ON THE SPARSITY OF LASSO MINIMIZERS IN SPARSE DATA RECOVERY

SIMON FOUCART, EITAN TADMOR, AND MING ZHONG

Dedicated to Ron DeVore with friendship and admiration

ABSTRACT. We present a detailed analysis of the unconstrained ℓ_1-weighted LASSO method for recovery of sparse data from its observation by randomly generated matrices, satisfying the Restricted Isometry Property (RIP) with constant $\delta < 1$, and subject to negligible measurement and compressibility errors. We prove that if the data is k-sparse, then the size of support of the LASSO minimizer, s, maintains a comparable sparsity, $s \leqslant C_\delta k$. For example, if $\delta = 0.7$ then $s < 11k$ and a slightly smaller $\delta = 0.4$ yields $s < 4k$. We also derive new ℓ_2/ℓ_1 error bounds which highlight precise dependence on k and on the LASSO parameter λ, before the error is driven below the scale of negligible measurement/ and compressiblity errors.

CONTENTS

1. INTRODUCTION

In 2006, the pioneering works of Candès, Romberg and Tao [13, 14] and of Donoho [21] suggested the framework of a constrained ℓ_1-method to recover a sparse unknown $\boldsymbol{x}_* \in \mathbb{R}^N$ from its observation $\boldsymbol{y}_* = A\boldsymbol{x}_* \in \mathbb{R}^m$[1]. The key point is that one can design observing matrices $A \in \mathbb{R}^{m\times N}$

Date: March 16, 2022.

2020 *Mathematics Subject Classification.* 94A12, 94A20.

Key words and phrases. Inverse problems, data recovery, compressive sensing, basis pursuit de-noising method, robust null space property.

Research was supported in part by NSF grants CCF-1934904, DMS-2053172 (SF), and DMS16-13911 (ET) and ONR grants N00014-2012787 (SF) and N00014-2112773 (ET).

[1]Earlier announcement of the works was presented in the workshops, organized together with Ron DeVore, which can be found at home.cscamm.umd.edu/programs/srs05/candes_srs05.pdf and home.cscamm.umd.edu/programs/srs05/donoho_srs05.htm.

with a relatively small number of observations, $m \ll N$, such that a constrained ℓ_1-method — also known as Basis Pursuit (BP) in [18, 16, 17] — finds a sparse solution as a minimizer of[2]

$$\boldsymbol{x}_{BP} := \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min} \left\{ |\boldsymbol{x}|_1 \;\middle|\; A\boldsymbol{x} = \boldsymbol{y}_* \right\}, \qquad A \in \mathbb{R}^{m\times N},\ m \ll N.$$

This is closely related to the well-known LASSO algorithm introduced in 1996 in the statistics literature [39], $\underset{|\boldsymbol{x}|_1\leqslant\delta}{\arg\min} \left\{ |\boldsymbol{y}_*^\epsilon - A\boldsymbol{x}|_2^2 \right\}$, which can be viewed as an ℓ_1-penalty relaxation of a least squares subject to (possibly noisy) observation $\boldsymbol{y}_*^\epsilon$.

The BP minimizer, $\boldsymbol{x}_{BP}$, recovers the sparse $\boldsymbol{x}_*$ when the observing matrix A satisfies an appropriate recoverability condition; we mention here the Restricted Isometry Property (RIP) introduced in [13], the ℓ_1-Coherence discussed in [40, 27, 22, 23], the restricted eigenvalue condition [6, §3], or the Null Space Property (NSP) of DeVore and his co-authors [19, 20], and related Robust Null Sparse Property (RNSP) of [26]. Important classes of such observing matrices with desired sparse recoverability conditions are randomly generated, e.g., [26, §9].

1.1. **Statement of main results.** Throughout the paper we will be using the two notions of sparsity and compressibility. A vector $\boldsymbol{x} \in \mathbb{R}^N$ is *sparse* if

$$s_{\boldsymbol{x}} := |\boldsymbol{x}|_0 \ll N.$$

In applications, sparsity is often difficult to acquire, and clean observations are not always available, since the observation process is inevitably and easily corrupted by errors — human and/or machine measurement errors. We turn our attention to the recovery of compressible unknown from its *noisy* observations. A vector $\boldsymbol{x} \in \mathbb{R}^N$ is *compressible* of order k, or simply k-compressible, if its content is faithfully captured by a k-sparse vector — specifically, if its ℓ_1-distance to the set of all k-sparse vectors,

$$\sigma_k(\boldsymbol{x}) := \inf_{\boldsymbol{z}\in\mathbb{R}^N} \left\{ |\boldsymbol{x} - \boldsymbol{z}|_1 : |\boldsymbol{z}|_0 \leqslant k \right\}, \tag{1.1}$$

is small relative to $|\boldsymbol{x}|_1$. We note that $\sigma_k(\boldsymbol{x})$ is realized by a (not necessarily unique) vector, denoted $\boldsymbol{x}(k)$, whose non-zero entries are the k largest of $\boldsymbol{x}$ in absolute value.

Let $\boldsymbol{x}_*$ be a compressible unknown of order k so that $\sigma_k(\boldsymbol{x}_*) \ll |\boldsymbol{x}_*|_1$, and assume we only have access to its measured observation $\boldsymbol{y}_*^\epsilon = A\boldsymbol{x}_* + \boldsymbol{\varepsilon}$. The term $\boldsymbol{\varepsilon}$ is the measurement error caused by a number of factors which are assumed statistically independent of the unknown $\boldsymbol{x}_*$ and the observing operator A. The details of $\boldsymbol{\varepsilon}$ remain untraceable except for its size which is assumed to be negligibly small. In this case, one should not expect an exact recovery of a sparse $\boldsymbol{x}_*$, but instead, accept an approximate solution, $\boldsymbol{y}_*^\epsilon = A\boldsymbol{x}_*(k) + \boldsymbol{\varepsilon}'$, where $\boldsymbol{\varepsilon}' = A(\boldsymbol{x}_* - \boldsymbol{x}_*(k)) + \boldsymbol{\varepsilon}$ is adapted to the small scale built into the problem, which consists of two contributions — the small measurement error, $\epsilon := |\boldsymbol{\varepsilon}|_2$, and the small compressibility error,[3] $|A(\boldsymbol{x}_* - \boldsymbol{x}_*(k))|_2 \leqslant \sigma_k(\boldsymbol{x}_*)$,

$$\boldsymbol{y}_*^\epsilon = A\boldsymbol{x}_*(k) + \boldsymbol{\varepsilon}', \qquad |\boldsymbol{\varepsilon}'|_2 \leqslant \mu,$$

such that $\mu = \sigma_k(\boldsymbol{x}_*) + \epsilon$ is much smaller relative to the unknown data, $\mu \ll |\boldsymbol{x}_*|_1$. Although the observing operator A is linear, the recovery of $\boldsymbol{x}_*$ by a direct "solution" of the linear problem $A\boldsymbol{x} = \boldsymbol{y}_*^\epsilon$ is ill-posed, unless additional conditions on A and $\boldsymbol{x}_*$ are enforced so that the unknown

[2]Given $x \in \mathbb{R}^N$ we let $|\boldsymbol{x}|_p$ denote its ℓ_p-norm, with the usual conventional limiting cases of $p = \infty$ and $p = 0$, where $|\boldsymbol{x}|_\infty := \max\limits_{1\leqslant i\leqslant N} |x_i|$, and respectively $|\boldsymbol{x}|_0 := |\mathrm{supp}(\boldsymbol{x})|$ where $|\cdot|$ is the cardinality of a finite set.

[3]The columns of A are assumed ℓ_2-normalized so that $|A|_{1\to 2} = 1$.

object $\boldsymbol{x}_*$, or at least a faithful approximation of it, is recovered by solving an augmented *well-posed* regularized minimization problem. On the way, the original linear problem is replaced by a nonlinear procedure. To capture the compressible information of $\boldsymbol{x}_*$ from its noisy observation $\boldsymbol{y}_*^\epsilon$, we seek a minimizer of the unconstrained ℓ_1-regularized Least Squares problem,

$$\boldsymbol{x}_\lambda := \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min}\Big\{\lambda|\boldsymbol{x}|_1 + \frac{1}{2}|\boldsymbol{y}_*^\epsilon - A\boldsymbol{x}|_2^2\Big\}, \qquad A\in\mathbb{R}^{m\times N},\ m\ll N. \tag{1.2}$$

The unconstrained variational statement (1.2) falls under the general class of Tikhonov regularization. The distinctive feature is the ℓ_1-regularization, leading to an approximate decomposition of the basis pursuit of Chen & Donoho [18], $\boldsymbol{y}_*^\epsilon = A\boldsymbol{x}_\lambda + \boldsymbol{r}_\lambda$ with (hopefully) small residual, $\boldsymbol{r}_\lambda = \boldsymbol{y}_*^\epsilon - A\boldsymbol{x}_\lambda$, depending on a parameter λ. This version of ℓ_1-regularization, called "Basis Pursuit De-Noising" in [16], which became known as the unconstrained ℓ_1-weighted LASSO, is the main focus of our work. As noted in the 1996 thesis [16], the work on this version of BP was motivated by a series of ideas using ℓ_0/ℓ_1-based regularization that appeared in early 1990s, primarily the empirical atomic decomposition of Donoho and Johnstone [24], $\underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min}\Big\{\lambda|\boldsymbol{x}|_0 + \frac{1}{2}|\boldsymbol{y}_*^\epsilon - A\boldsymbol{x}|_2^2\Big\}$, the multi-scale edge representation with wavelets of Hwang and Mallat [29] and the TV-based de-noising method of ROF [31], $\underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min}\Big\{\lambda|\boldsymbol{x}|_{TV} + \frac{1}{2}|\boldsymbol{y}_*^\epsilon - A\boldsymbol{x}|_2^2\Big\}$. These works were later further explored as the Lagrangian formulation of the quadratically constrained Basis Pursuit de-noising [17, 14] and the noise-aware ℓ_1-minimization [26].

Since $\lambda > 0$ controls the distance between $A\boldsymbol{x}_\lambda$ and $\boldsymbol{y}_*$, the parameter λ can be interpreted as a regularization *scale*. In a subsequent work, [37], we pursue a *multi-scale* generalization based on a ladder of hierarchical scales constructed by the Hierarchical Decomposition (HD) method [34, 35, 36, 33]. The goal of this work is to analyze the sparsity behavior of the *mono-scale* LASSO (1.2), observed by a sub-class of RIP matrices satisfying the Robust Null Space Property (RNSP) which is discussed in section 2. Our main results, outlined and proved in section 3, are summarized in the following. Our results involve three main parameters: the Restricted Isometry Constant (RIC) in (2.2) below, $\delta = \delta_k < 1$, the related RNSP constant, $\beta_\delta = \dfrac{\sqrt{1+\delta}}{\sqrt{1-\delta^2} - \delta/4}$, depending on the RIC δ, and the small scale of compressibility+measurement, $\mu = \sigma_k(\boldsymbol{x}_*) + \epsilon$.

