Sparse Processing
Tutorial for Swe-CTW 2014

Saikat Chatterjee

Communication Theory Lab, KTH

June 3, 2014

Motivation of the course

1. The necessity of linear set of equations
2. Elementary problem: Sparse solution of an under-determined linear set of equations

Figure: An example image with blurring

Under-determined setup (Our interest in the tutorial)

$$\mathbf{b} = \mathbf{A}\mathbf{x}, \tag{1}$$

where

$$\mathbf{b} \in \mathbb{R}^{n\times 1},\ \mathbf{x} \in \mathbb{R}^{m\times 1},\ \mathbf{A} \in \mathbb{R}^{n\times m}$$

and most importantly

$$n < m.$$

♠ Reference book: "Sparse and Redundant Representations", By Michael Elad, Springer.

Regularization

$$(P_J): \quad \arg\min_{\mathbf{x}\in\mathbb{R}^m} \; J(\mathbf{x}) \text{ subject to } \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{2}$$

If, we are interested for the minimum norm solution, then $J(\mathbf{x}) = \|\mathbf{x}\|_2^2$. So, we solve

$$\arg\min_{\mathbf{x}\in\mathbb{R}^m} \; \|\mathbf{x}\|_2^2 \text{ subject to } \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{3}$$

Solution of the above is right Pseudo-inverse:
$\mathbf{x}^\star = \mathbf{A}^\dagger\mathbf{b} = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{b}$.
Proof: ? Homework

Convexity

Definition

Convex set: A set Ω is convex if $\forall\, \mathbf{x}_1,\, \mathbf{x}_2 \in \Omega$ and $\forall t \in [0,1]$, the convex combination $\mathbf{x} = t\mathbf{x}_1 + (1-t)\mathbf{x}_2 \in \Omega$.

Definition

Convex function: A function $J(\mathbf{x}) : \Omega \to \mathbb{R}$ is convex if $\forall\, \mathbf{x}_1,\, \mathbf{x}_2 \in \Omega$ and $\forall t \in [0,1]$, the convex combination $\mathbf{x} = t\mathbf{x}_1 + (1-t)\mathbf{x}_2 \in \Omega$ satisfies

$$J(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tJ(\mathbf{x}_1) + (1-t)J(\mathbf{x}_2). \tag{4}$$

Remark

Convex function: If $\nabla^2 J(\mathbf{x})$ is positive definite. In fact, for $J(\mathbf{x}) = \|\mathbf{x}\|_2^2$, $\nabla^2 \|\mathbf{x}\|_2^2 = 2\mathbf{I} \geq \mathbf{0}$ is strictly positive definite.

Convex Vs. Non-convex

Remark

Carefully chosen $J(\mathbf{x})$ allows battery of convex optimization algorithms in use, leading to globally optimum solutions. But that does not make an impression that we should never seek non-convex solutions. We should never think that non-convex solutions/approaches do not have any value. The engineering choice resides in the problem.

Special interest

The l_p norm

$$J(\mathbf{x}) = \|\mathbf{x}\|_p^p = \sum_i |x_i|^p, \tag{5}$$

where $p \geq 1$. In system identification, $p = \infty$ is used. We are mainly interested in $p = 1$.

l_1-norm minimization

$$(P_1): \quad \underset{\mathbf{x}\in\mathbb{R}^m}{\arg\min} \ \|\mathbf{x}\|_1 \ \text{subject to } \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{6}$$

Some remarks:

1. l_1 norm is not strictly convex. Example.
2. So, there may be multiple global solutions

Important claims:

1. The solutions are gathered in a set that is bounded and convex
2. Among the solutions, there exists at-least one with at-most n non-zeros (i.e. the number of constraints)

Proof: ? Please see the reference

Remark

The l_1-norm minimization promotes sparsity

Conversion of (P_1) to linear program

The standard form LP:

$$\text{minimize } \mathbf{c}^T\mathbf{z} \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{b} \text{ and } \mathbf{z} \geq \mathbf{0}. \tag{7}$$

Support set notation: $I_{\mathbf{x}} = \{i : x_i \neq 0\}$
Proof sketch: Please see reference for details

1. Let $\mathbf{u}, \mathbf{v} \geq \mathbf{0}$ such that $\mathbf{x} = \mathbf{u} - \mathbf{v}$.
2. Note $\mathbf{u}^T\mathbf{v} = \mathbf{0}$ or $I_{\mathbf{u}} \cap I_{\mathbf{v}} = \emptyset$
3. Define $\mathbf{z} = \begin{bmatrix}\mathbf{u}^T & \mathbf{v}^T\end{bmatrix}^T$ and hence $\mathbf{z} \geq \mathbf{0}$
4. Show that $\|\mathbf{x}\|_1 = \mathbf{1}^T\mathbf{z}$ and $\mathbf{b} = \mathbf{A}\mathbf{x} = [\mathbf{A} \ -\mathbf{A}]\,\mathbf{z}$
5. Then show that LP is equivalent to (P_1) by proving that solution of LP never violates $\mathbf{u}^T\mathbf{v} = 0$

Promoting sparse solutions

1. Moving from l_2 to l_1 regularization leads to promoting sparsity
2. By this rationale, we can consider l_p norms where $p < 1$
3. But then the norms are not strictly norms (they do not satisfy triangular inequality) and things are not convex

Question

Does the norm with $p < 1$ leads to sparse solution? Yes it is.

A case study: Let $\mathbf{x}$ satisfies $\|\mathbf{x}\|_0 = a$, $\|\mathbf{x}\|_q^q = 1$ and $\mathbf{x} \geq 0$. Let $p < q$ and the problem is:

$$\min_{\mathbf{x} \in \mathbb{R}^m} \ \|\mathbf{x}\|_p^p \ \text{subject to} \ \|\mathbf{x}\|_q^q = 1 \ \text{and} \ \|\mathbf{x}\|_0 = a. \tag{8}$$

Proof: Using Lagrangian, we will get $\min \|\mathbf{x}\|_p^p = a^{(1-\frac{p}{q})}$

Question

What we get from the proof?
Since $p < q$, *this means that the shortest* ℓ_p*-norm is obtained for* $a = 1$, *having only one non-zero element in* $\mathbf{x}$.

