Pass on the message: recent insights in large-scale sparse recovery

Felix J. Herrmann

thanks to Xiang Li
Big data

“We are drowning in data but starving for understanding” USGS director Marcia McNutt

“Got data now what” Carlsson & Ghrist SIAM
Problem

"Data explosion is bigger than Moore's law"
Goals

Replace a ‘sluggish’ processing paradigm that

- relies on touching all data all the time

by an agile optimization paradigm that works on

- small randomized subsets of data iteratively

Confront “data explosion” by

- reducing acquisition costs
- removing IO & PDEs-solve bottlenecks
Compressive sensing

randomized sampling
detection + data-consistent amplitude recovery

\[
\begin{aligned}
\min_x \|x\|_1 & \quad \text{subject to} \\
A & = RF^H \\
b & = Ax
\end{aligned}
\]

\(A \in \mathbb{C}^{n \times N}\) with \(n \ll N\)
Convex optimization

Sparse recovery involves iterations of the type

$$x^{t+1} = \eta_t \left( A^* r^t + x^t \right)$$

$$r^t = b - Ax^t$$

Corresponds to vanilla “denoising” if $A$ is a Gaussian matrix.

But does the same hold for later ($t>1$) iterations...?
Iteration $t=1$

$$A^* r^t + x^t$$

$$\eta_t(A^* r^t + x^t)$$
Iteration $t=2$

$$A^* r^t + x^t$$

$$\eta_t(A^* r^t + x^t)$$
Iteration $t=3$

$$A^* r^t + x^t$$

$$\eta_t(A^* r^t + x^t)$$
Iteration $t=4$

\[ A^* r^t + x^t \]

\[ \eta_t(A^* r^t + x^t) \]
Approximate message passing

Add a term to iterative soft thresholding, i.e.,

\[
\begin{align*}
x^{t+1} &= \eta_t \left( A^* r^t + x^t \right) \\
r^t &= b - Ax^t + \frac{\|x^{t+1}\|_0}{n} r^{t-1}
\end{align*}
\]

Holds for

- normalized Gaussian matrices \( A_{ij} \in n^{-1/2} N(0, 1) \)
- large-scale limit and for specific thresholding strategy

[Donoho et. al, ’09–’12; Montanari, ’10–’12, Rangan, ’11]
Approximate message passing

Statistically equivalent to

\[
x^{t+1} = \eta_t \left( A^*_t r^t + x^t \right)
\]

\[
r^t = b_t - A_t x^t
\]

by drawing new independent pairs \( \{b_t, A_t\} \) for each iteration

Changes the story completely

- breaks correlation buildup between model iterate \( x^t \) & the matrix \( A \)
- faster convergence

[Montanari, ’12]
Iteration $t=1$

\[ \mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \eta_t (\mathbf{A}^* \mathbf{r}^t + \mathbf{x}^t) \]

\[ \mathbf{r}^t = \mathbf{b}_t - \mathbf{A}_t \mathbf{x}^{10^4}_t \]

\[ \eta_t (\mathbf{A}_t^* \mathbf{r}^t + \mathbf{x}^{10^4}_t) \]
Iteration $t=2$

$$r^t = b - Ax^t + \eta_t (A^*r^t + x^t)$$

Message passing

$\gamma_1$

W/O Message passing

$\gamma_2$

With renewals

$\gamma_3$

$\eta_t (A^*r^t + x^t)$
Iteration $t=3$

$$r^t = b - Ax^t + \frac{||x^{t+1}||_0}{\eta} r^{t-1}$$

Message passing

$$\eta_t (A^* r^t + x^t)$$

Message passing

With renewals

$$r^t = b_t - A_t x^{t_0}$$

W/O Message passing

$$r^t = \eta_t (A^* r^t + x^t)$$
Iteration $t=4$

$$\mathbf{r}^t = \mathbf{b} - \mathbf{A} \mathbf{x}^t + \frac{\mathbf{x}^{t+1}}{0} \mathbf{r}^{t-1}$$

$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$

Message passing

- $\gamma_1$
- $\gamma_2$
- $\gamma_3$

W/O Message passing

- $\chi_1$
- $\chi_2$

With renewals

- $\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)^{10^4}$

$\mathbf{r}^t = \mathbf{b}_t - \mathbf{A}_t \mathbf{x}^{t10^4}$
Missing-trace interpolation [SPGI1]

