WHEN CAN YOU TRUST FEATURE SELECTION? – II: ON THE EFFECTS OF RANDOM DATA ON CONDITION IN STATISTICS AND OPTIMISATION

ALEXANDER BASTOUNIS, FELIPE CUCKER, AND ANDERS C. HANSEN

ABSTRACT. In Part I, we defined a LASSO condition number and developed an algorithm – for computing support sets (feature selection) of the LASSO minimisation problem – that runs in polynomial time in the number of variables and the logarithm of the condition number. The algorithm is trustworthy in the sense that if the condition number is infinite, the algorithm will run forever and never produce an incorrect output. In this Part II article, we demonstrate how finite precision algorithms (for example algorithms running floating point arithmetic) will fail on open sets when the condition number is large – but still finite. This augments Part I's result: If an algorithm takes inputs from an open set that includes at least one point with an infinite condition number, it fails to compute the correct support set for all inputs within that set. Hence, for any finite precision algorithm working on open sets for the LASSO problem with random inputs, our LASSO condition number – as a random variable – will estimate the probability of success/failure of the algorithm. We show that a finite precision version of our algorithm works on traditional Gaussian data for LASSO with high probability. The algorithm is trustworthy, specifically, in the random cases where the algorithm fails, it will not produce an output. Finally, we demonstrate classical random ensembles for which the condition number will be large with high probability, and hence where any finite precision algorithm on open sets will fail. We show numerically how commercial software fails on these cases.

1. INTRODUCTION

This article is the continuation of [7], which in the sequel we will refer to as Part I. Both here and in Part I, the *unconstrained LASSO feature selection problem* [33,50] is the main focus. Specifically, we are interested in computing, for fixed $\lambda \in \mathbb{Q}$, $\lambda > 0$, an element in

$$\Xi(y, A) = \{\mathsf{supp}(x) \,|\, x \in \operatorname*{argmin}_{\hat{x}\in\mathbb{R}^N} \|A\hat{x} - y\|_2^2 + \lambda\|\hat{x}\|_1\}, \tag{1.1}$$

where $(y, A) \in \mathbb{R}^m \times \mathbb{R}^{m\times N}$. The rationale is as follows: given the many AI-based algorithms in the computational sciences, with the potential for hallucinations and non-robustness [5, 20, 20, 21, 28, 31, 34, 36, 38, 39], the question of trustworthiness of algorithms is now becoming a crucial topic. For example, the European Commission [44] has been particularly vocal about its demand for trust in algorithms. However, with this new focus on trust in algorithms comes an important question: Which of the classical (non-AI-based) approaches are trustworthy, such as LASSO feature selection?

Part I first defines a LASSO condition number $\mathscr{C}_{\mathrm{UL}}(b, U)$ (see Definition 5.2) for any pair $(b, U) \in \mathbb{R}^m \times \mathbb{R}^{m\times N}$, and then provides the following Theorem (Theorem 1.2 there[1]) below. The model of computation is that any algorithm reads variable-precision approximations of the input $(b, U) \in \mathbb{R}^m \times \mathbb{R}^{m\times N}$. Most importantly, the set of variable precision algorithms contains the set of finite precision algorithms, which is a typical way of modelling algorithms using floating point arithmetic.

Remark 1.1 (Model of computation – Inexact input)**.** In practice, when trying to compute an element of $\Xi(y, A)$ in (1.1), we must assume that the A and y are given inexactly. This is because either we have: (1) an irrational input; or (2) the input is rational (for example $1/3$), but our computer expresses numbers in a certain base (typically base-2); (3) the computer uses floating-point arithmetic for which – in many cases –

[1]In all what follows, for simplicity, we will use a prefix 'I.' in the references to objects in Part I not contained here. Thus, for instance, Theorem 1.2 or equation (7.19) there become Theorem I.1.2 and equation (I.7.19) here.

the common backward-error analysis (popularized by Wilkinson [53]) translates the accumulation of round-off in a computation into a single-perturbation of the input data. Hence, we assume that algorithms access the input to whatever finite precision desired and that all computational operations are done exactly.

Theorem 1.2 (Main theorem of Part I)**.** *Consider the condition number $\mathscr{C}_{\mathrm{UL}}(b,U)$ defined in* (5.1)*.*

(1) We exhibit an algorithm Γ which, for any input pair $(b,U)\in\mathbb{R}^m\times\mathbb{R}^{m\times N}$, reads variable-precision approximations of (b,U). If $\mathscr{C}_{\mathrm{UL}}(b,U)<\infty$ then the algorithm halts and returns a correct value in $\Xi(b,U)$. The cost of this computation is

$$\mathcal{O}\left\{N^3\left[\log_2\left(N^2[\![(b,U)]\!]^2_{\max}\mathscr{C}_{\mathrm{UL}}(b,U)\right)\right]^2\right\}.$$

If, instead, $\mathscr{C}_{\mathrm{UL}}(b,U)=\infty$ then the algorithm runs forever.

(2) The condition number $\mathscr{C}_{\mathrm{UL}}(b,U)$ can be estimated in the following sense: There exists an algorithm that provides an upper bound on $\mathscr{C}_{\mathrm{UL}}(b,U)$, when it is finite, and runs forever when $\mathscr{C}_{\mathrm{UL}}(b,U)=\infty$.

(3) If $\Omega\subseteq\mathbb{R}^m\times\mathbb{R}^{m\times N}$ is an open set and there is a $(b,U)\in\Omega$ with $\mathscr{C}_{\mathrm{UL}}(b,U)=\infty$ then there is no algorithm that, for all input $(y,A)\in\Omega$, computes an element of $\Xi(y,A)$ given approximations to $(y,A)\in\Omega$. Moreover, for any randomised algorithm Γ^{ran} that always halts and any $p>1/2$, there exists $(y,A)\in\Omega$ and an approximate representation $(\tilde{y},\tilde{A})$ (for a precise statement see §9 in Part I [7]) of (y,A) so that $\Gamma^{\mathrm{ran}}(\tilde{y},\tilde{A})\notin\Xi(y,A)$ with probability at least p.

If $(b,U)\in\Omega$ is computable, then the failure point $(y,A)\in\Omega$ above can be made computable.

Remark 1.3 (Trustworthiness of algorithms – No wrong outputs)**.** By 'trustworthy algorithm' for a computational problem, we mean the following. If the computational problem takes only discrete values (as is the case when computing support sets of minimisers of optimisation problems), a trustworthy algorithm will always produce a correct answer – if it halts.

Note that (1) implies – in view of the question on trustworthiness – that our algorithm will never output a wrong solution, and thus if it halts, the output is always trustworthy. However, there are inputs for which it will not produce an answer. In view of (3) this is optimal in terms of existence of algorithms on open sets: Every algorithm will fail on some inputs, although not necessarily on inputs (b,U) where $\mathscr{C}_{\mathrm{UL}}(b,U)=\infty$, which is where our algorithm fails.

1.1. **The LASSO problem with random inputs.** To motivate our main results in this paper – Part II – we begin by asking the basic question:

Can commercial software for the LASSO problem be trusted on random data? If not, why?
Is there a link to our LASSO condition number? And how can a lack of trust be mitigated?

To illustrate the motivation behind this question we consider the following example, using a variety of different probability distributions on the input data.

Example 1.4 (**Testing algorithms for LASSO with random inputs).** We set $m=1$, $b=1$, and $\lambda=10^{-2}$ and generate each entry of a matrix $U\in\mathbb{R}^{1\times N}$ iid from three distributions: the exponential distribution with parameter 1, the normal distribution with mean 1 variance 10^{-4}, and the uniform distribution on $(0,1)$. This purposefully simplified situation is considered because it is easy to use Lemma 6.3 to compute $\Xi(b,U)$ as defined in (1.1) (this is a singleton with probability 1). We compare this answer (the ground truth) with the following procedure: we use Matlab's lasso routine to attempt to compute an element x in

$$\mathsf{Sol}^{\mathsf{UL}}(b,U):=\operatorname*{argmin}_{\hat{x}\in\mathbb{R}^N}\|U\hat{x}-b\|_2^2+\lambda\|\hat{x}\|_1. \tag{1.2}$$

and then set any values of x larger in absolute value than a parameter 'threshold' to 0 and consider the resulting vector's support. We do this 500 times for each choice of $N\in\{10,20,\ldots,10010\}$ inclusive. In

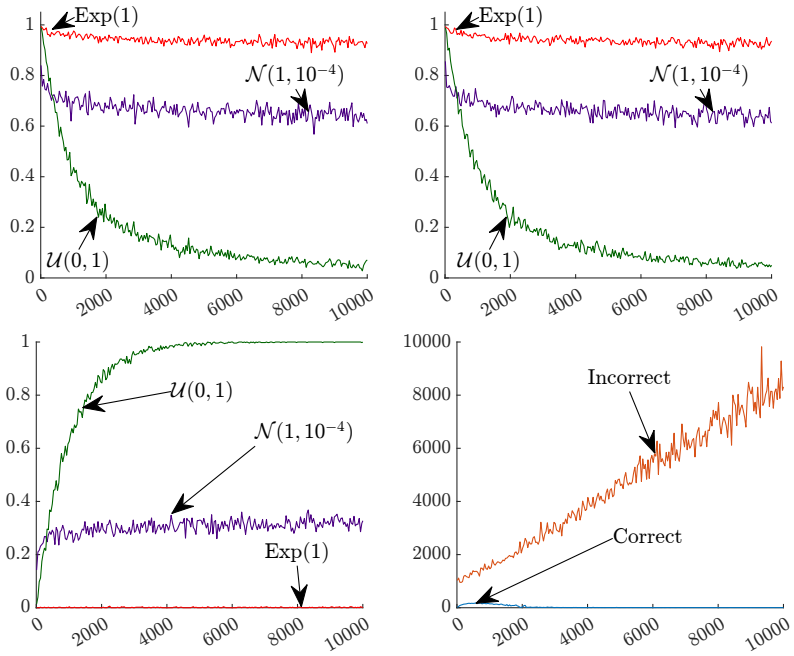

FIGURE 1. *(*$\mathbb{P}$**(LASSO has a unique minimiser)** $= 1$ **and** $\mathbb{P}(\mathscr{C}_{\mathrm{UL}}(b,U) < \infty) = 1$**, yet standard algorithms fail to compute the support set***).* Testing MATLAB's `lasso` on random iid inputs $(b, U) \in \mathbb{R}^1 \times \mathbb{R}^{1\times N}$ – according to the distributions $\mathcal{U}(a, b)$ (uniform), $\mathrm{Exp}(\nu)$ (exponential) and $\mathcal{N}(\mu, \sigma^2)$ (normal). The task is to compute the support set of a LASSO minimiser i.e. an element in $\Xi(b, U)$ from (1.1) with $\lambda = 10^{-2}$. All figures: the horizontal axis represents the dimension N. Top figures: the vertical axis represents the success rate $\frac{\#\text{ of successes}}{\#\text{ of trials}}$ with threshold value – see the text accompanying (1.2) – set to 10^{-3} and 10^{-12} for the left and right figures respectively. Bottom left figure: the proportion of trials (threshold = 10^{-12}) for each dimension which had a condition $\mathscr{C}_{\mathrm{UL}}(b, U)$ above $1,000$, for each distribution. Bottom right figure: the vertical axis is the median condition $\mathscr{C}_{\mathrm{UL}}(b, U)$ (across all distributions and trials) for that dimension for the cases where MATLAB was correct and MATLAB was incorrect (threshold = 10^{-12}). Note that for all the distributions considered $\mathbb{P}$(LASSO has a unique minimiser) $= 1$ and $\mathbb{P}(\mathscr{C}_{\mathrm{UL}}(b, U) < \infty) = 1$ for all N.

addition, we also compute the condition number $\mathscr{C}_{\mathrm{UL}}(b, U)$ and present the proportion of experiments for each dimension and distribution which had condition above $1,000$ as well as the median condition across all distributions of cases where Matlab fails to compute the correct support set and where Matlab computes the correct support set. The results are presented in Figure 1.

We see that for large N, MATLAB is more accurate when data are drawn from an exponential distribution instead of a normal distribution and similarly the algorithm is more accurate when data are drawn from a normal distribution instead of a uniform one. This also correlates with the size of the condition number: large median condition number correlates with low success rate.

2. MAIN RESULTS

Our main results can be described with three main theorems: Theorem 2.3 demonstrating a relation between failure of finite precision algorithms and our condition number $\mathscr{C}_{\mathrm{UL}}(b, U)$; Theorem 2.4, which shows asymptotic estimates of $\mathscr{C}_{\mathrm{UL}}(b, U)$, and thus helps explain Example 1.4; and Theorem 2.5 presenting conditions in classical statistics that allow us to obtain 'good/small' condition numbers $\mathscr{C}_{\mathrm{UL}}(b, U)$ – and

thus effective and trustworthy algorithms for these inputs (we have a stronger, yet more involved version presented as Theorem 3.3 in §3).

2.1. **Condition $\mathscr{C}_{\mathrm{UL}}(b,U)$ and failure of algorithms on random inputs.** The failure of MATLAB's `lasso` in Example 1.4 and Figure 1 yields the question: Why does the algorithm fail on these basic random LASSO problems? The key issue is that $\mathscr{C}_{\mathrm{UL}}(b,U)$ characterises failures of finite precision algorithms.

Remark 2.1 (Finite precision algorithms)**.** For a precise definition of a finite precision algorithm, see Definition 9.5. However, the concept can be explained simply: A finite precision algorithm with precision 2^{-k} (with $k \in \mathbb{N}$) can only read a dyadic approximation of the correct input with error bound 2^{-k}.

Remark 2.2 (Inputs that are represented exactly)**.** Given Remark 1.1, there are certain dyadic inputs for which a finite precision algorithm will have an exact representation. Hence, we can consider algorithms that are *correct on all inputs that can be represented exactly*. This concept is formally defined in Definition 9.6.

Theorem 2.3. *Let Γ_1 and Γ_2 be finite precision algorithms with precision 2^{-k} with an open domain $\Omega \subset \mathbb{R}^m \times \mathbb{R}^{m\times N}$ for the LASSO function Ξ defined in* (1.1)*. Suppose that Γ_1 is correct on all inputs that can be represented exactly in Ω (see Remarks 2.1 & 2.2) . Suppose that $\mathscr{C}_{\mathrm{UL}}(b,U) = \alpha$ with $0 < \frac{1}{\alpha} < 2^{-k-1}$ and that for some $r \in (\alpha^{-1}, 2^{-k-1})$ we have $\mathcal{B}_\infty(b,U,r) \subset \Omega$. Then, Γ_1 fails on a set F such that*

- *$\mathcal{B}_\infty(b,U,r) \setminus F$ has Lebesgue measure 0,*
- *$\mathcal{B}_\infty(b,U,\alpha^{-1}) \in F$.*

Moreover, Γ_2 either fails on the whole of $\mathcal{B}_\infty(b,U,\alpha^{-1})$ or on another open set $\theta \subset \mathcal{B}_\infty(b,U,r)$.

Theorem 2.3 demonstrates how finite precision algorithms will fail when $\mathscr{C}_{\mathrm{UL}}(b,U)$ is large relative to the precision of the algorithm. This means that if the probability that $\mathscr{C}_{\mathrm{UL}}(b,U)$ is large is high, it is highly likely that the algorithm will fail.

2.2. **Condition $\mathscr{C}_{\mathrm{UL}}(b,U)$ as a random variable.** Theorem 2.4 provides a theoretical explanation to the behaviours exhibited in Example 1.4 and the corresponding Figure 1.

Theorem 2.4. *Let $y \in \mathbb{R}$ be fixed and $A \in \mathbb{R}^{1\times N}$ be random with i.i.d. entries.*

- *If the entries of A follow an exponential distribution with parameter 1 then*

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{1}{t} < \mathscr{C}_{\mathrm{UL}}(y,A)\right) = \begin{cases} 1-\mathrm{e}^{-2t} & \text{for } t < |y| \\ 1 & \text{for } t \geq |y|. \end{cases}$$

- *If the entries of A follow the normal distribution $\mathcal{N}(\mu,\sigma^2)$, then $2\sqrt{2\ln(N)}[\mathscr{C}_{\mathrm{UL}}(y,A)]^{-1}$ converges in distribution to an exponential random variable with parameter $1/\sigma$.*
- *If the entries of A follow a uniform distribution $\mathcal{U}(0,1)$ on $(0,1)$, then $2N[\mathscr{C}_{\mathrm{UL}}(y,A)]^{-1}$ converges in distribution to an exponential random variable with parameter 1.*

This theorem can help us to explain Example 1.4. Indeed, a basic understanding of Theorem 2.4 is that, with high probability and for large N, the condition number (as a function of N) stays roughly constant when A is exponentially distributed, grows like $\sqrt{\ln(N)}/\sigma$ when A is normally distributed, and grows like N when A is uniformly distributed.