Some pictorial understanding

Let us assume $q = 2$ and $p = 1$.
Illustration: We work out by diagrams

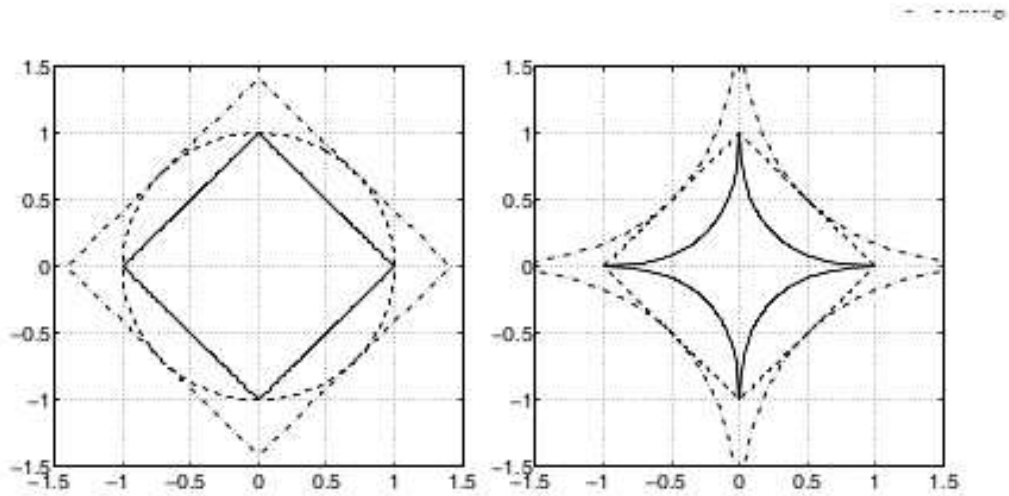

Figure: The Figure is copied from Reference book by Micheal Elad. It may be copyright violation.

The l_0-norm and implications

The l_0 norm

$$\|\mathbf{x}\|_0 = \lim_{p \to 0} \|\mathbf{x}\|_p^p = \lim_{p \to 0} \sum_{i=1}^{m} |x_i|^p = |I_{\mathbf{x}}|. \tag{9}$$

Definition: $I_{\mathbf{x}}$ is called support-set, defined as $I_{\mathbf{x}} = \{i : x_i \neq 0\}$

Remark

We check the norm properties as follows:

1. $\|\mathbf{u} + \mathbf{v}\|_0 \leq \|\mathbf{u}\|_0 + \|\mathbf{v}\|_0$. *So triangular inequality satisfies.*
2. $\|t\mathbf{u}\|_0 = \|\mathbf{u}\|_0 \neq t\|\mathbf{u}\|_0$. *Homogeneity does not hold*
3. *The problem of scaling will haunt us*

The (P_0) problem

$$(P_0): \quad \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{10}$$

Remark

While we have a clear understanding of (P_2) and a better understanding of (P_1), the (P_0) is difficult. Mainly because of its discrete nature.

1. *Can uniqueness of a solution be claimed? Under what conditions?*
2. *If a candidate solution is available, can we perform a simple test to verify that the solution is actually the global minimizer of (P_0)?*

Comments: On exhaustive search complexity

Question around the (P_0) problem

Remark

The complexity of exhaustive serach is exponential in m, and indeed, it has been proven that (P_0) is NP-hard in general. Thus, a mandatory and crucial set of questions arise:

1. *Can (P_0) be efficiently solved by some other means?*
2. *Can approximate solutions be accepted?*
3. *How accurate can those solutions be?*
4. *What kind of approximations will work?*

Uniqueness via Spark

♠ To characterize the null space of a general $\mathbf{A}$ using l_0 norm
♠ Donoho and Elad coined and defined 'spark' in 2003

Definition

Spark (Rank): The spark (rank) of a matrix $\mathbf{A}$ *is the smallest (largest) number of linearly dependent (independent) columns*

$$spark(\mathbf{A}) = \min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{0},\ \mathbf{x} \neq \mathbf{0}. \tag{11}$$

Note: By definition, the non-zero vectors in the null space of $\mathbf{A}$, i.e., $\{\mathbf{x} :\ \mathbf{A}\mathbf{x} = \mathbf{0},\ \mathbf{x} \neq \mathbf{0}\}$, must satisfy $\|\mathbf{x}\|_0 \geq spark(\mathbf{A})$

Uniqueness via Spark

Theorem

Uniqueness via Spark: *For* $\mathbf{Ax} = \mathbf{b}$, *if a candidate solution* $\mathbf{x}$ *satisfies* $\|\mathbf{x}\|_0 < \frac{1}{2} spark(\mathbf{A})$, *then it is necessarily the sparsest.*

Proof: Consider an alternative solution $\mathbf{y}$ satisfying $\mathbf{Ax} = \mathbf{b}$. So $\mathbf{Ax} - \mathbf{Ay} = \mathbf{A}(\mathbf{x} - \mathbf{y}) = \mathbf{0}$; $(\mathbf{x} - \mathbf{y})$ is in null space and $(\mathbf{x} - \mathbf{y}) \neq \mathbf{0}$. $\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \|\mathbf{x} - \mathbf{y}\|_0 \geq spark(\mathbf{A})$. So, if $\|\mathbf{x}\|_0 < \frac{1}{2} spark(\mathbf{A})$ then $\|\mathbf{y}\|_0 > \frac{1}{2} spark(\mathbf{A})$. Implies $\mathbf{x}$ is the sparsest.

Remark

Range of Spark: $2 \leq spark(\mathbf{A}) \leq n + 1$

Proof: We discuss

♠ Finding 'spark' is again NP hard. So comes mutual coherence.

Definition

Mutual coherence: Denoting the i-th column by $\mathbf{a}_i$

$$\mu(\mathbf{A}) = \max_{1 \leq i,j \leq m, i \neq j} \frac{|\mathbf{a}_i^T \mathbf{a}_j|}{\|\mathbf{a}_i\|_2 \ \|\mathbf{a}_j\|_2} \tag{12}$$

1. For an orthogonal matrix, $\mu(\mathbf{A}) = 0$
2. For two-ortho case: $\frac{1}{\sqrt{n}} \leq \mu(\mathbf{A}) \leq 1$
3. For a general $\mathbf{A} \in \mathbb{R}^{n \times m}$, we desire for a low $\mu(\mathbf{A})$ such that it exhibits a close behavior of an orthogonal matrix
4. Donoho and Huo's Work: For randomly constructed $\mathbf{A} \in \mathbb{R}^{n \times m}$ with full rank n, we have $\mu(\mathbf{A}) \geq \sqrt{\frac{m-n}{n(m-1)}}$.
   - When $m = n$, $\mu(\mathbf{A}) = 0$

Lemma

For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, *spark*$(\mathbf{A}) \geq 1 + \frac{1}{\mu(\mathbf{A})}$

Proof: We work out (by using Gersh-gorin disk theorem). We learn the Gersh-gorin disk theorem and its relation with positive-definite property. Then we proceed with formal proof.

