Statistical RIP sampling matrices and robust recovery

Arya Mazumdar[*]   Alexander Barg[*,§]

I. Introduction: RIP and sparse recovery

Compressive sampling is a technique of recovering sparse N-dimensional signals from low-dimensional projections, i.e., their linear images in $\mathbb{R}^m, m \ll N$. In formal terms the problem can be stated as follows. Let $\Phi : \mathbb{R}^m \to \mathbb{R}^N, m \ll N$ be a linear operator used to create a "sketch" of a signal represented by a real vector $\boldsymbol{x} \in \mathbb{R}^N$. In other words, we observe a compressed version of the signal, i.e., a vector $\boldsymbol{r} = \Phi\boldsymbol{x}$, where Φ is an $m \times N$ sampling matrix. Recovering $\boldsymbol{x}$ from $\boldsymbol{r}$ is generally impossible because the system of equations is under-determined, and the solutions form an affine subspace in $\mathbb{R}^N$. The problem becomes tractable if we seek an approximation of $\boldsymbol{x}$ by a vector $\hat{\boldsymbol{x}}$ that satisfies

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_{p_1} \le C \min_{\boldsymbol{x}' \text{ is } k\text{-sparse}} \|\boldsymbol{x} - \boldsymbol{x}'\|_{p_2} \qquad (1)$$

for some $p_1, p_2 \ge 1$, where a vector is called k-sparse if it contains k or fewer nonzero coordinates. Note that if $\boldsymbol{x}$ itself is k-sparse, then (1) implies that $\hat{\boldsymbol{x}} = \boldsymbol{x}$. Moreover it implies that the recovery is stable: if $\boldsymbol{x}$ is approximately k-sparse ($\boldsymbol{x}$ contains only k "significant" entries) then the recovery error is small.

A useful tool for the construction of sampling matrices is provided in the works of Candès et al. [6], [7] who showed that recovery is possible if the matrix Φ has the Restricted Isometry Property (RIP), i.e., it acts as near-isometry on all k-sparse vectors. Let $I \subseteq [N] := \{1, \dots, N\}$. Denote by $\Phi_I \in \mathbb{R}^{m\times|I|}$ the matrix formed by the columns of Φ with indices in I.

Definition 1 (The Restricted Isometry Property (RIP)): A matrix $\Phi \in \mathbb{R}^{m\times N}$ is said to satisfy the *restricted isometry property* (k, δ), or (k, δ)-RIP, $k \le m$, $0 \le \delta \le 1$, if for all $I \subset [N]$ such that $|I| = k$ and for all $\boldsymbol{u} \in \mathbb{R}^k$,

$$(1 - \delta)\|\boldsymbol{u}\|_2^2 \le \|\Phi_I \boldsymbol{u}\|_2^2 \le (1 + \delta)\|\boldsymbol{u}\|_2^2. \qquad (2)$$

This property is known to hold if the columns of Φ are near-orthogonal, i.e., $|\phi_i^T \phi_j| \le \alpha$ for all $i \ne j$ and some α. Such collections of vectors are also known as *incoherent dictionaries* (e.g., [18]).

As proved by Candès et al. [6], [8], if the sampling matrix Φ has the $(2k, \sqrt{2} - 1)$-RIP property, the signal recovered as the solution to the linear programming problem

$$\max_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \Phi\boldsymbol{x} = \boldsymbol{r} \qquad (3)$$

satisfies the $(p_1 = 2, p_2 = 1)$ error guarantee (1). Moreover, the recovery is robust in the following sense: Even in the case of noisy observations $\boldsymbol{r} = \Phi\boldsymbol{x} + \boldsymbol{z}$, where $\boldsymbol{z}$ is an unknown noise vector, the solution to

$$\max_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \|\Phi\boldsymbol{x} - \boldsymbol{r}\|_{\ell_2} \le \epsilon \qquad (4)$$

provides a recovery guarantee of the form (1) with a further additive error term that is proportional to ϵ.

