Revisiting Model Selection and Recovery of Sparse Signals Using One-Step Thresholding

Waheed U. Bajwa, Robert Calderbank, and Sina Jafarpour

***Abstract*—This paper studies non-asymptotic model selection and recovery of sparse signals in high-dimensional, linear inference problems. In contrast to the existing literature, the focus here is on the general case of arbitrary design matrices and arbitrary nonzero entries of the signal. In this regard, it utilizes two easily computable measures of coherence—termed as the worst-case coherence and the average coherence—among the columns of a design matrix to analyze a simple, model-order agnostic one-step thresholding (OST) algorithm. In particular, the paper establishes that if the design matrix has reasonably small worst-case and average coherence then OST performs near-optimal model selection when either (i) the energy of any nonzero entry of the signal is close to the average signal energy per nonzero entry or (ii) the signal-to-noise ratio (SNR) in the measurement system is not too high. Further, the paper shows that if the design matrix in addition has sufficiently small spectral norm then OST also exactly recovers most sparse signals whose nonzero entries have approximately the same magnitude even if the number of nonzero entries scales almost linearly with the number of rows of the design matrix. Finally, the paper also presents various classes of random and deterministic design matrices that can be used together with OST to successfully carry out near-optimal model selection and recovery of sparse signals under certain SNR regimes or for certain classes of signals.**

I. Introduction

Model selection and signal recovery are two of the most well-studied problems in the statistics and signal processing literature. In this paper, we revisit these two problems for the case when the measured data $y \in \mathbb{C}^n$ is characterized by the linear model $y = X\beta + \eta$ and the signal $\beta \in \mathbb{C}^p$ satisfies $\|\beta\|_0 \doteq \sum_{i=1}^{p} 1_{\{|\beta_i|>0\}} \leq k \ll p$. Here, X is an $n \times p$ matrix called the *measurement* or *design matrix*, while $\eta \in \mathbb{C}^n$ represents noise in the measurement system. In this setup, the assumption that β is "k-sparse" allows one to operate in the so-called "compressed" setting, $k < n \ll p$, thereby enabling tasks that might be deemed prohibitive otherwise.

The primary objective of this paper is to consider the general case of arbitrary (random or deterministic) design matrices and arbitrary nonzero entries of the signal and study in a compressed setting the problems of (i) *polynomial-time, model-order agnostic model selection* (also known as *variable selection* and *sparsity pattern recovery*) and (ii) *low-complexity, model-order agnostic recovery of sparse signals in the noiseless case*. In order to accomplish this task, we consider two fundamental measures of coherence among the (normalized) columns $\{\mathrm{x}_i \in \mathbb{C}^n\}$ of X, namely,[1]

- *Worst-Case Coherence*: $\mu(X) \doteq \max\limits_{i,j:i\neq j} \left|\langle \mathrm{x}_i, \mathrm{x}_j\rangle\right|$, and
- *Average Coherence*: $\nu(X) \doteq \frac{1}{p-1} \max\limits_{i} \left|\sum\limits_{j:j\neq i} \langle \mathrm{x}_i, \mathrm{x}_j\rangle\right|$.

Roughly speaking, worst-case coherence—which seems to have been introduced in the related literature in [1]—is a similarity measure between the columns of a design matrix: the smaller the worst-case coherence, the less similar the columns. On the other hand, average coherence—which was introduced in a prequel to this paper [2]—is a measure of the spread of the columns of a design matrix within the n-dimensional unit ball: the smaller the average coherence, the more spread out the column vectors.

The first main contribution of this paper is that we make use of these two measures of coherence to propose and analyze model-order agnostic thresholds for the *one-step thresholding* (OST) algorithm (see Algorithm 1) for model selection and recovery of sparse signals. Specifically, we strengthen in this paper our recent result concerning OST [2] by (i) characterizing in Section II-A both the exact and the partial model-selection performance of OST in a non-asymptotic setting in terms of μ and ν, and (ii) extending in Section II-B our results on model selection using OST to (noiseless) recovery of sparse signals by making use of a recent result by Tropp [3]. Two key implications of the reported results in this regard are as follows. First, if the design matrix X satisfies $\mu(X) \asymp n^{-1/2}$ and $\nu(X) \precsim n^{-1}$ then OST—despite being computationally primitive—performs near-optimal model selection when either (i) the energy of any nonzero entry of β is not too far away from the average signal energy per nonzero entry $\|\beta\|_2^2/k$ or (ii) the signal-to-noise ratio (SNR) in the measurement system is not too high.[2] Second, if X in addition satisfies $\|X\|_2 \asymp \sqrt{p/n}$ then OST also exactly recovers most k-sparse *unimodal signals* (defined as signals whose nonzero entries have approximately the same magnitude) as long as $k \precapprox n$.

The second main contribution of this paper is that we characterize in Section III the worst-case coherence μ, the

[1] Here, we assume without loss of generality that X has (approximately) unit ℓ_2-norm columns. This is because deviations to this assumption can always be accounted for by appropriately scaling the entries of β instead.

[2] Recall "*Big–O*" notation: $f(n) = O(g(n))$ (alternatively, $f(n) \precsim g(n)$) if $\exists\, c_o > 0, n_o : \forall\, n \geq n_o, f(n) \leq c_o g(n)$, $f(n) = \Omega(g(n))$ (alternatively, $f(n) \succsim g(n)$) if $g(n) = O(f(n))$, and $f(n) = \Theta(g(n))$ (alternatively, $f(n) \asymp g(n)$) if $g(n) \precsim f(n) \precsim g(n)$. In addition, we sometimes use the shorthand notation $f(n) \succapprox g(n)$ (resp. $f(n) \precapprox g(n)$) to indicate that $f(n) \succsim g(n)$ (resp. $f(n) \precsim g(n)$) modulo a logarithmic factor.

This work was completed at Princeton University. Two of the authors, WUB and RC, are now with Duke University. The work is supported in part by NSF under grant DMS-0914892, by ONR under grant N00014-08-1-1110, and by AFOSR under grants FA9550-09-1-0643 and FA9550-09-1-0551. (E-mails: {w.bajwa, robert.calderbank}@duke.edu, sina@cs.princeton.edu)

average coherence ν, and the spectral norm $\|X\|_2$ of various classes of random and deterministic design matrices. In particular, these results—in conjunction with the results of Section II—imply that OST can be used together with random design matrices, such as Gaussian matrices and (random) partial Fourier matrices, as well as with deterministic design matrices, such as Alltop Gabor frames [4], [5], discrete-chirp matrices [6], [7], Delsarte–Goethals frames [8], and dual BCH sensing matrices, to successfully carry out model selection and sparse-signal recovery under certain SNR regimes or for certain classes of signals as long as $k \precapprox n$.

