EE 381V: Large Scale Learning	Spring 2013
Lecture 21 — April 2	
Lecturer: Caramanis & Sanghavi	*Scribe: Megasthenis Asteris*

21.1 Review

In the past few lectures we have been studying the problem of sparse signal reconstruction from a small number of noiseless linear measurements. The objective is to recover an unknown, k-sparse, n-dimensional vector $\mathbf{x}$ from a measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$, which is a linear transformation of $\mathbf{x}$ by a known $m \times n$ matrix $\mathbf{A}$, where $m < n$.

We have presented various heuristics for sparse signal recovery, including greedy algorithms such as Orthogonal Matching Pursuit, and algorithms based on convex optimization such as Basis Pursuit. In order to provide performance guarantees, the analysis of these algorithms seeks conditions on the measurement matrix $\mathbf{A}$ and the sparsity k of $\mathbf{x}$, under which the solution is unique and/or the algorithm can successfully recover the sparsest solution. Examples of such conditions include *mutual coherence* (maximum coherence between columns of the measurement matrix $\mathbf{A}$), the *Restricted Isometry Property*, and the *Restricted Strong Convexity* (which may be the subject of a future lecture).

The previous lecture focused on the analysis of Basis Pursuit under the assumption that the measurement matrix satisfies an RIP condition. Specifically, if $\mathbf{A}$ satisfies the $(\epsilon, 2k)$-RIP for $\epsilon < \sqrt{2} - 1$, then we can recover the unique k-sparse vector $\mathbf{x}$ as the solution of the l_1-minimization problem

$$\begin{aligned} &\text{minimize } \|\mathbf{x}\|_1 \\ &\text{subject to: } \mathbf{y} = \mathbf{A}\mathbf{x}. \end{aligned}$$

21.2 Overview

Although the l_1-norm minimization is arguably the most famous approach to the sparse recovery problem, it is not empirically the best. In this lecture we revisit Orthogonal Matching Pursuit with an analysis based on the Restricted Isometry Property. Specifically, we describe a result by Davenport and Wakin according to which if the measurement matrix satisfies the RIP property of order $k + 1$ with an isometry constant $\delta < 1/3\sqrt{k}$, Orthogonal Matching Pursuit can recover any k-sparse signal. Recall that Orthogonal Matching Pursuit aims to discover the support of the sparse vector $\mathbf{x}$, gradually expanding a set of indices through an iterative procedure. If the correct support has been retrieved, the calculation of the solution $\mathbf{x}$ reduces to a simple least squares problem. The proof of the previous result essentially relies on showing that under the aforementioned condition the algorithm in every iteration selects an index that belongs to the true support of $\mathbf{x}$. We adhere to the argument outline of [1] and large portion of the following description has been adapted from that paper.

21.3 Introduction

Orthogonal Matching Pursuit, depicted in Algorithm 1, is a greedy iterative algorithm for the recovery of a k-sparse vector $\mathbf{x}$ from a vector of linear measurements $\mathbf{y} = \mathbf{\Phi x}$. Note that in the notation of [1], $\mathbf{\Phi}$ is used to denote the measurement matrix. The algorithm iteratively expands an initially empty set Λ of indices, that correspond to the estimated support of $\mathbf{x}$. Once the support Λ has been determined, the final estimate $\hat{\mathbf{x}}$ is obtained as the orthogonal of $\mathbf{y}$ on the subset of columns of Φ indexed by Λ. Let Λ^l denote the estimate of Λ in the l-th

