The Restricted Isometry Property and ℓ^p sparse recovery failure

Mike Davies
Institute for Digital Communications &
Joint Research Institute for Signal and Image Processing
Edinburgh University, King's Buildings,
Mayfield Road, Edinburgh EH9 3JL, UK
email: mike.davies@ed.ac.uk

Rémi Gribonval
METISS project-team, INRIA Rennes,
Campus de Beaulieu, F-35042
Rennes Cedex, France
email: remi.gribonval@irisa.fr

***Abstract*—This paper considers conditions based on the restricted isometry constant (RIC) under which the solution of an underdetermined linear system with minimal ℓ^p norm, $0 < p \leq 1$, is guaranteed to be also the sparsest one. Specifically matrices are identified that have RIC, δ_{2m}, arbitrarily close to $1/\sqrt{2} \approx 0.707$ where sparse recovery with $p = 1$ fails for at least one m-sparse vector. This indicates that there is limited room for improvement over the best known positive results of Foucart and Lai, which guarantee that ℓ^1-minimisation recovers all m-sparse vectors for any matrix with $\delta_{2m} < 2(3 - \sqrt{2})/7 \approx 0.4531$.**

We also present results that show, compared to ℓ^1 minimisation, ℓ^p minimisation recovery failure is only slightly delayed in terms of the RIC values. Furthermore when ℓ^p optimisation is attempted using an iterative reweighted ℓ^1 scheme, failure can still occur for δ_{2m} arbitrarily close to $1/\sqrt{2}$.

I. Introduction and state of the art

This paper considers conditions under which the solution $\hat{\mathbf{y}}$ of minimal ℓ^p norm, $0 < p \leq 1$, of an underdetermined linear system $\mathbf{x} = \mathbf{\Phi y}$ is guaranteed to be also the sparsest one. This is a central problem in sparse overcomplete signal representations, where $\mathbf{x}$ is a vector representing some signal or image, $\mathbf{\Phi}$ is an overcomplete signal dictionary, and $\mathbf{y}$ is a sparse representation of the signal. This problem is also at the core of compressed sensing, where $\mathbf{\Phi}$ is called a sensing matrix and $\mathbf{x}$ is a collection of M linear measurements of some ideally sparse data $\mathbf{y}$.

Given a vector $\mathbf{x} \in \mathbb{R}^M$ and a matrix

$$\mathbf{\Phi} \in \mathbb{R}^{M \times N}$$

with $M < N$, we are interested in sparse solutions to $\mathbf{x} = \mathbf{\Phi y}$. We will denote by $\|\mathbf{y}\|_p$ the ℓ^p sparsity measure defined as:

$$\|\mathbf{y}\|_p := \Big(\sum_{j=1}^{N} |y_j|^p\Big)^{1/p}$$

where $0 < p \leq 1$. When $p = 0$, $\|\mathbf{y}\|_0$ denotes the ℓ^0 pseudo-norm that counts the number of non-zero elements of $\mathbf{y}$. The coefficient vector $\mathbf{y}$ is said to be m-sparse if $\|\mathbf{y}\|_0 \leq m$. We will use $\mathcal{N}(\mathbf{\Phi})$ for the null space of $\mathbf{\Phi}$. We will also make use of the subscript notation $\mathbf{y}_\Omega$ to denote a vector that is equal to some $\mathbf{y}$ on the index set Ω and zero everywhere else. Denoting $|\Omega|$ the cardinality of Ω, the vector $\mathbf{y}_\Omega$ is $|\Omega|$-sparse and we will say that the support of the vector $\mathbf{y}$ lies within Ω whenever $\mathbf{y}_\Omega = \mathbf{y}$. For matrices the subscript notation $\mathbf{\Phi}_\Omega$ will denote a submatrix composed of the columns of $\mathbf{\Phi}$ that are indexed in the set Ω.