Theorem

Uniqueness via Mutual Coherence: For $\mathbf{Ax} = \mathbf{b}$, *if a candidate solution* $\mathbf{x}$ *satisfies* $\|\mathbf{x}\|_0 < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{A})})$, *then it is necessarily the sparsest.*

Proof: By simple arguments

Constructing Grassmannian matrices

Definition

A Grassmannian (real) matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ *with* $m \geq n$ *and normalized columns satisfies that its Gram matrix* $\mathbf{G} = \mathbf{A}^T \mathbf{A}$ *has the following property*

$$\forall i \neq j, \ \ |G_{i,j}| = \sqrt{\frac{m-n}{n(m-1)}}. \tag{13}$$

For a Grassmannian matrix, $\mu(\mathbf{A}) = \sqrt{\frac{m-n}{n(m-1)}}$.

1. Grassmannian is special as the angle between each and every pairs of columns is same and smallest possible.
2. It has strong connection with packing vectors/subspaces in $\mathbb{R}^n$.
3. Important in channel coding and wireless communication.
4. Hard to construct such matrix. A numerical method was proposed by Joel Tropp et al. The paper: "Designing structured tight frames via an alternating projection method," IEEE Trans Information Theory, January 2005

Home Work: Read the paper and simulate the algorithm. Reproduce the experimental and/or numerical results in the paper.

Algorithms : Attempt to Solve (P_0)

$$(P_0): \quad \underset{\mathbf{x}\in\mathbb{R}^m}{\arg\min} \ \|\mathbf{x}\|_0 \ \text{subject to } \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{14}$$

1. Finding support: Estimation of support is a discrete problem. As the support is discrete in nature, algorithms that seek it are discrete as well. This line of reasoning leads to greedy algorithms.
2. Smoothing penalty: Most talked about algo is l_1 norm minimization.

Greedy Algorithm

Let $\|\mathbf{x}\|_0 = k_0 < n$. Then what is our system?

$$\mathbf{b} = \mathbf{A}\mathbf{x} = \sum_{t \in I_{\mathbf{x}}} \mathbf{a}_t x_t = A_{I_{\mathbf{x}}} \mathbf{x}_{I_{\mathbf{x}}} \tag{15}$$

where $I_{\mathbf{x}}$ is the support. Then we can do 'm choose k_0' subspace searches (all possibilities) and can find the solution.

1. Greedy Strategy: A greedy strategy abandons exhaustive search in favor of a series of locally optimal single-term updates.
2. Generally two types: serial identification of atoms or parallel identification.

Greedy methods:

1. Serial: Mathcing pursuit (MP), Orthogonal MP (OMP), Weak MP, LS-OMP
2. Parallel: CoSaMP, Subspace pursuit (SP)

Normalization: Is there any difference between using $\mathbf{A}$ and its normalized version $\tilde{\mathbf{A}}$ where each column l_2 norm is one?

We can express $\tilde{\mathbf{A}} = \mathbf{AW}$ where $\mathbf{W} = diag\{\frac{1}{\|\mathbf{a}_i\|_2}\}$

Theorem

The greedy algorithms (OMP, MP and Weak MP) produce the same solution using either original matrix $\mathbf{A}$ *or its normalized version* $\tilde{\mathbf{A}}$*.*

Proof: Please see reference

Question: Is this theorem holds true for other greedy algorithms? Such as CoSaMP and SP? Possibly yes (still check it out).

Orthogonal matching pursuit

Input: $\mathbf{A}$, $\mathbf{b}$, k_0;
Initialization: Iteration counter $k \leftarrow 0$; $\mathbf{r}_0 \leftarrow \mathbf{b}$, $\mathcal{I}_0 \leftarrow \emptyset$;
repeat
$k \leftarrow k + 1$;
$i_k \leftarrow$ index of the highest amplitude of $\mathbf{A}^t \mathbf{r}_{k-1}$;
$\mathcal{I}_k \leftarrow \mathcal{I}_{k-1} \cup i_k$; (Note: $|\mathcal{I}_k| = k$)
$\mathbf{r}_k \leftarrow \mathbf{b} - \mathbf{A}_{\mathcal{I}_k} \mathbf{A}_{\mathcal{I}_k}^{\dagger} \mathbf{b}$; (Orthogonal projection)
until $((\|\mathbf{r}_k\|_2 > \|\mathbf{r}_{k-1}\|_2) \text{ or } (k > k_0))$
$k \leftarrow k - 1$; (Previous iteration)
Output: $\hat{\mathbf{x}} \in \mathbb{R}^N$, satisfying $\hat{\mathbf{x}}_{\mathcal{I}_k} = \mathbf{A}_{\mathcal{I}_k}^{\dagger} \mathbf{y}$ and $\hat{\mathbf{x}}_{\overline{\mathcal{I}}_k} = \mathbf{0}$.

Spbspace Pursuit

$\mathbf{A}$, $\mathbf{b}$, k_0;
Initialization: Iteration counter $k \leftarrow 0$;
$\mathcal{I}_0 \leftarrow$ indices of the k_0 highest amplitudes of $\mathbf{A}^t\mathbf{b}$;
$\mathbf{r}_0 \leftarrow \mathbf{b} - \mathbf{A}_{\mathcal{I}_0}\mathbf{A}^{\dagger}_{\mathcal{I}_0}\mathbf{b}$;
repeat
$k \leftarrow k + 1$;
$\mathcal{I}_{(p)} \leftarrow \{\text{indices of } k_0 \text{ highest amplitudes of } \mathbf{A}^t\mathbf{r}_{k-1}\}$;
$\mathcal{I}_{(u)} \leftarrow \mathcal{I}_{k-1} \cup \mathcal{I}_{(p)}$; $(k_0 \leq |\mathcal{I}_{(u)}| \leq 2K)$
$\hat{\mathbf{x}}_{\mathcal{I}_{(u)}} \leftarrow \mathbf{A}^{\dagger}_{\mathcal{I}_{(u)}}\mathbf{b}$; $\hat{\mathbf{x}}_{\overline{\mathcal{I}}_{(u)}} \leftarrow \mathbf{0}$; (Orthogonal projection)
$\mathcal{I}_k \leftarrow \{\text{indices of the } k_0 \text{ highest amplitudes of } \hat{\mathbf{x}}\}$;
$\mathbf{r}_k \leftarrow \mathbf{y} - \mathbf{A}_{\mathcal{I}_k}\mathbf{A}^{\dagger}_{\mathcal{I}_k}\mathbf{b}$; (Orthogonal projection)
until $(\|\mathbf{r}_k\|_2 > \|\mathbf{r}_{k-1}\|_2)$
$k \leftarrow k - 1$; (Previous iteration)
Output: $\hat{\mathbf{x}} \in \mathbb{R}^N$, satisfying $\hat{\mathbf{x}}_{\mathcal{I}_k} = \mathbf{A}^{\dagger}_{\mathcal{I}_k}\mathbf{b}$ and $\hat{\mathbf{x}}_{\overline{\mathcal{I}}_k} = \mathbf{0}$.