Recovery with 3D curvelets (N=1.12 × 10^9) 7.75 dB

50 % missing data

recovery

50 iterations

difference
Missing trace interpolation [AMP]

Recovery with 3D curvelets ($N=1.12 \times 10^9$) 9.75 dB

50% missing data  
recovery  
50 iterations  
difference
Observations

Message-pass term has the same effect as drawing independent experiments \( \{b_t, A_t\} \)

- “Gaussian” matrices
- delicate normalization and thresholding strategy
- renders proposed method impractical
- can lead to dramatically improved convergence

How can we still reap benefits from message passing in realistic less-than-ideal geophysical settings?
**Supercooled spectral-projected gradients**

![Lasso problem](min \|A_1 x - b_1\|_2 \ s.t. \|x\|_1 \leq \tau^1)

- [van den Berg & Friedlander, '08]
- [Hennefent et al., '08]
- [Lin & FJH, '09-]
Supercooled spectral-projected gradients

Fig. 6.1. Corrupted and interpolated images for problem seismic. Graph (c) shows the Pareto curve and the solution path taken by SPGL1.

However, as might be expected of an interior-point method based on a conjugate-gradient linear solver, it can require many matrix-vector products.

It may be progressively more difficult to solve \((\text{QP}_\lambda)\) as \(\lambda \to 0\) because the regularizing effect from the one-norm term tends to become negligible, and there is less control over the norm of the solution. In contrast, the \((\text{LS}_\tau)\) formulation is guaranteed to maintain a bounded solution norm for all values of \(\tau\).

6.4. Sampling the Pareto curve.

In situations where little is known about the noise level \(\sigma\), it may be useful to visualize the Pareto curve in order to understand the trade-offs between the norms of the residual and the solution. In this section we aim to obtain good approximations to the Pareto curve for cases in which it is prohibitively expensive to compute it in its entirety.

We test two approaches for interpolation through a small set of samples \(i = 1, \ldots, k\). In the first, we generate a uniform distribution of parameters \(\lambda_i = (i/k) \Vert A^T b \Vert_\infty\) and solve the corresponding problems \((\text{QP}_{\lambda_i})\). In the second, we generate a uniform distribution of parameters \(\sigma_i = (i/k) \Vert b \Vert_2\) and solve the corresponding problems \((\text{BP}_{\sigma_i})\).

We leverage the convexity and differentiability of the Pareto curve to approximate it with piecewise cubic polynomials that match function and derivative values at each end. When a nonconvex fit is detected, we switch to a quadratic interpolation.

\[
\text{Lasso problem} \quad \min_x \| A_2 x - b_2 \|_2 \quad \text{s.t.} \quad \|x\|_1 \leq \tau^2
\]
Supercooled
spectral-projected gradients

Lasso problem
\[
\min_x \|A_3 x - b_3\|_2 \quad s.t \quad \|x\|_1 \leq \tau^3
\]
**Supercooled spectral-projected gradients**

---

**Lasso problem**

\[
\begin{align*}
\min_x & \quad \|A_4 x - b_4\|_2 \\
\text{s.t.} & \quad \|x\|_1 \leq \tau_4
\end{align*}
\]
**Supercooling**

Break *correlations* between the model *iterate* and matrix $A$ by *rerandomization*

- draw new *independent* $\{b_t, A_t\}$ after each LASSO subproblem is solved
- brings in “extra” information *without* growing the system

- *minimal* extra computational & memory cost
Ideal ‘Seismic’ example
\[ n/N = 0.13; N = 248759; T = 500 \]
Ideal ‘Seismic’ example
\[ n/N=0.13; N=248759; T=500 \]

**Cooled**

*recovery*  
*error*  
*solution path*
Ideal ‘Seismic’ example

\[ n/N=0.13;N=248759;T=500 \]

10 X

- Recovery
- Error
- Solution path
Independent redraws of $\{b_t, A_t\}$ get rid of small difficult to remove interferences

- working only with subsets of the data

But, aren’t we fooling ourselves since proposed method

- defeats the premise of compressive sampling

Or, are there data-rich applications for this method?

