Near Ideal Behavior of a Modified Elastic Net Algorithm in Compressed Sensing

M. Vidyasagar

Cecil & Ida Green Chair
The University of Texas at Dallas
M.Vidyasagar@utdallas.edu
www.utdallas.edu/∼m.vidyasagar

Glover Fest Version 2.0
Cambridge, 24 September 2013

Outline

Outline

Compressed Sensing: Rough Formulation

Simple Version:

Knowing that an n-dimensional vector x has very few nonzero components (say k), but *not* knowing the *locations* of the nonzero components,

- Is it possible to recover x *exactly* by making $m \ll n$ *noise-free linear* measurements?
- Is it possible to recover x *approximately* by making $m \ll n$ *noisy linear* measurements?

Precise Formulation

Define the set of k-**sparse** vectors in $\mathbb{R}^n$:

$$\Sigma_k = \{x \in \mathbb{R}^n : |\text{supp}(x)| \leq k\},$$

where $\text{supp}(x) = \{i : x_i \neq 0\}$ is the **support** of x.

Is it possible to choose (a) an integer $m \ll n$, (ii) a matrix $A \in \mathbb{R}^{m \times n}$, and (iii) a "demodulation" map $\Delta : \mathbb{R}^m \to \mathbb{R}^n$, such that

- $\Delta(Ax) = x \; \forall x \in \Sigma_k$?
- $\|\Delta(Ax + \eta) - x\| \leq c\|\eta\| \; \forall x \in \Sigma_k$, where c is a "universal" constant that does not depend on x or η?

Note: Measurements are linear, but demodulation can be highly nonlinear.

Rough Formulation (Cont'd)

Suppose $x \in \mathbb{R}^n$ is "nearly k-sparse," though not exactly so. Suppose we have $m \ll n$ exact or noisy linear measurements of x. Is it possible to recover a k-sparse approximation of x?

Signal Compression Interpretation: Suppose x represents the Fourier coefficients of a periodic signal, and only k coefficients are "significant." Can we construct a good approximation of x *without* knowing *which* Fourier coefficients are significant?

Precise Formulation (Cont'd)

Define the k-**sparsity index** of x in the norm $\| \cdot \|$.

$$\sigma_k(x, \| \cdot \|) = \inf\{\|x - z\| : z \in \Sigma_k\}.$$

Note: $\sigma_k(x, \| \cdot \|)$ depends on the norm $\| \cdot \|$.

Question: Is it possible to choose an integer $m \ll n$, a matrix $A \in \mathbb{R}^{m \times n}, m \ll n$, and a "demodulation" map $\Delta : \mathbb{R}^m \to \mathbb{R}^n$, such that

$$\|\Delta(Ax) - x\|_2 \leq C_0 \sigma_k(x, \| \cdot \|_1) \ \forall x \in \mathbb{R}^n?$$

$$\|\Delta(Ax + \eta) - x\|_2 \leq C_0 \sigma_k(x, \| \cdot \|_1) + C_2 \|\eta\|_2?$$

for "universal" constants C_0 and C_2?

Note mixture of ℓ_1- and ℓ_2-norms! More on this later.

Outline

Outline

Restricted Isometry Property (RIP)

Note: *Not* most general result, but easy to state!

A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the RIP (Restricted Isometry Property) of order k with constant δ_k if

$$(1 - \delta_k)\|u\|_2^2 \leq \|Au\|_2^2 \leq (1 + \delta_k)\|u\|_2^2, \ \forall u \in \Sigma_k.$$

Interpretation: Every set of k or fewer columns of A is "nearly orthonormal."

Precisely, if we take columns of A from the set $J \subseteq \{1, \ldots, n\}$, call the submatrix A_J, then all eigenvalues of $A_J^t A_J$ lie in the interval $[1 - \delta_k, 1 + \delta_k]$ whenever $|J| \leq k$.

Candès-Tao Result on ℓ_1-Norm Minimization

Theorem: (Candès-Tao (2005); see also Donoho (2006)). Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the RIP of order δ_{2k} with constant $\delta_{2k} < \sqrt{2} - 1$, and that $y = Ax$ for some $x \in \Sigma_k$. Define

$$\hat{x} = \operatorname*{argmin}_z \|z\|_1 \text{ s.t. } y = Az.$$

Then $\hat{x} = x$.

Note: Problem at hand is a linear programming problem.

Exact recovery of sparse vectors, if only we can design a matrix A that satisfies RIP.