Theorem 1.1 (Main result). *Let $\boldsymbol{x}_*$ be k-compressible, and let $\boldsymbol{y}_*^\epsilon = A\boldsymbol{x}_* + \boldsymbol{\varepsilon}$ be its observation with observing matrix A satisfying the RIP (2.2) with constant δ large enough, $\delta > \delta_t$, such that (3.7) below holds. Let $\boldsymbol{x}_\lambda$ be the LASSO minimizer (1.2).*

(i) **(Sparsity).** *The sparsity of the LASSO minimizer, $s_\lambda = s_{\boldsymbol{x}_\lambda}$, does not exceed*

$$s_\lambda < (1+\delta)\Big(\beta_\delta\sqrt{k} + \frac{2\mu}{\lambda}\Big)^2.$$

(ii) **(ℓ_2-error bound).** *The following ℓ_2-error bound holds*

$$\frac{1}{\sqrt{1+\delta}}\Big(\frac{\sqrt{s_\lambda}\lambda}{\sqrt{1+\delta}} - \mu\Big) \leqslant |\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2 \leqslant \frac{1}{\sqrt{1-\delta}}\big(\beta_\delta\sqrt{k}\lambda + 3\mu\big).$$

(iii) **(ℓ_1-error bound).** *The following ℓ_1-error bound holds*

$$|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 < \frac{\sqrt{1+\delta}}{\sqrt{1-\delta}}\frac{1}{\lambda}\big((\beta_\delta + 1/2)\sqrt{k}\lambda + 2\mu\big)^2.$$

We interpret these bounds as follows. Set $\theta = 2\mu/\sqrt{k}\lambda$, then (i) reads

$$s_\lambda \leqslant \chi^2 k, \qquad \chi = \sqrt{1+\delta}(\beta_\delta + \theta).$$

Thus, if $\theta \leqslant 1$ — namely, as long as λ does not get exceedingly small so that $\lambda \geqslant 2\mu/\sqrt{k}$, then the sparsity of $\boldsymbol{x}_\lambda$ is comparable to the sparsity of $\boldsymbol{x}_*$. Furthermore, in (ii) we have the ℓ_2-error bound of order $\lesssim \sqrt{k}\lambda + \mu$ and in (iii), an ℓ_1-error bound of order $\lesssim k\lambda + \sqrt{k}\mu$.

We conclude with a few comments on theorem 1.1. The sparsity bound in (*i*) with RIC $\delta = 0.7$ yields $s_\lambda < 11k$, while a slightly smaller RIC $\delta = 0.4$ yields $s_\lambda < 4k$. This should be compared with the sparsity bounds in [5, Theorem 3] and [32]. The ℓ_2-upper bound on the right of (*ii*) is not new; here we recover the ℓ_2-bound, derived under appropriate assumptions, in [11], [6, Theorem 7.1] and [30, 38, 28]. This should be contrasted with the ℓ_2-error *lower-bound* on the left, derived in section 4.1 (see figure 4.1). Indeed, this ℓ_2 lower-bound is the essential ingredient in our proof of the sparsity bound in (*i*). Finally, the ℓ_1-bound in (*iii*) with RIC $\delta = 0.7$ yields $|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 < 16.97k\lambda + 24.04\sqrt{k}\mu$. Here, the linear decay with λ is not new and can be found for example, under various assumptions, in [6, Theorem 7.1] and [9, Theorem 6.1].

2. The Robust Null Space Property

Optimality of the minimizer. The variational problem (1.2) admits a minimizer, $\boldsymbol{x}_\lambda$, and at least for certain relevant classes of full row rank A's, the minimizer is unique, [41]. The minimizer is completely characterized by its residual, $\boldsymbol{r}_\lambda := \boldsymbol{y}_*^\epsilon - A\boldsymbol{x}_\lambda$ (to simplify notations we suppress the dependence of $\boldsymbol{r}_\lambda$ on ϵ). We summarize the results from [35, §2.1],[33, Appendix] where we distinguish between two cases.

- **(i)** If $\lambda \geqslant \lambda_\infty := |A^\top \boldsymbol{y}_*^\epsilon|_\infty$ then (1.2) admits only the trivial minimizer $\boldsymbol{x}_\lambda \equiv \boldsymbol{0}$. In this case, λ is too large to extract the compressibility information in $\boldsymbol{y}_*^\epsilon$.
- **(ii)** If $\lambda < \lambda_\infty = |A^\top \boldsymbol{y}_*^\epsilon|_\infty$ then (1.2) admits a non-trivial minimizer, $\boldsymbol{x}_\lambda$, with the corresponding residual, $\boldsymbol{r}_\lambda = \boldsymbol{y}_*^\epsilon - A\boldsymbol{x}_\lambda$, such that $(\boldsymbol{x}_\lambda, \boldsymbol{r}_\lambda)$ forms an *extremal pair* in the sense that

$$\langle A\boldsymbol{x}_\lambda, \boldsymbol{r}_\lambda\rangle = \lambda|\boldsymbol{x}_\lambda|_1 \ \text{ and } \ |A^\top \boldsymbol{r}_\lambda|_\infty = \lambda. \tag{2.1}$$

To proceed we will need the following notations. The restriction of a vector $\boldsymbol{w} \in \mathbb{R}^N$ on an index set $\mathcal{K} \subset \{1, 2, \ldots, N\}$ of size $k = |\mathcal{K}|$ is denoted $\boldsymbol{w}_\mathcal{K} := \{w_i,\ i \in \mathcal{K}\} \in \mathbb{R}^k$. Similarly, given a matrix $W \in \mathbb{R}^{m\times N}$ with columns $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots$, its restriction on an index set $\mathcal{K}$ of size $k = |\mathcal{K}|$ consists of the k columns $W_\mathcal{K} := \text{col}\{\boldsymbol{w}_i, i \in \mathcal{K}\}$. The size of W can be measured by its induced matrix norm, $\|W\|_p = \sup_{|\boldsymbol{w}|_p=1} |W\boldsymbol{w}|_p$. The signum vector is defined component-wise, $\mathbf{sgn}(\boldsymbol{w})_i = \text{sgn}(w_i)$, in terms of the usual signum function $\text{sgn}(w) = \left\{ \begin{array}{cc} -1, & w<0 \\ 1, & w>0 \end{array} \right\}$ for $w \neq 0$.

Restricted Isometry Poperty (RIP). A matrix A satisfies the Restricted Isometry Property (RIP) of order k with Restricted Isometry Constant (RIC) $\delta_k < 1$ if the following holds, [15, 21, 13, 7],

$$(1-\delta_k)|\boldsymbol{x}|_2^2 \leqslant |A\boldsymbol{x}|_2^2 \leqslant (1+\delta_k)|\boldsymbol{x}|_2^2, \qquad \forall |\boldsymbol{x}|_0 \leqslant k. \tag{2.2}$$

Throughout the paper we adopt the usual assumption that δ_k is measured for A's with ℓ^2-normalized *columns*[4]. There are two classes of matrices $A \in \mathbb{R}^{m\times N}$ satisfying the RIP of order k: deterministic

[4]The RIP of A asserts that for any subset of its k columns, $\{\boldsymbol{a}_i\}_{i\in\mathcal{K}}$, the entries $|\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle|_{i\neq j} \lesssim \delta_k$ while $|\boldsymbol{a}_i|_2^2 = 1+\epsilon_i$ such that $|\epsilon_i| \lesssim \delta_k$. Therefore, one can always re-normalize the columns of A by a factor$\lesssim (1-\delta_k)^{-1/2}$ yielding a new RIP matrix with ℓ_2-normalized columns and with possibly slightly larger RIP constant $\delta_k' \lesssim \delta_k/(1-\delta_k)$.

A's with number of observations $m \gtrsim k^2$ (the quadratic bottleneck is lessened in [8]); and a large class of randomly generated A's for which the restriction on the number of observations can be further lessened to having only m observations, [26, §9.4]

$$m \sim Const \cdot \delta^{-2} k \ln\left({}^{eN}/_{k}\right). \tag{2.3}$$

Candès proved the exactness of the constrained BP for RIP matrices with $\delta < \sqrt{2}-1$, [12]. Further refinements were reported in [25] before the definitve result of [10].

Robust Null Space Property (RNSP). A crucial step in quantifying the recovery error of $\boldsymbol{x}_*$ using (1.2) is to enforce a recoverability condition on the observing matrix A. This brings us to the Robust Null Sparse Property (RNSP) introduced in [26, §4.3]. A matrix $A \in \mathbb{R}^{m\times N}$ satisfies the RNSP of order k with constants $0 < \rho < 1$ and $\tau > 0$, if for all $\mathcal{K} \subset \{1, 2, \ldots, N\}$ of size $|\mathcal{K}| \leqslant k$, there holds

$$|\boldsymbol{x}_{\mathcal{K}}|_1 \leqslant \rho |\boldsymbol{x}_{\mathcal{K}^c}|_1 + \tau |A\boldsymbol{x}|_2, \qquad \forall \boldsymbol{x} \in \mathbb{R}^N. \tag{2.4}$$

We refer to the "RNSP$_{\rho,\tau}$ of order k", and unless needed, we suppress the dependence of (ρ, τ) on k. In particular, given a k-sparse $\boldsymbol{v}$ and any $\boldsymbol{u}$, we apply (2.4) to $\boldsymbol{x} = \boldsymbol{u} - \boldsymbol{v}$ with $\mathcal{K} = \mathrm{supp}(\boldsymbol{v})$, where $|\boldsymbol{x}_{\mathcal{K}}|_1 - |\boldsymbol{x}_{\mathcal{K}^c}|_1 \geqslant |\boldsymbol{v}|_1 - |\boldsymbol{u}|_1$ yields the following useful consequence of RNSP.

Lemma 2.1. *If $A \in \mathbb{R}^{m\times N}$ satisfies the RNSP$_{\rho,\tau}$ of order k, then for all k-sparse $\boldsymbol{v}$'s and any $\boldsymbol{u}$,*

$$|\boldsymbol{v}|_1 - |\boldsymbol{u}|_1 \leqslant \tau |A(\boldsymbol{u} - \boldsymbol{v})|_2, \qquad |\mathit{supp}(\boldsymbol{v})| \leqslant k. \tag{2.5}$$

As an example for the class of observation matrices satisfying the RNSP$_{\rho,\tau}$ of order k, we mention the class of randomly generated RIP matrices with RICs $\delta = \delta_{2k}$, [26, Theorem 6.13],

$$\rho = \frac{\delta}{\sqrt{1-\delta^2} - {}^{\delta}/_{4}} \quad \text{and} \quad \tau = \beta\sqrt{k}, \quad \beta := \frac{\sqrt{1+\delta}}{\sqrt{1-\delta^2} - {}^{\delta}/_{4}}, \qquad \delta = \delta_{2k}. \tag{2.6}$$

These RNSP parameters, (ρ, β), are dictated as increasing functions of the RIC $\delta < 1$. A smaller δ requires an increased number of observations. All proofs invoke different classes of observing matrices which are randomly generated so that they satisfy a desirable observing properties—RIP, RNSP, or Constrained Minimal Singular Values (CMSV) property. Accordingly, the error statements are probabilistic in nature, referring to the ensemble of these observations.