Note that – according to Theorem 2.3 – larger values of the condition number are unfavourable for finite precision algorithms. Thus, it is unsurprising that in Figure 1 we see that MATLAB commits more errors for large N when A is uniform than when A is normal and similarly that MATLAB commits more errors for large N when A is normal than when A is exponentially distributed.

2.3. $\mathscr{C}_{\mathrm{UL}}(b,U)$ **and trustworthy algorithms for classical statistics.** Our next set of major results is focused on the following question:

> *In view of Example 1.4, Figure 1 and Theorem 2.3, under which conditions do there exist efficient and trustworthy finite precision algorithms – for the LASSO feature selection problem with random data – that produce correct outputs with high probability?*

Our final main result (Theorem 3.3) gives a bound on the probability of $\mathscr{C}_{\mathrm{UL}}$ being large for normally distributed inputs (b,U) (as a function of the input's size). Its statement involves a number of technical conditions laid out in Section 3 below. To give an idea of its significance, however, we next state a variant of it in a simple, yet frequently considered, context. Consistently with the notation used in Part I, we write $[\![v]\!]_2 := \max\{1, \|v\|_2\}$.

Theorem 2.5. *Let $v \in \mathbb{R}^N$, $S = \mathsf{supp}(v)$, and $s = |S|$. Assume $\ln(N/2) \leq s \leq N/8$ and $m \leq N/9$. Let $U \in \mathbb{R}^{m\times M}$ be random with i.i.d. entries with standard Gaussian distribution and $b = Uv$. There exists a universal constant $\overline{c} \geq 1$ with the following property. Assume that*

(i) $m > (1+\epsilon)12s\ln(N-s)$ *for some $\epsilon \in (0,1/2)$ with $\epsilon > 8\sqrt{s/m}$,*

(ii) $\overline{c}\lambda < 2m \min_{j\in S} |v_j| - \dfrac{1}{N^2}$,

(iii) $\lambda \geq \frac{2}{N^2}$.

Then, the algorithm in Theorem I.1.2, which reads variable-precision approximations of input (b,U), returns $S = \mathsf{supp}(v)$, with a cost bounded by $\mathcal{O}\big(N^3(\log_2 N[\![v]\!]_2)^2\big)$ and maximum number of digits bounded by $\mathcal{O}(\lceil \log_2(N[\![v]\!]_2)\rceil)$ with probability at least $1 - \mathbf{C}_1 N^{-\mathbf{C}_2}$ for some positive universal constants $\mathbf{C}_1$ and $\mathbf{C}_2$.

Remark 2.6. The more general statement in Theorem 3.3 applies under the presence of noise —now $b = Uv + w$ with $w \in \mathbb{R}^m$ random with i.i.d entries drawn from $\mathcal{N}(0,\eta^2)$— however, this requires slightly more involved assumptions and is done in the next section.

3. Precise and general statement of Theorem 2.5

Recall, for a matrix $M \in \mathbb{R}^{m\times N}$, its ∞-norm is given by $\|M\|_\infty := \max_{i\in\{1,2,\ldots,m\}} \sum_{j=1}^N |M_{i,j}|$.

In this section we consider m feature points $a_i \in \mathbb{R}^N$ independently drawn from a normal distribution in $\mathbb{R}^N$. Moreover, we assume that the data y is a random corruption of a fixed linear predictor $v \in \mathbb{R}^N$. More precisely, we consider the setup given by the following assumptions:

(Si) $A \in \mathbb{R}^{m\times N}$ is a random matrix such that each row is an i.i.d. random vector with distribution $\mathcal{N}(0,\Sigma)$ for some covariance matrix Σ.

(Sii) The vector $w \in \mathbb{R}^m$, chosen independently from A, has i.i.d. entries with $\mathcal{N}(0,\eta^2)$ distribution.

(Siii) $y = Av + w$ where $v \in \mathbb{R}^N$ is non-random and has support S with $|S| = s$.

As it happens, we cannot expect a useful probabilistic bound on $\mathscr{C}_{\mathrm{UL}}$ for arbitrary m, N, and Σ even if there is no noise in the measurements (i.e. $\eta^2 = 0$). In order to prove a useful bound on the condition number we will therefore make some assumptions on the covariance matrix Σ, number of measurements m, and vectors v. Since our intention is to understand when the lasso can be successfully applied, it is sensible to use existing conditions that guarantee a low chance of a recovery error. Our goal will be to show that these conditions can also be used to give an upper bound on the condition number and thus a guarantee of a low chance of a numerical error. As a starting point, we therefore use the conditions defined in an important paper in understanding recovery errors for unconstrained lasso, [52]. More precisely, we make use of the following parameters taken from [52]:

(1) For sets $G = \{g_1, g_2, \ldots, g_{|G|}\} \subseteq \{1,2,\ldots,N\}$ and $H = \{h_1, h_2, \ldots, h_{|H|}\} \subseteq \{1,2,\ldots,N\}$ we define the matrix $\Sigma_{GH} \in \mathbb{R}^{|G|\times|H|}$ so that the i,jth entry $(\Sigma_{GH})_{i,j}$ is given by Σ_{g_i,h_j} In other words, Σ_{GH} is the restriction of Σ to the rows given by G and the columns given by H.

(2) Define $C_{\min}$ and $C_{\max}$ to be, respectively, the minimal and maximal eigenvalues of Σ_{SS}.
(3) Define the matrix $\Sigma_{S^c|S} \in \mathbb{R}^{(N-s)\times(N-s)}$ by $\Sigma_{S^c|S} := \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}$.
(4) We define $\gamma := 1 - \|\Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_\infty$.
(5) For a square, symmetric, matrix M we define $\rho_l(M) := \min_{i\neq j}(M_{ii} + M_{jj} - 2M_{ij})/2$ and $\rho_u(M) := \max\{M_{ii}\}$. When convenient, we will write for shorthand $\rho_u = \rho_u(\Sigma_{S^c|S})$ and $\rho_l = \rho_l(\Sigma_{S^c|S})$.
(6) Define $\theta_l = \theta_l(\Sigma) := \rho_l/(C_{\max}(2-\gamma^2))$ and $\theta_u = \theta_u(\Sigma) := \rho_u/(C_{\min}\gamma^2)$,
(7) Define

$$\phi_N := \frac{\lambda^2}{8\eta^2 \ln(N) C_{\min}\theta_u m}, \tag{3.1}$$

Remark 3.1. Although at first glance these parameters —$C_{\min}, C_{\max}, \theta_u, \theta_l$ and γ— seem somewhat complicated, it can be revealing to consider their values when the rows of A are drawn from a standard isotropic Gaussian, that is, when $\Sigma = \mathrm{I}_N$. In this case, for any S, $\Sigma_{SS} = \mathrm{I}_s$. Therefore clearly, $C_{\min} = C_{\max} = 1$. Also, $\Sigma_{S^cS} = 0$ and $\Sigma_{S^cS^c} = \mathrm{I}_{N-s}$ so that $\Sigma_{S^c|S} = \mathrm{I}_{N-s}$ and hence, $\gamma = \rho_l = \rho_u = 1$ and thus $\theta_l = \theta_u = 1$. This means that $\phi_N = \lambda^2(8\eta^2\log(N)m)^{-1}$.

We next consider the following assumption:

(a0): γ is strictly positive.

Our starting point is [52, Thm. 3], which when written in the notation of this paper becomes the following.

Theorem 3.2 (Theorem 3 [52])**.** *Assume that $\phi_N \geq 2$ and that*

$$\frac{m}{2s\ln(N-s)} > (1+\epsilon)\theta_u\left(1 + \frac{4m^2\eta^2 C_{\min}}{\lambda^2 s}\right) \tag{3.2}$$

for some $\epsilon \in (0, 1/2)$ with $\epsilon > \max\{8C_{\min}\sqrt{s/m}, \sqrt{s/m}\}$ as well as assumption (a0).

Then there exist constants c_1, c_2 and c_3 independent of all parameters such that the following is true. If

$$g(\lambda) := \frac{c_3\lambda\|\Sigma^{-1/2}\|_\infty^2}{2m} + 20\sqrt{\frac{\eta^2\ln(s)}{C_{\min}m}} < \min_{j\in S}|v_j|$$

then, with probability greater than $1 - c_1\mathrm{e}^{-c_2\min\{s,\ln(N-s)\}}$, the LASSO problem $\mathsf{Sol}^{\mathsf{UL}}(y, A)$ defined in (1.2) *has a unique solution x. Moreover x satisfies the following:*

(1) $\mathsf{supp}(x) = S$
(2) $\mathsf{sgn}(x_S) = \mathsf{sgn}(v_S) \in \{-1, 1\}^S$
(3) $\|x_S - v_S\|_\infty \leq g(\lambda)$. □

Theorem 3.2 is a very important result in the literature concerning the lasso. Firstly, the result applies both when $\Sigma = \mathrm{I}_N$ and $\eta = 0$ (giving the traditional bound from the compressed sensing literature that $m \geq Cs\log(N-s)$ for some constant C as a sufficient condition for recovering the set S from gaussian measurements) and when $\Sigma \neq \mathrm{I}_N$ to explain the more realistic scenario of trying to distinguish between correlated normally distributed features.

The second and perhaps more crucial reason for the importance of Theorem 3.2 is optimality. Indeed, [52, Thm. 4] also contains a corresponding lower bound: if instead of assuming (3.2), we assume that

$$\frac{m}{2s\ln(N-s)} < (1-\epsilon)\theta_l\left(1 + \frac{4m^2\eta^2 C_{\max}}{\lambda^2 s}\right)$$

then (with probability approaching 1 as N increases) no solution x of the lasso problem has $\mathsf{supp}(x) = S$ and $\mathsf{sgn}(x_S) = \mathsf{sgn}(v_S)$. Thus the hypotheses of Theorem 3.2 can be seen as necessary for the lasso to avoid recovery errors and are therefore a basic requirement for working with the lasso with normally distributed data.

It is therefore worth considering if the conditions in Theorem 3.2 and assumption (a0) are sufficient to also avoid numerical errors. The major result of this section will be to show that this is indeed the case, under slightly stronger assumptions. More precisely, we assume the following:

(ai): The number of measurements m satisfies

$$m\left(\frac{1}{1+\epsilon}-\frac{6}{\phi_N}\right) > 12s\ln(N-s)\theta_u \tag{3.3}$$

for some $\epsilon \in (0,1/2)$ with $\epsilon > \max\{8C_{\min}\sqrt{s/m}, \sqrt{s/m}\}$.

(aii): $g(\lambda) < \min_{s\in S}|v_s|$.

At first glance, it may seem that assumption (ai) is only loosely related to (3.2). However, because of (3.1), we can write (3.2) as

$$\frac{m}{2s\ln(N-s)} > (1+\epsilon)\theta_u\left(1+\frac{m}{2s\ln(N)\theta_u\phi_N}\right)$$

Replacing the constant $1/2$ in the left-hand side by the smaller value $1/12$ and replacing $1/\ln(N)$ by the larger value $1/\ln(N-s)$ in the right-hand side of this bound yields the simpler inequality (3.3). Thus (3.3) can be seen as a slightly stronger condition than (3.2) (but note that (3.3) is still optimal up to the change of constants). Moreover, for the left hand side of (3.3) to be positive we will require $\phi_N \geq 6(1+\epsilon)$ and hence assumption (ai) supersedes the requirement that $\phi_N \geq 2$.

To state our result under assumptions (ai–aii) we will use one additional parameter. Recall from (I.7.1) the function q in the variables $\nu, \xi > 0$,

$$q(\nu,\xi) := 96\nu^5 + 12\nu^3(1+\lambda\sqrt{N})\sqrt{\xi} + \xi\left(\frac{2\nu^3}{\lambda}+3\nu\right). \tag{3.4}$$

We are now in a position to state the major result of this section.

Theorem 3.3. *In the setup described by* (Si–iii) *and under the assumptions* (ai–aii)*, there exists constants* $\mathbf{c}_1, \mathbf{c}_2$ *such that if* $p = \mathbf{c}_1 \mathrm{e}^{-\mathbf{c}_2 \min\{\ln(N-s),s\}}$ *then*

$$\mathbb{P}(\mathscr{C}_{\mathrm{UL}}(y,A) \geq \widehat{K}) \leq p \text{ where } \widehat{K} := (mN)^{\frac{1}{2}} \max\left\{\frac{q(\hat{\alpha},\hat{\sigma})}{\hat{\sigma}^2}, \frac{6\hat{\alpha}}{\sqrt{\hat{\sigma}}}, 1\right\}$$

with

$$\hat{\sigma} = \min\left\{C_{\min}^2/(4(\sqrt{s}+\sqrt{m})^4), \lambda/2, \min_{j\in S}|v_j| - g(\lambda)\right\},$$
$$\hat{\alpha} = \max\left\{1, \sqrt{2m}(\eta + C_{\max}\|v\|_2), \|\Sigma\|_2^{1/2}(3\sqrt{m}+6\sqrt{N})\right\}.$$

In particular, the algorithm in Theorem I.1.2, which reads variable-precision approximations of input (y,A)*, returns* $S = \mathsf{supp}(v)$*, with a cost bounded by* $\mathcal{O}\big(N^3(\log_2 \hat{K}[\![v]\!]_2)^2\big)$ *and maximum number of digits bounded by* $\mathcal{O}(\left\lceil \log_2\left(\hat{K}[\![v]\!]_2\right)\right\rceil)$*. with probability at least* p*.*

As per Remark 3.1, when the rows of A are drawn from a standard isotropic Gaussian assumption (a0) is automatically satisfied whereas (ai–aii) reduce to the following

(ai'): The number of measurements m satisfies

$$m\left(\frac{1}{1+\epsilon}-\frac{6}{\phi_N}\right) > 12s\ln(N-s) \tag{3.5}$$

for some $\epsilon \in (0,1/2)$ with $\epsilon > 8\sqrt{s/m}$.

(aii'): $g(\lambda) < \min_{s\in S}|v_s|$ where $g(\lambda) = c_3\lambda/(2m) + 20\sqrt{\eta^2\ln(s)/m}$ and c_3 is as in Theorem 3.2.

and hence we obtain the following simpler form of Theorem 3.3 which we state in full.

Corollary 3.4. *Assume* $\Sigma = \mathrm{I}_N$. *In the setup above and under assumptions* (ai') *and* (aii')*, there exist absolute constants* $\mathbf{c}_1$ *and* $\mathbf{c}_2$ *such that the following holds true. Let*

$$\hat{\sigma} = \min\left\{\frac{1}{4(\sqrt{s}+\sqrt{m})^4}, \frac{\lambda}{2}, \min_{j\in S}|v_j| - g(\lambda)\right\}$$
$$\text{and} \quad \hat{\alpha} = \max\left\{\sqrt{2m}(\eta + \|v\|_2), 3\sqrt{m} + 6\sqrt{N}\right\}.$$

Then, $\mathbb{P}(\mathscr{C}_{\mathrm{UL}}(y, A) \geq \widehat{K}) \leq \mathbf{c}_1 \mathrm{e}^{-\mathbf{c}_2 \min\{\ln(N-s), s\}}$ *where* $\widehat{K} := (mN)^{\frac{1}{2}} \max\left\{\frac{q(\hat{\alpha},\hat{\sigma})}{\hat{\sigma}^2}, \frac{6\hat{\alpha}}{\sqrt{\hat{\sigma}}}\right\}$. *In particular, the algorithm in Theorem I.1.2, which reads variable-precision approximations of input* (y, A)*, returns* $S = \mathsf{supp}(v)$*, with a cost bounded by* $\mathcal{O}\big(N^3(\log_2 \hat{K}[\![v]\!]_2)^2\big)$ *and maximum number of digits bounded by* $\mathcal{O}(\left\lceil \log_2\left(\hat{K}[\![v]\!]_2\right)\right\rceil)$*. with probability at least* p*.*

4. Connection to previous work

Below follows an account of the connection to different areas and works that are crucial for the paper.