---

Algorithm 1 Orthogonal Matching Pursuit

Require: $\mathbf{\Phi}$ {The measurement matrix},
  $\mathbf{y}$ {The vector of measurements},
  stopping criterion
  Initialization: $\mathbf{r}^0 = \mathbf{y}$ {residual}, $\mathbf{x}^0 = \mathbf{0}$ {estimate}, $\mathbf{\Lambda}^0 = \emptyset$, $l = 0$ {Iteration counter},
  while not converged **do**
    match: $\mathbf{h}^l = \mathbf{\Phi}^T \mathbf{r}^l$
    identify: $\mathbf{\Lambda}^{l+1} = \mathbf{\Lambda}^l \cup \{\arg\max_j |\mathbf{h}^l(j)|\}$
    {If multiple maxima exist, choose only one arbitrarily.}
    update: $\mathbf{x}^{l+1} = \arg\min_{\mathbf{z}:\mathrm{supp}(\mathbf{z})\subseteq \mathbf{\Lambda}^{l+1}} \|\mathbf{y} - \mathbf{\Phi z}\|_2$
    $\mathbf{r}^{l+1} = \mathbf{y} - \mathbf{\Phi x}^{l+1}$
    $l = l + 1$.
  end while
  return $\hat{\mathbf{x}} = \mathbf{x}^l = \arg\min_{\mathbf{z}:\mathrm{supp}(\mathbf{z})\subseteq \mathbf{\Lambda}^{l}} \|\mathbf{y} - \mathbf{\Phi z}\|_2$

---

iteration. This set is expanded into Λ^{l+1} by including a new index that hopefully belongs to the true support of $\mathbf{x}$. Throughout the execution, the algorithm maintains the residual vector $\mathbf{r}^l$, *i.e.*, the component of the measurement $\mathbf{y}$ that cannot be *explained* by the columns of $\mathbf{\Phi}$ indexed by Λ^l. The index to be inserted in Λ^l yielding Λ^{l+1} is the one corresponding to the column of $\mathbf{\Phi}$ that has the largest inner product with $\mathbf{r}^l$, *i.e.*, the column that *best* explains the residual $\mathbf{r}^l$. Note that $\mathbf{r}^l$ is by construction orthogonal to the columns of $\mathbf{\Phi}$ indexed by Λ^l. Hence, $h(j) = 0$ for $j \in \Lambda^l$ and the new index $j^* = \arg\max_j |\mathbf{h}^l(j)|$ is guaranteed to belong to $\left(\Lambda^l\right)^c$, the complement of Λ^l, increasing $|\Lambda|$ by one in every iteration.

Let $\mathbf{x}_\Lambda$ denote the vector containing the entries of $\mathbf{x}$ indexed by Λ, and similarly, $\mathbf{\Phi}_\Lambda$ the $m \times |\Lambda|$ matrix formed by the $|\Lambda|$ columns of $\mathbf{\Phi}$ indexed by Λ. The estimate of the solution at the l-th iteration, $\mathbf{x}^l$, is obtained as the orthogonal projection of the measurement $\mathbf{y}$ on the columns of $\mathbf{\Phi}_{\Lambda^l}$. In other words,

$$\mathbf{x}^l_{\Lambda^l} = \mathbf{\Phi}^\dagger_{\Lambda^l}\mathbf{y}, \quad \text{and} \quad \mathbf{x}^l_{(\Lambda^l)^c} = 0,$$

where $\mathbf{\Phi}_{\Lambda^l}^{\dagger} = \left(\mathbf{\Phi}_{\Lambda}^T \mathbf{\Phi}_{\Lambda}\right)^{-1} \mathbf{\Phi}_{\Lambda}^T$ is the Moore-Penrose pseudoinverse of $\mathbf{\Phi}_{\Lambda}$. The residual in the l-th iteration is

$$\mathbf{r}^l = \mathbf{y} - \mathbf{\Phi}\mathbf{x}^l = \mathbf{y} - \mathbf{\Phi}_{\Lambda^l}\mathbf{\Phi}_{\Lambda^l}^{\dagger}\mathbf{y}.$$