A. Known conditions for ℓ^p sparse recovery

It has been shown in [15] that if:

$$\|\mathbf{z}_\Omega\|_p < \|\mathbf{z}_{\Omega^c}\|_p \tag{1}$$

holds for all nonzero $\mathbf{z} \in \mathcal{N}(\mathbf{\Phi})$ then any vector $\mathbf{y}^\star$ whose support lies within Ω, can be recovered as *the unique* solution of the following optimisation problem (which is non-convex for $0 \leq p < 1$):

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmin}} \|\mathbf{y}\|_p \text{ s.t. } \mathbf{\Phi y} = \mathbf{\Phi y}^\star. \tag{2}$$

Hence if (1) holds for all $\mathbf{z} \in \mathcal{N}(\mathbf{\Phi})$ and all index sets Ω of size m, then any m-sparse vector $\mathbf{y}^\star$ is recovered as the unique minimiser of (2). Furthermore this condition is tight [16], [17]

Using (2), particularly when $p = 1$, has become a popular mean of solving for sparse representations. An important characteristic of a dictionary that guarantees recovery in this case is the Restricted Isometry Property (RIP). For a matrix $\mathbf{\Phi}$ the restricted isometry constant, δ_k, is defined as the smallest number such that:

$$(1 - \delta_k) \leq \frac{\|\mathbf{\Phi y}_\Omega\|_2^2}{\|\mathbf{y}_\Omega\|_2^2} \leq (1 + \delta_k) \tag{3}$$

for every vector $\mathbf{y}$ and every index set Ω with $|\Omega| \leq k$. As the upper and lower bounds play fundamentally different roles we will further consider the matrix to have been appropriately rescaled so that both the upper and the lower bound are tight.

The RIC's importance can be linked with the following results:

1) Every m-sparse representation is unique if and only if [10]

$$\delta_{2m} < 1 \tag{4}$$

for an appropriately re-scaled dictionary. Furthermore almost every dictionary $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ with $M \geq 2m$ satisfies this condition (again with appropriate re-scaling).

Foucart and Lai [14] have also shown that for a given dictionary with $\delta_{2m+2} < 1$ there exists a sufficiently small p for which solving (2) is guaranteed to recover any m-sparse vector.

2) If

$$\delta_{2m} < 2(3-\sqrt{2})/7 \approx 0.4531 \tag{5}$$

then every m-sparse representation can be exactly recovered using linear programming to solve (2) with $p = 1$, [14]. Furthermore most dictionaries $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$ (sampled from an appropriate probability model) will have an RIC $\delta_{2m} < \delta$ as long as: $M \geq C\delta^{-1} m \log(N/m)$, where C is some constant [1].

The RIC also bounds the condition number, κ, of submatrices, $\mathbf{\Phi}_\Omega$, of a dictionary,

$$\kappa(\mathbf{\Phi}_\Omega) \leq \sqrt{\frac{1+\delta_k}{1-\delta_k}}, \qquad |\Omega| \leq k \tag{6}$$

(indeed, Foucart and Lai [14] formulated their results in asymmetric bounds $\alpha_k \leq \|\mathbf{\Phi}\mathbf{y}_\Omega\|_2^2/\|\mathbf{y}_\Omega\|_2^2 \leq \beta_k$ that provide a sharper bound on the maximal submatrix condition number, with less re-scaling issues). This in turn bounds the Lipschitz constant of the inverse mapping resulting from solving the optimisation problem (2). In this regard the RIC also plays an important role in the noisy recovery problems [4], [14]: $\mathbf{x} = \mathbf{\Phi}\mathbf{y} + \epsilon$ where ϵ is an unknown but bounded noise term.

Note that when (5) holds all the $2m$-submatrices have condition number $\kappa(\mathbf{\Phi}_\Omega) \leq 1.7$ when $|\Omega| \leq 2m$, so they are extremely well behaved. In contrast, $\delta_{2m} < 1$ imposes the finiteness of the condition number of the submatrices as the only constraint.

