The Sparsity of LASSO-type Minimizers

Simon Foucart*— Texas A&M University

Abstract

This note extends an attribute of the LASSO procedure to a whole class of related procedures, including square-root LASSO, square LASSO, LAD-LASSO, and an instance of generalized LASSO. Namely, under the assumption that the input matrix satisfies an ℓ_p-restricted isometry property (which in some sense is weaker than the standard ℓ_2-restricted isometry property assumption), it is shown that if the input vector comes from the exact measurement of a sparse vector, then the minimizer of any such LASSO-type procedure has sparsity comparable to the sparsity of the measured vector. The result remains valid in the presence of moderate measurement error when the regularization parameter is not too small.

Key words and phrases: Regularization, sparsity-promoting optimization, LASSO, compressive sensing, restricted isometry property.

AMS classification: 62J07, 65F22, 90C25, 94A12.

1 Introduction

The purpose of this note is to extend a result established in [8], namely that the LASSO output has sparsity at most proportional to the sparsity of the vector giving rise to the LASSO input. More precisely, consider a vector $x \in \mathbb{R}^N$ measured via $y = Ax \in \mathbb{R}^m$ for some matrix $A \in \mathbb{R}^{m \times N}$, $m \ll N$, and suppose that x is s-sparse, meaning that

$$\|x\|_0 := |\mathrm{supp}(x)| \le s, \qquad \text{where} \quad \mathrm{supp}(x) = \{j \in [1 : N] : x_j \ne 0\}.$$

Let $x^\lambda \in \mathbb{R}^N$ be the output of the LASSO procedure with regularization parameter $\lambda > 0$, i.e., let x^λ be a solution of

$$\underset{z \in \mathbb{R}^N}{\text{minimize}} \; \frac{1}{2}\|y - Az\|_2^2 + \lambda\|z\|_1.$$

Assume that the measurement matrix $A \in \mathbb{R}^{m \times N}$ satisfies the standard restricted isometry property of order t (proportional to s) and constant $\delta \in (0, 1)$, meaning that

$$(1) \qquad \sqrt{1 - \delta}\,\|z\|_2 \le \|Az\|_2 \le \sqrt{1 + \delta}\,\|z\|_2 \quad \text{whenever} \; \|z\|_0 \le t.$$

Then, as shown in [8], there is a constant C_δ depending on δ such that $\|x^\lambda\|_0 \le C_\delta \, s$. For instance, the choice $\delta = 0.4$ guarantees that $\|x^\lambda\|_0 \le 4s$.

In this note, a similar result is established for a generalization of the LASSO procedure branching out in four directions:

- the power of the norm on $y - Az$ can be any $q \ge 1$—in particular, the case $q = 1$ corresponds to the square-root LASSO (see [2]);

- in fact, the norm on $y - Az$ is not restricted to be the ℓ_2-norm, as any ℓ_p-norm with $p \in [1, 2]$ is possible—in particular, the case $p = 1$ and $q = 1$ corresponds to the least absolute deviation (LAD) LASSO (see [12]);

- the ℓ_1-norm on z can be raised to any power $r \ge 1$—in particular, the case $r = 2$ corresponds to the square LASSO, which can be recast as a nonnegative least squares problem (see [6]);

- the ℓ_1-norm can act not only on z but also on $B^{-1}z$ for some invertible matrix $B \in \mathbb{R}^{N \times N}$— this corresponds to the generalized LASSO (see [11], where B^{-1} can instead be a rectangular matrix).

As in [8], the full result incorporates measurement errors. However, for simplicity, a particularized version is stated here first.

Theorem 1. Let $p \in [1, 2]$, $q \ge 1$, and $r \ge 1$. Consider a vector $x \in \mathbb{R}^N$ such that $B^{-1}x$ is s-sparse for some invertible matrix $B \in \mathbb{R}^{N \times N}$ with condition number $\kappa_B := \|B\|_{2\to2}\|B^{-1}\|_{2\to2}$. Suppose that the vector x is measured via $y = Ax \in \mathbb{R}^m$ for some matrix $A \in \mathbb{R}^{m \times N}$ satisfying the nonstantard restricted isometry property of order $t = \lfloor (2\gamma\kappa_B)^2 s \rfloor + 1$ with ratio $\gamma = \beta/\alpha$, i.e.,

$$\alpha\|z\|_2 \le \|Az\|_p \le \beta\|z\|_2 \qquad \text{whenever } \|B^{-1}z\|_0 \le t.$$

Then, for any $\lambda > 0$, the solution x^λ of the LASSO-type procedure

$$\underset{z \in \mathbb{R}^N}{\text{minimize}} \; \frac{1}{q}\|y - Az\|_p^q + \lambda\frac{1}{r}\|B^{-1}z\|_1^r$$

has sparsity at most proportional to s, namely

$$\|x^\lambda\|_0 \le \lfloor \chi^2 s \rfloor, \qquad \chi := 2\gamma\kappa_B.$$

The restrictions $p, q, r \ge 1$ are here to make the optimization program convex. The restriction $p \le 2$ is here to make the nonstandard restricted isometry property assumption realizable in the regime $m \asymp s \ln(eN/s)$. This assumption is in fact weaker than the standard restricted isometry property assumption, as discussed in Subsection 2.2.

The remainder of this note is dedicated to the proof of Theorem 1, or rather of the more general version given in Theorem 5. The latter shows that the conclusion remains unchanged when the measurement vector $y = Ax + e$ features a nonzero error vector $e \in \mathbb{R}^m$, provided its magnitude $\|e\|_p$ is small enough compared to λ. From a reversed point of view, favored here, it can also be claimed that the LASSO-type minimizer has sparsity at most proportional to s, provided the regularization parameter λ is large enough compared to $\|e\|_p$.

The organization is now straightforward: Section 2 establishes side results needed in the said proof and Section 3 completes the arguments.

2 Preliminary results

This section collects some auxiliary results that will be invoked later. The author is not aware of earlier appearances of these results, but the vastness of the literature around compressive sensing makes such appearances not unlikely.

2.1 Characterization of LASSO-type minimizers

This subsection establishes a necessary and sufficient condition for a vector to be a solution of the general LASSO-type procedure. This condition is essential in the subsequent considerations. In the case of standard LASSO, i.e., $p = 2$, $q = 2$, and $r = 1$, it seems to be folklore—it can be found e.g. in [9, Section 3.4], [10, Section 2.1], and [7, Section 15.1]. The arguments below could be shortcut by exploiting the notion of subdifferential, but a more elementary proof has been preferred, at the expense of brevity.

