Random matrix theory questions arising in Compressed Sensing and related areas

Bubacarr Bah
with Prof. Jared Tanner

School of Mathematics, University of Edinburgh
and
Maxwell Institute

Foundations of Computational Mathematics
July 2011
Budapest, Hungary

1. Compressed Sensing (CS) and Related Areas
2. Restricted Isometry Property (2) - RIP_2
   - Restricted Isometry Constants (2) - RIC_2
   - Improved RIC_2 Bounds for the Gaussian Ensemble
   - Numerical Results
   - Performance Guarantees for CS Algorithms
   - Finite (k, n, N) interpretations
3. Asymptotic Approximation of RIC_2 Bounds
   - Approximation of Bounds
   - Implications for CS
4. Conclusion

Compressed Sensing & Spare Approximation

- Signal $x \in \mathbb{R}^N$, k-sparse.
- Sensing matrix $A \in \mathbb{R}^{n\times N}$; measurements $y = Ax, (n \ll N)$.
- Problem $(P_0^k): \min_{x\in\mathbb{R}^N} \|x\|_0 \quad \text{s.t.} \quad Ax = y.$
- Solution: l_q minimizations & Greedy Algorithms (OMP, IHT, ...)

Rank Minimization & Matrix Completion

- Matrix $X \in \mathbb{R}^{m\times n}$, low rank $\mathrm{rank}(X) \leq r$.
- Linear map $\mathcal{A}(X) : \mathbb{R}^{m\times n} \to \mathbb{R}^p$; measurements $y = \mathcal{A}(X) \in \mathbb{R}^p$.
- Problem $(P_0^r): \min_{X} \mathrm{rank}(X) \quad \text{s.t.} \quad \mathcal{A}(X) = y.$
- Solution: $\|\cdot\|_*$ minimizations, & Greedy Algorithms (SVT, ...)

CS Applications

- Medical Imaging: MRI, fMRI, Radiology, ...
- Infrared spectroscopy & Seismic imaging
- Single pixel camera & Analog-to-digital converters
- DNA micro-arrays, radar, wireless communications, ...

Tools of Analysis

- Coherence [Donoho & Huo; Elad & Bruckstein]
- Restricted isometry property [Candès & Tao]
- Nullspace property [Donoho & Huo]
- Stochastic geometry [Donoho; Donoho & Tanner]
- Message passing [Donoho, Maleki & Montanari]

- RIP certainly a popular tool of analysis; FoCM'11?
- With the introduction of RIP_1, we refer to the standard RIP as RIP_2 - the subscripts 1 & 2 refer to the norms used

Definition

RIC_2 of A of order k is the smallest number R_k, for all k-sparse x, such that

$$(1 - R_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + R_k)\|x\|_2^2$$

Definition (Rank Minimization Equivalence)

RIC_2 of $\mathcal{A}(X)$, the r-restricted isometry constant, is the smallest number R_r, for all matrices X of rank at most r, such that

$$(1 - R_r)\,\|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1 + R_r)\,\|X\|_F^2$$

- A having RIP_2 means that A is a near isometry for k-sparse x
- RIP_2 gives a sufficient guarantees for exact recovery

ℓ_1 minimization works if:

- $R_{3k} + 3R_{4k} < 2$, [Candès, Romberg & Tao, 2006]
- $R_{2k} < \sqrt{2} - 1$, [E. Candès, 2008]
- $R_{2k} < 2/(3 + \sqrt{7/4}) \approx 0.4627$, [S. Foucart, 2010]

Similarly for Greedy Algorithms:

- **IHT:** $R_{3k} < 1/\sqrt{3}$, [S. Foucart, 2011]
- **CoSaMP:** $R_{4k} < \sqrt{2/(5 + \sqrt{73})}$, [S. Foucart, 2011]
- **Subspace Pursuit (SP):** $R_{3k} \lesssim 0.06$, [Dai & Milenkovic, 2009]

- A more quantitative definition is the asymmetric RIP_2:

Definition

RIC_2 of A of order k is the smallest L & U, for all k-sparse x, s.t.

$$(1 - L(k, n, N; A))\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + U(k, n, N; A))\|x\|_2^2.$$

