Information and Inference: A Journal of the IMA (2017) **6**, 1–40
doi:10.1093/imaiai/iaw015
Advance Access Publication on 14 September 2016

High-dimensional estimation with geometric constraints

Y. PLAN† AND R. VERSHYNIN
Department of Mathematics, University of Michigan, 2074 East Hall, 530 Church Street, Ann Arbor, MI 48109, USA
†Corresponding author: yplan.work@gmail.com

AND

E. YUDOVINA
Department of Statistics, University of Michigan, 439 West Hall, 1085 South University Ave., Ann Arbor, MI 48109, USA

[Received on 6 October 2015; accepted on 26 April 2016]

Consider measuring a vector $x \in \mathbb{R}^n$ through the inner product with several measurement vectors, $a_1, a_2, \ldots, a_m$. It is common in both signal processing and statistics to assume the linear response model $y_i = \langle a_i, x \rangle + \varepsilon_i$, where ε_i is a noise term. However, in practice the precise relationship between the signal x and the observations y_i may not follow the linear model, and in some cases it may not even be known. To address this challenge, in this article we propose a general model where it is only assumed that *each observation y_i may depend on a_i only through $\langle a_i, x \rangle$*. We do not assume that the dependence is known. This is a form of the semiparametric-single index model, and it includes the linear model as well as many forms of the *generalized linear model* as special cases. We further assume that the signal x has some structure, and we formulate this as a general assumption that x belongs to some known (but arbitrary) feasible set $K \subseteq \mathbb{R}^n$. We carefully detail the benefit of using the signal structure to improve estimation. The theory is based on the *mean width* of K, a geometric parameter which can be used to understand its effective dimension in estimation problems. We determine a simple, efficient two-step procedure for estimating the signal based on this model—a linear estimation followed by metric projection onto K. We give general conditions under which the estimator is minimax optimal up to a constant. This leads to the intriguing conclusion that in the high noise regime, an unknown nonlinearity in the observations does not significantly reduce one's ability to determine the signal, even when the nonlinearity may be non-invertible. Our results may be specialized to understand the effect of nonlinearities in compressed sensing.

Keywords: high-dimensional inference; semiparametric single-index model; compressed sensing; matrix completion; mean width; dimension reduction.

1 Introduction

An important challenge in the analysis of high-dimensional data is to combine noisy observations—which individually give uncertain information—to give a precise estimate of a signal. One key to this endeavor is to utilize some form of structure of the signal. If this structure comes in a low-dimensional form, and the structure is aptly used, the resulting dimension reduction can find a needle of signal in a haystack of noise. This idea is behind many of the modern results in *compressed sensing* [19].

To make this more concrete, suppose one is given a series of measurement vectors a_i, paired with observations y_i. It is common to assume a linear data model $y_i = \langle a_i, x\rangle + \varepsilon_i$, relating the response to an unknown signal or parameter vector, x. However, in many real problems the linear model may not be justifiable or even plausible—consider binary observations. Such data can be approached with the *semiparametric single index model*, in which one models the data as

$$y_i = f(\langle a_i, x\rangle + \varepsilon_i)$$

for some unknown function $f: \mathbb{R} \to \mathbb{R}$. The goal is to estimate the signal x despite the unknown nonlinearity f. In fact, one may take a significantly more general approach as described in the next section.

It is often the case that some prior information is available on the structure of the signal x. Structure may come in many forms; special cases of interest include manifold structure, sparsity, low-rank matrices and compressible images. We consider a general model, assuming only that

$$x \in K \text{ for some closed star-shaped}^{[1]} \text{ set } K \subset \mathbb{R}^n.$$

This includes cones and convex sets containing the origin. We focus on three goals in this article:

(1) Determine general conditions under which the signal can be well estimated.

(2) Give an efficient method of estimation.

(3) Determine precisely what is gained by utilizing the feasible set K that encodes signal structure.

Let us outline the structure of this article. In the rest of this section, we carefully specify our model, describe the proposed estimator and analyze its performance for general K; the main result is Theorem 1.3 in Section 1.5. In Section 2, we specialize our results to a number of standard feasible sets K, including sparse vectors and low-rank matrices. In Section 3, we specialize our results to specific versions of the semiparametric single-index model. We include the linear model, the binary regression model and a model with explicit nonlinearity. In Section 4, we discuss the optimality of the estimator. We show that it is minimax optimal up to a constant under fairly general conditions. In Section 5.1, we illustrate the connection between our results and a deep classical estimate from *geometric functional analysis* called the *low M^* estimate*. This estimate tightly controls the diameter of a random section of an arbitrary set K. Section 6 gives an overview of the literature on the semiparametric single-index model, emphasizing the differences between the more geometric approach in this article and the classical approaches used in statistics and econometrics. We give some concluding remarks in Section 7. Sections 8–10 contain the technical proofs. In Section 8, we give the proofs of the results from Section 1, including our main result; in Section 9, we state and prove a version of our main result which holds with high probability rather than in expectation, and in Section 10 we give proofs of the optimality results from Section 4.

Throughout this article, we use notation common in the compressed sensing literature. However, because we expect our results to be of interest to a wider statistical audience, we provide a dictionary of notation in Section 6, Table 1.

[1] A set K is called star shaped if $\lambda K \subseteq K$, whenever $0 \le \lambda \le 1$.

1.1 *Model*

Let $x \in \mathbb{R}^n$ be a fixed (unknown) *signal* vector, and let $a_i \in \mathbb{R}^n$ be independent random measurement vectors. We assume that observations y_i are independent real-valued random variables such that

$$\text{each observation } y_i \text{ may depend on } a_i \text{ only through } \langle a_i, x \rangle. \tag{1.1}$$

In other words, we postulate that, given $\langle a_i, x \rangle$, the observation y_i and the measurement vector a_i are conditionally independent. We are interested in recovering the signal vector x using as few observations as possible. Furthermore, we will usually assume some *a priori* knowledge about x of the form $x \in K$ for some known set $K \subset \mathbb{R}^n$. Note that the norm of x is sacrificed in this model since it may be absorbed into the dependence of y_i on $\langle a_i, x \rangle$. Thus, it is of interest to estimate x up to a scaling factor.

Unless otherwise specified, we assume that a_i are independent standard normal vectors in $\mathbb{R}^n$. We lift this assumption in several places in the article: see the matrix completion problem in Section 2.5 and the lower bounds on all possible estimators in Section 4. We make further suggestions of how this assumption may be generalized in Section 7.

1.2 *Linear estimation*

The first and simplest approach to estimation is to ignore the feasible set K for a moment. One may then employ the following *linear estimator*

$$\widehat{x}_{\text{lin}} := \frac{1}{m} \sum_{i=1}^{m} y_i a_i.$$

It is not difficult to see that $\widehat{x}_{\text{lin}}$ is an unbiased estimator of the properly scaled vector x and to compute the mean squared error. This is the content of the following proposition whose proof we defer to Section 8.

PROPOSITION 1.1 (Linear estimation) Let $\bar{x} = x/\|x\|_2$. Then

$$\mathbb{E}\,\widehat{x}_{\text{lin}} = \mu \bar{x} \quad \text{and} \quad \mathbb{E}\,\|\widehat{x}_{\text{lin}} - \mu \bar{x}\|_2^2 = \frac{1}{m}\left[\sigma^2 + \eta^2(n-1)\right].$$

Here

$$\mu = \mathbb{E}\, y_1 \langle a_1, \bar{x} \rangle, \quad \sigma^2 = \text{Var}(y_1 \langle a_1, \bar{x} \rangle), \quad \eta^2 = \mathbb{E}\, y_1^2. \tag{1.2}$$

By rotation invariance of a_i, the parameters μ, σ and η depend on the magnitude $\|x\|_2$, but not on the direction $\bar{x} = x/\|x\|_2$ of the unknown vector x or on the number of observations m. These properties make it simple to compute or bound these parameters in many important cases, as will be clear from several examples below. For now, it is useful to think of these parameters as constants.

The second part of Proposition 1.1 essentially states that

$$\left[\mathbb{E}\,\|\widehat{x}_{\text{lin}} - \mu \bar{x}\|_2^2\right]^{1/2} \asymp \frac{1}{\sqrt{m}}\left[\sigma + \eta\sqrt{n}\right]. \tag{1.3}$$

We can express this informally as follows:

$$\text{Linear estimation is accurate for } m = O(n) \text{ observations.} \tag{1.4}$$

1.3 *Projection onto the feasible set*

Although estimate (1.3) on the accuracy of linear estimation is sharp, it can be significantly improved if some *prior information* is available about the vector x. A rigorous way to encode the prior information would be to assume that $x \in K$, where $K \subset \mathbb{R}^n$ is some fixed, closed and known *feasible set*. Because the scaling of x may be absorbed into the semiparametric single-index model, it is actually more natural to assume that

$$\mu\bar{x} \in K \tag{1.5}$$

in the notation of Proposition 1.1.[2]

Recall that $\widehat{x}_{\text{lin}}$ is an unbiased estimator of $\mu\bar{x}$, a vector that lies in K. So to incorporate the feasible set K into estimation, a natural step is to *metrically project* the linear estimator $\widehat{x}_{\text{lin}}$ onto K. In other words, we define

$$\widehat{x} = P_K(\widehat{x}_{\text{lin}}) = \arg\min_{z \in K} \|\widehat{x}_{\text{lin}} - z\|_2. \tag{1.6}$$

As we will see shortly, this nonlinear estimator outperforms the linear one, often by a big margin.

1.4 *Measuring the size of the feasible set by mean width*

The quality of the nonlinear estimator (1.6) should depend on the size of the feasible set K. It turns out that there is a simple geometric notion that captures the size of K for this purpose. This notion is the *local mean width*. At first reading, one may replace it with a slightly simpler concept of *global* mean width.

DEFINITION 1.2 (Mean width) The (global, Gaussian) *mean width* of a subset $K \subset \mathbb{R}^n$ is defined as

$$w(K) = \mathbb{E} \sup_{x,y \in K} \langle g, x - y \rangle,$$

where $g \sim N(0, I_n)$. The *local mean width* of a subset $K \subset \mathbb{R}^n$ is a function of scale $t \geq 0$ and is defined as

$$w_t(K) = \mathbb{E} \sup_{x,y \in K,\, \|x-y\|_2 \leq t} \langle g, x - y \rangle.$$

Note that $w_t(K) \leq w(K)$ trivially holds for all t. The concepts of mean width, both global and local, originate in geometric functional analysis and asymptotic convex geometry (see e.g. [23]). Quantities equivalent to mean width appear also in stochastic processes under the name of γ_2 functional (see [53]) and in statistical learning theory under the name of Gaussian complexity (see [5]).

More recently, the role of mean width (both local and global) was recognized in the area of signal recovery [1, 3, 37, 43, 45]. To interpret these developments as well as this article, it is often helpful to

[2] Passing between assumptions $x \in K$ and $\mu\bar{x} \in K$ in practice should be painless by rescaling K, as the scaling factor μ is usually easy to estimate. Further, if K is a cone, then there is no difference between the assumptions.

think of the square of the mean width of the properly scaled feasible set K as an *essential dimension* (as opposed to the algebraic dimension, which typically equals n). A notable benefit of this method of measuring dimension is that it is robust to perturbations: if K is slightly increased, the mean width only changes slightly.

1.5 *Main result*

THEOREM 1.3 (Non-linear estimation) Let $\bar{x} = x/\|x\|_2$. Assume that $\mu\bar{x} \in K$, where K is a fixed star-shaped closed subset of $\mathbb{R}^n$. Then the nonlinear estimator $\widehat{x}$ defined in (1.6) satisfies the following for every $t > 0$:

$$\mathbb{E}\,\|\widehat{x} - \mu\bar{x}\|_2 \le t + \frac{2}{\sqrt{m}}\left[\sigma + \eta\,\frac{w_t(K)}{t}\right]. \tag{1.7}$$

Here μ, σ and η are the numbers defined in (1.2).

In Section 4, we will give general conditions under which the error achieved in this theorem is minimax optimal up to a constant factor—thus the theorem gives a precise, non-asymptotic, characterization of the benefit of using the signal structure K to improve the estimation. In Section 9, we will state a version of Theorem 1.3 whose conclusion is valid with high probability rather than in expectation. In the next two sections we will simplify Theorem 1.3 in special cases, show how it is superior to Proposition 1.1 and illustrate it with a number of examples.

2 Feasible sets K: consequences and examples

In this section, we state a simpler version of Theorem 1.3, compare Theorem 1.3 with Proposition 1.1 and illustrate it with several classes of feasible sets K that may be of interest in applications.

2.1 *General sets: error bounds via global mean width*

Let us state a simpler conclusion of Theorem 1.3, in terms of global mean width $w(K)$ and without the parameter t.

Let K be an arbitrary compact star-shaped subset of $\mathbb{R}^n$. Replacing in (1.7) the local mean width $w_t(K)$ by the bigger quantity $w(K)$ and optimizing in t, we obtain

$$\mathbb{E}\,\|\widehat{x} - \mu\bar{x}\|_2 \le \frac{2\sigma}{\sqrt{m}} + 2\sqrt{2}\left[\frac{\eta\, w(K)}{\sqrt{m}}\right]^{1/2}. \tag{2.1}$$

While this conclusion is less precise than (1.7), it may be sufficient in some applications. We note the unusual rate $O(m^{-1/4})$ in the right-hand side, which nevertheless can be sharp for some signal structures such as the ℓ_1-ball. See Sections 2.6, 4.3 and also [48].

2.2 *General sets: comparison with linear estimation*

Let us compare the qualities of the linear and nonlinear estimators. Let K be an arbitrary star-shaped closed subset of $\mathbb{R}^n$. The definition of local mean width implies that

$$w_t(K) \le w(tB_2^n) = t\,w(B_2^n) = t\,\mathbb{E}\,\|g\|_2 \le t(\mathbb{E}\,\|g\|_2^2)^{1/2} = t\sqrt{n}. \tag{2.2}$$

(Here and below, B_2^n is the n-dimensional Euclidean ball.) Substituting this bound into (1.7) and letting $t \to 0$, we deduce the following estimate from Theorem 1.3:

$$\mathbb{E}\,\|\widehat{x} - \mu\bar{x}\|_2 \le \frac{2}{\sqrt{m}}\big[\sigma + \eta\sqrt{n}\big].$$

Note that this is the same upper bound as Proposition 1.1 gives for linear estimation, up to an absolute constant factor. We can express this conclusion informally as follows:

Projecting the linear estimator $\widehat{x}_{\text{lin}}$ onto the feasible set K can only improve the accuracy of estimation.

