Fiftieth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 1 - 5, 2012

Learning Sparse Boolean Polynomials

Sahand Negahban and Devavrat Shah
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
{sahandn,devavrat}mit.edu

***Abstract*— We are given a Boolean function $f : \{-1, 1\}^n \mapsto \mathbb{R}$ that can be written as a sparse linear combination of s polynomials. The *Junta* problem cf. [1] is an instance of such a setting. Our goal is to learn the function f by accessing its values at randomly sampled m elements from $\{-1, 1\}^n$. In this paper, we draw connections between the sparse polynomial learning problem and compressed sensing. As a result we provide a convex program that learns an s-sparse polynomial with high probability using $m = \mathcal{O}(s^2 n)$ observations. We contrast this result with the worst case sample-complexity which requires $\mathcal{O}(n2^n)$ random samples to learn the entire function f. Our results naturally extend to the setting where the data is noisy or f is well approximated by an s-sparse polynomial. Our results also show that the solution adapts to the number of observations and finds a natural approximation given the available information.**

I. Introduction.

The problem of learning a Boolean function is a simple learning task that allows us to still model complex interactions between the variables. It is one of the simplest settings of non-parametric regression and captures instances of learning problems where the features are categorical, for example: male or female? In contrast to many models that assume the linearity of the response on the features, we would like to understand models where the value of the Boolean function can depend on intricate Boolean operations between the input features. Concretely, we assume that our function $f \in \mathcal{F} := \{f : \{-1, 1\}^n \mapsto \mathbb{R}\}$ is *s-sparse Boolean function*, that is, it admits the decomposition

$$f(x) = \sum_{i=1}^{s} \alpha_{J_i} \chi_{J_i}(x), \tag{1}$$

where $J_i \subset [n] := \{1, 2, \ldots, n\}$, $\alpha_{J_i} \in \mathbb{R}$, and for any arbitrary $J \subset [n]$ we define $\chi_J \in \mathcal{F}_{\text{bool}}$ as

$$\chi_J(x) = \begin{cases} \Pi_{i \in J} x_i & \text{if } J \neq \emptyset \\ 1 & \text{if } J = \emptyset, \end{cases} \tag{2}$$

where x_i denotes the i^{th} component of x. The collection of functions $\{\chi_J(x)\}$ corresponds to the higher-order polynomials and allows us to model more complex interactions between the input features.

One specific instance of this problem is known as the d-Junta and was considered earlier by Littlestone [2] as well as Blum, Hellerstein and Littlestone [3][1]. In its current form, the problem was introduced by Blum and Langley [1]. The d-Junta problem is to learn an unknown Boolean function $f : \{-1, 1\}^n \mapsto \mathbb{R}$, that depends only on $d < n$ variables from labeled samples $(x, f(x))$, where $x = (x_1, \ldots, x_n)$ are sampled uniformly from $\{-1, 1\}^n$. Thus, there are d unknown indices $1 \leq i_1 < \cdots < i_d \leq n$ and a *hidden* function $g : \{-1, 1\}^d \mapsto \mathbb{R}$ so that for all $x \in \{-1, 1\}^n$

$$f(x_1, \ldots, x_n) = g(x_{i_1}, \ldots, x_{i_d}). \tag{3}$$

If we let $S = \{i_1, i_2, \ldots, i_d\}$ then we shall see in the sequel that f admits the decomposition in equation (1) such that $\alpha_{J_i} = 0$ for all J_i that are not subsets of S. Therefore, in the setting of the d-Junta, the sparsity index $s = 2^d$. In our setting, we are interested in more general sparse polynomial decompositions. Our functions may depend on any number of the input variables, however, we assume that there are few interactions between these variables.