- RIC_2 of A & eigenvalues of $A_K^* A_K$, for $\Omega = \{1, 2, 3, \dots, N\}$

$$1 + U(k, n, N; A) := \max_{K \subset \Omega, |K| = k} \lambda^{max}\left(A_K^* A_K\right)$$

$$1 - L(k, n, N; A) := \min_{K \subset \Omega, |K| = k} \lambda^{min}\left(A_K^* A_K\right)$$

- Thus L & U are smallest & largest deviation from unity of smallest & largest $\lambda\left(A_K^* A_K\right)$ respectively

- RIC_2 combinatorial, intractable for deterministic A, NP-hard
- Probabilistic bounds possible $\Rightarrow$ the use of random matrices
- Different approaches include:
  - Largest ensembles with bounded RIC_2, [Mandelson et. al.]
  - RIC_2 bounds for partial Fourier matrices, [H. Rauhut]
  - RIC_2 bounds for Gaussian matrices, [C. & Tao; B.C.T.]

Goal:

Calculate accurate RIC_2 bounds for Gaussian random matrices with entries drawn i.i.d. from $\mathcal{N}(0, 1/n)$

Motivation:

(1) Using the Gaussian to model ***zero mean i.i.d*** ensembles
(2) Easier! - Lot of literature available on the Gaussian ensemble

Linear growth or proportional-growth asymptotics(p.g.a)

Problem instances (k, n, N) considered is where the following ratios converge to nonzero bounded limits:
$\frac{k}{n} = \rho_n \to \rho$ and $\frac{n}{N} = \delta_n \to \delta$ for $(\delta, \rho) \in (0,1)^2$ as $(k, n, N) \to \infty$.

Theorem

Let A be a matrix of size $n \times N$ whose entries are drawn i.i.d. from $\mathcal{N}(0, 1/n)$. For any $\epsilon > 0$, in the proportional-growth asymptotics

$$\mathbf{P}(L(k,n,N) < \mathcal{L}^{BT}(\delta,\rho) + \epsilon) \to 1 \quad \& \quad \mathbf{P}(U(k,n,N) < \mathcal{U}^{BT}(\delta,\rho) + \epsilon) \to 1$$

exponentially in n.

- Prior bounds by Candès & Tao; and Blanchard et. al. (BCT)

- ***Derivation technique:*** finding smallest $\lambda^{max}(\delta,\rho) > 0$ such that in the p.g.a, for $U_k = U(k,n,N;A)$,

$$\mathbf{P}\left(1 + U_k > \lambda^{max}(\delta,\rho)\right) = \mathbf{P}\left(\max_{K\subset\Omega,|K|=k} \lambda^{max}\left(A_K^* A_K\right) > \lambda^{max}(\delta,\rho)\right) \to 0$$

- Candès & Tao used union bounds and concentration of measure bounds on the extreme eigenvalues of Wishart matrices - *valid for sub-gaussian matrices*

$$\mathbf{P}\left(\max_{K\subset\Omega,|K|=k} \lambda^{max}\left(A_K^* A_K\right) > \lambda^{max}(\delta,\rho)\right) \leq \binom{N}{k}\mathbf{P}\left(\lambda^{max}\left(A_K^* A_K\right) > \lambda^{max}(\delta,\rho)\right)$$

where $\lambda^{max}(\delta,\rho) := \left[1 + \sqrt{\rho} + \left(2\delta^{-1}H(\delta\rho)\right)^{1/2}\right]^2$

- Their upper bound is then $\mathcal{U}^{CT}(\delta,\rho) := \lambda^{max}(\delta,\rho) - 1$

- BCT achieved tighter bounds using union bounds and bounds of probability density functions of the extreme eigenvalues of Wishart matrices [A. Edelman, 1989]

$$\mathbf{P}\left(\max_{K\subset\Omega,|K|=k} \lambda^{max}\left(A_K^* A_K\right) > \lambda^{max}\left(\delta,\rho\right)\right) \leq \int_{\lambda^{\max}(\delta_n,\rho_n)}^{\infty} \binom{N}{k} f_{\max}(m,n;\lambda)d\lambda$$