In this formulation, the study of the compressed sensing problem has been focused on the design of good sampling matrices Φ in conjunction with low-complexity recovery algorithms that provide an error guarantee of the form (1) based on as few samples as possible. It is known that random Gaussian matrices and random Bernoulli matrices provide a $(p_1 = 2, p_2 = 1)$ error guarantee with $m = O(k \log(N/k))$ sketch length under the linear programming recovery algorithm [6]. It is also known that at least $m = \Omega(k \log(N/k))$ samples are required for any recovery algorithm with an error guarantee of the form (1) (see, for example, [12], [1]).

If we restrict matrices Φ to have entries ±1, then the RIP property can be guaranteed by relying on binary error-correcting codes and utilizing information about their distance distribution. This link has been used in a number of publications on explicit constructions of sampling matrices [2], [3], [17] At the same time, classic bounds on codes [13] preclude these constructions from attaining the optimal sketch length.

II. Statistical RIP of Sampling Matrices

There is another aspect of compressed sensing where randomization is built into the signal and recovery model [4], [5], [11], [18]. In this paper we analyze sparse recovery under the assumption that Φ acts as near-isometry for almost all rather than all sparse vectors. Analyzing the recovery properties under this relaxation is not immediate; however, a number of useful ideas in this direction have been suggested in earlier works [4], [11], [18]. The two properties desired from a sampling matrix that had been put forward by these works are the Statistical RIP (SRIP) and Statistical Unique Recovery Property (SURP).

This work was supported by NSF grants CCF0916919, CCF0830699, and DMS0807411.

[*]Department of ECE and Institute for Systems Research, University of Maryland, College Park, 20742, USA. E-mails: {arya,abarg}@umd.edu.
[§]Dobrushin Mathematical Laboratory, Institute for Problems of Information Transmission, Russian Academy of Science, Moscow, Russia.

A statistical counterpart of the RIP property (SRIP) was introduced in [4], [11]. Roughly speaking, the matrix Φ has a statistical RIP if the near-isometry condition (2) holds for almost all choices of the support I. The definition that we give is slightly stronger than the one in [4] and is close to the definition in [11].

Definition 2 (SRIP): An $m \times N$ sampling matrix Φ is said to satisfy the (k, δ, ϵ)-SRIP property $k \le m$, $0 \le \delta \le 1$, $0 \le \epsilon < 1$ (is (k, δ, ϵ)-SRIP), if (2) holds for at least $1 - \epsilon$ proportion of all subsets $I \subset [N]$ such that $|I| = k$.

The statistical unique recovery property (SURP) is another useful property for sparse signal recovery. Consider a product measure $P_k \times P_{\boldsymbol{z}}$ where P_k is the uniform distribution on the k-subsets of $[N]$ and where $P_{\boldsymbol{z}}$ is some probability measure on $\mathbb{R}^k$. In the following definition $\Pr(\cdot)$ refers to $P_k \times P_{\boldsymbol{z}}$.

Definition 3 (SURP): Let $k \le m$, $0 \le \epsilon < 1$. An $m \times N$ sampling matrix Φ is said to satisfy the (k, ϵ)-SURP if

$$\Pr(\{k\text{-sparse } \boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{y} \ne \boldsymbol{x} : \Phi\boldsymbol{y} = \Phi\boldsymbol{x}\}) < \epsilon.$$

This definition is close to the concepts discussed in [4], [18], but not equivalent to them in terms of the underlying probabilistic model.

The RIP property of Φ with respect to a particular k-subset $I \subset [N]$ allows a vector $\boldsymbol{x}$ supported on I to be recovered from its sketch $\Phi\boldsymbol{x}$ using, for instance, the basis pursuit algorithm of Candès et al. [7], with complexity $O(N^3)$. The SRIP and SURP are used by Calderbank et al. [4] to show recovery guarantee for k-sparse signals under their reconstruction algorithm. These properties are also used in [18] to show that exact recovery of signals under some random models is possible. As shown in [14], some specific sampling matrices with statistical recovery properties support efficient recovery of sparse signals.