A. Summary of results.

As alluded to above, the aim of this paper is to develop a framework linking sparse polynomial Boolean function learning and compressed sensing. In particular we show that a natural sensing matrix that arises in the Boolean function learning setting satisfies an incoherence type condition [4], [5], [6]. With that, we are then able to leverage some of the existing results in the compressed sensing literature as well as develop some new results in order to demonstrate particular error bounds for recovering the coefficients of the polynomial decomposition of f. We show that with order $s^2 n$ samples we are able to successfully recover an s-sparse Boolean function f. Our result is robust in the sense that it naturally extends to the setting where the observations are noisy. More generally, it extends for noisy instances of approximate s-sparse polynomial functions; that is our results are applicable in the setting where the function is not an exactly s-sparse Boolean function. Our final set of results demonstrate that the algorithm we will present in the sequel is adaptive to the number of observations available in that the method will find a natural s-sparse approximation to the function given the available data.

[1] An informative blog-post by Rick Lipton dated June 4, 2009 titled *The Junta Problem* is worth a read. Our approach utilizes techniques from compressive sensing literature and it is note worthy that RL had mentioned this as a plausible plan of attack in the Junta setting.

978-1-4673-4539-2/12/$31.00 ©2012 IEEE

Few words about the proof technique. The results of this paper, at some level, are simple observations. Specifically, we show that under the uniform sampling model, the induced 'Fourier Matrix' has an appropriate *incoherence* property. We are able to establish this fact by noting that under the uniform sampling model, the columns of such a matrix are pair-wise independent. Using this property along with the techniques developed in the *compressive sensing* literature for approximate sparse recovery [7], we establish our results.

II. Setup and problem statement.

This section describes the necessary background, setup and precise problem statement.

A. Notations.

For a vector $x \in \{-1,1\}^n$, let $x_i \in \{-1,1\}$ denote its i^{th} coordinate. Define $[n] = \{1,2,\dots,n\}$, and for any $J \subset [n]$, let $|J|$ denote its cardinality. For $x \in \{-1,1\}^n$ and $J \subset [n]$, let $x_J = (x_j)_{j\in J} \in \{-1,1\}^{|J|}$ denote sub-vector with its co-ordinates coming from J. For a set S, we denote its cardinality as $|S|$ and let 2^S to be the power-set of S,.

For a matrix $A = [A_{ij}] \in \mathbb{R}^{m\times N}$, let $A_{i\star}$ denote its i^{th} row and $A_{\star j}$ denote its j^{th} column for $1 \le i \le m$, $1 \le j \le N$. Given a matrix $A \in \mathbb{R}^{m\times N}$ we let $\|A\|_\infty = \max_{i,j} |A_{i,j}|$. Finally, for a vector $v \in \mathbb{R}^n$, let its ℓ_p norm, $p \ge 1$, be $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$.

We shall use $\mathbf{0}$ for the vector of all 0s and $\mathbf{1}$ for the vector of all 1s with dimension dependent on the context.

B. Fourier representation.

Let $\mathcal{F}$ be the space of all real-valued functions defined on $\{-1,1\}^n$, i.e. $\mathcal{F} = \{f : \{-1,1\}^n \mapsto \mathbb{R}\}$. This space of functions forms a Hilbert space under the following inner product: for any $f,\ g \in \mathcal{F}$, $\langle f, g\rangle = \frac{1}{2^n}\sum_{x\in\{-1,1\}^n} f(x)g(x)$. The induced norm is

$$\|f\|^2 = \langle f, f\rangle = \frac{1}{2^n}\sum_{x\in\{-1,1\}^n} f^2(x).$$

For the set of all Boolean functions, $\mathcal{F}_{\text{bool}} = \{f : \{-1,1\}^n \mapsto \{-1,1\}\}$, $\|f\| = 1$ since $f^2(x) = 1$ for all $x \in \{-1,1\}^n$.