- But $f_{\max}(m,n;\lambda) \leq p_{\max}(n,\lambda;\rho)\exp\left(n\cdot\psi_{\max}(\lambda,\rho)\right)$ where $\psi_{\max}\left(\lambda,\rho\right) := \frac{1}{2}\left[(1+\rho)\ln\lambda - \rho\ln\rho + 1 + \rho - \lambda\right]$
- Bounding $\binom{N}{k}$ by the Stirling's formula the exponent of the exponential term becomes $\delta\psi_{\max}\left(\lambda^{\max}\left(\delta,\rho\right),\rho\right) + H(\delta\rho)$
- In the p.g.a ***only the exponential term matters*** and $\lambda^{\max}\left(\delta,\rho\right)$ becomes a solution to when the exponent is zero
- Their upper bound is thus $\mathcal{U}^{BCT}(\delta,\rho) = \lambda^{\max}\left(\delta,\rho\right) - 1$

- Improvement on BCT bounds achieved by grouping submatrices, i.e. for A_K and $A_{K'}$ with $|K \cap K'| \gg 1$, hence decreasing the combinatorial term significantly

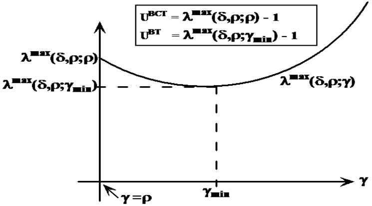
$$\mathbf{P}\left(\max_{K \subset \Omega, |K|=k} \lambda^{max}(A_K^* A_K) > \lambda^{max}(\delta, \rho) \right)$$
$$= \mathbf{P}\left(\max_{i=1,\dots,u} \max_{K \subset \mathcal{G}_i, |K|=k} \lambda^{max}(A_K^* A_K) > \lambda^{max}(\delta, \rho) \right)$$

- The RHS upper bounded using union bound over groups of $m \geq k$ distinct elements and controlling dependencies in $\lambda^{\max}(A_K^* A_K)$ for $K \subset \mathcal{G}_i$ by replacing the maximization over $K \subset \mathcal{G}_i$ by $\lambda^{\max}(A_M^* A_M)$, $M := \bigcup_{K \subset \mathcal{G}_i, |K|=k} K$, $|M| = m \geq k$.

$$\mathrm{RHS} \leq u \cdot \mathbf{P}\left(\lambda^{max}\left(A_M^* A_M\right) > \lambda^{\max}(\delta, \rho; \gamma) \right)$$

where u is the number of groups

- $\lambda^{\max}(\delta,\rho)$ in the BCT analysis becomes $\lambda^{\max}(\delta,\rho;\gamma)$ for each $\gamma := \frac{m}{n} \in [\rho, \delta^{-1})$ by substituting γ for ρ; $\gamma = \rho$ recovers BCT
- Larger values of m decrease the combinatorial term at the cost of increasing $\lambda^{\max}(A_M^* A_M)$
- Interplay btw number & size of groups, $\Rightarrow$ optimizing over γ
- There exist an optimal γ; shown below and proof trivial



- Consequently, $\lambda^{\max}(\delta,\rho) := \min_\gamma \lambda^{\max}(\delta,\rho;\gamma)$ and our upper bound is $\mathcal{U}^{BT}(\delta,\rho) = \lambda^{\max}(\delta,\rho) - 1$

- Form groups $\mathcal{G}_i := \{K\}$ for $K \subset \Omega := \{1, 2, \ldots, N\}$ and $M_i := \bigcup_{K \subset \mathcal{G}_i, |K|=k} K$ with $|M_i| = m \geq k$.
- Define $G := \bigcup_{i=1}^{u} \mathcal{G}_i$ such that $|G| \geq \binom{N}{k}$, to have a covering.

Lemma

Set $r = \binom{N}{k}\binom{m}{k}^{-1}$ and draw $u := rN$ M_i sets uniformly at random from the $\binom{N}{m}$ possible M_i sets. With G defined as above,

$$\mathbf{P}\left[|G| < \binom{N}{k}\right] < C(k/N)N^{-1/2}e^{-N(1-\ln 2)}, \text{ where } C(p) \leq \frac{5}{4}(2\pi p(1-p))^{(-1/2)}$$

Corollary

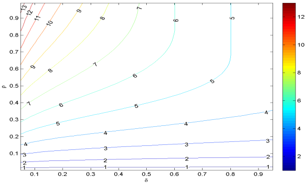
Given the above lemma, as $n \to \infty$ in the proportional-growth asymptotics, the probability that all the $\binom{N}{k}$ $K \subset \Omega$ are covered by G converges to one exponentially in n.