Proposition 2. Let $p > 1$, $q > 1$, and $r \geq 1$. Given $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times N}$, and $B \in \mathbb{R}^{N \times N}$, the matrix B being invertible, a LASSO-type minimizer

$$x^\lambda = \operatorname*{argmin}_{z \in \mathbb{R}^N} \frac{1}{q}\|y - Az\|_p^q + \lambda \frac{1}{r}\|B^{-1}z\|_1^r$$

is characterized, with $\nu^\lambda := \lambda\|y - Ax^\lambda\|_p^{p-q}\|B^{-1}x^\lambda\|_1^{r-1}$ and $S^\lambda := \operatorname{supp}(B^{-1}x^\lambda)$, by

(2) $\qquad \left(B^\top A^\top \left(\operatorname{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}\right)\right)_j = \nu^\lambda \operatorname{sgn}((B^{-1}x^\lambda)_j), \qquad j \in S^\lambda,$

(3) $\qquad \left|\left(B^\top A^\top \left(\operatorname{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}\right)\right)_\ell\right| \leq \nu^\lambda, \qquad \ell \notin S^\lambda.$

Proof. Notice first that it is enough to consider the case $B = \mathrm{I}_N$ by setting $z' = B^{-1}z$ and $A' = AB$. Second, thanks to the convexity of the objective function, temporarily denoted by F, notice that a

global minimizer x^λ is equivalently a local minimizer, so it is characterized by $F(x^\lambda) \leq F(x^\lambda + tu)$ for all $u \in \mathbb{R}^N$ and all $t \in \mathbb{R}$ small enough in absolute value. Fixing $u \in \mathbb{R}^N$, one shall expand $F(x^\lambda + tu)$ around $t = 0$. To start, setting aside the eventuality that $y - Ax^\lambda = 0$ for now and considering $I := \{i \in [1 : m] : (y - Ax^\lambda)_i \neq 0\}$, one writes

$$\|y - A(x^\lambda + tu)\|_p^p = \sum_{i=1}^m |(y - Ax^\lambda)_i - t(Au)_i|^p$$

$$= \sum_{i \in I} |(y - Ax^\lambda)_i|^p \left(1 - t\frac{(Au)_i}{(y - Ax^\lambda)_i}\right)^p + |t|^p \sum_{i \notin I} |(Au)_i|^p.$$

In view of $|t|^p = o(t)$ (since $p > 1$) and of $(1 + t\xi)^\eta = 1 + t\eta\xi + o(t)$, used repeatedly below for several choices of η (here for $\eta = p$), one obtains

$$\|y - A(x^\lambda + tu)\|_p^p = \sum_{i \in I} |(y - Ax^\lambda)_i|^p \left(1 - tp\frac{(Au)_i}{(y - Ax^\lambda)_i}\right) + o(t)$$

$$= \|y - Ax^\lambda\|_p^p - tp \sum_{i \in I} \operatorname{sgn}((y - Ax^\lambda)_i)|(y - Ax^\lambda)_i|^{p-1}(Au)_i + o(t)$$

$$= \|y - Ax^\lambda\|_p^p - tp \sum_{i=1}^m \operatorname{sgn}((y - Ax^\lambda)_i)|(y - Ax^\lambda)_i|^{p-1}(Au)_i + o(t)$$

$$= \|y - Ax^\lambda\|_p^p - tp\langle A^\top\left(\operatorname{sgn}(y - Ax^\lambda)|y - Ax^\lambda|^{p-1}\right), u\rangle + o(t),$$

where the last-but-one step relied on $p > 1$ through $|(y - Ax^\lambda)_i|^{p-1} = 0$ for $i \notin I$. It follows that

$$\|y - A(x^\lambda + tu)\|_p^q = \left[\|y - Ax^\lambda\|_p^p\left(1 - tp\frac{\langle A^\top\left(\operatorname{sgn}(y - Ax^\lambda)|y - Ax^\lambda|^{p-1}\right), u\rangle}{\|y - Ax^\lambda\|_p^p} + o(t)\right)\right]^{q/p}$$

$$= \|y - Ax^\lambda\|_p^q\left(1 - tq\frac{\langle A^\top\left(\operatorname{sgn}(y - Ax^\lambda)|y - Ax^\lambda|^{p-1}\right), u\rangle}{\|y - Ax^\lambda\|_p^p} + o(t)\right)$$

$$(4) \qquad = \|y - Ax^\lambda\|_p^q - tq\|y - Ax^\lambda\|_p^{q-p}\langle A^\top\left(\operatorname{sgn}(y - Ax^\lambda)|y - Ax^\lambda|^{p-1}\right), u\rangle + o(t).$$

To continue, with $S^\lambda := \{j \in [1 : N] : x_j^\lambda \neq 0\}$ (recall that $B = \mathrm{I}_N$ is assumed here), one writes

$$\|x^\lambda + tu\|_1 = \sum_{j \in S^\lambda} \operatorname{sgn}((x^\lambda + tu)_j)(x^\lambda + tu)_j + |t| \sum_{\ell \notin S^\lambda} |u_\ell| = \sum_{j \in S^\lambda} \operatorname{sgn}(x_j^\lambda)(x^\lambda + tu)_j + |t| \sum_{\ell \notin S^\lambda} |u_\ell|$$

$$= \|x^\lambda\|_1 + t\sum_{j \in S^\lambda} \operatorname{sgn}(x_j^\lambda)u_j + |t| \sum_{\ell \notin S^\lambda} |u_\ell|,$$

where the last-but-one step used the fact that, for $j \in S^\lambda$, $\operatorname{sgn}((x^\lambda + tu)_j) = \operatorname{sgn}(x_j^\lambda)$ when $|t|$ is small enough. It follows that

$$\|x^\lambda + tu\|_1^r = \left[\|x^\lambda\|_1\left(1 + t\frac{\sum_{j \in S^\lambda} \operatorname{sgn}(x_j^\lambda)u_j}{\|x^\lambda\|_1} + |t|\frac{\sum_{\ell \notin S^\lambda} |u_\ell|}{\|x^\lambda\|_1}\right)\right]^r$$

$$= \|x^\lambda\|_1^r\left(1 + tr\frac{\sum_{j \in S^\lambda} \operatorname{sgn}(x_j^\lambda)u_j}{\|x^\lambda\|_1} + |t|r\frac{\sum_{\ell \notin S^\lambda} |u_\ell|}{\|x^\lambda\|_1} + o(t)\right)$$

$$(5) \qquad = \|x^\lambda\|_1^r + tr\|x^\lambda\|_1^{r-1} \sum_{j \in S^\lambda} \operatorname{sgn}(x_j^\lambda)u_j + |t|r\|x^\lambda\|_1^{r-1} \sum_{\ell \notin S^\lambda} |u_\ell| + o(t).$$