A. Relationship to Previous Work

In the context of model selection in the compressed setting, Mallow's C_p selection procedure [9], Akaike's information criterion [10], and their variants [11], [12] are known to perform well empirically as well as theoretically. Solving these model-selection procedures, however, is known to be an NP-hard problem. In order to overcome the computational intractability of [9]–[12], several methods based on convex optimization have been proposed in recent years. Among these proposed methods, the lasso [13] has arguably become the standard tool for model selection, which can be partly attributed to the theoretical guarantees provided for the lasso in [14]–[16]. Despite the recent theoretical triumphs of the lasso, however, it is still desirable to study alternative solutions to the problem of polynomial-time, model-order agnostic model selection in a compressed setting. This is because: (i) Lasso requires the minimum singular values of the submatrices of X corresponding to the true models to be bounded away from zero [14]–[16]. While this is a plausible condition for the case when one is interested in recovering β, it is arguable whether this condition is necessary for the case of model selection; (ii) The current literature on model selection using the lasso lacks guarantees beyond $k \succsim \mu^{-1}$ for the case of generic design matrices and arbitrary nonzero entries. In particular, given an arbitrary design matrix X [14]–[16] do not provide any guarantees beyond $k \succsim \sqrt{n}$ for even the simple case of $\beta \in \mathbb{R}^p_+$; and (iii) The computational complexity of the lasso for generic design matrices tends to be $O(p^3 + np^2)$ [17]. This makes the lasso computationally demanding for large-scale model-selection problems.

Recently, a few researchers have raised somewhat similar concerns about the lasso and revisited the much older (and oft-forgotten) method of thresholding for model selection [17]–[20], which has computational complexity of $O(np)$ only and which is known to be nearly optimal for $p \times p$ orthonormal design matrices [21]. Algorithmically, this makes our approach to model selection similar to that of [17]–[20]. Nevertheless, the OST algorithm presented in this paper differs from [17]–[20] in five key aspects: (i) Unlike [17]–[20], the OST algorithm presented in this paper is completely agnostic to the true model order k; (ii) The results reported in this paper hold for arbitrary (random or deterministic) design matrices and do not assume any statistical prior on the values of the nonzero entries of β even when k scales linearly with n. In contrast, [19] only studies the problem of Gaussian design matrices whereas the most influential results reported in [17], [18], [20] assume that the values of the nonzero entries of β are independent and statistically symmetric around zero; (iii) In contrast to [17]–[20], we relate the model-selection performance of OST to two global parameters of X, namely, μ and ν, which are trivially computable in polynomial time; (iv) Similar to [17], [19], [20], the analysis in this paper can be used to establish that OST achieves (asymptotically) consistent model selection under certain conditions. However, the results reported in this paper are completely non-asymptotic in nature (with explicit constants) and thereby shed light on the rate at which OST achieves consistent model selection; and (v) In addition to the exact model-selection performance of OST, we also characterize in the paper its partial model-selection performance. In this regard, we establish that the *universal threshold* proposed in Section II-A for OST guarantees $\widehat{\mathcal{S}} \subset \mathcal{S} \doteq \{i \in \{1, \dots, p\} : |\beta_i| > 0\}$ with high probability and we quantify the cardinality of the estimate $\widehat{\mathcal{S}}$. On the other hand, both [18] and [19] study only exact model selection, whereas [17], [20] study approximate (though not partial) model selection only for Gaussian design matrices [17] and assuming Gaussian (resp. statistical) priors on the nonzero entries of β [20] (resp. [17]).

Finally, in the context of sparse-signal recovery in the compressed setting, there exists a large body of literature that studies this problem under the rubric of *compressed sensing*. However, low-complexity iterative algorithms such as matching pursuit [22], subspace pursuit [23], CoSaMP [24], and iterative hard thresholding [25], and combinatorial algorithms based on group testing such as HHS pursuit [26] and Fourier samplers [27], [28] have been shown to perform well either only for some special classes of design matrices [26]–[28] or for design matrices that satisfy the *restricted isometry property* (RIP) [29]. Nevertheless, explicitly verifying that X satisfies the RIP of order $k \succsim \mu^{-1}$ is computationally intractable; in particular, since we have from the Welch bound [30] that $\mu^{-1} \precsim \sqrt{n}$ for $p \gg 1$, the guarantees provided in [23]–[25] for the case of generic design matrices at best hold only for k-sparse signals with $k \precsim \sqrt{n}$. On the other hand, convex optimization procedures such as basis pursuit [31] and lasso are ill-suited for large-scale problems because of their computational complexity and because they too lack guarantees beyond $k \succsim \mu^{-1}$ for the case of generic design matrices and arbitrary nonzero entries [16], [32]. In contrast, and motivated by the need to have verifiable sufficient conditions for low-complexity algorithms and arbitrary values of the nonzero entries of β even when $k \succsim \sqrt{n}$, we extend in Section II-B our results on model selection using OST and characterize the sparse-signal recovery performance of Algorithm 1 in terms of three global parameters of X: $\mu(X)$, $\nu(X)$, and $\|X\|_2$. In particular, a key implication of this part of the paper is that any design matrix that satisfies $\mu(X) \asymp n^{-1/2}$, $\nu(X) \precsim n^{-1}$, and $\|X\|_2 \asymp \sqrt{\frac{p}{n}}$ can be used along with OST to recover most k-sparse unimodal signals with arbitrary nonzero entries even when k scales almost linearly with n.

Algorithm 1 The One-Step Thresholding (OST) Algorithm for Model Selection and Recovery of Sparse Signals

```
Input: An $n \times p$ matrix $X$, a vector $y \in \mathbb{C}^n$, and a threshold $\lambda > 0$
Output: An estimate $\widehat{\mathcal{S}} \subset \{1, \dots, p\}$ of the model $\mathcal{S}$ and an estimate $\widehat{\beta} \in \mathbb{C}^p$ of the signal $\beta$
  $\widehat{\beta} \leftarrow \mathbf{0}$                                              {Initialize}
  $f \leftarrow X^{\mathrm{H}} y$                                             {Form signal proxy}
  $\widehat{\mathcal{S}} \leftarrow \{i \in \{1, \dots, p\} : |f_i| > \lambda\}$              {Select model via OST}
  $\widehat{\beta}_{\widehat{\mathcal{S}}} \leftarrow (X_{\widehat{\mathcal{S}}})^{\dagger} y$                          {Recover signal via least-squares}
```

II. Main Results

A. Model Selection Using One-Step Thresholding

We begin by reconsidering the model $y = X\beta + \eta$ and assume that X is an $n \times p$ design matrix having unit ℓ_2-norm columns, $\beta \in \mathbb{C}^p$ is a k-sparse signal ($\|\beta\|_0 \leq k$), and $k < n \leq p$. Here, we allow X to be either a random or a deterministic design matrix, while we take η to be a complex additive white Gaussian noise vector that is distributed as $\mathcal{CN}(\mathbf{0}, \sigma^2 I)$.[3] Finally, the main assumption that we make here is that the true model $\mathcal{S} \doteq \{i \in \{1, \dots, p\} : |\beta_i| > 0\}$ is a uniformly random k-subset of $\{1, \dots, p\}$. In other words, we have a uniform prior on the *support* of the signal β.