3. On the sparsity of the unconstrained LASSO minimizer

We analyze the sparsity and ℓ_1/ℓ_2-error bounds of the minimizer (1.2) in recovering $\boldsymbol{x}_*(k)$ from the observation $\boldsymbol{y}_*^{\epsilon} = A\boldsymbol{x}_* + \boldsymbol{\varepsilon}$, with small measurement error, $\epsilon = |\boldsymbol{\varepsilon}|_2$, and — assuming that $\boldsymbol{x}_*$ is k-compressible — with small compressibility error, $\sigma_k(\boldsymbol{x}_*) = |\boldsymbol{x}_* - \boldsymbol{x}_*(k)|_1$. Set

$$\mu := \sigma_k(\boldsymbol{x}_*) + \epsilon.$$

Clearly, since the exact solution is observed up to ℓ_2 residual error of order $|\boldsymbol{y}_*^{\epsilon} - A\boldsymbol{x}_*|_2 \leqslant \mu$, we do not have much to say when the computed residual error $|\boldsymbol{r}_\lambda|_2$ is of order μ and we will therefore limit ourselves to the parametric regime where $|\boldsymbol{r}_\lambda|_2 \gg \mu$. Below we show that $|\boldsymbol{r}_\lambda|_2 \sim \lambda\sqrt{k}$ and therefore throughout the paper we make the assumption

$$\theta := \frac{2\mu}{\lambda\sqrt{k}} \leqslant 1, \qquad \mu = \sigma_k(\boldsymbol{x}_*) + \epsilon. \tag{3.1}$$

Thus, we assume the LASSO weight, λ, does not get exceedingly small, $\lambda \geqslant {}^{2\mu}/_{\sqrt{k}}$. In concrete examples demonstrating the sparsity and error bounds reported below we use $\theta = 0.1$, corresponding to $\lambda \geqslant {}^{20\mu}/_{\sqrt{k}}$.

Lemma 3.1 (The re-scaled residual — an upper-bound). *Fix $\lambda < \lambda_\infty := |A^\top \boldsymbol{y}_*^\epsilon|_\infty$. Let $\boldsymbol{y}_*^\epsilon = A\boldsymbol{x}_* + \boldsymbol{\varepsilon}$ be the observation of a k-compressible unknown $\boldsymbol{x}_* \in \mathbb{R}^N$, observed by $A \in \mathbb{R}^{m\times N}$ satisfying the RNSP$_{\rho,\tau}$ of order k, (2.6). Let μ denote the small scale of k-compressiblity and measurement errors, see (3.1). Then the residual of the LASSO (1.2), $\boldsymbol{r}_\lambda = \boldsymbol{y}_*^\epsilon - A\boldsymbol{x}_\lambda$, satisfies*

$$\frac{|\boldsymbol{r}_\lambda|_2}{\lambda} \leqslant (\beta_\delta + \theta)\sqrt{k}, \qquad \beta_\delta = \frac{\sqrt{1+\delta}}{\sqrt{1-\delta^2} - \delta/4}. \tag{3.2}$$

Proof. Clearly, $|A(\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k))|_2 \leqslant |\boldsymbol{r}_\lambda|_2 + \mu$. Using (2.5) with the k-sparse $\boldsymbol{v} = \boldsymbol{x}_*(k)$ and $\boldsymbol{u} = \boldsymbol{x}_\lambda$ yields

$$|\boldsymbol{x}_*(k)|_1 - |\boldsymbol{x}_\lambda|_1 \leqslant \tau |A(\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k))|_2 \leqslant \tau |\boldsymbol{r}_\lambda|_2 + \tau\mu. \tag{3.3}$$

Next, a lower-bound for the quantity on the left follows. Recall that $\boldsymbol{x}_\lambda$, being the LASSO minimizer (1.2), is characterized by the extremal property that its scaled residual $\boldsymbol{z} = \dfrac{\boldsymbol{r}_\lambda}{\lambda}$ satisfies (2.1),

$$|\boldsymbol{x}_\lambda|_1 = \langle A\boldsymbol{x}_\lambda, \boldsymbol{z}\rangle \ \text{ and } \ |A^\top \boldsymbol{z}|_\infty = 1, \qquad \boldsymbol{z} := \frac{\boldsymbol{r}_\lambda}{\lambda}.$$

Hence

$$\begin{aligned} |\boldsymbol{x}_*(k)|_1 - |\boldsymbol{x}_\lambda|_1 &\geqslant \langle \boldsymbol{x}_*(k), A^\top \boldsymbol{z}\rangle - \langle A\boldsymbol{x}_\lambda, \boldsymbol{z}\rangle = \langle A\boldsymbol{x}_*(k) - A\boldsymbol{x}_\lambda, \boldsymbol{z}\rangle \\ &= \langle \boldsymbol{r}_\lambda, \boldsymbol{z}\rangle + \langle A\boldsymbol{x}_*(k) - \boldsymbol{y}_*^\epsilon, \boldsymbol{z}\rangle \\ &\geqslant \frac{|\boldsymbol{r}_\lambda|_2^2}{\lambda} - |A\boldsymbol{x}_*(k) - \boldsymbol{y}_*^\epsilon|_2 \frac{|\boldsymbol{r}_\lambda|_2}{\lambda}. \end{aligned}$$

Now, assumption (3.1) and the fact that $|A|_{1\to 2} \leqslant 1$ imply,

$$|A\boldsymbol{x}_*(k) - \boldsymbol{y}_*^\epsilon|_2 \leqslant |A\boldsymbol{x}_* - \boldsymbol{y}_*^\epsilon|_2 + |A(\boldsymbol{x}_*(k) - \boldsymbol{x}_*)|_2 \leqslant \epsilon + \sigma_k(\boldsymbol{x}_*) = \mu,$$

and we end with the desired lower-bound

$$|\boldsymbol{x}_*(k)|_1 - |\boldsymbol{x}_\lambda|_1 \geqslant \frac{|\boldsymbol{r}_\lambda|_2^2}{\lambda} - \mu\frac{|\boldsymbol{r}_\lambda|_2}{\lambda}. \tag{3.4}$$

Combining (3.3) and (3.4) we conclude that $|\boldsymbol{z}|_2 = \dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda}$ satisfies the quadratic inequality, $|\boldsymbol{z}|_2^2 \leqslant \left(\tau + \dfrac{\mu}{\lambda}\right)|\boldsymbol{z}|_2 + \tau\dfrac{\mu}{\lambda}$, and therefore

$$\frac{|\boldsymbol{r}_\lambda|_2}{\lambda} = |\boldsymbol{z}|_2 < \tau + \frac{2\mu}{\lambda} = (\beta_\delta + \theta)\sqrt{k}, \tag{3.5}$$

proving (3.2). □

3.1. **Bounds of the sparsity.** We now come to the main point of the lower-bound on the scaled residual in terms of the size of the support of $\boldsymbol{x}_\lambda$, $\dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda} \gtrsim \sqrt{s_\lambda}$. Fix $\lambda < \lambda_\infty := |A^\top \boldsymbol{y}_*^\epsilon|_\infty$. Recall that if $\boldsymbol{x}_\lambda$ is the LASSO minimizer (1.2) then by the extremal property (2.1), the scaled residual $\boldsymbol{z} = \dfrac{\boldsymbol{r}_\lambda}{\lambda}$ satisfies the two properties $\langle A\boldsymbol{x}_\lambda, \boldsymbol{z}\rangle = |\boldsymbol{x}_\lambda|_1$ and $|A^\top \boldsymbol{z}|_\infty = 1$. Thus, the extremal $\boldsymbol{x}_\lambda$ with support $\mathcal{S} = \operatorname{supp}(\boldsymbol{x}_\lambda)$ of size $s_\lambda = |\boldsymbol{x}_\lambda|_0$, is identified by a re-scaled residual satisfying

$$(A^\top \boldsymbol{z})_\mathcal{S} = \mathbf{sgn}(\boldsymbol{x}_{\lambda,\mathcal{S}}), \qquad \boldsymbol{z} = \frac{\boldsymbol{r}_\lambda}{\lambda}, \quad \mathcal{S} = \operatorname{supp}(\boldsymbol{x}_\lambda). \tag{3.6}$$

Fix the integer t,

$$t := [(1+\delta)(\beta_\delta + \theta)^2 k] + 1 \ \text{ with constant } \ \delta > \delta_t. \tag{3.7}$$

Since the RIC δ_t is increasing with the order t, there is no need to trace a precise fixed point associated with (3.7), $t = [(1+\delta_t)(\beta_{\delta_t}+\theta)^2 k] + 1$. Instead, we can use a priori bounds of δ_t; for example, if we restrict ourselves to the range $\delta < 0.7$, we can set the integer upper bound $t = 11k$. Below, we demonstrate refined versions of this bound.
We claim that

$$s_\lambda < t = [(1+\delta)(\beta_\delta+\theta)^2 k] + 1. \tag{3.8}$$

To this end we proceed by contradiction. Assume $s_\lambda \geqslant t$. Then the support of $\boldsymbol{x}_\lambda$ has a subset $\mathcal{T}$ of size t for which the extremal property (3.6) reads $(A^\top \boldsymbol{z})_\mathcal{T} = \mathbf{sgn}(\boldsymbol{x}_{\lambda,\mathcal{T}})$, and the RIP (2.2) for such set $\mathcal{T}$ implies

$$|A(A^\top \boldsymbol{z})_\mathcal{T}|_2^2 \leqslant (1+\delta_t)|(A^\top \boldsymbol{z})_\mathcal{T}|_2^2 = (1+\delta_t)|\,\mathbf{sgn}(\boldsymbol{x}_{\lambda,\mathcal{T}})|_2^2 = (1+\delta_t)t. \tag{3.9}$$

On the other hand, we have

$$|A(A^\top \boldsymbol{z})_\mathcal{T}|_2^2 \geqslant \frac{1}{|\boldsymbol{z}|_2^2}\big\langle A(A^\top \boldsymbol{z})_\mathcal{T}, \boldsymbol{z}\big\rangle^2 = \frac{1}{|\boldsymbol{z}|_2^2}\big\langle (A^\top \boldsymbol{z})_\mathcal{T}, A^\top \boldsymbol{z}\big\rangle^2 = \frac{1}{|\boldsymbol{z}|_2^2}|(A^\top \boldsymbol{z})_\mathcal{T}|_2^4 = \frac{t^2}{|\boldsymbol{z}|_2^2}.$$

The last two inequalities followed by Lemma 3.1 imply $t \leqslant (1+\delta_t)|\boldsymbol{z}|_2^2 < (1+\delta)(\beta_\delta+\theta)^2 k$, which contradicts the definition of t,

$$t = [(1+\delta)(\beta_\delta+\theta)^2 k] + 1 \geqslant (1+\delta_t)(\beta_\delta+\theta)^2 k.$$

Thus, (3.8) holds.
In fact, a refined statement follows. Now that we know $|\mathcal{S}| \leqslant [(1+\delta)(\beta_\delta+\theta)^2 k]$ we can argue along the same line as above with $\mathcal{T} = \mathcal{S}$, obtaining $s_\lambda \leqslant (1+\delta)|\boldsymbol{z}|_2^2$.