Condition in optimisation: The concept of condition numbers has proven a crucial to computational mathematics and numerical analysis for securing trustworthy algorithms that are accurate and stable [24, 35]. J. Renegar's contributions in optimization and condition are particularly noteworthy, with [45–49] important for understanding stability, accuracy, and efficiency of optimization algorithms. This is extensively discussed in [15]. Furthermore, the following have important links to condition in optimization: J. Peña [42, 43] as well as D. Amelunxen, M. Lotz, J. Walvin [37], and D. Amelunxen, M. Lotz, M. McCoy, J. Tropp [3] see also [17–19].

GHA and robust optimisation: GHA [2, 4, 8, 23, 27] plays an instrumental role in establishing some of the computational barriers that this paper introduces. The topic is related (although mathematically very different) to hardness of approximation in computer science [6]. Importantly, GHA in optimisation can be viewed as a part of the broader program on robust optimisation (A. Ben-Tal, L. El Ghaoui & Nemirovski [12, 13, 40]) for computing minimisers. It is also a part of broader efforts to establishthe mathematics behind the Solvability Complexity Index (SCI) hierarchy, see for example the work by J. Ben-Artzi, M. Colbrook, M. Marletta [10, 11, 22, 32].

Trustworthy algorithms and computer assisted proofs: The importance of finding trustworthy algorithms for optimisation goes beyond scientific computing: it has important implications in computer assisted proofs. T. Hales' proof of Kepler's conjecture [29, 30] is a prime example. Hales' computer assisted proof of this famous conjecture relied on solving around 50,000 linear programs with irrational inputs, thus highlighting the importance of understanding computation with inexact inputs across all of mathematics. For other examples, see [26]. Of particular interest to this work are Problem 2 (T. Hou) and Problem 5 (J. Lagarias) which discusses results on developing algorithms that are 100% trustworthy and thus appropriate for computer assisted proofs.

Algorithms for computing minimisers of LASSO: Numerous algorithms are available for solving the LASSO problem, as detailed in the review articles by Nesterov & Nemirovski [41] and Chambolle & Pock [16], which include additional references for a comprehensive understanding. See also the work by Beck & Teboulle [9] and Wright, Nowak & Figueiredo [55], and the references therein, as well as [2, 14, 54]. However, while these algorithms are capable of approximating the objective function, they cannot – in general – determine the support sets of minimisers of LASSO.

5. Definitions and results from [7]

In this section we recall some definitions and results that are taken from [7]. These are provided without proofs, as the proofs are contained in [7], and are included for the sake of ensuring that this paper can be read as self contained material.

In addition to $[\![(y,A)]\!]_{\max}$ we will also use the ℓ^p-norm $\|y\|_p$ of y, the operator norms $\|A\|_{qr} = \sup_{\|x\|_q=1} \|Ax\|_r$ (writing $\|A\|_q$ when $q=r$), and the truncated norms

$$[\![(y,A)]\!]_{\mathrm{S}} = \max\left\{\sum_{i=1}^{m}\sum_{j=1}^{N}|A_{ij}|,\ \sum_{i=1}^{m}|y_i|,\ 1\right\}$$

and $[\![(y,A)]\!]_2 := \max\{\|A\|_2, \|y\|_2, 1\}$.

Definition 5.1. The *stability support* of a pair (y,A) is defined as

$$\begin{aligned}\mathsf{stsp}(y,A) := \inf\Big\{\delta \geq 0 \,\Big|\, \exists\, \tilde{y}\in\mathbb{R}^m, \tilde{A}\in\mathbb{R}^{m\times N}, x\in \mathsf{Sol}^{\mathsf{UL}}(y,A),\ \text{and } \tilde{x}\in\mathsf{Sol}^{\mathsf{UL}}(\tilde{y},\tilde{A})\\ \text{such that } \|\tilde{y}-y\|_\infty, \|A-\tilde{A}\|_{\max}\leq\delta \text{ and } \mathsf{supp}(x)\neq\mathsf{supp}(\tilde{x})\Big\}.\end{aligned}$$

The stability support is therefore the *distance to support change*. If $\mathsf{stsp}(y,A) > 0$ then there exists $S\in\mathbb{B}^N$ such that $\Xi(y,A) = \{S\}$. Furthermore, for all pairs (y',A') in a ball (w.r.t. the max distance) of radius $\mathsf{stsp}(y,A)$ around (y,A) we have $\Xi(y',A') = \{S\}$. If, instead, $\mathsf{stsp}(y,A) = 0$ then there are arbitrarily small perturbations of (y,A) which yield LASSO solutions with different support. We use stability support to define the condition:

Definition 5.2. For an input (y,A) to UL feature selection we define the *condition number* $\mathscr{C}_{\mathrm{UL}}(y,A)$ to be

$$\mathscr{C}_{\mathrm{UL}}(y,A) = \begin{cases}(\mathsf{stsp}(y,A))^{-1} & \text{if } \mathsf{stsp}(y,A)\neq 0\\ \infty & \text{otherwise.}\end{cases} \tag{5.1}$$

The set $\Sigma_{\mathrm{UL}} := \{(y,A) \mid \mathsf{stsp}(y,A) = 0\}$ is the *set of ill-posed inputs*.

To make this easier to work with, we define the following:

Definition 5.3. For a pair (y,A), we write (with the convention that if M is non-invertible, $\|M^{-1}\|_2 := \infty$ and so $\|M^{-1}\|_2^{-1} = 0$),

$$\begin{aligned}\sigma_1(y,A) &:= \inf\{t \mid \exists x\in\mathsf{Sol}^{\mathsf{UL}}(y,A) \text{ with } \|A^*_{S^{\mathrm{c}}}(Ax-y)\|_\infty = \lambda/2 - t, S = \mathsf{supp}(x)\},\\ \sigma_2(y,A) &:= \inf\{\|(A_S^*A_S)^{-1}\|_2^{-1} \mid \exists x\in\mathsf{Sol}^{\mathsf{UL}}(y,A) \text{ with } S=\mathsf{supp}(x)\},\\ \sigma_3(y,A) &:= \inf\{t \mid \exists i\in\{1,2,\dots,N\} \text{ and } x\in\mathsf{Sol}^{\mathsf{UL}}(y,A) \text{ such that } 0<|x_i|\leq t\}.\end{aligned}$$

where, for the empty-set $\varnothing$, we interpret $\|A^*_\varnothing(Ax-y)\|_\infty = 0$, we treat $A^*_\varnothing A_\varnothing$ as invertible with $\|(A^*_\varnothing A_\varnothing)^{-1}\|_2^{-1} = \infty$, and we set $\inf\varnothing = \infty$.

We combine each of σ_1,σ_2 and σ_3 into a single quantity as follows,

$$\sigma(y,A) := \min\left\{\sigma_1(y,A), \sigma_2(y,A)^2, \sigma_3(y,A)\right\}$$

The next proposition provides a lower bound for stsp which makes use of the polynomial (3.4):

Proposition 5.4. *Set* $\alpha = [\![(y,A)]\!]_2$ *and* $\sigma = \sigma(y,A)$. *Then* $\mathsf{stsp}(y,A) \geq (mN)^{-\frac{1}{2}}\min\left\{\frac{\sigma^2}{q(\alpha,\sigma)}, \frac{\sqrt{\sigma}}{6\alpha}, \alpha\right\}$.

6. Proof of Theorem 2.4

To prove Theorem 2.4, we first relate the condition number to some quantities that will be easier to deal with. Specifically, we define the event Z and the quantity δ as follows.

Definition 6.1. For pairs $(y,A)\in\mathbb{R}^{1+N}$ and $\epsilon>0$, we define $Z(\epsilon) = Z(y,A,\epsilon)$ to be the following event: if $\|(y,A)-(\tilde{y},\tilde{A})\|_{\max}\leq\epsilon$, then $0\notin\mathsf{Sol}^{\mathsf{UL}}(\tilde{y},\tilde{A})$.

Definition 6.2. For a vector $A\in\mathbb{R}^N$ we let $\delta\geq 0$ denote the difference between the largest entry of $|A|$ and the second largest entry of $|A|$, where $|A|\in\mathbb{R}^N$ is the vector with entries $|A_i|$ for $i=1,\dots,N$.

The proof of Theorem 2.4 easily follows from the following four lemmas, the second of which relates the condition number to Z and δ.

Lemma 6.3. *Let $A \in \mathbb{R}^N$ and $i \in \{1, 2, \dots, N\}$ be such that $|A_i| > |A_j|$ for all $j \in \{1, 2, \dots, N\}$ with $j \neq i$. Let $y \in \mathbb{R}$ and $x \in \mathsf{Sol}^{\mathsf{UL}}(y, A)$. Then $\mathsf{supp}(x) = \{i\}$ if $|A_i y| > \lambda/2$ and $\mathsf{supp}(x) = \varnothing$ (i.e. $x = 0$) if $|A_i y| \leq \lambda/2$.*

Lemma 6.4. *Let $A \in \mathbb{R}^N$ be randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^N$. Let $y \in \mathbb{R}$ and $\epsilon > 0$. Then, $\mathbb{P}[Z(\epsilon) \cap (\mathsf{stsp} < \epsilon)] = \mathbb{P}[Z(\epsilon) \cap (\delta < 2\epsilon)]$.*

Thus to understand stsp it suffices to analyse Z and δ. This is what we do in the next two lemmas under the assumptions that A is exponentially, gaussian, or uniformly distributed.

Lemma 6.5. *Let F be the cumulative distribution of a non-negative random variable X which is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}$ and $c > 0$ be such that $F(c) < 1$. Suppose that $A^{(N)} \in \mathbb{R}^N$ is a random vector with i.i.d. entries such that $\left|A_i^{(N)}\right|$ is distributed as X. Let $y \in \mathbb{R}$ and $(b_N)_{N=1}^{\infty}$ be a sequence of non-negative reals satisfying that $b_N < |y|$ and $\lambda/(2|y| - 2b_N) + b_N \leq c$ for all N sufficiently large. Then $\lim_{N\to\infty} \mathbb{P}(Z(y, A^{(N)}, b_N)) = 1$.*

In particular, for all $y \in \mathbb{R}$, and $b > 0$ and all random $A \in \mathbb{R}^N$,

1. *$\lim_{N\to\infty} \mathbb{P}(Z(y, A, b)) = 1$ for $b < |y|$ and $\lim_{N\to\infty} \mathbb{P}(Z(y, A, b)) = 0$ for $b \geq |y|$, if each entry of A is exponentially distributed with parameter 1.*
2. *$\lim_{N\to\infty} \mathbb{P}(Z(y, A, b\sigma/\sqrt{2\ln(N)})) = 1$, if $y \neq 0$ and each entry of A is Gaussian with mean 1, variance σ^2.*
3. *$\lim_{N\to\infty} \mathbb{P}(Z(y, A, b/N)) = 1$, if $|y| > \lambda/2$ and each entry of A is uniformly distributed on $[0, 1]$.*

Lemma 6.6. *Suppose that $A \in \mathbb{R}^N$ is a random vector with i.i.d. entries. Then for each $t > 0$*

1. *$\lim_{N\to\infty} \mathbb{P}(\delta > t) = \mathrm{e}^{-t}$ if each entry of A is exponentially distributed.*
2. *$\lim_{N\to\infty} \mathbb{P}(\delta > t\sigma/\sqrt{2\ln(N)}) = \mathrm{e}^{-t}$ if each entry of A is Gaussian.*
3. *$\lim_{N\to\infty} \mathbb{P}(\delta > t/N) = \mathrm{e}^{-t}$ if each entry of A is uniformly distributed on $[0, 1]$.*

Assuming (for now) that these Lemmas hold, we proceed to prove Theorem 2.4.

Proof of Theorem 2.4. For any sequence of positive real valued functions $(f_n)_{n=1}^{\infty}$, we split $\mathbb{P}(\mathsf{stsp} < f_N(t)) = \mathbb{P}[(\mathsf{stsp} < f_N(t)) \cap Z(f_N(t))] + \mathbb{P}[(\mathsf{stsp} < f_N(t)) \cap Z(f_N(t))^{\mathsf{c}}]$. Using Lemma 6.4 with $\epsilon = f_N(t)$ on the first term this splitting becomes

$$\mathbb{P}(\mathsf{stsp} < f_N(t)) = \mathbb{P}[(\delta < 2f_N(t)) \cap Z(f_N(t))] + \mathbb{P}[(\mathsf{stsp} < f_N(t)) \cap Z(f_N(t))^{\mathsf{c}}] \tag{6.1}$$

Now take $f_N(t) = t$ when the entries of A are exponentially distributed and $t < |y|$, $f_N(t) = \frac{t\sigma}{\sqrt{2\ln N}}$ if they are Gaussian, and $f_N(t) = \frac{t}{N}$ if they are uniformly distributed on $[0, 1]$. In the three cases Lemma 6.5 shows that $\mathbb{P}(Z(f_N(t))) \to 1$ when $N \to \infty$. Consequently, in all three cases we have

$$\lim_{N\to\infty} \mathbb{P}(\mathsf{stsp} < f_N(t)) = \lim_{N\to\infty} \mathbb{P}[(\delta < 2f_N(t)) \cap Z(f_N(t))] = \lim_{N\to\infty} \mathbb{P}[\delta < 2f_N(t)].$$

We can then apply Lemma 6.6 to deduce that

- $\lim_{N\to\infty} \mathbb{P}(\mathsf{stsp} < t) = 1 - \mathrm{e}^{-2t}$ for $t < |y|$ if each entry of A is exponentially distributed with parameter 1.
- $\lim_{N\to\infty} \mathbb{P}(\mathsf{stsp} < t/\sqrt{2\ln(N)}) = 1 - \mathrm{e}^{-2t/\sigma}$ if each entry of A is a standard Gaussian.
- $\lim_{N\to\infty} \mathbb{P}(\mathsf{stsp} < t/N) = 1 - \mathrm{e}^{-2t}$ if $|y| > \lambda/2$ and each entry of A is uniformly distributed on $[0, 1]$.

It only remains to deal with $\mathbb{P}(\mathsf{stsp} < t)$ when $t \geq |y|$ and each entry of A is exponentially distributed with parameter 1. We start by examining $\mathbb{P}[(\mathsf{stsp} \geq t) \cap Z(t)^{\mathsf{c}}]$ in this setting.

Assume that $Z(t)^{\mathsf{c}}$ occurs. Then $0 \in \mathsf{Sol}^{\mathsf{UL}}(\tilde{y}, \tilde{A})$ for some perturbation $(\tilde{y}, \tilde{A})$ of (y, A) satisfying that $d[(\tilde{y}, \tilde{A}), (y, A)]_\infty < t$. If, in addition, $\mathsf{stsp} \geq t$ then $0 \in \mathsf{Sol}^{\mathsf{UL}}(y, A)$ as well. Now, Lemma 6.3 and the fact that $|y| > 0$ show that this occurs if and only $\max_{i=1,2,\ldots,N} |A_i| \leq \lambda/(2|y|)$. Since the entries of A are independent, we conclude that

$$\mathbb{P}(0 \in \mathsf{Sol}^{\mathsf{UL}}(y, A)) = \Pi_{i=1}^{N} \mathbb{P}(|A_i| \leq \lambda/(2|y|)) = [1 - e{-}\lambda/(2y)]^N$$

a quantity tending to 0 as $N \to \infty$. Thus $\lim_{N\to\infty} \mathbb{P}[(\mathsf{stsp} \geq t) \cap Z(t)^{\mathsf{c}}] \leq \lim_{N\to\infty} \mathbb{P}(0 \in \mathsf{Sol}^{\mathsf{UL}}(y, A)) = 0$.