Proof.

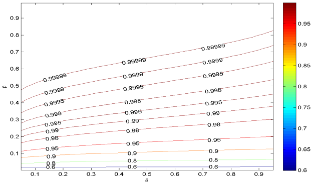
- Groups with $m \geq k$ distinct elements, contains $\binom{m}{k}$ $K \subset \Omega$
  $\Rightarrow$ at least $\binom{N}{k}\binom{m}{k}^{-1} =: r$ groups to cover each K
- For any random group, M_i & K, $\mathbf{P}\left(M_i \supset K\right) = 1/r$ and
$$\mathbf{P}\left(G \not\supset K\right) = (1 - 1/r)^u \leq \exp\left(-u/r\right)$$
- A union bound over $\binom{N}{k}$ K, yields $\quad \mathbf{P}\left[|G| < \binom{N}{k}\right] < \binom{N}{k} e^{-u/r}$
- The RHS of Sterling's Inequality gives
$$\binom{N}{pN} \leq \frac{5}{4}(2\pi p(1-p)N)^{(-1/2)} e^{NH(p)}, \quad H(p) \leq \ln 2 \; \textit{for} \; p \in [0, 1]$$
- Choosing $u = rN$ completes proof of lemma.
- Letting $n \to \infty$ proves the corollary.

$\square$

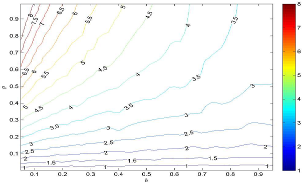
$U(\delta,\rho)$



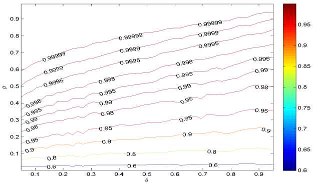
$L(\delta,\rho)$



- Algorithms for calculating lower bounds of $L(k,n,N;A)$ & $U(k,n,N;A)$ by Dossal et. al. and Journée et. al. respectively

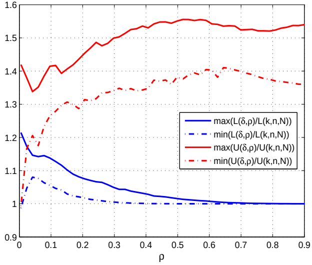


$U(k,n,N;A)$

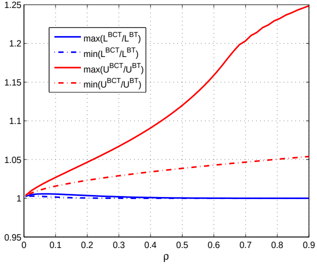


$L(k,n,N;A)$

Sharpness ratios



Improvement ratios

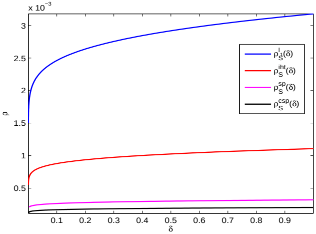


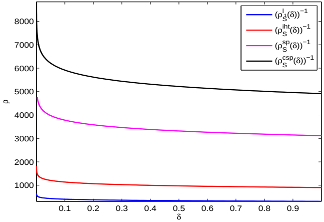
RIC bounds	C.T.	B.C.T.	B.T.
Factor of empirical data	2.74	1.83	1.57

- Factor 1.57 decreases to about 1.05 for $\rho < \frac{1}{100}$, where the CS results are applicable
- BCT suffers from excessive overestimation when $\delta\rho \approx 1/2$