Lemma 3.2 (The re-scaled residual — a lower-bound). *Fix $\lambda < \lambda_\infty := |A^\top \boldsymbol{y}_*^\epsilon|_\infty$ and let $\boldsymbol{x}_\lambda$ be the s_λ-sparse minimizer of the corresponding LASSO (1.2), observed with RIC δ such that (3.7) holds. Then the residual, $\boldsymbol{r}_\lambda = \boldsymbol{y}_*^\epsilon - A\boldsymbol{x}_\lambda$, satisfies*

$$\frac{|\boldsymbol{r}_\lambda|_2^2}{\lambda^2} \geqslant \frac{s_\lambda}{1+\delta}. \tag{3.10}$$

Combining the lower- and upper-bounds of $\dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda}$ we conclude the following.

Theorem 3.3 (Sparsity bound). *Fix $\lambda < \lambda_\infty := |A^\top \boldsymbol{y}_*^\epsilon|_\infty$ and let $\boldsymbol{x}_\lambda$ be the s_λ-sparse minimizer of the corresponding LASSO (1.2), observed with RIC δ such that (3.7) holds. Then*

$$\frac{s_\lambda}{1+\delta} \leqslant \frac{|\boldsymbol{r}_\lambda|_2^2}{\lambda^2} \leqslant (\beta_\delta+\theta)^2 k, \qquad \delta > \delta_t, \quad t = [(1+\delta_t)(\beta_{\delta_t}+\theta)^2 k] + 1. \tag{3.11}$$

In particular, we recover (3.8), $s_\lambda \leqslant [(1+\delta)(\beta_\delta+\theta)^2 k]$.

We demonstrate the application of corollary 3.3 for different choices of RICs. In all cases, we set $\theta = 0.1$. We begin with the RIC $\delta = 0.7$, obtaining $(\beta_\delta+\theta) = 2.52 \rightsquigarrow s_\lambda \leqslant (1+\delta)(\beta_\delta+\theta)^2 k < 11k$. Thus, with $t = 11k$ we require $\delta_{11k} < 0.7$ which in turn, by (2.3), set the number of required observations

$$s_\lambda < 11k: \qquad m \approx Const.\frac{11k}{0.7^2}\ln(eN/k) \approx Const.22.4\,k\ln(eN/k).$$

For a second example we choose a smaller RIC $\delta = 0.4$ and $\theta = 0.1$. Recall, that a smaller δ requires more observations yet in the number of observations in the present context depends on δ_t.

In this case $(\beta_\delta + \theta) = 1.55 \rightsquigarrow s_\lambda \leqslant (1+\delta)(\beta_\delta+\theta)^2 k = 3.36k < 4k$. This requires a slightly larger number of observations (or at least a smaller bound (2.3))

$$s_\lambda < 4k: \qquad m \approx Const. \frac{4k}{0.4^2} \ln(eN/k) \approx Const. 25\, k \ln(eN/k).$$

Finally, as a third example we choose an even smaller the RIC $\delta = 0.26$ and the same $\theta = 0.1$. In this case $(\beta_\delta + \theta) = 1.35 \rightsquigarrow s_\lambda \leqslant (1+\delta)(\beta_\delta+\theta)^2 k = 2.28k < 3k$, and this yields the number of required observations

$$s_\lambda < 3k: \qquad m \approx Const. \frac{3k}{0.26^2} \ln(eN/k) \approx Const. 44.4\, k \ln(eN/k).$$

Remark 3.4 (On the threshold parameter χ). *Observe that the sparsity bound s_λ is uniform in the small scale μ throughout the parametric regime assumed in (3.1). Thus, in the range of $\lambda \gg 2\mu/\sqrt{k}$, the support of the computed solution $|\boldsymbol{x}_\lambda|_0$ can grow at most by a fixed factor relative to the k-support of underlying unknown $\boldsymbol{x}_*$,* [38, Appendix A]. *We write*

$$s_\lambda < ([\chi^2]+1)k, \qquad \chi := \sqrt{1+\delta}(\beta_\delta+\theta) = \frac{1+\delta}{\sqrt{1-\delta^2-\delta/4}} + \sqrt{1+\delta}\theta. \tag{3.12}$$

We have the theoretical bounds $[\chi^2]+1 = 11$ corresponding to $\delta = 0.7$ and $[\chi^2]+1 = 4$ corresponding to $\delta \approx 0.4$.

3.2. **Numerical simulations.** We report here on our simulations of the unconstrained LASSO (1.2), applied to the recovery of k-sparse data, $\sigma_k = 0$, that is $\mu = \epsilon$, with $(k, m, N) = (160, 1024, 4096)$. We consider different levels of noise $\epsilon = 10^{-3}, 10^{-2}, 10^{-1}$, in the corresponding parametric regime (3.1), $\lambda > 2\mu/\sqrt{k} = 0.16\epsilon$. The results are obtained by averaging 100 observations using randomly generated RNSP$_{\rho,\tau}$ matrices based on Gaussian distributions. A simple proof of the RIP for such matrices can be found in [4]. The results are compared with the sparsity bound of theorem 3.3. We note that our sparsity bound depends in an essential manner on the the RICs, $1 \pm \delta$, in (2.2). The parametric regime in (2.3) provides only a rough estimate on the range of allowable RICs, and in particular, does not cover the parameters used in the simulations below, [26, §9.4]. A detailed study which traces the sharp RICs can be found in [2, 3], but is beyond the scope of our work. We compare the simulations with our sparsity bound based on the RIC $\delta = 0.7$. This is partly motivated by the result of [10] in which the authors prove an exact BP recovery of k-sparse data from the RIP with $\delta_{tk} < \sqrt{(t-1)/t}$. In our case, the computation reported in figure 3.1 indicates the *actual* sparsity $s_\lambda < tk$ with $t = 2$ which is consistent with $\delta < \sqrt{1/2} \approx 0.7$. Although the RIC $\delta = 0.7$ does not provide a tight bound, $s_\lambda < 11k$, it suffices to detect the correct *behavior* of the LASSO minimizer, reported in figures 3.1–3.3 and 4.1–4.2.
We record here the corresponding parameters involved in our bounds:

$$\beta_{\delta}{}_{|\delta=0.7} = \frac{\sqrt{1+\delta}}{\sqrt{1-\delta^2-\delta/4}} = 2.42, \quad \chi_{|\delta=0.7} = \sqrt{1+\delta}(\beta_\delta+\theta) = 3.16, \quad \eta_{|\delta=0.7} = \frac{1}{1+\delta} = 0.59$$

Our main result on the sparsity of the LASSO minimizer in theorem 3.3 provides a reasonably accurate information about the behavior of the unconstrained LASSO minimization (1.2). Figure 3.1 shows the behavior of the support, $s_\lambda = |\boldsymbol{x}_\lambda|_0$, starting with $s_\lambda = 0$ for $\lambda > \lambda_\infty$ and monotonically increasing as λ decreases all the way to a critical value, $\lambda_c \sim 0.11$, at which point s_{λ_c} reaches its maximal value of 215. This should be compared with our bound $s_\lambda \leqslant (1+\delta)(\beta_\delta+\theta)^2 k$. For $\delta = 0.7$ we have $s_\lambda \leqslant 11k$, which is a rough sparsity bound, relative to the actual $s_\lambda \sim 215$. A smaller RIC $\delta \sim 0.2$ yields a tighter sparsity bound $1.66k \sim 313$.

