The generalized Lasso with non-linear observations

Yaniv Plan Roman Vershynin

***Abstract*—We study the problem of signal estimation from non-linear observations when the signal belongs to a low-dimensional set buried in a high-dimensional space. A rough heuristic often used in practice postulates that *non-linear* observations may be treated as *noisy linear* observations, and thus the signal may be estimated using the generalized Lasso. This is appealing because of the abundance of efficient, specialized solvers for this program. Just as noise may be diminished by projecting onto the lower dimensional space, the error from modeling non-linear observations with linear observations will be greatly reduced when using the signal structure in the reconstruction. We allow general signal structure, only assuming that the signal belongs to some set $K \subset \mathbb{R}^n$. We consider the single-index model of non-linearity. Our theory allows the non-linearity to be discontinuous, not one-to-one and even unknown. We assume a random Gaussian model for the measurement matrix, but allow the rows to have an unknown covariance matrix. As special cases of our results, we recover near-optimal theory for noisy linear observations, and also give the first theoretical accuracy guarantee for 1-bit compressed sensing with unknown covariance matrix of the measurement vectors.**

I. Introduction

Before describing to the non-linear setting which is the main theme of this paper, let us first consider the structured linear model

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$$

where an unknown vector $\boldsymbol{x}$ belongs to some known set $K \subset \mathbb{R}^n$. The goal is to reconstruct the signal $\boldsymbol{x}$ from the noisy measurement vector $\boldsymbol{y} \in \mathbb{R}^m$. A common method is to minimize the ℓ_2 loss subject to a structural constraint:

$$\text{minimize } \|\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{y}\|_2 \quad \text{subject to} \quad \boldsymbol{x}' \in K. \tag{I.1}$$

We shall refer to this generalized Lasso as the *K-Lasso* for the rest of the paper. The set K is meant to capture structure of the signal. In many cases of interest K behaves as if it were a *low-dimensional* set, although it often has full linear algebraic dimension. For example, to promote sparsity of the solution, one can choose K to be a scaled ℓ_1 ball, and this gives the vanilla Lasso as proposed by R. Tibshirani [44]. When the signals are matrices, to promote low rank one can choose K to be a scaled ball in the nuclear norm, and this is referred to as the matrix Lasso [9] or trace Lasso [22].

Y. Plan is with the Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, BC V6T 1Z2, Canada (e-mail: yaniv@math.ubc.ca).

R. Vershynin is with the Department of Mathematics, University of Michigan, 530 Church St., Ann Arbor, MI 48109, U.S.A (e-mail: romanv@umich.edu).

Manuscript received February 19, 2015. YP is partially supported by NSERC grant 22R23068. R. V. is partially supported by NSF grant 1265782 and U.S. Air Force grant FA9550-14-1-0009.

How well can the signal be reconstructed based on the complexity of the set K? Under the linear model, the last two decades have seen the development of a strong theoretical backing for the Lasso from the statistical community, mostly based on a sparsity assumption. See, e.g., [8], [23], [6], [29], [32], [46], [10]. Further, recent results developed from the compressed sensing community give a clean, comprehensive theory for arbitrary signal structure. See Section II.

Consider the more challenging situation, in which there is an unknown non-linearity in the observations. We ask:

> *What happens when the K-Lasso is used to reconstruct a signal based on non-linear observations?*

On the one hand, Lasso is by design a method for linear regression, and it is dubious to expect it to work if $\boldsymbol{y}$ depends non-linearly on $\boldsymbol{A}\boldsymbol{x}$. On the other hand, practitioners have been successfully using Lasso for non-linear (especially binary) observations without theoretical backing.

In this paper we demonstrate that K-Lasso can be used for non-linear observations. We will see that from Lasso's point of view, *non-linear* observations behave as scaled and *noisy linear* observations, and we will characterize the scaling and the noise. Furthermore, we assume $\boldsymbol{A}$ to be Gaussian, but in contrast to much of the literature, we allow *unknown covariance* of rows. A particular non-linearity of interest in signal processing is 1-bit quantization, which, when combined with sparse signal structure, leads to the model of 1-bit compressed sensing. We believe all previous theoretical results in this area have required knowledge of the covariance of rows for the recovery algorithm to be accurate; our work broadens the theory by removing this requirement. We will describe related literature regarding non-linear observations in Section II below.

A. Model

We will work with semiparametric single-index model of a similar form to the one in [37]. Let $\boldsymbol{x} \in K \subset \mathbb{R}^n$ be a fixed (unknown) *signal* vector, let $\boldsymbol{a}_i \sim \mathcal{N}(0, \Sigma)$ be independent random measurement vectors, and let $\boldsymbol{A}$ be the matrix whose i-th row is $\boldsymbol{a}_i^{\mathsf{T}}$. Let $f_i : \mathbb{R} \to \mathbb{R}$ be independent copies of an *unknown*, random function f modeling the non-linearity (it also may be deterministic), which are independent of $\boldsymbol{A}$. We assume that the m observations y_i that form the vector $\boldsymbol{y} = (y_1, \dots, y_m)$ take the form

$$y_i = f_i(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle). \tag{I.2}$$

Note that the norm of $\boldsymbol{x}$ is sacrificed in this model since it may be absorbed into the unknown random function f_i. Thus, to simplify presentation, we will assume that $\|\sqrt{\Sigma}\boldsymbol{x}\|_2 = 1$. We will remark on how to remove this assumption by a rescaling argument.

B. Examples

We now give two concrete examples of the above model: quantized and binary observations.

A first non-linearity of interest is quantization applied to linear observations. Then the function f maps $\langle a_i, x\rangle$ to a finite alphabet of real numbers. In this case, the non-linearity is known, and furthermore, it is designed. Thus, the theoretical error bounds we develop below may be tuned to optimize the error. This observation has been made in [43].

On the extreme end, one may consider *1-bit quantization:* $f(\langle a_i, x\rangle) = \text{sign}(\langle a_i, x\rangle)$. Measurements of this kind are of special interest due to the simplicity of hardware implementation, and the robustness to multiplicative errors. We further discuss 1-bit quantization in Section III below.

Interestingly, binary statistical models are quite similar. For example, $f(\langle a_i, x\rangle) = \text{sign}(\langle a_i, x\rangle + z_i)$ gives the *logistic regression* model, provided that z_i is logit noise. Other binary models are available by adjusting the distribution of z_i. The classical approach in these models is (regularized) maximum likelihood estimation [32], [14]. However, it requires knowledge of the form of the nonlinearity, which is equivalent to knowledge of the distribution of z_i, and in practice one would often not expect this to be known. Further, the theory requires the log-likelihood to be *strongly convex*, which ceases to hold when z_i is small compared to $||x||_2$. Ironically, the noise needs to be roughly larger than the signal in the theoretical treatment of maximum-likelihood estimation (see [14] for a discussion of this point). In contrast, as we show, the K-Lasso does not need knowledge of the non-linearity, and is accurate even when the noise z_i disappears, as in the 1-bit compressed sensing model.

