Condition-Number-Independent Sparse Linear Regression on Random Supports

Ying Feng
MIT
yingfeng@mit.edu

Abstract

Sparse linear regression is a widely-studied question in statistics and machine learning. Given a design matrix $X \in \mathbb{R}^{N \times d}$ and labels $y = Xw^* + \xi$, where $w^* \in \mathbb{R}^d$ is an unknown vector with at most $k \ll d$ nonzero entries and $\xi \in \mathbb{R}^N$ is measurement noise, the goal is to output a vector $\widehat{w}$ with small prediction error $\varepsilon = \frac{1}{N} \|X(w^* - \widehat{w})\|_2^2$. For this problem, polynomial-time algorithms with sublinear-in-d sample complexity are known only under strong structural assumptions on the design matrix, such as incoherence, the restricted isometry property, or a restricted eigenvalue condition.

Recently, Chandrasekaran, Meka, and Stavropoulos gave the first efficient algorithm for highly ill-conditioned design matrices under a random-support model. When the support of w^* is chosen uniformly at random, their algorithm achieves ε error with $1 - \delta$ probability over the support, using $N = \mathrm{poly}(k, \log d, 1/\varepsilon, 1/\delta, \log\log \kappa)$ samples in polynomial time, where κ is the condition number of X. Their key ingredient is a good preconditioner: after a change of basis, all but a small exceptional set of coordinates have bounded mass. They find such a preconditioner with exceptional-set size $O(k^2(\log\log \kappa)^2/\delta)$ through an iterative convex-optimization procedure.

We give a simple one-shot construction of a good preconditioner with exceptional-set size $O(k^2/\delta)$, subject only to the qualitative nondegeneracy condition $\kappa < \infty$. This size is optimal within the preconditioner framework. Combining with the regression theorem of Chandrasekaran, Meka, and Stavropoulos, this leads to $N = \mathrm{poly}(k, \log d, 1/\varepsilon, 1/\delta)$ sample complexity for ε prediction error, with no quantitative dependence on the condition number.

1 Introduction

Sparse linear regression is one of the most basic problems in high-dimensional statistics. In this problem, we are given a design matrix $X \in \mathbb{R}^{N\times d}$ and labels

$$y = Xw^* + \xi,$$

where $w^* \in \mathbb{R}^d$ is an unknown coefficient vector with at most $k \ll d$ nonzero entries and $\xi \in \mathbb{R}^N$ is measurement noise. The goal is to output a vector $\widehat{w}$ with small prediction error

$$\varepsilon := \frac{1}{N}\,\|X(w^* - \widehat{w})\|_2^2\,.$$

Information-theoretically, for $\xi = O(1)$, prediction error ε is achievable from roughly $k\log(d)/\varepsilon$ measurements. The natural exhaustive-search algorithm over the k-sparse support of w^* attains this statistical rate but takes time about d^k, which remains the state of the arts. At the other extreme, ordinary least squares runs in polynomial time when N is comparable to d. Whether polynomial time and $o(d)$ samples can be achieved for worst-case supports and general design matrices remains a central computational-statistical question; see, for example, [RWY11, Nat95, ZWJ14, FKT15, GV21, GVV24, BDT24].

Most efficient estimators with sublinear sample complexity impose a quantitative assumption on the design matrix X. Classical examples include incoherence, restricted isometry, restricted eigenvalue, and compatibility conditions [Tib96, DS89, CT05, CT07, BRT09, vdGB09]. Preconditioning can substantially enlarge the tractable family: a change of basis may convert an ill-conditioned instance into one on which an ℓ_1-based estimator succeeds [DHL17, FS11, KKMM20, KKMR22, KKMR23, KKMR24]. These results are powerful, but the available preconditioners are usually tied to *structural assumptions on the design matrix*.

Recently, Chandrasekaran, Meka, and Stavropoulos [CMS26] take a different perspective. They consider an arbitrary, potentially highly ill-conditioned matrix and instead assume that the support of w^* is a uniformly random k-subset of $[d]$. Under this model, they show that, for all but a δ fraction of k-sparse supports, prediction error ε can be achieved in polynomial time using $\mathrm{poly}(k, \log d, 1/\varepsilon, 1/\delta, \log\log\kappa_k(X))$ samples, where

$$\kappa_k(X) := \sup_{\substack{w\neq 0\\ \|w\|_0\le k}} \frac{\sqrt{N}\,\|w\|_2}{\|Xw\|_2}$$

is the k-sparse condition number of X. Their central abstraction is a good preconditioner. After a basis transformation $Z = XB^\top$, the transformed coefficient vector $u^* = B^{-\top}w^*$ has bounded ℓ_1-norm outside a known exceptional set I. Regression can then be performed by leaving the coordinates in I unconstrained and imposing an ℓ_1 constraint on the remaining coordinates.