Smoothing Penalty

1. FOCUSS, Iteratively reweighted least squares
   Difficult to analyze
2. Using l_1 norm: Basis Pursuit (BP)

$$(P_1): \quad \underset{\mathbf{x}\in\mathbb{R}^m}{\arg\min} \ \|\mathbf{x}\|_1 \ \text{subject to } \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{16}$$

Candes and Tao: If $\mathbf{A}$ holds Restricted Isometry Property (RIP), then we need $n = \mathcal{O}(\|\mathbf{x}\|_0 \log m)$ for perfect recovery.

Evaluating Some Algorithms

1. Write the codes for some algorithms and experiment with them. Find by your own that it really works.
2. Design a formal experiment setup and evaluate the algorithms.
3. The project is assigned through the tutorial website.

Theory - The Guarantee Question

Remark

Question: Assume that $\mathbf{Ax} = \mathbf{b}$ *has a sparse solution with* k_0 *non-zeros, i.e.,* $\|\mathbf{x}\|_0 = k_0$*. Furthermore, assume that* $k_0 < \frac{1}{2} spark(\mathbf{A})$*. Will OMP, BP succeed in recovering the sparsest solution?*

1. Note that, such success for any k_0 and $\mathbf{A}$ is not possible due to the known conflict of NP-hardness.
2. However, if the solution is "sufficiently sparse", the success of the some algos is guaranteed.

OMP Performance Guarantee

Theorem

Equivalence - OMP : For $\mathbf{Ax} = \mathbf{b}$ *where* $\mathbf{A}$ *is full row rank, if a solution* $\mathbf{x}$ *exists such that*

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{A})}\right), \tag{17}$$

OMP is guaranteed to find the solution. That means the solution is both the unique solution of OMP and the unique solution of (P_0)

Proof: We work out.

Proof: w.l.o.g let the true $\mathbf{x}$ be in such a way that all its k_0 non-zero elements are at the beginning of the vector, in decreasing order of the values x_j.
First iteration: to choose one from k_0 entries in the vector, also we assumed that x_1 is the highest. So it should choose 1st entry.
That means, we must need

$$\forall i > k_0,\ |\mathbf{a}_1^T \mathbf{b}| > |\mathbf{a}_i^T \mathbf{b}|$$

$$|\sum_{t=1}^{k_0} x_t \mathbf{a}_1^T \mathbf{a}_t| > |\sum_{t=1}^{k_0} x_t \mathbf{a}_i^T \mathbf{a}_t|$$

Proof technique: A game of using lower bound and upper bound judiciously.

$$
\begin{aligned}
L.H.S &= |\sum_{t=1}^{k_0} x_t \mathbf{a}_1^T \mathbf{a}_t| \\
&\geq |x_1| - \sum_{t=2}^{k_0} |x_t| |\mathbf{a}_1^T \mathbf{a}_t| && Use |a+b| \geq |a| - |b| \\
&\geq |x_1| - \sum_{t=2}^{k_0} |x_t| \ \mu(\mathbf{A}) && Use |\mathbf{a}_1^T \mathbf{a}_t| \leq \mu(\mathbf{A}) \\
&\geq |x_1| - |x_1| (k_0 - 1) \ \mu(\mathbf{A})
\end{aligned}
$$

$$
\begin{aligned}
R.H.S &= |\sum_{t=1}^{k_0} x_t \mathbf{a}_i^T \mathbf{a}_t| \leq \sum_{t=1}^{k_0} |x_t| |\mathbf{a}_i^T \mathbf{a}_t| \leq \sum_{t=1}^{k_0} |x_t| \ \mu(\mathbf{A}) \\
&\leq |x_1| \, k_0 \ \mu(\mathbf{A})
\end{aligned}
$$

So if the lower bound $>$ upper bound, then we get $\|\mathbf{x}\|_0 \triangleq k_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{A})}\right)$. Equivalence of OMP proved.

BP Performance Guarantee

Theorem

Equivalence - BP : *For* $\mathbf{Ax} = \mathbf{b}$, *if a solution* $\mathbf{x}$ *exists such that*

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{A})}\right), \tag{18}$$

then that solution is both the unique solution of (P_1) *and the unique solution of* (P_0).

Proof: We skip it for now, if time permits we come back.

1. Note: In general case, both OMP and BP have same worst case bounds. This is not alright.
2. Can we do something more? Different kind of tools and analysis? Yes, we can, like the Tropp's Exact Recovery Condition. But, we mostly skip them. More interested readers should go by themselves.

Exact to Approximate (with Noise) - General Motivation

1. The exact constraint $\mathbf{Ax} = \mathbf{b}$ is often relaxed, with a quadratic penalty $Q(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\|_2^2$.
2. Such relaxation allows us to
   - define a quasi-solution in case no exact solution exists (even in the case of an over-determined setup),
   - exploit ideas from optimization theory, and
   - measure the quality of a candidate solution.

Therefore, we relax the (P_0) problem with the use of an error tolerance $\epsilon > 0$,

$$(P_0^\epsilon): \quad \underset{\mathbf{x}\in\mathbb{R}^m}{\arg\min} \ \|\mathbf{x}\|_0 \ \text{subject to} \ \|\mathbf{b} - \mathbf{Ax}\|_2 \leq \epsilon. \tag{19}$$

Remark

A comment: When (P_0) *and* (P_0^ϵ) *are applied on the same problem instance, the error-tolerent problem* (P_0^ϵ) *must always provide results at-least as sparse as those arising in the exact constrained problem* (P_0)*, since the feasible solution set is wider.*

Remark

An alternative interpretation: Interpreting the problem (P_0^ϵ) *as a noise removal scheme. Consider a sufficiently sparse vector* $\mathbf{x}_0$*, and assume that* $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{e}$*, where* $\mathbf{e}$ *is a nuisance vector of finite energy* $\|\mathbf{e}\|_2^2 = \epsilon^2$*. Rougly speaking* (P_0^ϵ) *aims to find* $\mathbf{x}_0$*, i.e., to do rougly the same thing as* (P_0) *would be on noiseless data* $\mathbf{b} = \mathbf{A}\mathbf{x}_0$*.*

Our rational thought process

1. We can have a rationale that the results for (P_0^ϵ) are some ways parallel to those in the noiseless case (P_0).
2. Specifically, we should discuss the uniqueness property - conditions under which a sparse solution is known to be the global minimizer of (P_0^ϵ) and hence the true solution.