Designing Matrices with RIP

(Candès-Tao (2005)): Choose columns of A to be realizations of m-dimensional zero-mean Gaussians. Then with "high probability" (which can be computed), A satisfies RIP.

Difficulty: Resulting A matrix has all nonzero entries with probability one – *implementation issues!*.

(Achlioptas (2003)): Choose columns of A to be realizations of i.i.d. (independent and identically distributed) random process $\{X_t\}$ assuming values in $\{-1, 0, +1\}$, with

$$\Pr\{X_t = -1\} = \Pr\{X_t = +1\} = \epsilon, \Pr\{X_t = 0\} = 1 - 2\epsilon.$$

Benefit: Resulting A matrix is very sparse, *no implementation issues,* and also satisfies RIP with "high probability."

Outline

1. Introduction
2. Some Known Results
   - Exact Measurements
   - Noisy Measurements
3. New Results
4. An Open Problem(?)
5. References

Defining Near Ideal Behavior

Suppose $x \in \Sigma_k$, and we measure $y = Ax + \eta$, where $\|\eta\|_2 \leq \epsilon$, where ϵ is known. An "oracle" would know the support set J of x, and then (in obvious notation)

$$y = A_J x_J + \eta.$$

So estimate and estimation error of the oracle are

$$\hat{x} = (A_J^t A_J)^{-1} A_J^t y,$$

$$\hat{x} - x = (A_J^t A_J)^{-1} A_J^t \eta,$$

$$\|\hat{x} - x\|_2 \leq \text{const.}\epsilon.$$

An algorithm is **near ideal** if, *without knowing the support of x*, it achieves an error proportional to ϵ, for all $x \in \Sigma_k$.

A General Theorem

Theorem: (Candès-Plan (2009); see also DDEK (2012)). Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the RIP of order δ_{2k} with constant $\delta_{2k} < \sqrt{2} - 1$, and that $y = Ax + \eta$ for some $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$ with $\|\eta\|_2 \leq \epsilon$.

$$\hat{x} = \operatorname*{argmin}_z \|z\|_1 \text{ s.t. } \|y - Az\|_2 \leq \epsilon.$$

Then

$$\|\hat{x} - x\|_2 \leq C_0 \frac{\sigma_k(x, \|\cdot\|_1)}{\sqrt{k}} + C_2 \epsilon,$$

where

$$C_0 = 2 \frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (\sqrt{2} + 1)\delta_{2k}}, C_2 = \frac{4\sqrt{1 + \delta_{2k}}}{1 - (\sqrt{2} + 1)\delta_{2k}}.$$

Outline

LASSO and the Elastic Net Algorithms

The problem

$$\min_z \|z\|_1 \text{ s.t. } \|y - Az\|_2 \leq \epsilon$$

is roughly equivalent to LASSO (Tibshirani (1998)). Candès-Plan result shows that "LASSO exhibits near ideal behavior."

Better numerical behavior compared to LASSO results from the Elastic Net (EN) algorithm (Zou-Hastie (2005)):

$$\min_z [(1-\mu)\|z\|_1 + \mu\|z\|_2^2] \text{ s.t. } \|y - Az\|_2 \leq \epsilon.$$

Question: Does EN algorithm also have "near ideal behavior"?

A Modified Elastic Net Algorithm

Difficulty: The quantity

$$(1-\mu)\|z\|_1 + \mu\|z\|_2^2$$

isn't a norm!

Modified Elastic Net (MEN) Algorithm:

$$\min_z [(1-\mu)\|z\|_1 + \mu\|z\|_2] \text{ s.t. } \|y - Az\|_2 \leq \epsilon.$$

Compare with EN:

$$\min_z [(1-\mu)\|z\|_1 + \mu\|z\|_2^2] \text{ s.t. } \|y - Az\|_2 \leq \epsilon.$$

Near Ideal Behavior of MEN Algorithm

Theorem (MV CDC 2013): Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the RIP of order $2k$ with constant $\delta_{2k} < \sqrt{2} - 1$, and that $y = Ax + \eta$ for some $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$ with $\|\eta\|_2 \leq \epsilon$. Define

$$\hat{x}_{\mathrm{MEN}} := \operatorname*{argmin}_z \|z\|_\mu \text{ s.t. } \|y - Az\|_2 \leq \epsilon.$$