The above Hilbert space naturally has an orthonormal basis. As alluded to above in equation (2), a particular choice of it, utilized popularly in the literature, is as follows. For each $J \subset [n]$, define a basis function $\chi_J \in \mathcal{F}_{\text{bool}}$ as

$$\chi_J(x) = \begin{cases} \Pi_{i\in J} x_i & \text{if } J \neq \emptyset \\ 1 & \text{if } J = \emptyset. \end{cases}$$

Now $\|\chi_J\| = 1$ for all $J \subset [n]$, since $\chi_J \in \mathcal{F}_{\text{bool}}$. It can be checked easily that for $J \neq J'$ and $J, J' \subset [n]$,

$$\langle \chi_J, \chi_{J'}\rangle = 0.$$

Finally, the size of collection $\{\chi_J : J \subset [n]\}$ is 2^n, the dimension of $\mathcal{F}$. Therefore, it is indeed an orthonormal basis of $\mathcal{F}$. Given this, for any $f \in \mathcal{F}$, it can be represented as

$$f(x) = \sum_{J\subset[n]} \alpha_J(f)\chi_J(x), \tag{4}$$

where the 'Fourier' coefficient $\alpha_J(f)$ is given by

$$\alpha_J(f) = \langle f, \chi_J\rangle.$$

When clear from context, we shall drop the reference to f in the notation $\alpha_J(f)$ and instead simply use α_J. Finally, we recall the Parseval's identity

$$\begin{aligned}\|f\|_2^2 &= \langle f, f\rangle \\ &= \sum_{J,J'\subset[n]} \alpha_J\alpha_{J'}\langle \chi_J, \chi_{J'}\rangle \\ &= \sum_{J\subset[n]} \alpha_J^2. \end{aligned} \tag{5}$$

C. Sparse polynomials.

A function $f \in \mathcal{F}$ (not necessarily ± 1 valued) can be decomposed as a sparse polynomial if we assume that the set $\{\alpha_J(f) \neq 0\}$ has cardinality $s \ll 2^n$. The goal of this paper will be to effectively exploit this sparse structure of the set of coefficients $\{\alpha_J(f)\}$. As an example, we again recall the d-Junta described above. Let $K \subset [n]$ be the subset of $d = |K|$ variables that determine the function f. For such a function, it can be verified that for $J \subset [n]$ such that $J\backslash K \neq \emptyset$ we have $\alpha_J(f) = \langle f, \chi_J\rangle\ =\ 0$. Therefore,

$$f(x) = \sum_{J\subset K} \alpha_J\chi_J(x). \tag{6}$$

Thus, learning f boils down to learning 2^d coefficients, $\alpha_J = \alpha_J(f)$ for $J \subset K$. A number of authors have developed techniques for solving the Junta-problem, however, those techniques do not necessarily lend themselves to learning generic sparse polynomials.

D. Observation model.

We assume that we are given m labeled observations $(x^i, f(x^i))$ to learn a sparse-polynomial function f. The $x^i \in \{-1,1\}^n$ are chosen independently and uniformly at random. That is, we observe $y \in \mathbb{R}^m$, where the i^{th} component of y,

$$\begin{aligned} y_i &= f(x^i) \\ &= \sum_{J\subset[n]} \alpha_J\chi_J(x^i). \end{aligned} \tag{7}$$

We shall call learning f with respect to this observation model, *exact* recovery of f since we observe the exact value of f evaluated at the sample x^i. In contrast, in a *noisy* observation model, observations are captured by $y \in \mathbb{R}^m$ with the i^{th} component of y being

$$y_i = f(x_i) + \varepsilon_i, \tag{8}$$

with noise ε_i being such that for $\epsilon > 0$,

$$\left(\frac{1}{m}\sum_i \varepsilon_i^2\right)^{\frac{1}{2}} \le \epsilon. \tag{9}$$

We shall call this the ϵ-noisy version. Clearly, for the special case of $\varepsilon = \mathbf{0}$, the problem is equivalent to the task of exact recovery.