C. Simplified results when K is a subspace

To begin in a simpler setting, let us assume that the covariance matrix Σ is identity, K is a d-dimensional subspace, and there is no non-linearity, just an unknown rescaling and noise. Thus, we assume that $f_i(u) = \mu u + z_i$ for $u \in \mathbb{R}$, where $\mu > 0$ and $z_i \sim \mathcal{N}(0, \sigma^2)$. Then the observations take the form

$$y_i = \mu\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle + z_i. \tag{I.3}$$

The K-Lasso (I.1) becomes the *least squares estimator* whose behavior is well known. Let $\widehat{\boldsymbol{x}}$ be the solution to the K-Lasso. Then, the conditional expectation of the squared error with respect to $\boldsymbol{A}$ satisfies

$$\mathbb{E}\,\|\widehat{\boldsymbol{x}} - \mu\boldsymbol{x}\|_2^2 = \sigma^2 \cdot \sum_{i=1}^{d} \frac{1}{\sigma_i^2(\boldsymbol{A}_K)}$$

where $\sigma_i(\boldsymbol{A}_K)$ is the i-th singular value of A restricted to the subspace K. Since $\boldsymbol{A}$ is Gaussian, it is well conditioned with high probability as long as the number of observations m is significantly larger than the dimension d of K [48]. In this case, with high probability, each singular value does not deviate significantly from $\sqrt{m}$ [48] and thus

$$\mathbb{E}\,\|\widehat{\boldsymbol{x}} - \mu\boldsymbol{x}\|_2^2 \approx \frac{d}{m}\sigma^2.$$

Let us make a few observations about the ingredients involved in the above calculation. First, the K-Lasso gives an estimate of a scaled version of $\boldsymbol{x}$. Second, note the vital requirement that the number of observations m exceeds the dimension of the subspace d. Third, observe that the size of the scaling and the noise satisfy

$$\mu = \mathbb{E}(f(g)\cdot g) \quad \text{and} \quad \sigma^2 = \mathbb{E}(f(g) - \mu g)^2 = \mathbb{E}\, f(g)^2 - \mu^2,$$

where g is a standard normal random variable.

Our main result states that up to a small extra summand, the K-Lasso gives the same accuracy for *non-linear* observations, with σ and μ measured in the same way. To easily compare, we first state this result when K is a subspace. Here and in the rest of the paper, a statement is said to hold with high probability if it holds with probability at least 0.99. Further, the symbol $\lesssim$ hides an absolute constant.

Proposition 1.1 (Non-linear estimation on a subspace): Suppose that $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I})$, and that $\boldsymbol{y}$ follows the semi-parametric single index model of Section I-A. Let K be a d-dimensional subspace and assume $\boldsymbol{x} \in K \cap S^{n-1}$. Suppose that

$$m \gtrsim d.$$

Then, with high probability, the non-linear estimator $\widehat{\boldsymbol{x}}$ which minimizes the K-Lasso (I.1) satisfies

$$\|\widehat{\boldsymbol{x}} - \mu\boldsymbol{x}\|_2 \lesssim \frac{\sqrt{d}\,\sigma + \eta}{\sqrt{m}} \tag{I.4}$$

where

$$\begin{aligned} &\mu := \mathbb{E}[f(g)\cdot g], \qquad \sigma^2 := \mathbb{E}(f(g) - \mu g)^2, \\ &\eta^2 := \mathbb{E}(f(g) - \mu g)^2 g^2. \end{aligned} \tag{I.5}$$

One sees that this mirrors the result for linear observations aside from the extra summand $\eta/\sqrt{m}$, which becomes quite small with a moderate number of observations m. For example, in the noisy linear model (I.3) one has $\eta = \sigma$, so this result gives the classic error rate as a special case.

Results of the above flavour have been rigorously proven in the statistics literature [7], with a focus on asymptotic behaviour of the error. In this paper, we extend these ideas to modern trends in signal processing and statistics, in which it is assumed that the signal belongs to some non-linear low-dimensional signal structure, such as the set of sparse vectors or low-rank matrices. We now proceed to our main results in which K will be allowed to be a general set.

D. Main results

We will give two results below, one specialized to the case when the scaled signal $\mu\boldsymbol{x}$ lies at an extreme point of K with (small) *tangent cone*, and one which only assumes that $\mu\boldsymbol{x}$ lies in K.

Definition 1.2 (Tangent cone): The tangent cone[1] of K at $\boldsymbol{x}$ is

$$D(K, \boldsymbol{x}) := \{\tau\boldsymbol{h} : \tau \geq 0, \boldsymbol{h} \in K - \boldsymbol{x}\}.$$

For sets with non-smooth boundary, such as the ℓ_1 ball or the nuclear norm ball, the tangent cone at a boundary point

[1]To allow non-convex K, the above is slight variation on the standard definition of tangent cone [26]. The tangent cone may also be called the *descent cone*.

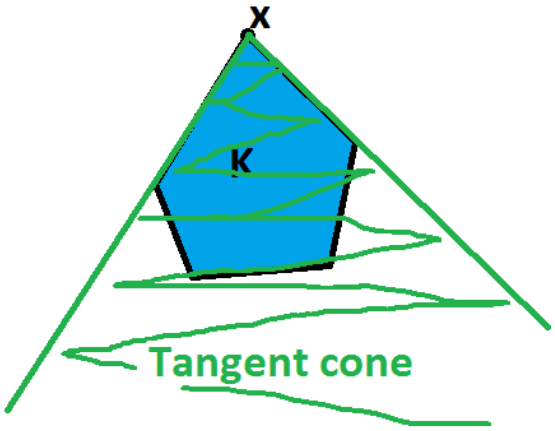

Fig. 1: The tangent cone

can be quite narrow, and intuitively should behave like a low-dimensional subspace. We give an illustrative example of a tangent cone in Figure 1, although, in a two-dimensional representation, we cannot do justice to the high-dimensional effects which allow convex sets to have extremely narrow tangent cones.

While $\boldsymbol{A}$ may be singular, it can be quite well conditioned when restricted to the tangent cone; it is not surprising that this restricted conditioning of $\boldsymbol{A}$ can determine the accuracy of the solution to (I.1). Further, this restricted condition number can be well understood via Gordon's escape through the mesh theorem (see Theorem 4.2). It states that the restriction of $\boldsymbol{A}$ onto K is well conditioned provided that the number of observations m exceeds the *effective dimension* of K. The effective dimension is measured in Gordon's theorem by the notion of Gaussian mean width. Let us recall the notion of the local (Gaussian) mean width; see [36], [37], [48] for further discussion of the mean width and how is serves as a measure of effective dimension.

Definition 1.3 (Local mean width): The *local mean width* of a subset $K \subset \mathbb{R}^n$ is a function of scale $t \geq 0$ defined as

$$w_t(K) = \mathbb{E} \sup_{\boldsymbol{x} \in K \cap tB_2} \langle \boldsymbol{x}, \boldsymbol{g} \rangle,$$

where B_2 denotes the unit Euclidean ball in $\mathbb{R}^n$.

Let us pause to explain the heuristic meaning of the local mean width of a cone D. The square of the mean width, $w_1(D)^2$, can be described as a measure of *effective dimension* of D. This can be seen on the following two examples. First, let D be d-dimensional subspace in $\mathbb{R}^n$. It is not difficult to check that

$$w_1(D)^2 \sim d,$$

up to a absolute multiplicative constants. Thus in this case, the square of the mean width is equivalent to the algebraic dimension d.

A deeper example is where $D = D(B_1^n, \boldsymbol{x})$ is the tangent cone of the unit ℓ_1 ball $B_1^n = \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u}\|_1 \leq 1\}$ at some point $\boldsymbol{x}$ on the boundary of B_1^n. Suppose $\boldsymbol{x}$ is s-sparse, meaning that $\boldsymbol{x}$ has s non-zero coordinates. It should be clear that the smaller sparsity s, the thinner the tangent cone D is. Quantitatively, this is captured by the notion of local mean width, which can be shown (see e.g. [11]) to behave as follows:

$$w_1(D)^2 \sim s \log(n/s).$$

Thus, up to a logarithmic factor, the square of the mean width is again equivalent to the dimensionality of the signal $\boldsymbol{x}$, which is its sparsity s.

We refer the reader to [36, Section 2] where the notion of mean width is discussed in more detail, as well as to [4] where an equivalent concept of *statistical dimension* is introduced.

Let us first state our first main result specialized to the case when $\Sigma = \boldsymbol{I}$ and to tangent-cone structure.