B. New results

It is an open question as to how much better we could expect to do, i.e. how large can we set $\delta \leq 1$ while still guaranteeing ℓ^1 recovery of any m-sparse vector for any dictionary with $\delta_{2m} < \delta$? This question is partially addressed by the following result:

Theorem 1: For any $\epsilon > 0$ there exists an integer m and a dictionary $\mathbf{\Phi}$ with a restricted isometry constant $\delta_{2m} \leq 1/\sqrt{2}+\epsilon$ for which ℓ^1 recovery fails on some m-sparse vector.

This and the other results presented in this paper are derived in full in [7]. The main idea of the proof is to first reduce the search for a failing dictionary to so-called *minimally redundant* (i.e., $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$ with $M = N-1$) unit spectral norm dictionaries (i.e. $|\!|\!|\mathbf{\Phi}|\!|\!| = 1$). Such dictionaries have one-dimensional kernels which simplifies the calculation considerably (they are of course of little practical interest for compressed sensing). Moreover their RIC, δ_k, is related to the smallest singular value of any k-column submatrix, denoted $\sigma_k(\mathbf{\Phi})$ by:

$$\delta_k(\mathbf{\Phi}) \leq \frac{1-\sigma_k^2(\mathbf{\Phi})}{1+\sigma_k^2(\mathbf{\Phi})} \tag{7}$$

Through (7) Theorem 1 can be seen as a special case of the following theorem concerning general ℓ^p-minimization for $p \leq 1$:

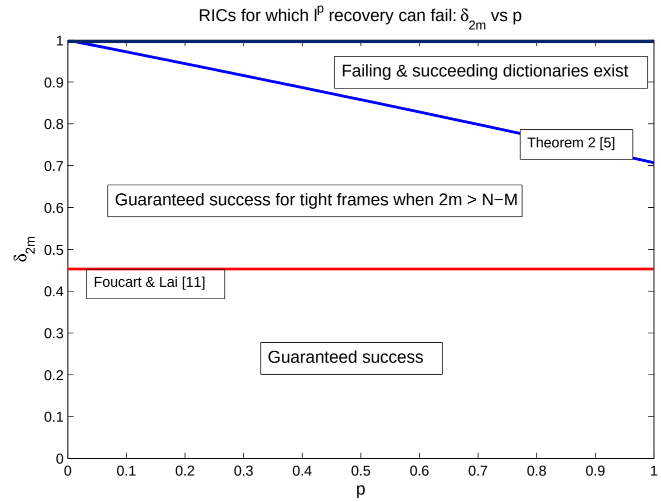

Fig. 1. A summary of known results from [14] and [7] relating the restricted isometry constant to ℓ^p recovery.

Theorem 2: Consider $0 < p \leq 1$ and let $0 < \eta_p < 1$ be the unique positive solution to $\eta_p^{2/p} + 1 = \frac{2}{p}(1-\eta_p)$.

- If $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$ is a *unit spectral norm* dictionary and $2m \leq M < N$ and

$$\sigma_{2m}^2(\mathbf{\Phi}) > 1 - \frac{2}{2-p}\eta_p \tag{8}$$

then all m-sparse vectors can be uniquely recovered by solving (2).

- For every $\epsilon > 0$, there exist integers $m \geq 1, N \geq 2m+1$ and a minimally redundant row orthonormal dictionary $\mathbf{\Phi} \in \mathbb{R}^{(N-1)\times N}$ with:

$$\sigma_{2m}^2(\mathbf{\Phi}) \geq 1 - \frac{2}{2-p}\eta_p - \epsilon \tag{9}$$

for which there exists an m-sparse vector which cannot be uniquely recovered by solving (2).

Whenever η_p is irrational the inequality in (8) can be replaced with $\geq$. Whenever η_p is rational, equality can be achieved with $\epsilon = 0$ in (9).