We introduce some notations for an alternative representation for observations y. Consider an ordering of 2^n subsets $J \subset [n]$, denoted by $\pi : [2^n] \to 2^{[n]}$. Thus, index $j \in [2^n]$ corresponds to the subset $J = \pi(j) \subset [n]$. With this ordering in mind, we shall (ab)use notation $\alpha_j \equiv \alpha_j(f) = \alpha_{\pi(j)}(f)$ and $\chi_j \equiv \chi_{\pi(j)}$. Therefore, for any given function f with Fourier coefficient vector $\alpha \in \mathbb{R}^{2^n}$, the observation vector $y \in \mathbb{R}^m$, under the exact and noisy model, can be compactly represented as

$$y = A\alpha + \varepsilon. \tag{10}$$

In above, $\varepsilon \in \mathbb{R}^m$ is the noise vector, which is the vector of 0s in the exact model and we have $\|\varepsilon\|_2 \leq \epsilon$ for the ϵ-noisy model. The matrix $A \in \{-1, 1\}^{m \times 2^n}$ is such that $A_{i,j} = \chi_{\pi(j)}(x^i)$.

E. Problem statement.

In summary, given observations $y \in \mathbb{R}^m$ related to the sparse-polynomial f as per (10), the interest is in recovering $\alpha \in \mathbb{R}^{2^n}$ with $\|\alpha\|_0 \leq s$ from *independent* samples. The goal is to do so with high probability for as small m as possible with error in the produced estimation $\hat{\alpha}$, $\|\hat{\alpha} - \alpha\|_2 = O(\epsilon)$.

We note that this setting is exactly the same as sparse vector recovery from (noisy) linear measurements considered in the compressed sensing literature, cf. [8][9]. In a generic result in compressed sensing literature, the considered measurement matrices (here A) usually have independent and identically distributed entries (with distributions like Gaussian, Bernoulli or Rademacher, etc.). In our setting, though matrix A is random (due to randomness of samples, x^i), its entries are strongly correlated and more related to the problems in compressed sensing involving the Fourier ensemble [10] or correlated Gaussian design matrices [11]. However, as we shall see, any two distinct columns of A are independent. This observation turns out to be sufficient to establish guarantees similar to those obtained in the compressive sensing literature.

III. Results.

This section describes the recovery algorithm, as well as its sample complexity.

A. Recovery algorithm.

Given the similarity with stable compressive sensing [9], we propose to estimate the unknown α by solving the following convex program:

$$\widehat{\alpha} \in \arg\min_{\beta \in \mathbb{R}^{2^n}} \|\beta\|_1 \quad \text{such that} \quad \frac{1}{\sqrt{m}}\|A\beta - y\|_2 \leq \epsilon. \tag{11}$$

Here we assume that we know the bound ϵ on the normalized ℓ_2-norm of the error vector. In case of the exact observation model, $\epsilon = 0$, and hence the above program becomes a linear program (also known as basis pursuit):

$$\widehat{\alpha} \in \arg\min \|\beta\|_1 \quad \text{such that} \quad y = A\beta \quad \text{over} \quad \beta \in \mathbb{R}^{2^n}. \tag{12}$$

a) Two remarks:: First, we note that in the setting where α is not exactly s-sparse, but approximately s-sparse, we may still use the above algorithm and obtain an estimate $\widehat{\alpha}$ that approximates the true α. Second, we may incorporate additional information about any restrictions on the support of α by changing the condition $\beta \in \mathbb{R}^{2^n}$ to the appropriate sets of indices, for example, enforcing certain coefficients to be zero.

B. Sample complexity.

In this section we discuss some of the consequences of this paper. We show that given m observations, we may recover an estimate $\widehat{\alpha}$ of α that will satisfy certain desirable properties with high probability, in the noiseless and noisy settings for both approximately and exactly sparse polynomials. We establish the following guarantees about the estimator (11).

Theorem 1: Let constants $c = 4096$, $c_1 = 4$ and $c_2 = 8$. Then, for an arbitrary subset S with $s = |S|$ and

$$m \geq c\, s^2 n \tag{13}$$

we have that with probability at least $1 - O\left(\frac{1}{4^n}\right)$, the solution $\widehat{\alpha}$ of (11) is such that

$$\|\widehat{\alpha} - \alpha\|_2 \leq c_1\epsilon + c_2\|\alpha_{S^c}\|_1 \left(\frac{n}{m}\right)^{\frac{1}{4}}. \tag{14}$$