We will shortly see that such improvement is often significant.

2.3 *General cones*

Let K be a fixed closed cone in $\mathbb{R}^n$, so $tK = K$ is satisfied for all $t \ge 0$. Then

$$w_t(K) = t\, w_1(K).$$

Substituting this into (1.7) and letting $t \to 0$, we can state Theorem 1.3 in this case as follows.

THEOREM 2.1 (Estimation in a cone) Assume that $x \in K$, where K is a fixed closed cone in $\mathbb{R}^n$. Let $\bar{x} = x/\|x\|_2$. Then

$$\mathbb{E}\,\|\widehat{x} - \mu\bar{x}\|_2 \le \frac{2}{\sqrt{m}}\,[\sigma + \eta\, w_1(K)]. \tag{2.3}$$

Here μ, σ and η are the numbers defined in (1.2).

To further simplify the bound (2.3), note that term containing $w_1(K)$ essentially dominates there. This is based on the following observation.

LEMMA 2.2 For a non-empty cone K, one has $\mathbb{E}\, w_1(K) \ge \sqrt{2/\pi}$.

Proof. Note that $K - K$ contains a line, so $(K - K) \cap B_2^n$ contains a pair of antipodal points on the unit sphere. Therefore, $w_1(K) = w((K - K) \cap B_2^n)$ is bounded below by the first absolute moment of the standard normal distribution, which equals $\sqrt{2/\pi}$. □

Using Lemma 2.2, we see that (2.3) implies that

$$\mathbb{E}\,\|\widehat{x} - \mu\bar{x}\|_2 \le \gamma\,\frac{w_1(K)}{\sqrt{m}}, \quad \text{where} \quad \gamma = \sqrt{2\pi}\,\sigma + 2\eta. \tag{2.4}$$

In particular, we arrive at the following informal conclusion:

Linear estimation followed by projection onto the feasible cone K is accurate for $m = O(w_1(K)^2)$ observations.

REMARK 2.3 (Essential dimension) This result becomes especially transparent if we think of $w_1(K)^2$ as the *essential dimension* of the cone K. A good estimation is then guaranteed for *the number of observations proportional to the essential dimension* of K. This paradigm manifested itself in a number of recent results in the area of signal recovery [1, 3, 37, 43, 45]. Among these we especially note [3], where the results are explicitly stated in terms of a 'statistical dimension', a very close relative of $w_1(K)$.

REMARK 2.4 (Projecting onto S^{n-1}) Since magnitude information about x is irrecoverable in the semi-parametric model, and K has a conic, scale-invariant structure in this subsection, it is natural to rephrase our theorem in terms of the estimation of $\bar{x}$, unscaled. One may then normalize $\widehat{x}$ to come to the following conclusion:

$$\left\| \frac{\widehat{x}}{\|\widehat{x}\|_2} - \bar{x} \right\|_2 \leq \frac{2\gamma}{\mu} \cdot \frac{w_1(K)}{\sqrt{m}}. \tag{2.5}$$

Indeed, this follows because

$$\left\| \frac{\widehat{x}}{\|\widehat{x}\|_2} - \bar{x} \right\|_2 \leq \left\| \frac{\widehat{x}}{\|\widehat{x}\|_2} - \frac{\widehat{x}}{\mu} \right\|_2 + \left\| \frac{\widehat{x}}{\mu} - \bar{x} \right\|_2 \leq 2 \left\| \frac{\widehat{x}}{\mu} - \bar{x} \right\|_2,$$

where the last step follows since $\widehat{x}/\|\widehat{x}\|_2$ is the closest point in S^{n-1} to $\widehat{x}/\mu$.

A quick computation similar to (2.2) shows that the essential dimension of a cone is always bounded by its algebraic dimension. Thus the nonlinear estimator $\widehat{x}$ outperforms the linear estimator $\widehat{x}_{\text{lin}}$ discussed in (1.4), and the improvement is dramatic in cases where $w_1(K)^2 \ll n$. We give examples of such situations below.

2.4 *The set of sparse vectors*

Sparsity is a key signal structure used in many modern applications. For example, it is common to assume that only a small subset of coefficients are significant in a regression model. As another example, images are generally compressible in some dictionary, i.e. up to a linear transformation an image is sparse. The *compressed sensing* model (see the book by Eldar & Kutyniok [19]) is based on sparsity, and one may use the results in this paper as a treatment of unknown nonlinearities in compressed sensing.

The sparsity model takes $K = \{x \in \mathbb{R}^n : |\text{supp}(x)| \leq s\}$, i.e. K is the set of vectors with at most s non-zero entries. Fortunately, projecting onto K is computationally efficient: one only needs to retain the s largest entries of the vector, replacing all other entries by 0. This is referred to as *hard thresholding*. Thus, the projection estimator is quite amenable to large data.

Further, when s is significantly smaller than n, projecting onto K gives a fruitful dimension reduction. This is evident from the calculation

$$c\sqrt{s\log(2n/s)} \leq w_1(K) \leq C\sqrt{s\log(2n/s)}$$

which is given in Plan & Vershynin [45, Lemma 2.3]. Thus, $m \sim s\log(2n/s)$ observations are sufficient to estimate an s-sparse vector in $\mathbb{R}^n$.

Sparse in a dictionary. In many applications, the signal may be sparse in a dictionary, rather than canonically sparse, e.g. images are often compressible in a wavelet basis. In this case, the feasible set $K \subseteq \mathbb{R}^q$ is defined as

$$K = \{Dv : |\text{supp}(v)| \leq s\},$$

where the matrix $D \in \mathbb{R}^{q \times n}$ is referred to as a dictionary. A computation of the essential dimension of K can be made through a slight generalization of Plan & Vershynin [45, Lemma 2.3] (via a covering argument), which shows that once again

$$w_1(K) \leq C\sqrt{s \log(2n/s)}.$$

While projection onto K can pose a challenge, we will consider signals which are approximately sparse in a dictionary in Section 2.6; in this case projection can be done efficiently through convex programming.

2.5 *The set of low-rank matrices*

The low-rank signal structure is prevalent in statistical applications, thus leading to the popularity of *principal component analysis*; for many examples, see Kannan & Vempala [30]. Reconstructing a low-rank matrix from linear combinations of entries arises in various applications such as quantum state tomography or recommender systems [26, 49]. Recently, there has been quite a bit of interest and new theory on the problem of *matrix completion*: reconstruction of a low-rank matrix from a subset of its entries [10, 11], with binary nonlinearities given special consideration [18, 51]. In this subsection, we first specialize our theory to the low-rank structure and then adjust our theory to the matrix completion model.

Here we take the cone K to be the set of $d_1 \times d_2$ matrices with rank bounded by some small r. Projection onto K requires only taking the singular value decomposition and keeping the largest r singular values and singular vectors. This procedure is referred to as *hard thresholding of singular values*. The mean width of K is easy to estimate (see e.g. [45]); one has

$$w_1(K) \leq \sqrt{2r(d_1 + d_2)}.$$

Thus, $m \sim r(d_1 + d_2)$ observations are sufficient to estimate a $d_1 \times d_2$ matrix of rank r.

We now move away from the assumption of Gaussian measurement vectors in order to accommodate the matrix completion model. In this case, one observes a random sample of the entries of the matrix X (linear observations); that is, the measurement 'vectors' (matrices) a_i are uniformly distributed on the set of matrices with exactly one non-zero entry, and we take $y_i = \langle a_i, X \rangle$. Projection-based estimators have been considered for matrix completion in [12, 31, 32]. This article takes general models, which require a somewhat more complicated theory. We consider a simple projection-based estimator and prove that it is quite accurate under a simple model, recovering best-known theory for this model. The proof takes much in common with known matrix completion theory; the purpose of putting it here is to emphasize its shortness under a simple model. Further, this gives an intuition as to how to extend our results in general to non-Gaussian measurements.

Let us consider the following simple *matrix completion model.* Take $d_1 = d_2 = d$ and consider a random subset $\Omega \subset \{1, \ldots, d\} \times \{1, \ldots, d\}$, which includes each entry (i, j) with probability p, independently of all other entries. Let the observations y_i give the entries of the matrix X contained in

Ω. Further, assume that X satisfies the following *incoherence condition*: each entry of X is bounded by a parameter ζ.

Let us specify the *estimator* of X. Define the mask matrix Δ_Ω with 0, 1 entries by

$$(\Delta_\Omega)_{ij} = \mathbf{1}_{\{(i,j)\in\Omega\}}.$$

Then the linear estimator discussed in Section 1.2, $\frac{1}{m}\sum_{i=1}^m y_i a_i$, is naturally replaced with the Hadamard (entry-wise) product $\frac{1}{p}\Delta_\Omega \circ X$. Clearly, this is a rough estimator of X, but nevertheless it is an unbiased estimator. Fortunately, the projection step can significantly improve the rough estimator. As above, project $\frac{1}{p}\Delta_\Omega \circ X$ onto the set of rank-r matrices and call this $\widehat{X}$.

Under this model, we prove that the estimator is accurate, thereby recovering best-known theory for this model.

PROPOSITION 2.5 (Matrix completion accuracy) Consider the model described above. Let $m := \mathbb{E}\,|\Omega| = pd^2$ and assume that $m \geq d\log d$. The estimator $\widehat{X}$ has the following average error per entry:

$$\frac{1}{d}\,\mathbb{E}\,\|\widehat{X} - X\|_F \leq C\sqrt{\frac{rd}{m}}\,\zeta.$$

REMARK 2.6 (The benefit of projection) Despite the non-Gaussian form of the measurements, the estimation error in matrix completion is proportional to $\sqrt{w_1(K)/m} = \sqrt{rd/m}$.

REMARK 2.7 As is evident from the proof below, if the observations are corrupted with i.i.d. $N(0,\nu^2)$ noise, the error bound becomes

$$\frac{1}{d}\,\mathbb{E}\,\|\widehat{X} - X\|_F \leq C\sqrt{\frac{rd}{m}}\,(\zeta + \nu).$$

When $\nu \geq \zeta$, this error is known to be minimax optimal up to a constant.

The noise is easily incorporated and carried through the proof. At the end, it is simply necessary to replace the scaled Rademacher matrix $\zeta \cdot R$ by the sum of $\zeta \cdot R$ and a noise matrix.

Proof of Proposition 2.5. Note that $\operatorname{rank}(\widehat{X} - X) \leq 2r$. It follows that

$$\|\widehat{X} - X\|_F \leq \sqrt{2r}\|\widehat{X} - X\|.$$

The operator norm may be bounded as follows:

$$\|\widehat{X} - X\| \leq \|\widehat{X} - p^{-1}\Delta_\Omega \circ X\| + \|p^{-1}\Delta_\Omega \circ X - X\| \leq 2p^{-1}\|\Delta_\Omega \circ X - pX\|,$$

where the last step follows since $\widehat{X}$ is the closest rank-r matrix to $p^{-1}\Delta_\Omega \circ X$ in operator norm. The right-hand side is the operator norm of a matrix with independent, mean-zero entries, and may be controlled directly in many ways. To give an optimal result without log factors, one may use techniques from empirical process theory and random matrix theory. By *symmetrization* [35, Lemma 6.3] followed by the *contraction principle* [35, Theorem 4.4],

$$\mathbb{E}\,\|\Delta_\Omega \circ X - pX\| \leq 2\,\mathbb{E}\,\|R \circ \Delta_\Omega \circ X\| \leq 2\zeta\,\mathbb{E}\,\|R \circ \Delta_\Omega\|,$$

where R is a matrix whose entries are independent Rademacher random variables (i.e. $R_{ij} = 1$ or -1 with probability $1/2$). Thus, $R \circ \Delta_\Omega$ has independent, identically distributed entries. Seginer's [50] theorem then gives

$$\mathbb{E}\,\|R \circ \Delta_\Omega\| \leq C\sqrt{pd},$$

where C is a numerical constant. For the above inequality to hold, we must have $p \geq \log(d)/d$, i.e. on average at least $\log(d)$ observations per row. Putting all of this together gives

$$\mathbb{E}\,\|\widehat{X} - X\|_F \leq C\sqrt{\frac{rd}{p}}\zeta.$$

Divide through by d and recall $m = pd^2$ to conclude. □

2.6 *The set of approximately sparse vectors: ℓ_1-ball*

In real applications, exact sparsity is unusual and is usually replaced with approximate sparsity. A clean way to do this is to assume that the signal belongs to a scaled ℓ_p ball in $\mathbb{R}^n$ with $0 < p \leq 1$. The ℓ_1-ball, denoted B_1^n, is an especially useful signal structure due to its convexity [see, e.g. 45].

To see the connection between sparsity and the ℓ_1-ball, consider an s-sparse vector x, with Euclidean norm bounded by 1. Then, by Cauchy–Schwartz inequality, $\|x\|_1 \leq \sqrt{s}\,\|x\|_2 \leq \sqrt{s}$, and thus $x \in \sqrt{s}B_1^n$. However, one may take a slight perturbation of x, without significantly changing the ℓ_1 norm, and thus the set accommodates approximately sparse vectors.

Let $K = \sqrt{s}B_1^n$. For simplicity, we compute the global (as opposed to local) mean width, which is

$$w(K) = \mathbb{E}\sup_{x \in 2\sqrt{s}B_1^n} \langle x, g\rangle = 2\sqrt{s}\,\mathbb{E}\,\|g\|_\infty \leq 4\sqrt{2s\log n}.$$

The last inequality comes from the well-known fact $\mathbb{E}\,\|g\|_\infty \leq 2\sqrt{2\log n}$. Plugging into (2.1) gives

$$\mathbb{E}\,\|\widehat{x} - \mu\bar{x}\|_2 \leq \frac{2\sigma}{\sqrt{m}} + 16\left[\frac{\eta\sqrt{s\log n}}{\sqrt{m}}\right]^{1/2}.$$

Thus, $m \sim s\log n$ observations are sufficient to estimate an approximately s-sparse vector in $\mathbb{R}^n$.