This last limit implies that $\lim_{N\to\infty} \mathbb{P}[(\mathsf{stsp} < t) \cap Z(t)^{\mathsf{c}}] = \lim_{N\to\infty} \mathbb{P}(Z(t)^{\mathsf{c}})$. As $t \geq |y|$, Lemma 6.5 part (1) shows that $\lim_{N\to\infty} \mathbb{P}(Z(t)^{\mathsf{c}}) = 1$. Hence $\lim_{N\to\infty} \mathbb{P}(\mathsf{stsp} < t) \geq \lim_{N\to\infty} \mathbb{P}[(\mathsf{stsp} < t) \cap Z(t)^{\mathsf{c}}] = \lim_{N\to\infty} \mathbb{P}(Z(t)^{\mathsf{c}}) = 1$, completing the proof. □

We now prove Lemmas 6.3 to 6.6.

Proof of Lemma 6.3. The proof follows from the KKT conditions (UL5) in Section I.4. Suppose firstly that $|A_i y| > \lambda/2$. Then 0 is not a solution to the lasso problem. Indeed, if it were the KKT conditions would imply that $|A_i y| = |(A^*(A0 - y))_i| = \lambda/2$, contradicting our assumption. Moreover for all $j \in \mathsf{supp}(x)$ we have that j is in the equicorrelation set and thus $|A_j||(Ax - y)| = |A_j^*(Ax - y)| = \lambda/2$. This means that the value of $|A_j|$ is the same for all $j \in \mathsf{supp}(x)$. Our hypothesis then imply that $\mathsf{supp}(x) = \{i\}$.

Suppose now that, instead, $|A_i y| \leq \lambda/2$. Then $\|A^*(A0 - y)\|_\infty = |A_i^*(A0 - y)| \leq \lambda/2$ and thus 0 is a solution as it satisfies the KKT conditions. We claim that 0 is in fact the only solution. Indeed, by (UL4), all lasso solutions have the same ℓ^1 norm and hence all solutions have norm 0. Thus the only possible solution is 0. □

Proof of Lemma 6.4. Write, for simplicity, $\xi := \mathsf{stsp}(y, A)$. It suffices to show that

$$\mathbb{P}[(\xi < \epsilon) \cap Z(\epsilon)] \leq \mathbb{P}[(\delta < 2\epsilon) \cap Z(\epsilon)] \text{ and } \mathbb{P}[(\xi \geq \epsilon) \cap Z(\epsilon)] \leq \mathbb{P}[(\delta \geq 2\epsilon) \cap Z(\epsilon)]. \tag{6.2}$$

As the distribution for the entries of A is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^N$, we can assume that each entry of A is unique (this is true with probability 1). Assume that $Z(\epsilon)$ occurs and let $x \in \mathsf{Sol}^{\mathsf{UL}}(y, A)$. Because $Z(\epsilon)$ holds, $x \neq 0$. This implies the existence of $i \leq N$ with $\mathsf{supp}(x) = \{i\}$ and such that $|A_i| > |A_k|$ for all $k \neq i$ by Lemma 6.3.

Let us now prove the first of the inequalities in (6.2). For this, assume that, in addition to $Z(\epsilon)$, we have $\xi < \epsilon$. Then there exists $(\tilde{y}, \tilde{A})$ with $d_{\max}[(\tilde{y}, \tilde{A}), (y, A)] < \epsilon$ such that there is an $\tilde{x} \in \mathsf{Sol}^{\mathsf{UL}}(\tilde{y}, \tilde{A})$ with $\mathsf{supp}(\tilde{x}) \neq \mathsf{supp}(x)$. Under the assumption that $Z(\epsilon)$ occurs, $\mathsf{supp}(\tilde{x}) \neq \varnothing$ and so there must exist a j such that $|\tilde{A}_j| \geq |\tilde{A}_i|$ (otherwise Lemma 6.3 applied to $(\tilde{y}, \tilde{A})$ implies that $\mathsf{supp}(\tilde{x}) = \{i\}$, contradicting the fact that $\mathsf{supp}(\tilde{x}) \neq \mathsf{supp}(x)$). Since $\tilde{A}$ is an ϵ-perturbation of A, the condition $|\tilde{A}_j| \geq |\tilde{A}_i|$ implies $|A_j| \geq |A_i| - 2\epsilon$. In particular, $\delta \leq 2\epsilon$, and we conclude that $\mathbb{P}(\xi < \epsilon \cap Z(\epsilon)) \leq \mathbb{P}(\delta < 2\epsilon \cap Z(\epsilon))$.

We proceed with the second inequality. For this, assume that, in addition to $Z(\epsilon)$, we have $\xi \geq \epsilon$. Then for any $j \neq i$, the perturbation $\tilde{A}$ defined by $\tilde{A}_i = A_i - \epsilon\,\mathsf{sgn}(A_i)$, $\tilde{A}_j = A_j + \epsilon\,\mathsf{sgn}(A_j)$ and $\tilde{A}_k = A_k$ whenever both $k \neq i$ and $k \neq j$ must be such that if $\tilde{x} \in \mathsf{Sol}^{\mathsf{UL}}(y, \tilde{A})$ then $\mathsf{supp}(\tilde{x}) = \{i\}$ by the definition of $\xi = \mathsf{stsp}(y, A)$. But for this to occur we must have $|\tilde{A}_i| \geq |\tilde{A}_j|$, otherwise Lemma 6.3 applies to yield either $\tilde{x} = 0$ or $\mathsf{supp}(\tilde{x}) = \{j\}$. The condition $|\tilde{A}_i| \geq |\tilde{A}_j|$ reduces to $|A_i| - \epsilon \geq |A_j| + \epsilon$. Since this occurs for any $j \neq i$, we must have $\delta \geq 2\epsilon$. We conclude that if both $Z(\epsilon)$ and $\xi \geq \epsilon$ then we must have $\delta \geq 2\epsilon$ and thus $\mathbb{P}[(\xi \geq \epsilon) \cap Z(\epsilon)] \leq \mathbb{P}[(\delta \geq 2\epsilon) \cap Z(\epsilon)]$. □

Proof of Lemma 6.5. Assume $Z\big(y, A^{(N)}, b_N\big)$ does not hold. Then there exists $(\tilde{y}, \tilde{A})$ in the ball of radius b_N (w.r.t. $d_{\max}$) about $\big(y, A^{(N)}\big)$ such that $0 \in \mathsf{Sol}^{\mathsf{UL}}\big(\tilde{y}, \tilde{A}\big)$. This implies that

$$\big\|A^{(N)}\big\|_\infty < \|\tilde{A}\|_\infty + b_N \qquad \text{and} \qquad |\tilde{y}| > |y| - b_N \tag{6.3}$$

and that (by Lemma 6.3) $\|\tilde{A}\|_\infty |\tilde{y}| \leq \lambda/2$. Together with (6.3) this inequality implies that

$$\left\|A^{(N)}\right\|_\infty < \|\tilde{A}\|_\infty + b_N \leq \frac{\lambda}{2|\tilde{y}|} + b_N < \frac{\lambda}{2(|y| - b_N)} + b_N.$$

Consequently, for sufficiently large N,

$$\begin{aligned}\mathbb{P}\big(Z\big(y, A^{(N)}, b_N\big)\big) &\geq \mathbb{P}\left(\left\|A^{(N)}\right\|_\infty > \frac{\lambda}{2|y| + 2b_N} - b_N\right) \\ &= 1 - F\left(\frac{\lambda}{2|y| + 2b_N} - b_N\right)^N \geq 1 - F(c)^N\end{aligned}$$

the equality because the entries of $A^{(N)}$ are i.i.d. and our hypothsesis on their distribution, and the last inequality by our hypothesis. We conclude (since we assume that $F(c) < 1$) that

$$\lim_{N\to\infty} \mathbb{P}\big(Z\big(y, A^{(N)}, b_N\big)\big) = 1.$$

We use this result to easily prove each of (1), (2) and (3). Starting with (1), suppose first that $b < |y|$. The result follows from setting $b_N = b$ for all $N \geq 0$ and $c = \lambda/(2|y| - 2b) + b + 1$. In the case $b \geq |y|$ consider the perturbation $\tilde{y} = 0$, $\tilde{A} = A$. Then $|y - \tilde{y}| = |y| \leq b$ and, obviously, $\|\tilde{A} - A\|_\infty = 0 \leq b$. But 0 is the unique solution to lasso for the pair $(\tilde{y}, \tilde{A})$, so $Z(y, A, b)$ does not hold and, consequently, $\mathbb{P}(Z(y, A, b)) = 0$.

Similarly, in the Gaussian case (2) we set $c := \frac{\lambda}{|y|}$ and $b_N := \frac{b\sigma}{\sqrt{2\ln N}}$. Clearly, $c < \infty$ so that $F(c) < 1$. In addition, for N sufficiently large, $b_N < |y|$ and $\lambda/(2|y| - 2b_N) + b_N < c$, both because $b_N \to 0$ when $N \to \infty$. The desired convergence follows.

Finally, in the uniform case (3) we set $c = 1/2 + \lambda/(4|y|)$ (i.e. c is the midpoint between $\lambda/(2|y|)$ and 1). Note that since $c < 1$, we must have $F(c) < 1$. Furthermore, since $1/N \to 0$ as $N \to \infty$ we must eventually have $\lambda(2|y| - 2b/N)^{-1} + b/N < c$. The result follows. □

Proof of Lemma 6.6. If the entries of $|A|$ are i.i.d. with distribution with density function f and cumulative distribution F, the formula for order statistics (see [25, Equation (2.3.1)] which we use with $s = n$ and $r = n - 1$) gives

$$\mathbb{P}(\delta < t) = N(N-1)\int_0^\infty (F(u))^{N-2} f(u)(F(u+t) - F(u))\,\mathrm{d}u. \tag{6.4}$$

We will use this formula to study the case where A has exponential, uniform and Gaussian entries respectively. Starting with the exponential distribution, we observe that $A = |A|$ in this case and therefore we have $F(u) = 1 - \mathrm{e}^{-u}$ and $f(u) = \mathrm{e}^{-u}$. Integrating (6.4) by parts yields

$$\mathbb{P}(\delta < t) = N \lim_{u\to\infty} (F(u))^{N-1}[F(u+t) - F(u)] - N\int_0^\infty (F(u))^{N-1}[f(u+t) - f(u)]\,\mathrm{d}u$$

and since $F(u+t) - F(u) = \mathrm{e}^{-u}(1 - \mathrm{e}^{-t})$, which tends to 0 when $u \to \infty$, and $f(u+t) - f(u) = -\mathrm{e}^{-u}(1 - \mathrm{e}^{-t}) = -f(u)(1 - \mathrm{e}^{-t})$ we obtain

$$\begin{aligned}\mathbb{P}(\delta < t) &= N(1 - \mathrm{e}^{-t})\int_0^\infty (F(u))^{N-1} f(u)\,\mathrm{d}u \\ &= (1 - \mathrm{e}^{-t}) \lim_{u\to\infty} [(F(u))^N - (F(0))^N] = 1 - \mathrm{e}^{-t}.\end{aligned}$$

For the uniform distribution, assume that N is large enough so that $t/N < 1$. Integrating by parts (6.4) again we obtain

$$\begin{aligned}\mathbb{P}(\delta < t/N) =& N \lim_{u\uparrow 1} F(u)^{N-1}[F(u + t/N) - F(u)] \\ &- N\int_0^1 F(u)^{N-1}[f(u + t/N) - f(u)]\,\mathrm{d}u \\ =& - N\int_0^1 F(u)^{N-1}[f(u + t/N) - f(u)]\,\mathrm{d}u = N\int_{1-t/N}^1 F(u)^{N-1} f(u)\,\mathrm{d}u\end{aligned}$$

$$=1-(F(1-t/N))^N$$

where the second to last equality follows because $f(u+t/N)=f(u)$ provided u is in $(0,1-t/N)$ and for u larger than $1-t/N$ we have $f(u+t/N)=0$. Furthermore, for the uniform distribution $F(1-t/N)=1-t/N$ and thus $\lim_{N\to\infty}F(1-t/N)=\lim_{N\to\infty}(1-t/N)^N=\mathrm{e}^{-t}$.

The process for the Gaussian is more involved. Without loss of generality, we can assume that $\sigma=1$: other σ reduce to this case by considering δ/σ. Let us calculate $\lim\limits_{N\to\infty}\mathbb{P}(\delta\sqrt{2\ln(N)}<t)$. From (6.4), we must have

$$\mathbb{P}(\delta\sqrt{2\ln(N)}<t)=N(N-1)\int_0^\infty (F(u))^{N-2}f(u)\int_u^{u+\frac{t}{\sqrt{2\ln(N)}}} f(v)\,\mathrm{d}v\,\mathrm{d}u \tag{6.5}$$

where $f(u)=\sqrt{\frac{2}{\pi}}e^{-\frac{\left(u^2+\mu^2\right)}{2}}\cosh(\mu u)$ for $u\geq 0$ is the density function for the absolute value of a normal random variable. By the mean value theorem, for each u we have

$$\int_u^{u+\frac{t}{\sqrt{2\ln(N)}}} f(v)\,\mathrm{d}v=\frac{tf(c_u)}{\sqrt{2\ln(N)}}$$

for some $c_u\in(u,u+\frac{t}{\sqrt{2\ln(N)}})$. Hence, integrating by parts in (6.5) and noting that $\lim_{u\to\infty}f(c_u)=0$ yields

$$\begin{aligned}\mathbb{P}(\delta\sqrt{2\ln(N)}<t)&=N\left[\lim_{u\to\infty}(F(u))^{N-1}f(c_u)-(F(0))^{N-1}f(c_0)\right]\\&\quad-N\int_0^\infty(F(u))^{N-1}\left(f\left(u+\frac{t}{\sqrt{2\ln(N)}}\right)-f(u)\right)\mathrm{d}u\\&=-N\int_0^\infty(F(u))^{N-1}\left(f\left(u+\frac{t}{\sqrt{2\ln(N)}}\right)-f(u)\right)\mathrm{d}u.\end{aligned}$$

Because of the form of f and using $\cosh((u+t\alpha)\mu)=\cosh(u\mu)\cosh(t\alpha\mu)+\sinh(u\mu)\sinh(t\alpha\mu)$ we can write $f\left(u+\frac{t}{\sqrt{2\ln(N)}}\right)$ as

$$\begin{aligned}&\sqrt{\frac{2}{\pi}}\mathrm{e}^{-(u^2+\mu^2)/2}\mathrm{e}^{\frac{-tu}{\sqrt{2\ln(N)}}}\mathrm{e}^{\frac{-t^2}{4\ln(N)}}\cosh(\mu u)\left(\cosh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)+\tanh(\mu u)\sinh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)\right)\\&=f(u)\mathrm{e}^{\frac{-tu}{\sqrt{2\ln(N)}}}\mathrm{e}^{\frac{-t^2}{4\ln(N)}}\left(\cosh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)+\tanh(\mu u)\sinh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)\right)=f(u)g(u)\end{aligned}$$

where

$$g(u):=\mathrm{e}^{\frac{-tu}{\sqrt{2\ln(N)}}}\mathrm{e}^{\frac{-t^2}{4\ln(N)}}\left(\cosh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)+\tanh(\mu u)\sinh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)\right)$$

and thus

$$g'(u)=\frac{-tg(u)}{\sqrt{2\ln(N)}}+\mathrm{e}^{\frac{-tu}{\sqrt{2\ln(N)}}}\mathrm{e}^{\frac{-t^2}{4\ln(N)}}\left(\mu\,\mathrm{sech}^2(\mu u)\sinh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)\right)=\frac{-tg(u)}{\sqrt{2\ln(N)}}+c_Nh(u)$$

with

$$c_N=\mathrm{e}^{\frac{-t^2}{4\ln(N)}}\left(\mu\sinh\left(\frac{t\mu}{\sqrt{2\ln(N)}}\right)\right)\quad h(u)=\mathrm{e}^{\frac{-tu}{\sqrt{2\ln(N)}}}\,\mathrm{sech}^2(\mu u)$$