III. Lasso estimator

The SURP property is useful if instead of recovering the signal $\hat{x}$ we are interested in finding the locations of its largest coordiantes. In statistics, this task is known as model selection and is handled by procedures such as Lasso. The Lasso estimator seeks a solution to the optimization problem

$$\min_{\hat{\boldsymbol{x}} \in \mathbb{R}^N} \frac{1}{2}\|\boldsymbol{r} - \Phi\hat{\boldsymbol{x}}\|_{\ell_2} + \lambda\sigma\|\hat{\boldsymbol{x}}\|_{\ell_1}$$

where σ^2 is the variance of the coordinates of the random noise $\boldsymbol{z} = r - \Phi\boldsymbol{x}$ (see (4)) and λ is a regularization parameter. In an important contribution, Candès and Plan [9] establish several results about the error guarantee of the Lasso estimate if the columns of Φ form a sufficiently incoherent dictionary. The results in [9] and a number of related papers are established for *generic k-sparse* signals, defined as follows:

1. The support $I \subset [N]$ of the k nonzero coordinates is a uniformly random k-subset of [N];

2. Conditional on I, the signs of the nonzero entries of $\boldsymbol{z}$ are independent and uniformly random.

The following result is proved as Theorem 1.2 in [9].

Theorem 1: Let $\boldsymbol{x}$ be a generic k-sparse signal, where $k \le c_0 N/(\|\Phi\|^2 \log N)$ for some positive constant c_0, where $\|\Phi\|$ is the operator 2-norm. Assume that $\boldsymbol{r} = \Phi\boldsymbol{x} + \boldsymbol{z}$ where the coordinates of $\boldsymbol{z}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ Gaussian random variables. Then the Lasso estimate computed with $\lambda = 2\sqrt{2\log N}$ obeys

$$\|\Phi\boldsymbol{x} - \Phi\hat{\boldsymbol{x}}\|_{\ell_2}^2 \le 16(1 + \sqrt{2})^2 k \log N \sigma^2$$

with probability at least $1 - 6N^{-2\log 2} - N^{-1}(2\pi \log N)^{-1/2}$.

If the signal $\boldsymbol{x}$ is approximately k-sparse, then the error of the Lasso estimator can be bounded above in terms of the minimum mean square error $\min_{I \subset [N]} \|\Phi\boldsymbol{x} - P[I]\Phi\boldsymbol{x}\|_{\ell_2}^2$, where $P[I]$ is the projection on the subspace spanned by the columns $\phi_i, i \in I$. This is the contents of Theorem 1.4 in [9] which proves that the error estimate holds with high probability over the choice of supports of k-sparse linear approximations of $\boldsymbol{x}$. Proofs of these theorems make essential use of the results about norms of random submatrices of a matrix obtained by Tropp [19].

IV. Statistical RIP matrices from Codes

In this section, we address the task of constructing matrices with statistical recovery properties using binary error-correcting codes.

Let $\mathcal{C} \subset \{0,1\}^m$ be a subset of cardinality N. Below we call $\mathcal{C}$ an (m, N) code. To construct an $m \times N$ sampling matrix from it, we use the mapping $0 \to 1/\sqrt{m}, 1 \to -1/\sqrt{m}$.

We rely on the concepts of the distance distribution and dual distance of codes. To remind ourselves, the distance distribution of an (m, N) code $\mathcal{C}$ is the set of numbers $(A_0 = 1, A_1, \ldots, A_m)$ such that

$$A_w = \frac{1}{N}|\{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{C}^2 : d_H(\boldsymbol{x}_1, \boldsymbol{x}_2) = w\}|$$

where d_H is the Hamming distance.

The MacWilliams transform of the distance distribution is the set $(A_0^\perp, A_1^\perp, \ldots, A_m^\perp)$, where for all w

$$A_w^\perp = \frac{1}{N}\sum_{i=0}^m A_i K_i(w),$$

where $K_i(t)$ is a Krawtchouk polynomial of degree i. It is known that $A_0^\perp = 1, A_w^\perp \ge 0$ for all w [15, p. 139]. The number $d^\perp$ such that $A_1^\perp = \cdots = A_{d^\perp - 1}^\perp = 0, A_{d^\perp}^\perp > 0$ is called the dual distance of the code $\mathcal{C}$. If $\mathcal{C}$ is a linear code then (A_w) is the weight distribution of $\mathcal{C}$ and $(A_w^\perp)$ is the weight distribution of the dual code $\mathcal{C}^\perp$.

We are now ready to state one of the main theorems of this paper.