Stability of the sparsest solution

Remark

A fundamental question: *Suppose that a sparse vector* $\mathbf{x}_0$ *is pre-multiplied by* $\mathbf{A}$*, and we obtain a noise version as* $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{e}$ *with* $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon$*. Let*

$$\mathbf{x}_0^\epsilon = \arg\min_{\mathbf{x}} \ \|\mathbf{x}\|_0 \ \text{subject to} \ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon. \tag{20}$$

1. *How good shall this approximation be?*
2. *How its accuracy is affected by the sparsity of* $\mathbf{x}_0$*?*

These questions are the natural extension from the uniqueness porperty of (P_0)*.*

Uniqueness versus stability - Gaining intuition

A question: Can we have the sparsest solution for (P_0^ϵ) unique?
Answer: No, we can not claim uniqueness for the solution of(P_0^ϵ).

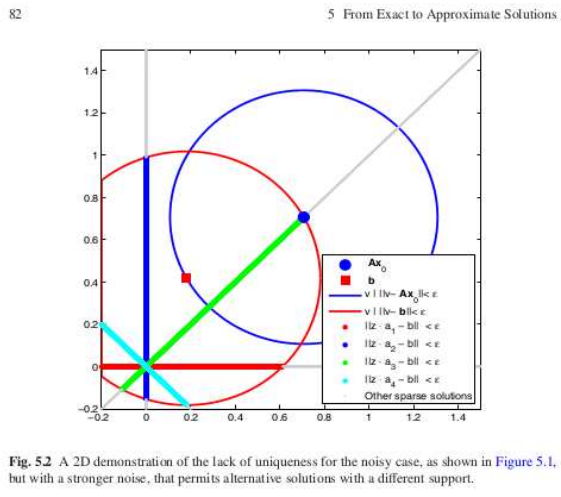

Fig. 5.2 A 2D demonstration of the lack of uniqueness for the noisy case, as shown in Figure 5.1, but with a stronger noise, that permits alternative solutions with a different support.

Figure: Note ANSWER. (The figure is copied from the reference book by Michael Elad. Possibly this is a copyright violation)

1. A different support solution with same sparsity level.
2. Even null solution can be a solution.

Theoretical Study of stability of (P_0^ϵ)

- Instead of claiming uniqueness of a sparse solution, we try to be happy with a notion of stability - a claim that if a sufficiently spase solution is found, then all alternative solutions necessarily resides (very) close to it.
- Starting point: Extending the notion of 'spark' by considering a relaxed notion of linear dependency.
- Recall that *spark*($\mathbf{A}$) is the minimum number of linearly dependent columns. Mathematically, it was defined as

$$spark(\mathbf{A}) = \min_{\mathbf{x}\in\mathbb{R}^m} \ \|\mathbf{x}\|_0 \ \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{0}, \ \ \mathbf{x} \neq \mathbf{0}. \qquad (21)$$

- If $\mathbf{x}_1$ and $\mathbf{x}_2$ are two solutions for noiseless case $\mathbf{A}\mathbf{x} = \mathbf{b}$, then we can have $\mathbf{d} = \mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{A}\mathbf{d} = \mathbf{0}$. This motivates the null space characterization.

- Following the same rationale, if there exists two feasible solutions $\mathbf{x}_1$ and $\mathbf{x}_2$ satisfying $\|\mathbf{A}\mathbf{x}_i - \mathbf{b}\|_2 \leq \epsilon$, $i = 1, 2$, then we can have $\mathbf{d} = \mathbf{x}_1 - \mathbf{x}_2$ and $\|\mathbf{A}\mathbf{d}\|_2 = \|\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)\|_2 \leq 2\epsilon$.
- Therefore, we may generalize the spark to allow for ϵ-proximity to the null-space.

Definition

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$*, we consider all possible sub-sets of* s *columns, each such set forrms a sub-matrix* $\mathbf{A}_s \in \mathbb{R}^{n \times s}$*. We define* $spark_\eta(\mathbf{A})$ *as the smallest possible* s *(number of columns) that guarantees*

$$\min_s \ \sigma_s(\mathbf{A}_s) \leq \eta. \tag{22}$$

In Words: This is the smallest (integer) number of columns that can be gathered from $\mathbf{A}$*, such that the smallest singular-value of* $\mathbf{A}_s$ *is no larger than* η*.*

Question: How this new stuff is connected with the spark definition? For $\eta = 0$, $spark_0(\mathbf{A}) = spark(\mathbf{A})$.

Theorem

Stability of (P_0^ϵ): Consider the instance of problem (P_0^ϵ) defined by the triplet $(\mathbf{A}, \mathbf{b}, \epsilon)$. Suppose that a sparse vector $\mathbf{x}_0 \in \mathbb{R}^m$ satisfies that sparsity constraint $\|\mathbf{x}_0\|_0 < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{A})})$, and gives a representation of $\mathbf{b}$ within error tolerance ϵ (i.e., $\|\mathbf{b} - \mathbf{A}\mathbf{x}_0\|_2 \leq \epsilon$). Every solution $\mathbf{x}_0^\epsilon$ of (P_0^ϵ) must obey

$$\|\mathbf{x}_0^\epsilon - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathbf{A})(2\|\mathbf{x}_0\|_0 - 1)} \tag{23}$$

Proof: We will see ....

Remark

Note that this result parallels the uniqueness result for (P_0) problem, and indeed it reduces to it exactly for the case of $\epsilon = 0$.

RIP and Stability Analysis

- We now introduce a new measure of quality for a given matrix $\mathbf{A}$ that replaces mutual-coherence and spark.
- The property introduced by Candes and Tao - Restricted Isometry Property (RIP)

Definition

For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ with l_2-normalized columns, and for an integer scalar $s \leq n$, consider sub-matrices $\mathbf{A}_s$ containing s-columns from $\mathbf{A}$. Define δ_s as the smallest quantity such that

$$\forall \mathbf{c} \in \mathbb{R}^s, \;\; (1 - \delta_s)\|\mathbf{c}\|_2^2 \leq \|\mathbf{A}_s\mathbf{c}\|_2^2 \leq (1 + \delta_s)\|\mathbf{c}\|_2^2 \tag{24}$$

holds true for any choice of s columns. Then $\mathbf{A}$ is said to have an s-RIP with a constant δ_s.

- Note: The above definition is only informative when $\delta_s < 1$.
- Key idea: The key idea is that any subset of 's' columns from '$\mathbf{A}$' behave like an orthogonal transform that loses/gains almost no energy.
- Explain: What 'Restricted Isometry' means? We will have some pictorial illustration.

Some important points:

- There is a close resemblance between RIP and $spark_{\eta}(\mathbf{A})$.
- $spark_{\eta}(\mathbf{A})$ is the minimum number of required columns 's' such that the lowest singular value of $\mathbf{A}_s$ is η away from singularity. This follows from $spark_{\eta}(\mathbf{A})$ definition.
- RIP fixes 's' and seeks the maximum of $(1 - \delta_s)$, again implying that any set of 's' columns or the matrix $\mathbf{A}_s$ is $(1 - \delta_s)$ away from singularity.
- However, RIP is richer than $spark_{\eta}(\mathbf{A})$ as it is also bounded from above. That means it is defined by lower and upper bounds.