Intuitively speaking, successful model selection requires the columns of the design matrix to be *incoherent*. In this paper, we formulate this notion in terms of the *coherence property*.

Definition 1 (The Coherence Property)**.** An $n \times p$ design matrix X having unit ℓ_2-norm columns is said to obey the coherence property if the following two conditions hold:

$$\text{(CP-1)} \quad \mu(X) \leq \frac{0.1}{\sqrt{2 \log p}}\,, \quad \text{and} \quad \text{(CP-2)} \quad \nu(X) \leq \frac{\mu}{\sqrt{n}}\,.$$

Note that the coherence property is superior to other measures of incoherence such as the *irrepresentable condition* [14] in two key aspects. First, it does not require the singular values of the submatrices of X to be bounded away from zero. Second, it can be easily verified in polynomial time since it simply requires checking that $\|X^{\mathrm{H}}X - I\|_{\max} \leq (200 \log p)^{-1/2}$ and $\|(X^{\mathrm{H}}X - I)\mathbf{1}\|_\infty \leq (p-1)n^{-1/2}\|X^{\mathrm{H}}X - I\|_{\max}$.

Below, we describe the implications of the coherence property for both the exact and the partial model-selection performance of OST. Before proceeding further, however, it is instructive to first define some fundamental quantities pertaining to the problem of model selection as follows:

$$\beta_{\min} \doteq \min_{i \in \mathcal{S}} |\beta_i|\,, \qquad \mathsf{MAR} \doteq \frac{\beta_{\min}^2}{\|\beta\|_2^2/k}\,,$$

$$\mathsf{SNR}_{\min} \doteq \frac{\beta_{\min}^2}{\mathbb{E}[\|\eta\|_2^2]/k}\,, \qquad \mathsf{SNR} \doteq \frac{\|\beta\|_2^2}{\mathbb{E}[\|\eta\|_2^2]}\,.$$

In words, $\beta_{\min}$ is the magnitude of the smallest nonzero entry of β, while MAR—which is termed as *minimum-to-average ratio* [19]—is the ratio of the *energy in the smallest nonzero entry* of β and the *average signal energy per nonzero entry* of β. Likewise, $\mathsf{SNR}_{\min}$ is the ratio of the energy in the smallest nonzero entry of β and the average *noise* energy per nonzero entry, while SNR simply denotes the usual signal-to-noise ratio in the system. We are now ready to state the first main result of this paper that concerns the performance of OST in terms of exact model selection.

Theorem 1 (Exact Model Selection Using OST)**.** *Suppose that X satisfies the coherence property and choose the threshold $\lambda = \max\left\{\frac{1}{t}10\mu\sqrt{n \cdot \mathsf{SNR}}, \frac{1}{1-t}\sqrt{2}\right\}\sqrt{2\sigma^2 \log p}$ for any $t \in (0,1)$. Then, if we write $\mu(X)$ as $\mu = c_1 n^{-1/\gamma}$ for some $c_1 > 0$ (which may depend on p) and $\gamma \in \{0\} \cup [2, \infty)$, the OST algorithm (Algorithm 1) satisfies $\Pr(\widehat{\mathcal{S}} \neq \mathcal{S}) \leq 6p^{-1}$ provided $p \geq 128$ and the number of measurements satisfies*

$$n > \max\left\{2k \log p, \frac{c_2 k \log p}{\mathsf{SNR}_{\min}}, \left(\frac{c_3 k \log p}{\mathsf{MAR}}\right)^{\gamma/2}\right\}. \qquad (1)$$

Here, the quantities $c_2, c_3 > 0$ are defined as $c_2 \doteq 16(1-t)^{-2}$ and $c_3 \doteq 800c_1^2 t^{-2}$, while the probability of failure is with respect to the true model $\mathcal{S}$ and the noise vector η.

The proof of this theorem is provided in [33, Theorem 1]. Note that the parameter 't' in Theorem 1 can always be fixed a priori (say $t = 1/2$) without affecting the scaling relation in (1). In practice, however, t should be chosen so as to reduce the total number of measurements needed to ensure successful model selection. There are a few important remarks that need to be made at this point. First, it is easy to see that the proposed threshold in Theorem 1 is completely agnostic to the model order k and only requires knowledge of the SNR and the noise variance. Second, some of the bounds in the proof of [33, Theorem 1] and extensive simulations suggest that the absolute constant 10 in the proposed threshold is somewhat conservative and can be reduced through the use of more sophisticated analytical tools (this constant was 24 in [2]). Finally, while estimating the true model order k tends to be harder than estimating the SNR and the noise variance σ^2 in majority of the situations, it might be the case that estimating k is easier in some applications. It is better in such situations to work with a sorted variant of the OST algorithm that relies on knowledge of the model order k instead and returns an estimate $\widehat{\mathcal{S}}$ corresponding to the k largest (in magnitude) entries of $f \doteq X^H y$. We characterize the performance of this algorithm—which we term as *sorted one-step thresholding* (SOST) algorithm—in terms of the following theorem.

[3] Note that Gaussianity of η is just a simplified assumption for the sake of this exposition; in particular, the results presented here are generalizable to other noise distributions as well as perturbations having bounded ℓ_2-norms.

Theorem 2 (Exact Model Selection Using SOST)**.** *Suppose that X satisfies the coherence property and write $\mu(X)$ as $\mu = c_1 n^{-1/\gamma}$ for some $c_1 > 0$ (which may depend on p) and $\gamma \in \{0\} \cup [2, \infty)$. Then the sorted variant of the OST algorithm satisfies $\Pr(\widehat{\mathcal{S}} \neq \mathcal{S}) \leq 6p^{-1}$ as long as $p \geq 128$ and*

$$n > \min_{t \in (0,1)} \max\left\{2k\log p, \frac{c_2 k \log p}{\mathsf{SNR}_{\min}}, \left(\frac{c_3 k \log p}{\mathsf{MAR}}\right)^{\gamma/2}\right\}. \quad (2)$$

Here, the quantities $c_2, c_3 > 0$ are as defined in Theorem 1.

The proof of this theorem is just a slight variant of the proof of Theorem 1. A few remarks are in order now concerning OST and SOST. First, the computational complexity of SOST is comparable with that of OST since efficient sorting algorithms (such as heap sort) tend to have computational complexity of $O(p \log p)$ only. Second, (1) and (2) suggest that knowledge of the true model order k allows SOST to perform better than OST in situations where the threshold parameter t is fixed a priori (cf. Theorem 1). In this sense, SOST should be preferred over OST for exact model selection *provided* one has accurate knowledge of the true model order k. On the other hand, OST should be the algorithm of choice for model-selection problems where it is difficult to obtain a reliable estimate of the true model order.