In Section 4.3, we give a careful treatment of the ℓ_1-ball using local mean width. We will do a (well-known) calculation showing that this improves the above error bound slightly (roughly, $\log n$ can be replaced by $\log(2n/s)$). We will find that the true error rate is minimax optimal for many models of the functional dependence of y_i on $\langle a_i, x\rangle$ when m is not too large ($m \leq C\sqrt{n}$). We emphasize the unusual dependence $m^{-1/4}$ in this non-asymptotic regime.

Approximately sparse in a dictionary. Consider a dictionary D (see Section 2.4). By replacing B_1^n, with DB_1^n, one has a set encoding approximate sparsity in the dictionary D. Fortunately, because of its convexity, projection onto this set may be performed efficiently. Further, it is straightforward to bound the mean width. Indeed, Slepian's inequality [35] gives

$$w(DB_1^n) \leq \|D\| \cdot w(B_1^n) \leq 4\|D\|\sqrt{2s\log n}.$$

In other words, if the operator norm of D is bounded, then the set DB_1^n has similar essential dimension—and therefore error bound—as the set B_1^n. Summarizing, we find that $m \sim s \log n$ observations are sufficient to estimate a vector that is s-sparse in a dictionary of n elements.

3 Observations y_i: consequences and examples

We specialize our results to some standard observation models in this section.

3.1 *Linear observations*

The simplest example of y_i are linear observations

$$y_i = \langle a_i, x \rangle. \tag{3.1}$$

The parameters in (1.2) are then

$$\mu = \|x\|_2, \quad \sigma = \sqrt{2}\,\|x\|_2, \quad \eta = \|x\|_2.$$

Substituting them into (1.7) and (2.4), we obtain the following result.

COROLLARY 3.1 (Estimation from linear observations) Assume that $x \in K$ where K is a star-shaped set in $\mathbb{R}^n$. Assume the observations y_i are given by (3.1). Then for every $t > 0$,

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le t + \frac{2\,\|x\|_2}{\sqrt{m}} \left[\sqrt{2} + \eta \frac{w_t(K)}{t} \right].$$

If K is a cone, then

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le C\,\frac{w_1(K)}{\sqrt{m}}\,\|x\|_2,$$

where $C = 2(\sqrt{\pi} + 1) \approx 5.54$.

Note that here we estimate the signal x itself, rather than its scaled version as in our previous results.

3.2 *Noisy linear observations*

A more general class of examples includes noisy linear observations of the form

$$y_i = \langle a_i, x \rangle + \varepsilon_i, \tag{3.2}$$

where ε_i are mean zero, variance ν^2, random variables which are independent of each other and of a_i. Clearly, such observations follow the single-index model (1.1).

A straightforward computation of the parameters in (1.2) shows that

$$\mu = \|x\|_2, \quad \sigma = \sqrt{2\|x\|_2^2 + \nu^2} \le \sqrt{2}\,\|x\|_2 + \nu, \quad \eta = \sqrt{\|x\|_2^2 + \nu^2} \le \|x\|_2 + \nu.$$

Substituting them into (1.7) and (2.4), we obtain the following result.

COROLLARY 3.2 (Estimation from noisy linear observations) Assume that $x \in K$ where K is a star-shaped subset of $\mathbb{R}^n$. Assume that the observations y_i are given by (3.2). Then

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le t + C\frac{\|x\|_2 + \nu}{\sqrt{m}}\left[1 + \frac{w_t(K)}{t}\right], \tag{3.3}$$

where $C = 2\sqrt{2} \approx 2.83$. If K is a cone then

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le C'(\|x\|_2 + \nu)\,\frac{w_1(K)}{\sqrt{m}},$$

where $C' = 2(\sqrt{\pi} + 1) \approx 5.54$.

REMARK 3.3 (Signal buried in noise) One observes that both the size of the signal, $\|x\|_2$, and the size of the noise, ν, contribute to the error bound. When the noise is larger than the signal, the estimate may still be quite accurate because of the dimension reduction gained by projecting onto K. In fact, we will show that the error is minimax optimal up to a multiplicative constant under general conditions on K in Section 4. When the signal is larger than the noise, the proposed estimator may not be optimal, depending on K. Indeed, if $K = \mathbb{R}^n$, $m \ge n$ and $\nu = 0$, the minimax error is 0. However, as a general theme in this article we concentrate on rough observations for which 0 error is impossible—noiseless linear observations being the exception to this rule. The case of noiseless linear observations are handled well by the generalized Lasso estimator, as analyzed in the follow-up paper [46]. In that case, the size of the signal does not appear in the error bound.

REMARK 3.4 (Generalizing assumptions) We also note that the corollary could be adapted to the case when the random variables ε_i expressing the noise are not mean zero and do depend on x (but only through $\langle a_i, x\rangle$).

3.3 *Non-linear observations*

A general class of examples satisfying the single-index model (1.1) consists of nonlinear observations of the form

$$y_i = f(\langle a_i, x\rangle). \tag{3.4}$$

Here $f : \mathbb{R} \to \mathbb{R}$ is a fixed link function, which may be unknown. In particular, f may be discontinuous and not one-to-one (like the sign function) or even non-monotonic in its argument (like the sine function).

The rotation invariance of a_i allows us to express the parameters in (1.2) as functions of f and $\|x\|_2$ only. Indeed,

$$\mu = \mathbb{E}f(g\|x\|_2)g, \quad \sigma^2 = \mathrm{Var}\left[f(g\|x\|_2)g\right], \quad \eta^2 = \mathbb{E}f(g\|x\|_2)^2, \tag{3.5}$$

where $g \sim N(0,1)$. Substituting this into Theorem 1.3 or Theorem 2.1, we can bound the error of estimation in terms of f, the magnitude of signal $\|x\|_2$ and the local mean width of K; we are leaving this to the interested reader.

It is important and surprising that our estimation procedure—computing $\widehat{x}$ from the observations y_i—does not depend on the function f. In other words,

Any knowledge of the nonlinearity f defining the observations y_i is not needed to estimate x from these observations.

The quality of estimation must of course depend on f, and the exact dependence is encoded in the parameters μ, σ and η in (3.5). However, one does not need to know f exactly to compute reasonable bounds for these parameters.

Roughly, σ and η, which play a similar role to a noise level, are well upper bounded if $f(g)$ does not have extremely heavy tails. On the other hand, μ which plays a role similar to signal strength, can be lower bounded when f is monotonically increasing (although this is not necessary). A notable exception is when f is an even function and $\mu = 0$. In that case, the method of signal estimation described in this article would be inappropriate. In Section 4.2, we give a more precise treatment bounding these parameters and thus derive uniform error bounds for a fairly large class of functions f.

3.4 *Generalized linear models*

An even more general class of examples for the single-index model (1.1) consists of the noisy nonlinear observations

$$y_i = f(\langle a_i, x\rangle + \varepsilon_i) + \delta_i.$$

Here ε_i are random variables independent of each other and of a_i, and the same is assumed about δ_i. (However, ε_i may be correlated with δ_i.) This is a variation on the *generalized linear model* (GLM), which assumes

$$\mathbb{E}[y_i|a_i] = f(\langle a_i, x\rangle)$$

for some known function f.[3] In contrast, we take the semiparametric approach and assume that the nonlinearity, f, is unknown. (On the other hand, GLMs typically assume more general noise models than we do; we discuss the connections between single-index models and GLMs further in Section 6.) To specialize Theorem 1.3 to this case, we need only control μ, σ and η.

3.5 *Binary observations and logistic regression*

An important class of examples is formed by binary observations—those satisfying the model (1.1) and such that

$$y_i \in \{-1, 1\}^n, \quad \mathbb{E}\, y_i = f(\langle a_i, x\rangle). \tag{3.6}$$

[3] The GLM also usually assumes that y_i belongs to an *exponential family*.

Denoting $g \sim N(0,1)$ (the standard normal variable), we can compute the parameters in (1.2) as follows:

$$\mu = \mathbb{E} f(g\|x\|)g, \quad \sigma^2 \leq 1, \quad \eta = 1.$$

(The computation for σ follows by replacing the variance by the second moment.) Substituting this into (2.4), we obtain the following result.

COROLLARY 3.5 (Estimating from binary observations) Assume that $x \in K$ where K is a fixed cone in $\mathbb{R}^n$. Assume the observations y_i satisfy (1.1) and (3.6). Let $\bar{x} = x/\|x\|_2$. Then

$$\mathbb{E}\,\|\widehat{x} - \mu\bar{x}\|_2 \leq C\,\frac{w_1(K)}{\sqrt{m}},$$

where $\mu = \mathbb{E} f(g\|x\|)g$ and $C = \sqrt{2\pi - 4} + 2 \approx 3.51$.

Binary observations are important in both signal processing and statistics. We outline the connections between our model and the literature below.

1-bit compressed sensing. The noiseless *1-bit compressed sensing* model [6] takes observations

$$y_i = \operatorname{sign}(\langle a_i, x\rangle).$$

These form a particular case for which Corollary 3.5 applies with $\mu = \mathbb{E}\,|g| = \sqrt{2/\pi}$. This type of observations is motivated as a way of understanding the combination of extreme quantization with low-dimensional or structured signals. The s-sparse signal set described in Section 2.4 is of key interest. We note that the magnitude of x is completely lost in the observations (even with the nonlinearity known), and thus it is standard in this literature to estimate the direction $\bar{x}$. We may normalize our estimator to give

$$\mathbb{E}\left\|\frac{\widehat{x}}{\|\widehat{x}\|_2} - \bar{x}\right\|_2 \leq C\sqrt{\frac{s\log(2n/s)}{m}}.$$

This recovers the following result from the literature [see 29, 45]: $m \sim s\log(2n/s)$ 1-bit observations are sufficient to estimate an s-sparse vector in $\mathbb{R}^n$.

Logistic regression. *Logistic regression* is a common statistical model for binary data and takes the form

$$y_i = \operatorname{sign}(\langle a_i, x\rangle + \varepsilon_i),$$

where ε_i is logit noise. We note that other forms of noise lead to other binary statistical models. For example, if ε_i is Gaussian, this recovers the *probit* model. There is a recent influx of statistical literature on combining sparsity with binary observations, see [4] and references therein. The standard method is ℓ_1-penalized maximum likelihood estimation. To perform maximum likelihood estimation, it is vital to know the likelihood function—this is equivalent to knowing the form of the noise. However, in practice, given binary observations, it is often unclear which binary model to choose, and one is chosen arbitrarily. In this article, we emphasize that estimation can be done accurately without precise knowledge of relationship between $\langle a_i, x\rangle$ and y_i.

4 Optimality

In this section, we determine general conditions under which the projection estimator gives an optimal result up to a numerical constant, i.e. there is no estimator which significantly improves on the projection estimator. We begin by considering the noisy linear model described in (3.2). This will automatically give a lower bound in the case when the observations include a nonlinearity. We will come to the following intriguing conclusion.

When the measurements are noisy, an unknown, non-invertible nonlinearity in the measurements often does not significantly decrease one's ability to estimate the signal.

4.1 *Lower bound in the linear model*

We begin by considering the linear model with Gaussian noise

$$y_i = \langle a_i, x \rangle + \varepsilon_i, \tag{4.1}$$

where the noise variables $\varepsilon_i \sim N(0, \nu^2)$ are independent of each other and the a_i. (The parameter ν is the level of noise.) For compact notation, let $y \in \mathbb{R}^m$ be the vector of observations and $A \in \mathbb{R}^{m \times n}$ be the matrix whose ith row is a_i^{T}.

Our goal is to determine conditions on K and the noise which imply that the projection estimator is minimax up to a numerical constant. For simplicity, we refer to numerical constants as C if they are greater than 1 and c if they are less than 1. These constants may change from instance to instance, but each is bounded by an absolute numerical value. We refer to the projection estimator (1.6) as $\widehat{x}_{\text{proj}}$ in this section.

The *local packing number*, which we define as follows, plays a key role in estimation error.

DEFINITION 4.1 (Local packing number, P_t) Given a set $K \subset \mathbb{R}^n$, the local packing number P_t is the packing number[4] of $K \cap tB_2^n$ with balls of radius $t/10$.

We now give a lower bound on the minimax error in the noisy linear model.

THEOREM 4.2 Assume that $x \in K$, where K is a star-shaped subset of $\mathbb{R}^n$. Assume that the observations y_i are given by 4.1. Let

$$\delta_* := \inf_{t>0} \left\{ t + \frac{\nu}{\sqrt{m}} \left[1 + \sqrt{\log P_t} \right] \right\}. \tag{4.2}$$

Then there exists an absolute constant $c > 0$ such that any estimator $\widehat{x}$ which depends only on the observations y_i and measurements a_i satisfies

$$\sup_{x \in K} \mathbb{E} \, \|\widehat{x} - x\|_2 \geq c \min(\delta_*, \operatorname{diam}(K)).$$

[4] Given a set $K \subset \mathbb{R}^n$ and a scalar $t > 0$, a packing of K with balls of radius t is a set $\mathcal{X} \subset K$ satisfying $\|v - w\|_2 \geq t$ for each pair of distinct vectors $v, w \in \mathcal{X}$. The packing number is the cardinality of the largest such packing.

Let us compare this to the upper bound achieved by the projection estimator, as derived in Corollary 3.2:

$$\mathbb{E}\left\|\widehat{x}_{\text{proj}} - x\right\|_2 \leq C \inf_{t>0}\left\{t + \frac{\nu + \|x\|_2}{\sqrt{m}}\left[1 + \frac{w_t(K)}{t}\right]\right\}.$$

We now determine conditions under which the two match up to a constant. Observe that the two match most closely when the noise is large. In particular, when $\nu \geq \|x\|_2$, we have the simplified error bound:

$$\mathbb{E}\left\|\widehat{x}_{\text{proj}} - x\right\|_2 \leq C' \inf_{t>0}\left\{t + \frac{\nu}{\sqrt{m}}\left[1 + \frac{w_t(K)}{t}\right]\right\} =: C'\delta^*.$$

Further, since the estimator projects onto K, the upper bound may be tightened:

$$\mathbb{E}\left\|\widehat{x}_{\text{proj}} - x\right\|_2 \leq C' \min(\delta^*, \text{diam}(K)).$$

The only difference between lower and upper bounds is that $\sqrt{\log P_t}$ is replaced by $w_t(K)/t$. While these quantities may be in general different, for many sets they are comparable. To compare them, let us introduce the following geometric parameter:

DEFINITION 4.3 (Ratio: packing to local mean width) Define α as

$$\alpha = \alpha(K) = \sup_{t>0} \frac{w_t(K)}{t\sqrt{\log P_t}}. \tag{4.3}$$

It follows by definition of α that the lower bound of Theorem 4.2 satisfies

$$\delta_* \geq \delta^*/\alpha.$$

The following corollary follows immediately.