According to (4) and (5), the validity of $F(x^\lambda) \leq F(x^\lambda + tu)$ for all t small enough in absolute value is equivalent to

$$0 \leq t\Big(-\|y - Ax^\lambda\|_p^{q-p} \langle A^\top \big(\mathrm{sgn}(y - Ax^\lambda)\, |y - Ax^\lambda|^{p-1}\big), u \rangle + \lambda \|x^\lambda\|_1^{r-1} \sum_{j \in S^\lambda} \mathrm{sgn}(x_j^\lambda) u_j \Big)$$

$$+ |t|\Big(\lambda \|x^\lambda\|_1^{r-1} \sum_{\ell \notin S^\lambda} |u_\ell| \Big)$$

$$= t\Big\{ \sum_{j \in S^\lambda} \big(-\|y - Ax^\lambda\|_p^{q-p} \big(A^\top \big(\mathrm{sgn}(y - Ax^\lambda)\, |y - Ax^\lambda|^{p-1}\big)\big)_j + \lambda \|x^\lambda\|_1^{r-1} \mathrm{sgn}(x_j^\lambda) \big) u_j \Big\}$$

$$+ |t|\Big\{ \sum_{\ell \notin S^\lambda} \big(-\mathrm{sgn}(t)\mathrm{sgn}(u_\ell)\|y - Ax^\lambda\|_p^{q-p} \big(A^\top \big(\mathrm{sgn}(y - Ax^\lambda)\, |y - Ax^\lambda|^{p-1}\big)\big)_\ell + \lambda \|x^\lambda\|_1^{r-1} \big) |u_\ell| \Big\}$$

for all t small enough in absolute value. The latter inequality holds for all such t's and all $u \in \mathbb{R}^N$ if and only if the first term in curly brackets equals zero for all u supported on S^λ and the second term in curly brackets is nonnegative for all u supported on the complement of S^λ. This first condition holds if and only if (2) is satisfied and this second condition holds if and only if (3) is satisfied. The full characterization (2)–(3) of the LASSO-type minimizer has now been established in the sufficient case $B = \mathrm{I}_N$, save for the eventuality that $y - Ax^\lambda = 0$. But this cannot occur (justifying in retrospect that ν^λ is always well defined). Indeed, if $y - Ax^\lambda = 0$, then (4) would simply be replaced by $\|y - A(x^\lambda + tu)\|_p^q = |t|^q \|Au\|_p^q = o(t)$ (since $q > 1$) and, for all $u \in \mathbb{R}^N$ supported on S^λ and all t small enough in absolute value, $F(x^\lambda) \leq F(x^\lambda + tu)$ would read $0 \leq t\{\lambda \|x^\lambda\|_1^{r-1} \mathrm{sgn}(x_j^\lambda) u_j\}$, which is of course impossible. $\qquad\square$

From the above characterization, one easily sees that $x^\lambda = 0$ for $\lambda \geq \|B^\top A^\top \big(\mathrm{sgn}(y)\, |y|^{p-1}\big)\|_\infty \|y\|_p^{q-p}$ in the usual case $r = 1$. In other words, the output of the LASSO-type procedure is as sparse as can be for λ large enough. In the case $r > 1$, however, $x^\lambda = 0$ would force $\nu^\lambda = 0$, so (2)-(3) do not hold in general, and the output of the LASSO-type procedure is not the zero vector (although it converges to it) even for λ large enough. This remark reinforces the significance of the main result that the output of the LASSO-type procedure is sparse when λ is large enough.

2.2 Variation on the restricted isometry property

The standard restricted isometry property (1) is not appropriate to the present setting, where the sparsity does not concern the vector x itself, but $B^{-1}x$. Thus, the restricted isometry property

$$(6) \qquad \sqrt{1 - \delta}\, \|z\|_2 \leq \|Az\|_2 \leq \sqrt{1 + \delta}\, \|z\|_2 \qquad \text{whenever } \|B^{-1}z\|_0 \leq t$$

would be more appropriate. This property—a special case of the D-restricted isometry property [3] where the dictionary D is the invertible matrix B—holds for properly scaled subgaussian random matrices $A \in \mathbb{R}^{m \times n}$ with failure probability at most $2\exp(-c\delta^2 m)$ provided $m \geq C\delta^{-2} t \ln(eN/t)$.

5

Still, this restricted isometry property is not yet appropriate to the present setting, which puts an ℓ_p-norm on the measurement vectors Az. Thus, the nonstandard restricted isometry property utilized here reads

$$(7) \qquad\qquad \alpha\|z\|_2 \leq \|Az\|_p \leq \beta\|z\|_2 \qquad \text{whenever } \|B^{-1}z\|_0 \leq t.$$

It is to be noted that the constants α and β need not be close to one—they could e.g. scale with m. The only requirement is that the restricted isometry ratio $\gamma := \beta/\gamma$ is bounded by an absolute constant. For the main result to be nonvacuous, it is important to point out that this nonstandard restricted isometry property holds for subgaussian matrices $A \in \mathbb{R}^{m \times N}$ with failure probability at most $2\exp(-cm)$ provided $m \geq Ct\ln(eN/t)$, with constants c, C depending on γ, hence being absolute. To see this, one could follow typical proofs of the D-restricted isometry property while making suitable modifications, or realize that (7) is actually a consequence of the usual D-restricted isometry property, as revealed by the observation below. Thus, the nonstandard restricted isometry propery (7) is a weaker assumption than the standard restricted isometry property, in the sense that it holds with overwhelming probability for at least as many ensembles as (6). In fact, it holds for more: with $B = I_N$, the ℓ_1-version of (7) holds for Laplace ensembles in the optimal regime $m \asymp t\ln(eN/t)$, see [5], while (6) necessitates $m = \Omega(t\ln^2(eN/t))$, see [1].

Proposition 3. Given $p' \leq p \leq 2$, if an ensemble of random matrices satisfies the ℓ_p-version of (7) with failure probability at most $2\exp(-cm)$ provided $m \geq Ct\ln(eN/t)$, then it satisfies the $\ell_{p'}$-version of (7) with failure probability at most $2\exp(-c'm)$ provided $m \geq C't\ln(eN/t)$.