The final result that we present here concerns the partial model-selection performance of OST. Specifically, note that our focus in this section has so far been on specifying conditions for the number of measurements that ensure exact model selection. In many real-world applications, however, the parameters of the problem are fixed and it is not always possible to ensure that n satisfies the aforementioned conditions. A natural question to ask then is whether the OST algorithm completely fails in such circumstances or whether any guarantees can still be provided for its performance. We address this aspect of the OST algorithm in the following and show that OST has the ability to identify the locations of the nonzero entries of β whose energies are greater than both the noise power and the average signal energy per nonzero entry. In order to make this notion mathematically precise, we first define the *m-th largest-to-average ratio* (LAR_m) of β as the ratio of the *energy in the m-th largest nonzero entry of* β and the average signal energy per nonzero entry of β; that is,

$$\mathsf{LAR}_m \doteq \frac{|\beta_{(m)}|^2}{\|\beta\|_2^2/k}$$

where $\beta_{(m)}$ denotes the m-th largest nonzero entry of β (note that $\mathsf{MAR} \equiv \mathsf{LAR}_k$). We are now ready to specify the partial model-selection performance of the OST algorithm.

Theorem 3 (Partial Model Selection Using OST)**.** *Suppose that X satisfies the coherence property and let $p \geq 128$. Next, fix a parameter $t \in (0,1)$ and choose the threshold $\lambda = \max\left\{\frac{1}{t}10\mu\sqrt{n \cdot \mathsf{SNR}}, \frac{1}{1-t}\sqrt{2}\right\}\sqrt{2\sigma^2 \log p}$. Then, under the assumption that $k \leq n/(2\log p)$, the OST algorithm (Algorithm 1) guarantees with probability exceeding $1 - 6p^{-1}$ that $\widehat{\mathcal{S}} \subset \mathcal{S}$ and $\left|\mathcal{S} - \widehat{\mathcal{S}}\right| \leq (k - M)$, where M is the largest integer for which the following inequality holds:*

$$\mathsf{LAR}_M > \max\left\{\frac{c_2 k \log p}{n \cdot \mathsf{SNR}}, \frac{c_3' k \log p}{\mu^{-2}}\right\}. \quad (3)$$

Here, $c_2 > 0$ is as defined in Theorem 1, $c_3' > 0$ is defined as $c_3' \doteq 800t^{-2}$, and the probability of failure is with respect to the true model $\mathcal{S}$ and the noise vector η.

The proof of this theorem is provided in [33, Theorem 5]. We conclude here by pointing out that no counterpart of Theorem 3 exists for the SOST algorithm since we can never have $\widehat{\mathcal{S}} \subset \mathcal{S}$ in that case because of the nature of the algorithm.

B. Recovery of Sparse Signals Using One-Step Thresholding

In this section, we extend our results on model selection using OST to model-order agnostic recovery of k-sparse signals. Note that we limit ourselves in this exposition to recovery of k-sparse signals in a noiseless setting; extensions of these results to reconstruction of k-sparse signals in noisy settings would be reported in a sequel to this paper. In other words, the measurement model that we study in here is $y = X\beta$ and the goal is to recover the k-sparse β using OST under the assumption that the true model $\mathcal{S} \doteq \{i \in \{1, \dots, p\} : |\beta_i| > 0\}$ is a uniformly random k-subset of $\{1, \dots, p\}$.

Intuitively speaking, the problem of sparse-signal recovery is inherently more difficult than the problem of model selection. We capture part of this intuitive notion in the following in terms of the *strong coherence property*.

Definition 2 (The Strong Coherence Property)**.** An $n \times p$ design matrix X having unit ℓ_2-norm columns is said to obey the strong coherence property if the following hold:

$$\text{(SCP-1) } \mu(X) \leq \frac{1}{60e \log p}, \quad \text{and} \quad \text{(SCP-2) } \nu(X) \leq \frac{\mu}{\sqrt{n}}.$$

In order to better illustrate the difference between the coherence property and the strong coherence property, note that the worst-case and the average coherence results reported for Gaussian design matrices in Section III-A show that Gaussian matrices satisfy the coherence property with high probability as long as $n \gtrsim (\log p)^2$. The same results, however, suggest that Gaussian matrices satisfy the strong coherence property with high probability as long as $n \gtrsim (\log p)^4$. In other words, there are scaling regimes in which Gaussian design matrices satisfy the coherence property but are not guaranteed to satisfy the strong coherence property. We are now ready to state the main result of this section.

Theorem 4 (Sparse-Signal Recovery Using OST)**.** *Suppose that X satisfies the strong coherence property and let $p \geq 128$. Next, choose the threshold $\lambda = 10\mu\|y\|_2\sqrt{\frac{2\log p}{1-e^{-1/2}}}$. Then the OST algorithm satisfies $\Pr(\widehat{\beta} \neq \beta) \leq 6p^{-1}$ provided*

$$k \leq \min\left\{\frac{p}{c_4^2 \|X\|_2^2 \log p}, \frac{\mu^{-2}\mathsf{MAR}}{c_5^2 \log p}\right\}. \quad (4)$$

Here, the probability of failure is only with respect to the true model $\mathcal{S}$, while c_4, c_5 are defined as $c_4 \doteq 37e$ and $c_5 \doteq 43$.

The proof of this theorem is provided in [33, Theorem 6]. We conclude this section by pointing out that if one does have knowledge of the true model order then it can be shown through a slight variation of the proof of [33, Theorem 6] that SOST (the sorted variant of the OST) can also recover sparse signals with high probability—the only difference in that case being that c_5 in Theorem 4 gets replaced with $c_5' \doteq \sqrt{800}$.

III. Near-Optimal Design Matrices for One-Step Thresholding: Some Examples

Section II establishes that design matrices with small worst-case coherence (and consequently small average coherence) and small spectral norm are particularly well-suited for model selection and recovery of sparse signals using OST (cf. Theorems 1–4). Further, since the Welch bound [30] dictates that $\mu \gtrsim n^{-1/2}$ for $p \gg 1$ and since we have from elementary linear algebra that $\|X\|_2 \geq \sqrt{p/n}$, we are in particular interested in design matrices that approximately satisfy the scaling relations $\mu(X) \asymp n^{-1/2}$, $\nu(X) \precsim n^{-1}$, and $\|X\|_2 \asymp \sqrt{p/n}$. In the following, we provide some examples of both random and deterministic design matrices that are nearly-optimal in terms of these requisite conditions (also, see Table I for an overview of the results reported in here).