Evaluation of RIP constant δ_s

- Given a matrix $\mathbf{A}$, it is hard or impossible to evaluate δ_s
- However, just like $spark_\eta(\mathbf{A})$ was bounded by mutual coherence $\mu(\mathbf{A})$, we can also bound δ_s

Remark

Important bound: $\delta_s \leq (s-1)\mu(\mathbf{A})$

Proof: This proof takes a style of using lower bound and upper bound. Check reference.

An alternating proof: By Gershgorin Disk Theorem and Eigenvalues of Gram matrix $\mathbf{A}_s^T\mathbf{A}_s$. This proof is more aligned with System Identification studies.

Coming back to stability of (P_0^ϵ):

Theorem

Stability of (P_0^ϵ): Consider the instance of problem (P_0^ϵ) defined by the triplet $(\mathbf{A}, \mathbf{b}, \epsilon)$. Suppose that a sparse vector $\mathbf{x}_0 \in \mathbb{R}^m$ satisfies that sparsity constraint $\|\mathbf{x}_0\|_0 < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{A})})$, and gives a representation of $\mathbf{b}$ within error tolerance ϵ (i.e., $\|\mathbf{b} - \mathbf{A}\mathbf{x}_0\|_2 \leq \epsilon$). Every solution $\mathbf{x}_0^\epsilon$ of (P_0^ϵ) must obey

$$\|\mathbf{x}_0^\epsilon - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathbf{A})(2\|\mathbf{x}_0\|_0 - 1)} \tag{25}$$

Proof: We now work out. A better style of proving by RIP.

In the proof, illustrate where is the requirement $\|\mathbf{x}_0\|_0 < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{A})})$?

- Let $\mathbf{x}_0$ is the true vector with $\|\mathbf{x}_0\|_0 = s_0$ and $\|\mathbf{b} - \mathbf{A}\mathbf{x}_0\|_2 \leq \epsilon$
- Let $\tilde{\mathbf{x}}$ is a feasible soln of (P_0^ϵ). So $\|\tilde{\mathbf{x}}\|_0 \leq \|\mathbf{x}_0\|_0$ and $\|\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}\|_2 \leq \epsilon$
- Let $\mathbf{d} = \tilde{\mathbf{x}} - \mathbf{x}_0$. So $\|\mathbf{A}\mathbf{d}\|_2 = \|(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}) + (\mathbf{b} - \mathbf{A}\mathbf{x}_0)\|_2 \leq 2\epsilon$ and $\|\mathbf{d}\|_0 \leq 2s_0$
- Let assume that $\mathbf{A}$ satisfies RIP for $2s_0$, with the constant $\delta_{2s_0} < 1$. So by LB of RIP, we have $(1 - \delta_{2s_0})\|\mathbf{d}\|_2^2 \leq \|\mathbf{A}\mathbf{d}\|_2^2$ and $\|\mathbf{A}\mathbf{d}\|_2^2 \leq 4\epsilon^2$.
- So $\|\mathbf{d}\|_2^2 = \|\tilde{\mathbf{x}} - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1-\delta_{2s_0}}$
- Recall $\delta_{2s_0} \leq (2s_0 - 1)\mu(\mathbf{A})$. Plug this to the right side of above and we get $\|\mathbf{d}\|_2^2 \leq \frac{4\epsilon^2}{1-\delta_{2s_0}} \leq \frac{4\epsilon^2}{1-(2s_0-1)\mu(\mathbf{A})}$
- To remain valid $1 - (2s_0 - 1)\mu(\mathbf{A}) > 0$ and so we require $s_0 < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{A})})$

The style is more general

- The analysis (or proof) can be generalized more.
- Let us say a feasible solution $\mathbf{x}_0^\epsilon$ follows $\|\mathbf{x}_0^\epsilon\|_0 = s_1$. Then, we can proof $\|\mathbf{x}_0^\epsilon - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1-\mu(\mathbf{A})(\|\mathbf{x}_0\|_0+\|\mathbf{x}_0^\epsilon\|_0-1)}$.

What Happens for Practical Algos?

- OMP : Very easy to implement. Just choose $\epsilon_0 = \epsilon$. With this minor change, the OMP is ready, and this possibly explains it popularity. Problem: Who will supply ϵ?
- Basis pursuit denoising (BPDN) : Second order cone program

$$(P_1^\epsilon): \quad \mathbf{x}_1^\epsilon = \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon. \quad (26)$$

  But, again - Problem: Who will supply ϵ?

- Least absolute shrinkage and selection operator (LASSO) : For an appropriate Lagrange multiplier λ, the solution of BPDN is precisely the solution of the unconstrained problem

$$(Q_1^\lambda): \quad \arg\min_{\mathbf{x}\in\mathbb{R}^m} \left\{\lambda\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2\right\} \quad (27)$$

  But, again - Problem: Who will supply λ?

What we can do?

- We can check many λ and choose the best.
- Well, for the optimal minimizer of (Q_1^λ), the solution should lead to a sub-gradient set that contains the zero vector. Our cost function $f(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$.
- The sub-gradient set is given by all the vectors

$$\partial f(\mathbf{x}) = \left\{\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) + \lambda\mathbf{z}\right\}, \ \forall \mathbf{z}, z_i = 1 \ \textit{if} \ x(i) > 0, \quad (28)$$
$$\textit{or} \ [-1, 1] \ \textit{if} \ x(i) = 0, \ \textit{or} \ -1 \ \textit{if} \ x(i) < 0.$$

When searching the minimizer of $f(\mathbf{x})$, we should seek both $\mathbf{x}$ and $\mathbf{z}$ such that $\mathbf{0} \in \partial f(\mathbf{x})$.

- The goal in LASSO: Can we solve LASSO for all possible choice of λ at once?
- Surprising answer: Yes, there exists that kind of Algorithm: Least Angle Regression Stagewise (LARS).

Performance Guarantee

Theorem

BPDN Stability Guarantee: For (P_1^ϵ), *suppose that* $\mathbf{x}_0$ *is a feasible solution satisfying* $\|\mathbf{x}_0\|_0 < \frac{1}{4}\left(1 + \frac{1}{\mu(\mathbf{A})}\right)$. *The solution* $\mathbf{x}_1^\epsilon$ *of* (P_1^ϵ) *must obey*

$$\|\mathbf{x}_1^\epsilon - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathbf{A})(4\|\mathbf{x}_0\|_0 - 1)} \tag{29}$$

Proof: We work out.

