EE546/STAT593C Lecture 9
Greedy recovery algorithms

Maryam Fazel and Marina Meila
April 23, 2010

These notes supplement the reading of [Tropp and Wright, 2010], [Tropp and Gilbert, 2007], [Needell and Tropp, 2008].

1 Notation

$\Phi x = y$ with $\Phi \in \mathbb{R}^{n \times N}$ is an underdetermined linear system, which admits a sparse solution $\tilde{x}$, with support $(x) = T$ and sparsity $k = |T| = ||x||_0$. $||.||$ represents the Euclidean norm.

We denote by $\Phi_i$ the $i$-th column of $\Phi$, and for simplicity we assume that $||\Phi_i|| = 1$ for all $i$.

2 Incoherence and RIP

Another property that guarantees sparse recovery is the incoherence of the columns of $\Phi$. Denote

$$\mu = \max_{i \neq j} |<\Phi_i, \Phi_j>|$$

$\mu$ is called the incoherence of $\Phi$ and it should be as small as possible; $\mu = 0$ (ideal but not achievable) would mean that the columns of $\Phi$ are orthogonal.

It is easy to show that incoherence and RIP imply each other.

For any set $T$ of columns of $\Phi$ we have

$$\min \lambda(\Phi_T^T \Phi_T) \geq \min_i \left[ ||\Phi_i||^2 - \sum_{i \neq j} |<\Phi_i, \Phi_j>| \right] \geq 1 - (k-1)\mu \quad (2)$$

The first inequality holds by the Gershgorin theorem. The above implies $\delta_k \leq (k-1)\mu$. 

1
The converse is proved taking \( x \) to be a vector with \( x_i = 1, x_j = -1 \) and 0 otherwise. Then, \( ||\Phi x||^2 = ||\Phi_i - \Phi_j||^2 = 2 - 2 < \Phi_i, \Phi_j > = (1 - \delta_2)^2 \). It follows that \( < \Phi_i, \Phi_j > \geq \delta_2 \). The inequality \( < \Phi_i, \Phi_j > \geq -2 \) is proved similarly.

A more general “converse” is proved by e.g [Candes, 2009] and is central to the recovery proofs by \( l_1 \) minimization. Let \( S, T \) be disjoint subsets of cardinalities \( k, k' \) respectively. Then

\[
| < \Phi_T x_T, \Phi_S x_S > | \leq \delta_{k+k'} ||x_T|| ||x_S||
\] (3)

3 The OMP, CoSAMP algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Orthogonal Matching Pursuit (OMP)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initialize</strong></td>
<td>Residual ( r_0 = v ), support ( \Omega_0 = \emptyset ), counter ( t = 1 )</td>
</tr>
<tr>
<td><strong>Repeat</strong></td>
<td></td>
</tr>
<tr>
<td>1. Find the column most correlated with ( r ): ( \lambda_t = \arg \max_j</td>
<td>&lt; r_{t-1}, \Phi_j &gt;</td>
</tr>
<tr>
<td>2. Add it to the index set ( \Omega_t = \Omega_{t-1} \cup { \lambda_t } )</td>
<td></td>
</tr>
<tr>
<td>3. Re-evaluate solution ( x_t = \arg \min_x</td>
<td></td>
</tr>
<tr>
<td>4. ( r_t = y - \Phi_{\Omega} x_t ) (note ( x_t \in \mathbb{R}^t ))</td>
<td></td>
</tr>
<tr>
<td>until ( t = m ) (if we know that ( x ) is ( m )-sparse) OR</td>
<td></td>
</tr>
<tr>
<td>until (</td>
<td></td>
</tr>
</tbody>
</table>

The CoSAMP algorithm is similar to OMP, but does a limited “search” at each step, in the sense that it adds more than one coordinate at a time, then it discards the least useful coordinates. Note also that, in both OMP and CoSAMP, the values of \( x_i, i \in \Omega \) are re-estimated at each step, even if for the \( i \)’s that were added to \( \Omega \) in past steps. Without this back-fitting, which is a characteristic of projection pursuit algorithms in general, the algorithm would be very similar to boosting.