A. Random Design Matrices

Random matrices are perhaps the most well-studied design matrices in the literature on high-dimensional, linear inference problems. This is in part due to the fact that geometric concepts such as the irrepresentable condition [14] and the restricted isometry property (RIP) [29] have, to date, been shown to hold near-optimally only for the case of random matrices. The following two lemmas specify that traditional random design matrices such as Gaussian matrices and (random) partial Fourier matrices also tend to be near-optimal in terms of the geometric measures of μ, ν, and/or $\|X\|_2$.

Lemma 1 (Geometry of Gaussian Matrices)**.** *Let X be an $n \times p$ design matrix with independent and identically distributed (i.i.d.) $\mathcal{N}(0, 1/n)$ entries and let $n \geq 60 \log p$. Then, we have that X satisfies (i) $\mu(X) \leq \sqrt{\frac{15 \log p}{n}}$, (ii) $\nu(X) \leq \frac{\sqrt{15 \log p}}{n}$, and (iii) $\|X\|_2 \leq 1 + 2\sqrt{\frac{p}{n}}$ with probability exceeding $1 - 2(p^{-1} + p^{-2} + e^{-p/2})$.*[4]

Note that the worst-case coherence bound in this lemma follows from bounds on the inner product of independent Gaussian vectors (see, e.g., [33, Appendix A]) and a simple union bound argument, the proof of the average coherence bound is provided in [33, Lemma 2], and the spectral norm bound follows from [34, (2.3)]. It is worth pointing out here that similar results can also be obtained for sub-Gaussian design matrices using standard concentration inequalities and [34, Proposition 2.4].

Lemma 2 (Geometry of Partial Fourier Matrices)**.** *Let U be a p-point (non-normalized) discrete Fourier transform matrix such that $U^{\mathrm{H}} U = pI$. Next, populate Ω by sampling n times with replacement from the set $\{1, \dots, p\}$ and construct X by collecting the rows of U corresponding to the indices in Ω and normalizing the resulting matrix by $1/\sqrt{n}$. Then X satisfies (i) $\mu(X) \leq \sqrt{\frac{12 \log p}{n}}$ and (ii) $\nu(X) \leq \max\left\{\frac{1}{p-1}, \frac{p-n}{n(p-1)}\right\}$ with probability exceeding $1 - 2p^{-1}$.*

In this lemma, the worst-case coherence bound follows by noting that the columns of U form a group under pointwise multiplication and then making use of the complex Hoeffding inequality [35]. On the other hand, the average coherence expression in it follows from the definition of the average coherence and the fact that $\mathbf{1}$ is in the null space of any partial Fourier matrix that does not include the first row of U. Finally, note that the fact that sampling in Lemma 2 is carried out with replacement makes it difficult to specify the spectral norm of X. In practice, however, one would not construct partial Fourier matrices with identical rows and the spectral norm of partial Fourier matrices in such cases would be $\sqrt{\frac{p}{n}}$ for the simple reason that the rows of U are mutually orthogonal.

B. Deterministic Design Matrices

Having described the geometry of Gaussian matrices and partial Fourier matrices, we now show that there in fact exist many classes of deterministic design matrices that are quite similar to these random design matrices in terms of the geometric measures of μ, ν, and $\|X\|_2$. This is in stark contrast to the best known results for the RIP of deterministic matrices and has important implications from an implementation viewpoint since multiplications with the deterministic matrices described below (and their adjoints) can be efficiently carried out using algorithms such as the *fast Fourier transform* (FFT) and the *fast Hadamard transform* (FHT).

1) Geometry of Alltop Gabor Frames: Gabor frames for $\mathbb{C}^n$ constitute an important class of frames, which are constructed from time- and frequency-shifts of a nonzero seed vector in $\mathbb{C}^n$. Specifically, let $g \in \mathbb{C}^n$ be a unit-norm seed vector and define T to be an $n \times n$ *time-shift matrix* that is generated from g as follows

$$T(g) \doteq \begin{bmatrix} g_1 & g_n & & g_2 \\ g_2 & g_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & g_n \\ g_n & g_{n-1} & & g_1 \end{bmatrix} \qquad (5)$$

where we write $T = T(g)$ to emphasize that T is a matrix-valued function on $\mathbb{C}^n$. Next, denote the collection of n samples of a discrete sinusoid with frequency $2\pi\frac{m}{n}, m \in \mathbb{Z}_n$ as $\omega_m \doteq \begin{bmatrix} e^{j2\pi\frac{m}{n}0} & \dots & e^{j2\pi\frac{m}{n}(n-1)} \end{bmatrix}^{\mathrm{T}}$. Finally, define the corresponding $n \times n$ diagonal *modulation matrices* as $W_m = \mathrm{diag}(\omega_m)$. Then the Gabor frame generated from g is an $n \times n^2$ block matrix of the form

$$X = \begin{bmatrix} W_0 T & W_1 T & \dots & W_{n-1} T \end{bmatrix}. \qquad (6)$$

In words, columns of the Gabor frame X are given by downward circular shifts and modulations (frequency shifts)

[4] Note that the results (and the definition of the coherence property) presented earlier remain valid if $\mu(X)$ is replaced with an upperbound $\bar{\mu}(X)$.

TABLE I
COMPARISONS BETWEEN DIFFERENT CLASSES OF RANDOM AND DETERMINISTIC DESIGN MATRICES

Design Matrices	p	$\mu(X)$	$\nu(X)$	$\lVert X\rVert_2$	Randomness	Complexity
Gaussian Matrices	–	$O\left(\sqrt{\frac{\log p}{n}}\right)$	$O\left(\frac{\sqrt{\log p}}{n}\right)$	$\Theta\left(\sqrt{\frac{p}{n}}\right)$	$\Theta\left(np\right)$	$O\left(np\right)$
Partial Fourier Matrices	–	$O\left(\sqrt{\frac{\log p}{n}}\right)$	$O\left(\max\left\{\frac{1}{p}, \frac{p-n}{n(p-1)}\right\}\right)$	–	$\Theta\left(n\right)$	$O\left(p\log p\right)$
Alltop Gabor Frames	n^2	$\frac{1}{\sqrt{n}}$	$O\left(\frac{1}{n}\right)$	$\sqrt{\frac{p}{n}}$	–	$O\left(p\log p\right)$
Discrete-Chirp Matrices	n^2	$\frac{1}{\sqrt{n}}$	$\frac{p-n}{n(p-1)}$	$\sqrt{\frac{p}{n}}$	–	$O\left(p\log p\right)$
Delsarte–Goethals Frames	n^{2+r} $\left(0 \le r \le \frac{\log n-1}{2}\right)$	$O\left(\frac{2^r}{\sqrt{n}}\right)$	$\frac{1}{p-1}$	$\sqrt{\frac{p}{n}}$	–	$O\left(p\log p\right)$
Dual BCH Sensing Matrices	n^2	$\sqrt{\frac{2}{n}}$	$\frac{p-n}{n(p-1)}$	$\sqrt{\frac{p}{n}}$	–	$O\left(p\log p\right)$

of the seed vector g. Therefore, Gabor frames are completely specified by a total of n numbers that describe the seed vector and matrix–vector multiplications $X\beta$ and $X^{\mathrm{H}}y$ can be carried out using the FFT in $O(p\log p)$ time. The following lemma characterizes the geometry of one specific class of Gabor frames, termed as Alltop Gabor frames [4], [5].