This identifies values of σ_{2m}, and hence δ_{2m}, for which ℓ^p recovery can fail. The complete results for RIC recovery conditions from [7] along with the result of [14] are summarised in Figure 1. Although there is a gap between the positive result of Foucart and Lai [14] for $p = 1$ and the negative result presented here, it is not a large one. The plot suggests that there might still be some benefit in using $p \ll 1$ to improve sparse recovery. However, at this juncture it is important to consider the practical aspects of ℓ^p minimization. As noted earlier, (2) with $p < 1$ is a non-convex optimisation problem. Furthermore it is easy to show that there can be many local minima associated with incorect sparse recovery. Thus even the potential improvements in using $p \ll 1$ suggested by Theorem 2 may not be realizable in practice.

This is an often missed point: an optimization algorithm is not necessarily equivalent to the associated optimization problem (an exception here is when the cost function is guaranteed to have a unique minimum, as in the ℓ^1 case). We therefore need to consider *algorithm specific* recovery results. We do this next.

II. Reweighted ℓ^1 implementations for ℓ^p optimisation

One approach to solving (2) for $p < 1$ is to attempt to solve a sequence of reweighted ℓ^1 optimisation problems of the form [12], [13], [5], [14]:

$$\hat{\mathbf{y}}^{(n)} = \operatorname*{argmin}_{\mathbf{y}} \|\mathbf{W}_n \mathbf{y}\|_1 \text{ s.t. } \mathbf{\Phi y} = \mathbf{\Phi y}^\star, \tag{10}$$

where the initial weight matrix is set to the identity, $\mathbf{W}_1 = \mathbf{Id}$, and then subsequently $\mathbf{W}_n$ is selected as a diagonal positive definite weight matrix that is a function of the previous solution vector, $\mathbf{W}_n = f_n(\mathbf{y}^{(n-1)})$. In order to consider the widest possible set of reweighting schemes we define the following reweighting that encompasses *all* 'reasonable' reweighting schemes.

Definition 1 (Admissible reweighting schemes): A reweighting scheme is considered to be admissible if, $\mathbf{W}_1 = \mathbf{Id}$ and if, for each n, there exists a $w^n_{\max} < \infty$ such that for all k, $0 \le W_n(k,k) \le w^n_{\max}$ and $W_n(k,k) = w^n_{\max} \Leftarrow \hat{y}_k^{(n-1)} = 0$.

While it can be shown [7] that such reweighting strategies cannot damage an already successful solution we also have the following negative result.

Proposition 1 (Iteratively reweighted ℓ^1 performance): Let $\mathbf{\Phi} \in \mathbb{R}^{(N-1)\times N}$ be a minimally redundant dictionary of maximal rank $N-1$. Let Ω be a support set for which ℓ^1 recovery fails. Then *any* iteratively reweighted ℓ^1 algorithm with an admissible reweighting scheme will also fail for some vector $\mathbf{y}$ with support Ω.

Proof: Let $\mathbf{\Phi} \in \mathbb{R}^{(N-1)\times N}$ be a minimally redundant dictionary with maximal rank and let $\mathbf{z} \in \mathcal{N}(\mathbf{\Phi})$ be an arbitrary generator of its null space. Consider any set Ω for which ℓ^1 recovery can fail, i.e., $\|\mathbf{z}_\Omega\|_1 \ge \|\mathbf{z}_{\Omega^c}\|_1$. Let $\mathbf{y}^\star = \mathbf{z}_\Omega$. Because of the dimensionality of the null space, any representation satisfying $\mathbf{\Phi y} = \mathbf{\Phi y}^\star$ takes the form $\mathbf{y} = \mathbf{z}_\Omega - \alpha \mathbf{z} = (1-\alpha)\mathbf{z}_\Omega - \alpha \mathbf{z}_{\Omega^c}$. For any weight