Then, for μ sufficiently small, there exist constants $C_{0,\mu}$ and $C_{2,\mu}$ such that Then

$$\|\hat{x}_{\mathrm{MEN}} - x\|_2 \leq C_{0,\mu} \frac{\sigma_k(x, \| \cdot \|_1)}{\sqrt{k}} + C_{2,\mu}\epsilon.$$

Moreover when $\mu = 0$ these reduce to earlier constants.

$$C_0 = 2\frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (\sqrt{2} + 1)\delta_{2k}}, C_2 = \frac{4\sqrt{1 + \delta_{2k}}}{1 - (\sqrt{2} + 1)\delta_{2k}}.$$

A Useful Corollary

Theorem: Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the RIP of order δ_{2k} with constant $\delta_{2k} < \sqrt{2} - 1$, and that $y = Ax$ for some $x \in \Sigma_k$. Define

$$\hat{x} = \operatorname*{argmin}_z \|z\|_\mu \text{ s.t. } y = Az.$$

Then $\hat{x} = x$ provided μ is sufficiently small.

In short, there are *infinitely many norms* $\| \cdot \|_\mu$ that permit exact recovery of sparse signals.

Advantages of MEN Algorithm

Minimizing $\| \cdot \|_1$ is a quadratic program. What are the advantages of minimizing$\| \cdot \|_\mu$?

- $\| \cdot \|_\mu$ is *strictly convex*, whereas $\| \cdot \|_1$ is not. So MEN algorithm *always* produces a unique solution.
- EN has better numerical behavior than LASSO.
  - LASSO uses fewer features.
  - EN produces lower errors.

Does MEN outperform LASSO?

No theoretical results as yet, but on lung and ovarian cancer data, MEN combines *accuracy* of EN with *sparsity* of LASSO.

Outline

Sensing with ℓ_2-Norm Sparsity Index

With exact measurements, earlier conclusion becomes

$$\|\Delta(Ax) - x\|_2 \leq C_0 \sigma_k(x, \|\cdot\|_1).$$

Theorem: (CDD (2009)) Suppose there exist an integer m, a matrix $A \in R^{m \times n}$ and a function $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ such that, for some constant C_0, we have

$$\|\Delta(Ax) - x\|_2 \leq C_0 \sigma_k(x, \|\cdot\|_2).$$

Then $m \geq C_0^2 n$.

No compression is possible using ℓ_2-norm sparsity index.

Open Problem: Can we replace $\|\cdot\|_2$ on right side by $\|\cdot\|_\mu$?

Outline

References

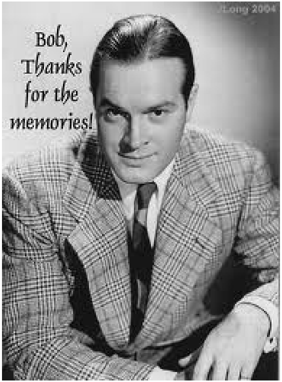
- D. Achlioptas, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *J. Control and Sys. Sci.*, 66, 681-687, 2003.
- E. J. Candès, J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Thy.*, 52(2), 489-509, 2006.
- E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Thy.*, 51(12), 4203-4215, 2005.
- E. J. Candès and Y. Plan, "Near ideal model selection by ℓ_1 minimization," *Annals of Statistics*, 37(5A), 2145-2177, 2009.

References (Cont'd)

- A. Cohen, W. Dahmen and R. Devore, "Compressed sensing and best k-term approximation," *J. Amer. Math. Soc.*, 22(1), 211-231, 2009.
- M. Davenport, *Random observations on random observations: Sparse signal acquisition and processing*, Ph.D. thesis, Rice University, 2009.
- M. A. Davenport, M. F. Duarte, Y. C. Eldar and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok (Eds.), Cambridge, 1-68, 2012.
- D. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution," *Comm. Pure and Appl. Math.*, 59(6), 797-829, 2006.

References (Cont'd)

- Y. C. Eldar and G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications*, Cambridge, 2012.
- S. Negabhan, P. Ravikumar, M. J. Wainwright and B. Yu, "A unified framework for high-imensional analysis of m-estimators with decomposable regularizers," *Statistical Science*, 27(4), 538-557, 2012.
- R. Tibshirani, "Regression shrinkage and selectioon via the lasso," *J. Royal Stat. Soc. B*, 58(1), 267-288, 1996.
- M. Vidyasagar, "Near ideal behavior of a modified elastic net algorithm," *IEEE Conference on Decision and Control*, (to be published), 2013.
- H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Stat. Soc. B*, 67, 301-320, 2005.



Thanks for the Memories!