Theorem 1.4 (Accuracy with tangent cone structure): Suppose that $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I})$, $\boldsymbol{x} \in S^{n-1}$, and that $\boldsymbol{y}$ follows the semi-parametric single index model of Section I-A. Assume that $\mu \boldsymbol{x} \in K$, and let $d(K) := w_1(D(K, \mu \boldsymbol{x}))^2$. Suppose that

$$m \gtrsim d(K).$$

Then, with high probability, the solution $\widehat{\boldsymbol{x}}$ of the K-Lasso (I.1) satisfies

$$\|\widehat{\boldsymbol{x}} - \mu \boldsymbol{x}\|_2 \lesssim \frac{\sqrt{d(K)}\,\sigma + \eta}{\sqrt{m}} \tag{I.6}$$

where μ, η, and σ are defined in (I.5).

It should be clear that this result extends Proposition 1.1 from linear to non-linear observations, and from subspaces to general sets. To see this, recall our observation that if K is a d-dimensional subspace, then $d(K) \sim d$ up to an absolute constant factor.

Remark 1.5 (Boundary of K): For the above theorem to be especially useful, $\mu \boldsymbol{x}$ needs to lie on the boundary of K. Otherwise, the tangent cone is the entire $\mathbb{R}^n$, and the effective dimension $d(K)$ is of order of n. In this case, the estimate becomes accurate only when the number of observations m exceeds the ambient dimension n rather than the effective dimension of the cone, which may be significantly smaller. Thus, in practice, one would like to rescale K to put $\mu \boldsymbol{x}$ on the boundary. If $\mu \boldsymbol{x}$ does not lie precisely on the boundary, we may appeal to our more general Theorem 1.9 below. Further, we note that the unconstrained version of the K-Lasso overcomes this obstacle. This has been proven in the asymptotic setting in [43], which built upon the ideas in this papper.

A substitution argument generalizes the above result to allow an unknown covariance matrix.

Corollary 1.6 (Anisotropic measurement vectors): Suppose that $\boldsymbol{a}_i \sim \mathcal{N}(0, \Sigma)$, $\sqrt{\Sigma}\boldsymbol{x} \in S^{n-1}$, and that $\boldsymbol{y}$ follows the semi-parametric single index model of Section I-A. Assume that $\mu \boldsymbol{x} \in K$, and let $d(K, \Sigma) := w_1(\sqrt{\Sigma} D(K, \mu \boldsymbol{x}))^2$. Suppose that

$$m \gtrsim d(K, \Sigma).$$

Then, with high probability, the non-linear estimator $\widehat{\boldsymbol{x}}$ which minimizes the K-Lasso (I.1) satisfies

$$\|\sqrt{\Sigma}(\widehat{\boldsymbol{x}} - \mu \boldsymbol{x})\|_2 \lesssim \frac{\sqrt{d(K, \Sigma)}\,\sigma + \eta}{\sqrt{m}} \tag{I.7}$$

where μ, η, and σ are defined in (I.5).

Proof: We may set $\boldsymbol{a}_i := \sqrt{\Sigma}\boldsymbol{g}_i$ where $\boldsymbol{g}_i \sim \mathcal{N}(0, \boldsymbol{I})$. Then $\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle = \langle \boldsymbol{g}_i, \sqrt{\Sigma}\boldsymbol{x}\rangle$. Thus, by replacing $\boldsymbol{x}$ with $\sqrt{\Sigma}\boldsymbol{x}$, we recover the model in which $\Sigma = \boldsymbol{I}$. Further, we may substitute $\boldsymbol{x}'$ with $\sqrt{\Sigma}\boldsymbol{x}'$ in the K-Lasso to arrive at the $\sqrt{\Sigma}K$-Lasso:

$$\text{minimize } \|\boldsymbol{G}x' - y\|_2 \text{ subject to } x' \in \sqrt{\Sigma}K \tag{I.8}$$

where $\boldsymbol{G}$ is a matrix which contains $\boldsymbol{g}_i^{\mathsf{T}}$ as its i-th row. We have now completely reduced to the setup of Theorem 1.4, with the caveat that we have substituted $\boldsymbol{x}, \boldsymbol{x}'$, and K by $\sqrt{\Sigma}\boldsymbol{x}$, $\sqrt{\Sigma}\boldsymbol{x}'$, and $\sqrt{\Sigma}K$. Apply the theorem to finish the proof of the corollary. ∎

Remark 1.7 (Removing Σ from the mean width): If the covariance matrix Σ is well conditioned, its effect on the error (I.7) can be easily evaluated using the inequality

$$d(K, \Sigma) \le \text{cond}(\Sigma) \cdot d(K). \tag{I.9}$$

where $\text{cond}(\Sigma) = \|\Sigma\| \cdot \|\Sigma^{-1}\|$ denotes the condition number and $d(K) = d(K, \boldsymbol{I})$ is the same as Theorem 1.4. Before we prove this bound, let us mention that in some situations the effect of Σ is much smaller than it predicts – for example, if K is a subspace, then $d(K, \Sigma) = d(K)$.

To check (I.9), note that for the tangent cone $D = D(K, \mu x)$ we have

$$\begin{aligned} w_1(\sqrt{\Sigma}D) &= \mathbb{E} \sup_{\boldsymbol{x} \in \sqrt{\Sigma}D \cap B_2} \langle \boldsymbol{x}, \boldsymbol{g}\rangle \\ &\le \|\sqrt{\Sigma}^{-1}\| \cdot \mathbb{E} \sup_{\boldsymbol{x} \in \sqrt{\Sigma}(D \cap B_2)} \langle \boldsymbol{g}, \boldsymbol{x}\rangle, \end{aligned} \tag{I.10}$$

where the inequality follows from the elementary containment $\sqrt{\Sigma}D \cap B_2 \subset \|\sqrt{\Sigma}^{-1}\| \cdot \sqrt{\Sigma}(D \cap B_2)$. A straightforward application of Slepians inequality [48] then bounds the quantity in (I.10) by $\|\sqrt{\Sigma}^{-1}\| \cdot \|\sqrt{\Sigma}\| \cdot w_1(D)$. Thus, we conclude (I.9).

Remark 1.8 (Removing assumption that $\|\sqrt{\Sigma}\boldsymbol{x}\|_2 = 1$): The theory may be generalized to the case when $\|\Sigma\boldsymbol{x}\|_2 \neq 1$ with a simple rescaling argument. Let $\delta = \|\sqrt{\Sigma}\boldsymbol{x}\|_2$ and let $\tilde{\boldsymbol{x}} := \boldsymbol{x}/\delta$. Observe that

$$f(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle) = f(\delta\langle \boldsymbol{a}_i, \tilde{\boldsymbol{x}}\rangle) =: \tilde{f}(\langle \boldsymbol{a}_i, \tilde{\boldsymbol{x}}\rangle).$$

Thus, the theorem applies to the estimation of $\tilde{x}$ with parameters

$$\begin{aligned} &\mu := \mathbb{E}[\tilde{f}(g) \cdot g], \qquad \sigma^2 := \mathbb{E}(\tilde{f}(g) - \mu g)^2, \\ &\eta^2 := \mathbb{E}(\tilde{f}(g) - \mu g)^2 g^2. \end{aligned}$$

In some cases, one does not expect the tangent cone to have especially small mean width. As a motivating example, in the field of compressed sensing, it is standard to call $\boldsymbol{x}$ *compressible* if it belongs to a scaled ℓ_p ball for $p \in (0, 1)$, or if the ratio $\|\boldsymbol{x}\|_1/\|\boldsymbol{x}\|_2$ is small. In this case, which contrasts with the case of exact sparsity, the tangent cone may have mean width comparable to the ambient dimension. However, the set K itself can still behave in a low-dimensional fashion. Since K is not necessarily a cone, and is not scale invariant, it is necessary to characterize dimension with a scaling parameter. Fortunately, the local mean width accomplishes this task with t as the scaling parameter, and $w_t(K - \mu\boldsymbol{x})^2/t^2$ serving as a measure of the dimension at scale t.

The next theorem considers a general signal structure.