Proof. Let $\theta \in (0, 1)$ to be chosen soon. Assuming that $(1-\theta)m \geq Ct\ln(eN/t)$, the event that all the row-submatrices A_I, $|I| \geq (1-\theta)m$, of a matrix $A \in \mathbb{R}^{m \times N}$ from the random ensemble satisfy the ℓ_p-version of (7) occurs with failure probability at most

$$\sum_{k=(1-\theta)m}^{m} \binom{m}{k} \times 2\exp(-ck) \leq \sum_{\ell=0}^{\theta m} \binom{m}{\ell} \times 2\exp(-c(1-\theta)m) \leq \left(\frac{em}{\theta m}\right)^{\theta m} \times 2\exp(-c(1-\theta)m)$$

$$= 2\exp\big(-(c(1-\theta) - \theta\ln(e/\theta))m\big).$$

In this event, for any $z \in \mathbb{R}^N$ with $\|B^{-1}z\|_0 \leq t$, one has $\|Az\|_{p'} \leq m^{1/p'-1/p}\|Az\|_p$, so that

$$\|Az\|_{p'} \leq \beta'\|z\|_2, \qquad \beta' := \beta m^{1/p'-1/p}.$$

Moreover, according to Stechkin's estimate for the error of best sparse approximation in ℓ_p in term of the norm in $\ell_{p'}$, see e..g. [7, Theorem 2.5], there is some $I \subseteq [1:m]$ of size $|I| = (1-\theta)m$ such that $\|A_I z\|_p \leq (1/(\theta m)^{1/p'-1/p})\|Az\|_{p'}$, Since $\|A_I z\|_p \geq \alpha\|z\|_2$, one derives that

$$\|Az\|_{p'} \geq \alpha'\|z\|_2, \qquad \alpha' := \alpha(\theta m)^{1/p'-1/p}.$$

Choosing the constant $\theta \in (0, 1)$ such that $c(1-\theta) - \theta\ln(e/\theta) = c/2 =: c'$ and setting $C' = C/(1-\theta)$, it has been proved that, provided $m \geq C't\ln(eN/t)$, the $\ell_{p'}$-version of (7) holds with failure probability at most $2\exp(-c'm)$, while the restricted isometry ratio $\gamma' = \beta'/\alpha' = \gamma(1/\theta)^{1/p'-1/p}$ is bounded by an absolute constant. $\qquad \square$

2.3 Robust null space property from nonstandard restricted isometry property

The robust null space property, introduced in [7, Section 4.3], is a sufficient—and somewhat necessary—condition for robust sparse recovery by ℓ_1-minimization (which corresponds to $\lambda = 0$). The standard robust null space property of order s with constants $\rho \in (0, 1)$ and $\tau > 0$, with respect to a norm $\|\cdot\|$ on $\mathbb{R}^m$, reads:

$$\|v_S\|_1 \leq \rho\,\|v_{S^c}\|_1 + \tau\,\|Av\| \qquad \text{for all } v \in \mathbb{R}^N \text{ and } S \subseteq [1\colon N] \text{ with } |S| = s.$$

It is well-known that the standard restricted isometry property (1) implies this standard robust null space property (for $\|\cdot\| = \sqrt{s}\|\cdot\|_2$). Here, it is shown that the nonstandard restricted isometry property (7) implies a nonstandard robust null space property order s with constants $\rho \in (0, 1)$ and $\tau > 0$, which reads:

$$(8) \quad \|(B^{-1}v)_S\|_1 \leq \rho\,\|(B^{-1}v)_{S^c}\|_1 + \tau\,\sqrt{s}\|Av\|_p \qquad \text{for all } v \in \mathbb{R}^N \text{ and } S \subseteq [1\colon N] \text{ with } |S| = s.$$

Special cases of the argument are routine, see e.g. [4, Theorem 14.4] for the case $p = 1$ and $B = \mathrm{I}_N$.

Proposition 4. Given $p \geq 1$, if a matrix $A \in \mathbb{R}^{m \times N}$ satisfies the restricetd isometry property (7) of order $t = \lceil ((1+\rho)\gamma\kappa_B/\rho)^2 s \rceil$ with ratio $\gamma = \beta/\alpha$, then it satisfies the robust null space property (8) of order s with constants ρ and $\tau = (1 + \rho)\|B^{-1}\|_{2\to 2}\sqrt{s}/\alpha$ with respect to the ℓ_p-norm on $\mathbb{R}^m$.

Proof. Fixing $v \in \mathbb{R}^N$, note that it is enough to prove (8) for an index set S of s largest absolute entries of $B^{-1}v$. Let $T_0 \supseteq S$ be an index set of t largest absolute entries of $B^{-1}v$, and define also index sets $T_1, T_2, \ldots$ such that

$$T_1 \text{ is an index set of next } t \text{ largest absolute entries of } B^{-1}v,$$
$$T_2 \text{ is an index set of next } t \text{ largest absolute entries of } B^{-1}v,$$
$$\vdots$$

For any $k \geq 1$, the inequality $\|(B^{-1}v)_{T_k}\|_2 \leq \|(B^{-1}v)_{T_{k-1}}\|_1/\sqrt{t}$ is easily obtained by comparing averages. Now, using (among other things) the leftmost inequality of (7), one can write

$$\|B(B^{-1}v)_{T_0}\|_2 \leq \frac{1}{\alpha}\|AB(B^{-1}v)_{T_0}\|_p = \frac{1}{\alpha}\Big\|AB\Big(B^{-1}v - \sum_{k\geq 1}(B^{-1}v)_{T_k}\Big)\Big\|_p$$

$$\leq \frac{1}{\alpha}\Big(\sum_{k\geq 1}\|AB(B^{-1}v)_{T_k}\|_p + \|Av\|_p\Big).$$

Next, using (among other things) the rightmost inequality of (7), one continues with

$$\|B(B^{-1}v)_{T_0}\|_2 \leq \frac{\beta}{\alpha} \sum_{k \geq 1} \|B(B^{-1}v)_{T_k}\|_2 + \frac{1}{\alpha}\|Av\|_p$$

$$\leq \gamma\|B\|_{2\to 2} \sum_{k \geq 1} \|(B^{-1}v)_{T_k}\|_2 + \frac{1}{\alpha}\|Av\|_p$$

$$\leq \gamma\|B\|_{2\to 2} \sum_{k \geq 1} \frac{1}{\sqrt{t}}\|(B^{-1}v)_{T_{k-1}}\|_1 + \frac{1}{\alpha}\|Av\|_p$$

$$\leq \frac{\gamma\|B\|_{2\to 2}}{\sqrt{t}}\|B^{-1}v\|_1 + \frac{1}{\alpha}\|Av\|_p.$$