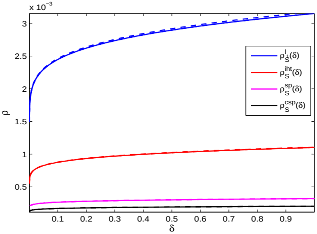
Phase Transitions

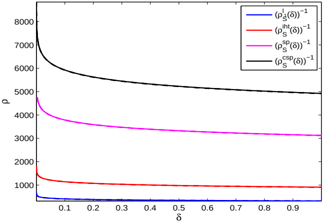
Inverse Phase Transitions





- Phase Transitions based on BCT bounds [B.C.T.T., 2011]





- Small improvement on phase transitions, $\approx 0.5 - 1\%$ higher

Minimum measurements comparison for the algorithms				
Bounds	ℓ_1	IHT	SP	CoSaMP
Blanchard et. al.	$317k$	$907k$	$3124k$	$4923k$
Bah & Tanner	$314k$	$902k$	$3116k$	$4913k$

- Conditions giving the phase transitions are driven by $\mathcal{L}(\delta,\rho)$; precisely depending on $1 - \mathcal{L}(\delta,\rho)$
- Our improvement has been greater in $\mathcal{U}(\delta,\rho)$ than in $\mathcal{L}(\delta,\rho)$
- Even with improvements in $\mathcal{U}(\delta,\rho)$ tightening from a max factor of 1.57 to about 1.05 in this regime of δ and ρ

- Our bounds valid for finite (k, n, N), satisfying specified probs.
- Probabilities extremely small, even for small (k, n, N) and ϵ

Bounds on $\mathbf{P}\left(L(k,n,N) > \mathcal{L}^{BT}(\delta_n,\rho_n) + \epsilon\right)$				
k	n	N	ϵ	*Prob*
100	200	2000	10^{-3}	2.9×10^{-2}
200	400	4000	10^{-3}	9.5×10^{-3}
400	800	8000	10^{-3}	2.9×10^{-3}

Bounds on $\mathbf{P}\left(U(k,n,N) > \mathcal{U}^{BT}(\delta_n,\rho_n) + \epsilon\right)$				
k	n	N	ϵ	*Prob*
100	200	2000	10^{-5}	2.8×10^{-18}
200	400	4000	10^{-5}	9.1×10^{-32}
400	800	8000	10^{-5}	2.8×10^{-58}

Theorem (Fixed δ and $\rho \to 0$)

Let $\widetilde{\mathcal{U}}(\delta,\rho)$ and $\widetilde{\mathcal{L}}(\delta,\rho)$ be the approximations of $\mathcal{U}(\delta,\rho)$ and $\mathcal{L}(\delta,\rho)$ respectively. For a fixed δ as $\rho \to 0$,

$$\widetilde{\mathcal{L}}(\delta,\rho) = \widetilde{\mathcal{U}}(\delta,\rho) = \sqrt{2\rho \log\left(\delta^{-2}\rho^{-3}\right) + 6\rho}$$

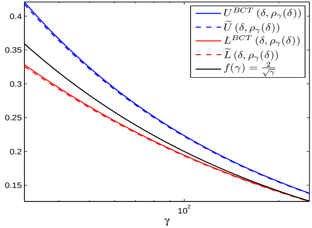
Theorem (ρ as a function of δ)

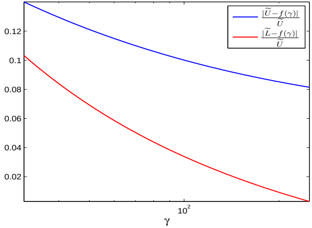
Let $\rho_\gamma(\delta) = \frac{1}{\gamma \log(\delta^{-1})}$ and let the approximations of $\mathcal{U}(\delta,\rho)$ & $\mathcal{L}(\delta,\rho)$ be $\widetilde{\mathcal{U}}\left(\delta,\rho_\gamma(\delta)\right)$ & $\widetilde{\mathcal{L}}\left(\delta,\rho_\gamma(\delta)\right)$ respectively. For a fixed γ as $\delta \to 0$,

$$\begin{aligned}
\widetilde{\mathcal{U}}\left(\delta,\rho_\gamma(\delta)\right) &= \sqrt{2\rho \log\left(\delta^{-2}\rho^{-3}\right)} + \frac{2}{3}\rho \log\left(\delta^{-2}\rho^{-3}\right) \\
\widetilde{\mathcal{L}}\left(\delta,\rho_\gamma(\delta)\right) &= \sqrt{2\rho \log\left(\delta^{-2}\rho^{-3}\right)} - \frac{2}{3}\rho \log\left(\delta^{-2}\rho^{-3}\right)
\end{aligned}$$