Theorem 1.9 (Accuracy without tangent cone structure): Suppose that $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I})$, $\boldsymbol{x} \in S^{n-1}$, and that $\boldsymbol{y}$ follows the semi-parametric single index model of Section I-A. Assume that $\mu\boldsymbol{x} \in K$ where K is convex,[2] and let $d_t(K) := w_t(K - \mu\boldsymbol{x})^2/t^2$. Then, the following holds with high probability. For any $t > 0$ such that

$$m \gtrsim d_t(K),$$

the non-linear estimator $\widehat{\boldsymbol{x}}$ which minimizes the K-Lasso (I.1) satisfies

$$\|\widehat{\boldsymbol{x}} - \mu\boldsymbol{x}\|_2 \lesssim \frac{\sqrt{d_t(K)}\,\sigma + \eta}{\sqrt{m}} + t \tag{I.11}$$

where μ, η, and σ are defined in (I.5).

Note that one may derive Theorem 1.4 by taking the limit as t goes to zero in the above theorem. However, in the proofs we will give a simpler and more straightforward route to the proof of Theorem 1.4.

Remark 1.10 (Non-trivial covariance matrix): As above, this result can be generalized to the case when the covariance matrix of the rows is $\Sigma \neq \boldsymbol{I}$. One would just define $d_t(K, \Sigma)$ in a straightforward way similar to that in Corollary 1.6.

E. Key idea in the proof

While it may be surprising that the K-Lasso is provably accurate even under the (non-linear) single-index model, it becomes much clearer when one observes that the expected loss, $\mathbb{E}\,||\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{y}||_2^2$, is minimized by $\mu\boldsymbol{x}$. In other words, regardless of the form of the non-linearity, the expected squared error is minimized by a multiple of the original signal. See Section IV for a proof.

In fact, one may transform the single-index model into a scaled linear model with an unusual noise term. Define an *induced* noise vector $\boldsymbol{z}$ to satisfy

$$\boldsymbol{y} = \boldsymbol{A}\mu\boldsymbol{x} + \boldsymbol{z}.$$

One may not expect $\boldsymbol{z}$ to play the role of noise, since it generally does not have zero mean, and is not independent of $\boldsymbol{A}$. However, z_i is uncorrelated with $\boldsymbol{a}_i$ (see Section IV).

We note that under this scaled linear model, one could use standard techniques to derive error bounds if $\boldsymbol{z}$ were deterministic, or independent of $\boldsymbol{A}$ [33], or if $\boldsymbol{z}$ were sub-Gaussian. However, since we make quite mild assumptions in our single-index model, only implicitly assuming that the parameters μ, σ, and η are well-defined, this induced noise may have heavy tails and requires novel analysis. Some of the tools for this analysis are available in the recent work [37] by the current authors and Yudovina. However, this earlier paper did not apply to the K-Lasso, and there were many technical details needed to extend these results. In particular, the extra steps in the proof of Theorem 1.9 are new ideas, as well as the method to give results with non-trivial covariance matrix. We give a detailed comparison with this earlier work and others in the next section.

[2]More generally, the proof only requires that $K - \mu\boldsymbol{x}$ be contained in a star shaped set. This star shaped set can take the place of $K - \mu\boldsymbol{x}$ in the results of this theorem.

II. Related literature

There is now a precise and comprehensive theory of signal reconstruction from *linear* observations, which takes into account signal structure. While it is largely motivated by the quite modern area of *compressed sensing* [18], [19], it is rooted in results developed in the older areas of *geometric functional analysis* [47], [21] and *convex integral geometry* [40]. To leverage these tools, it is vital to assume that the measurement matrix $\boldsymbol{A}$ is random. We give a brief overview of the results most closely aligned with this work. The literature that we describe below takes $\boldsymbol{A}$ to be a matrix with independent Gaussian or sub-Gaussian entries.

In the noiseless case, signal reconstruction is possible as soon as the number of observations exceeds the manifold dimension [17]. Even in the noisy case, there is a large pool of theory addressing signal reconstruction based on manifold dimension [5], [50], [51], [16]. However, in the noisy case, it is necessary to make extra structural assumption of the set K beyond assuming that it has small manifold dimension. Otherwise, signal reconstruction based on a number of observations comparable to the manifold dimension can be unstable [20].

The Gaussian mean width gives an alternative measure of dimension. When it is applicable, it leads to simpler assumptions. Indeed, as described above, the Gaussian mean width controls the conditioning of $\boldsymbol{A}$ when restricted to a cone, as proved in Gordon's escape through the mesh theorem. Rudelson and Vershynin [38] leveraged this result in the compressed sensing setup, showing that the signal could be reconstructed as long as the number of observations exceeded the squared Gaussian mean width of the tangent cone; Stojnic continued in this line of research [41]. Chandrasekaran et al. [11] extended this result to general convex bodies K. Amelunxen et al. [4] took a different route, synthesizing tools from conic integral geometry to give a precise phase transition for the number of observations needed to reconstruct $\boldsymbol{x}$. There work is based on the *statistical dimension*, which is roughly equivalent to the mean width, but has some extra convenient properties (see [4]). This showed that previous results were tight. A line of work by Thrampoulidis, Oymak, and Hassibi [33], [34], [42] concentrated on the precise reconstruction error from noisy observations, and also considered unconstrained versions of the K-Lasso. Our theoretical results in the non-linear case can be seen to mirror Theorem [33, Theorem 1] in the linear case. We state a simplified version of this theorem, specialized to Gaussian noise (see the original theorem for a very careful treatment of constants).

Theorem 2.1: Suppose that $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I})$, $\boldsymbol{x} \in S^{n-1}$, and that $\boldsymbol{y}$ follows the noisy linear model (I.3). Assume that $\mu \boldsymbol{x} \in K$, and let $d(K) := w_1(D(K, \boldsymbol{x}))^2$. Suppose that

$$m \gtrsim d(K).$$

Then, with high probability, the solution $\widehat{\boldsymbol{x}}$ of the K-Lasso (I.1) satisfies

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \lesssim \frac{\sqrt{d(K)}\,\sigma}{\sqrt{m}}.$$

Thus, one sees that our theorem 1.4, when specialized to linear observations, recovers this modern theory up to an absolute constant.

A. Prior work addressing non-linearity of the observations

There are also numerous works, and fields of study, addressing non-linearity. We describe the work that is most closely related to the present paper.

The semiparametric single-index model that we take in this paper is well studied in econometrics; see the monograph [24]. Most work in this area is asymptotic, although recent works have considered the finite case [25], [3], [13]. However, we believe that this literature does not address, from a theoretical standpoint, the gains that can be made by utilizing a general low-dimensional structure. See [37, Section 6] for a more thorough discussion of this literature.

In contrast, our work precisely characterizes the benefits from taking into account low-dimensional signal structure. For example, consider the sparse signal structure assumed in compressed sensing, in which $\boldsymbol{x}$ contains at most s non-zero entries. The effective dimension is $O(s \log(n/s))$ which can be significantly smaller than the ambient dimension, n. Thus, we show only $O(s \log(n/s))$ measurements are needed to estimate $\boldsymbol{x}$. Specialized to the case of linear, noiseless measurements, our theory recovers the classic result that $\boldsymbol{x}$ may be exactly reconstructed from this number of measurements. When non-linearity is present, the "noise" induced by modeling non-linear measurements with linear measurements is reduced proportionally to $s \log(n/s)/m$.

The area of *1-bit compressed sensing* [1] concentrates on the case when the non-linearity is 1-bit quantization. In other words, for $q \in \mathbb{R}$, $f(q) = \operatorname{sign}(q)$ or $f(q) = \operatorname{sign}(q + z)$ where z is noise. This has been a lively field of research for several years, in part due to a wide range of applicability in both signal processing problems and also statistical models in which the data is inherently binary. The discrete nature of this problem has led to new challenges that were not inherent in unquantized compressed sensing. Indeed, even the method of reconstruction of the signal has posed a challenge, and some of the proposed methods, such as the approach of [37] require knowledge of the covariance of the rows to be accurate. We believe our paper provides the first analysis of the K-Lasso for this problem, and the first theoretical result which allows non-trivial covariance of the rows of $\boldsymbol{A}$. In the next section, we specialize our work to the 1-bit compressed sensing model.