Therefore,

$$\begin{aligned}\mathbb{P}(\delta\sqrt{2\ln(N)}<t)&=-N\int_0^\infty(F(u))^{N-1}f(u)\,(g(u)-1)\,\mathrm{d}u\\&=\lim_{u\to\infty}-(F(u))^N\,(g(u)-1)\\&\quad+\int_0^\infty\frac{-tg(u)}{\sqrt{2\ln(N)}}(F(u))^N\,\mathrm{d}u+c_N\int_0^\infty h(u)(F(u))^N\,\mathrm{d}u\end{aligned}$$

where the last equality holds by a further integration by parts. Note that as $u \to \infty$, $F(u) \to 1$ and $g(u) \to 0$. Thus

$$\mathbb{P}(\delta\sqrt{2\ln(N)} < t) = 1 + \int_0^\infty \frac{-tg(u)}{\sqrt{2\ln(N)}}(F(u))^N \,\mathrm{d}u + c_N \int_0^\infty h(u)(F(u))^N \,\mathrm{d}u$$

For the second of these two integrals, note $c_N \to 0$ as $N \to \infty$ and so

$$\left| c_N \int_0^\infty h(u)(F(u))^N \,\mathrm{d}u \right| \leq c_N \int_0^\infty \operatorname{sech}^2(\mu u)\,\mathrm{d}u = c_N/\mu \to 0.$$

For the first integral $\int_0^\infty \frac{tg(u)}{\sqrt{2\ln(N)}}(F(u))^N \,\mathrm{d}u = I_1 + I_2$ where (using the change of variables $v = tu/\sqrt{2\ln(N)}$)

$$I_1 := \mathbf{C}_N \int_0^\infty \mathrm{e}^{-v} \left(F\left(\frac{v\sqrt{2\ln(N)}}{t} \right) \right)^N \mathrm{d}v$$

$$I_2 := \mathbf{S}_N \int_0^\infty \mathrm{e}^{-v} \tanh\left(\frac{\mu v\sqrt{2\ln(N)}}{t} \right) \left(F\left(\frac{v\sqrt{2\ln(N)}}{t} \right) \right)^N \mathrm{d}v$$

$$\mathbf{C}_N := \mathrm{e}^{\frac{-t^2}{4\ln(N)}} \left(\cosh\left(\frac{t\mu}{\sqrt{2\ln(N)}} \right) \right), \quad \mathbf{S}_N := \mathrm{e}^{\frac{-t^2}{4\ln(N)}} \left(\sinh\left(\frac{t\mu}{\sqrt{2\ln(N)}} \right) \right)$$

Now we note that since F is the cumulative density function of the absolute value of the normal distribution we have $F(x) = [\operatorname{erf}((x+\mu)/\sqrt{2}) + \operatorname{erf}((x-\mu)/\sqrt{2})]/2$. Thus we can use (e.g. [1, Inequality 7.1.13]) to obtain, for any $x > \mu$, that $F(x) \in [1 - h(x, 8/\pi), 1 - h(x, 4)]$, with

$$h(x,y) = \sqrt{\frac{2}{\pi}} \left[\frac{\mathrm{e}^{-(\frac{x+\mu}{2})^2}}{x+\mu+\sqrt{(x+\mu)^2+y}} + \frac{\mathrm{e}^{-(\frac{x-\mu}{2})^2}}{x-\mu+\sqrt{(x-\mu)^2+y}} \right]$$

In particular, for $x > 0$ large enough and y fixed it is easy to see that there exist positive constants C_1,C_2 so that $C_2\mathrm{e}^{-\frac{x^2}{2}}\mathrm{e}^{x|\mu|}/x \geq h(x,y) \geq C_1\mathrm{e}^{-\frac{x^2}{2}}/x$ and therefore when N is sufficiently large

$$\left(1 - \frac{C_2 t N^{-v^2/t^2} \mathrm{e}^{\frac{v\sqrt{2\ln(N)}|\mu|}{t}}}{v\sqrt{2\ln(N)}} \right)^N \leq \left(F(v\sqrt{2\ln(N)}/t) \right)^N \leq \left(1 - \frac{C_1 t N^{-v^2/t^2}}{v\sqrt{2\ln(N)}} \right)^N.$$

Hence, pointwise in v,

$$\lim_{N\to\infty} \left(F(v\sqrt{2\ln(N)}/t) \right)^N = \begin{cases} 1 & \text{if } v^2/t^2 > 1 \\ 0 & \text{if } v^2/t^2 < 1. \end{cases}$$

Thus the integrands of I_1 and I_2 defined for non-negative v converge pointwise as $N \to \infty$ to the function $g(v) = \mathrm{e}^{-v}$ for $v > t$ and $g(v) = 0$ otherwise. Furthermore the integrands of I_1 and I_2 are dominated (uniformly in N) by the integrable function e^{-v} Thus we can apply the dominated convergence theorem to see that I_1/C_N, I_2/S_N both tend to e^{-t} as $N \to \infty$. Since $C_N \to 1$ and $S_N \to 0$ as $N \to \infty$, we conclude that

$$\lim_{N\to\infty} \mathbb{P}(\delta\sqrt{2\ln(N)} < t) = 1 - \mathrm{e}^{-t}.$$

□

7. Proof of Theorem 3.3

To prove Theorem 3.3 our strategy is to use Proposition 5.4 to get a bound on $\mathsf{stsp}(y, A)$ that depends on the random quantities $\alpha := [\![(y, A)]\!]_2$ and $\sigma := \sigma(y, A)$. The following two lemmas yield probabilistic bounds on α and σ for the randomly generated data (y, A) according to the setup described by (Si–iii). We assume throughout the next two statements that (a0)-(aii) are satisfied.

Lemma 7.1. *Set* $\hat{\alpha} = \max\left\{1, \frac{5}{2}\sqrt{m}(\eta + C_{\max}\|v\|_2), \|\Sigma\|_2^{1/2}(3\sqrt{m} + 6\sqrt{N})\right\}$. *Then* $\mathbb{P}(\alpha \geq \hat{\alpha}) \leq 3\mathrm{e}^{-m}$.

Lemma 7.2. *There exists universal constants* c_4, c_5 *and* c_6 *such that the following is true. For any* $t > 0$ *such that*

$$t < \min\left\{C_{\min}^2(\sqrt{s} + \sqrt{m})^{-4}, \frac{\gamma\lambda}{2}, \min_{j\in S}|v_j| - g(\lambda)\right\}$$

we have

$$\mathbb{P}(\sigma \leq t) \leq 4\mathrm{e}^{-c_6\min\{m\epsilon^2, s\}} + 2(N-s)\mathrm{e}^{-\frac{(\gamma - 2t/\lambda)^2}{2\rho_u \overline{M}_N(\epsilon)}} + c_4\mathrm{e}^{-c_5\min\{s, \ln(N-s)\}}] + 2\mathrm{e}^{-\frac{(\sqrt{C_{\min}} - (\sqrt{s}+\sqrt{m})t^{\frac{1}{4}})^2}{2t^{\frac{1}{2}}}} \tag{7.1}$$

where

$$\overline{M}_N(\epsilon) := \frac{(1+\epsilon)}{C_{\min}\theta_u m}\left(s\theta_u + \frac{m}{2\ln(N)\phi_N}\right). \tag{7.2}$$

In particular, there exists constants c_7, c_8 *such that*

$$\mathbb{P}(\sigma \leq \hat{\sigma}) \leq c_7\mathrm{e}^{-c_8\min\{s, \ln(N-s)\}} \tag{7.3}$$

where $\hat{\sigma} = \min\left\{\frac{C_{\min}^2}{4(\sqrt{s}+\sqrt{m})^4}, \frac{\gamma\lambda}{4}, \min_{j\in S}|v_j| - g(\lambda)\right\}$.

Assuming for now Lemmas 7.1 and 7.2, the proof of Theorem 3.3 is simple.

Proof of Theorem 3.3. Recall the function $q(\nu, \xi)$ defined in (3.4) (and I.7.1). Note that for any fixed $\nu \geq 1$ the function $f(\xi) = \xi^2/q(\nu, \xi)$ is increasing on $[0, \infty)$. Indeed, it suffices to show that $g(\xi) := f(\xi^2)$ is increasing on $[0, \infty)$ and this can be done by noting that $g(\xi)$ takes the form $g(\xi) = \xi^4/(C + D\xi + E\xi^2)$ for positive C, D, E and thus, for $\xi \geq 0$,

$$g'(\xi) = \frac{4\xi^3(C + D\xi + E\xi^2) - \xi^4(D + 2E\xi)}{(C + D\xi + E\xi^2)^2} \geq \frac{3D\xi^4 + 2E\xi^5}{(C + D\xi + E\xi^2)^2} \geq 0.$$

It is also clear that if we instead fix ξ then $\xi^2/q(\nu, \xi)$ is decreasing in ν. Similarly, the function $\sqrt{\xi}/6\nu$ is increasing in ξ and decreasing in ν for $\nu \in [1, \infty)$.

Thus if we take $\sigma = \sigma(y, A) \geq \hat{\sigma}$ and $1 \leq \alpha = [\![(y, A)]\!]_2 < \hat{\alpha}$, we have (by Proposition 5.4)

$$\mathsf{stsp}(y, A) \geq (mN)^{-\frac{1}{2}}\min\left(\frac{\sigma^2}{q(\alpha, \sigma)}, \frac{\sqrt{\sigma}}{6\alpha}, \alpha\right) \geq (mN)^{-\frac{1}{2}}\min\left(\frac{\hat{\sigma}^2}{q(\hat{\alpha}, \hat{\sigma})}, \frac{\sqrt{\hat{\sigma}}}{6\hat{\alpha}}, 1\right) = \hat{\xi}$$

where we set $\hat{\xi} = \widehat{K}^{-1}$.

We conclude that for $\mathsf{stsp}(y, A) \leq \hat{\xi}$ we must have either $\sigma < \hat{\sigma}$ or $\alpha > \hat{\alpha}$. Therefore $\mathbb{P}(\mathscr{C}_{\mathrm{UL}}(y, A) \geq \widehat{K}) = \mathbb{P}(\mathsf{stsp}(y, A) \leq \hat{\xi}) \leq \mathbb{P}(\sigma < \hat{\sigma} \cup \alpha > \hat{\alpha}) \leq \mathbb{P}(\sigma < \hat{\sigma}) + \mathbb{P}(\alpha > \hat{\alpha})$. By Lemmas 7.1 and 7.2, this is bounded above by $3\mathrm{e}^{-m} + c_7\mathrm{e}^{-c_8\min\{s, \ln(N-s)\}} \leq \mathbf{c}_1\mathrm{e}^{-\mathbf{c}_2\min\{s, \ln(N-s)\}}$ for some universal constants $\mathbf{c}_1, \mathbf{c}_2$, the last inequality by (ai), completing the proof. □

The remainder of this section is dedicated to proving Lemmas 7.1 and 7.2. Since $\alpha = \max\{\|y\|_2, \|A\|_2, 1\}$, we devise probabilistic bounds for $\|y\|_2$ and $\|A\|_2$ separately.

Lemma 7.3. *We have* $\mathbb{P}\big(\|A\|_2 \geq \|\Sigma\|_2^{\frac{1}{2}}(3\sqrt{m} + 6\sqrt{N})\big) \leq \mathrm{e}^{-m}$.

Proof. Write $A = U\sqrt{\Sigma}$, where $\sqrt{\ }$ represents the matrix square root. Then, the entries of $U \in \mathbb{R}^{m\times N}$ are i.i.d. with distribution $\mathcal{N}(0, 1)$. Furthermore, for any $t > 0$,

$$\begin{aligned}\mathbb{P}\big(\|A\|_2 \geq \|\Sigma\|_2^{\frac{1}{2}}(\sqrt{m}(1+t) + 6\sqrt{N})\big) &\leq \mathbb{P}\big(\|U\|_2 \geq \sqrt{m}(1+t) + 6\sqrt{N}\big) \\ &\leq \mathrm{e}^{-\frac{2(\sqrt{m}(1+t))^2}{\pi^2}} \;=\; \mathrm{e}^{-m\frac{2(1+t)^2}{\pi^2}}\end{aligned}$$

the second line by [15, (4.6) and Lemma 4.14]. The result follows by setting $t = 2$. □

Lemma 7.4. $\mathbb{P}\left(\|y\|_2 \geq \frac{5}{2}\sqrt{m}(\eta + C_{\max}\|v\|_2)\right) \leq 2\mathrm{e}^{-m}$.

Proof. As $y = Av + w$, for all $t > 0$, we have

$$\begin{aligned}\mathbb{P}(\|y\|_2 \geq t) &\leq \mathbb{P}\left(\|Av\|_2 \geq \frac{t\sqrt{v^{\mathrm{T}}\Sigma v}}{\eta + \sqrt{v^{\mathrm{T}}\Sigma v}}\right) + \mathbb{P}\left(\|w\|_2 \geq \frac{t\eta}{\eta + \sqrt{v^{\mathrm{T}}\Sigma v}}\right) \\ &= \mathbb{P}\left(\frac{\|Av\|_2}{\sqrt{v^{\mathrm{T}}\Sigma v}} \geq \frac{t}{\eta + \sqrt{v^{\mathrm{T}}\Sigma v}}\right) + \mathbb{P}\left(\frac{\|w\|_2}{\eta} \geq \frac{t}{\eta + \sqrt{v^{\mathrm{T}}\Sigma v}}\right).\end{aligned}$$

We next observe that $\|w\|_2/\eta$ and $\|Av\|_2/\sqrt{v^{\mathrm{T}}\Sigma v}$ are independent random variables, both Chi distributed with parameter m (this is obvious for $\|w\|_2/\eta$ whereas for $\|Av\|_2/\sqrt{v^{\mathrm{T}}\Sigma v}$, note that for each $i = 1, 2, \ldots, m$ the random variable $\Sigma_{j=1}^{N} A_{ij}v_j$ is normally distributed with variance $v^{\mathrm{T}}\Sigma v$ and the claim follows by the independence of the rows of A). Let ζ denote any of these two random variables. By [15, Corollary 4.6], $\mathbb{P}(\zeta \geq \sqrt{m} + u) \leq \mathrm{e}^{-\frac{u^2}{2}}$. Hence, taking $t = \frac{5}{2}\sqrt{m}(\eta + C_{\max}\|v\|_2)$ and noting that $\sqrt{v^{\mathrm{T}}\Sigma v} \leq C_{\max}\|v\|_2$ we get

$$\mathbb{P}(\|y\|_2 \geq t) \leq 2\mathbb{P}\left(\zeta \geq \frac{\frac{5}{2}\sqrt{m}(\eta + C_{\max}\|v\|_2)}{(\eta + \sqrt{v^{\mathrm{T}}\Sigma v})}\right) \leq 2\mathbb{P}\left(\zeta \geq \frac{5}{2}\sqrt{m}\right) \leq 2\mathrm{e}^{-\frac{9}{8}m} \leq 2\mathrm{e}^{-m}.$$

□

We can now prove Lemma 7.1.

Proof of Lemma 7.1. By the definition of α,

$$\begin{aligned}\mathbb{P}(\alpha \geq \hat{\alpha}) &\leq \mathbb{P}(\|y\|_2 \geq \hat{\alpha}) + \mathbb{P}(\|A\|_2 \geq \hat{\alpha}) \\ &\leq \mathbb{P}\Big(\|y\|_2 \geq \frac{5}{2}\sqrt{m}(\eta + C_{\max}\|v\|_2)\Big) + \mathbb{P}\big(\|A\|_2 \geq \|\Sigma\|_2^{\frac{1}{2}}(3\sqrt{m} + 6\sqrt{N})\big) \\ &\leq\ 3\mathrm{e}^{-m}\end{aligned}$$

where the last inequality holds by Lemmas 7.3 and 7.4. □

The proof of Lemma 7.2 relies on three results from Wainwright [52]: Theorem 3.2 and the following two lemmas.

Lemma 7.5. ([52, p. 2193]) *Under the setup in Section 3 and assumptions* (a0) *and* (ai)*, there exists a universal constant c_0 such that, for any $t \leq \gamma$ we have*

$$\mathbb{P}(\max_{j \in S^c} |Z_j| \geq 1 - t) \leq 4\mathrm{e}^{-c_0 \min\{m\epsilon^2, s\}} + 2(N - s)\mathrm{e}^{-\frac{(\gamma - t)^2}{2\rho_u \overline{M}_N(\epsilon)}}$$

where $\overline{M}_N(\epsilon)$ is defined as in (7.2)*, Z_j is defined by*

$$Z_j = A_j^* \left[A_S(A_S^*A_S)^{-1}\,\mathsf{sgn}(\hat{z}) + R\left(\frac{2w}{\lambda}\right)\right], \quad R = I_m - A_S(A_S^*A_S)^{-1}A_S^* \tag{7.4}$$

and where $\hat{z}$ solves a restricted lasso problem*, specifically, $\hat{z}$ is a solution of*

$$\operatorname*{argmin}_{\hat{x} \in \mathbb{R}^{|S|}} \|A_S\hat{x} - y\|_2^2 + \lambda\|\hat{x}\|_1. \tag{7.5}$$

□

Note that by (UL4) and (UL5), the definition of Z_j is independent of the choice of $\hat{z}$ in the lemma above.