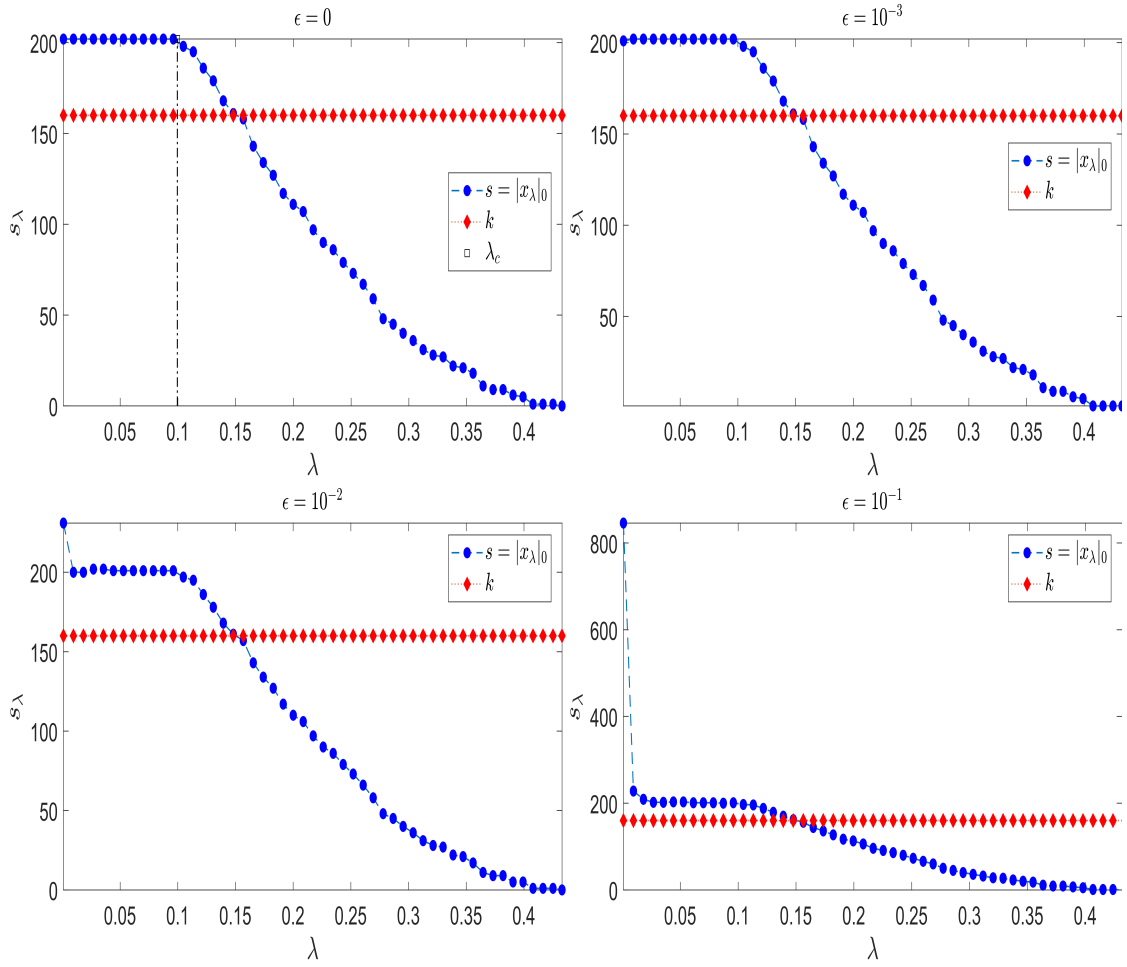

FIGURE 3.1. The support for computed minimizer $s_\lambda = |\boldsymbol{x}_\lambda|_0$ of k-sparse data, $k = 160$, peaks at the threshold value of $k_{\max} \sim 215$ when λ reaches $\lambda_c \sim 0.11$. This should be compared with the rough upper bound $(1+\delta)(\beta_\delta + \theta)^2 k \leqslant 11k$ corresponding to the RIC $\delta = 0.7$, and the more realistic bound $4k$ corresponding to $\delta = 0.4$. Observe (lower figures) that for exceedingly small $\lambda \ll \epsilon$, there is an additional growth of order $\dfrac{\epsilon}{\lambda}$.

Observe that according to (3.4), the ℓ_1-size of the LASSO minimizer $\boldsymbol{x}_\lambda$ remains smaller than the target $|\boldsymbol{x}_*(k)|_1$, Indeed, as long as the residual $|\boldsymbol{r}_\lambda|_2 > \mu$, then

$$|\boldsymbol{x}_*(k)|_1 - |\boldsymbol{x}_\lambda|_1 \geqslant (|\boldsymbol{r}_\lambda|_2 - \mu)\frac{|\boldsymbol{r}_\lambda|_2}{\lambda}. \tag{3.13}$$

This is depicted in figure 3.2: as λ decreases, the ratio $\dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda} \gtrsim \sqrt{s_\lambda}$ is increasing until $|\boldsymbol{x}_\lambda|_1$ reaches its upper bound of $|\boldsymbol{x}_*(k)|_1$.

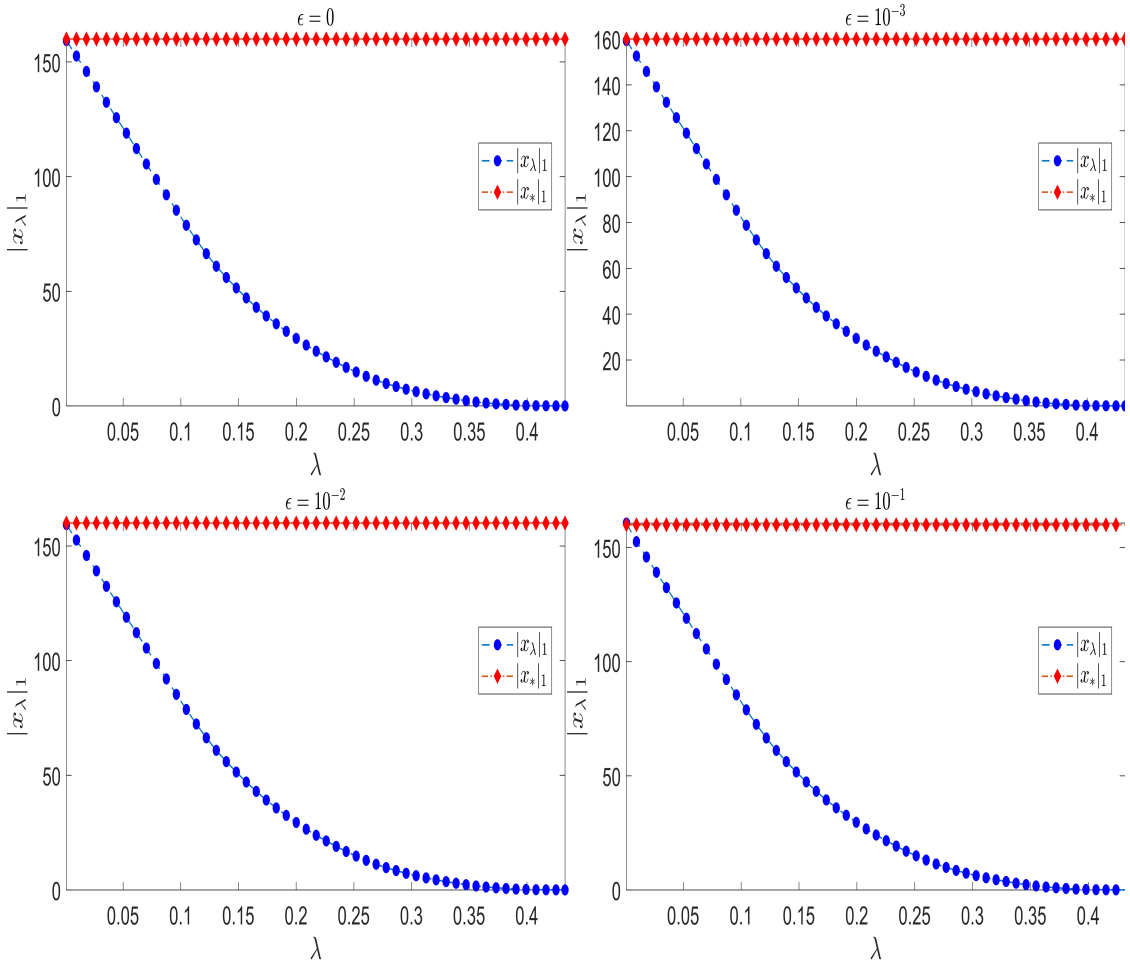

FIGURE 3.2. ℓ_1 norm of $\boldsymbol{x}_\lambda$ approaches its upper-bound $|\boldsymbol{x}_*(k)|_1$ as λ decreases.

Figure 3.3 shows the aptitude of the lower- and upper-bounds of the re-scaled residual (3.11), in capturing the re-scaled residual $\dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda}$. Again, the three quantities increase with deceasing λ, until λ reaches the threshold λ_c at which point the re-scaled residual, $\dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda}$, peaks at its maximal value ~ 27, in agreement with the upper-bound (3.2), $\dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda} < \beta_\delta\sqrt{k} + \dfrac{2\epsilon}{\lambda} < 30.61 + \dfrac{2\epsilon}{\lambda}$.

4. ERROR BOUNDS

4.1. **ℓ_2-error bounds.** The sparsity bound (3.11) was derived based on a two-sided ℓ_2-bound of the scaled residual. The latter can be converted into a two-sided ℓ_2 error bound of $|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2$. Note that since $\big||\boldsymbol{r}_\lambda|_2 - |A(\boldsymbol{x}_*(k) - \boldsymbol{x}_\lambda)|_2\big| \leqslant \mu$, then the upper-bound on $|\boldsymbol{r}_\lambda|_2$, see (3.5), also bounds the 'observed error' $A(\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k))$,

$$|A(\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k))|_2 \leqslant (\beta_\delta + \theta)\sqrt{k}\lambda + \mu \leqslant \beta_\delta\sqrt{k}\lambda + 3\mu. \tag{4.1}$$

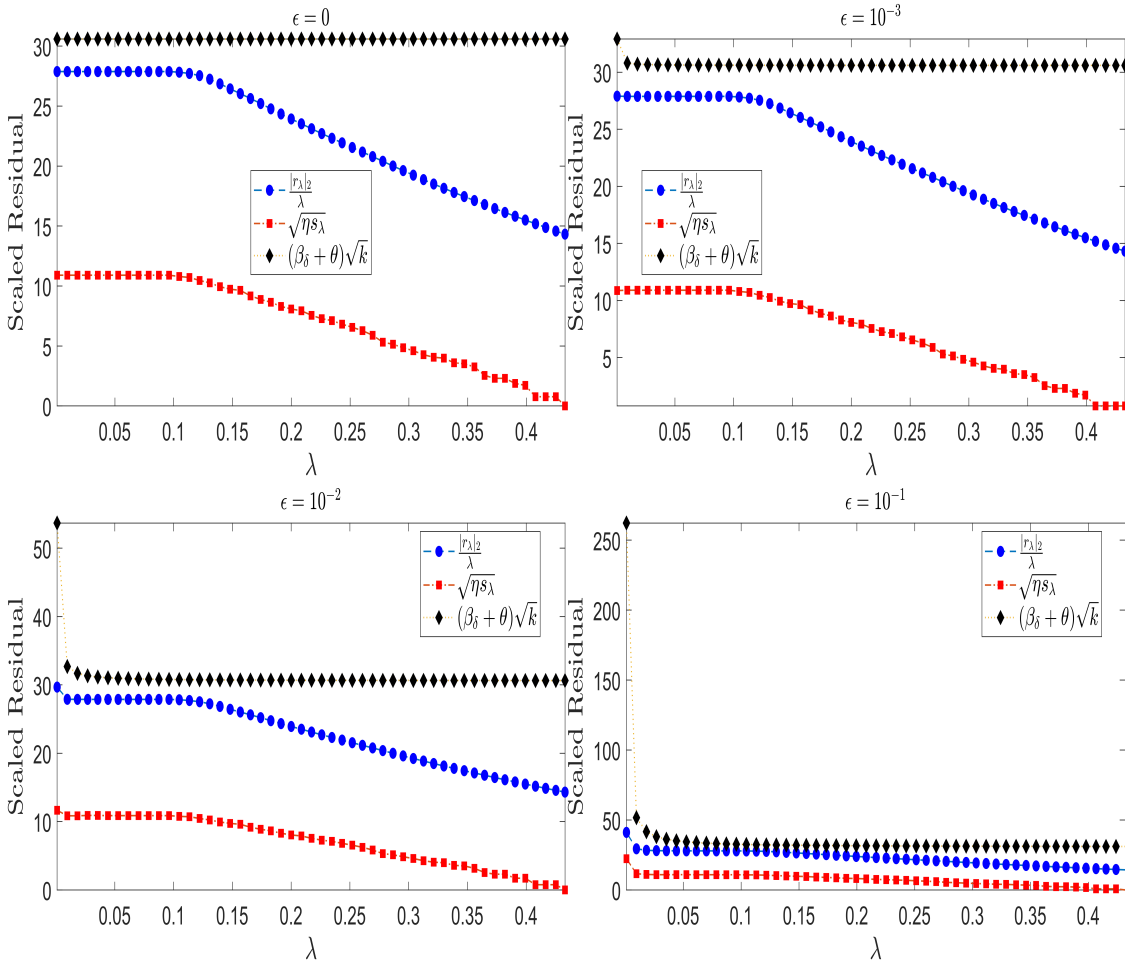

FIGURE 3.3. Re-scaled residual $\frac{|\boldsymbol{r}_\lambda|_2}{\lambda}$ captured between its lower- and upper-bounds (3.10) and (3.2), $\sqrt{\eta s_\lambda} \leqslant \frac{|\boldsymbol{r}_\lambda|_2}{\lambda} \leqslant \beta_\delta\sqrt{k} + \frac{2\epsilon}{\lambda} \approx 30.61$ with $(\eta, \beta, \theta) = (0.59, 2.42, 0.1)$ corresponding to $\delta = 0.7$. It peaks at a threshold value of 27, independent of the level of noise. When $\lambda \ll \epsilon$, there is an additional large term of order $\frac{2\epsilon}{\lambda}$.