The above result holds for any arbitrary set S and any function f. Since the choice of $\widehat{\alpha}$ is independent of S, the result holds for S (with $|S| \leq s$) that optimizes (14). Specifically, if f is indeed approximately s-sparse or approximately observed cf. equation (10), then as per (14), algorithm (11) recovers a function $\hat{f}$ with $\|\hat{f} - f\|_2 = O(\epsilon)$ by Parseval's identity, where $\hat{f}(x) = \sum_{k=1}^{2^n} \widehat{\alpha}_k \chi_k(x)$. Indeed, if f were exactly s-sparse and $\varepsilon = 0$, our algorithm recovers f exactly with sample complexity $O(s^2 n)$ with high probability. The computational cost of the algorithm (11)–a linear program–scales polynomially in the optimization problem size, i.e. $O(\exp(\Theta(n)))$.

Now let us contrast the above performance with the naive algorithm: the algorithm that will search through all possible sets of coefficients and find the set that best fits the observations. Suppose we know a priori that the underlying function f is s-sparse, i.e. there exists a set S, $|S| \leq s$, corresponding to the indices of non-zero coefficients in the polynomial decomposition of f, cf. equation (1). Then the naive method is one algorithm for finding the maximum likelihood estimate of f. Note that there are $\binom{2^n}{s} \approx O(2^{ns})$ such possible sets of indices of size s, which we will denote as the *models*. Therefore, by a standard argument, to confidently learn the function from a model class of size $O(2^{ns})$, one needs at least $\Omega(\log 2^{ns}) = \Omega(ns)$ samples. However, we must enumerate through all possible sets which is approximately 2^{ns}. Therefore, the computational cost of the above algorithm scales proportionally to the size of the model class, i.e. $O(2^{ns})$. In summary, our algorithm achieves sample complexity that is nearly optimal ($O(s^2 n)$ vs. $O(ns)$); and has computation cost that scales polynomial with ambient dimension $O(2^n)$ without any dependence on s, in contrast to the computational cost of $O(2^{ns})$ of the naive algorithm. We may now proceed with the proof of Theorem 1 deferring certain technical details to the appendix.

IV. Proof of Theorem 1.

We now establish the proof of Theorem 1. To that end, let $\widehat{\alpha}$ be the solution of (11), which is an estimate of α. Define $\Delta = \widehat{\alpha} - \alpha$ and recall that $\|\alpha_S\|_0 \le s$. Let S with be subset of indices of α that are non-zero. We state the following property of Δ, known as the *cone condition* [7]:

Proposition 1: For $\Delta \in \mathbb{R}^{2^n}$,

$$\|\Delta_{S^c}\|_1 \le \|\Delta_S\|_1 + 2\|\alpha_{S^c}\|_1. \tag{15}$$

Note that in the setting that α is exactly supported on S we have $\alpha_{S^c} = 0$. For completeness, we have included its proof in the Appendix. An important implication of equation (15) yields that

$$\begin{aligned} \|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \ &\le 2\|\Delta_S\|_1 + 2\|\alpha_{S^c}\|_1 \\ &\le 2\sqrt{|S|}\|\Delta_S\|_2 + 2\|\alpha_{S^c}\|_1 \\ &\le 2\sqrt{|S|}\|\Delta\|_2 + 2\|\alpha_{S^c}\|_1. \end{aligned} \tag{16}$$

In the above, we have used the fact that for any $v \in \mathbb{R}^N$, $\|v\|_1 \le \sqrt{N}\|v\|_2$. Using the fact that $\widehat{\alpha}$ is a feasible solution of (11) and $\|A\Delta\|_2 = \|A\alpha - A\widehat{\alpha}\|_2$ we have

$$\begin{aligned} \|A\alpha - A\widehat{\alpha}\|_2 &\le \|y - A\widehat{\alpha}\|_2 + \|\varepsilon\|_2 \\ &\le 2\sqrt{m}\epsilon. \end{aligned} \tag{17}$$