Then, a bound on $\|(B^{-1}v)_S\|_1$ is obtained via

$$\|(B^{-1}v)_S\|_1 \leq \sqrt{s}\|(B^{-1}v)_S\|_2 \leq \sqrt{s}\|(B^{-1}v)_{T_0}\|_2 \leq \sqrt{s}\|B^{-1}\|_{2\to 2}\|B(B^{-1}v)_{T_0}\|_2$$

$$\leq \gamma\kappa_B\sqrt{\frac{s}{t}}\|B^{-1}v\|_1 + \frac{\sqrt{s}\|B^{-1}\|_{2\to 2}}{\alpha}\|Av\|_p \leq \frac{\rho}{1+\rho}\|B^{-1}v\|_1 + \frac{\sqrt{s}\|B^{-1}\|_{2\to 2}}{\alpha}\|Av\|_p,$$

where the value $t = \lceil ((1+\rho)\gamma\kappa_B/\rho)^2 s\rceil$ was used in the last step. Finally, decomposing $\|B^{-1}v\|_1$ as $\|(B^{-1}v)_S\|_1 + \|(B^{-1}v)_{S^c}\|_1$ and rearranging, one deduces that

$$\|(B^{-1}v)_S\|_1 \leq \rho\|(B^{-1}v)_{S^c}\|_1 + \frac{(1+\rho)\|B^{-1}\|_{2\to 2}\sqrt{s}}{\alpha}\|Av\|_p,$$

which is the announced robust null space property. $\qquad\square$

3   The sparsity bound and its proof

In general, without any assumption that $y \in \mathbb{R}^m$ comes from a sparse vector, one can guarantee that, for (one of) the output(s) x^λ of the LASSO-type procedure, the vector $B^{-1}x^\lambda$ is m-sparse. Indeed, assuming uniqueness of x^λ for simplicity, if $S^\lambda = \mathrm{supp}(B^{-1}x^\lambda)$ had size $|S^\lambda| > m$, then the columns of $AB \in \mathbb{R}^{m \times N}$ indexed by S^λ would be linearly dependent, hence there would exist some $v \in \mathbb{R}^N$ supported on S^λ such that $ABv = 0$. Thus, for $u := Bv \in \mathbb{R}^N$ and $t \in \mathbb{R}$ small enough in absolute value, the vector $x^\lambda + tu$ would be another LASSO-type minimizer, in view of $A(x^\lambda + tu) = Ax^\lambda$ and of

$$\|B^{-1}(x^\lambda + tu)\|_1 = \langle \mathrm{sgn}(B^{-1}(x^\lambda + tu)), B^{-1}(x^\lambda + tu)\rangle = \langle \mathrm{sgn}(B^{-1}x^\lambda), B^{-1}x^\lambda + tB^{-1}u\rangle$$

$$= \|B^{-1}x^\lambda\|_1 + \frac{t}{\nu^\lambda}\langle B^\top A^\top\big(\mathrm{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}\big), B^{-1}u\rangle$$

$$= \|B^{-1}x^\lambda\|_1 + \frac{t}{\nu^\lambda}\langle \mathrm{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}, Au\rangle$$

$$= \|B^{-1}x^\lambda\|_1.$$

The main result below shows that, under a favorable condition on $A \in \mathbb{R}^{m \times N}$, if $y \in \mathbb{R}^m$ does come from a sparse vector, then B^{-1} applied to the output of the LASSO-type procedure is not simply m-sparse, but Cs-sparse, even with measurement error such that $\|e\|_p \leq c \min\{\|y\|_p, \lambda^{1/(q-r)}\}$. The constants C and c, appearing under different names below, have not been optimized.

Theorem 5. Let $p \in [1, 2]$, $q \geq 1$, and $r \geq 1$. Consider a vector $x \in \mathbb{R}^N$ such that $B^{-1}x$ is s-sparse for some invertible matrix $B \in \mathbb{R}^{N \times N}$ with condition number $\kappa_B := \|B\|_{2 \to 2} \|B^{-1}\|_{2 \to 2}$. Suppose that the vector x is measured via $y = Ax + e \in \mathbb{R}^m$ for some error vector $e \in \mathbb{R}^m$ with $\|e\|_p \leq (1/3)\|y\|_p$ and some matrix $A \in \mathbb{R}^{m \times N}$ satisfying the nonstandard restricted isometry property of order $t := \lfloor (6\gamma\kappa_B)^2 s \rfloor + 1$ with ratio $\gamma = \beta/\alpha$, i.e.,

$$\alpha\|z\|_2 \leq \|Az\|_p \leq \beta\|z\|_2 \qquad \text{whenever } \|B^{-1}z\|_0 \leq t.$$

Then, for any $\lambda > \lambda^* = 2^{q-1}\beta^r \|B\|_{2 \to 2}^r \|e\|_p^{q-r}$, the solution x^λ of the LASSO-type procedure

$$\underset{z \in \mathbb{R}^N}{\text{minimize}} \; \frac{1}{q}\|y - Az\|_p^q + \lambda \frac{1}{r}\|B^{-1}z\|_1^r$$

has sparsity at most proportional to s, namely

$$\|B^{-1}x^\lambda\|_0 \leq \lfloor \chi^2 s \rfloor, \qquad \chi := 6\gamma\kappa_B.$$

Proof (Part I). The content of this part of the proof is valid for any $e \in \mathbb{R}^m$, but its conclusion only pertains to $e = 0$. The case $p = 1$ or $q = 1$ is treated apart, as it is not covered by Proposition 2.

The case $p > 1$ and $q > 1$. The first step consists in bounding $\|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1$ from above. To do so, according to Proposition 4 specified with $\rho = 1$, the robust null space property (8) of order s holds. When applied to $v = x - x^\lambda$ and $S = \text{supp}(B^{-1}x)$, while taking $\|(B^{-1}v)_{S^c}\|_1 = \|(B^{-1}x^\lambda)_{S^c}\|_1$ and $\|(B^{-1}v)_S\|_1 = \|B^{-1}x - (B^{-1}x^\lambda)_S\|_1 \geq \|B^{-1}x\|_1 - \|(B^{-1}x^\lambda)_S\|_1$ into account, it implies that $\|B^{-1}x\|_1 - \|(B^{-1}x^\lambda)_S\|_1 \leq \|(B^{-1}x^\lambda)_{S^c}\|_1 + (2\|B^{-1}\|_{2 \to 2}\sqrt{s}/\alpha)\|A(x - x^\lambda)\|_p$. In other words, one has

$$\|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1 \leq \frac{2\|B^{-1}\|_{2 \to 2}\sqrt{s}}{\alpha}\|A(x - x^\lambda)\|_p$$

(9)
$$\leq \frac{2\|B^{-1}\|_{2 \to 2}\sqrt{s}}{\alpha}\left(\|y - Ax^\lambda\|_p + \|e\|_p\right).$$

The second step consists in bounding $\|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1$ from below. To do so, one relies on (2)–(3) to remark that $\|B^{-1}x^\lambda\|_1 = (1/\nu^\lambda)\langle B^\top A^\top\left(\text{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}\right), B^{-1}x^\lambda\rangle$ and that $\|B^{-1}x\|_1 \geq (1/\nu^\lambda)\langle B^\top A^\top\left(\text{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}\right), B^{-1}x\rangle$. Thus, one deduces that

$$\|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1 \geq \frac{1}{\nu^\lambda}\langle B^\top A^\top\left(\text{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}\right), B^{-1}x - B^{-1}x^\lambda\rangle$$

$$= \frac{1}{\nu^\lambda}\langle \text{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}, A(x - x^\lambda)\rangle$$

$$= \frac{1}{\nu^\lambda}\langle \text{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}, y - Ax^\lambda - e\rangle$$