The sparsity of $\boldsymbol{x}_\lambda$ does not exceed $s_\lambda \leqslant ([\chi^2] + 1)k$ hence $\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)$ has sparsity $([\chi^2] + 2)k$, and the RIP (2.2) implies the ℓ_2-error upper-bound

$$|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2 \leqslant \frac{1}{\sqrt{1-\delta}}(\beta_\delta\sqrt{k}\lambda + 3\mu), \qquad \delta = \delta_{([\chi^2]+2)k}. \tag{4.2}$$

In particular, (4.2) with $\frac{1}{\sqrt{1-\delta}} \leqslant 1.83$ and $\beta_\delta \leqslant 2.42$ corresponding to $\delta = 0.7$ yields

$$|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2 \lesssim 4.43\sqrt{k}\lambda + 5.48\mu. \tag{4.3}$$

This recovers a quantitative version of the ℓ_2 upper bound proved under additional condition of an incoherence design assumption in [30, Theorem 1], an ℓ_1-CMSV assumption[5] [38], or restricted eigenvalue bound in [28, Theorem 11.1].

[5]In fact, we slightly improve the quadratic dependence of the bound in [38, (23)] on the ℓ_1-CMSV constant $\sim \rho_{4k}^{-2}$, mentioned in (4.6) below.

The upper-bound (4.2) is sharp in the sense of having a tight ℓ_2-*lower bound*: since the error $\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)$ is at most $([\chi^2]+2)k$-sparse, we can use the RIP to translate the lower bound (3.10) into an ℓ_2 lower-bound,

$$|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2 \geqslant \frac{1}{\sqrt{1+\delta}}|A(\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k))|_2 \geqslant \frac{1}{\sqrt{1+\delta}}\,(|\boldsymbol{r}_\lambda|_2 - \mu) \geqslant \frac{\sqrt{s_\lambda}\lambda}{1+\delta} - \frac{\mu}{\sqrt{1+\delta}}.$$

We summarize these bounds in the following form.

Theorem 4.1 (ℓ_2**-bound).** *Fix $\lambda < \lambda_\infty := |A^\top \boldsymbol{y}_*^\epsilon|_\infty$ and let $\boldsymbol{x}_\lambda$ be the s_λ-sparse minimizer of the corresponding LASSO (1.2), observed with RIP matrix A. Then*

$$\frac{1}{\sqrt{1+\delta}}\Big(\frac{\sqrt{s_\lambda}\lambda}{\sqrt{1+\delta}} - \mu\Big) \leqslant |\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2 \leqslant \frac{1}{\sqrt{1-\delta}}(\beta_\delta\sqrt{k}\lambda + 3\mu), \qquad \delta = \delta_{([\chi^2]+2)k}. \tag{4.4}$$

Remark 4.2 (Compared with the ℓ_1-entropy bound). *The extremal relation $\langle A_{\mathcal{S}}\boldsymbol{x}_{\lambda,\mathcal{S}}, \boldsymbol{r}_\lambda\rangle = \lambda|\boldsymbol{x}_{\lambda,\mathcal{S}}|_1$ and the RIP (2.2) yield $\lambda|\boldsymbol{x}_{\lambda,S}|_1 \leqslant \sqrt{1+\delta}|\boldsymbol{x}_{\lambda,S}|_2|\boldsymbol{r}_\lambda|_2$, and hence we end up with a lower-bound involving the ℓ_1-entropy of $\{\boldsymbol{x}_{\lambda,S}\}$,*

$$\frac{|\boldsymbol{r}_\lambda|_2^2}{\lambda^2} \geqslant \frac{1}{1+\delta}\mathit{Ent}(\boldsymbol{x}_{\lambda,S}) \qquad \mathit{Ent}(\boldsymbol{x}) := \frac{|\boldsymbol{x}|_1^2}{|\boldsymbol{x}|_2^2}. \tag{4.5}$$

This bound is tied to a Null Entropy Property of A [1, §3.2] *or the ℓ_1-CMSV constant $\rho_s(A)$ introduced in* [38][6]

$$\frac{|\boldsymbol{r}_\lambda|_2^2}{\lambda^2} \geqslant \frac{\mathit{Ent}(\boldsymbol{x}_{\lambda,S})}{\rho_s(A)}, \qquad \rho_s(A) := \min_{|\boldsymbol{x}|_2=1}\Big\{|A\boldsymbol{x}|_2 \,:\, \mathit{Ent}(\boldsymbol{x}) \leqslant s\Big\}. \tag{4.6}$$

[6]Which is not to be confused with the RNSP parameter in (2.6)

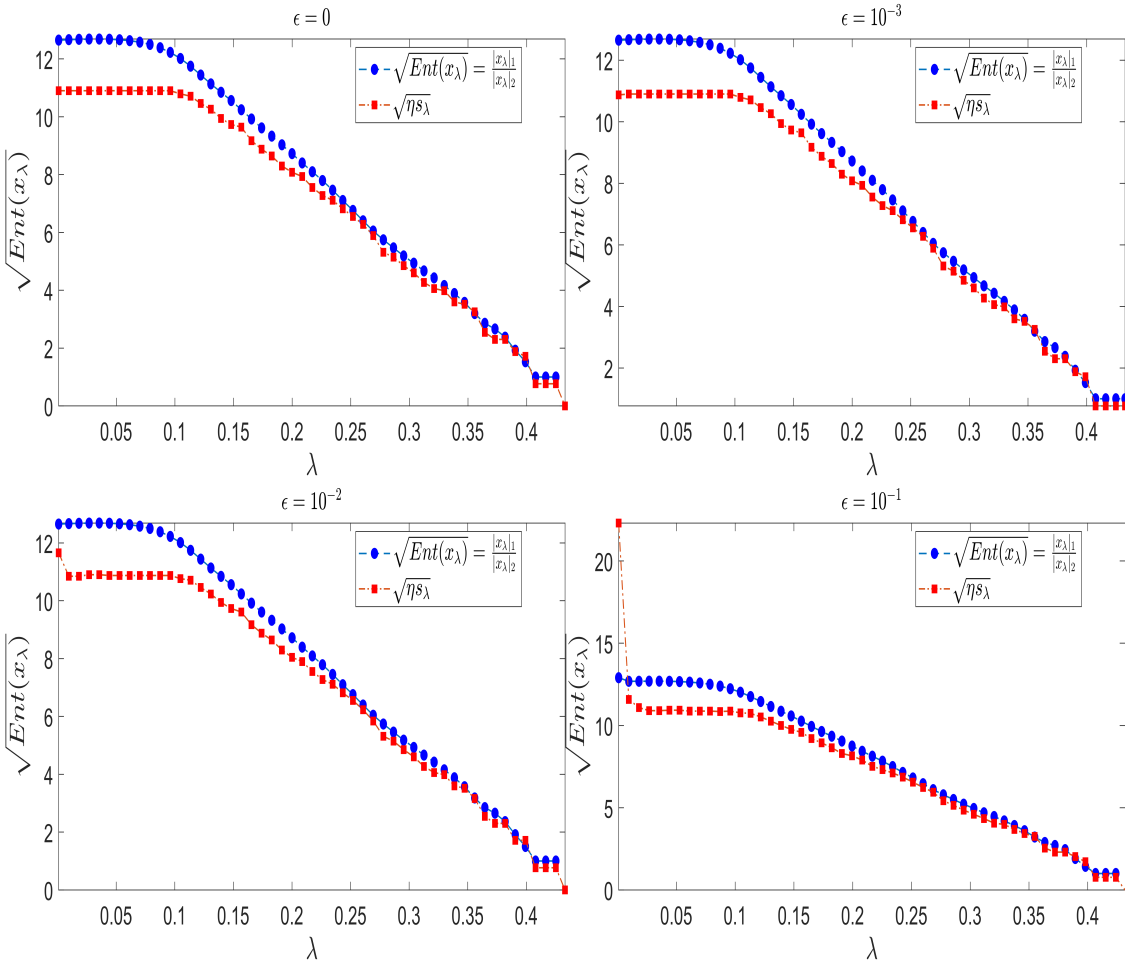

FIGURE 4.1. Lower bounds of the re-scaled residual: (3.10) with $\eta := \frac{1}{1+\delta} = 0.592$ vs. the ℓ_1-entropy based (4.5).

Clearly, if $\boldsymbol{x}_\lambda$ has the sparsity s_λ then $Ent(\boldsymbol{x}_{\lambda,S}) \leqslant s_\lambda$. Here we note about the reverse implication, namely — if the reverse inequality holds, $Ent(\boldsymbol{x}_{\lambda,S}) \gtrsim s_\lambda$, then it would yield our sparsity result of lemma 3.2, based on the lower bound $\frac{|\boldsymbol{r}_\lambda|_2}{\lambda} \gtrsim \frac{\sqrt{s_\lambda}}{\rho_{s_\lambda}}$. Theorem 3.3 suggests the lower-entropy bound for the minimizers $\boldsymbol{x}_\lambda$. Indeed, figure 4.1 shows a remarkable agreement between the lower bound (3.10) with $\delta = 0.7$ and the ℓ_1-entropy bound (4.5), $Ent(\boldsymbol{x}_\lambda)$, at least before the support of $\boldsymbol{x}_\lambda$ reaches its peak at k_{max}.