Lemma 3. *Let $n \geq 5$ be a prime number and construct the Alltop seed vector $g \in \mathbb{C}^n$ as follows*

$$g = \begin{bmatrix} \frac{1}{\sqrt{n}}e^{j2\pi\frac{0^3}{n}} & \frac{1}{\sqrt{n}}e^{j2\pi\frac{1^3}{n}} & \dots & \frac{1}{\sqrt{n}}e^{j2\pi\frac{(n-1)^3}{n}} \end{bmatrix}^{\mathrm{T}}. \quad (7)$$

Then the $n \times n^2$ Gabor frame X generated from g satisfies (i) $\mu(X) \equiv \frac{1}{\sqrt{n}}$, (ii) $\nu(X) \leq \frac{1}{n+1}$, and (iii) $\|X\|_2 \equiv \sqrt{\frac{p}{n}}$.

Here, the worst-case coherence expression follows from [5], the spectral norm expression is due to [36], while the proof of the average coherence bound is provided in [33, Theorem 7].

2) Geometry of Discrete-Chirp Matrices: Discrete-chirp matrices are $n \times n^2$ matrices that are constructed by collecting all possible chirp signals into columns [6]. Specifically, an n-length chirp signal for any prime n takes the form

$$\mathrm{x}_{m,r}(\ell) = \frac{1}{\sqrt{n}}e^{j2\pi\frac{m\ell}{n}+j2\pi\frac{r\ell^2}{n}}, \ \ell = 0, \dots, n-1 \quad (8)$$

where m is the base frequency and r is the chirp rate of the signal, and the columns of the $n \times n^2$ discrete-chirp matrix X are the n^2 distinct chirp signals corresponding to the n^2 possible pairs $(m,r) \in \mathbb{Z}_n \times \mathbb{Z}_n$. The following lemma characterizes the geometry of discrete-chirp matrices.

Lemma 4. *Let X be an $n \times n^2$ discrete-chirp matrix for any prime n. Then X satisfies (i) $\mu(X) \equiv \frac{1}{\sqrt{n}}$, (ii) $\nu(X) \equiv \frac{p-n}{n(p-1)}$, and (iii) $\|X\|_2 \equiv \sqrt{\frac{p}{n}}$.*

Here, the worst-case coherence bound and the spectral norm expression follow from [7], while the average coherence expression follows from the fact that $X^{\mathrm{H}}X\mathbf{1} \equiv \frac{p}{n}\mathbf{1}$. Finally, note that the structure of the discrete-chirp matrix X implies that the multiplications $X\beta$ and $X^{\mathrm{H}}y$ can be carried out using the FFT in $O(p\log p)$ time in this case also.

3) Geometry of Delsarte–Goethals Frames: The Delsarte–Goethals (DG) frames is a class of design matrices that has been recently introduced in the literature by Calderbank and Jafarpour [8]. Specifically, take $m \in \mathbb{Z}_+$ to be an odd number and $r \in \mathbb{Z}_+$ to be smaller than $\frac{m-1}{2}$. Next, use $DG(m,r)$ to denote the Delsarte–Goethals set of binary symmetric matrices, as described in [37]. Then, given $n = 2^m$ and $p = n2^{(r+1)m}$, the $n \times p$ DG frame X is constructed from $DG(m,r)$ in the following way. Index the rows of X by binary vectors $z \in \mathbb{F}_2^m$ and index the columns of X by pairs (P,b), where P ranges over all $2^{(r+1)m}$ binary symmetric matrices of $DG(m,r)$ and b ranges over all members of $\mathbb{F}_2^m$. Then the entries of X are given by

$$x_{(P,b)}(z) = \frac{1}{\sqrt{n}}\,\imath^{\mathsf{wt}(d_P)+2\mathsf{wt}(b)}\,\imath^{\langle z,Pz\rangle+2\langle z,b\rangle} \quad (9)$$

where $\imath \doteq \sqrt{-1}$, d_P denotes the principal diagonal of the matrix P, and $\mathsf{wt}(v)$ denotes the hamming weight (the number of nonzero entries) of a given vector v. It is easy to see from this description that (i) DG frames are unions of orthonormal bases and (ii) rows of DG frames correspond to Kronecker product of each row of an $n \times n$ Hadamard matrix with the corresponding row of an $n \times 2^{(r+1)m}$ matrix whose entries are given by $\{\imath^{\langle z,Pz\rangle}\}$, where P and z range over all the possible choices. This implies that the multiplications $X\beta$ and $X^{\mathrm{H}}y$ in the case of DG frames can be carried out using the FHT in $O(p\log p)$ time. Finally, the following lemma borrows results from [8] to characterize the geometry of DG frames.

Lemma 5. *Let X be an $n \times n^{2+r}$ Delsarte–Goethals frame obtained from $DG(m,r)$ set for some odd m. Then X satisfies (i) $\mu(X) \leq \frac{2^r}{\sqrt{n}}$, (ii) $\nu(X) \equiv \frac{1}{p-1}$, and (iii) $\|X\|_2 \equiv \sqrt{\frac{p}{n}}$.*

4) Geometry of Dual BCH Sensing Matrices: Dual BCH sensing matrices constitute another class of design matrices that corresponds to exponentiating the codewords of an algebraic code. Specifically, take $m \in \mathbb{Z}_+$ to be an odd number and use $BCH(m,2)$ to denote the extended 2-error correcting, binary BCH code of length $n = 2^m$ [38]. Then the dual of $BCH(m,2)$ is a code of length n and dimension $2m+1$ that is the union of n cosets of the first-order Reed–Muller code $RM(1,m)$ of dimension $m+1$; see [35] for further details. The important thing to point out here is that exponentiating codewords in the dual of $BCH(m,2)$ and scaling the resulting $n \times n^2$ matrix X by $1/\sqrt{n}$ gives a union of n orthonormal basis. This can be seen by noting that exponentiating codewords in $RM(1,m)$ gives Walsh basis vectors (and their negatives, which we discard in here). We also note because of the very same reason that the multiplications $X\beta$ and $X^{\mathrm{H}}y$ in the case of dual BCH sensing matrices can also be carried out using the FHT in $O(p\log p)$ time. The following lemma characterizes the geometry of dual $BCH(m,2)$ sensing matrices (a proof of this lemma would be provided in a sequel to this paper).