(10)
$$= \frac{1}{\nu^\lambda}\left(\|y - Ax^\lambda\|_p^p - \langle \text{sgn}(y - Ax^\lambda)\,|y - Ax^\lambda|^{p-1}, e\rangle\right).$$

9

By contradiction to the desired conclusion $\|B^{-1}x^\lambda\|_0 \leq \lfloor \chi^2 s \rfloor$, assume now that $S^\lambda := \mathrm{supp}(B^{-1}x^\lambda)$ has size $|S^\lambda| \geq \lfloor \chi^2 s \rfloor + 1$. One can then consider a subset T of S^λ having size $|T| = \lfloor \chi^2 s \rfloor + 1$. The vector $u := \left[B^\top A^\top \left(\mathrm{sgn}(y - Ax_\lambda) \, |y - Ax_\lambda|^{p-1} \right) \right]_T$ is then t-sparse and the nonstandard restricted isometry property yields $\|ABu\|_p \leq \beta \|Bu\|_2 \leq \beta \|B\|_{2\to2} \|u\|_2$. Moreover,

$$
\begin{aligned}
\|u\|_2^2 = \langle B^\top A^\top \left(\mathrm{sgn}(y - Ax_\lambda) \, |y - Ax_\lambda|^{p-1} \right), u \rangle &= \langle \mathrm{sgn}(y - Ax_\lambda) \, |y - Ax_\lambda|^{p-1}, ABu \rangle \\
&\leq \| \mathrm{sgn}(y - Ax_\lambda) \, |y - Ax_\lambda|^{p-1} \|_{p'} \|ABu\|_p,
\end{aligned}
$$

where $p' = p/(p-1)$ is the conjugate exponent of p. From the estimation of $\|ABu\|_p$ and the observation that $\| \mathrm{sgn}(y - Ax_\lambda) \, |y - Ax_\lambda|^{p-1} \|_{p'} = \|y - Ax^\lambda\|_p^{p-1}$, the above inequality rearranges as $\|y - Ax^\lambda\|_p^{p-1} \geq (1/(\beta \|B\|_{2\to2}))\|u\|_2$. But according to (2) and $T \subseteq S^\lambda$, one has $\|u\|_2 = \nu^\lambda \sqrt{|T|}$, and in turn

$$
\tag{11} \|y - Ax^\lambda\|_p^{p-1} \geq \frac{\nu^\lambda \sqrt{|T|}}{\beta \|B\|_{2\to2}}.
$$

With all of this in place, one can easily conclude in the case $e = 0$—thus proving Theorem 1. Indeed, (10) combined with (11) and with $|T| > \chi^2 s$ yields

$$
\|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1 > \frac{\chi \sqrt{s}}{\beta \|B\|_{2\to2}} \|y - Ax^\lambda\|_p,
$$

where one used $y - Ax^\lambda \neq 0$ (see the end of the proof of Proposition 2), while (9) reads

$$
\|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1 \leq \frac{2\|B^{-1}\|_{2\to2} \sqrt{s}}{\alpha} \|y - Ax^\lambda\|_p.
$$

Making the choice $\chi = 2\gamma \kappa_B$, one obtains a contradiction to conclude that $\|B^{-1}x^\lambda\|_0 \leq \lfloor \chi^2 s \rfloor$.

The case $p = 1$ or $q = 1$. A limiting argument does the trick here. Fixing $r \geq 1$ and $\lambda > 0$, consider decreasing sequences $(p_n)_{n\geq1}$ and $(q_n)_{n\geq1}$ of indices greater than 1 but converging to p and q, respectively. For each $n \geq 1$, denote by $x^{(n)} \in \mathbb{R}^N$ the LASSO-type minimizer corresponding to p_n and q_n, which is now known to be $\lfloor \chi^2 s \rfloor$-sparse. Note that the sequence $(x^{(n)})_{n\geq1}$ is bounded, as can be seen by evaluating the LASSO-type objective function at $z = 0$ to write

$$
\lambda \frac{1}{r} \|B^{-1}x^{(n)}\|_1^r \leq \frac{1}{q_n} \|y - Ax^{(n)}\|_{p_n}^{q_n} + \lambda \frac{1}{r} \|B^{-1}x^{(n)}\|_1^r \leq \frac{1}{q_n} \|y\|_{p_n}^{q_n} \leq \|y\|_1^{q_n} \leq \max\{\|y\|_1^{q_1}, \|y\|_1^{q}\}.
$$

Therefore, there is a subsequence $(x^{(n_k)})_{k\geq1}$ extracted from $(x^{(n)})_{n\geq1}$ and converging to some $x^* \in \mathbb{R}^N$, say. With $\widetilde{x} \in \mathbb{R}^N$ denoting (one of) the LASSO-type minimizer(s) corresponding to p and q, one has, for any $k \geq 1$,

$$
\frac{1}{q_{n_k}} \|y - Ax^{(n_k)}\|_{p_{n_k}}^{q_{n_k}} + \lambda \frac{1}{r} \|B^{-1}x^{(n_k)}\|_1^r \leq \frac{1}{q_{n_k}} \|y - A\widetilde{x}\|_{p_{n_k}}^{q_{n_k}} + \lambda \frac{1}{r} \|B^{-1}\widetilde{x}\|_1^r,
$$

which passes to the limit as $k \to +\infty$ to give

$$
\frac{1}{q} \|y - Ax^*\|_p^q + \lambda \frac{1}{r} \|B^{-1}x^*\|_1^r \leq \frac{1}{q} \|y - A\widetilde{x}\|_p^q + \lambda \frac{1}{r} \|B^{-1}\widetilde{x}\|_1^r,
$$

10

showing that x^* is a LASSO-type minimizer corresponding to p and q. But, the vector x^* being the limit of $\lfloor \chi^2 s \rfloor$-sparse vectors, it must itself be a $\lfloor \chi^2 s \rfloor$-sparse vector. □

Dealing with the general situation where $e \neq 0$ is not much more complicated. One shall use the following observation, which may be folklore knowledge.

Lemma 6. The function $\lambda \mapsto \|y - Ax^\lambda\|_p$ is continuous (for $p > 1$) and nondecreasing on $(0, +\infty)$, with $\lim_{\lambda \to 0} \|y - Ax^\lambda\|_p \leq \|e\|_p$ and $\lim_{\lambda \to +\infty} \|y - Ax^\lambda\|_p = \|y\|_p$.