Next, with $\mathbf{I}$ denoting the identity matrix and $\tilde{A} = \frac{1}{\sqrt{m}}A$ we have

$$\begin{aligned} \frac{1}{m}\|A\Delta\|_2^2 = \Delta^T\tilde{A}^T\tilde{A}\Delta \ &= \Delta^T\Big(\tilde{A}^T\tilde{A} - \mathbf{I}\Big)\Delta + \Delta^T\Delta \\ &\ge \|\Delta\|_2^2 - \|\Delta\|_1\Big\|\Big(\tilde{A}^T\tilde{A} - \mathbf{I}\Big)\Delta\Big\|_\infty \\ &\ge \|\Delta\|_2^2 - \|\Delta\|_1^2\Big\|\Big(\tilde{A}^T\tilde{A} - \mathbf{I}\Big)\Big\|_\infty. \end{aligned} \tag{18}$$

Finally, substituting equation (16) into equation (18) yields

$$\begin{aligned} \frac{1}{m}\|A\Delta\|_2^2 \ \ge \|\Delta\|_2^2 - 8|S|\|\Delta\|_2^2\Big\|\Big(\tilde{A}^T\tilde{A} - \mathbf{I}\Big)\Big\|_\infty \\ - 8\|\alpha_{S^c}\|_1^2\Big\|\Big(\tilde{A}^T\tilde{A} - \mathbf{I}\Big)\Big\|_\infty. \end{aligned} \tag{19}$$

In above, we have used inequalities: (a) for any two vectors $u, v \in \mathbb{R}^N$, $|u^Tv| \le \|u\|_1\|v\|_\infty$, and (b) for any matrix $Q \in \mathbb{R}^{L\times N}$ and $v \in \mathbb{R}^N$, $\|Qv\|_\infty \le \|Q\|_\infty\|v\|_1$. We wish to further lower bound (19) as $1/2\|\Delta\|_2^2$. Such a bound is known as a restricted eigenvalue condition [12], [7]. To that end, we state the following lemma, which shows that the matrix $\tilde{A}$ satisfies an incoherence property [4], [5], [6].

Lemma 1: The normalized observation matrix $\tilde{A} = \frac{1}{\sqrt{m}}A$ is such that

$$\Big\|\tilde{A}^T\tilde{A} - \mathbf{I}\Big\|_\infty \le 4\sqrt{\frac{n}{m}}, \tag{20}$$

with probability at least $1 - \frac{2}{4^n}$.

We provide the proof of the above lemma in the Appendix. From Lemma 1, for $m \ge 4096|S|^2n$, it follows that[2]

$$\Big\|\tilde{A}^T\tilde{A} - \mathbf{I}\Big\|_\infty \le \frac{1}{16|S|}. \tag{21}$$

Therefore, combining equations (19) and (21) yields

$$\frac{1}{m}\|A\Delta\|_2^2 \ge \frac{1}{2}\|\Delta\|_2^2 - 32\|\alpha_{S^c}\|_1^2\sqrt{\frac{n}{m}}. \tag{22}$$

From equations (17) and (22); and the fact that $|S| \le s$, it follows that for $m \ge 4096\, s^2n$, with probability $1 - O(1/2^n)$, $\|\Delta\|_2^2 \ \le\ 8\,\epsilon^2 \ +\ 64\|\alpha_{S^c}\|_1^2\sqrt{\frac{n}{m}}$, which is further upper bounded as $\Big(4\epsilon + 8\|\alpha_{S^c}\|_1\big(\frac{n}{m}\big)^{\frac{1}{4}}\Big)^2$. This completes the proof of Theorem 1.

V. Discussion.

In this paper, we considered learning s-sparse polynomial functions under a uniform sampling model. Inspired by results from compressive sensing, we presented a convex optimization based recovery algorithm. The algorithm requires $m = O\Big(s^2n\Big)$ samples where the produced estimate is within error $O(\epsilon)$ where ϵ is a bound on the rescaled ℓ_2-norm of the per-sample error. Our results naturally extend to the setting where the function f is well approximated by an s-sparse polynomial. We further note that the entire space of 2^n possible subsets need not necessarily be considered. Indeed, if it is known a priori that a smaller set of indices of α are non-zero, then we may restrict our attention to that smaller set. Namely, if we know that there are at most N possible locations where the s non-zero components lie then we can restrict our attention to those N coefficients. In such a setting, our results allow us to replace any occurrence of 2^n simply with N. Hence, the number of required samples becomes $O\Big(s^2\log N\Big)$ for learning an s-sparse Boolean polynomial.