Observe that $\mathbf{\Phi}_{\Lambda^l}\mathbf{\Phi}_{\Lambda^l}^{\dagger}$ is a projection matrix; it projects $\mathbf{y}$ on $\mathbf{R}(\mathbf{\Phi}_{\Lambda^l})$, the column space of $\mathbf{\Phi}_{\Lambda^l}$. Denoting this orthogonal projection operator onto $\mathbf{R}(\mathbf{\Phi}_{\Lambda^l})$ by $\mathbf{P}_{\Lambda^l}$, and similarly the projection operator on the complement space by $\mathbf{P}_{\Lambda^l}^{\perp} = (\mathbf{I} - \mathbf{P}_{\Lambda^l})$, the residual vector can be expressed as follows

$$\mathbf{r}^l = \mathbf{y} - \mathbf{\Phi}_{\Lambda^l}\mathbf{\Phi}_{\Lambda^l}^{\dagger}\mathbf{y} = \left(\mathbf{I} - \mathbf{P}_{\Lambda^l}\right)\mathbf{y} = \mathbf{P}_{\Lambda^l}^{\perp}\mathbf{y}. \tag{21.1}$$

The residual is correlated against all columns of $\mathbf{\Phi}$ to determine the index to be included in the estimate of $\mathbf{x}$'s support, Λ^{l+1}. Note that every column of $\mathbf{\Phi}$ can be written as the sum of two orthogonal components: the projection on $\mathbf{R}(\mathbf{\Phi}_{\Lambda^l})$, and the projection on its orthogonal complement $\mathbf{R}(\mathbf{\Phi}_{\Lambda^l})^{\perp}$. The vector of inner products is, hence,

$$\begin{aligned}
\mathbf{h}^l &= \mathbf{\Phi}^T\mathbf{r}^l && (21.2)\\
&= \left(\mathbf{P}_{\Lambda^l}\mathbf{\Phi} + \mathbf{P}_{\Lambda^l}^{\perp}\mathbf{\Phi}\right)^T \mathbf{P}_{\Lambda^l}^{\perp}\mathbf{y} \\
&= \mathbf{0} + \mathbf{\Phi}^T \left(\mathbf{P}_{\Lambda^l}^{\perp}\right)^T \mathbf{P}_{\Lambda^l}^{\perp}\mathbf{y} && (21.3)\\
&= \mathbf{\Phi}^T \left(\mathbf{P}_{\Lambda^l}^{\perp}\right)^T \left(\mathbf{P}_{\Lambda^l}^{\perp}\right)^T \mathbf{y} \\
&= \mathbf{\Phi}^T \left(\mathbf{P}_{\Lambda^l}^{\perp}\right)^T \mathbf{y} \\
&= \left(\mathbf{P}_{\Lambda^l}^{\perp}\mathbf{\Phi}\right)^T \mathbf{y} = \mathbf{A}_{\Lambda^l}^T\mathbf{y}, && (21.4)
\end{aligned}$$

where we have defined[1] $\mathbf{A}_{\Lambda^l} = \mathbf{P}_{\Lambda^l}^{\perp}\mathbf{\Phi}$. In the intermediate steps we have used the fact that for an orthogonal projection matrix $\mathbf{P} = \mathbf{P}^T = \mathbf{P}^2$. Equation (21.2) implies that for the *matching* step of the algorithm, the residual need not be explicitly calculated. Instead, at the l-th iteration, the columns of $\mathbf{\Phi}$ can be orthogonalized against $\mathbf{R}(\mathbf{\Phi}_{\Lambda^l})$ and then the inner products of the matching step can be directly calculated correlating $\mathbf{y}$ with the resulting columns. Although the two approaches are equivalent, the new perspective will prove useful in the subsequent analysis.

One final observation is that matrix $\mathbf{A}_{\Lambda^l}$ plays a significant role in the construction of the residual $\mathbf{r}^l$:

$$\mathbf{r}^l = \mathbf{P}_{\Lambda^l}^{\perp}\mathbf{y} = \mathbf{P}_{\Lambda^l}^{\perp}\mathbf{\Phi}\mathbf{x} = \mathbf{A}_{\Lambda^l}\mathbf{x}. \tag{21.5}$$

Recalling that the columns of $\mathbf{A}_{\Lambda^l}$ in Λ^l are zero, we can write

$$\mathbf{r}^l = \mathbf{A}_{\Lambda^l}\tilde{\mathbf{x}}^l, \tag{21.6}$$

[1] Note that this notation is a little confusing: contrary to $\mathbf{\Phi}_{\Lambda^l}$, $\mathbf{A}_{\Lambda^l}$ is a matrix with the same dimensions as Φ and not only a subset of the columns.

where

$$\tilde{\mathbf{x}}^l_{\Lambda^l} = 0 \quad \text{and} \quad \tilde{\mathbf{x}}^l_{(\Lambda^l)^c} = \mathbf{x}_{(\Lambda^l)^c}. \tag{21.7}$$

The benefit is that the support of $\tilde{\mathbf{x}}^l$ shrinks as Λ^l grows, an observation that will also prove useful in the sequel.