Proof. For $\mu > \lambda > 0$, using the optimality of x^μ and the optimality of x^λ, one can write

$$
\frac{1}{q}\|y - Ax^\mu\|_p^q + \mu\frac{1}{r}\|B^{-1}x^\mu\|_1^r \leq \frac{1}{q}\|y - Ax^\lambda\|_p^q + \mu\frac{1}{r}\|B^{-1}x^\lambda\|_1^r
$$

$$
= \frac{\mu}{\lambda}\left[\frac{1}{q}\|y - Ax^\lambda\|_p^q + \lambda\frac{1}{r}\|B^{-1}x^\lambda\|_1^r\right] - \left(\frac{\mu}{\lambda} - 1\right)\frac{1}{q}\|y - Ax^\lambda\|_p^q
$$

$$
\leq \frac{\mu}{\lambda}\left[\frac{1}{q}\|y - Ax^\mu\|_p^q + \lambda\frac{1}{r}\|B^{-1}x^\mu\|_1^r\right] - \left(\frac{\mu}{\lambda} - 1\right)\frac{1}{q}\|y - Ax^\lambda\|_p^q.
$$

Rearranging the latter yields

$$
\left(\frac{\mu}{\lambda} - 1\right)\frac{1}{q}\|y - Ax^\lambda\|_p^q \leq \left(\frac{\mu}{\lambda} - 1\right)\frac{1}{q}\|y - Ax^\mu\|_p^q,
$$

i.e., $\|y - Ax^\lambda\|_p \leq \|y - Ax^\mu\|_p$ thanks to $\mu/\lambda - 1 > 0$. Thus, the monotonicity claim is proved.

Next, for $\lambda > 0$, since the LASSO-type objective function at x^μ is no larger than at x and at 0, one has

$$
\frac{1}{q}\|y - Ax^\lambda\|_p^q + \lambda\frac{1}{r}\|B^{-1}x^\lambda\|_1^r \leq \begin{cases} \dfrac{1}{q}\|e\|_p^q + \lambda\dfrac{1}{r}\|B^{-1}x\|_1^r, \\[2mm] \dfrac{1}{q}\|y\|_p^q. \end{cases}
$$

The first inequality gives $\|y - Ax^\lambda\|_p^q \leq \|e\|_p^q + (\lambda q/r)\|B^{-1}x\|_1^r$. Letting $\lambda \to 0$ yields the claim about the first limit. The second inequality gives $\|B^{-1}x^\lambda\|_1^r \leq (r/(\lambda q))\|y\|_p^q$. Letting $\lambda \to +\infty$ yields $B^{-1}x^\lambda \to 0$, hence $x^\lambda \to 0$, and the claim about the second limit follows.

Finally, to establish the continuity for $p > 1$, consider a sequence $(\lambda_n)_{n \geq 1}$ of positive numbers converging to some $\lambda > 0$. It is enough to show that the sequence $(Ax^{\lambda_n})_{n \geq 1}$ converges to Ax^λ. By contradiction, if this is not so, then there exists $\varepsilon > 0$ and a subsequence $(Ax^{\lambda_{n_k}})_{k \geq 1}$ such that $\|Ax^{\lambda_{n_k}} - Ax^\lambda\|_p \geq \varepsilon$ for all $k \geq 1$. By the boundedness of the sequence $(x^{\lambda_{n_k}})_{k \geq 1}$ (recall that $\|B^{-1}x^{\lambda_{n_k}}\|_1^r \leq (r/(\lambda_{n_k}q))\|y\|_p^q$), there is also a subsequence $(x^{\lambda_{n_{k_\ell}}})_{\ell \geq 1}$ converging to some $\tilde{x} \in \mathbb{R}^N$. Passing to the limit as $\ell \to +\infty$ in $\|Ax^{\lambda_{n_{k_\ell}}} - Ax^\lambda\|_p \geq \varepsilon$ shows in particular that $A\tilde{x} \neq Ax^\lambda$. Passing to the limit in

$$
\frac{1}{q}\|y - Ax^{\lambda_{n_{k_\ell}}}\|_p^q + \lambda_{n_{k_\ell}}\frac{1}{r}\|B^{-1}x^{\lambda_{n_{k_\ell}}}\|_1^r \leq \frac{1}{q}\|y - Ax^\lambda\|_p^q + \lambda_{n_{k_\ell}}\frac{1}{r}\|B^{-1}x^\lambda\|_1^r
$$

also shows that $\widetilde{x}$ is a LASSO-type minimizer for the parameter λ, and hence so is $(\widetilde{x} + x^\lambda)/2$. But, by strict convexity of the ℓ_p-norm for $p \geq 1$, one deduces that $A\widetilde{x} = Ax^\lambda$. This provides the required contradiction to conclude the proof. $\qquad\square$

Proof of Theorem 5 (Part II). To finish with the general situation where $e \neq 0$, one separates the proof in three subparts: the first two subparts assume $p > 1$ and $q > 1$ and establish the existence of λ^* before estimating it, and the third subpart treats the case $p = 1$ or $q = 1$.

Existence of λ^*. By the assumption $\|e\|_p \leq (1/3)\|y\|_p$, one notices that $2\|e\|_p$ belongs to the interval $\left[\lim_{\lambda \to 0} \|y - Ax^\lambda\|_p, \lim_{\lambda \to +\infty} \|y - Ax^\lambda\|_p\right]$, so Lemma 6 guarantees the existence of $\lambda^\flat > 0$ such that $\|y - Ax^{\lambda^\flat}\|_p = 2\|e\|_p$. Considering from now on any $\lambda > \lambda^\flat$, Lemma 6 also guarantees that

$$\|y - Ax^\lambda\|_p \geq 2\|e\|_p, \qquad \text{i.e.,} \quad \|e\|_p \leq \frac{1}{2}\|y - Ax^\lambda\|_p.$$

Similarly to Part I of the proof but with χ to be chosen differently, it is assumed by contradiction that $|S^\lambda| \geq \lfloor \chi^2 s \rfloor + 1$ and one again picks some $T \subseteq S^\lambda$ with $|T| = \lfloor \chi^2 s \rfloor + 1$. One shall still exploit the inequalities (9), (10), and (11), which remain valid. Starting from (10), in view of

$$\langle \mathrm{sgn}(y - Ax^\lambda) \, |y - Ax^\lambda|^{p-1}, e \rangle \leq \|\mathrm{sgn}(y - Ax^\lambda) \, |y - Ax^\lambda|^{p-1}\|_{p'} \|e\|_p = \|y - Ax^\lambda\|_p^{p-1} \|e\|_p,$$

one obtains

$$\|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1 \geq \frac{1}{\nu^\lambda} \|y - Ax^\lambda\|_p^{p-1} \left(\|y - Ax^\lambda\|_p - \|e\|_p\right)$$

$$\geq \frac{1}{\nu^\lambda} \|y - Ax^\lambda\|_p^{p-1} \frac{1}{2} \|y - Ax^\lambda\|_p$$

$$\text{(12)} \qquad\qquad \geq \frac{\sqrt{|T|}}{2\beta\|B\|_{2\to2}} \|y - Ax^\lambda\|_p,$$

where (11) was used in the last step. As for (9), it now implies

$$\text{(13)} \qquad\qquad \|B^{-1}x\|_1 - \|B^{-1}x^\lambda\|_1 \leq \frac{3\|B^{-1}\|_{2\to2}\sqrt{s}}{\alpha} \|y - Ax^\lambda\|_p.$$