Theorem 2: Let $\mathcal{C}$ be an (m, N) code with $d^\perp(\mathcal{C}) > l$, l even, and let Φ be the sampling matrix constructed from it. Suppose that

$$m \ge \frac{2lk^{2+2/l}}{\delta^2 e \epsilon^{2/l}}.$$

holds for m sufficiently large. Then Φ is (k, δ, ϵ)-SRIP.

We say that the code $\mathcal{C}$ has width $\bar{w}$ if its distance distribution $\{A_w\}$ satisfies $A_w = 0$ for $|w - (m/2)| \ge \bar{w}/2$. If we have control of the width of the code $\mathcal{C}$ then the number of samples m can be made proportional to $\log {}^1\!/\!_\epsilon$ rather than to $\epsilon^{-2/l}$.

Theorem 3: Let k satisfies $k < 2 \ln N \log(k/\epsilon)$ and $m \geq 8(k/\delta^2) \log(k/\epsilon) \ln N$. Suppose that $\mathcal{C}$ is a linear (m, N) code of width $\frac{m\delta}{2\sqrt{2k \log(k/\epsilon)}}$ and $d^\perp(\mathcal{C}) > 2$, and let Φ be the sampling matrix constructed from it. Then Φ is (k, δ, ϵ)-SRIP.

We remark that the matrix Φ of the above theorem can be constructed deterministically with complexity polynomial in N and k [16].

In regards to the SURP property, we prove the following.

Theorem 4: Let $\mathcal{C}$ be an (m, N) code and let $d^\perp(\mathcal{C}) > l$ for some even l. Suppose that for m sufficiently large

$$m \geq \max\Big(\frac{6lk^{2+2/l}}{\delta^2 e \epsilon^{2/l}}, \frac{2kl}{\epsilon^{2/l}(1-\delta)}\Big).$$

Then the sampling matrix Φ constructed from $\mathcal{C}$ is (k, ϵ)-SURP and $(k, \delta, \epsilon/3)$-SRIP.

V. Algorithmic implications

Algorithmic consequences of SURP and SRIP properties for signal recovery and model selection have been studied [4], [11] for special choices of the matrix Φ (such matrices constructed from Delsarte-Goethals codes or from pairs of orthogonal bases in $\mathbb{R}^m$). We study a general relation between statistical properties of the sampling matrix Φ and the probability of signal recovery under Lasso and ℓ_1 minimization, assuming the signal model defined in the end of Sect. 2.

We prove that if Φ is constructed from a binary (m, N) code that forms an orthogonal array of strength l (has dual distance $\geq l + 1$) and satisfies

$$m \geq \max\Big\{\frac{lk^{2+\frac{2}{l}}}{2e\epsilon^{2/l}}, \frac{lk}{2e}\Big(\frac{\sqrt{2}N}{e}\Big)^{2/l} \log\Big(\frac{2N}{e}\Big)\Big\}$$

then the sufficient conditions for Theorem 1 [9] are satisfied with probability $1 - \epsilon$. These conditions are as follows:

1) $\|(\Phi_I^T \Phi_I)^{-1}\| \leq 2$,
2) The vector $\boldsymbol{z}$ obeys $\|\Phi^T \boldsymbol{z}\|_{\ell_\infty} \leq \sqrt{2}\lambda$,
3) The following inequality holds:

$$\begin{aligned}&\|\Phi_{I^c}^T \Phi_I (\Phi_I^T \Phi_I)^{-1} \Phi_I^T \boldsymbol{z}\|_{\ell_\infty}\\ &\quad + 2\lambda \|\Phi_{I^c}^T \Phi_I (\Phi_I^T \Phi_I)^{-1} \Phi_I^T \operatorname{sgn}(\boldsymbol{x}_I)\|_{\ell_\infty} \leq (2 - \sqrt{2})\lambda.\end{aligned}$$

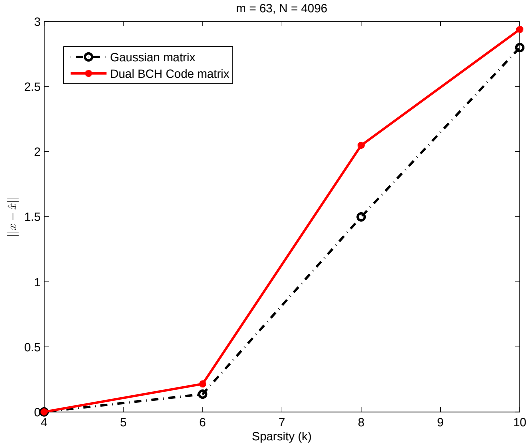
In our context, condition 1) follows from the $(k, {}^1\!/\!_2, \epsilon)$-SRIP property of the matrix Φ. The other two properties are proved by estimating norms of submatrices of Φ relying on properties of the underlying code.