4.2. **ℓ_1-error bound.** We recall the ℓ_2-bound (4.2) which we express in the form $|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2 \leqslant \frac{1}{\sqrt{1-\delta}}(\beta_\delta + 3/2\theta)\sqrt{k}\lambda$. Since $\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)$ has sparsity of order $\leqslant k + \chi^2 k$, we derive the following ℓ_1-bound.

Theorem 4.3 (**ℓ_1-error bound**). *Fix $\lambda < \lambda_\infty := |A^\top \boldsymbol{y}^\epsilon_*|_\infty$ and let $\boldsymbol{x}_\lambda$ be the LASSO minimizer of (1.2), observed with RIP matrix A with RIC δ such that (3.7) holds. Then the following ℓ_1-error*

bound holds,

$$
\begin{aligned}
|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 &\leqslant \sqrt{\big(1+(1+\delta)(\beta_\delta+\theta)^2\big)k}\,|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_2 \\
&< \sqrt{1+\delta}(\beta_\delta + 1/2 + \theta)\sqrt{k}\frac{1}{\sqrt{1-\delta}}\big(\beta_\delta + 3/2\theta\big)\sqrt{k}\lambda \\
&< \frac{\sqrt{1+\delta}}{\sqrt{1-\delta}}\frac{1}{\lambda}\Big(\big(\beta_\delta+1/2\big)\sqrt{k}\lambda + 2\mu\Big)^2.
\end{aligned} \tag{4.7}
$$

The amplitude of $k\lambda$ in the ℓ_1-error bound (4.7) is not sharp. For example, with RIC $\delta = \delta_{11k} < 0.7$ we have $\beta_\delta > 2$ in which case, omitting the negligibly small μ^2/λ terms, one ends up with the improved bound

$$
|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 < \frac{\sqrt{1+\delta}}{\sqrt{1-\delta}}\Big(\big(\beta_\delta + 1/4\big)^2 k\lambda + (4\beta_\delta+1)\sqrt{k}\mu\Big) < 16.97k\lambda + 24.04\sqrt{k}\mu. \tag{4.8}
$$

We conclude with an alternative derivation of an ℓ_1-error bound. To this end, we recall the RNSP bound [26, Theorem 4.20], which states that for all $\mathcal{K} \subset \{1, 2 \ldots, N\}$ of size$\leqslant k$ and for any $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^N$, the following holds,

$$
|\boldsymbol{u} - \boldsymbol{v}|_1 \leqslant \frac{1+\rho}{1-\rho}\big(|\boldsymbol{u}|_1 - |\boldsymbol{v}|_1 + 2|\boldsymbol{v}_{\mathcal{K}^c}|_1\big) + \frac{2\tau}{1-\rho}|A(\boldsymbol{u}-\boldsymbol{v})|_2, \qquad |\mathcal{K}| \leqslant k.
$$

Using it with $(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{x}_\lambda, \boldsymbol{x}_*(k))$ and $\mathcal{K} = \operatorname{supp}(\boldsymbol{x}_*(k))$ yields

$$
|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 \leqslant \frac{1+\rho}{1-\rho}\Big(|\boldsymbol{x}_\lambda|_1 - |\boldsymbol{x}_*(k)|_1\Big) + \frac{2\tau}{1-\rho}|A(\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k))|_2. \tag{4.9}
$$

Now, using (3.4) to bound the term inside the first parenthesis on the right, and as before, noting that the second term does not exceed $|A(\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k))|_2 \leqslant |\boldsymbol{r}_\lambda|_2 + \mu$, we find

$$
|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 \leqslant \frac{1+\rho}{1-\rho}\Big\{|\boldsymbol{r}_\lambda|_2\Big(\frac{2\tau}{1+\rho} + \frac{\mu}{\lambda} - \frac{|\boldsymbol{r}_\lambda|_2}{\lambda}\Big) + \frac{2\tau}{1+\rho}\mu\Big\}.
$$

Given the RNSP parameters (2.6), $2\tau = 2\beta\sqrt{k}$ and $\dfrac{\mu}{\lambda} < \dfrac{\theta}{1+\rho}\sqrt{k}$, the last bound yields

$$
|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 \leqslant \frac{1}{1-\rho}\left\{|\boldsymbol{r}_\lambda|_2\left(2(\beta_\delta+\theta)\sqrt{k} - (1+\rho)\frac{|\boldsymbol{r}_\lambda|_2}{\lambda}\right) + \beta_\delta\theta k\lambda\right\}. \tag{4.10}
$$

Viewed as quadratic in $\dfrac{|\boldsymbol{r}_\lambda|_2}{\lambda}$, the first expression on the right admits a maximal value $\dfrac{(\beta_\delta+\theta)^2}{1+\rho}k\lambda$, and we finally end up with

$$
|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 \leqslant \frac{1}{1-\rho^2}\left((\beta_\delta+\theta)^2 k\lambda + (1+\rho)\beta_\delta\theta k\lambda\right) \leqslant \frac{(\beta_\delta+2\theta)^2}{1-\rho^2}k\lambda. \tag{4.11}
$$

This recovers the ℓ_1-bound of order $\mathcal{O}(k\lambda)$ as in (4.7). However, since the ℓ_1 bound (4.11) involves the value of $1/(1-\rho)^2$, it is therefore limited to the RIC $\delta < 4/\sqrt{41}$ where ρ approaches 1.

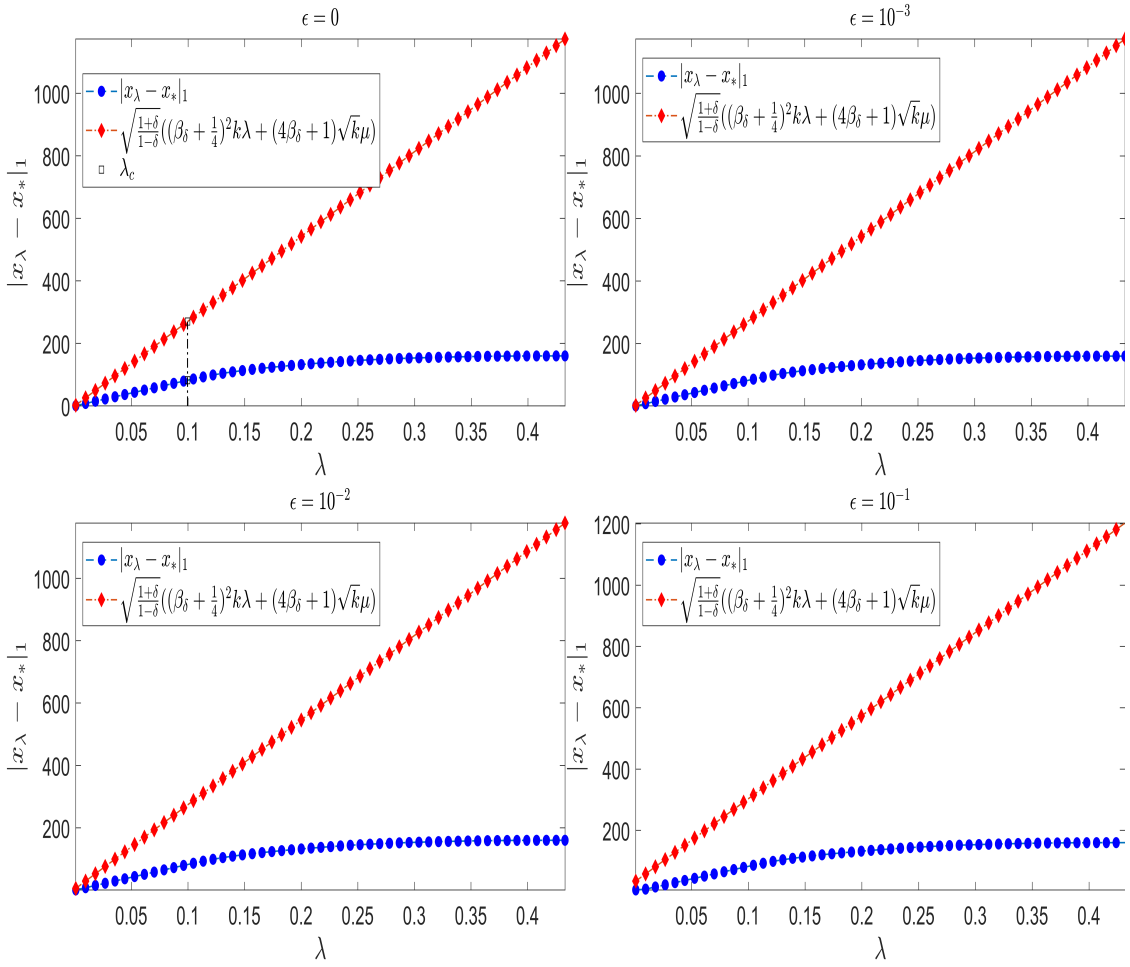

FIGURE 4.2. ℓ_1-error for recovery of sparse data compared with the upper-bound (4.12).

4.3. **Numerical simulations.** We report on the error behavior in our simulations of the unconstrained LASSO (1.2), applied to the recovery of k-sparse data, $\sigma_k = 0$, that is $\mu = \epsilon$, with $(k, m, N) = (160, 1024, 4096)$. The results are obtained by averaging 100 observations using randomly generated RNSP$_{\rho,\tau}$ matrices based on Gaussian distributions. We compare the ℓ_1-error with the error bound (4.8)

$$|\boldsymbol{x}_\lambda - \boldsymbol{x}_*(k)|_1 \leqslant 16.97 * 160\lambda + 24.04\sqrt{160}\,\epsilon. \tag{4.12}$$