- Both $\mathbf{x}_0$ and $\mathbf{x}_1^\epsilon$ are feasible solutions. Let $\mathbf{d} = \mathbf{x}_1^\epsilon - \mathbf{x}_0$. So $4\epsilon^2 \geq \|\mathbf{Ad}\|_2^2$. Now define Gram Matrix $\mathbf{G} = \mathbf{A}^T\mathbf{A}$. $\mathbf{1}$ is a square matrix with all elements are one.
- $4\epsilon^2 \geq \|\mathbf{Ad}\|_2^2 = \mathbf{d}^T\mathbf{A}^T\mathbf{Ad} = \mathbf{d}^T\mathbf{Gd} = \mathbf{d}^T\mathbf{Id} + \mathbf{d}^T(\mathbf{G} - \mathbf{I})\mathbf{d} \geq \|\mathbf{d}\|_2^2 - |\mathbf{d}|^T|\mathbf{G} - \mathbf{I}||\mathbf{d}| \geq \|\mathbf{d}\|_2^2 - \mu(\mathbf{A})|\mathbf{d}|^T|\mathbf{1} - \mathbf{I}||\mathbf{d}| = (1+\mu(\mathbf{A}))\|\mathbf{d}\|_2^2 - \mu(\mathbf{A})|\mathbf{d}|^T\mathbf{1}|\mathbf{d}| = (1+\mu(\mathbf{A}))\|\mathbf{d}\|_2^2 - \mu(\mathbf{A})\|\mathbf{d}\|_1^2$
- $\mathbf{x}_1^\epsilon$ is the lowest ℓ_1 norm solution. So $\|\mathbf{x}_1^\epsilon\|_1 = \|\mathbf{d} + \mathbf{x}_0\|_1 \leq \|\mathbf{x}_0\|_1$.
- Let $\mathbf{x}_0$ has k_0 non-zeros elements, its support is denoted by S and the k_0 non-zeros elements are first k_0 elements (we can always do that by column permutation)

- $0 \geq \|\mathbf{d}+\mathbf{x}_0\|_1 - \|\mathbf{x}_0\|_1 = \sum_{j=1}^{k_0} |d_j + x_{0j}| - |x_{0j}| + \sum_{j>k_0} |d_j| \geq -\sum_{j=1}^{k_0} |d_j| + \sum_{j>k_0} |d_j| = -2\sum_{j=1}^{k_0} |d_j| + \sum_{\forall j} |d_j| = -2\mathbf{1}_S^T|\mathbf{d}| + \|\mathbf{d}\|_1$. Or $\|\mathbf{d}\|_1 \leq 2\mathbf{1}_S^T|\mathbf{d}|$. Here $\mathbf{1}_S$ is a vector where elements on support S are one, others zeros.
- $\forall \mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\|_1 \leq \sqrt{n}\|\mathbf{v}\|_2$. So $\|\mathbf{d}\|_1 \leq 2\mathbf{1}_S^T|\mathbf{d}| \leq 2\sqrt{|S|}\|\mathbf{d}_S\|_2 \leq 2\sqrt{|S|}\|\mathbf{d}\|_2 = 2\sqrt{k_0}\|\mathbf{d}\|_2$
- $4\epsilon^2 \geq (1+\mu(\mathbf{A}))\|\mathbf{d}\|_2^2 - \mu(\mathbf{A})\|\mathbf{d}\|_1^2 \geq (1+\mu(\mathbf{A}))\|\mathbf{d}\|_2^2 - \mu(\mathbf{A})4k_0\|\mathbf{d}\|_2^2$.
- Or $\|\mathbf{d}\|_2^2 \leq \frac{4\epsilon^2}{1-\mu(\mathbf{A})(4k_0-1)}$

Further results (by Candes and Tao using RIP)

Theorem

Let us define $\mathcal{T}_s$ to be the set of all strictly s-sparse signals, i.e.,

$$\mathcal{T}_s = \{\mathbf{x} \in \mathbb{R}^m \ : \ \|\mathbf{x}\|_0 = s\}. \tag{30}$$

Then let us denote $\mathbf{x}_s$ as the best s-term approximation of a compressible signal $\mathbf{x}$ according to

$$\mathbf{x}_s = \underset{\mathbf{x}' \in \mathcal{T}_s}{\arg\min} \|\mathbf{x} - \mathbf{x}'\|_1. \tag{31}$$

Suppose that $\mathbf{A}$ holds the RIP of order $2s$ with isometry constant $\delta_{2s} < \sqrt{2} - 1$. Given a noisy measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$ and $\|\mathbf{w}\|_2 \leq \epsilon$, the solution to (P_1^ϵ) obeys $\|\mathbf{x} - \mathbf{x}_1^\epsilon\|_2 \leq C_0\epsilon + C_1 \frac{\|\mathbf{x}-\mathbf{x}_s\|_1}{\sqrt{s}}$, where C_0 and C_1 are typically small constants.

Iteratively-Reweighted-Least-Squares (IRLS)

Main Idea

- Setting $\mathbf{X} = diag(|\mathbf{x}|)$, we have $\|\mathbf{x}\|_1 = \mathbf{x}^T\mathbf{X}^{-1}\mathbf{x}$
- We can view the l_1-norm as an adaptively weighted l_2 norm
- In kth iteration, given a current approximate solution $\mathbf{x}_{k-1}$, set $\mathbf{X}_{k-1} = diag(|\mathbf{x}_{k-1}|)$ and attempt to solve for

$$(M_k): \quad \min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \lambda\, \mathbf{x}^T\mathbf{X}_{k-1}^{-1}\mathbf{x} + \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\} \tag{32}$$

  The above problem is regularized LS. The solution is assigned to form $\mathbf{x}_k$

- Initialization: $\mathbf{x}_0 = \mathbf{1}$.

Summary

- Noise modeling: Till now we performed worst case study by assuming that the noise always has the highest l_2 strength ϵ. So, we always use the worst realization of noise in a deterministic sense. There exists better results by shifting to random noise model, accompanied by a near one probability to the claimed bounds.
- Allowing rare failures: Worst case analysis does not allow failure. By allowing a small fraction of failure, it is possible to get better bounds.
- Worst case characterization of $\mathbf{A}$: The mutual coherence, spark, RIP are all leads to worst case analysis - over pessimistic results. Can we introduce more relaxed measures, such as a probabilistic RIP / coherence? Statistical mechanics based approach?

Applications

- Source coding
- Wireless channel estimation
- Wireless sensor network
- Information Theory based proofs for Sparse Reconstruction
- Cognitive radio

Check the review paper: K. Hayashi, M. Nagahara and T Tanaka, A Users Guide to Compressed Sensing for Communications Systems, IEICE Trans Communications, March 2013.

