PhD Course in KTH - Sparse Signal Processing
Slides 6
Discussion Topic - Exact to Approximate (Continuing...)

Saikat Chatterjee

Communication Theory Lab, KTH

April 17, 2014

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

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

Proof: We work out

Remark

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

Remark

In proving the previous theorem (Stability of (P_0^ϵ)), we found certain anomalies. So, we go for another path, by introducing RIP and using it.

RIP and Stability Analysis

- We now introduce a new measure of quality for a given matrix **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*

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

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.
- Illustration: Pictorial connection between RIP constant δ_s and $spark_\eta(\mathbf{A})$.

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: We work out. This proof takes a style of using lower bound and upper bound

An alternating proof: By Gershgorin Disk Theorem and Eigenvalues of Gram matrix $\mathbf{A}_s^T\mathbf{A}_s$. We discuss the steps of the proof.

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

$$\|\mathbf{x}_0^\epsilon - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathbf{A})(2\|\mathbf{x}_0\|_0 - 1)} \tag{3}$$

Proof: We now work out. A better style of proving by RIP.
In the proof, illustrate where is the requirement $\|\mathbf{x}_0\|_0 < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{A})})$?

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

$$(P_1^\epsilon): \quad \mathbf{x}_1^\epsilon = \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon. \quad (4)$$

  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

$$(Q_1^\lambda): \quad \arg\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \lambda\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\} \quad (5)$$

  But, again - Problem: Who will supply λ?

What we can do?

- We can check many λ and choose the best. But, by which is basis?
- 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

$$\partial f(\mathbf{x}) = \left\{\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) + \lambda\mathbf{z}\right\},\ \forall \mathbf{z}, z_i = 1 \ \textit{if}\ \ x(i) > 0, \quad (6)$$
$$\textit{or}\ [-1, 1]\ \textit{if}\ \ x(i) = 0,\ \textit{or}\ \ -1\ \textit{if}\ \ x(i) < 0.$$

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*

$$\|\mathbf{x}_1^\epsilon - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathbf{A})(4\|\mathbf{x}_0\|_0 - 1)} \tag{7}$$

Proof: We work out.

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.,

$$\mathcal{T}_s = \{\mathbf{x} \in \mathbb{R}^m \ : \ \|\mathbf{x}\|_0 = s\}. \tag{8}$$

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

$$\mathbf{x}_s = \arg\min_{\mathbf{x}' \in \mathcal{T}_s} \|\mathbf{x} - \mathbf{x}'\|_1. \tag{9}$$

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.

Proof: ?

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

$$(M_k): \quad \min_{\mathbf{x}\in\mathbb{R}^m} \left\{\lambda\, \mathbf{x}^T\mathbf{X}_{k-1}^{-1}\mathbf{x} + \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2\right\} \tag{10}$$

  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 analaysis - over pessimistic results. Can we introduce more relaxed measures, such as a probabilistic RIP / coherence?