While there are numerous other publications which relate to various forms of non-linearity and low-dimensionality, there are three papers which we believe are most closely related to our results [37], [32], [28]. All three papers address general low-dimensional signal set K combined with general non-linearity. Our current result builds on the work in [37], which considers a very similar model. There are two significant extensions that we make beyond this work. First, our results are tighter in the sense that when specialized to the linear model, they match modern theory which is developed specifically for the linear model (see above). This is only true in [37] when the noise is larger than the signal. Further, as discussed above, the method espoused in [37] is not the K-Lasso, and requires knowledge of Σ to be effective.

The other two related works [32], [28] give a very general framework, which does not focus on the K-Lasso, but can be specialized to this recovery method. We believe that using the framework of [32], a theorem similar to our Theorem 1.4 could be derived. A key statistical idea, which is put rigorously in [32], is that the solution to the K-Lasso is a good estimate of the minimizer of the expected loss. In other words, mis-specification of the model is tolerable provided that the true signal minimizes the expected loss. See [49, Theorem 1] for a simplified version of this result. As we noted in Section I-E, $\mu\boldsymbol{x}$ is indeed the minimizer of the expected loss—this is the first step in our proofs, and could be used as a first step to derive error bounds from the framework of [32]. However, the results of [32] are general enough that such a derivation is non-trivial. Furthermore, we do not require *restricted strong convexity* in our Theorem 1.9 or *decomposability* in any of our theorems, which are two strong requirements of [32]. Similarly, by observing that $\mu\boldsymbol{x}$ minimizes expected loss, the results of [28] could be specialized to the K-Lasso. This would give a result similar to our Theorem 1.9. However, our result expands upon this in two ways: 1) In [28] it is assumed that y_i is sub-Gaussian, whereas we make almost no assumption on y_i—roughly, it only needs a bounded second moment; 2) In contrast to [28], our theory takes advantage of local structure of K around $\mu\boldsymbol{x}$, thus allowing, for example, the consideration of tangent cones. By doing this, our theory re-creates classical compressed sensing results as a special case, for example.

Finally, we would like to point to the new work [43] which considers the unconstrained version of the K-Lasso. By considering the asymptotic regime and adopting a stochastic model for signals $\boldsymbol{x}$, the authors of [43] were able to give a precise treatment of constants involved in the error bounds.

III. Specialization to 1-bit compressed sensing

As discussed above, the simplest 1-bit compressed sensing model takes the following form: For $q \in \mathbb{R}$, $f(q) = \operatorname{sign}(q)$, i.e., we just observe the sign of the linear observations. Let K be a scaling of the ℓ_1 ball and $\boldsymbol{x}$ is assumed to be s-sparse, i.e., to contain only s non-zero entries. This latter requirement implies that the tangent cone has small mean width. Indeed, as can be seen from [11] for instance, for the appropriate scaling of K, one has

$$d(K) = w_1(K - \mu\boldsymbol{x})^2 \lesssim s\log(n/s).$$

A straightforward calculation shows that

$$\mu = \sqrt{\frac{2}{\pi}}, \quad \sigma^2 = 1 - \frac{2}{\pi}, \quad \eta^2 = 1 - \frac{2}{\pi}.$$

Thus, Theorem 1.4 states that as long as $m = O(s\log(n/s))$ observations are observed, the K-Lasso gives accuracy

$$\|\widehat{\boldsymbol{x}} - \sqrt{\tfrac{2}{\pi}}\,\boldsymbol{x}\|_2 \lesssim \sqrt{\frac{s\log(n/s)}{m}}. \tag{III.1}$$

Moreover, this bound holds for observations with general covariance structure. Indeed, Corollary 1.6 combined with (I.9) imply that (III.1) remains true as long as Σ is reasonably well conditioned.

This yields the following surprising conclusion:

Even for highly non-linear observations, such as 1-bit quantization, the K-Lasso is quite accurate as long as the number of observations significantly exceeds the effective dimension of the signal.

IV. Proof of main results

We begin by setting

$$\boldsymbol{z} := \boldsymbol{y} - \boldsymbol{A}\mu\boldsymbol{x}.$$

While $\boldsymbol{z}$ is not independent of $\boldsymbol{A}$ or $\boldsymbol{x}$, and generally does not have mean 0, it will nevertheless play the role of noise. As shown in [37], $\boldsymbol{z}$ satisfies

$$\mathbb{E}\,\boldsymbol{A}^{\mathsf{T}}\boldsymbol{z} = 0. \tag{IV.1}$$

We repeat the derivation here to keep the paper self contained. It suffices to show that for any $\boldsymbol{v} \in S^{n-1}$, $\mathbb{E}\,\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{z} = 0$, which in turn would follow from

$$\mathbb{E}\,y_i\langle\boldsymbol{a}_i,\boldsymbol{v}\rangle - \mathbb{E}\,\mu\langle\boldsymbol{a}_i,\boldsymbol{v}\rangle\langle\boldsymbol{a}_i,\boldsymbol{x}\rangle = 0.$$

Since the covariance of $\boldsymbol{a}_i$ is identity, the second term is equal to $\mu\langle\boldsymbol{x},\boldsymbol{v}\rangle$. To calculate the first term, note that $g_i := \langle\boldsymbol{a}_i,\boldsymbol{x}\rangle$ has distribution $\mathcal{N}(0,1)$. Then make the Gaussian decomposition $\langle\boldsymbol{a}_i,\boldsymbol{v}\rangle = \langle\boldsymbol{x},\boldsymbol{v}\rangle g_i + g_i^{\perp}$ where $g_i^{\perp}$ is independent of g_i. By independence, the first term above is equal to

$$\begin{aligned}\mathbb{E}\,y_i\langle\boldsymbol{a}_i,\boldsymbol{v}\rangle &= \mathbb{E}\,f(g_i)[\langle\boldsymbol{x},\boldsymbol{v}\rangle g_i + g_i^{\perp}]\\ &= \langle\boldsymbol{x},\boldsymbol{v}\rangle\,\mathbb{E}\,f(g_i)g_i = \mu\langle\boldsymbol{x},\boldsymbol{v}\rangle\end{aligned}$$

where the first equality follows from our model assumption (I.2) that $y_i = f(g_i)$, and the last equality follows by definition of μ in (I.5). This completes the derivation of (IV.1).

Now let $\widehat{\boldsymbol{x}}$ be the solution of the K-Lasso (I.1), that is the minimizer of the loss function $\|\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{y}\|_2$ on K. We may replace this loss function by

$$L(\boldsymbol{x}') := \frac{1}{m}\left(\|\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{y}\|_2^2 - \|\boldsymbol{A}\mu\boldsymbol{x} - \boldsymbol{y}\|_2^2\right)$$

without affecting the minimizer $\widehat{\boldsymbol{x}}$. Indeed, $\mu\boldsymbol{x}$ is a fixed scalar multiple of a fixed signal, and thus we have only squared the loss function, subtracted a constant and multiplied by $1/m$. Now, the new loss function is very well-behaved in expectation.

Lemma 4.1 (Expected loss):

$$\mathbb{E}\,L(\boldsymbol{x}') = \|\boldsymbol{x}' - \mu\boldsymbol{x}\|_2^2.$$

Proof: Expanding $L(\boldsymbol{x}')$, we can express it more conveniently as

$$L(\boldsymbol{x}') = \frac{1}{m}\|\boldsymbol{A}\boldsymbol{h}\|^2 - \frac{2}{m}\langle\boldsymbol{h},\boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}\rangle \quad \text{where} \quad \boldsymbol{h} := \boldsymbol{x}' - \mu\boldsymbol{x}. \tag{IV.2}$$

The second term has zero mean according to (IV.1). Since the covariance matrix of $\boldsymbol{a}_i$ is identity, the first term is $\|\boldsymbol{h}\|_2^2$ in expectation, as desired. ∎

Lemma 4.1 implies that $\mu\boldsymbol{x}$ minimizes the *expected* loss. In order to prove the main theorem, we need to control the deviation from expectation of the two terms in the loss function (IV.2).