- e.g. efficient imaging with random source encoding
Compressive imaging
[with message passing]

Select independent random source encodings after each LASSO subproblem is solved

- calculate corresponding supershots
- redefine demigration operator (and its adjoint)
  (select independent simultaneous sources & supershots)

Promote sparsity in the curvelet domain
Imaging results

Time-harmonic Helmholtz:

- 409 X 1401 with mesh size of 5m
- 9 point stencil [C. Jo et. al., ’96]
- absorbing boundary condition with damping layer with thickness proportional to wavelength
- solve wavefields on the fly with direct solver
Imaging results
[background model]
Migration results

[true perturbation]
Migration results
[migration with “all” data]
Imaging results

Reduced setup:

- 10 random frequencies (versus 300 frequencies) (20Hz-50Hz)
- 3 random simultaneous shots (versus 350 sequential shots)

Significant dimensionality reduction of

\[ \frac{K'}{K} = 0.0003 \]

[Herrmann & Li, 2011]
Imaging results

Least-squares migration with randomized supershots:

\[
\delta \tilde{m} = S^* \arg \min_{\delta x} \| \delta x \|_2 \quad \text{subject to} \quad \| \delta d - \nabla F [m_0; Q] S^* \delta x \|_2 \leq \sigma
\]

- \( \delta x \) = Sparse curvelet-coefficient vector
- \( S^* \) = Curvelet synthesis
- \( Q \) = Simultaneous sources
- \( \delta d \) = Super shots
Imaging results

Sparsity-promoting migration with *randomized supershots*:

\[
\delta \tilde{m} = S^* \arg \min_{\delta x} \| \delta x \|_{\ell_1} \quad \text{subject to} \quad \| \delta d - \nabla F[m_0; Q] S^* \delta x \|_2 \leq \sigma
\]

\[
\begin{align*}
\delta x &= \text{Sparse curvelet-coefficient vector} \\
S^* &= \text{Curvelet synthesis} \\
Q &= \text{Simultaneous sources} \\
\delta d &= \text{Super shots}
\end{align*}
\]
Migration results

$[\ell_2 \text{ without renewals}]$
Imaging results

[$\ell_1$ without renewals]
Migration results

\[ \ell_2 \text{ with renewals} \]
Migration results

[ $\ell_1$ with renewals]
Migration results

[true perturbation]
Migration results
[migration with “all” data]
Migration results

[model errors]

![Graph showing migration results with different norms and with or without renewals. The x-axis represents the number of passing through all the data, and the y-axis shows the relative two norm of model error. The graph includes lines for L2 with renewals, L1 with renewals, L2 without renewals, and L1 without renewals. The lines demonstrate the reduction in model error as the number of passes increases.]
Conclusions

Message passing improves image quality
  - computationally feasible one-norm regularization

Message passing via rerandomization
  - small system size with small IO and memory imprints

Possibility to exploit new computer architectures that employ model space parallelism to speed up wavefield simulations...
Acknowledgments

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Collaborative Research and Development Grant DNOISE II (375142-08).

We also would like to thank the authors of CurveLab.

This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BGP, BP, Chevron, ConocoPhillips, Petrobras, PGS, Total SA, and WesternGeco.
Further reading

Simultaneous & continuous acquisition:

– A new look at simultaneous sources by Beasley et. al., ’98.
– Changing the mindset in seismic data acquisition by Berkhout ’08.

Simultaneous simulations, imaging, and full-wave inversion:

– Phase encoding of shot records in prestack migration by Romero et. al., ’00.
– Efficient Seismic Forward Modeling using Simultaneous Random Sources and Sparsity by N. Neelamani et. al., ’08.
– Compressive simultaneous full-waveform simulation by FJH et. al., ’09.
– Randomized dimensionality reduction for full-waveform inversion by FJH & X. Li, ’10
– Fast full-wavefield seismic inversion using encoded sources by Krebs et. al., ’09
– An effective method for parameter estimation with PDE constraints with multiple right hand sides. by Eldad Haber, Matthias Chung, and Felix J. Herrmann. ’10
– Efficient least-squares imaging with sparsity promotion and compressive sensing by FJH & Li, ’12
– Fast randomized full-waveform inversion with compressive sensing by Xiang Li et. al., ’12
Further reading

Compressive sensing & sparse solvers
- Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information by Candes, ’06.
- Compressed Sensing by D. Donoho, ’06
- Probing the Pareto frontier for basis pursuit solutions by E. van den Berg and M. Friedlander, ’08

Machine learning & message passing
- Message passing algorithms for compressed sensing by David Donoho et. al., 2009
Thank you

www.slim.eos.ubc.ca