Lemma 7.6. [52, Lemma 9] *Let X be an $m \times s$ matrix such that each row is an i.i.d random column vector with distribution $\mathcal{N}(0, W)$ for some covariance matrix W. Then, for all $\tau > 0$*

$$\mathbb{P}\left(\left\|\left(\frac{X^*X}{m}\right)^{-1} - W^{-1}\right\|_2 \geq \frac{u(m, s, \tau)}{C_{\min}}\right) \leq 2\mathrm{e}^{-\frac{m\tau^2}{2}}$$

where $u(m, s, \tau) = 2(\sqrt{s/m} + \tau) + (\sqrt{s/m} + \tau)^2$. □

Proof of Lemma 7.2. We define E to be the event that $\sigma(y, A) \leq t$ and E_1 to be the event that the conclusion of Theorem 3.2 does not hold. Note that the matrix $A_S^* A_S$ is invertible with probability 1, so we assume throughout that it is indeed invertible.

On the event E_1^{c}, there is a unique minimiser x in $\mathsf{Sol}^{\mathsf{UL}}(y, A)$. Furthermore, x has support S and $\mathsf{sgn}(x_S) = \mathsf{sgn}(v_S)$. Thus if E_1^{c} occurs,

(1) $\sigma_1(y, A) = \lambda/2 - \|A_{S^{\mathsf{c}}}^*(Ax - y)\|_\infty$,
(2) $\sigma_2(y, A) := (\|(A_S^* A_S)^{-1}\|_2)^{-1}$,
(3) $\sigma_3(y, A) := \min_{i \in S} |x_i|$.

Our proof starts from the observation that

$$\mathbb{P}(E) \leq \mathbb{P}(E \cup E_1) = \mathbb{P}(E_1) + \mathbb{P}[E_1^{\mathsf{c}} \cap (\sigma_1 \leq t)] + \mathbb{P}[E_1^{\mathsf{c}} \cap (\sigma_2 \leq \sqrt{t})] + \mathbb{P}[E_1^{\mathsf{c}} \cap (\sigma_3 \leq t)] \tag{7.6}$$

together with the fact that Theorem 3.2 gives us a bound for $\mathbb{P}(E_1)$. We must thus analyse the events $E_1^{\mathsf{c}} \cap (\sigma_1 \leq t)$, $E_1^{\mathsf{c}} \cap (\sigma_2 \leq \sqrt{t})$ and $E_1^{\mathsf{c}} \cap (\sigma_3 \leq t)$.

Starting with $E_1^{\mathsf{c}} \cap (\sigma_1 \leq t)$, if E_1^{c} occurs then the only solution to the lasso problem is the vector x with support exactly S. Thus the solution to the restricted lasso problem (7.5) is given by x_S.

By the KKT conditions for (7.5) at x_S, we have $A_S^*(A_S x_S - A_S v_S - w) = A_S^*(A_S x_S - y) = -\lambda\, \mathsf{sgn}(x_S)/2$ so that $x_S - v_S = (A_S^* A_S)^{-1}[A_S^* w - \lambda\, \mathsf{sgn}(x_S)/2]$. Therefore for $j \in S^{\mathsf{c}}$,

$$\begin{aligned} A_j^*(Ax - y) &= A_j^*(A_S(x_S - v_S) - w) \\ &= A_j^*(A_S(A_S^* A_S)^{-1}[A_S^* w - \lambda\, \mathsf{sgn}(x_S)/2] - w) \\ &= -\frac{\lambda A_j^* \left[R\left(\frac{2w}{\lambda}\right) + A_S(A_S^* A_S)^{-1}\, \mathsf{sgn}(x_S)\right]}{2} = -\frac{\lambda Z_j}{2} \end{aligned}$$

where Z_j is defined as in (7.4).

Thus if $\sigma_1 \leq t$ then we must have $\lambda/2 - t \leq \|A_{S^{\mathsf{c}}}^*(Ax - y)\|_\infty = \lambda \max_{j \in S^{\mathsf{c}}} |Z_j|/2$ and it follows that for any $\gamma \geq t$,

$$\begin{aligned} \mathbb{P}(\sigma_1 \leq t) &\leq \mathbb{P}\left(\max_{j \in S^{\mathsf{c}}} |Z_j| > 1 - \frac{2t}{\lambda}\right) \\ &\leq 4\mathrm{e}^{-c_0 \min\{m\epsilon^2, s\}} + 2(N - s)\mathrm{e}^{-\frac{(\gamma - 2t/\lambda)^2}{2\rho_u \overline{M}_N(\epsilon)}} \end{aligned} \tag{7.7}$$

the last inequality from Lemma 7.5.

Next, let us analyse σ_2 assuming that the event E_1^{c} occurs. If $\sigma_2 \leq \sqrt{t}$ then since x is supported on S, we must have $\|(A_S^* A_S)^{-1}\|_2 \geq 1/\sqrt{t}$. In particular, if this occurs then

$$\left\| \frac{(A_S^* A_S)}{m}^{-1} - (\Sigma_{SS})^{-1} \right\|_2 \geq (\sqrt{t} m)^{-1} - \|(\Sigma_{SS})^{-1}\|_2 = \frac{1}{m\sqrt{t}} - \frac{1}{C_{\min}} = \frac{C_{\min} - m\sqrt{t}}{C_{\min}\, m\sqrt{t}}$$

Because $t < C_{\min}^2 (\sqrt{s} + \sqrt{m})^{-4}$, some simple algebra gives $\tau := \sqrt{\frac{C_{\min}}{m\sqrt{t}}} - \sqrt{\frac{s}{m}} - 1 > 0$ and moreover if $u(m, s, \tau) = 2(\sqrt{s/m} + \tau) + (\sqrt{s/m} + \tau)^2$ then some more simple algebra yields $u(m, s, \tau) + 1 = \frac{C_{\min}}{m\sqrt{t}}$. Thus

$$\begin{aligned} \mathbb{P}(\|(A_S^* A_S)^{-1}\|_2 \geq 1/\sqrt{t}) &\leq \mathbb{P}\left(\left\| \left(\frac{A_S^* A_S}{m}\right)^{-1} - (\Sigma_{SS})^{-1} \right\|_2 \geq \frac{C_{\min} - m\sqrt{t}}{C_{\min}\, m\sqrt{t}}\right) \\ &= \mathbb{P}\left(\left\| \left(\frac{A_S^* A_S}{m}\right)^{-1} - (\Sigma_{SS})^{-1} \right\|_2 \geq \frac{u(m, s, \tau)}{C_{\min}}\right) \\ &\leq 2\mathrm{e}^{-\frac{\left(\sqrt{C_{\min}} - (\sqrt{s} + \sqrt{m}) t^{1/4}\right)^2}{2t^{1/2}}} \end{aligned} \tag{7.8}$$

where the final inequality follows from Lemma 7.6.

We can immediately see that $\sigma_3(y,A) \leq t$ and E_1^{c} cannot occur simultaneously: if E_1^{c} occurs then for $j \in S$ we have $|v_j| - |x_j| \leq |v_j - x_j| \leq g(\lambda)$ and so $\sigma_3(y,A) = \min_{j\in S}|x_j| \geq \min_{j\in S}|v_j| - g(\lambda) > t$. We conclude that

$$\mathbb{P}[E_1^{\mathsf{c}} \cap (\sigma_3 \leq t)] = 0. \tag{7.9}$$

Equation (7.1) follows from (7.6), Theorem 3.2, (7.7), (7.8) and (7.9).

All that remains is to prove (7.3). We first claim that with the specific choice of $\hat{\sigma}$ we have

$$(N-s)\mathrm{e}^{-\frac{(\gamma-2\hat{\sigma}/\lambda)^2}{2\rho_u \overline{M}_N(\epsilon)}} \leq \mathrm{e}^{\frac{-\ln(N-s)}{2}}. \tag{7.10}$$

Indeed, since $\hat{\sigma} < \gamma\lambda/4$ we have $-(\gamma - 2\hat{\sigma}/\lambda)^2/(2\rho_u\overline{M}_N(\epsilon)) \leq -\gamma^2/(8\rho_u\overline{M}_N(\epsilon))$. Using the definition of ρ_u gives $\gamma^2/(8\rho_u\overline{M}_N(\epsilon)) = (8C_{\min}\theta_u\overline{M}_N(\epsilon))^{-1}$. Also, equation (3.3) implies

$$\frac{m}{1+\epsilon} > 12s\ln(N-s)\theta_u + \frac{6m\ln(N-s)}{\ln(N)\phi_N} \tag{7.11}$$

and hence,

$$\begin{aligned}
\frac{1}{8C_{\min}\theta_u\overline{M}_N(\epsilon)} &\underset{(7.2)}{=} \frac{m}{8(1+\epsilon)}\left(s\theta_u + \frac{m}{2\ln(N)\phi_N}\right)^{-1} \\
&\underset{(7.11)}{>} \ln(N-s)\left(\frac{1}{8}\left(12s\theta_u + \frac{6m}{\ln(N)\phi_N}\right)\right)\left(s\theta_u + \frac{m}{2\ln(N)\phi_N}\right)^{-1} \\
&= \frac{3}{2}\ln(N-s).
\end{aligned}$$

It follows that

$$\ln(N-s) - \frac{1}{8C_{\min}\theta_u\overline{M}_N(\epsilon)} < \ln(N-s) - \frac{3}{2}\ln(N-s) = -\frac{\ln(N-s)}{2}$$

and thus

$$\mathrm{e}^{\ln(N-s)-\frac{(\gamma-2\hat{\sigma}/\lambda)^2}{2\rho_u\overline{M}_N(\epsilon)}} \leq \mathrm{e}^{\ln(N-s)-\frac{1}{8C_{\min}\theta_u\overline{M}_N(\epsilon)}} \leq \mathrm{e}^{\frac{-\ln(N-s)}{2}}$$

proving the claim.

We further claim that

$$2\mathrm{e}^{-\frac{\left(\sqrt{C_{\min}}-(\sqrt{s}+\sqrt{m})t^{\frac{1}{4}}\right)^2}{t^{\frac{1}{2}}}}$$

is increasing in t when $(\sqrt{s}+\sqrt{m})t^{\frac{1}{4}} \leq \sqrt{C_{\min}}$. Indeed, the function $f(t^4) := -(\sqrt{C_{\min}} - (\sqrt{s}+\sqrt{m})t)^2/t^2 = -(\sqrt{C_{\min}}/t - (\sqrt{s}+\sqrt{m}))^2$ is increasing in t provided that $t < \sqrt{C_{\min}}/(\sqrt{s}+\sqrt{m})$. With this result and using that $\hat{\sigma} \leq \frac{C_{\min}^2}{4(\sqrt{s}+\sqrt{m})^4}$ (see Lemma 7.2), we see that

$$\begin{aligned}
2\mathrm{e}^{-\frac{\left(\sqrt{C_{\min}}-(\sqrt{s}+\sqrt{m})\hat{\sigma}^{\frac{1}{4}}\right)^2}{\hat{\sigma}^{\frac{1}{2}}}} &\leq 2\mathrm{e}^{-2(\sqrt{s}+\sqrt{m})^2\frac{\left[\sqrt{C_{\min}}-\sqrt{C_{\min}}/(4^{\frac{1}{4}})\right]^2}{C_{\min}}} \\
&= 2\mathrm{e}^{-2(1-4^{-1/4})^2(\sqrt{s}+\sqrt{m})^2} \leq 2\mathrm{e}^{-s/3}
\end{aligned} \tag{7.12}$$

since $2(1-4^{-1/4})^2(\sqrt{s}+\sqrt{m})^2 \geq (\sqrt{2}-1)^2 s \geq s/3$.

We can now conclude (7.3) in the following way, starting from (7.1),

$$\begin{aligned}
\mathbb{P}(\sigma \leq t) &\leq 4\mathrm{e}^{-c_6\min\{m\epsilon^2,s\}} + 2(N-s)\mathrm{e}^{-\frac{(\gamma-2t/\lambda)^2}{2\rho_u\overline{M}_N(\epsilon)}} + c_4\mathrm{e}^{-c_5\min\{s,\ln(N-s)\}} + 2\mathrm{e}^{-\frac{\left(\sqrt{C_{\min}}-(\sqrt{s}+\sqrt{m})t^{\frac{1}{4}}\right)^2}{t^{\frac{1}{2}}}} \\
&\underset{(7.10)}{\leq} 4\mathrm{e}^{-c_6\min\{m\epsilon^2,s\}} + 2\mathrm{e}^{-\ln(N-s)/2} + c_4\mathrm{e}^{-c_5\min\{s,\ln(N-s)\}} + 2\mathrm{e}^{-\frac{\left(\sqrt{C_{\min}}-(\sqrt{s}+\sqrt{m})t^{\frac{1}{4}}\right)^2}{t^{\frac{1}{2}}}} \\
&\underset{(7.12)}{\leq} 4\mathrm{e}^{-c_6\min\{m\epsilon^2,s\}} + 2\mathrm{e}^{-\ln(N-s)/2} + c_4\mathrm{e}^{-c_5\min\{s,\ln(N-s)\}} + 2\mathrm{e}^{-s/3} \\
&\underset{(\mathrm{ai})}{\leq} \max\{4,c_4\}\mathrm{e}^{-\min\{c_6,1/2,c_5,1/6\}\min\{s,\ln(N-s)\}}
\end{aligned}$$

and thus we have shown both (7.1) and (7.3), completing the proof. □

8. Proof of Theorem 2.5

Lemma 8.1. *Let $w \in \mathbb{R}^q$ be random with i.i.d. components distributed as $\mathcal{N}(0,1)$. Then, for all $t > 0$, $\mathbb{P}(\|w\|_\infty \leq t) \geq 1 - \frac{2q}{t\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$.*

Proof. We have $\mathbb{P}(|w_i| \geq t) \leq \frac{2}{t\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$ by [15, Lemma 2.16]. Hence,

$$\mathbb{P}(\|w\|_\infty \geq t) = \mathbb{P}(\vee_{i\leq q}|w_i| \geq t) \leq \frac{2q}{t\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$$

and the result follows. □

Proof of Corollary 2.5. We will use the notations in Theorem 3.3 and Corollary 3.4. Also, we take $\overline{c} := \max\{c_3, 1\}$ where c_3 is the universal constant in (aii').