Corollary (From Theorem of ρ as a function of δ)

Let $\rho_\gamma(\delta) = \frac{1}{\gamma \log(\delta^{-1})}$ *and let* $\widetilde{\mathcal{U}}\left(\delta, \rho_\gamma(\delta)\right)$ *and* $\widetilde{\mathcal{L}}\left(\delta, \rho_\gamma(\delta)\right)$ *be the approximations of* $\mathcal{U}(\delta,\rho)$ *and* $\mathcal{L}(\delta,\rho)$ *respectively. In the limit,* $\delta \to 0$ *and* $\gamma \to \infty$*, both* $\widetilde{\mathcal{U}}\left(\delta, \rho_\gamma(\delta)\right)$ *and* $\widetilde{\mathcal{L}}\left(\delta, \rho_\gamma(\delta)\right)$ *converge to* $f(\gamma) := \sqrt{2}\gamma^{-1/2}$.





Corollary (Sampling Theorem for ℓ_1, IHT, CoSaMP & SP)

Given a sensing matrix, A, of size $n \times N$ whose entries are drawn i.i.d. from $\mathcal{N}(0, 1/n)$, in the limit as $n/N \to 0$ the sufficient number of measurements for CS algorithms is $n \geq \gamma k \log(N/n)$, with

- $\gamma = 36$ *for ℓ_1-minimization,*
- $\gamma = 93$ *for Iterative Hard Thresholding (IHT),*
- $\gamma = 272$ *for Subspace Pursuit (SP) and*
- $\gamma = 365$ *for Compressed Sampling Matching Pursuit (CoSaMP).*

- Derivation uses approximations and recovery conditions for greedy algorithms $\mu^{alg}\left(\delta, \rho_\gamma(\delta)\right) = 1$ [B.C.T.T., 2011]
- $\gamma = 2e$ is known to be tight for ℓ_1 [Donoho & Tanner, 2009]

Corollary (Sampling Theorem for OMP)

Given a sensing matrix, A, of size $n \times N$ whose entries are drawn i.i.d. from $\mathcal{N}(0, 1/n)$, in the limit as $k/n \to 0$ the sufficient number of measurements for Orthogonal Matching Pursuit (OMP) is

$$n \geq 16k^2 \log(N/k) + 8k^2 \log(n/k) + 24k^2.$$

- OMP requires $O\left(k^2 \log(N/k)\right)$ measurements to guarantee exact recovery [Davenport & Wakin, 2010]
- Derivation uses approximation for fixed δ with $\rho \to 0$ and recovery condition $\mathcal{U}(\delta, \rho) < \sqrt{k} - \sqrt{k-1}$ [Huang et. al., '10]

Summary:

- The random matrix quantity, RIC is an important tool of analysis in CS and related areas
- Improvement is earlier bounds achieved by grouping of submatrices with significant column overlap
- Bounds clear improvements on prior bounds and are consistent with empirically observed data
- However, improvements on bounds led to very little improvements in phase transitions for CS algorithms
- Finite representation of bounds shows remarkable accuracy
- Asymptotic approximation of bounds lead to sampling theorems consistent with CS literature

Note:

In the spirit of reproducible research, software and web forms that evaluate $\mathcal{L}^{BT}(\delta,\rho)$ and $\mathcal{U}^{BT}(\delta,\rho)$ are publicly available at *http://ecos.maths.ed.ac.uk/ric_bounds.shtml*

References:

- B. Bah and J. Tanner. Improved Bounds on Restricted Isometry Constants for Gaussian Matrices. *SIAM J. on Matrix Analysis,* Vol. 31(5) (2010) 2882-2898.
- Paper on asymptotic approximations in preparation.
- J. D. Blanchard, C. Cartis, and J. Tanner. Compressed sensing: How sharp is the RIP? *SIAM Review,* Vol. 53(1) (2011) 105-125.
- E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):42034215, 2005.

THANK YOU