21.4 Main Result

Theorem 21.1. *Suppose that $\mathbf{\Phi}$ satisfies the RIP of order $k+1$ with isometry constant $\delta < \frac{1}{3\sqrt{k}}$. Then for any $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_0 \leq k$, Orthogonal Matching Pursuit will recover $\mathbf{x}$ exactly from $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$ in k iterations.*

In other words, if $\mathbf{\Phi}$ satisfies the conditions of the theorem, then Orthogonal Matching Pursuit will recover the correct set of indices, *i.e.*, $\Lambda = \text{supp}(\mathbf{x})$, which in turn implies that in every iteration the maximum (by absolute value) element of $\mathbf{h}$ corresponds to an index in the true support of $\mathbf{x}$. The analysis will therefore evolve around vector $\mathbf{h}$.

It is interesting to note that equation (21.6) essentially describes a sparse signal recovery problem. In the first iteration, $l = 0$, it corresponds to the original problem $\mathbf{y} = \mathbf{\Phi}\tilde{\mathbf{x}}^0$. In the l-th iteration the objective is to recover a sparse vector $\tilde{\mathbf{x}}^l$, from an observation $\mathbf{r}^l$ with a measurement matrix $\mathbf{A}_{\Lambda^l}$. In successive iterations, increasingly more zero columns appear in $\mathbf{A}_{\Lambda^l}$, but (assuming that the algorithm expands Λ^l without mistakes) the support of $\tilde{\mathbf{x}}^l$ shrinks accordingly, with nonzero entries only outside Λ^l. Assume that $\mathbf{\Phi} = \mathbf{A}_{\Lambda^0}$ has an RIP property that guarantees that the algorithm will select a correct index in the first iteration. Then, it would suffice the *measurement* matrices $\mathbf{A}_{\Lambda^l}$ of subsequent iterations to *inherit* a similar property for $\left(k - |\Lambda^l|\right)$-sparse vector supported on $(\Lambda^l)^c$, revealing the inductive nature of the proof to follow. Before we get there, however, we will prove a series of lemmata that will eventually be combined in the proof of Theorem 21.1.

Lemma 21.2. *If $\mathbf{\Phi}$ satisfied the RIP of order k with constant δ, then for any set Λ with $|\Lambda| < k$*

$$\left(1 - \frac{\delta}{1-\delta}\right)\|\mathbf{u}\|_2^2 \leq \|\mathbf{A}_\Lambda \mathbf{u}\|_2^2 \leq (1+\delta)\|\mathbf{u}\|_2^2, \tag{21.8}$$

for all $\mathbf{u}$ such that supp$(\mathbf{u}) \cap \Lambda = \emptyset$ and $\|\mathbf{u}\|_0 \leq k - |\Lambda|$.

The previous Lemma (see [3], [4] for proof) relates the RIP of the original $\mathbf{\Phi}$ matrix to the RIP of $\mathbf{A}_{\Lambda^l}$. If $\mathbf{\Phi}$ satisfies the RIP with a constant δ on all k-sparse vectors, then $\mathbf{A}_\Lambda$ satisfies a modified version of the RIP on every $(k - |\Lambda|)$-sparse vector supported on Λ^c.