Lemma 6. *Let X be an $n \times n^2$ dual BCH sensing matrix obtained from the dual of $BCH(m,2)$ for some odd m. Then the matrix X satisfies (i) $\mu(X) \equiv \sqrt{\frac{2}{n}}$, (ii) $\nu(X) \equiv \frac{p-n}{n(p-1)}$, and (iii) $\|X\|_2 \equiv \sqrt{\frac{p}{n}}$.*

IV. Discussion

We conclude this paper by discussing our results in light of some of the results reported in previous works on model selection and recovery of sparse signals.

A. Model Selection Using One-Step Thresholding

1) Gaussian Design Matrices: Gaussian matrices are perhaps the most widely assumed design matrices in the model-selection literature. In order to specialize our results to Gaussian design matrices, recall from Lemma 1 that Gaussian matrices satisfy the coherence property with high probability as long as $n \gtrsim (\log p)^2$. Further, notice the following relation between $\mathsf{SNR}_{\min}$ and SNR and MAR: $\mathsf{SNR}_{\min} = \mathsf{SNR}\cdot\mathsf{MAR}$. Theorem 1 (resp. Theorem 2) then implies that OST (resp. SOST) correctly identifies the exact model with probability exceeding $1 - O(p^{-1})$ as long as $n \gtrsim \max\left\{1, \frac{1}{\mathsf{SNR}\cdot\mathsf{MAR}}, \frac{\log p}{\mathsf{MAR}}\right\} k\log p$. In particular, this suggests that if either $\mathsf{MAR}(\beta) = \Theta(1)$ or $\mathsf{SNR} = O(1)$ then OST leads to successful model selection with high probability provided $n \gtrapprox \max\left\{1, \frac{1}{\mathsf{SNR}\cdot\mathsf{MAR}}\right\} k\log p$. On the other hand, one of the best known results for model selection using the maximum likelihood algorithm requires that $n \gtrsim \max\left\{\frac{k\log(p-k)}{\mathsf{SNR}\cdot\mathsf{MAR}}, k\log(p/k)\right\}$ [39] (also see [19]). This establishes that OST performs near-optimally for model selection using Gaussian design matrices provided (i) the SNR in the measurement system is not too high or (ii) the energy of any nonzero entry of β is not too far away from the average energy $\|\beta\|_2^2/k$ and k scales sublinearly with p.

2) Lasso versus OST: Historically, OST is preferred over the lasso for model selection because of its low computational complexity. The results reported earlier, however, bring forth another important aspect of OST (also see [17]): *OST can lead to successful model selection even when the lasso fails*. Specifically, note that the lasso solution is not even guaranteed to be unique if the minimum singular value of the submatrix of X corresponding to the true model is not bounded away from zero (see, e.g., [14], [15]). On the other hand, OST does not require this condition for model selection. This is in part due to the fact that model selection using the lasso is in fact a byproduct of signal reconstruction, whereas the OST results for model selection do not guarantee signal reconstruction without imposing additional constraints on X. In other words, we have established here that *model selection is inherently an easier problem than sparse-signal reconstruction*.

Finally, it is worth comparing the model-selection performance of OST with that of the lasso for the cases when the lasso does succeed. In this regard, the most general result for model selection using the lasso states that if X is close to being a tight frame in the sense that $\|X\|_2 \approx \sqrt{p/n}$ then the lasso identifies the correct model with probability exceeding $1 - O(p^{-1})$ as long as (i) the nonzero entries of β are independent and statistically symmetric around zero, (ii) $k \lesssim n/\log p$, and (iii) $\mathsf{MAR} \gtrsim \frac{k\log p}{n\cdot\mathsf{SNR}}$ [16, Theorem 1.3]. On the other hand, assume now that the design matrix X has $\mu(X) \asymp n^{-1/2}$ and $\nu(X) \lesssim n^{-1}$ (Section III shows that there indeed exist many such matrices). We can then make use of the relation $\mathsf{SNR}_{\min} = \mathsf{SNR}\cdot\mathsf{MAR}$ to conclude from Theorem 1 (resp. Theorem 2) that OST (resp. SOST) identifies the correct model with probability exceeding $1 - O(p^{-1})$ as long as $k \lesssim n/\log p$ and $\mathsf{MAR} \gtrsim \max\left\{\frac{1}{\mathsf{SNR}}, 1\right\}\frac{k\log p}{n}$. This suggests that, even for the cases in which the lasso succeeds, OST can be guaranteed to perform as well as the lasso in situations where either the energy of any nonzero entry of β is not too far away from the average energy ($\mathsf{MAR} = \Theta(1)$) or the SNR is not too high ($\mathsf{SNR} = O(1)$). Equally importantly, and in contrast to the lasso results reported in [16], OST is guaranteed to attain this performance *irrespective* of the values of the nonzero entries of the signal β.

3) Near-Optimality of OST: We have concluded up to this point that—under certain conditions on MAR and SNR—the OST algorithm for model selection can perform as well as the lasso and it performs near-optimally for Gaussian design matrices. We conclude this discussion by arguing that OST in fact performs near-optimal model selection for *any* design matrix that satisfies $\mu(X) \asymp n^{-1/2}$ and $\nu(X) \lesssim n^{-1}$ as long as $\mathsf{MAR} = \Theta(1)$ or $\mathsf{SNR} = O(1)$. In order to accomplish this, we first recall the thresholding results obtained by Donoho and Johnstone [21]—which form the basis of ideas such as the wavelet denoising—for the case of orthonormal design matrices. Specifically, it was shown in [21] that if X is an orthonormal basis then thresholding the entries of $X^H y$ at $\lambda \asymp \sqrt{\sigma^2 \log p}$ results in oracle-like performance in the sense that one recovers (with high probability) the locations of all the nonzero entries of β that are above the noise floor.

Now the first thing to note regarding the results presented earlier is the intuitively pleasing nature of the threshold proposed for model selection using OST. Specifically, assume that X is an orthonormal matrix and notice that, since $\mu(X) = 0$ in this case, the threshold $\lambda \asymp \max\left\{\mu\sqrt{n\cdot\mathsf{SNR}}, 1\right\}\sqrt{\sigma^2\log p}$ proposed earlier reduces to the threshold proposed in [21] *and* Theorem 3 guarantees that thresholding recovers (with high probability) the locations of all the nonzero entries of β that are above the noise floor: $\mathsf{LAR}_m \gtrsim \frac{k\log p}{n\cdot\mathsf{SNR}} \;\Rightarrow\; m \in \widehat{\mathcal{S}}$. Now consider instead design matrices that are not necessarily orthonormal but which have $\mu(X) \asymp n^{-1/2}$ and $\nu(X) \lesssim n^{-1}$ (cf. Table I). Then we have from Theorem 3 that OST identifies (with high probability) the locations of the nonzero entries of β whose energies are greater than both the noise power and the average signal energy per nonzero entry: $\mathsf{LAR}_m \gtrsim \max\left\{\frac{1}{\mathsf{SNR}}, 1\right\}\frac{k\log p}{n} \;\; \Rightarrow \;\; m \in \widehat{\mathcal{S}}$. In particular, under the assumption that either $\mathsf{MAR} = \Theta(1)$ (and since $\mathsf{MAR} \leq \mathsf{LAR}_m$) or $\mathsf{SNR} = O(1)$, this suggests that the OST in such situations performs in a near-optimal (oracle-like) fashion in the sense that it recovers (with high probability) the locations of all the nonzero entries of β that are above the noise floor *without* requiring the design matrix X to be an orthonormal basis.