First, we lower bound the ratio of $\frac{1}{m}\|\boldsymbol{A}\boldsymbol{h}\|_2^2$ to its expectation value of $\|\boldsymbol{h}\|_2^2$. This can be done by applying the classical result from the work of Gordon [21].

Theorem 4.2 (Escape through the mesh): Let $D \subset \mathbb{R}^n$ be a cone. Then

$$\inf_{\boldsymbol{v}\in D\cap S^{n-1}} \|\boldsymbol{A}\boldsymbol{v}\|_2 \geq \sqrt{m-1} - w_1(D) - r \qquad \text{(IV.3)}$$

with probability at least $1 - e^{-r^2/2}$.

Next, we control the size of $\langle \boldsymbol{h}, \boldsymbol{A}^\mathsf{T}\boldsymbol{z}\rangle$.

Lemma 4.3: Let $D \subset tB_2^n$, and let $\boldsymbol{z} := \boldsymbol{y} - \boldsymbol{A}\mu\boldsymbol{x}$ as before. Then

$$\mathbb{E}\sup_{\boldsymbol{v}\in D} \langle \boldsymbol{v}, \boldsymbol{A}^\mathsf{T}\boldsymbol{z}\rangle \leq C\left(w(D)\sigma + t\eta\right)\sqrt{m}. \qquad \text{(IV.4)}$$

Here and in the rest of the argument, C, c refer to numerical constants; their values may differ from instance to instance. Before proving Lemma 4.3, we pause to show how the lemma and Theorem 4.2 imply our main result.

Proof of Theorem 1.4: For convenience, let us denote the spherical part of the tangent cone by $D = D(K, \mu\boldsymbol{x}) \cap S^{n-1}$. We begin by recording two events which occur with high probability. First, under the assumptions of our main Theorem 1.4, the escape through the mesh Theorem 4.2 implies that the following event holds with probability at least 0.995:

$$\text{Event 1:} \quad \inf_{\boldsymbol{v}\in D} \frac{1}{\sqrt{m}}\|\boldsymbol{A}\boldsymbol{v}\|_2 \geq c.$$

Second, Markov's inequality combined with Lemma 4.3 implies that the following event holds with probability at least 0.995:

$$\text{Event 2:} \quad \sup_{\boldsymbol{v}\in D} \langle \boldsymbol{v}, \boldsymbol{A}^\mathsf{T}\boldsymbol{z}\rangle \leq C\left(w(D)\sigma + \eta\right)\sqrt{m}.$$

By the union bound, both events hold together with probability at least 0.99. (We note in passing that the probability of success, and also the constant C in the bound of Event 2 could be sharpened using concentration inequalities. However, this would not change our final presentation.)

We now show how to bound the error vector $\boldsymbol{h} := \widehat{\boldsymbol{x}} - \mu\boldsymbol{x}$ in the intersection of these events. Since $\widehat{\boldsymbol{x}}$ minimizes the loss, we have

$$L(\widehat{\boldsymbol{x}}) \leq L(\mu\boldsymbol{x}) = 0.$$

Combine this with Equation (IV.2) to give

$$\frac{1}{m}\|\boldsymbol{A}\boldsymbol{h}\|^2 \leq \frac{2}{m}\langle \boldsymbol{h}, \boldsymbol{A}^\mathsf{T}\boldsymbol{z}\rangle. \qquad \text{(IV.5)}$$

On the other hand, $\boldsymbol{h}$ belongs to the tangent cone $D(K, \mu\boldsymbol{x})$, so $\boldsymbol{v} := \boldsymbol{h}/\|\boldsymbol{h}\|_2$ belongs to its spherical part $D = D(K, \mu\boldsymbol{x}) \cap S^{n-1}$. Then, by Events 1 and 2, we have

$$\frac{1}{m}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \geq c\|\boldsymbol{h}\|_2^2 \quad \text{and}$$
$$\langle \boldsymbol{h}, \boldsymbol{A}^\mathsf{T}\boldsymbol{z}\rangle \leq \|\boldsymbol{h}\|_2 \cdot C\left(w(D)\sigma + \eta\right)\sqrt{m}.$$

Combining these two inequalities with (IV.5), we obtain

$$c\|\boldsymbol{h}\|_2^2 \leq \frac{2}{m} \cdot \|\boldsymbol{h}\|_2 \cdot C\left(w(D)\sigma + \eta\right)\sqrt{m}.$$

Simplifying this bound, we complete the proof. ∎

We now prove Lemma 4.3.

Proof of Lemma 4.3: This proof has similar steps to the proof of Theorem 1.3 in [37]. We begin with a projection argument to (mostly) decouple $\boldsymbol{z}$ from $\boldsymbol{A}$. Let $\boldsymbol{P} := \boldsymbol{x}\boldsymbol{x}^\mathsf{T}$ be the orthogonal projection onto the span of $\boldsymbol{x}$ and let $\boldsymbol{P}^\perp := \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^\mathsf{T}$ be the projection onto the orthogonal complement. Then, convexity of the functional $\|\boldsymbol{u}\|_{D^\circ} := \sup_{\boldsymbol{v}\in D}\langle \boldsymbol{v}, \boldsymbol{u}\rangle$ leads to the following decomposition:

$$\mathbb{E}\,\|\boldsymbol{A}^\mathsf{T}\boldsymbol{z}\|_{D^\circ} \leq \mathbb{E}\,\|\boldsymbol{P}^\perp\boldsymbol{A}^\mathsf{T}\boldsymbol{z}\|_{D^\circ} + \mathbb{E}\,\|\boldsymbol{P}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}\|_{D^\circ} =: I + II.$$

We first control I. Note that, since $\boldsymbol{A}$ is Gaussian, $\boldsymbol{P}^\perp\boldsymbol{A}^\mathsf{T}$ is independent from $\boldsymbol{P}\boldsymbol{A}^\mathsf{T}$. It follows that $\boldsymbol{P}^\perp\boldsymbol{A}^\mathsf{T}$ is also independent of $\boldsymbol{z}$. Indeed, to obtain the latter conclusion, simply note that the columns of $\boldsymbol{P}\boldsymbol{A}^\mathsf{T}$ are $\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\boldsymbol{x}$, and the coordinates of $\boldsymbol{z}$ are

$$z_i = f(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle) - \mu\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle. \qquad \text{(IV.6)}$$

Therefore, $\boldsymbol{P}^\perp\boldsymbol{A}^\mathsf{T}\boldsymbol{z}$ is distributed identically with $\boldsymbol{P}^\perp\tilde{\boldsymbol{A}}^\mathsf{T}\boldsymbol{z}$, where $\tilde{\boldsymbol{A}}$ is an independent copy of $\boldsymbol{A}$ (independent also of $\boldsymbol{z}$). Thus

$$I = \mathbb{E}\,\|\boldsymbol{P}^\perp\boldsymbol{A}^\mathsf{T}\boldsymbol{z}\|_{D^\circ} = \mathbb{E}\,\|\boldsymbol{P}^\perp\tilde{\boldsymbol{A}}^\mathsf{T}\boldsymbol{z}\|_{D^\circ}$$
$$= \mathbb{E}\,\|(\boldsymbol{P}^\perp\tilde{\boldsymbol{A}}^\mathsf{T} + \mathbb{E}[\boldsymbol{P}\tilde{\boldsymbol{A}}^\mathsf{T}])\boldsymbol{z}\|_{D^\circ}.$$

Now, by Jensen's inequality, the last quantity is bounded by

$$\mathbb{E}\,\|(\boldsymbol{P}^\perp\tilde{\boldsymbol{A}}^\mathsf{T} + \boldsymbol{P}\tilde{\boldsymbol{A}}^\mathsf{T})\boldsymbol{z}\|_{D^\circ} = \mathbb{E}\,\|\tilde{\boldsymbol{A}}^\mathsf{T}\boldsymbol{z}\|_{D^\circ}.$$

Now condition on $\boldsymbol{z}$. Then $\tilde{\boldsymbol{A}}^\mathsf{T}\boldsymbol{z}$ has distribution $\|\boldsymbol{z}\|_2 \cdot \mathcal{N}(0, \boldsymbol{I})$. Thus

$$I \leq \mathbb{E}\,\|\tilde{\boldsymbol{A}}^\mathsf{T}\boldsymbol{z}\|_{D^\circ} = \mathbb{E}\,\|\boldsymbol{z}\|_2 \cdot w(D) \leq \sqrt{\mathbb{E}\,\|\boldsymbol{z}\|_2^2} \cdot w(D)$$
$$= \sqrt{m}\,\sigma \cdot w(D).$$

Here in the first equality we used the definition of $w(D)$; in the last equality, we recall (IV.6) and definition of σ from (I.5).