Because $|T| > \chi^2 s$, the inequalities (12) and (13) become contradictory for the choice $\chi = 6\gamma\kappa_B$. In summary, for any $\lambda \geq \lambda^\flat$, it has been proved that $\|B^{-1}x^\lambda\|_0 \leq \lfloor \chi^2 s \rfloor$. One can therefore select $\lambda^* = \lambda^\flat$ in the statement of Theorem 5.

Estimation of λ^*. The objective here is to show that $\lambda^\flat \leq \lambda^\sharp$ for a more explicit $\lambda^\sharp$, so that $\|B^{-1}x^\lambda\|_0 \leq \lfloor \chi^2 s \rfloor$ holds whenever λ is larger than $\lambda^\sharp$, which can then be selected as λ^*. Since it is now known that $s^\flat := \|B^{-1}x^{\lambda^\flat}\|_0 \leq \lfloor \chi^2 s \rfloor$, taking $T = \mathrm{supp}(B^{-1}x^{\lambda^\flat})$ in (11) gives

$$\|y - Ax^{\lambda^\flat}\|_p^{p-1} \geq \frac{\nu^{\lambda^\flat}\sqrt{s^\flat}}{\beta\|B\|_{2\to2}} = \frac{\lambda^\flat \|y - Ax^{\lambda^\flat}\|_p^{p-q} \|B^{-1}x^{\lambda^\flat}\|_1^{r-1} \sqrt{s^\flat}}{\beta\|B\|_{2\to2}},$$

which in turn yields

$$(14) \qquad \|y - Ax^{\lambda^\flat}\|_p^{q-1} \geq \frac{\lambda^\flat \|B^{-1}x^{\lambda^\flat}\|_1^{r-1} \sqrt{s^\flat}}{\beta \|B\|_{2\to 2}}, \qquad \text{hence} \quad (2\|e\|_p)^{q-1} \geq \frac{\lambda^\flat \|B^{-1}x^{\lambda^\flat}\|_1^{r-1}}{\beta \|B\|_{2\to 2}}.$$

From here, one bounds from below the ℓ_1-norm of $B^{-1}x^{\lambda^\flat}$ as

$$\begin{aligned}
\|B^{-1}x^{\lambda^\flat}\|_1 &\geq \|B^{-1}x^{\lambda^\flat}\|_2 \geq \frac{1}{\|B\|_{2\to 2}}\|x^{\lambda^\flat}\|_2 \geq \frac{1}{\beta\|B\|_{2\to 2}}\|Ax^{\lambda^\flat}\|_p \\
&\geq \frac{1}{\beta\|B\|_{2\to 2}}(\|y\|_p - \|y - Ax^{\lambda^\flat}\|_p) \geq \frac{1}{\beta\|B\|_{2\to 2}}(3\|e\|_p - 2\|e\|_p) \\
&= \frac{1}{\beta\|B\|_{2\to 2}}\|e\|_p.
\end{aligned}$$

Substituting the latter into (14), one derives that $\lambda^\flat \leq 2^{q-1}\beta^r\|B\|_{2\to 2}^r\|e\|_p^{q-r} =: \lambda^\sharp$. This justifies the announced expression for λ^*.

The case $p = 1$ or $q = 1$. Let $(p_n)_{n\geq 1}$ and $(q_n)_{n\geq 1}$ be decreasing sequences of indices greater than 1 but converging to p and q, respectively. Any regularization parameter λ in the admissible range for p and q is also in the admissible range for p_n and q_n when n is large enough. The limiting argument of Part I then applies *mutatis mutandis*, since the $\lfloor \chi^2 s \rfloor$-sparsity of all the corresponding $x^{(n)}$'s is ensured. $\qquad\square$

Remark. The expression $\lambda^* = 2^{q-1}\beta^r\|B\|_{2\to 2}^r\|e\|_p^{q-r}$ is certainly improvable, especially when $q \leq r$. For instance, when $q = r$, which includes square-root LASSO and square LASSO, the threshold λ^* reduces to a positive quantity independent of the magnitude of e. But this does not transition continuously to the case $e = 0$, where one can take $\lambda^* = 0$, as revealed in Theorem 1.

References

[1] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. *Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling.* Constructive Approximation 34.1 (2011): 61–88.

[2] A. Belloni, V. Chernozhukov, and L. Wang. *Square-root Lasso: pivotal recovery of sparse signals via conic programming.* Biometrika 98.4 (2011): 791–806.

[3] E. J. Candès, Y. C. , Eldar, D. Needell, and P. Randall. *Compressed sensing with coherent and redundant dictionaries.* Applied and Computational Harmonic Analysis 31.1 (2011): 59–73.

[4] S. Foucart. Mathematical Pictures at a Data Science Exhibition. Cambridge University Press, 2022.

[5] S. Foucart and M.-J. Lai. *Sparse recovery with pre-Gaussian random matrices.* Studia Math 200.1 (2010): 91–102.

[6] S. Foucart and D. Koslicki. *Sparse recovery by means of nonnegative least squares.* IEEE Signal Processing Letters 21.4 (2014): 498–502.

[7] S. Foucart and H. Rauhut. A Mathematical Introduction to Compressive Sensing. Birkhäuser, 2013.

[8] S. Foucart, E. Tadmor, and M. Zhong. *On the sparsity of LASSO minimizers in sparse data recovery.* Constructive Approximation, accepted. arXiv preprint arXiv:2004.04348.

[9] T. Hastie, R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning (2nd edition). Springer, 2009.

[10] E. Tadmor, S. Nezzar, and L. Vese. *Multiscale hierarchical decomposition of images with applications to deblurring, denoising, and segmentation.* Communications in Mathematical Sciences 6.2 (2008): 281–307.

[11] R. J. Tibshirani and J. Taylor. *The solution path of the generalized LASSO.* The Annals of Statistics 39.3 (2011): 1335–1371.

[12] H. Wang, G. Li, and G. Jiang. *Robust regression shrinkage and consistent variable selection through the LAD-Lasso.* Journal of Business & Economic Statistics 25.3 (2007): 347–355.