Appendix

We begin with the Proof of Proposition 1. This proposition is known in the literature, cf. [9], [7]. We provide its proof for completeness. Let S be some set of indices–typically we set S to be the set of indices over which the components of α are non-zero. Both α and $\widehat{\alpha}$ are feasible solutions of (11) and $\widehat{\alpha}$ is its solution. Therefore,

$$\begin{aligned} \|\widehat{\alpha}_S\|_1 + \|\widehat{\alpha}_{S^c}\|_1 &= \|\widehat{\alpha}\|_1 \\ &\le \|\alpha\|_1 \ \le \|\alpha_S\|_1 + \|\alpha_{S^c}\|_1. \end{aligned} \tag{23}$$

Whence,

$$\begin{aligned} \|\widehat{\alpha}_{S^c}\|_1 &\le \|\alpha_S\|_1 - \|\widehat{\alpha}_S\|_1 + \|\alpha_{S^c}\|_1 \\ &\le \|\alpha_S - \widehat{\alpha}_S\|_1 + \|\alpha_{S^c}\|_1 \\ &= \|\Delta_S\|_1 + \|\alpha_{S^c}\|_1. \end{aligned} \tag{24}$$

Note that $\|\widehat{\alpha}_{S^c}\|_1 = \|\Delta_{S^c} + \alpha_{S^c}\|_1 \ge \|\Delta_{S^c}\|_1 - \|\alpha_{S^c}\|_1$. Rearranging terms completes the proof of Proposition 1.

[2] These constants can be improved, however, to simplify the exposition and dependency on the problem parameters we have opted not to optimize numeric constants.

We now proceed with the Proof of Lemma 1. The proof relies on the following simple observation. Since x^i are chosen uniformly at random from $\{-1, 1\}^n$, all rows of $\tilde{A}$ are independent and identically distributed. The columns of $\tilde{A}$, however are not mutually independent. But, we show that they are pair-wise independent and this is sufficient to establish Lemma 1.

Now $\tilde{A}$ has a total of 2^n columns. Each such column is a $\pm\frac{1}{\sqrt{m}}$ valued vector of length m, and hence is normalized, i.e. the ℓ_2 norm is 1. For that reason, the diagonal entries of $\tilde{A}^T\tilde{A}$ are equal to 1. Therefore, to establish the Lemma 1, we need to show that the absolute values of the non-diagonal entries of $\tilde{A}^T\tilde{A}$ are at most $4\sqrt{\frac{n}{m}}$.

In order to establish this claim, let us inspect the columns of $\tilde{A}$ more carefully. The column of $\tilde{A}$ corresponding to the empty set has all entries $\frac{1}{\sqrt{m}}$; all other columns each have entries distributed uniformly (not necessarily independently) over $\pm\frac{1}{\sqrt{m}}$. We state the following property about the inner product of any two distinct columns of $\tilde{A}$ that will establish the desired result:

Lemma 2: Consider two columns of $\tilde{A}$, corresponding to sets $J, J' \subset [n]$. Let $a = [a_i], b = [b_i] \in \{-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\}^m$ denote these columns of $\tilde{A}$ corresponding to J and J', respectively. Let $z_i = a_i b_i$ for $1 \leq i \leq m$. Then $z_1, \ldots, z_m$ are independent and identically distributed random variables and each of them is uniformly distributed over $\{-\frac{1}{m}, \frac{1}{m}\}$.