COROLLARY 4.4 Assume that $x \in K$ where K is a star-shaped subset of $\mathbb{R}^n$. Assume that the observations y_i are given by 4.1. Let

$$\delta^* := \inf_t \left\{t + \frac{\nu}{\sqrt{m}}\left[1 + \frac{w_t(K)}{t}\right]\right\}. \tag{4.4}$$

Then any estimator $\widehat{x}$ satisfies

$$\sup_{x\in K} \mathbb{E}\,\|\widehat{x} - x\|_2 \geq c_\alpha \min(\delta^*, \text{diam}(K)),$$

where $c_\alpha = c/\alpha$ for the numerical constant c defined in Theorem 4.2 and α defined in (4.3).

Thus, in the high-noise regime, the difference between upper and lower bounds is a factor of (numerical constant times) α. Fortunately, for many sets of interest, α is itself bounded (on both sides) by a

numerical constant. In particular, this holds true for sparse vectors, low-rank matrices and the ℓ_1-ball, as described in Section 4.3. Thus, in this case, the minimax error satisfies

$$c \min(\delta^*, \operatorname{diam}(K)) \leq \inf_{\widehat{x}} \sup_{x \in K} \mathbb{E} \, \|\widehat{x} - x\|_2 \leq C \min(\delta^*, \operatorname{diam}(K)).$$

REMARK 4.5 (Defining α at scale) If desired, α can be tightened by defining it at the relevant scale. Let $\delta_2^* = \frac{1}{2}\delta^*$. Then one can redefine α as

$$\alpha(K) = \frac{w_{\delta_2^*}(K)}{\delta_2^* \sqrt{\log P_{\delta_2^*}}} \tag{4.5}$$

and the result of the Corollary still holds.

REMARK 4.6 (General measurement vectors) Theorem 4.2 and Corollary 4.4 hold for a general class of random (or deterministic) measurement vectors. The theory only requires that A does not stretch out signals too much in any one direction, that is

$$\frac{1}{m} \| \mathbb{E} A^* A \| \leq 1.$$

Note that for the special case of a Gaussian measurement matrix $\frac{1}{m} \mathbb{E} A^* A = I_n$ and the above requirement is met; similarly if A contains i.i.d. entries with variance 1 or, more generally, if the rows are in *isotropic position*. If one wishes to generalize further, a look at the proof implies that we only require

$$\mathbb{E} \sup_{x \in K - K} \frac{\|Ax\|_2}{\|x\|_2} \leq 1.$$

Further, the number 1 on the right-hand side could be replaced by any numerical constant without significantly affecting the result.

REMARK 4.7 (Signal-to-noise ratio assumption) Our optimality assumption requires $\|x\|_2 \leq \nu$, which places a bound on the norm of the signal. To give a precise minimax accounting of the error, this assumption can be incorporated into Theorem 4.2 by replacing K by $K \cap \nu B_2^n$. In this case, the theorem implies the lower bound

$$\mathbb{E} \, \|\widehat{x} - x\|_2 \geq c \min(\delta, \operatorname{diam}(K), \nu).$$

Of course, if ν is known then this information can be taken into account in the projection estimator, giving it a mirroring error bound. One may check that the geometric assumption on K need not be adjusted. This follows from a rescaling argument and since K is star shaped.

4.2 *Optimality in the nonlinear model*

Now we extend our discussion of optimality to nonlinear observations. Let us now assume that the noisy linear data are passed through an unknown nonlinearity:

$$y_i = f(\langle a_i, x \rangle + \varepsilon_i), \tag{4.6}$$

where $\varepsilon_i \sim N(0, \nu^2)$. Note that the nonlinearity can only decrease one's ability to estimate x, and thus our lower bound still holds. It remains to determine conditions under which our upper bound matches. For easiest comparison, we rescale Theorem 1.3 to give the reconstruction error in estimating x.

THEOREM 4.8 (Non-linear estimation rescaled) Let $\lambda := \|x\|_2 / \mu$, so that $\mathbb{E}\, \lambda \widehat{x}_{\text{lin}} = \lambda \mu \bar{x} = x$. Assume that $x \in K' = \lambda K$, where K is a fixed star-shaped closed subset of $\mathbb{R}^n$. Then the nonlinear estimator $\widehat{x}$ defined in (1.6) satisfies the following for every $t > 0$:

$$\mathbb{E}\, \|\lambda \widehat{x} - x\|_2 \leq t + \frac{2\lambda}{\sqrt{m}} \left[\sigma + \eta \frac{w_t(K')}{t} \right]. \tag{4.7}$$

Here μ, σ and η are the numbers defined in (1.2).

Proof. The unscaled version, Theorem 1.3, gives

$$\mathbb{E}\, \|\widehat{x} - \mu \bar{x}\|_2 \leq t + \frac{2}{\sqrt{m}} \left[\sigma + \eta \frac{w_t(K)}{t} \right].$$

Now multiply both sides of the inequality by λ, substitute t for λt and note that

$$w_{\lambda t}(\lambda K) = \lambda w_t(K)$$

to complete the rescaling argument. □

Comparing to the estimation error from noisy linear observations (3.3), we find that the nonlinearity increases the error by at most a constant factor provided

$$\lambda(\sigma + \eta) \leq C(\|x\|_2 + \nu).$$

Below, we give some general conditions on f which imply this inequality for all signals x and noise levels ν.

LEMMA 4.9 (Conditions on the nonlinearity) Suppose that f is odd and non-decreasing. Further, suppose that f is sub-multiplicative on $\mathbb{R}^+$, that is $f(a \cdot b) \leq f(a) \cdot f(b)$ for all $a, b > 0$. Then, for the model defined in (4.6), the parameters $\lambda = \mu^{-1} \|x\|_2$, σ, η defined in (1.2) and the noise level ν satisfy

$$\lambda(\sigma + \eta) \leq C_f(\|x\|_2 + \nu), \tag{4.8}$$

where

$$C_f = C\, \mathbb{E}[f^4(g)]^{1/4}, \quad C = 48^{1/4} \mathbb{P}\left\{|g| \geq 1\right\} \approx 1.8,$$

with $g \sim N(0, 1)$. In particular, C_f does not depend on $\|x\|_2$ or ν.

Below are a few examples of nonlinearities f that satisfy the conditions of Lemma 4.9:

(1) The identity function, $f(x) = x$; $C_f \approx 2.36$.
(2) The sign function, $f(x) = \text{sign}(x)$; $C_f \approx 1.8$.
(3) Monomials of the form $f(x) = x^k$ for some odd natural number k; $C_f \approx 1.8M_{4k}$, where $M_{4k} = [(4k-1)!!]^{1/4k}$ is the $4k$th moment of the standard normal distribution.

REMARK 4.10 (Linear combinations) Note that the left-hand side of (4.8) remains unchanged if f is multiplied by a constant. Moreover, take k functions $f_1, \ldots, f_k$ satisfying the conditions of Lemma 4.9 and consider a nonlinear model of the form

$$y_i = \sum_{j=1}^{k} C_j f_j(\langle a_i, x\rangle + \varepsilon_i) \quad C_j > 0 \,\forall j,$$

where all C_j are positive (so that $C_j f_j$ is still monotonic). It is not hard to see that λ, σ, η and ν satisfy

$$\lambda(\sigma + \eta) \le \sqrt{k} \max_j C_{f_j}(\|x\|_2 + \nu),$$

where C_{f_j} are defined as in Lemma 4.9. This allows us to obtain uniform error bounds for finite linear combinations of any of the functions satisfying the conditions of Lemma 4.9, such as odd non-decreasing polynomials of degree at most k.

We summarize our findings in the following corollary:

COROLLARY 4.11 (Summary or optimality results) Consider the nonlinear model (4.6), and suppose the nonlinearity takes the form $f = \sum_{j=1}^{k} C_j f_j$, where the functions f_j are odd, non-decreasing and sub-multiplicative. Let C_{f_j} be the constants defined in Lemma 4.9. Let the signal set K satisfy $\alpha(K) < C_1$, for α defined in (4.5). Suppose the noise level ν is comparable to the signal level $\|x\|_2$: $\nu \ge c_2 \|x\|_2$. Define

$$\delta^* = \inf_t \left\{ t + \frac{\nu}{\sqrt{m}} \left[1 + \frac{w_t(K)}{t} \right] \right\}.$$

Let $\widehat{x}_{\text{proj}}$ be the projection estimator defined in (1.6), and let $\widehat{x}_{\text{opt}}$ be the optimal estimator of x. Then there exists an absolute numeric constant c and constants $C(f_j)$ depending only on f_j, such that

$$\left\|\widehat{x}_{\text{opt}} - x\right\|_2^2 \ge c \frac{c_2}{C_1} \min(\delta^*, \text{diam}(K))$$

and

$$\left\|\widehat{x}_{\text{proj}} - x\right\|_2^2 \le \sqrt{k} \left(\sum_{j=1}^{k} C_{f_j}^2 \right) \min(\delta^*, \text{diam}(K)).$$

4.3 *Sets satisfying the geometric condition*

In this section we show that for the three commonly used signal sets discussed in Sections 2.4–2.6, the relationship

$$\frac{w_t(K)}{t\sqrt{\log P_t}} \leq C \tag{4.9}$$

holds for all resolutions t (with $t < 1$ in the case of the ℓ_1-ball), for some universal constant C. Thus, $\alpha(K) \leq C$ for these sets. The condition (4.9) is essentially asserting that *Sudakov's minoration inequality* (see Ledoux & Talagrand [35, Theorem 3.18]) is *reversible* for $(K - K) \cap tB_2^n$ at scale $t/10$. Indeed, applying Sudakov's minoration inequality to the Gaussian process $X_x = \langle g, x\rangle$, $g \sim N(0, I_n)$, defined on $(K - K) \cap tB_2^n$ shows $\mathbb{E}\sup X_x = w_t(K) \geq ct\sqrt{\log P_t}$ for some numeric constant c. (Note that the packing number of $(K - K) \cap tB_2^n$ is at least as great as the packing number of $K \cap tB_2^n$.)

4.3.1 *Sparse vectors* Let $S_{n,s} \subset \mathbb{R}^n$ be the set of s-sparse vectors, i.e. the set of vectors in $\mathbb{R}^n$ with at most s non-zero entries. Note that $S_{n,s}$ is a cone, so if (4.9) is satisfied for $S_{n,s}$ at some level t, then it is satisfied at all levels. Thus, without loss of generality we take $t = 1$.

For $K = S_{n,s}$ we have $K - K = S_{n,2s}$. As shown in Section 2.4,

$$w_1^2(K) = w^2(S_{n,2s} \cap B_2^n) \leq Cs\log\frac{2n}{s},$$

for some constant C. On the other hand, we show that there exists a $1/10$-packing $\mathcal{X}_1$ of $S_{n,s} \cap B_2$ such that

$$\log|\mathcal{X}_1| \geq c\sqrt{s\log\frac{2n}{s}},$$

for some (possibly different) constant c. Such a packing exists when $s > n/4$, because then $S_{n,s}$ contains an $(n/4)$-dimensional unit ball, whose packing number is exponential in n. Thus, we may assume $s < n/4$. Consider the set Q of vectors x, where each $x \in Q$ has exactly s non-zero coordinates, and the non-zero coordinates are equal to exactly $s^{-1/2}$. Thus, $Q \subset S_{n,s} \cap B_2$ and $|Q| = \binom{n}{s}$. Q itself is not the packing, but we will show that we can pick a large subset $\mathcal{X}_1$ of Q such that any two vectors $x, y \in \mathcal{X}_1$ satisfy $\|x - y\|_2 > 1/10$.

Consider picking vectors from Q uniformly at random. For two uniformly chosen vectors $x, y \in Q$, we have

$$\mathbb{P}\left\{\|x - y\|_2^2 \leq \frac{1}{100}\right\} = \mathbb{P}\left\{x, y \text{ disagree on at most } \frac{s}{100} \text{ coordinates}\right\}.$$

This requires y to have at least $\frac{99}{100}s$ of the same non-zero coordinates as x. Given x, and assuming without loss of generality that $0.01s$ is an integer (rounding will not make a significant effect in the end), this happens for exactly

$$\binom{s}{0.99s}\binom{n - 0.99s}{0.01s} \text{ out of } \binom{n}{s}$$

values of $y \in Q$. Using Stirling's approximation (and the assumption $s < n/4$),

$$\mathbb{P}\left\{\|x - y\|_2 \leq \frac{1}{10}\right\} \leq \exp\left(-Cs\log\frac{2n}{s}\right).$$

Now let $\mathcal{X}_1 \subseteq Q$ contain $c\exp(s\log\frac{2n}{s})$ uniformly chosen elements of Q. Then with positive probability $\|x-y\|_2 > 1/10$ for all pairs $x, y \in \mathcal{X}_1$. In particular, there exists at least one $1/10$-packing $\mathcal{X}_1$ of $S_{n,s} \cap B_2$ of size $c\exp(s\log\frac{2n}{s})$.

4.3.2 *Low-rank matrices* Let $M_{d_1,d_2,r}$ be the set of $d_1 \times d_2$ matrices with rank (at most) r; clearly $r \leq \min(d_1, d_2)$. This is again a cone, so it suffices to show that (4.9) is satisfied at a single level $t = 1$. Now, for $K = M_{d_1,d_2,r}$, we have $K - K = M_{d_1,d_2,2r}$, so (see Plan & Vershynin [45, Section 3.3] or Section 2.5 of this article)

$$w_1^2(K-K) = w^2(M_{d_1,d_2,2r} \cap B_2^{d_1 d_2}) \leq C(d_1+d_2)r.$$

For the packing number, note that $M_{d_1,d_2,r}$ contains all matrices whose last $d_2 - r$ rows are identically zero, and also all matrices whose last $d_1 - r$ rows are identically zero. Thus, $M_{d_1,d_2,r}$ contains a copy of $\mathbb{R}^{d_1 r}$ and a copy of $\mathbb{R}^{d_2 r}$. Since the packing number of a unit ball in Euclidean space is exponential in its dimension, we conclude for the packing number of K

$$\log P_1 \geq c\max(d_1,d_2)r \geq \frac{c}{2}(d_1+d_2)r,$$

as required.