B. Recovery of Sparse Signals Using One-Step Thresholding

The significance of the sparse-signal recovery results reported in this paper for OST can be best put into perspective by considering the case of the design matrix X being an approximately tight frame in the sense that $\|X\|_2 \approx \sqrt{p/n}$ (Section III provides a small list of many such design matrices). It then follows from Theorem 4 that if X satisfies the strong coherence property then OST exactly recovers any k-sparse vector β with high probability as long as $k \lessapprox \mu^{-2}\mathsf{MAR}$; in particular, if we assume that $\mathsf{MAR} = \Theta(1)$ then this condition reduces to $k \lessapprox \mu^{-2}$. On the other hand, low-complexity sparse-recovery algorithms such as subspace pursuit [23], CoSaMP [24], and iterative hard thresholding [25] all rely on the restricted isometry property [29]. Therefore, the guarantees provided in [23]–[25] for the case of generic design matrices are limited to k-sparse signals that satisfy $k \lesssim \mu^{-1}$, which is much weaker than the $k \lessapprox \mu^{-2}$ scaling claimed here.

We conclude this discussion by pointing out that it is established in [32] that basis pursuit [31] also recovers most k-sparse signals—albeit in $O(p^3 + np^2)$ time—using arbitrary tight frames as long as $k \lessapprox \mu^{-2}$. Nevertheless, the basic difference between that result and Theorem 4 is that [32] requires the phases of the nonzero entries of β to be statistically independent and uniformly distributed on the unit torus whereas we do not assume any statistical prior on the values of the nonzero entries of β. Because of this reason, note that the basis pursuit result does not provide any guarantees beyond $k \gtrsim \mu^{-1}$ for even the simple case of $\beta \in \mathbb{R}^p_+$.

References

[1] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," in *Constructive Approximation*, Mar. 1997, pp. 57–98.

[2] W. U. Bajwa, R. Calderbank, and S. Jafarpour, "Model selection: Two fundamental measures of coherence and their algorithmic significance," in *Proc. IEEE Intl. Symp. Information Theory (ISIT '10)*, Austin, TX, Jun. 2010, pp. 1568–1572.

[3] J. A. Tropp, "Norms of random submatrices and sparse approximation," in *C. R. Acad. Sci., Ser. I*, Paris, 2008, vol. 346, pp. 1271–1274.

[4] W. Alltop, "Complex sequences with low periodic correlations," *IEEE Trans. Inform. Theory*, vol. 26, no. 3, pp. 350–354, May 1980.

[5] T. Strohmer and R. Heath, "Grassmanian frames with applications to coding and communication," *Appl. Comput. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.

[6] L. Applebaum, S. D. Howard, S. Searle, and R. Calderbank, "Chirp sensing codes: Deterministic compressed sensing measurements for fast recovery," *Appl. Comput. Harmon. Anal.*, pp. 283–290, 2009.

[7] P. G. Casazza and M. Fickus, "Fourier transforms of finite chirps," *EURASIP J. Appl. Signal Process.*, pp. 1–7, 2006.

[8] R. Calderbank and S. Jafarpour, "Reed Muller sensing matrices and the LASSO," in *Proc. Sequences and Their Applications (SETA '10)*, 2010, pp. 442–463.

[9] C. L. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, no. 4, pp. 661–675, Nov. 1973.

[10] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[11] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[12] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *Ann. Statist.*, vol. 22, no. 4, pp. 1947–1975, 1994.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[14] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Machine Learning Res.*, vol. 7, pp. 2541–2563, 2006.

[15] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

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

[17] C. Genovese, J. Jin, and L. Wasserman, "Revisiting marginal regression," submitted. [Online]. Available: arXiv:0911.4080v1

[18] K. Schnass and P. Vandergheynst, "Average performance analysis for thresholding," *IEEE Signal Processing Lett.*, pp. 828–831, Nov. 2007.

[19] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.

[20] G. Reeves and M. Gastpar, "A note on optimal support recovery in compressed sensing," in *Proc. 43rd Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2009.

[21] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[22] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, pp. 3397–3415, Dec. 1993.

[23] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Th.*, pp. 2230–2249, May 2009.

[24] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[25] T. Blumensath and M. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, pp. 265–274, 2009.

[26] A. C. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin, "One sketch for all: Fast algorithms for compressed sensing," in *Proc. ACM Symp. Theory Computing (STOC '07)*, Jun. 2007, pp. 237–246.

[27] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss, "Near-optimal sparse Fourier representations via sampling," in *Proc. ACM Symp. Theory Computing (STOC '02)*, May 2002, pp. 152–161.

[28] M. Iwen, "A deterministic sub-linear time sparse Fourier algorithm via non-adaptive compressed sensing methods," in *Proc. 19tth Annu. ACM-SIAM Symp. Discrete Algorithms (SODA '08)*, Jan. 2008, pp. 20–29.

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

[30] L. Welch, "Lower bounds on the maximum cross correlation of signals," *IEEE Trans. Inform. Theory*, vol. 20, no. 3, pp. 397–399, May 1974.

[31] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Comput.*, pp. 33–61, Jan. 1998.

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

[33] W. U. Bajwa, R. Calderbank, and S. Jafarpour, "Why Gabor frames? Two fundamental measures of coherence and their role in model selection," *J. Commun. Netw.*, vol. 12, no. 4, pp. 289–307, Aug. 2010.

[34] M. Rudelson and R. Vershynin, "Non-asymptotic theory of random matrices: Extreme singular values," in *Proc. Int. Congr. of Mathematicians*, Hyderabad, India, Aug. 2010.

[35] R. Calderbank, S. Howard, and S. Jafarpour, "Construction of a large class of matrices satisfying a statistical isometry propery," *IEEE J. Select. Topics Signal Processing*, vol. 4, no. 2, pp. 358–374, Apr. 2010.

[36] J. Lawrence, G. E. Pfander, and D. Walnut, "Linear independence of Gabor systems in finite dimensional vector spaces," *J. Fourier Anal. Appl.*, vol. 11, no. 6, pp. 715–726, Dec. 2005.

[37] P. Delsarte and J. M. Goethals, "Alternating bilinear forms over GF(q)," *J. Combinatorial Theory, Ser. A*, vol. 19, pp. 26–50, 1975.

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

[39] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.