Proof: Given that $J \neq J'$, their symmetric difference $J\Delta J' \neq \emptyset$. Recall that

$$a_i = \frac{1}{\sqrt{m}} \prod_{j \in J} x^i_j, \qquad b_i = \frac{1}{\sqrt{m}} \prod_{j' \in J'} x^i_{j'}. \tag{25}$$

Since x^i are chosen independently and uniformly over $\{-1, 1\}^n$, the random variables x^i_j are independent and identically distributed with distribution being uniform over $\{1, -1\}$ for fixed j and varying i. Therefore,

$$\begin{aligned} z_i &= \frac{1}{m}\Big(\prod_{j \in J} x^i_j\Big) \times \Big(\prod_{j' \in J'} x^i_{j'}\Big) \\ &= \frac{1}{m}\Big(\prod_{j \in J\Delta J'} x^i_j\Big) \times \Big(\prod_{j' \in J \cap J'} (x^i_{j'})^2\Big) \\ &= \frac{1}{m}\Big(\prod_{j \in J\Delta J'} x^i_j\Big). \end{aligned} \tag{26}$$

Since $J \neq J'$ and hence $J\Delta J' \neq \emptyset$, from above it follows that z_i is distributed uniformly over $\{-\frac{1}{m}, \frac{1}{m}\}$. Also since z_i depends on x^i, they are independent across $1 \leq i \leq m$. This completes the proof of Lemma 2. ■

From Lemma 2, it follows that the inner product of columns of $\tilde{A}$ corresponding to any two different sets $J \neq J'$ is the sum of m independent and identically distributed random variables, $z_1, \ldots, z_m$, with each z_i distributed uniformly over $\Big\{-\frac{1}{m}, \frac{1}{m}\Big\}$. By standard Azuma-Hoeffding's bound, it follows that for any $t > 0$,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^m z_i\Big| \geq t\Big) \leq 2\exp\Big(-\frac{mt^2}{8}\Big). \tag{27}$$

Therefore, by selecting $t = 4\sqrt{\frac{n}{m}}$, it follows that

$$\mathbb{P}\Big(\Big|\sum_{i=1}^m z_i\Big| \geq 4\sqrt{\frac{n}{m}}\Big) \leq \frac{2}{256^n}. \tag{28}$$

Thus, the absolute values of all non-diagonal entries of $\tilde{A}^T\tilde{A}$ are at most $4\sqrt{\frac{n}{m}}$ with probability at least $1 - \frac{2}{256^n}$. Therefore, by union bound (over at most 4^n non-diagonal possible entries), it follows that the maximum of the absolute values of all non-diagonal entries of $\tilde{A}^T\tilde{A}$ is at most $4\sqrt{\frac{n}{m}}$ with probability at least $1 - \frac{2}{64^n}$. In summary, it follows that with probability at least $1 - O\big(1/4^n\big)$,

$$\Big\|\tilde{A}^T\tilde{A} - \mathbf{I}\Big\|_\infty \leq 4\sqrt{\frac{n}{m}}. \tag{29}$$

This completes the proof of Lemma 1.

References

[1] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[2] N. Littlestone, "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm," *Machine learning*, vol. 2, no. 4, pp. 285–318, 1988.

[3] A. Blum, L. Hellerstein, and N. Littlestone, "Learning in the presence of finitely or infinitely many irrelevant attributes," *Journal of Computer and System Sciences*, vol. 50, no. 1, pp. 32–40, 1995.

[4] D. L. Donoho, M. Elad, and V. M. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Info Theory*, vol. 52, no. 1, pp. 6–18, January 2006.

[5] F. Bunea, A. Tsybakov, and M. Wegkamp, "Sparsity oracle inequalities for the lasso," *Electronic Journal of Statistics*, vol. 1, pp. 169–194, 2007.

[6] K. Lounici, "Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators," *Electronic Journal of Statistics*, vol. 2, pp. 90–102, 2008.

[7] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," in *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2009, to appear in Statistical Science.

[8] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, August 2006.

[10] ——, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Theory*, vol. 52, no. 2, pp. 489–509, February 2004.

[11] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue conditions for correlated Gaussian designs," *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, August 2010.

[12] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.