Lemma 21.3. $\Lambda \subseteq [n]$ *and suppose $\tilde{\mathbf{x}} \in \mathbb{R}^n$ with supp$(\tilde{\mathbf{x}}) \cap \Lambda = \emptyset$. Let*

$$\mathbf{h} = \mathbf{A}_\Lambda^T \mathbf{A}_\Lambda \tilde{\mathbf{x}}. \tag{21.9}$$

If $\mathbf{\Phi}$ satisfies the RIP of order $\|\tilde{\mathbf{x}}\|_0 + |\Lambda| + 1$ with isometry constant δ, we have

$$|\mathbf{h}(j) - \tilde{\mathbf{x}}(j)| \leq \frac{\delta}{1-\delta}\|\tilde{\mathbf{x}}\|_2, \tag{21.10}$$

for all $j \notin \Lambda$.

Proof: From Lemma 21.2 we have that the restriction of $\mathbf{A}_\Lambda$ to the columns indexed by Λ^c satisfies the RIP of order $(\|\tilde{\mathbf{x}}\|_0 + |\Lambda| + 1) - |\Lambda| = \|\tilde{\mathbf{x}}\|_0 + 1$ with isometry constant $\delta/(1-\delta)$. From equations (21.1), (21.3) and (21.5), we have $\mathbf{h} = \mathbf{A}_\Lambda^T \mathbf{A}_\Lambda \tilde{\mathbf{x}}$, which can be alternatively written as

$$h(j) = \langle \mathbf{A}_\Lambda \tilde{\mathbf{x}}, \mathbf{A}_\Lambda \mathbf{e}_j \rangle, \tag{21.11}$$

where $\mathbf{e}_j$ denotes the j-th vector of the standard basis. Now suppose $j \notin \Lambda$. Then, since $\|\tilde{\mathbf{x}} \pm \mathbf{e}_j\|_0 \leq \|\tilde{\mathbf{x}}\|_0 + 1$ and $\text{supp}(\tilde{\mathbf{x}} \pm \mathbf{e}_j) \cap \Lambda = \emptyset$, we conclude from that

$$|h(j) - \tilde{x}(j)| = |\langle \mathbf{A}_\Lambda \tilde{\mathbf{x}}, \mathbf{A}_\Lambda \mathbf{e}_j \rangle - \langle \tilde{\mathbf{x}}, \mathbf{e}_j \rangle| \leq \frac{\delta}{1-\delta}\|\tilde{\mathbf{x}}\|_2 \|\mathbf{e}_j\|_2. \tag{21.12}$$

The inequality is a result of the fact that $\mathbf{A}_\Lambda$ satisfies the RIP with constant $\delta/(1-\delta)$. Noting that $\|\mathbf{e}_j\|_2 = 1$, we reach the desired conclusion. □

Lemma 21.4. *Suppose that Λ, $\mathbf{\Phi}$, $\tilde{\mathbf{x}}$ meet the assumptions specified in Lemma 21.3, and let $\mathbf{h}$ be as defined in* (21.9). *If*

$$\|\tilde{\mathbf{x}}\|_\infty > \frac{2\delta}{1-\delta}\|\tilde{\mathbf{x}}\|_2, \tag{21.13}$$

we are guaranteed that $\arg\max_j |h(j)| \in supp(\tilde{\mathbf{x}})$.

Proof: For $j \notin \text{supp}(\tilde{\mathbf{x}})$, we clearly have $\tilde{x}(j) = 0$. Therefore, by (21.10) we have

$$|h(j)| \leq \frac{\delta}{1-\delta}\|\tilde{\mathbf{x}}\|_2, \quad \text{for } j \notin \text{supp}(\tilde{\mathbf{x}}). \tag{21.14}$$

On the other hand, if (21.13) is satisfied, then there must exists an element $\tilde{x}(j)$, clearly in the support of $\tilde{\mathbf{x}}$, such that

$$|\tilde{x}(j)| > \frac{2\delta}{1-\delta}\|\tilde{\mathbf{x}}\|_2. \tag{21.15}$$

For that j, from (21.10) and the triangle inequality, we obtain

$$|h(j)| > \frac{\delta}{1-\delta}\|\tilde{\mathbf{x}}\|_2. \tag{21.16}$$

The bottom line is that there must exist a j in the support of $\tilde{\mathbf{x}}$ for which $|h(j)|$ is greater than all $|h(j')|$ for $j' \notin \text{supp}(\tilde{\mathbf{x}})$. Thus, the index j for which $|h(j)|$ is maximized is guaranteed to belong to the support of $\tilde{\mathbf{x}}$. □

Finally, without a proof, note the following.