We now control II. Note that

$$\boldsymbol{P}\boldsymbol{A}^\mathsf{T}\boldsymbol{z} = \sum_{i=1}^m z_i\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\boldsymbol{x} = \sum_{i=1}^m \xi_i \cdot \boldsymbol{x}$$

where $\xi_i := z_i\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle = \big[f(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle) - \mu\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\big]\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle$. Thus,

$$II \leq \|\boldsymbol{x}\|_{D^\circ} \cdot \mathbb{E}\left|\sum_{i=1}^m \xi_i\right|.$$

Since $D \subset tB_2^n$, we have $\|\boldsymbol{x}\|_{D^\circ} \leq t$. Substituting this, we obtain

$$II \leq t\,\mathbb{E}\left|\sum_{i=1}^m \xi_i\right| \leq t\sqrt{\sum_{i=1}^m \mathbb{E}\,\xi_i^2} = t\sqrt{m\,\mathbb{E}\,\xi_1^2} = t\sqrt{m}\cdot\eta$$

where the last equality follows by definition of η from (I.5). The proof is complete. ∎

A. Proof of Theorem 1.9

When the error vector $\boldsymbol{h} = \widehat{\boldsymbol{x}} - \mu\boldsymbol{x}$ is not known to belong to a cone, but rather a general set, it can no longer be guaranteed that $\boldsymbol{h}$ is not in the null space of $\boldsymbol{A}$ (which was true for cones via Gordon's Theorem 4.2.) Nevertheless, such bad behaviour generally only occurs at tiny scales, and at large scales $\boldsymbol{A}$ may be quite well conditioned even on general sets. This idea

is made rigorous in the following lemma, which is known in the geometric functional analysis community even in more generality, see [39], [27], [31], [30], [45]. For the sake of the reader, we will include a proof below.

Lemma 4.4: Let $K \subset \mathbb{R}^n$ be a star shaped set.[3] Let $t > 0$ and suppose that $m \gtrsim w_t(K)^2/t^2$. Then, with probability at least $1 - 2\exp(-m/8)$, the following holds for all $\boldsymbol{v} \in K$ satisfying $\|\boldsymbol{v}\|_2 \geq t$:

$$\|\boldsymbol{A}\boldsymbol{v}\|_2 \geq c\sqrt{m}\|\boldsymbol{v}\|_2.$$

Before proving this lemma, let us combine it with Lemma 4.3 to prove the second main result.

Proof of Theorem 1.9: For convenience, let us denote $K_x := K - \mu\boldsymbol{x}$. As before, we begin by considering two good events, whose intersection holds with probability at least 0.99, based on Lemma 4.4 and Lemma 4.3.

$$\text{Event 1:} \quad \inf_{\boldsymbol{v}\in K_x \cap tB_2^c} \frac{1}{\sqrt{m}} \frac{\|\boldsymbol{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2} \geq c.$$

$$\text{Event 2:} \quad \sup_{\boldsymbol{v}\in K_x \cap tB_2} \langle \boldsymbol{v}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}\rangle \leq C\left(w_t(K_x)\sigma + t\eta\right)\sqrt{m}.$$

We now show how to bound the error vector $\boldsymbol{h} := \widehat{\boldsymbol{x}} - \mu\boldsymbol{x}$ in the intersection of these events. As in the proof of Theorem 1.4, the fact that $\widehat{\boldsymbol{x}}$ minimizes the loss implies that

$$\frac{1}{m}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \leq \frac{2}{m}\langle \boldsymbol{h}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}\rangle. \tag{IV.7}$$

We can assume that $\|\boldsymbol{h}\| \geq t$, since in the opposite case the error bound of Theorem 1.9 holds trivially. Since $\boldsymbol{h} \in K_x$, the inequality of Event 1 followed by (IV.7) gives

$$c^2\|\boldsymbol{h}\|_2^2 \leq \frac{2}{m}\langle \boldsymbol{h}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}\rangle. \tag{IV.8}$$

We would like to apply the inequality of Event 2, but cannot do this directly because $\|\boldsymbol{h}\|_2$ is not bounded above by t. Fortunately, since K_x is convex and contains the origin, K_x is star shaped. Using this fact, we may massage our bound into the form of Event 2 via a monotonicity argument.

Divide both sides of (IV.8) by $\delta := \|\boldsymbol{h}\|_2$. This gives

$$c^2\delta \leq \frac{2}{m}\delta^{-1}\langle \boldsymbol{h}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}\rangle \leq \frac{2}{m}\sup_{\boldsymbol{u}\in\delta^{-1}K_x\cap B_2}\langle \boldsymbol{u}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}\rangle =: f(\delta), \tag{IV.9}$$

where in the second inequality we set $\boldsymbol{u} = \delta^{-1}\boldsymbol{h}$ and used that $\boldsymbol{h} \in K_x$ again. Now, since K_x is star shaped, $f(\delta)$ is a monotonically decreasing function. Thus, by assumption $\delta \geq t$, we may replace δ by t in our bound, giving

$$c^2\|\boldsymbol{h}\|_2 \leq f(t) = \frac{2}{mt}\sup_{\boldsymbol{v}\in K_x\cap tB_2}\langle \boldsymbol{v}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}\rangle.$$

The proof is completed by applying the inequality of Event 2. ∎

It remains to prove Lemma 4.4.

Proof: We begin with the following simple comparison, which follows from the Cauchy-Schwartz inequality for all $\boldsymbol{v} \in \mathbb{R}^n$:

$$\|\boldsymbol{A}\boldsymbol{v}\|_2 \geq \frac{\|\boldsymbol{A}\boldsymbol{v}\|_1}{\sqrt{m}}. \tag{IV.10}$$

[3] K is a star shaped set if it satisfies $\lambda K \subset K$ for any $0 \leq \lambda \leq 1$.

Furthermore, since K is star shaped, we have

$$\inf_{\boldsymbol{v}\in K\cap tB_2^c}\frac{\|\boldsymbol{A}\boldsymbol{v}\|_1}{\|\boldsymbol{v}\|_2} = \inf_{\boldsymbol{u}\in K\cap tS^{n-1}}\frac{\|\boldsymbol{A}\boldsymbol{u}\|_1}{t}. \tag{IV.11}$$

(Indeed, $\boldsymbol{u} = t\boldsymbol{v}/\|\boldsymbol{v}\|_2$ lies in K since $t/\|\boldsymbol{v}\| \leq 1$ and K is star shaped.)

Next, we will control $\|\boldsymbol{A}\boldsymbol{u}\|_1$ with an application of the following uniform deviation inequality, which we proved in [35].

