{x}\|_1 \quad \text{subject to} \quad \Delta\underline{\mathbf{b}}_i = \underline{\mathbf{A}}_i\Delta\mathbf{x}, \qquad (6)$$

$i = 1 \cdots K'$. Here, $\Delta\mathbf{b}_i$ and $\Delta\underline{\mathbf{b}}_i$ are the vectorized sequential and simultaneous shot records with linearized data (residue), $\mathbf{A}_i$ and $\underline{\mathbf{A}}_i$ are the linearized Born scattering matrices for sequential and simultaneous sources, and $\Delta\mathbf{x}$ is the unknown seismic image (or its curvelet representation). The sparsity-promoting formulation is more efficient because (i) we work with fewer numer of simultaneous sources or super shots ($K' \ll K$) and (ii) we can use the fact that

$$\begin{aligned} \mathbf{x}^{t+1} &= \eta\left(\mathbf{A}^H(t)\mathbf{r}^t + \mathbf{x}^t; \theta_t\right) \\ \mathbf{r}^t &= \mathbf{b}(t) - \mathbf{A}(t)\mathbf{x}^t, \end{aligned} \qquad (7)$$

is statistically equivalent to Equation 2 . New 'sensing' matrices $\mathbf{A}(t)$ and data $\mathbf{b}(t)$ are drawn for each iteration, which motivated us [2] to adapt SPGℓ_1 by selecting new simultaneous shots after solving each LASSO subproblem. Fig. 3 shows that these renewals have a similar positive effect as the message term in Eq. 3. The renewals break correlations that build up between the model iterate and the 'sensing' matrix.

Acknowledgements

5. REFERENCES

[1] D.L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.

[2] Felix J. Herrmann and Xiang Li, "Efficient least-squares imaging with sparsity promotion and compressive sensing," *Geophysical Prospecting*, 2012, 10.1111/j.1365-2478.2011.01041.x.

[3] F.J. Herrmann, M.P. Friedlander, and O. Yilmaz, "Fighting the curse of dimensionality: Compressive sensing in exploration seismology," *Signal Processing Magazine, IEEE*, vol. 29, no. 3, pp. 88 –100, may 2012.

[4] G. Hampson, J. Stefani, and F. Herkenhoff, "Acquisition using simultaneous sources," *The Leading Edge*, vol. 27, no. 7, pp. 918–923, 2008.

[5] H. Mansour, H. Wason, T. T.Y. Lin, and F. J. Herrmann, "Simultaneous-source marine acquisition with compressive sampling matrices," To appear in Geophysical Prospecting, 2011.

[6] F. J. Herrmann, "Randomized sampling and sparsity: Getting more information from fewer samples," *Geophysics*, vol. 75, no. 6, pp. WB173–WB187, 2010.

[7] E.J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[8] D.L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[10] A. Montanari, *Graphical Models Concepts in Compressed Sensing*, Cambridge University press, 2012, Compressed Sensing: Theory and Applications.

[11] C. Bekas, E. Kokiopoulou, and Y. Saad, "An estimator for the diagonal of a matrix," *Applied Numerical Mathematics*, vol. 57, no. 1112, pp. 1214 – 1229, 2007.
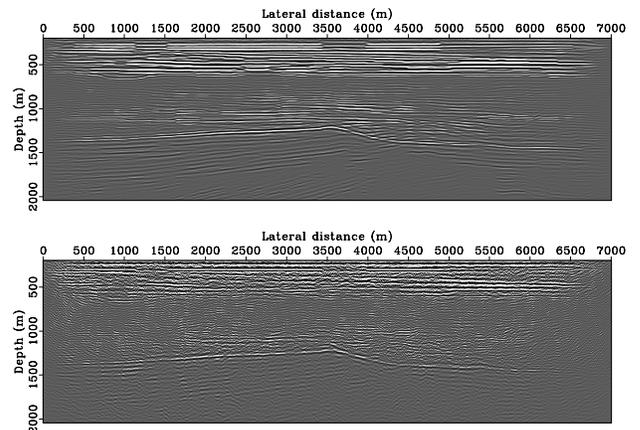
Fig. 3. Imaging with and without 'messaging' for 3 super shots.