4.3.3 *Approximately sparse signals: ℓ_1-ball* For the set of approximately sparse signals contained in the ℓ_1-ball B_1^n, we will show that condition (4.9) is satisfied at all levels $t < 1$. Note that for $K = B_1^n$, $K - K \subset 2B_1^n$. We first derive precise bounds on the local mean width of K,

$$w_t^2(B_1^n) \leq w^2(2B_1^n \cap tB_2^n) \leq t^2 w^2(2t^{-1}B_1^n \cap B_2^n).$$

It is easy to see that, for all t, $w_t^2(B_1^n) \leq C'\min(t^2 n, \log n)$. The former will apply when t is very small ($t \leq Cn^{-1/2}$); the latter will apply when t is very large ($t \geq c$). In the intermediate regime $Cn^{-1/2} < t < c$, let $\sqrt{s} = 2t^{-1}$; then $c^{-2} < s < C^{-2}n$. Assume for simplicity that s is an integer; rounding s will not affect the results substantially. Now, by [44, Lemma 3.1]

$$\frac{1}{2}(\sqrt{s}B_1^n \cap B_2^n) \subset \operatorname{conv}(S_{n,s} \cap B_2^n),$$

and by [45, Lemma 2.3]

$$w^2\operatorname{conv}(S_{n,s} \cap B_2^n) = w^2(S_{n,s} \cap B_2^n) \leq C' s\log\frac{2n}{s}.$$

Consequently,

$$t^{-2}w_t^2(B_1^n) \leq \begin{cases} C'n, & t \leq Cn^{-1/2}; \\ C't^{-2}\log(nt), & Cn^{-1/2} < t < c; \\ C't^{-2}\log n, & c \leq t \leq 1. \end{cases}$$

We now consider packings of $B_1^n \cap tB_2^n$ by balls of radius $t/10$. Clearly, for $t < n^{-1/2}$, the packing number is at least exponential in n, because the set contains an n-dimensional ball of radius t; and for $c < t < 1$ the packing number is at least $2n$, because we can simply pack the vertices. It remains to treat the intermediate regime $n^{-1/2} < t < c$. Note that the problem is equivalent to packing $t^{-1}B_1^n \cap B_2^n$ by balls of radius $1/10$. Let $\sqrt{s} = t^{-1}$; assuming s is an integer, we have $S_{n,s} \subset t^{-1}B_1^n \cap B_2^n$, so we may apply the packing bounds for sparse vectors:

$$\log |\mathcal{X}_t| \geq cs \log \frac{2n}{s} = ct^{-2} \log(nt).$$

It is not hard to put these bounds together to conclude that indeed,

$$\log |\mathcal{X}_t| \geq \begin{cases} cn, & t \leq Cn^{-1/2}; \\ ct^{-2} \log(nt), & Cn^{-1/2} < t < c; \\ ct^{-2} \log n, & c \leq t \leq 1. \end{cases}$$

5 Related results: Low M^* estimate, Chatterjee's least squares estimate

5.1 *Low M^* estimate from geometric functional analysis*

The function

$$t \mapsto \frac{w_t(K)}{t}$$

that appears in Theorem 1.3 has been studied in geometric functional analysis. It is known to tightly control the *diameter of random sections* of K. An upper bound on the diameter is known as the *low M^* estimate* (see [25, 38, 39, 42]) and lower bounds have been established in [20–22, 24]. The following is a simplified version of the low M^* estimate, see [37].

THEOREM 5.1 (Low M^* estimate) There exists an absolute constant $c > 0$ such that the following holds. Let K be a star-shaped subset of $\mathbb{R}^n$. Let A be an $n \times m$ random matrix whose entries are independent $N(0, 1)$ random variables. Assume that $t > 0$ is such that

$$\frac{w_t(K)}{t} \leq c\sqrt{m}.$$

Then with probability at least $1 - \exp(-cm)$, we have

$$\|u - v\|_2 \leq t \quad \text{for any } u, v \in K,\ u - v \in \ker(A). \tag{5.1}$$

The conclusion (5.1) can be interpreted in a geometric way as follows. All sections of K parallel to the random subspace $E = \ker(A)$ have diameter at most t, i.e.

$$\operatorname{diam}(K \cap (E + v)) \leq t \quad \text{for all } v \in \mathbb{R}^n.$$

The relevance of the low M^* estimate to the estimation problems was first observed in [37]. Suppose one wants to estimate a vector $x \in K$ from linear observations as in Section 3.1, i.e.

$$y_i = \langle a_i, x \rangle \quad i = 1, \dots, m.$$

Choose an arbitrary vector $\widehat{x}_0 \in K$ which is consistent with the observations, i.e. such that

$$\langle a_i, \widehat{x}_0 \rangle = y_i \quad i = 1, \dots, m.$$

For convex feasible sets K, computing a vector $\widehat{x}_0$ this way is an algorithmically tractable convex feasibility problem.

To bound the error of this estimate, we can apply the low M^* estimate (5.1) for A the $m \times n$ matrix with rows a_i^{T}. Since $x, \widehat{x}_0 \in K$ and $x - \widehat{x}_0 \in \ker(A)$, it follows that with high probability, $\|\widehat{x}_0 - x\|_2 \le t$. To summarize,

$$\frac{w_t(K)}{t} \le c\sqrt{m} \quad \text{implies} \quad \|\widehat{x}_0 - x\|_2 \le t \text{ with high probability.} \tag{5.2}$$

In particular, if K is a cone then $w_t(K)/t = w_1(K)$, and we can let $t \to 0$. In this case, the low M^* estimate guarantees *exact recovery* once $w_1(K) \le c\sqrt{m}$.

To compare (5.2) with the error bound of the nonlinear estimation, we can state the conclusion of Theorem 1.3 as follows:

$$\frac{w_t(K)}{t} \le \varepsilon\sqrt{m} \quad \text{implies} \quad \|\widehat{x} - x\|_2 \le t + 2\varepsilon + \frac{2\sigma}{\sqrt{m}} \text{ with high probability.}$$

The two additional terms in the right-hand side can be explained by the fact that the exact recovery is impossible in the single-index model (1.1), in particular because of the noise and due to the unknown nonlinear dependence of y_i on $\langle a_i, x \rangle$.

5.2 *Chatterjee's least squares under convex constraint*

As this article was being written, [13] proposed a solution to a closely related problem. Suppose that an unknown vector $x \in \mathbb{R}^n$ lies in a known closed convex set K. We observe the noisy vector

$$y = x + g, \quad \text{where} \quad g \in N(0, I_n).$$

We would like to estimate x from y. This is very similar to the linear estimation problem (3.2), which can be written in the form $y = Ax + g$. An important difference is that in Chatterjee's model, A is the identity matrix; so one needs to take n observations of an n-dimensional vector x (given as coordinates y_i of y). In contrast, this article assumes that A is an $m \times n$ Gaussian matrix (essentially a projection), so one is allowed to take m observations where usually $m \ll n$.

The estimator proposed in [13] is the least-squares estimator, which clearly equals the metric projection of y onto K, that is

$$\widehat{x} = P_K(y).$$

Note that this coincides with the second step of our estimator. (The first step—a linear estimator—is clearly not needed in Chatterjee's model where A is identity.)

The performance of the least-squares estimator is quantified in [13] using a version of the local mean width of K; this highlights one more similarity with this article.

6 Connections to statistics and econometrics

The single-index model we consider is widely used in econometrics; the monograph of [27] gives a good introduction to the subject. In this section, we discuss connections and differences between our method and some of the literature. For this section only, we will use notation more common in statistics literature; Table 1 below summarizes the translation between the two worlds.

We note that much of the work on single-index models considers a broader formulation, namely

$$Y_i = f(\langle \beta, X_i \rangle) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i | X_i] = 0.$$

Our work relies crucially on the additional structural assumption that ε_i and X_i are conditionally independent given $\langle \beta, X_i \rangle$.

The single-index model is similar to GLMs in regression analysis and contains many of the widely used regression models; for example, linear, logistic or Poisson regressions. Our results apply directly to those models, with the added advantage that we do not need to know in advance which of the models is appropriate for a specific dataset. (Note especially that Corollary 4.11 guarantees uniform error bounds, e.g. for most cubic functions f.) In addition, Section 3.5 demonstrates the optimality of our approach for a rather general model of sparse binary-response data, which encompasses the logit or probit models for specific distributions of ε_i.

However, we remark that there is no containment between single-index models and GLMs; for example, the negative binomial model is a commonly used GLM which does not fit in the single-index framework. (There is also no reverse containment, as single-index models do not require the distribution of Y_i to come from an exponential family.)

The literature on single-index and more general semiparametric models is large and growing, so we can only outline some of the connections here. The primary appeal of single-index models to statisticians has been that it allows estimation of β at the parametric rate ($\sqrt{n}(\hat{\beta}_n - \beta)$ is tight), as well as estimation of f at the same rate as if the one-dimensional input $\langle \beta, X_i \rangle$ were observable. Estimation of an unknown function is notoriously difficult when the domain is high dimensional, so the latter feature is particularly attractive when p is large. We note that much of the analysis of these results is focused on the asymptotic

TABLE 1 *Dictionary of notation*

Quantity	Rest of paper	This section
Independent variable	$a_i \in \mathbb{R}^n$	$X_i \in \mathbb{R}^p$
Observation	$y_i \in \mathbb{R}$	$Y_i \in \mathbb{R}$
Index	$x \in \mathbb{R}^n$	$\beta \in \mathbb{R}^p$
Dimension of parameter space	n	p
Number of observations	m	n

rates of convergence, although some recent papers [2, 17, 28] also provide non-asymptotic (finite n) guarantees.

However, in contexts where the primary concern is with the index β, the classical approaches (such as [28, 47, 52]) can be unduly restrictive with regards to the link function. This is because they tend to be based on the *average derivative* method, i.e. the observation that

$$\frac{\partial\, \mathbb{E}[Y_1|X_1 = x]}{\partial x} = C\beta.$$

Naturally, methods based on this idea require the link function f to be (at least) differentiable, whereas we can allow $f(\cdot) = \text{sign}(\cdot)$.

To our knowledge, the only fundamentally different approach to recovery of β alone (without f) was taken by Li & Duan [36]. The authors considered observations X_i with an elliptically symmetric distribution (e.g. correlated normal). They then demonstrated under some mild conditions that any method based on minimization of a convex criterion function (e.g. least squares) would achieve one of its minima at $\beta^* \propto \beta$. Thus, essentially any method for which the minimizer happens to be unique can be used to consistently recover (a multiple of) β; under some additional assumptions, $\sqrt{n}(\hat{\beta}_n - \beta)$ is asymptotically normal. The major advantage of this method is that it does not in any way rely on the smoothness of the link function f.

The above discussion applies to low-dimensional settings, in which all that is known about β is $\beta \in \mathbb{R}^n$ (perhaps with $\|\beta\|_2 = 1$). Recently, a lot of work has been devoted to the analysis of high-dimensional data ($p > n$) with additional structural assumptions about β, for example sparsity. In the case of linear link function f, sparse high-dimensional regression is well studied (see Bühlmann & Van De Geer [7] and references therein). On the other end of the complexity spectrum, the work of Comminges and Dalayan [15] considers a sparse non-parametric model

$$y_i = f(X_i) = f_J(\{(X_i)_j : j \in J\}) \quad |J| \leq s,$$

where at most s of the components of the X_i are relevant to the model. They ask the question of when the set J of relevant components can be recovered and find that the necessary number of measurements is

$$n \geq C_1 \exp(C_2 s) \log(p/s).$$

We compare this to our result on recovering a single index β belonging to the cone K of s-sparse vectors, for which

$$(w_1(K))^2 \leq s \log(p/s).$$

We see that going from a single-index to a non-parametric model involves an exponentially larger number of measurements (but the number is exponential in the underlying sparsity not in the full dimension of the parameter space).

We mention also some of the recent work on the estimation of the link function f in high-dimensional settings. Alquier and Biau [2] and Cohen *et al.* [14] consider the model of a link function f which depends on at most s of the components of x. More generally, Dalalyan *et al.* [17] allow all link functions that can be decomposed as the sum of at most m 'nice' functions f_V, with each f_V depending on at most s coordinates of x. (A special case of all of these models, including ours, is $f(x) = \tilde{f}(\langle x, \beta\rangle)$, where β s-sparse and $\tilde{f}$ comes from a nice function class.) The recent work of Dalalyan *et al.* [17] obtain general bounds on the quality of estimation of f, uniformly over classes containing f_V.

6.1 *Connection to statistical learning theory*

In statistical (supervised) learning theory, the data are used to determine a function that can be used for future prediction. Thus the data $\{a_i, y_i, i = 1, \dots, m\}$ are used to generate a function which can predict y_{m+1} given a_{m+1}. Such a predicting function is usually sought in a given function class. A main focus of the research is to develop oracle inequalities, which state that the predicting function estimated from the data performs nearly as well as the optimal function in that class.

By restricting to classes of linear functionals, such results can (with some work) be specialized and translated to give error bounds on the estimation of x, as in our paper. We focus on the results of Lecué & Mendelson [33], which gives general error bounds under mild conditions. Nevertheless, there are important differences between these and the results of our paper. First, statistical learning theory concentrates on *empirical risk minimization* Lecué & Mendelson [33], which, when combined with a *squared loss function*, and specialized to linear functionals, recovers a generalized version of the Lasso. In contrast, we consider a simpler, and often more efficient, projection method in this article. (We note in passing that, after writing this paper, the first two authors studied the behavior of the Lasso under the single-index model [46].) Furthermore, we make especially mild assumptions on the data. In contrast to Lecué & Mendelson [33], our Theorem 1.3 does not require y_i to be sub-Gaussian. Instead, we roughly only require y_i to have finite second moment so that the parameters μ, σ and η are well defined. On the other hand, the Lasso can return an exact solution in the noiseless model, whereas the project-based estimator proposed in this article always has some error attached. The Lasso can also tolerate the case when the measurement vectors are anisotropic [46].

We also point to the work of [40], which can also be specialized to linear functionals to give an error bound similar to the one in our paper. However, the simple projection method espoused in our paper is often more computationally efficient than the optimization programs suggested in [40]. Further, the conditions on the signal structure are much milder. Indeed, in contrast to [40], there is no need for *decomposability* or *restricted strong convexity*.