Source coding

The best example is Compressed Sensing.

$$\mathbf{b} = \mathbf{A}\mathbf{x} \in \mathbb{R}^{n \times 1}, \tag{33}$$

where $\mathbf{x} \in \mathbb{R}^{m \times 1}$ and $n < m$. The assumption is $\mathbf{x}$ is sparse in some basis, for example in Fourier transform $\mathbf{F}$ or DCT, and then we have $\mathbf{y} = \mathbf{F}\mathbf{x}$, where $\mathbf{y}$ is sparse. The effective measurement is $\mathbf{b} = \mathbf{A}\mathbf{F}^{-1}\mathbf{y} = \mathbf{C}\mathbf{y}$.

- We have $\mathbf{b}$, but further $\mathbf{b}$ needs to be quantized and sent over the channel - Source coding[1].
- Compressed sensing has connection with channel coding[2].
- If $\mathbf{x}$ is not in one place, but in a distributed setup. Distributed source and joint source-channel coding.

[1] V. K. Goyal, A. K. Fletcher, and S. Rangan, Compressive Sampling and Lossy Compression, IEEE Signal Processing Magazine, 25(2):48-56, March 2008.

[2] A.G. Dimakis et al, LDPC Codes for Compressed Sensing, IEEE Trans. Information Theory, 2012

- We have $\mathbf{b}$, but further $\mathbf{b}$ needs to be quantized and sent over the channel - Source coding[3].
- Compressed sensing has connection with channel coding[4].
- If $\mathbf{x}$ is not in one place, but in a distributed setup. Distributed source and joint source-channel coding [5].

---

[3] V. K. Goyal, A. K. Fletcher, and S. Rangan, Compressive Sampling and Lossy Compression, IEEE Signal Processing Magazine, 25(2):48-56, March 2008.

[4] A.G. Dimakis et al, LDPC Codes for Compressed Sensing, IEEE Trans. Information Theory, 2012

[5] A. Shirazinia, S. Chatterjee and M. Skoglund, Joint source-channel vector quantization for compressed sensing, Accepted for IEEE Trans. Signal Proc., 2014

Wireless channel estimation

- Sparsity in channel impulse response, for example, larger bandwidth wireless channel [6], Underwater acoustic channel[7]
- using training signals (pilots) $\mathbf{t}$, most of the times we can form a system where received signal $\mathbf{y} = \mathbf{T}\mathbf{x}$ and typically $\mathbf{T}$ is Toeplitz and $\mathbf{x}$ is channel impulse response (sparse) [8].

---

[6] Raghavan V., Hariharan G., Sayeed A.M., Capacity of Sparse Multipath Channels in the Ultra-Wideband Regime, Selected Topics in Signal Processing, IEEE Journal of , vol.1, no.3, pp.357,371, Oct. 2007

[7] Berger C.R., Shengli Zhou, Preisig J.C., Willett P., Sparse Channel Estimation for Multicarrier Underwater Acoustic Communication: From Subspace Methods to Compressed Sensing, Signal Processing, IEEE Transactions on , vol.58, no.3, pp.1708,1721, March 2010

[8] Berger C.R., Zhaohui W., J. Huang, Shengli Z., Application of compressive sensing to sparse channel estimation, Communications Magazine, IEEE , vol.48, no.11, pp.164,174, November 2010

Wireless sensor network

- Let m and x_j, $j = 1, 2, ..., m$ denote the number of sensors and the measured signal at the jth node [9].
- Each sensor node sends x_j to the fusion center using n time slots ($n < m$) with n random coeeficients $\{A_{i,j}\}_{i=1}^{n}$. Since m sensor nodes send $A_{i,j}x_j$ in the ith slot simultaneously, the recived signal at fusion center at ith slot is $y_i = \sum_{j=1}^{m} A_{i,j}x_j + noise$
- Assume $\mathbf{x} = \{x_j\}$ is sparse in some domain, for example differential domain, or DFT, etc.

[9] Haupt J., Bajwa W.U., Rabbat M., Nowak R., Compressed Sensing for Networked Data, Signal Processing Magazine, IEEE , vol.25, no.2, pp.92,101, March 2008

Information Theory based proofs for Sparse Reconstruction

- A specific example: Connection the Gaussian Multiple Access Channel (MAC) model for Sparse Reconstruction [10]

[10] Yuzhe J., Young-Han Kim, Rao B.D. Limits on Support Recovery of Sparse Signals via Multiple-Access Communication Techniques, IEEE Trans on Information Theory, vol.57, no.12, pp.7877,7892, Dec. 2011

Cogitive Radio

- A specific example: Distributed spectrum sensing
- Lets say there exists lots of sensors that observe spectrum, and spectrum is sparse. So the question is how to detect the unoccupied spectrum bands in consensus? [11]

[11] Zeng Fanzi, Chen Li, Zhi Tian, Distributed Compressive Spectrum Sensing in Cooperative Multihop Cognitive Networks, IEEE Jour. of Selected Topics in Signal Processing, vol.5, no.1, pp.37,48, Feb. 2011

Orthogonal Application: Bioinformatics

- In metagenomics where the task is to identify the proportions of species from large scale metagenomic data.
- Philosophy: In any sample, number of known species are small compared to all known species. So, IF WE ASSUME A SPECIES space, then a species in a sample is sparse. [12]

[12] S. Chatterjee, D. Koslicki, S. Dong, N. Innocenti, L Cheng, Y Lan, M Vehkapera, M. Skoglund, L.K. Rasmussen, E. Aurell and J. Corander, SEK: Sparsity exploiting k-mer-based estimation of bacterial community composition, Bioinformatics, 2014

Lets begin .... Low rank matrix system

Let us consider a matrix $\mathbf{X} \in \left\{\mathbf{X} \in \mathbb{C}^{P\times N} : \mathrm{rank}(\mathbf{X}) = r\right\}$, where $r \ll \min(P, N)$. The linear measurement

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) \in \mathbb{R}^{M\times 1}, \tag{34}$$

where $\mathcal{A} : \mathbb{R}^{P\times N} \rightarrow \mathbb{R}^{M\times 1}$ and $M < PN$. Further

$$\mathcal{A}(\mathbf{X}) = \begin{bmatrix} \langle \mathbf{X}, \mathbf{A}_1 \rangle \\ \vdots \\ \langle \mathbf{X}, \mathbf{A}_M \rangle \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_M \end{bmatrix} vec(\mathbf{X}) = \mathbf{B}\, vec(\mathbf{X}), \tag{35}$$

where $\langle \mathbf{X}, \mathbf{A}_m \rangle \triangleq \mathrm{trace}(\mathbf{A}_m^t \mathbf{X}) = \mathbf{b}_m^t\, vec(\mathbf{X})$.

- Note $vec(\mathbf{X})$ is not sparse in any linear transform. But sparse in SVD.
- So far this new system has limited investigation for communication technologies.

Thank You