Because $m, s \leq N/9 < N/8$ we have $\frac{1}{4(\sqrt{s}+\sqrt{m})^4} > \frac{1}{N^2}$. Also, by hypothesis (ii) and since $\eta = 0$, $\min_{j\in S}|v_j| - g(\lambda) \geq \frac{1}{N^2}$. Along with hypothesis (iii) it follows that

$$\hat{\sigma} \geq \frac{1}{N^2}. \tag{8.1}$$

Similarly, using that $m \leq N/9$,

$$\hat{\alpha} \leq \max\left\{\sqrt{2N}\|v\|_2, \sqrt{N} + 6\sqrt{N}\right\} \leq 7\,[\![v]\!]_2\sqrt{N}. \tag{8.2}$$

Finally, because $\lambda \leq \frac{2N[\![v]\!]_2}{\overline{c}}$ (by (ii)) and $\overline{c} \geq 1$, $\lambda\sqrt{N} \leq 2N^{1.5}[\![v]\!]_2$ and, hence,

$$(1 + \lambda\sqrt{N}) \leq 3N^{1.5}[\![v]\!]_2. \tag{8.3}$$

It follows that

$$\begin{aligned}
\frac{q(\hat{\alpha},\hat{\sigma})}{\hat{\sigma}^2} &= \frac{96\hat{\alpha}^5}{\hat{\sigma}^2} + \frac{12\hat{\alpha}^3(1+\lambda\sqrt{N})}{\hat{\sigma}^{1.5}} + \left(\frac{2\hat{\alpha}^3}{\lambda} + 3\hat{\alpha}\right)\frac{1}{\hat{\sigma}} \\
&\underset{(8.3),(iii)}{\leq} \frac{96\hat{\alpha}^5}{\hat{\sigma}^2} + \frac{36\hat{\alpha}^3 N^{1.5}[\![v]\!]_2}{\hat{\sigma}^{1.5}} + \frac{\hat{\alpha}^3N^2 + 3\hat{\alpha}}{\hat{\sigma}} \\
&\underset{(8.2)}{\leq} \frac{96\cdot 7^5\cdot N^{2.5}[\![v]\!]_2^5}{\hat{\sigma}^2} + \frac{36\cdot 7^3\cdot N^3[\![v]\!]_2^4}{\hat{\sigma}^{1.5}} + \frac{7^3N^{3.5}[\![v]\!]_2^3 + 21\,N^{0.5}[\![v]\!]_2}{\hat{\sigma}} \\
&\underset{(8.1)}{\leq} 96\cdot 7^5\cdot N^{6.5}[\![v]\!]_2^5 + 36\cdot 7^3\cdot N^6[\![v]\!]_2^4 + 7^3N^{5.5}[\![v]\!]_2^3 + 21\,N^{2.5}[\![v]\!]_2 \\
&\leq 1626184\,N^{6.5}[\![v]\!]_2^5.
\end{aligned}$$

Also, $\frac{6\hat{\alpha}}{\sqrt{\hat{\sigma}}} \leq 42N^{1.5}[\![v]\!]_2$. We conclude that

$$\hat{K} \leq (mN)^{\frac{1}{2}}\max\left\{\frac{q(\hat{\alpha},\hat{\sigma})}{\hat{\sigma}^2}, \frac{6\hat{\alpha}}{\sqrt{\hat{\sigma}}}\right\} \leq \frac{1626184}{3}N^{7.5}[\![v]\!]_2^5 < 542062N^{7.5}[\![v]\!]_2^5. \tag{8.4}$$

We next note that hypotheses (i) and (ii) in our statement imply that (ai') and (aii') are satisfied. Hence, Corollary 3.4 (with $\eta = 0$) may be applied to deduce that

$$\mathbb{P}(\mathscr{C}_{\mathrm{UL}}(b,U) \geq \widehat{K}) \leq \mathbf{c}_1\mathrm{e}^{-\mathbf{c}_2\min\{\ln(N-s),s\}} \leq \mathbf{c}_1\mathrm{e}^{-\mathbf{c}_2\ln(N/2)} \leq \mathbf{c}_1\left(\frac{N}{2}\right)^{-\mathbf{c}_2}$$

the second inequality by our hypothesis on s.

Because of Lemma 8.1 with $q = mN$ and $t = N^2$, we have

$$\mathbb{P}(\|U\|_{\max} \leq N^2) \geq 1 - \frac{2mN}{N^2\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}} \geq 1 - \mathrm{e}^{-\frac{N^4}{2}}.$$

Also, $\|b\|_\infty \leq \|U\|_{2\infty}\|v\|_2 \leq \sqrt{N}\|U\|_{\max}\|v\|_2$. Therefore, $[\![(b,U)]\!]_{\max} \leq N^{2.5}[\![v]\!]_2$ and

$$\lambda + \lambda^{-1} \leq \frac{2N\|v\|_2}{\overline{c}} + N^2/2 \leq 2N^2[\![v]\!]_2, \tag{8.5}$$

$$[\![(b,U)]\!]_{\mathrm{S}} \leq mN[\![(b,U)]\!]_{\max} \leq N^{4.5}[\![v]\!]_2, \tag{8.6}$$

with probability at least $1-\mathrm{e}^{-\frac{N^4}{2}}$.

Using (8.4) and Theorem I.1.2, we conclude that the cost of the algorithm for random (b,U) satisfying our assumptions is bounded by $\mathcal{O}\big(N^3(\log_2 N^{14.5}[\![v]\!]_2^7)^2\big)$ and using (8.4), (8.5) and (8.6), the maximum number of digits accessed by the algorithm is bounded by

$$\mathcal{O}(\lceil \log_2\left(\max\{\lambda+\lambda^{-1}, N, [\![(b,U)]\!]_{\mathrm{S}}, \mathscr{C}_{\mathrm{UL}}(b,U)\}\right)\rceil) = \mathcal{O}(\log_2(N[\![v]\!]_2))$$

with probability at least $1-(\mathbf{c}_1\left(\frac{N}{2}\right)^{-\mathbf{c}_2}+\mathrm{e}^{-\frac{N^4}{2}}) \geq 1-\mathbf{C}_1N^{-\mathbf{C}_2}$ for some appropriately chosen constants $\mathbf{C}_1,\mathbf{C}_2$. Because the hypotheses of Theorem 3.2 hold it follows that with probability greater than $1-c_1\mathrm{e}^{-c_2\min\{\ln(N-s),s\}} \geq 1-c_1\left(\frac{N}{2}\right)^{-c_2}$ we also have that $\mathsf{supp}(x)=\mathsf{supp}(v)$ (here $\mathsf{Sol}^{\mathsf{UL}}(b,U)=\{x\}$). The result follows by changing, if necessary, the values of $\mathbf{C}_1$ and $\mathbf{C}_2$. □

9. Proof of Theorem 2.3

9.1. **Definitions for the computational problem.** We start by recalling some definitions from [7]:

Definition 9.1 (The LASSO computational problem)**.** For some set $\Omega \subset \mathbb{R}^m \times \mathbb{R}^{m\times N}$, which we call the *input* set, the *LASSO computational problem on* Ω is the collection $\{\Xi,\Omega,\mathbb{B}^N,\Lambda\}$ where $\Xi:\Omega\to 2^{\mathbb{B}^N}$ is defined as in (1.1) and

$$\Lambda=\{f^{\mathrm{vec}},f^{\mathrm{mat}}\} \text{ with } f^{\mathrm{vec}}:\Omega\to\mathbb{R}^m, f^{\mathrm{mat}}:\Omega\to\mathbb{R}^{m\times N}$$

are defined by $f^{\mathrm{vec}}(y,A)=y$ and $f^{\mathrm{mat}}(y,A)=A$ for all $(y,A)\in\Omega$.

We want to generalise the LASSO computational problem so that we work with *inexact inputs*. To do so, we will consider the collection of all functions $f_n^{\mathrm{vec}}:\Omega\to\mathbb{R}^m$ and $f_n^{\mathrm{mat}}:\Omega\to\mathbb{R}^{m\times N}$ satisfying

$$\|f_n^{\mathrm{vec}}(y,A)-y\|_\infty \leq 2^{-n}, \quad \|f_n^{\mathrm{mat}}(y,A)-A\|_{\max}\leq 2^{-n} \tag{9.1}$$

Definition 9.2 (Inexact LASSO computational problem)**.** The *inexact LASSO computational problem on* Ω (ILCP) is the quadruple $\{\tilde{\Xi},\tilde{\Omega},\mathbb{B}^N,\tilde{\Lambda}\}$, where

$$\begin{aligned}\tilde{\Omega}=\big\{\tilde{\iota}=\{(f_n^{\mathrm{vec}}(\iota),f_n^{\mathrm{mat}}(\iota)\}_{n\in\mathbb{N}} \mid \iota=(y,A)\in\Omega \text{ and}\\ f_n^{\mathrm{vec}}:\Omega\to\mathbb{R}^m, f_n^{\mathrm{mat}}:\Omega\to\mathbb{R}^{m\times N} \text{ satisfy (9.1) respectively}\big\}\end{aligned} \tag{9.2}$$

It follows from (9.1) that there is a unique $\iota=(y,A)\in\Omega$ for which $\tilde{\iota}=\big\{(f_n^{\mathrm{vec}}(\iota),f_n^{\mathrm{mat}}(\iota))\big\}_{n\in\mathbb{N}}$. We say that this $\iota\in\Omega$ *corresponds* to $\tilde{\iota}\in\tilde{\Omega}$ and we set $\tilde{\Xi}:\tilde{\Omega}\rightrightarrows\mathbb{B}^N$ so that $\Xi(\tilde{\iota})=\Xi(\iota)$, and $\tilde{\Lambda}=\{\tilde{f}_n^{\mathrm{vec}},\tilde{f}_n^{\mathrm{mat}}\}_{n\in\mathbb{N}}$, with $\tilde{f}_n^{\mathrm{vec}}(\tilde{\iota})=f_n(\iota)$, $\tilde{f}_n^{\mathrm{mat}}(\tilde{\iota})=f_n^{\mathrm{mat}}(\iota)$ where ι corresponds to $\tilde{\iota}$.

Definition 9.3 (General Algorithms for the ILCP)**.** A *general algorithm* for $\{\tilde{\Xi},\tilde{\Omega},\mathbb{B}^N,\tilde{\Lambda}\}$, is a mapping $\Gamma:\tilde{\Omega}\to\mathbb{B}^N$ such that, for every $\tilde{\iota}\in\tilde{\Omega}$, the following conditions hold:

(i) there exists a nonempty subset of evaluations $\Lambda_\Gamma(\tilde{\iota})\subset\tilde{\Lambda}$ with $|\Lambda_\Gamma(\tilde{\iota})|<\infty$,
(ii) the action of Γ on $\tilde{\iota}$ is uniquely determined by $\{f(\tilde{\iota})\}_{f\in\Lambda_\Gamma(\tilde{\iota})}$,
(iii) for every $\iota'\in\Omega$ such that $f(\iota')=f(\tilde{\iota})$ for all $f\in\Lambda_\Gamma(\tilde{\iota})$, it holds that $\Lambda_\Gamma(\iota')=\Lambda_\Gamma(\tilde{\iota})$.

Specific to this paper is the study of *finite precision algorithms* and *algorithms that are correct on all inputs that can be represented exactly*. To define these concepts, we first define the following, which is similar to [8, Definition 8.23].

Definition 9.4 (Number of correct 'digits' for the ILCP)**.** Given a general algorithm Γ for the inexact LASSO computational problem, we define the *'number of digits' required on the input* according to

$$D_\Gamma(\tilde{\iota}) := \sup\{m\in\mathbb{N} \mid \text{at least one of } f_m^{\mathrm{vec}}, f_m^{\mathrm{mat}} \in \Lambda_\Gamma(\tilde{\iota})\}.$$

Definition 9.5 (Finite precision algorithm)**.** A *finite precision general algorithm* with precision 2^{-k} for $\{\tilde{\Xi}, \tilde{\Omega}, \mathbb{B}^N, \tilde{\Lambda}\}$, is a general algorithm $\Gamma : \tilde{\Omega} \to \mathbb{B}^N$ such that, for every $\tilde{\iota} \in \tilde{\Omega}$, the number of correct digits $D_\Gamma(\tilde{\iota}) \leq k$.

Definition 9.6. A general algorithm Γ for the ILCP defined on $\Omega \subset \mathbb{R}^m \times \mathbb{R}^{m\times N}$ is said to be *correct on all inputs that can be represented exactly in* Ω if for all $y \in \mathbf{D}^m$ and $A \in \mathbf{D}^{m\times N}$ with $(y, A) \in \Omega$, we have $\Gamma(\tilde{\iota}) \in \Xi(y, A)$ whenever

$$\tilde{\iota} = ((f_1^{\mathrm{vec}}(y, A), f_1^{\mathrm{mat}}(y, A))(f_2^{\mathrm{vec}}(y, A), f_2^{\mathrm{mat}}(y, A)), \dots)$$

with $f_n^{\mathrm{vec}} : \Omega \to \mathbb{R}^m$ and $f_n^{\mathrm{mat}} : \Omega \to \mathbb{R}^{m\times N}$ defined for all $n \in \mathbb{N}$ so that $f_n^{\mathrm{vec}}(y, A) = y \quad f_n^{\mathrm{mat}}(y, A) = A$.

9.2. **Proof of Theorem 2.3.** We begin by stating and proving two preliminary lemmas:

Lemma 9.7. *Suppose that the open set $T \in \mathbb{R}^m \times \mathbb{R}^{m\times N}$ contains a point with condition infinity. Then for any $S_1 \subseteq \{1, 2, \dots, N\}$ there exists a point $P = (y^P, A^p) \in T$ and an $\epsilon > 0$ so that if $\|y - y^P\|_\infty \leq \epsilon$, $\|A - A^p\|_\infty \leq \epsilon$ then $\Xi(y, A) = \{S_2\}$ with $S_2 \neq S_1$.*

Proof. Since there is a point with condition infinity inside T, there must exist some point y^1, A^1 so that $\{S_1\} \neq \Xi(y^1, A^1)$ (otherwise $\Xi(b, U) = \{S_1\}$ for all $(b, U) \in T$ and hence $\mathrm{cond}(b, U) < \infty$ since T is open). In particular, either $\Xi(y^1, A^1) = \{W\} \neq \{S_1\}$ in which case $\Xi^{\mathrm{ms}}(y^1, A^1) = \Xi(y^1, A^1) = \{W\} \neq \{S_1\}$ or, by Lemma I.8.2 $|\mathsf{Sol}^{\mathsf{UL}}(y^1, A^1)| \geq 2$ in which case there must exist $W \in \Xi^{\mathrm{ms}}(y^1, A^1)$ with $W \neq S$.

From Lemma I.8.8 if we set $x \in \mathsf{Sol}^{\mathsf{UL}}(y^1, A^1)$ with $\mathsf{supp}(x) = W$ and E to be the diagonal matrix with entries $(\mathbb{1}_{1\in W^c}, \mathbb{1}_{2\in W^c}, \mathbb{1}_{3\in W^c}, \dots \mathbb{1}_{N\in W^c})$ on the diagonal, then $x = \mathsf{Sol}^{\mathsf{UL}}(y^1, A^1(I - \delta E))$ for $\delta > 0$ sufficiently small. Since x^1 is the unique vector in $\mathsf{Sol}(y^1, A^1(I-\delta E))$, we must have $\sigma_2(y^1, A^1(1-\delta E)) > 0$. We also have

$$\begin{aligned}\|[A^1(I - \delta E)]^*_{S^c}(A^1x - y^1)\|_\infty &= \|(I_{S^c} - \delta E_{S^c})(A^1)^*_{S^c}(A^1x - y^1)\|_\infty \\ &= \|(1 - \delta)I(A^1)^*_{S^c}(A^1x - y^1)\|_\infty \\ &= (1 - \delta)\|(A^1)^*_{S^c}(A^1x - y^1)\|_\infty \leq (1 - \delta)\lambda/2\end{aligned}$$

and thus $\sigma_1(y^1, A^1(I - \delta E)) \geq \delta\lambda/2$. Finally, $\sigma_3(y^1, A^1(I - \delta E)) > 0$ since $\mathsf{Sol}^{\mathsf{UL}}(y^1, A^1(I - \delta E))$ is a singleton. We have thus shown that $\sigma(y^1, A^1(I - \delta E)) > 0$. Thus by Proposition 5.4 we must have $\mathsf{stsp}(y^1, A^1(I - \delta E)) > 0$ and thus if we choose δ sufficiently small and positive so that $P = (y^1, A^1(I - \delta E)) \in S$ then there exists an $\epsilon > 0$ so that if $\|y - y^P\|_\infty \leq \epsilon$, $\|A - A^p\|_\infty \leq \epsilon$ then $\mathsf{Sol}(y, A) = x^1$ and in particular $\Xi(y, A) = \mathsf{supp}(x) = W \neq S$. □

Lemma 9.8. *Suppose that an inexact algorithm $\Gamma : \tilde{\Omega} \to \mathcal{M}$ is a finite precision algorithm with precision k. If there are sequences $\{s_n^1\}_{n=1}^\infty \in \tilde{\Omega}$ and $\{s_n^2\}_{n=1}^\infty \in \tilde{\Omega}$ such that $s_i^1 = s_i^2$ for $i = 1, 2, \dots, k$, then $\Gamma(\{s_n^1\}_{n=1}^\infty) = \Gamma(\{s_n^2\}_{n=1}^\infty)$.*

Proof. By the definition of a finite precision algorithm, the set $D_\Gamma(\{s_n^1\}_{n=1}^\infty)$ has cardinality at most k. In particular, $f(\{s_n^1\}_{n=1}^\infty) = f(\{s_n^2\}_{n=1}^\infty)$ for all $f \in \Lambda_\Gamma(\{s_n^1\}_{n=1}^\infty)$ and thus $\Lambda_\Gamma(\{s_n^1\}_{n=1}^\infty) = \Lambda_\Gamma(\{s_n^2\}_{n=1}^\infty)$ by Definition 9.3 (iii). We conclude that $\Gamma(\{s_n^1\}_{n=1}^\infty) = \Gamma(\{s_n^2\}_{n=1}^\infty)$ by by Definition 9.3 (ii). □

Proof of Theorem 2.3. For shorthand, we write $B = \mathcal{B}_\infty(b, U, r)$. Since $\mathrm{cond}(b, U) = \alpha$, we have $\mathsf{stsp}(b, U) = \alpha^{-1}$. In particular, $\Xi(b, U) = \{S_1\}$ for some set $S_1 \subseteq \{1, 2, \dots N\}$, otherwise, this would contradict Definition 5.1. Furthermore there exists a point $(\hat{y}, \hat{A})$ with $\mathrm{cond}(\hat{y}, \hat{A}) = \infty$ and $(\hat{y}, \hat{A}) \in \overline{\mathcal{B}}_\infty(b, U, \alpha^{-1}) \subset B$ since $r > \alpha^{-1}$.