Lemma 21.5. *For any $\mathbf{u} \in \mathbb{R}^n$, $\|\mathbf{u}\|_\infty \geq \frac{\|\mathbf{u}\|_2}{\sqrt{\|\mathbf{u}\|_0}}$.*

21.4.1 Proof of main theorem

Combining the above results, we can finally establish the proof of Theorem 21.1 In the first, iteration ($l = 0$) we have:

$$\mathbf{h}^0 = \mathbf{\Phi}^T \mathbf{r}^0 = \mathbf{\Phi}^T \mathbf{y} = \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{x}.$$

By assumption $\|\mathbf{x}\|_0 \leq k$ and therefore, by Lemma (21.5), $\|\mathbf{x}\|_\infty \geq \|\tilde{\mathbf{x}}\|_2/\sqrt{k}$. The conditions of Lemma (21.13) are satisfied. In order to satisfy the remaining conditions of Lemma 21.4, namely equation (21.13), it suffices to ask

$$\frac{2\delta}{1-\delta} \geq \frac{1}{\sqrt{k}} \Leftrightarrow \delta < \frac{1}{2\sqrt{k}+1}.$$

Clearly, if $\delta \leq 1/3\sqrt{k}$, as in the assumption, the above is satisfied, and

$$\arg\max_j |h^0(j)| \in \mathrm{supp}(\mathbf{x}),$$

i.e., the algorithm is guaranteed to choose a valid index in the first iteration. For the induction step, assume that $\Lambda^l \subseteq \mathrm{supp}(\mathbf{x})$, *i.e.*, all iterations up to the $(l-1)$-th have succeeded meaning that all indices in Λ^l belong to the support of $\mathbf{x}$, . Recall that by construction of $\tilde{\mathbf{x}}^l$, we have $\mathrm{supp}(\tilde{\mathbf{x}}^l) \cap \Lambda^l = \emptyset$ and hence,

$$\|\tilde{\mathbf{x}}^l\|_0 \leq k - |\Lambda^l|.$$

By assumption, $\mathbf{\Phi}$ satisfies the RIP of order $k+1$. Taking the previous inequality under consideration, we have

$$k + 1 = (k - |\Lambda^l|) + |\Lambda^l| + 1 \geq \|\tilde{\mathbf{x}}^l\|_0 + |\Lambda^l| + 1.$$

Finally, we have

$$\|\tilde{\mathbf{x}}\|_\infty \overset{\text{Lemma 21.5}}{\geq} \frac{\|\tilde{\mathbf{x}}\|_2}{\sqrt{k - |\Lambda^l|}} \geq \frac{\|\tilde{\mathbf{x}}\|_2}{\sqrt{k}} \overset{\text{for } \delta < 1/\sqrt{k}}{\geq} \frac{2\delta}{1-\delta}\|\tilde{\mathbf{x}}\|_2. \tag{21.17}$$

From Lemma 21.4, we conclude that

$$\arg\max_j |h^l(j)| \in \mathrm{supp}(\tilde{\mathbf{x}}^l), \tag{21.18}$$

and hence, $\Lambda^{l+1} \subseteq \mathrm{supp}(\mathbf{x})$, which completes the proof.

Bibliography

[1] Davenport, Mark A., and Michael B. Wakin, *Analysis of orthogonal matching pursuit using the restricted isometry property*. Information Theory, IEEE Transactions on 56.9 (2010): 4395-4401.

[2] Shai Shalev-Shwartz, *Compressed Sensing: Basic results and self contained proofs*. (2009).

[3] Dai, Wei, and Olgica Milenkovic. *Subspace pursuit for compressive sensing signal reconstruction*. Information Theory, IEEE Transactions on 55.5 (2009): 2230-2249.

[4] M. A. Davenport, P. T. Boufounos, and R. G. Baraniuk, *Compressive domain interference cancellation*. Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS), Saint-Malo, France, Apr., 2009.