In fact, in other circumstances, typically for non-sparse \( x \) vectors, boosting type approaches, without back-fitting, have been shown to work better.
**Algorithm** Compressive Sampling Matching Pursuit (CoSaMP)

**Initialize** Residual $r_0 = y$, support $\Omega = \emptyset$, counter $t = 1$

Repeat

1. Find the $2s$ columns most correlated with $r$: $\lambda_t = \arg\min_{|T|\leq 2s} \sum_{j \in T} |r_{t-1}, \Phi_j| < r_{t-1}, \Phi_j$

2. Add them to the index set $\Omega = \Omega \cup T$

3. Re-evaluate solution $x_t = \arg\min_x ||\Phi_\Omega x - y||_2$ by least squares

4. Prune: $\Omega = \{k$ largest coefficients of $x_t\}$, $x_t \leftarrow x_\Omega$

5. $r_t = y - \Phi_\Omega x_t$ (note $x_t \in \mathbb{R}^t$)

until stopping criterion

In the above, the $s \leq k$ and in the standard setting $s = k$.

The most expensive step in both OMP and CoSaMP is step ??, that is, finding the column(s) most aligned with the residual. This step takes $O(nN)$ multiplications. The least-squares estimation step ?? is on a $t \times n$ problem, and is typically $O(t^3 + mt^2)$ with the constants depending on the structure in $\Phi$.

**Theoretical guarantees for OMP** With requirements on the matrix $\Phi$ similar to those that guarantee RIP (e.g. Gaussian ensemble), if the unknown $x$ is $k$-sparse, with $k$ known, and no measurement noise, OMP is guaranteed to find $x$ in $k$ iterations with high probability [Tropp and Gilbert, 2007]. However, OMP can fail for some signals, so it offers weaker guarantees than LP recovery. Moreover, by testing if the residual is 0, one can detect when OMP has failed [Tropp and Gilbert, 2007]. The minimum number of measurements is (of the order)

$$n \leq \frac{k}{2 \log N}$$  \hspace{1cm} (4)

comparable with RIP-based recovery results.

For approximately sparse signals, the guarantees are much weaker (see [Tropp and Wright, 2010] for a discussion. The strongest and most interesting results is that, if $3\mu k \leq 1$

$$||r_k|| \leq \sqrt{1 + 6k}||y - y_k^*||$$  \hspace{1cm} (5)
where \( y_k^* \) is the best \( l_2 \) approximation of \( y \) with \( k \) columns from \( \Phi \). Note that this is a convergence result in the observation space, showing that we can approximate \( y \) accurately. It is not known if OMP recovers compressible non-sparse signals or if it works with noisy measurements.

**Theoretical results for CoSaMP**
Assume \( \Phi \) has the \( 2k-RIP \) property, with \( \delta_{2k} << 1 \) and \( k \) satisfying

\[
k \leq \frac{n}{2 \log (1 + N/k)}
\]

and that \( y = \Phi \bar{x} + e \). If \( t \) is sufficiently large, then

\[
||x_t - \bar{x}|| \leq C \frac{1}{\sqrt{k}} ||\bar{x} - \bar{x}_{k/2}||_1 + C ||e||
\]

Hence CoSaMP achieves recovery of sparse and compressible signals in noise.

**Iterative Hard Thresholding (IHT)** [Blumensath and Davies, 2008] also satisfies a bound of the form (7) under the RIP conditions.

**Algorithm** **Iterative Hard Thresholding (IHT)**

- **Initialize** Initial estimate \( x_0 = 0 \), counter \( t = 1 \)
- **Repeat**
  1. Estimate the residual \( r_t = y - \Phi x_{t-1} \)
  2. Re-evaluate solution \( x_t = (x_{t-1} + \Phi^+ r_t)_k \) (this involves a least-square solution)
- until stopping criterion

In the above \( \Phi^+ \) denotes the pseudo-inverse of \( \Phi \).

**4 Solving the least-squares problem**

This is described well by [Needell and Tropp, 2008]. For large problems, iterative methods are preferred to the direct solution by pseudo-inverse. Suppose the system of equations is \( Az = u \) with \( A \) a tall matrix.