We can thus apply Lemma 9.7 to the open set B and the support set S_1 to obtain a point (y^P, A^P) and an $\epsilon > 0$ so that $\mathcal{B}_\infty(y^P, A^P, \epsilon) \subseteq B$ and if $(y, A) \in \mathcal{B}_\infty(y^P, A^P, \epsilon)$ then $S_1 \notin \Xi(y, A) = \{S_2\}$.

We next define three sets

$$F_1 := \{(y,A) \in B \,|\, \Xi(y,A) = \{S_1\}\}, \quad F_2 := \{(y,A) \in B \,|\, \Xi(y,A) = \{S_2\}\},$$
$$F_3 := \{(y,A) \in B \,|\, |\Xi(y,A)| = 1,\, \Xi(y,A) \neq \{S_1\},\, \Xi(y,A) \neq \{S_2\}\}.$$

and note that the above argument shows that both F_1 and F_2 each contain an open ball. We now proceed to argue separately for Γ_1 and Γ_2

The result for Γ_1*:* We will first construct Δ_1 information for $F := F_1 \cup F_2 \cup F_3$ so that Γ_1 fails on this Δ_1 information for F. Note that F is the set of points (y,A) in B so that $|\Xi(y,A)| = 1$. We have already noted that $\mathsf{stsp}(b,U) = \alpha^{-1}$ implies that $\mathcal{B}_\infty(b,U,\alpha^{-1}) \in F_1$. Furthermore, by [51, Lemma 4], $B \setminus F$ has measure 0. so this will suffice to show that Γ_1 fails on F.

We thus define the Δ_1 information requireed. Since F_1 and F_2 each contain an open set they each must contain at least one point in $\mathbf{D}^m \times \mathbf{D}^{m\times N}$, say, $(d_1^{\mathrm{vec}}, d_1^{\mathrm{mat}})$ and $(d_2^{\mathrm{vec}}, d_2^{\mathrm{mat}})$ respectively.

Take arbitrary $\hat{f}_n^{\mathrm{vec}} : B \to \mathbf{D}^m$ and $\hat{f}_n^{\mathrm{mat}} : B \to \mathbf{D}^{m\times N}$ so that for all $(y,A) \in B$ we have

$$\|\hat{f}_n^{\mathrm{vec}}(y,A) - y\|_\infty \leq 2^{-n}, \quad \|\hat{f}_n^{\mathrm{mat}}(y,A) - A\|_{\max} \leq 2^{-n}$$

and $\hat{f}_n^{\mathrm{vec}}((d_i^{\mathrm{vec}}, d_i^{\mathrm{mat}})) = d_i^{\mathrm{vec}} \quad \hat{f}_n^{\mathrm{mat}}((d_i^{\mathrm{vec}}, d_i^{\mathrm{mat}})) = d_i^{\mathrm{mat}}$ for $i \in \{1,2\}$.

We now define the following Δ_1 information:

$$f_n^{\mathrm{vec}}(\iota) = \begin{cases} d_1^{\mathrm{vec}} & \text{if } n \leq k \\ \hat{f}_n^{\mathrm{vec}}(\iota) & \text{if } n > k \end{cases}, \quad f_n^{\mathrm{mat}}(\iota) = \begin{cases} d_1^{\mathrm{mat}} & \text{if } n \leq k \\ \hat{f}_n^{\mathrm{mat}}(\iota) & \text{if } n > k \end{cases} \quad \text{for } \iota \in F_2 \cup F_3$$

and

$$f_n^{\mathrm{vec}}(\iota) = \begin{cases} d_2^{\mathrm{vec}} & \text{if } n \leq k \\ \hat{f}_n^{\mathrm{vec}}(\iota) & \text{if } n > k \end{cases}, \quad f_n^{\mathrm{mat}}(\iota) = \begin{cases} d_2^{\mathrm{mat}} & \text{if } n \leq k \\ \hat{f}_n^{\mathrm{mat}}(\iota) & \text{if } n > k \end{cases} \quad \text{for } \iota \in F_1$$

and $f_n(\iota) = \hat{f}_n(\iota)$ whenever $\iota \in B \setminus F$.

This defines Δ_1 information for $(y,A) = \iota \in B$: this is clear for $\iota \in B \setminus F$. For $\iota \in F_2 \cup F_3$ this holds because $\|d_1^{\mathrm{vec}} - y\|_\infty \leq \|d_1^{\mathrm{vec}} - b\|_\infty + \|b - y\|_\infty \leq 2^{-r+1} \leq 2^{-k} \leq 2^{-n}$ and $\|d_1^{\mathrm{mat}} - A\|_{\max} \leq \|d_1^{\mathrm{mat}} - U\|_{\max} + \|U - A\|_\infty \leq 2^{-r+1} \leq 2^{-k} \leq 2^{-n}$ for $n \leq k$ and similarly for $\iota \in F_1$ we have $\|d_2^{\mathrm{vec}} - y\|_\infty, \|d_2^{\mathrm{mat}} - A\|_\infty \leq 2^{-n}$. Slightly abusing notation, we denote $\Gamma(\iota) = \Gamma(\{f_n^{\mathrm{vec}}(\iota), f_n^{\mathrm{mat}}(\iota)\}_{n=1}^\infty)$.

With this Δ_1 information, we now use the precondition that Γ is correct on all inputs that can be represented exactly in Ω to see that if $\{g_n^i\}_{n=1}^\infty$ is the constant sequence given by $g_n^i = (d_i^{\mathrm{vec}}, d_i^{\mathrm{mat}})$ then $\Gamma(\{g_n^i\}_{n=1}^\infty) = S_i$. We now apply Lemma 9.8 to obtain $\Gamma(\iota) = S_1$ when $\iota \in F_2 \cup F_3$ and $\Gamma(\iota) = S_2$ when $\iota \in F_1$. By the definitions of F_1, F_2, F_3 and F this proves that $\Gamma(\iota) \notin \Xi(\iota)$ for all $\iota \in F$.

The result for Γ_2*:* As before, there exists a point $(d_1^{\mathrm{vec}}, d_1^{\mathrm{mat}}) \in B$ with $\Xi(d_1^{\mathrm{vec}}, d_1^{\mathrm{mat}}) = S_1 = \Xi(b,U)$.

Take arbitrary $\hat{f}_n^{\mathrm{vec}} : B \to \mathbf{D}^m$ and $\hat{f}_n^{\mathrm{mat}} : B \to \mathbf{D}^{m\times N}$ so that for all $(y,A) \in B$ we have

$$\|\hat{f}_n^{\mathrm{vec}}(y,A) - y\|_\infty \leq 2^{-n}, \quad \|\hat{f}_n^{\mathrm{mat}}(y,A) - A\|_{\max} \leq 2^{-n}$$

We now define the following Δ_1 information:

$$f_n^{\mathrm{vec}}(\iota) = \begin{cases} d_1^{\mathrm{vec}} & \text{if } n \leq k \\ \hat{f}_n^{\mathrm{vec}}(\iota) & \text{if } n > k \end{cases}, \quad f_n^{\mathrm{mat}}(\iota) = \begin{cases} d_1^{\mathrm{mat}} & \text{if } n \leq k \\ \hat{f}_n^{\mathrm{mat}}(\iota) & \text{if } n > k \end{cases}$$

The same arguments as in the previous part show that this provides Δ_1 information for every $\iota \in B$. We again slightly abuse notation by setting $\Gamma(\iota) = \Gamma(\{f_n^{\mathrm{vec}}(\iota), f_n^{\mathrm{mat}}(\iota)\}_{n=1}^\infty)$.

By Lemma 9.8, if we define $\{g_n^i\}_{n=1}^\infty$ as before we must have that $\Gamma_2(\iota) = \Gamma_2(\{g_n^1\}_{n=1}^\infty)$ for all $\iota \in B$. Now if $\Gamma_2(\{g_n^1\}_{n=1}^\infty) = S_1$ then $\Gamma_2(\iota) \neq \Xi(\iota)$ for all $\iota \in F_2 \cup F_3$. Similarly, if $\Gamma_2(\{g_n^1\}_{n=1}^\infty) \neq S_1$ then $\Gamma_2(\iota) \neq \Xi(\iota)$ for all $\iota \in F_1$. Either way, since both F_1 and F_2 contain an open set, we have shown that Γ_2 fails on an open set. $\square$

REFERENCES

[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, 1964.

[2] B. Adcock and A. C. Hansen. *Compressive Imaging: Structure, Sampling, Learning*. Cambridge University Press, 2021.

[3] D. Amelunzen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge : phase transitions in convex programs with random data. *Information and Inference*, 3(3):224–294, June 2014.

[4] V. Antun, M. J. Colbrook, and A. C. Hansen. Proving existence is not enough: Mathematical paradoxes unravel the limits of neural networks in artificial intelligence. *SIAM News*, 55(04):1–4, May 2022.

[5] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc. Natl. Acad. Sci. USA*, 117(48):30088–30095, 2020.

[6] S. Arora and B. Barak. *Computational Complexity - A Modern Approach*. Princeton University Press, 2009.

[7] A. Bastounis, F. Cucker, and A. C. Hansen. When can you trust feature selection? – i: A condition-based analysis of lasso and generalised hardness of approximation. *Preprint*, 2023.

[8] A. Bastounis, A. C. Hansen, and V. Vlačić. The extended Smale's 9th problem – On computational barriers and paradoxes in estimation, regularisation, computer-assisted proofs and learning. *arXiv:2110.15734*, 2021.

[9] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[10] J. Ben-Artzi, M. J. Colbrook, A. C. Hansen, O. Nevanlinna, and M. Seidel. Computing spectra – On the solvability complexity index hierarchy and towers of algorithms. *arXiv:1508.03280v5*, 2020.

[11] J. Ben-Artzi, M. Marletta, and F. Rösler. Computing scattering resonances. *Journal of the European Mathematical Society*, 2023.

[12] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[13] A. Ben-Tal and A. Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88(3):411–424, 2000.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[15] P. Bürgisser and F. Cucker. *Condition: The Geometry of Numerical Algorithms*. Number 349 in Grundlehren der matematischen Wissenschaften. Springer Verlag, 2013.

[16] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[17] D. Cheung and F. Cucker. A new condition number for linear programming. *Mathematical Programming*, 91(1):163–174, 2001.

[18] D. Cheung and F. Cucker. Solving linear programs with finite precision: I. Condition numbers and random programs. *Math. Programming*, 99:175–196, 2004.

[19] D. Cheung, F. Cucker, and J. Peña. On strata of degenerate polyhedral cones. I: Condition and distance to stratae. *European Journal of Operational Research*, 198:23–28, 2009.

[20] C. Choi. 7 revealing ways AIs fail. *IEEE Spectrum*, September, 2021.

[21] C. Choi. Some AI systems may be impossible to compute. *IEEE Spectrum*, March, 2022.

[22] M. Colbrook and A. C. Hansen. The foundations of spectral computations via the solvability complexity index hierarchy. *Journal of the European Mathematical Society*, 2022 (online).

[23] M. J. Colbrook, V. Antun, and A. C. Hansen. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and smale's 18th problem. *Proc. Natl. Acad. Sci. USA*, 119(12):e2107151119, 2022.

[24] F. Cucker and S. Smale. Complexity estimates depending on condition and round-off error. *Journal of the ACM*, 46(1):113–184, 1999.

[25] H. David and H. Nagaraja. *Order statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

[26] C. Fefferman, A. C. Hansen, and S. Jitomirskaya, editors. *Computational mathematics in computer assisted proofs*, American Institute of Mathematics Workshops. American Institute of Mathematics, 2022. Available online at `https://aimath.org/pastworkshops/compproofsvrep.pdf`.

[27] L. E. Gazdag and A. C. Hansen. Generalised hardness of approximation and the SCI hierarchy – On determining the boundaries of training algorithms in AI. *arXiv:2209.06715*, 2022.

[28] N. M. Gottschling, V. Antun, A. C. Hansen, and B. Adcock. The troublesome kernel – on hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems. 2023.

[29] T. C. Hales. A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185, 2005.

[30] T. C. Hales and et al. A formal proof of the kepler conjecture. *Forum of Mathematics, Pi*, 5:e2, 2017.

[31] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

[32] A. C. Hansen. On the solvability complexity index, the n-pseudospectrum and approximations of spectra of operators. *Journal of the American Mathematical Society*, 24(1):81–124, 2011.

[33] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, May 2015.

[34] D. Heaven et al. Why deep-learning AIs are so easy to fool. *Nature*, 574(7777):163–166, 2019.

[35] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2002.

[36] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[37] M. Lotz, D. Amelunxen, and J. Walvin. Effective condition number bounds for convex regularization. *IEEE Transactions on Information Theory*, Jan. 2020.

[38] M. T. McCann, K. H. Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Process Magazine*, 34(6):85–95, 2017.

[39] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *IEEE Conference on computer vision and pattern recognition*, pages 86–94, July 2017.

[40] A. Nemirovski. Lectures on Robust Convex Optimization. Available online at `https://www2.isye.gatech.edu/~nemirovs/`, 2009.

[41] Y. E. Nesterov and A. Nemirovski. On first-order algorithms for l1/nuclear norm minimization. *Acta Numer.*, 22:509–575, 2013.

[42] J. Peña. Conditioning of convex programs from a primal-dual perspective. *Mathematics of Operations Research*, 26(2):206–220, 2001.

[43] J. Peña. Two properties of condition numbers for convex programs via implicitly defined barrier functions. *Mathematical Programming*, 93(1):55–75, 2002.

[44] H. R, J. H, and S. M. JI. Robustness and explainability of artificial intelligence. (KJ-NA-30040-EN-N (online)), 2020.

[45] J. Renegar. A polynomial-time algorithm, based on newton's method, for linear programming. *Mathematical Programming*, 40(1-3):59–93, 1988.

[46] J. Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM Journal on Optimization*, 5(3):506–524, 1995.

[47] J. Renegar. Linear programming, complexity theory and elementary functional analysis. *Mathematical Programming*, 70(1):279–351, 1995.

[48] J. Renegar. Condition numbers, the barrier method, and the conjugate-gradient method. *SIAM Journal on Optimization*, 6:879–912, 1996.

[49] J. Renegar. *A mathematical view of interior-point methods in convex optimization*, volume 3. Siam, 2001.

[50] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[51] R. J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7:1456–1490, 2013.

[52] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[53] J. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice Hall, 1963.

[54] S. Wright and B. Recht. *Optimization for Data Analysis*. Cambridge University Press, 2022.

[55] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

UNIVERSITY OF LEICESTER, UK
Email address: ajb177@leicester.ac.uk

CITY UNIVERSITY OF HONG KONG, HONG KONG
Email address: macucker@gmail.com

UNIVERSITY OF CAMBRIDGE, UK
Email address: ach70@cam.ac.uk