Finally, we note that the behavior of the generalized Lasso under the linear model is well studied, and error bounds are known with sharp constants [41]. Recently, building on the work of [46], sharp constants have been given for the nonlinear model in the asymptotic regime [54]. It is an interesting open question to see if this can also be done in the non-asymptotic regime.

7 Discussion

We have analyzed the effectiveness of a simple, projection-based estimator in the semiparametric single index model. We showed that by projecting onto the set encoding the signal structure, rough, unspecified observations could be combined to give an accurate description of the signal. The gain from this dimension reduction was described by a geometric parameter—the mean width. When the noise is large and under mild assumptions on the rest of the model, we showed that the estimator is minimax optimal up to a constant. We came to the surprising conclusion that an unknown, non-invertible nonlinearity often does not significantly affect one's ability to estimate the signal, aside from a loss of scaling information and an extra numerical constant in the error bound.

By comparing to what was known in the classic literature on the semiparametric single-index model, we believe our results (a) give a simple method of testing whether estimation is possible based on easily computable parameters of the model, (b) allow for non-invertible, discontinuous and unknown linearities, and (c) give a careful accounting of the benefit of using a low-dimensional signal structure.

Nevertheless, our model takes the measurement vectors to be standard normal, and it is important to understand whether this assumption may be generalized, and whether the theory in this paper matches practice. We discuss this challenge in the rest of this section. We pause to note that the lower bounds take a very general model for the measurement vectors, which may even be deterministic, thus it is the model used for the upper bounds which requires generalization.

There is a simple key observation that gives the first step to the theory in this paper: The linear estimator described in Section 1.2 is an unbiased estimate of the signal, up to scaling. It may be surprising at first that this holds regardless of a nonlinearity in the observations. This fortunate fact follows from the Gaussianity of the measurements, as described in the next section.

It is straightforward to generalize the theory to the case $a_i \sim N(0, \Sigma)$, provided Σ is known or can be estimated. One only needs to multiply the linear estimator by Σ^{-1} to give an unbiased estimator; as long as Σ is well conditioned, our theory remains essentially unchanged. The next question is the applicability of other random measurement ensembles. We see two avenues toward this research: (1) As we illustrated with the matrix completion example of Section 2.5, if the observations are linear, then the linear estimator can be unbiased for non-Gaussian measurement vectors. All that is needed is for the measurement vectors to be in isotropic position, that is, $\mathbb{E}\, a_i a_i^* = I_n$. This is a common assumption in the compressed sensing literature [9]. Of course, a slight nonlinearity can be accounted for by a Taylor series approximation. (2) A central idea in high-dimensional estimation is that non-Gaussian measurement vectors often give similar behavior to Gaussian measurement vectors. For example, this is made explicit with the universality phenomenon in compressed sensing. Such arguments have been applied to a special case of this semiparametric single-index model in [1] focused on binary observations with non-Gaussian measurement vectors. We believe that similar comparison arguments may be applied in the more general setting of this article.

In conclusion, we emphasize that the assumption of Gaussian measurements has allowed a very clean theory, with relatively few assumptions in a general model, and we hope that it can be a first step toward such theory with other models of the measurement vectors.

8 Proofs of Proposition 1.1 and Theorem 1.3

8.1 *Orthogonal decomposition*

The proof of Proposition 1.1 as well as of our main result, Theorem 1.3, will be based on the orthogonal decomposition of the vectors a_i along the direction of x and the hyperplane $x^\perp$. More precisely, we express

$$a_i = \langle a_i, \bar{x} \rangle \, \bar{x} + b_i. \tag{8.1}$$

The vectors b_i defined this way are orthogonal to x. Let us record a few elementary properties of this orthogonal decomposition.

LEMMA 8.1 (Properties of the orthogonal decomposition) The orthogonal decomposition (8.1) and the observations y_i satisfy the following properties:

(1) $\langle a_i, \bar{x} \rangle \sim N(0, 1)$;

(2) $b_i \sim N(0, I_{x^\perp})$;

(3) $\langle a_i, \bar{x} \rangle$ and b_i are independent;

(4) y_i and b_i are independent.

Proof. Properties (i), (ii) and (iii) follow from the orthogonal decomposition and the rotational invariance of the normal distribution.

Property (iv) follows from a contraction property of conditional independence. Let us denote by $Y \perp B$ the independence of random variables (or vectors) Y and B, and by $(Y \perp B) \mid H$ the conditional independence Y and B given H. The contraction property states that $(Y \perp B) \mid H$ and $B \perp H$ imply $Y \perp B$. In our situation, we have $(y_i \perp b_i) \mid \langle a_i, \bar{x} \rangle$ by the assumption on y_i and since b_i is uniquely determined by a_i. Moreover, $b_i \perp \langle a_i, \bar{x} \rangle$ by property (iii). The contraction property yields $y_i \perp b_i$. This proves (iv). □

8.2 *Proof of Proposition* 1.1

By the identical distribution of a_i and using the orthogonal decomposition (8.1), we have

$$\mathbb{E}\,\widehat{x}_{\text{lin}} = \mathbb{E}\, y_1 a_1 = \mathbb{E}\, y_1 \langle a_1, \bar{x} \rangle \bar{x} + \mathbb{E}\, y_1 b_1. \tag{8.2}$$

The first term in the right-hand side equals $\mu\bar{x}$ by definition of μ. The second term equals zero, since by the independence property (iv) in Lemma 8.1 and since $\mathbb{E}\, b_1 = 0$. We proved the first part of the proposition, $\mathbb{E}\,\widehat{x}_{\text{lin}} = \mu\bar{x}$.

To prove the second part, we express

$$\mathbb{E}\,\|\widehat{x}_{\text{lin}} - \mu\bar{x}\|_2^2 = \mathbb{E}\left\| \frac{1}{m} \sum_{i=1}^{m} Z_i \right\|_2^2,$$

where $Z_i = y_i a_i - \mu\bar{x}$ are independent and identically distributed random vectors with zero mean. Thus

$$\mathbb{E}\,\|\widehat{x}_{\text{lin}} - \mu\bar{x}\|_2^2 = \frac{1}{m^2} \sum_{i=1}^{m} \mathbb{E}\,\|Z_i\|_2^2 = \frac{1}{m}\,\mathbb{E}\,\|Z_1\|_2^2. \tag{8.3}$$

Using orthogonal decomposition (8.1) again, we can express Z_1 as follows:

$$Z_1 = y_1 \langle a_1, \bar{x} \rangle \bar{x} + y_1 b_1 - \mu\bar{x} = X + Y,$$

where

$$X = \big[y_1 \langle a_1, \bar{x} \rangle - \mu \big] \bar{x}, \quad Y = y_1 b_1.$$

Note that

$$\mathbb{E}\,\langle X, Y \rangle = 0. \tag{8.4}$$

To see this, $\langle X, Y \rangle = [y_1^2 \langle a_1, \bar{x} \rangle - \mu y_1] \cdot \langle b_1, \bar{x} \rangle$. The two terms forming this product are independent by properties (iii) and (iv) in Lemma 8.1. Moreover, $\mathbb{E}\, b_1 = 0$, which yields $\mathbb{E}\,\langle b_1, \bar{x} \rangle = 0$, and as a consequence, (8.4) follows.

Property (8.4) implies that

$$\mathbb{E}\,\|Z_1\|_2^2 = \mathbb{E}\,\|X+Y\|_2^2 = \mathbb{E}\,\|X\|_2^2 + \mathbb{E}\,\|Y\|_2^2.$$

We have $\mathbb{E}\,\|X\|_2^2 = \sigma^2$ by definition of σ and since $\|\bar{x}\|_2 = 1$. Next, $\mathbb{E}\,\|Y\|_2^2 = \eta^2\,\mathbb{E}\,\|b_1\|_2^2$ by the independence property (iv) in Lemma 8.1 and by definition of η. Recalling that b_1 is a standard normal random variable in the hyperplane $x^\perp$, we get $\mathbb{E}\,\|b_1\|_2^2 = n-1$. It follows that

$$\mathbb{E}\,\|Z_1\|_2^2 = \sigma^2 + \eta^2(n-1).$$

Putting this into (8.3), we complete the proof. □

8.3 *Metric projection and dual norm*

For a subset T of $\mathbb{R}^n$, we define

$$\|x\|_{T^\circ} = \sup_{u\in T} \langle x, u\rangle \qquad x \in \mathbb{R}^n. \tag{8.5}$$

It is a seminorm if T is symmetric. We define also $T_d = (T-T)\cap dB_2^n$.

LEMMA 8.2 Let K be an arbitrary subset of $\mathbb{R}^n$ and let $z \in K$, $w \in \mathbb{R}^n$. Then the distance

$$d := \|P_K(w) - z\|_2$$

satisfies the inequality

$$d^2 \le 2\|w-z\|_{K_d^\circ}.$$

Proof. By definition, $P_K(w)$ is the closest vector to w in K, so

$$\|P_K(w) - w\|_2 \le \|z-w\|_2.$$

We write this as

$$\|(P_K(w)-z) + (z-w)\|_2^2 \le \|z-w\|_2^2,$$

expand the left-hand side and cancel the terms $\|z-w\|_2^2$ on both sides. This gives

$$\|P_K(w)-z\|_2^2 \le 2\,\langle P_K(w)-z, w-z\rangle. \tag{8.6}$$

The left-hand side of (8.6) equals d^2 by definition. To estimate the right-hand side, note that vectors $P_K(w)$ and z lie in K and they are d apart in Euclidean distance. Therefore $P_K(w)-z \in (K-K)\cap dB_2^n = K_d$. Thus the right-hand side of (8.6) is bounded by $2\|w-z\|_{K_d^\circ}$ as claimed. □

COROLLARY 8.3 Let K be a star-shaped set and let $z \in K$, $w \in \mathbb{R}^n$. Then for every $t > 0$, we have

$$\|P_K(w) - z\|_2 \le \max\left(t, \frac{2}{t}\|w - z\|_{K_t^\circ}\right). \tag{8.7}$$

Proof. The proof relies on the following fact, which is well known in geometric functional analysis.

CLAIM For any fixed $x \in \mathbb{R}^n$, the function $\frac{1}{t}\|x\|_{K_t^\circ}$ is non-increasing in $t \in \mathbb{R}_+$.

To prove this claim, we express the function as

$$\frac{1}{t}\|x\|_{K_t^\circ} = \frac{1}{t}\,\mathbb{E}\sup_{u \in (K-K)\cap tB_2^n}\langle x, u\rangle = \mathbb{E}\sup_{v \in \frac{1}{t}(K-K)\cap B_2^n}\langle x, v\rangle. \tag{8.8}$$

Since K is star shaped, $K - K$ is star shaped, too. Then the set $\frac{1}{t}(K - K)$ is non-increasing (with respect to inclusion) in $t \in \mathbb{R}_+$. Thus the same is true about $\frac{1}{t}(K - K) \cap B_2^n$. This and the identity (8.8) prove the claim.

To deduce (8.7), denote $d = \|P_K(w) - z\|_2$. If $d \le t$ then (8.7) holds. If $d > t$, we apply Lemma 8.2 followed by the claim above. We obtain

$$d \le \frac{2}{d}\|w - z\|_{K_d^\circ} \le \frac{2}{t}\|w - z\|_{K_t^\circ}.$$

This implies (8.7). □

8.4 *Proof of Theorem 1.3*

8.4.1 *Decomposition of the error* We apply Corollary 8.3 for $w = \widehat{x}_{\text{lin}}$ and $z = \mu\bar{x}$; note that the requirement that $z \in K$ is satisfied by assumption. We obtain

$$\|\widehat{x} - \mu\bar{x}\|_2 = \|P_K(\widehat{x}_{\text{lin}}) - \mu\bar{x}\|_2 \le t + \frac{2}{t}\|\widehat{x}_{\text{lin}} - \mu\bar{x}\|_{K_t^\circ}. \tag{8.9}$$

Recall that $\widehat{x}_{\text{lin}} = \frac{1}{m}\sum_{i=1}^m y_i a_i$ and use the orthogonal decomposition of a_i from (8.1). By triangle inequality, we obtain

$$\begin{aligned}\|\widehat{x}_{\text{lin}} - \mu\bar{x}\|_{K_t^\circ} &= \left\|\frac{1}{m}\sum_{i=1}^m \left[y_i\,\langle a_i, \bar{x}\rangle\,\bar{x} + y_i b_i - \mu\bar{x}\right]\right\|_{K_t^\circ} \\ &\le \left\|\frac{1}{m}\sum_{i=1}^m \left[y_i\,\langle a_i, \bar{x}\rangle - \mu\right]\bar{x}\right\|_{K_t^\circ} + \left\|\frac{1}{m}\sum_{i=1}^m y_i b_i\right\|_{K_t^\circ} =: E_1 + E_2. \end{aligned} \tag{8.10}$$

8.4.2 *Estimating* E_1 Denoting $\xi_i = y_i \langle a_i, \bar{x} \rangle - \mu$, we have

$$E_1 = \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \right| \cdot \|\bar{x}\|_{K_t^\circ}.$$

By definition, $K_t \subseteq tB_2^n$ and $\bar{x} \in B_2^n$, so $\|\bar{x}\|_{K_t^\circ} \leq t$. Further, ξ_i are independent and identically distributed random variables with zero mean. Therefore

$$\mathbb{E}\, E_1^2 \leq \frac{1}{m} \mathbb{E}[\xi_1^2] \cdot t^2 = \frac{\sigma^2 t^2}{m}, \tag{8.11}$$

where the last equality follows by the definition of ξ_1 above and the definition of σ in (1.2).

8.4.3 *Estimating* E_2 We will estimate

$$\mathbb{E}\, E_2 = \mathbb{E}\, \|h\|_{K_t^\circ}, \quad \text{where} \quad h := \frac{1}{m} \sum_{i=1}^{m} y_i b_i.$$

Recall that y_i and b_i are independent by property (iv) in Lemma 8.1. Let us condition on y_i. Rotation invariance of the normal distribution implies that

$$h \sim N(0, sI_{x^\perp}), \quad \text{where} \quad s^2 = \frac{1}{m^2} \sum_{i=1}^{m} y_i^2.$$

In order to extend the support of the distribution of h from $x^\perp$ to $\mathbb{R}^n$, we recall the following elementary and well-known fact.