The iterative methods solve it by minimizing the function \( f(z) = ||Az - u||^2 \).

The **conjugate gradient** method uses the conjugate gradient descent applied to \( f \). Since the function is quadratic, it will stop after a number of iterations equal to the dimension of \( z \).
Richardson’s method is gradient descent (with step size 1) applied to \( f \). Denote \( M = A^T A - I \). The gradient descent is defined by

\[
z^{k+1} = z^k - \nabla f(z^k) = z^k - A^T (Az - u) = (I - A^T A)z + A^T u = A^T u - Mz
\]

(8)

One can prove that

\[
z^{k+1} - A^+ u = A^T u - Mz - (A^T A)^{-1} A^T u \\
= (A^T A)(A^T A)^{-1} A^T u - Mz - (A^T A)^{-1} A^T u \\
= (A^T A - I)A^+ u - Mz = M(A^+ u - z)
\]

(9)

(10)

Therefore, \( ||z^{k+1} - A^+ u|| \leq ||M||||z^k - A^+ u|| \) and the norm of \( M \) is bounded using the RIP or the incoherence coefficient \( \mu \).

When \( M \) is as such, and \( A^T A \) is very well conditioned (also by RIP), and a good initial approximation is available (the previous CoSaMP iteration), then the Richardson method converges very fast, and it is more advantageous than the direct method. Additional advantages would appear if \( A \) was sparse or structured, since the algorithm involves only multiplications with \( A \) or \( A^T \).

In [Needell and Tropp, 2008] it is proved that a good error bound is obtained after only three iterations of the Richardson method.

5 Proof ideas for recovery with CoSaMP

1. The irrecoverable energy

\[
\nu = ||\bar{x} - \bar{x}_k|| + \frac{1}{\sqrt{s}}||\bar{x} - \bar{x}_k||_1 + ||e||
\]

(12)

The above terms represent: the sparsity residual, an additional residual produced by the Gelfand width, and the noise energy.

2. Theorem 2.1. The iteration invariant

\[
||\bar{x} - x_{t+1}|| \leq \frac{1}{2} ||\bar{x} - x_t|| + 10\nu
\]

(13)

from which it follows that

\[
||\bar{x} - x_t|| \leq \frac{1}{2t} ||\bar{x}|| + 20\nu
\]

(14)
3. First, assume $\bar{x}$ is $k$-sparse. Then, using RIP, each step of the algorithm provides a bound on the error (or on a part of it). All the steps but the LS problem use the usual properties of RIP. For the LS problem, in addition to an infinite precision error bound, we can compute another term corresponding to the error incurred for stopping the iterative solver early. Putting all the bounds together proves the invariant.

4. Second, if $\bar{x}$ is compressible but not sparse, the problem is reduced to the previous case, by analyzing the problem $y = \Phi \bar{x}_k + \tilde{e}$, where $\tilde{e} = e + \Phi(\bar{x} - \bar{x}_k)$. It can be shown that under $RIP$,

$$||\tilde{e}|| \leq c \left( ||\bar{x} - \bar{x}_k|| + \frac{1}{\sqrt{k}}||\bar{x} - \bar{x}_k||_1 \right) + ||e||$$

The sparse invariant result and the triangle inequality prove the invariant in this case too.

5. The result in (??) is obtained by bounding $\nu$ with

$$\nu \leq \frac{c'}{\sqrt{k}}||\bar{x} - \bar{x}_{k/2}||_1 + ||e||$$

which follows from the standard $||z - z_k||_1 \leq \frac{1}{2}||z||_1$ and $||\|\|\sqrt{\text{dim}} \geq ||\||_1$ applied with $k \leftarrow k/2 = \text{dim}$.

6 Conclusions

Greedy pursuit methods are in general faster than optimization (LP) algorithms, but are theoretically less understood, and have weaker theoretical guarantees. Empirically, the LP methods are in general more robust to noise and more accurate when the signal is not sparse. However, this is an area under development.

References