Lemma 4.5 (Uniform deviation for the ℓ_1 norm): Let $K \subset \mathbb{R}^n$ and let $r, t > 0$. Then, with probability at least $1 - 2\exp(-mr^2/t^2)$, the following holds for all $\boldsymbol{u} \in K$ satisfying $\|\boldsymbol{u}\|_2 \leq t$:

$$\left|\frac{1}{m}\|\boldsymbol{A}\boldsymbol{u}\|_1 - \sqrt{\frac{2}{\pi}}\,t\right| \leq \frac{4w_t(K)}{\sqrt{m}} + r.$$

Choosing $r = t/2$ in this lemma, we conclude that with probability at least $1 - \exp(-m/8)$, one has

$$\inf_{\boldsymbol{u}\in K\cap tS^{n-1}}\frac{1}{m}\|\boldsymbol{A}\boldsymbol{u}\|_1 \geq ct \quad \text{where} \quad c = \sqrt{\frac{2}{\pi}} - \frac{1}{2} - \frac{4w_t(K)}{t\sqrt{m}}. \tag{IV.12}$$

Recalling the assumption of Lemma 4.4 that $m \gtrsim w_t(K)^2/t^2$, we see that c is bounded below by a positive absolute constant. In this case, we can substitute the bound into (IV.11) to obtain

$$\inf_{\boldsymbol{v}\in K\cap tB_2^c}\frac{\|\boldsymbol{A}\boldsymbol{v}\|_1}{\|\boldsymbol{v}\|_2} \geq cm.$$

We finish the proof by an application of inequality (IV.10). ∎

V. Discussion

We have analyzed the K-Lasso for signal reconstruction from the semiparametric single-index model. We showed that the K-Lasso solution under the non-linear model $y_i = f(\langle a_i, x\rangle)$ behaves roughly like the K-Lasso solution under the noisy linear model $y_i = \mu x + \sigma z_i$ with $z_i \sim N(0,1)$, where $\mu = \mu(f)$ and $\sigma = \sigma(f)$ have simple expressions; the error of the K-Lasso is controlled by the local mean width of K. We hope this theoretical result may aid researchers who use the K-Lasso in situations when the response may not be linear. See [12] for one such implementation.

We have made some idealized assumptions in this paper thus allowing theoretical results that are simple to state and understand. There are many future directions of research both of theoretical and practical interest, particularly in softening assumptions, which we describe below.

We considered a Gaussian design matrix, A, and this allowed for a clean theoretical result. It is of interest to determine whether these results have some universality properties. Can the same kind of accuracy be expected for random non-Gaussian matrices? Under the linear model, universality results have been shown in the compressed sensing literature [15], that is, theoretical performance based on a Gaussian matrix is shown to empirically match the performance for many other kinds of matrices. However, there is an extra wrinkle under the single-index model: a universality result is impossible when x

is extremely sparse [2]. When A has independent sub-Gaussian entries, we conjecture that the results of our paper should still hold, although with an extra error term that becomes large when x is very sparse, and shrinks towards zero if x is spread out. It is of interest to iron out this theory and also to determine, both theoretically and empirically, how far these results may extend towards general design matrices.

Another direction of interest is robustness of the K-Lasso to model inaccuracies. Will the K-Lasso solution remain accurate if the single-index model is only approximately true, or if μx does not quite reside in K?

Finally, these results lead to new opportunities in signal processing problems in which the scientist has some control over the non-linearity f, e.g., for quantization (see [42]). In that case, the explicit expressions for $\mu(f)$ and $\sigma(f)$ may be tuned to optimize the error. It is of interest to identify other such problems, aside from quantization, that can benefit from this.

REFERENCES

[1] 1-bit compressive sensing webpage. http://dsp.rice.edu/1bitCS/.

[2] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[3] Pierre Alquier and Gérard Biau. Sparse single-index model. *The Journal of Machine Learning Research*, 14(1):243–280, 2013.

[4] Dennis Amelunxen, Martin Lotz, Michael McCoy, and Joel Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. 2013. Available at http://arxiv.org/abs/1303.6672.

[5] Richard Baraniuk and Michael Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[6] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[7] David R Brillinger. A generalized linear model with gaussian regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.

[8] Florentina Bunea, Alexandre B Tsybakov, Marten H Wegkamp, et al. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[9] Emmanuel Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[10] Emmanuel J Candès, Yaniv Plan, et al. Near-ideal model selection by 1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[11] Venkat Chandrasekaran, Benjamin Recht, Pablo Parrilo, and Alan Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[12] Stéphane Chrétien, Christophe Guyeux, Michael Boyer-Guittaut, Régis Delage-Mouroux, and Françoise Descôtes. Using the lasso for gene selection in bladder cancer data. 2015. Available at http://arxiv.org/abs/1504.05004.

[13] Arnak Dalalyan, Yuri Ingster, and Alexandre Tsybakov. Statistical inference in compound functional models. *Probability Theory and Related Fields*, pages 1–20, 2013.

[14] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[15] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[16] Armin Eftekhari and Michael Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. 2013. Available at http://arxiv.org/abs/1306.4748.

[17] Yonina Eldar, Deanna Needell, and Yaniv Plan. Uniqueness conditions for low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 33(2):309–314, 2012.

[18] Yonina C Eldar and Gitta Kutyniok, editors. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[19] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[20] Raja Giryes, Yaniv Plan, and Roman Vershynin. On the effective measure of dimension in the analysis cosparse model. 2014. Available at http://arxiv.org/abs/1410.0989.

[21] Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in r^n. *Geometric Aspects of Functional Analysis*, pages 84–106, 1988.

[22] Edouard Grave, Guillaume Obozinski, and Francis Bach. Trace lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems*, pages 2187–2195, 2011.

[23] Eitan Greenshtein et al. Best subset selection, persistence in high-dimensional statistical learning and optimization under l1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

[24] Joel Horowitz. *Semiparametric and nonparametric methods in econometrics*, volume 692. Springer, 2010.

[25] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

[26] Johannes Jahn. *Introduction to the theory of nonlinear optimization*. Springer Science & Business Media, 2007.

[27] B Klartag and S Mendelson. Empirical processes and random projections. *Journal of Functional Analysis*, 225(1):229–245, 2005.

[28] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds.

[29] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[30] Shahar Mendelson. Learning without concentration. 2014. Available at http://arxiv.org/abs/1401.0304.

[31] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

[32] Sahand Negahban, Pradeep Ravikumar, Martin Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[33] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. Simple bounds for noisy linear inverse problems with exact side information. 2013. Available at http://arxiv.org/abs/1312.0641.

[34] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. 2013. Available at http://arxiv.org/abs/1311.0830.

[35] Yaniv Plan and Roman Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, pages 1–24, 2011.

[36] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on*, 59(1):482–494, 2013.

[37] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. 2014.

[38] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[39] Gideon Schechtman. Two observations regarding embedding subsets of euclidean spaces in normed spaces. *Advances in Mathematics*, 200(1):125–135, 2006.

[40] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Springer, 2008.

[41] Mihailo Stojnic. Various thresholds for ℓ_1 optimization in compressed sensing. 2009. Available at http://arxiv.org/abs/0907.3666.

[42] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Simple error bounds for regularized noisy linear inverse problems. 2014. Available at http://arxiv.org/abs/1401.6578.

[43] Chrtistos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. The lasso with non-linear measurements is equivalent to one with linear measurements. 2015. Available at http://arxiv.org/abs/1506.02181.

[44] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[45] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. 2014. Available at http://arxiv.org/abs/1405.1102.

[46] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[47] R Vershynin. Lectures in geometric functional analysis. *Unpublished manuscript. Available at http://www-personal.umich.edu/~romanv/papers/GFA-book/GFA-book.pdf, year=2011*.

[48] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, Cambridge, 2012.

[49] Martin J Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

[50] Michael Wakin. Manifold-based signal recovery and parameter estimation from compressive measurements. 2010. Available at http://arxiv.org/abs/1002.1247.

[51] Han Lun Yap, Michael Wakin, and Christopher J Rozell. Stable manifold embeddings with operators satisfying the restricted isometry property. In *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, pages 1–6. IEEE, 2011.