$$\|\mathbf{W}_n \mathbf{y}\|_1 = |1-\alpha| \cdot \|\mathbf{W}_n \mathbf{z}_\Omega\|_1 + |\alpha| \cdot \|\mathbf{W}_n \mathbf{z}_{\Omega^c}\|_1, \alpha \in \mathbb{R}$$

hence there are only two possible unique solutions to (10), corresponding to $\alpha = 0$ and $\alpha = 1$. Since ℓ^1 fails to recover $\mathbf{y}^\star$, we have $\hat{\mathbf{y}}^{(1)} = -\mathbf{z}_{\Omega^c}$, therefore $\Omega \subset \Gamma_1$ and $\mathbf{W}_2(k,k) = w^2_{\max}$, $k \in \Omega$. [1] It follows that

$$\begin{aligned} |\langle \mathbf{W}_2 \mathbf{z}, \operatorname{sign}(\hat{\mathbf{y}}^{(1)})\rangle| &\le w^2_{\max} \|\mathbf{z}_{\Omega^c}\|_1 \\ &\le w^2_{\max} \|\mathbf{z}_T\|_1 \\ &\le \|(\mathbf{W}_2 \mathbf{z})_{\Gamma_1}\|_1 \end{aligned}$$

and we obtain that $\hat{\mathbf{y}}^{(n)} = -\mathbf{z}_{\Omega^c}$ for all n. ■

[1] If the solution to (10) is not unique then all values of α between 0 and 1 result in valid solutions and the algorithm has no means for determining the correct one. We therefore make the pessimistic assumption that the algorithm will select the incorrect representation associated with $\alpha = 1$.

This immediately leads to the following statement:

Theorem 3: For any $\epsilon > 0$ there exists an integer m and a dictionary $\mathbf{\Phi}$ with a restricted isometry constant $\delta_{2m} \le 1/\sqrt{2} + \epsilon$ for which recovery using any iteratively reweighted ℓ^1 algorithm fails on some m-sparse vector.

This is a somewhat surprising result since one would suspect that the adaptivity built into the choice of the weight should enable improved performance. However it does not necessarily imply that the uniform performance of iterative reweighted ℓ^1 techniques is no better than ℓ^1 alone. Instead the result highlights the danger of characterising sparse recovery uniformly only in terms of the RIP.

III. RIP rest in peace?

RIP recovery conditions, be they for ℓ^1 or ℓ^p, come from a worst case analysis with respect to several parameters: worst case over all coefficients for a given sign pattern; worst case over all sign patterns for a given support; worst case over all supports of a given size; and worst case over all dictionaries with a given RIC. Our results emphasize the pessimism of such a worst case analysis. In the context of compressed sensing [9], [3], there is also the desire to characterise the degree of undersampling (M/N) possible while still achieving exact recovery. Here RIP has been used to show that certain random matrices with high probability guarantee exact recovery with an undersampling of the order $(m/N)\log(N/m)$. However this result is indirect, firstly due to the worst case analysis discussed above and then secondly through the application of the concentration of measure [1]. A more direct approach seems to provide a much clearer indication of the relationship between undersampling and recovery [11]. Of course, deriving expressions for such phase transitions when $p \neq 1$ is likely to be a very challenging problem. Reducing $p < 1$ also introduces other issues since the cost function is no longer convex. Thus recovery results should necessarily be derived in an *algorithm dependent* manner.

Using RIP to characterize matrices has been extremely useful in understanding when ℓ^1 sparse recovery can be achieved and indeed opened the door to the possibility of compressed sensing. However to further explore the possibilities of better recovery using ℓ^p minimisation techniques we require new more refined tools. An obvious contender for sharper techniques is to work directly with the null space property, (1) and its relative the *robust null space property* [8], particularly as these are both in a sense tight. For example an interesting non-RIP based analysis of ℓ^1 sparse recovery is presented in [18]. Like RIP analysis, this also presents a worst case analysis but based around typical properties of the null spaces of random matrices. Unlike RIP, such an analysis is naturally invariant to the observation metric and so is more in line with the minimum ℓ^p estimators which are what in statistics are called *equivariant estimators*[2].