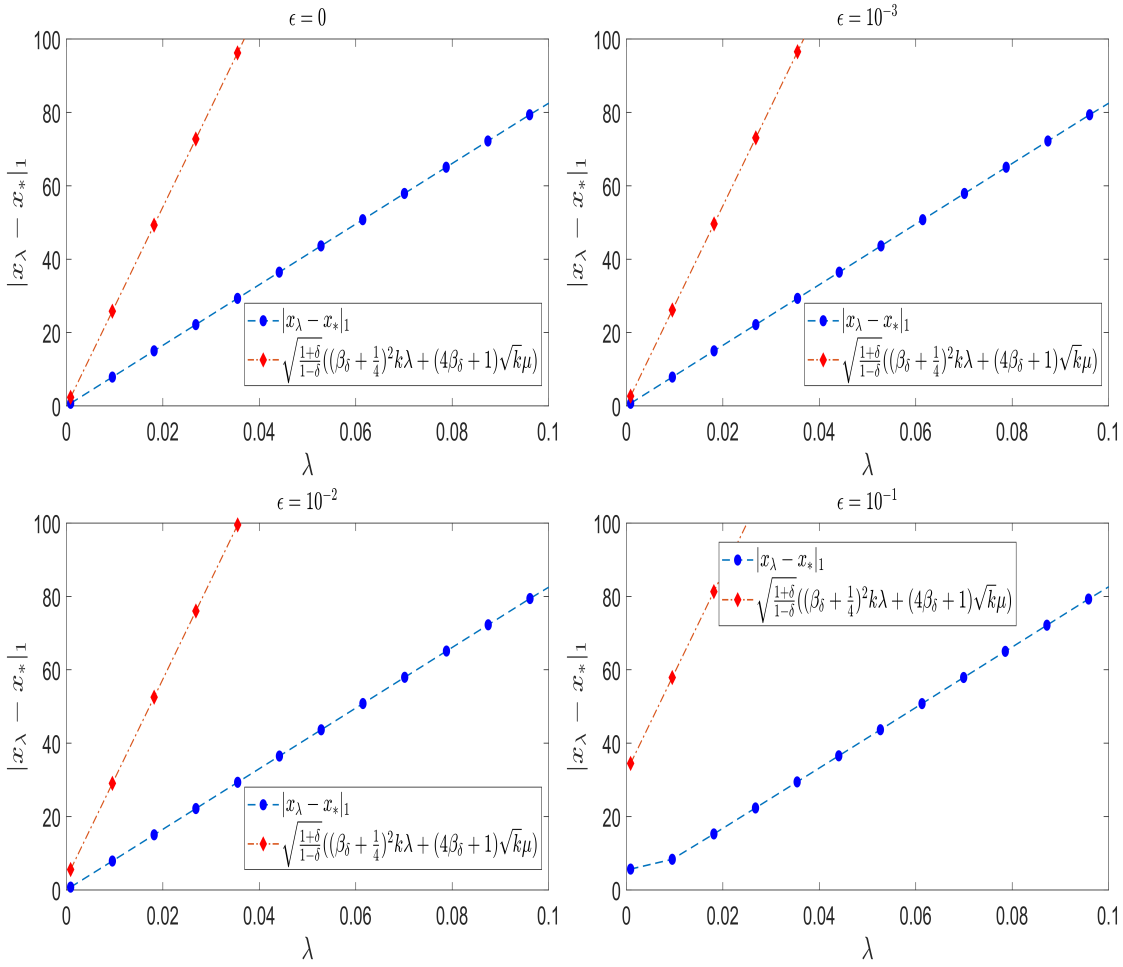

FIGURE 4.3. The ℓ_1-error compared with the upper-bound (4.12) zoomed near $\lambda = 0$.

Acknowledgment. We are indebted to Wolfgang Dahmen for comments which greatly improved the presentation of material of this paper.

REFERENCES

[1] J. ANDERSSON AND J.-O. STRÖMBERG, *On the theorem of uniform recovery of random sampling matrices*, IEEE Transactions on Information Theory, 60 (2014), pp. 1700–1710.

[2] B. BAH AND J. TANNER, *Improved bounds on restricted isometry constants for Gaussian matrices*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2882–2898.

[3] ———, *Bounds of restricted isometry constants in extreme asymptotics: formulae for Gaussian matrices*, Linear Algebra and its Applications, 441 (2014), pp. 88–109.

[4] R. BARANIUK, M. DAVENPORT, R. DEVORE, AND M. WAKIN, *A simple proof of the restricted isometry property for random matrices*, Constructive Approximation, 28 (2008), pp. 253–263.

[5] A. BELLONI AND V. CHERNOZHUKOV, *Least squares after model selection in high-dimensional sparse models*, Bernoulli, 19 (2013), pp. 521–547.

[6] P. J. BICKEL, Y. RITOV, AND A. B. TSYBAKOV, *Simultaneous analysis of Lasso and Dantzig selector*, The Annals of Statistics, 37 (2009), pp. 1705–1732.

[7] J. D. BLANCHARD, C. CARTIS, AND J. TANNER, *Compressed sensing: How sharp is the restricted isometry property?*, SIAM Review, 53 (2011), pp. 105–125.

[8] J. BOURGAIN, S. DILWORTH, K. FORD, S. KONYAGIN, AND D. KUTZAROVA, *Explicit constructions of RIP matrices and related problems*, Duke Mathematical Journal, 159 (2011), pp. 145–185.
[9] P. BÜHLMANN AND S. VAN DE GEER, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media, 2011.
[10] T. T. CAI AND A. ZHANG, *Sparse representation of a polytope and recovery of sparse signals and low-rank matrices*, IEEE transactions on information theory, 60 (2013), pp. 122–132.
[11] E. CANDÈS AND T. TAO, *The Dantzig selector: Statistical estimation when p is much larger than n*, The Annals of Statistics, 35 (2007), pp. 2313–2351.
[12] E. J. CANDÈS, *The restricted isometry property and its implications for compressed sensing*, Comptes Rendus Mathematique, 346 (2008), pp. 589–592.
[13] E. J. CANDÈS, J. K. ROMBERG, AND T. TAO, *Robust uncertainty principles: exact signal reconstruction from high incomplete frequency information*, IEEE Trans. Infom. Theor., 52 (2006), pp. 489 – 509.
[14] ——, *Stable signal recovery from incomplete and inaccurate measurements*, Commun. Pure Appl. Math., 59 (2006), pp. 1207 – 1233.
[15] E. J. CANDÈS AND T. TAO, *The Dantzig selector: statistical estimation when p is much larger than n*, Ann. Stat., 35 (2007), pp. 2313 – 2351.
[16] S. S. CHEN, *Basis pursuit*, PhD thesis, Stanford University, 1996.
[17] S. S. CHEN, D. L. DOHONO, AND M. A. SAUNDERS, *Atomic decomposition by Basis Pursuit*, SIAM J. Sci. Comput., 20 (1999), pp. 33 – 61.
[18] S. S. CHEN AND D. L. DONOHO, *Basis pursuit*, in Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers, vol. 1, IEEE, 1994, pp. 41–44.
[19] A. COHEN, W. DAHMEN, AND R. A. DEVORE, *Compressed sensing and the best k-term approximation*, J. Amer. Math. Soc., 22 (2009), pp. 211 – 231.
[20] R. DEVORE, G. PETROVA, AND P. WOJTASZCZYK, *Instance-optimality in probability with an ℓ_1-minimization decoder*, Appl. Comput. Harmon. Anal., 27 (2009), pp. 275–288.
[21] D. L. DONOHO, *Compressed sensing*, IEEE Trans. Infom. Theor., 52 (2006), pp. 1289 – 1306.
[22] D. L. DONOHO AND M. ELAD, *Optimally sparse representations in general (non-orthogonal) dictionaries via ℓ^1 minimization*, Proc. Nat. Acad. Sci., 100 (2003), pp. 2197 – 2202.
[23] D. L. DONOHO AND X. HUO, *Uncertainty principles and ideal atomic decompositions*, IEEE Trans. Inform. Theor., 47 (2001), pp. 2845 – 2862.
[24] D. L. DONOHO AND I. M. JOHNSTONE, *Empirical atomic decomposition*, Unpublished manuscript, (1995).
[25] S. FOUCART, *Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants*, in Approximation Theory XIII: San Antonio 2010, Springer, 2012, pp. 65–77.
[26] S. FOUCART AND H. RAUHUT, *A Mathematical Introduction to Compressive Sensing*, Applied and Numerical Harmonic Analysis, Birkhäuser, 2013.
[27] R. GRIBONVAL AND M. NIELSEN, *Sparse representations in unions of bases*, IEEE Trans. Inform. Theor., 49 (2003), pp. 3320 – 3325.
[28] T. HASTIE, R. TIBSHIRANI, AND M. WAINWRIGHT, *Statistical Learning with Sparsity: the LASSO and generalizations*, CRC Press, 2015.
[29] S. MALLAT AND W. L. HWANG, *Singularity detection and processing with wavelets*, IEEE Transactions on Information Theory, 38 (1992), pp. 617–643.
[30] N. MEINSHAUSEN, B. YU, ET AL., *Lasso-type recovery of sparse representations for high-dimensional data*, The Annals of Statistics, 37 (2009), pp. 246–270.
[31] L. I. RUDIN, S. OSHER, AND E. FATEMI, *Nonlinear total variation based noise removal algorithms*, Physica D: nonlinear phenomena, 60 (1992), pp. 259–268.
[32] W. SU, M. BOGDAN, AND E. CANDÈS, *False discoveries occur early on the Lasso path*, The Annals of Statistics, (2017), pp. 2133–2150.
[33] E. TADMOR, *Hierarchical construction of boudned solutions in regularity spaces*, Communications in Pure & Applied Mathematics, 69(6) (2015), pp. 1087–1109.
[34] E. TADMOR, S. NEZZAR, AND L. VESSE, *A multiscale image representation using hierarchical (BV, L^2) decompositions*, Multiscale Model. Simul., 2 (2004), pp. 554 – 579.
[35] ——, *Multiscale hierarchical decomposition of images with applications to deblurring, denoising and segmentation*, Commun. Math. Sci., 6 (2008), pp. 281 – 307.

[36] E. TADMOR AND C. TAN, *Hierarchical construction of bounded solutions of* $\operatorname{div} U = F$ *in critical regularity spaces*, in Nonlinear Partial Differential Equations, H. Holden and K. Karlsen, eds., Abel Symposia 7, Oslo, Sep. 2010, pp. 255 – 269. 2010 Abel Symposium.

[37] E. TADMOR AND M. ZHONG, *The multi-scale hierarchical reconstruction of ill-posed inverse problems*, In preparation, (2021).

[38] G. TANG AND A. NEHORAI, *Performance analysis of sparse recovery based on constrained minimal singular values*, IEEE Transactions on Signal Processing, 59 (2011), pp. 5734–5745.

[39] R. TIBSHIRANI, *Regression shrinkage and selection via the lasso*, J. Roy. Stat. Soc. B, 58 (1996), pp. 267 – 288.

[40] J. A. TROPP, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inform. Theor., 50 (2004), pp. 2231 – 2242.

[41] H. ZHANG, W. YIN, AND L. CHENG, *Necessary and sufficient conditions on solution uniqueness in* ℓ_1 *minimization*, Journal of Optimization Theory and Applications, 164 (2015), pp. 109 – 122.

DEPARTMENT OF MATHEMATICS AND TEXAS A&M INSTITUTE OF DATA SCIENCE
TEXAS A&M UNIVERSITY, COLLEGE STATION, TX 77843
Email address: foucart@tamu.edu

DEPARTMENT OF MATHEMATICS AND INSTITUTE FOR PHYSICAL SCIENCES & TECHNOLOGY (IPST)
UNIVERSITY OF MARYLAND, COLLEGE PARK, MD 20742
Email address: tadmor@umd.edu

TEXAS A&M INSTITUTE OF DATA SCIENCE
TEXAS A&M UNIVERSITY, COLLEGE STATION, TX 77843
Email address: mingzhong@tamu.edu