CLAIM For a subspace E of $\mathbb{R}^n$, let g_E denote a random vector with distribution $N(0, I_E)$. Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then

$$\mathbb{E}\, \Phi(g_E) \leq \mathbb{E}\, \Phi(g_F) \quad \text{whenever} \quad E \subseteq F.$$

To prove the claim, consider the orthogonal decomposition $F = E \oplus L$ for an appropriate subspace L. Then $g_F = g_E + g_L$ in distribution, where $g_L \sim N(0, I_L)$ is independent of g_E. Jensen's inequality yields

$$\mathbb{E}\, \Phi(g_E) = \mathbb{E}\, \Phi(g_E + \mathbb{E}\, g_L) \leq \mathbb{E}\, \Phi(g_E + g_L) = \mathbb{E}\, \Phi(g_F)$$

as claimed.

In the notation of the claim, we can estimate the expectation of $\|h\|_{K_t^\circ}$ (still conditionally on the y_is) as follows:

$$\mathbb{E}\, \|h\|_{K_t^\circ} = \mathbb{E}\, \|s g_{x^\perp}\|_{K_t^\circ} = s\, \mathbb{E}\, \|g_{x^\perp}\|_{K_t^\circ} \leq s\, \mathbb{E}\, \|g_{\mathbb{R}^n}\|_{K_t^\circ} = s \cdot w_t(K),$$

where the last equality follows by the definitions of the dual norm and local mean width. Therefore, unconditionally we have

$$\begin{aligned}\mathbb{E}\, E_2 &= \mathbb{E}\, \|h\|_{K_t^\circ} = \mathbb{E}[s] \cdot w_t(K) \le (\mathbb{E}\, s^2)^{1/2} \cdot w_t(K) \\ &= \left[\frac{1}{m}\, \mathbb{E}\, y_1^2\right]^{1/2} \cdot w_t(K) = \frac{\eta}{\sqrt{m}} \cdot w_t(K), \end{aligned} \tag{8.12}$$

where the last equality follows from the definition of η in (1.2).

It remains to plug estimates (8.11) for E_1 and (8.12) for E_2 into (8.10), and we obtain

$$\mathbb{E}\, \|\widehat{x}_{\text{lin}} - \mu\bar{x}\|_{K_t^\circ} = \mathbb{E}(E_1 + E_2) \le (\mathbb{E}\, E_1^2)^{1/2} + \mathbb{E}\, E_2 \le \frac{1}{\sqrt{m}}\, [\sigma t + \eta\, w_t(K)].$$

Finally, substituting this estimate into (8.9) and taking expectation of both sides completes the proof of Theorem 1.3. □

9 High-probability version of Theorem 1.3

Using standard concentration inequalities, we can complement Theorem 1.3 with a similar result that is valid with high probability rather than in expectation.

To this end, we will assume that the observations y_i have *sub-Gaussian* distributions. This is usually expressed by requiring that

$$\mathbb{E} \exp(y_i^2/\psi^2) \le 2 \quad \text{for some } \psi > 0. \tag{9.1}$$

Basic facts about sub-Gaussian random variables can be found, e.g. in [55].

To understand this assumption, consider the case when the nonlinearity is given by an explicit function, i.e. $y_i = f(\langle a_i, x\rangle)$. Since $\langle a_i, x\rangle$ is Gaussian, y_i will be sub-Gaussian provided that f does not grow faster than linearly, i.e. provided that $f(x) \le a + b|x|$ for some scalars a and b.

THEOREM 9.1 (Non-linear estimation with high probability) Let $\bar{x} = x/\|x\|_2$. Assume that $\mu\bar{x} \in K$, where K is a fixed star-shaped closed subset of $\mathbb{R}^n$. Assume that observations y_i satisfy the sub-Gaussian bound (9.1). Let $t > 0$ and $0 < s \le \sqrt{m}$. Then the nonlinear estimator $\widehat{x}$ defined in (1.6) satisfies

$$\|\widehat{x} - \mu\bar{x}\|_2 \le t + \frac{4\eta}{\sqrt{m}} \left[s + \frac{w_t(K)}{t}\right]$$

with probability at least $1 - 2\exp(-cs^2\eta^4/\psi^4)$. Here μ and η are the numbers defined in (1.2), and $c > 0$ is an absolute constant.

Proof. The conclusion follows by combining the proof of Theorem 1.3 and standard concentration techniques, which can be found in [55]. So we will only outline the argument. Let $\varepsilon = s/(2\sqrt{m})$; then $\varepsilon < 1/2$ by assumption.

First we bound E_1. Since y_i are sub-Gaussian as in (9.1) and $\langle a_i, \bar{x}\rangle$ is $N(0, 1)$, the random variables ξ_i are sub-exponential. More precisely, we have $\mathbb{E} \exp(\xi_i/C\psi) \le 2$, where $C > 0$ denotes an absolute

constant here and thereafter. A Bernstein-type inequality (see Vershynin [55, Corollary 5.17]) implies that

$$\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i\right| \le \varepsilon\eta \quad \text{and thus} \quad E_1 \le \varepsilon\eta t, \tag{9.2}$$

with probability at least

$$1 - 2\exp\left[-c\min\left(\frac{\varepsilon^2\eta^2}{\psi^2}, \frac{\varepsilon\eta}{\psi}\right)m\right] \ge 1 - 2\exp\left(-\frac{cm\varepsilon^2\eta^2}{\psi^2}\right). \tag{9.3}$$

In the last inequality we used that $\eta = \mathbb{E}\, y_i^2 \le C\psi$ by definition of ψ and that $\varepsilon \le 1$ by assumption.

Turning to E_2, we need to bound $\sum_{i=1}^m y_i^2$ and $\|g\|_{K_t^\circ}$. A similar application of a Bernstein-like inequality for the sub-exponential random variables $y_i^2 - \eta^2$ shows that

$$\sum_{i=1}^{m} y_i^2 \le 4\eta^2 m$$

with probability at least

$$1 - 2\exp\left[-c\min\left(\frac{\eta^4}{\psi^4}, \frac{\eta^2}{\psi^2}\right)m\right] \ge 1 - 2\exp\left(-\frac{cm\eta^4}{\psi^4}\right). \tag{9.4}$$

Next, we bound $\|g\|_{K_t^\circ}$ using Gaussian concentration. Since $K_t \subseteq tB_2^n$, the function $x \mapsto \|x\|_{K_t^\circ}$ on $\mathbb{R}^n$ has Lipschitz norm at most t. Therefore, the Gaussian concentration inequality (see e.g. [34]) implies that

$$\|g\|_{K_t^\circ} \le \mathbb{E}\,\|g\|_{K_t^\circ} + t\varepsilon\sqrt{m} = w_t(K) + t\varepsilon\sqrt{m}$$

with probability at least

$$1 - \exp(-c\varepsilon^2 m). \tag{9.5}$$

If both $\sum_{i=1}^m y_i^2$ and $\|g\|_{K_t^\circ}$ are bounded as above, then

$$E_2 \le \frac{1}{m}\left(\sum_{i=1}^{m} y_i^2\right)^{1/2}\|g\|_{K_t^\circ} \le \frac{2\eta}{\sqrt{m}}\left(w_t(K) + t\varepsilon\sqrt{m}\right).$$

If also E_1 is bounded as in (9.2), then we conclude as in the proof of Theorem 1.3 that

$$\|\widehat{x} - \mu\bar{x}\|_2 \le t + \frac{2}{t}(E_1 + E_2) \le t + \frac{4\eta}{\sqrt{m}}\cdot\frac{w_t(K)}{t} + 6\eta\varepsilon.$$

Recalling (9.3), (9.4) and (9.5), we see that this happens with probability at least

$$1 - 2\exp\left(-\frac{cm\varepsilon^2\eta^2}{\psi^2}\right) - 2\exp\left(-\frac{cm\eta^4}{\psi^4}\right) - \exp(-c\varepsilon^2 m) \geq 1 - 2\exp\left(-\frac{c'm\varepsilon^2\eta^4}{\psi^4}\right)$$

for an appropriate absolute constant $c' > 0$. (Here we used again that $\eta/\psi < 1$ and $\varepsilon < 1$.) The conclusion of Theorem 9.1 follows by definition of ε. □

10 Proofs of Theorem 4.2, Corollary 4.4, and Lemma 4.9

In this section, for a set $K \subset \mathbb{R}^n$ we use the notation $K_t = K \cap tB_2^n$. (Note that we do not symmetrize K first.)

10.1 *Proof of Theorem 4.2*

The theorem is proven with a careful application of Fano's inequality, an information theoretic method which is useful for determining minimax estimation rates. We state a version of Fano's inequality synthesized for lower-bounding expected error in the linear model (the steps needed to specialize Fano's inequality can be seen in [8, 48], for example). For the general version see [16]. In all that follows, $\mathcal{X}_t$ is a $t/10$-packing of K_t with maximal cardinality. Thus, $|\mathcal{X}_t| = P_t$.

THEOREM 10.1 (Fano's Inequality for expected error) Consider a fixed matrix A in the linear model of Section 4.1, and let $t > 0$. Suppose that

$$\frac{1}{v^2|\mathcal{X}_t|^2} \sum_{\substack{w \neq v \\ w,v \in \mathcal{X}_t}} \|A(w-v)\|_2^2 + \log(2) \leq \frac{1}{2}\log(|\mathcal{X}_t|). \tag{10.1}$$

Then there exists a constant $c > 0$ such that for any estimator $\widehat{x}(A, y)$,

$$\sup_{x \in \mathcal{X}_t} \mathbb{E}\,\|\widehat{x} - x\|_2 \geq ct.$$

Theorem 4.2 will be proven by conditioning on A and applying Fano's inequality. We begin to control the left-hand side of (10.1) by showing that A does not significantly increase the average distance between points in our packing. Indeed,

$$\mathbb{E} \sum_{\substack{w \neq v \\ w,v \in \mathcal{X}_t}} \|A(w-v)\|_2^2 \leq \|\,\mathbb{E}A^*A\| \sum_{\substack{w \neq v \\ w,v \in \mathcal{X}_t}} \|(w-v)\|_2^2 \leq m \sum_{\substack{w \neq v \\ w,v \in \mathcal{X}_t}} \|(w-v)\|_2^2 \leq 4m|\mathcal{X}_t|^2 t^2.$$

The last step follows since $\mathcal{X}_t \subset tB_2^n$. Thus, by Markov's inequality,

$$\sum_{\substack{w \neq v \\ w,v \in \mathcal{X}_t}} \|A(w-v)\|_2^2 \leq 16m|\mathcal{X}_t|^2 t^2 \qquad \text{with probability at least } 3/4. \tag{10.2}$$

We pause to note that the above inequality is the only property of A that is needed.

By conditioning on the good event that the Inequality (10.2) holds, the following corollary gives a significant simplification of Theorem 10.1.

COROLLARY 10.2 (Simplified Fano Bound) Consider the linear model of Section 4.1, suppose $\|\mathbb{E}\, AA^*\| \leq m$ and let $t > 0$. Set $\bar{\nu} := \nu/\sqrt{m}$. Suppose that

$$\frac{t^2}{\bar{\nu}^2} + 1 \leq c \log(|\mathcal{X}_t|) \tag{10.3}$$

for a small enough constant c. Then there exists a constant $c' > 0$ such that for any estimator $\widehat{x}(A, y)$,

$$\sup_{x \in \mathcal{X}_t} \mathbb{E}\, \|\widehat{x} - x\|_2 \geq c' t.$$

We are now in position to prove Theorem 4.2.

Proof. Let c_0 be a small numerical constant, whose value will be chosen at the end of the proof, and let

$$\delta := c_0 \delta_* = c_0 \inf_t \left\{ t + \bar{\nu} \left[1 + \sqrt{\log P_t} \right] \right\}. \tag{10.4}$$

We split the proof into two cases, depending on the relative size of δ and $\bar{\nu}$.

Case 1: $\delta \geq \bar{\nu}$.

This case contains the meat of the proof, but nevertheless it can be proven quite quickly with the tools we have gathered. Since δ is the infimum in (10.4), it satisfies

$$\delta \leq c_0 \left(\delta + \bar{\nu} \left[1 + \sqrt{\log P_\delta} \right] \right).$$

Further, $\bar{\nu} \leq \delta$ in Case 1. Thus, we may massage the equation to give

$$\delta \leq c_1 \bar{\nu} \sqrt{\log P_\delta},$$

where $c_1 = c_0/(1 - 2c_0)$ and we take $c_0 < 1/2$.

We now check that δ satisfies Inequality (10.3) as required for Fano's inequality. Since $\delta \geq \bar{\nu}$, one has

$$\frac{\delta^2}{\bar{\nu}^2} + 1 \leq 2 \frac{\delta^2}{\bar{\nu}^2} \leq 2 c_1^2 \log P_\delta.$$

The inequality now follows if we take c_0 such that $2c_1^2 < c$.

We now move to the second case.

Case 2: $\delta \leq \bar{\nu}$. This is an unusual case; it only occurs when K is quite small, e.g. a one-dimensional subspace. Fano's inequality, which is more applicable for higher dimensional sets, may not give the optimal answer, but we may use much simpler methods. We give a bare-bones argument.

Consider any two signals, $w, v \in K$ satisfying $\|w - v\|_2 \leq c\bar{\nu}$. We will show that these two points are indistinguishable based on the noisy observations y. Consider the hypothesis testing problem of determining which of the two points generate y, i.e. either $y = Aw + \varepsilon$ or $y = Av + \varepsilon$. As seen from (10.2), with probability at least $3/4$, $\|A(v - w)\|_2 \leq c\sqrt{m}\bar{\nu} \leq c\nu$. Thus, the difference between the two candidate signals is much smaller than the noise level. It follows from a quick analysis of the *likelihood ratio test*, which is optimal by *Neyman–Pearson lemma*, that the hypotheses are indistinguishable, i.e. there is no estimator which has a more than $3/4$ chance of determining the original signal. This gives the lower bound

$$\sup_{x \in K} \mathbb{E}\, \|\widehat{x} - x\|_2 \geq c\, \|v - w\|_2.$$

By taking the maximum distance between signals $v, w \in K \cap \bar{\nu} B_2^n$, this gives

$$\sup_{x \in K} \mathbb{E}\, \|\widehat{x} - x\|_2 \geq c \min(\operatorname{diam}(K), \bar{\nu}) \geq c \min(\operatorname{diam}(K), \delta). \qquad \square$$

10.2 *Proof of Corollary 4.4 incorporating Remark 4.5*

We prove a generalization of Corollary 4.4 that takes the tighter version of α from Remark 4.5.