[2] an equivariant estimator is one that is invariant to specific group actions - in the case of compressed sensing this is the set of linear invertible transformations of the observation space

Finally we note that sparse recovery in the presence of noise and/or quantization errors, $\mathbf{x} = \Phi\mathbf{y} + \epsilon$, is a different type of problem, and one that is not metric invariant. As the RIP clearly controls the forward and inverse Lipschitz constants of the embedding of the m-sparse set [2] it will still play an important role in the performance of sparse approximation in the presence of noise. So perhaps we should not be ready to bury RIP just yet.

Acknowledgement

This work was supported in part by the European Union through the project SMALL (Sparse Models, Algorithms and Learning for Large-Scale data). The project SMALL acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 225913.

MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices.” *To appear in Constructive Approximation*, 2008.

[2] T. Blumensath and M. E. Davies,”Sampling theorems for signals from the union of linear subspaces.”*Awaiting Publication, IEEE Transactions on Information Theory*, 2009.

[3] E. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.” *IEEE Trans. Info. Theory*, vol. 52, pp. 489–509, Feb 2006.

[4] E. Candès, “The Restricted Isometry Property and its implications for Compressed Sensing.” *Comptes Rendus de l’Académie des Sciences, Paris, Série I*, 346, 589–592, 2008.

[5] E. J. Candés, M. B. Wakin and S. P. Boyd, “Enhancing sparsity by reweighted ell-1 minimization”, to appear in J. Fourier Anal. Appl., 2008.

[6] S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci Comp.*, vol. 20, no. 1, pp. 33-61, 1999.

[7] M. E. Davies and R. Gribonval, “Restricted Isometry Constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$,” To appear in IEEE Trans. Inf. Theory, 2009.

[8] M. E. Davies and R. Gribonval, “ℓ^p minimization and sparse approximation failure for compressible signals,” *Submitted to SAMPTA, 2009*.

[9] D. Donoho,“Compressed Sensing.” *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp 1289-1306, April, 2006.

[10] D. L. Donoho and M. Elad, “Optimally sparse representation from overcomplete dictionaries via ℓ^1 norm minimization,” *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197-2002, Mar. 2002.

[11] D. Donoho and J. Tanner, “Counting faces of randomly-projected polytopes when the projection radically lowers dimension”, *to appear in Journal of the AMS*, 2008.

[12] M. Figueiredo and R. Nowak, “A bound optimization approach to wavelet-based image deconvolution”, *In Proc. IEEE International Conference on Image Processing — ICIP’2005*, 2005.

[13] M. Figueiredo, J. Bioucas-Dias and R. Nowak, “Majorization-Minimization Algorithms for Wavelet-Based Image Restoration”, *IEEE Trans. Image Processing*, vol. 16, no. 12, pp. 2980–2991, Dec. 2007.

[14] S. Foucart and M.-J. Lai,“Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$.” *Submitted to Applied and Computational Harmonic Analysis*, 2008.

[15] R. Gribonval and M. Nielsen, “Sparse decompositions in unions of bases.” *IEEE Trans. Info. Theory*, vol. 49, no. 12, pp 3320-3325, Dec 2003.

[16] R. Gribonval and M. Nielsen, “On the strong uniqueness of highly sparse expansions from redundant dictionaries.” *In Proc. Int Conf. Independent Component Analysis (ICA’04)*, Sep 2004.

[17] R. Gribonval and M. Nielsen, “Highly sparse representations from dictionaries are unique and independent of the sparseness measure.” *Applied and Computational Harmonic Analysis*, vol. 22, no. 3, pp 335–355, May 2007. [Technical report October 2003].

[18] Y. Zhang, “On the Theory of Compressive Sensing via ℓ^1-minimization: simple derivations and extensions,” *CAAM Technical report TR08-11, Rice University*, Sept. 2008.