A modification of these calculations also enables us to show that the Lasso estimate is stable, in similarity to the modification of Theorem 1 discussed in the end of Section III.

Similar considerations also enable us to find conditions for recovery with high probability using the linear programming decoder (3).

Conditions 1)-3) are used in [9], [19] where they are shown to hold because of the coherence property of the matrix Φ. (A matrix Φ is said to satisfy the coherence condition if its columns are sufficiently uncorrelated, for instance if $\max_{i \neq j} |(\phi_i, \phi_j)| \leq O(1/\log N)$.) Our contribution is to replace the coherence condition in the error estimates of the recovery algorithms considered with properties of binary codes controlled by their dual distance.

We include a small example with simulated performance of the matrix Φ constructed from the $(63, 4096)$ dual BCH code with $l = 4$ vs random Gaussian Φ. The recovery procedure used to generate it is the linear programming decoder.



References

[1] K. D. Ba, P. Indyk, E. Price, and D. Woodruff, "Lower bounds for sparse recovery," *Proc. 21th ACM-SIAM Sympos. Discrete Algorithms*, Austin, TX, 2010, pp. 1190–1197.

[2] A. Barg and A. Mazumdar, "Small ensembles of sampling matrices constructed from coding theory," *IEEE International Symposium on Information Theory (ISIT)*, Austin, Jul 13–18, 2010, pp. 1963–1967.

[3] J. Bourgain, S. J. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova, "Explicit constructions of RIP matrices and related problems," arXiv:1008.4535.

[4] R. Calderbank, S. Howard, and S. Jafarpour "Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, April, 2010, pp. 358–374.

[5] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *The Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, 2009.

[6] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, 2006, pp. 489–509.

[7] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, 2005, pp. 4203–4215.

[8] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Acad. Sci. Paris*, Ser. I 346, 2008, pp. 589–592.

[9] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Ann. Statist.*, vol. 37, no. 5A, 2009, pp. 2145–2177.

[10] W. Dai, O. Milenkovic and H. V. Pham, "Structured sublinear compressive sensing via dense belief propagation," arXiv:1101.3348.

[11] A. Gurevich and R. Hadani, "The statistical restricted isometry property and the Wigner semicircle distribution of incoherent dictionaries," arXiv:0903.3627.

[12] B. Kashin and V. Temlyakov, "A remark on compressed sensing," *Math. Notes*, vol. 82, no. 5-6, 2007, pp. 748–755.

[13] V. I. Levenshtein, "Bounds on the maximum cardinality of codes with a bounded modulus of the scalar product," Soviet Mathematics Doklady, vol. 263, no. 6, 1982, pp. 1303–1308.

[14] K. Li, L. Gan, and C. Ling, "Orthogonal symmetric Toeplitz matrices for compressed sensing: statistical isometry property," arxiv:1012.5947, 2010.

[15] F. J. MacWilliams and N.J.A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[16] A. Mazumdar and A. Barg, “General constructions of deterministic (S)RIP matrices for compressive sampling,” Proc. 2011 IEEE International Symposium on Information Theory, St. Petersburg, Russia, August 1-5, 2011.

[17] H. Pham, W. Dai, and O. Milenkovic, “Sublinear compressive sensing reconstruction via belief propagation decoding,” *Proc. IEEE International Symposium on Information Theory*, Seoul, South Korea, 2009, pp. 674–678.

[18] J. A. Tropp, “On the conditioning of random subdictionaries,” *Appl. Comput. Harmon. Anal.*, vol. 25, 2008, pp. 1–24.

[19] J. A. Tropp, “Norms of random submatrices and sparse approximation,” *C. R. Acad. Sci. Paris, Ser I*, vol. 346, 2008, pp. 1271-1274.