COROLLARY 10.3 Assume that $x \in K$ where K is a star-shaped subset of $\mathbb{R}^n$. Assume that the observations y_i are given by 4.1. Let

$$\delta^* := \inf_t \left\{ t + \frac{\nu}{\sqrt{m}} \left[1 + \frac{w_t(K)}{t} \right] \right\}. \tag{10.5}$$

Then any estimator $\widehat{x}$ satisfies

$$\sup_{x \in K} \mathbb{E}\, \|\widehat{x} - x\|_2 \geq c_\alpha \min(\delta^*, \operatorname{diam}(K)),$$

where $c_\alpha = c/\alpha$ for a numerical constant c and α defined in (4.5).

Proof. The proof proceeds simply by showing that $\delta_2^* = \frac{1}{2}\delta^* \leq \alpha\delta_*$, with δ_* defined in (4.2). One may then apply Theorem 4.2. First, observe that by definition

$$\delta_2^* \leq \frac{1}{2} \left(\delta_2^* + \frac{\nu}{\sqrt{m}} \left[1 + \frac{w_{\delta_2^*}(K)}{\delta_2^*} \right] \right).$$

Massage the equation to give

$$\delta_2^* \leq \frac{\nu}{\sqrt{m}} \left[1 + \frac{w_{\delta_2^*}(K)}{\delta_2^*} \right].$$

The geometric assumption then implies that

$$\delta_2^* \leq \frac{\nu}{\sqrt{m}} \left[1 + \alpha \sqrt{\log P_{\delta_2^*}} \right] \leq \alpha \cdot \frac{\nu}{\sqrt{m}} \left[1 + \sqrt{\log P_{\delta_2^*}} \right].$$

Note that since K is star shaped, P_t is decreasing in t and thus,

$$\delta_2^*/\alpha \le \inf_{t\le\delta_2^*} \frac{\nu}{\sqrt{m}}\left[1+\sqrt{\log P_{\delta_2^*}}\right] \le \inf_{t\le\delta_2^*}\left\{t+\frac{\nu}{\sqrt{m}}\left[1+\sqrt{\log P_{\delta_2^*}}\right]\right\}.$$

Now, if $\delta_2^* < \delta_*$, then the proof is done, so we may restrict to the case $\delta_2^* \ge \delta_*$. However, in this case, by definition of δ_*, the restriction $t \le \delta_2^*$ may be removed from the infimum just above, and so

$$\delta_2^*/\alpha \le \inf_t \left\{t+\frac{\nu}{\sqrt{m}}\left[1+\sqrt{\log P_{\delta_2^*}}\right]\right\} = \delta_*. \qquad \square$$

10.3 *Proof of Lemma* 4.9

Proof of Lemma 4.9. The proof will follow from several manipulations and decompositions of Gaussian random variables. By definition, in the nonlinear model

$$\mu = \mathbb{E} f(\langle a_1, x\rangle + \varepsilon_1)\,\langle a_1, \bar{x}\rangle = \mathbb{E} f(\|x\|_2 \cdot g + \nu z) g,$$

where $g, z \sim N(0,1)$ are independent. Let $\beta^2 = \mathrm{Var}(\|x\|_2 \cdot g + \nu z) = \|x\|_2^2 + \eta^2$ and let

$$w := \frac{\|x\|_2 \cdot g + \nu z}{\beta} \sim N(0,1).$$

Further, decompose g as $g = (\|x\|_2/\beta) w + w_\perp$, where $w_\perp$ is independent of w. Thus,

$$\mu\beta/\|x\|_2 = \mathbb{E} f(\beta w) w = \mathbb{E}\left[f(\beta|w|)\cdot|w|\right].$$

The second equality follows since f is assumed to be odd. Now, since f is non-decreasing, we have

$$\mathbb{E} f(\beta|w|)\cdot|w| \ge f(\beta)\cdot\mathbb{P}\left\{|w|\ge 1\right\} = C f(\beta),$$

where $C = \mathbb{P}\left\{|w|\ge 1\right\} \approx 0.683$.

Putting pieces together, we have

$$\mu \ge C\frac{\|x\|_2 f(\beta)}{\beta}.$$

We now give an upper bound for σ and η.

$$\sigma^2 = \mathrm{Var}(f(\beta w)\cdot g) \le \mathbb{E}(f(\beta w)\cdot g)^2.$$

By Cauchy–Schwarz inequality, the right-hand side is bounded by

$$\sqrt{\mathbb{E} f^4(\beta w)}\cdot\sqrt{\mathbb{E} g^4} = 3^{1/2}\sqrt{\mathbb{E} f^4(\beta w)}. \tag{10.6}$$

The sub-multiplicative assumption implies that

$$\mathbb{E} f^4(\beta w) \le f^4(\beta)\,\mathbb{E} f^4(w) = C_f \cdot f^4(\beta).$$

Thus, $\sigma \leq C_f f(\beta)$. As a bonus, we bounded η by the same quantity, since η^2 is bounded by the right-hand side of (10.6). Thus,

$$\lambda(\sigma + \eta) = \frac{\|x\|_2}{\mu}(\sigma + \eta) \leq \frac{\|x\|_2 f(\beta)\beta}{\|x\|_2} C_f f(\beta) \leq C_f(\|x\|_2 + \eta),$$

where

$$C_f = C\,\mathbb{E}[f^4(w)]^{1/4} \quad C = 48^{1/4}\mathbb{P}\left\{|w| \geq 1\right\} \approx 1.8. \qquad \square$$

Funding

NSF Postdoctoral Research Fellowship (award No. 1103909, to Y.P.); NSF grants (DMS 1161372, 1001829, 1265782, to R.V.); USAF Grant (FA9550-14-1-0009, to R.V.); NSF Postdoctoral Research Fellowship (award No. 1204311, to E.Y.).

References

1. Ai, A., Lapanowski, A., Plan, Y. & Vershynin, R. (2014) One-bit compressed sensing with non-Gaussian measurements. *Lin. Algebra Appl.*, **441**, 222–239.
2. Alquier, P. & Biau, G. (2013) Sparse single-index model. *J. Mach. Learn. Res.*, **14**, 243–280.
3. Amelunxen, D., Lotz, M., McCoy, M. B. & Tropp, J. A. (2014) Living on the edge: A geometric theory of phase transitions in convex optimization. *Inf. inference*, **3**, 224–294.
4. Bach, F. (2010) Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, **4**, 384–414.
5. Bartlett, P. L. & Mendelson, S. (2003) Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, **3**, 463–482.
6. Boufounos, P. T. & Baraniuk, R. G. (2008) 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, IEEE, pp. 16–21.
7. Bühlmann, P. & Van De Geer, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.
8. Candès, E. J. & Davenport, M. A. (2013) How well can we estimate a sparse vector? *Appl. Comput. Harmonic Ana.*, **34**, 317–323.
9. Cande, E. J. & Plan, Y. (2011a) A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory*, **57**, 7235–7254.
10. Cande, E. J. & Plan, Y. (2011b) Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory*, **57**, 2342–2359.
11. Candès, E. J. & Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations Comput. Math.*, **9**, 717–772.
12. Chatterjee, S. (2015) Matrix estimation by universal singular value thresholding. *Ann. Statist.*, **43**, 177–214.
13. Chatterjee, S. (2014) A new perspective on least squares under convex constraint. *Ann. Statist.*, **42**, 2340–2381.
14. Cohen, A., Daubechies, I., DeVore, R., Kerkyacharian, G. & Picard, D. (2012) Capturing ridge functions in high dimensions from point queries. *Constructive Approximation*, **35**, 225–243.
15. Comminges, L. & Dalalyan, A. S. (2012) Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Stat.*, **40**, 2667–2696.
16. Cover, T. M. & Thomas, J. A. (2012) *Elements of Information Theory*. New Jersey: John Wiley & Sons.
17. Dalalyan, A., Ingster, Y. & Tsybakov, A. B. (2013) Statistical inference in compound functional models. *Prob. Theory Rela. Fields*, **5**, 1–20.
18. Davenport, M. A., Plan, Y., Berg, E. v. d. & Wootters, M. (2014) 1-bit matrix completion. *Inf. inference*, **3.3**, 189–223.

19. ELDAR, Y. C. & KUTYNIOK, G. (eds) (2012) *Compressed Sensing: Theory and Applications*. Cambridge: Cambridge University Press.

20. GIANNOPOULO, A. A. & MILMAN, V. D. (1997a) How small can the intersection of a few rotations of a symmetric convex body be? *C. R. Acad. Sci. Paris Sér. I Math.*, **325**, 389–394.

21. GIANNOPOULO, A. A. & MILMAN, V. D. (1997b) On the diameter of proportional sections of a symmetric convex body. *Int. Math. Res. Notices*, **1**, 5–19.

22. GIANNOPOULO, A. A. & MILMAN, V. D. (1998) Mean width and diameter of proportional sections of a symmetric convex body. *J. Reine Angew. Math.*, **497**, 113–139.

23. GIANNOPOULO, A. A. & MILMAN, V. D. (2004) Asymptotic convex geometry: Short overview. *Different faces of geometry*, International Mathematical Series, (S. Donaldson, Y. Eliashberg & M. Gromov eds), 87–162. New York: Springer.

24. GIANNOPOULO, A. A. & MILMAN, V. D. (2005) Asymptotic formulas for the diameter of sections of symmetric convex bodies. *J. Funct. Anal.*, **223**, 86–108.

25. GORDON, Y. (1988) On Milman's Inequality and Random Subspaces which Escape through a Mesh in $\mathbb{R}^n$. *Lecture Notes in Mathematics*, vol. 1317.

26. GROSS, D., LIU, Y.-K., FLAMMIA, S. T., BECKER, S. & EISERT, J. (2010) Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, **105**, 150401.

27. HOROWITZ, J. L. (2010) *Semiparametric and Nonparametric Methods in Econometrics*, vol. 692. New York: Springer.

28. HRISTACHE, M., JUDITSKY, A. & SPOKOINY, V. (2001) Direct estimation of the index coefficient in a single-index model. *Ann. Stat.*, **29**, 595–623.

29. JACQUES, L., LASKA, J., BOUFOUNOS, P. & BARANIUK, R. (2013) Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Trans. Inf. Theory*, **59**, 2082–2102.

30. KANNAN, R. & VEMPALA, S. (2009) *Spectral Algorithms*. Boston: Now Publishers Inc.

31. KESHAVAN, R. H., MONTANARI, A. & OH, S. (2010) Matrix completion from a few entries. *IEEE Trans. Inf. Theory*, **56**, 2980–2998.

32. KOLTCHINSKII, V., LOUNICI, K. & TSYBAKOV, A. B. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.*, **39**, 2302–2329.

33. LECUÉ, G. & MENDELSON, S. (2013) Learning subgaussian classes: upper and minimax bounds. Available at http://arxiv.org/abs/1305.4825.

34. LEDOUX, M. (2005) *The Concentration of Measure Phenomenon*, vol. 89. Providence: American Mathematical Society.

35. LEDOUX, M. & TALAGRAND, M. (1991) *Probability in Banach Spaces: Isoperimetry and Processes*, vol. 23. Berlin: Springer.

36. LI, K.-C. & DUAN, N. (1989) Regression analysis under link violation. *Ann. Stat.*, **17**, 1009–1052.

37. MENDELSON, S., PAJOR, A. & TOMCZAK-JAEGERMANN, N. (2007) Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric Funct. Anal.*, **17**, 1248–1282.

38. MILMAN, V. D. (1985a) Geometrical inequalities and mixed volumes in the local theory of Banach spaces. *Astérisque*, **131**, 373–400.

39. MILMAN, V. D. (1985) Random Subspaces of Proportional Dimension of Finite-Dimensional Normed Spaces: Approach through the Isoperimetric Inequality. *Lecture Notes in Mathematics*, vol. 1166.

40. NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M. & YU, B. (2012) A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Stat. Sci.*, **27**, 538–557.

41. OYMAK, S., THRAMPOULIDIS, C. & HASSIBI, B. (2013) The squared-error of generalized lasso: a precise analysis. *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pp. 1002–1009. IEEE.

42. PAJOR, A. & TOMCZAK-JAEGERMANN, N. (1986) Subspaces of small codimension of finite-dimensional Banach subspaces of small codimension of finite-dimensional Banach spaces. *Proc. Amer. Math. Soc.*, **97**, 637–642.

43. PLAN, Y. & VERSHYNIN, R. (2011) Dimension reduction by random hyperplane tessellations. *Discrete Comput. Geometry*, **51**, 1–24.

44. PLAN, Y. & VERSHYNIN, R. (2013a) One-bit compressed sensing by linear programming. *Commun. Pure Appl. Math.*, **66**, 1275–1297.
45. PLAN, Y. & VERSHYNIN, R. (2013b) Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. *IEEE Trans. Inf. Theory*, **59**, 482–494.
46. PLAN, Y. & VERSHYNIN, R. (2015) The generalized lasso with nonlinear observations. Available at http://arxiv.org/abs/1502.04071.
47. POWELL, J. L., STOCK, J. H. & STOKER, T. M. (1989) Semiparametric estimation of index coefficients. *Econo. J. Econ. Soc.*, **57**, 1403–1430.
48. RASKUTTI, G., WAINWRIGHT, M. J. & YU, B. (2011) Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory*, **57**, 6976–6994.
49. RECHT, B., FAZEL, M. & PARRILO, P. A. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**, 471–501.
50. SEGINER, Y. (2000) The expected norm of random matrices. *Comb. Prob. Comput.*, **9**, 149–166.
51. SREBRO, N., ALON, N. & JAAKKOLA, T. S. (2004) Generalization error bounds for collaborative prediction with low-rank matrices. In Advances In Neural Information Processing Systems, pp. 1321–1328.
52. STOKER, T. M. (1986) Consistent estimation of scaled coefficients. *Econ. J. Econ. Soc.*, **54**, 1461–1481.
53. TALAGRAND, M. (2005) *The Generic Chaining. Upper and Lower Bounds of Stochastic Processes*, vol. 154. Berlin: Springer.
54. THRAMPOULIDIS, C., ABBASI, E. & HASSIBI, B. (2015) Lasso with nonlinear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, pp. 3402–3410.
55. VERSHYNIN, R. (2012) Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing*, (Y. C. Elda & G. Kutyniok eds), Cambridge: Cambridge University Press, pp. 210–268.