To construct such a preconditioner, [CMS26] uses an iterative sampling-and-update procedure. Starting from a bound $\tau = \kappa_k(X)$ on the nonexceptional coefficients, given by the identity basis, each improvement step improves the current bound τ to $\sqrt{\tau}$, while enlarging the exceptional set. Repeating this improvement $\Theta(\log\log\kappa_k(X))$ times reduces the coefficient bound to a constant and yields an exceptional set of size

$$O\left(\frac{k^2(\log\log\kappa_k(X))^2}{\delta}\right).$$

Thus, despite applying to highly ill-conditioned designs, the resulting prediction and sample-complexity bounds retain a quantitative dependence on $\kappa_k(X)$. The preconditioner construction is also technically involved: each improvement stage itself requires repeated sampling of candidate sets, coordinate-wise convex optimization, and basis updates. Our construction addresses both aspects: it replaces the iterative procedure with a one-shot argument and removes the condition-number dependence. We discuss the relationship between the two constructions in more detail in Section 1.2.

1.1 Our results

We give a one-shot good preconditioner whose exceptional-set size is $O(k^2/\delta)$. Throughout this paper, we assume $\kappa_k(X) < \infty$, equivalently, the nullspace of X does not contain nonzero k-sparse vector. By rescaling the columns and applying the inverse rescaling to the corresponding coefficients, we may assume without loss of generality that

$$\max_{j\in[d]} \left\|X^{(j)}\right\|_2 \le \sqrt{N}$$

where $X^{(j)}$ denotes the j-th column. This transformation preserves the prediction vector Xw^*, the support of w^*, and the qualitative guarantee $\kappa_k(X) < \infty$.

Our main regression theorem is the following.

Theorem 1.1. *Let $X \in \mathbb{R}^{N\times d}$ satisfy $\max_{j\in[d]} \left\|X^{(j)}\right\|_2 \le \sqrt{N}$ and $\kappa_k(X) < \infty$, and let $2 \le k \le d$ and $\delta, \gamma \in (0, 1/2)$. With probability at least 0.9 over the randomness of the algorithm, there is a family*

$$\mathcal{G} \subseteq \binom{[d]}{k}, \qquad |\mathcal{G}| \ge (1-\delta)\binom{d}{k},$$

such that the following holds for every $S \in \mathcal{G}$ and every w^* supported on S with $\|Xw^*\|_2 \le \sqrt{N}$.

Given labels $y = Xw^* + \xi$, where ξ is a mean-zero σ-sub-gaussian random vector independent of X, the regression algorithm outputs $\widehat{w} \in \mathbb{R}^d$ satisfying

$$\frac{1}{N}\|X(w^* - \widehat{w})\|_2^2 \le \frac{C\sigma k}{\sqrt{N}}\left(\frac{k}{\sqrt{\delta}} + \sqrt{\log(d/\gamma)}\right) \tag{1}$$

with probability at least $1-\gamma$ over the noise. Here $C > 0$ is a universal constant.

The preconditioner construction uses $O\left(\frac{Ndk^2}{\delta}\right)$ real-arithmetic operations. The regression step is a convex program and runs in polynomial time[1].

$\|Xw^*\|_2 \le \sqrt{N}$ is a normalization condition that can be replaced by a general radius parameter at the cost of a corresponding scaling of the error bound.

As an immediate consequence of theorem 1.1, an ε prediction error is achieved once

$$N \ge C\frac{\sigma^2 k^2}{\varepsilon^2}\left(\frac{k^2}{\delta} + \log(2d/\gamma)\right). \tag{2}$$

The preconditioner theorem underlying theorem 1.1 may be useful independently:

[1] Concretely, in the real-arithmetic model, an η-accurate solution of the regression step can be computed in $\mathrm{poly}(N, d, \log(1/\eta))$ arithmetic operations

Theorem 1.2 (One-shot good preconditioner, informal)**.** *For $X \in \mathbb{R}^{N\times d}$ satisfying $\max_{j\in[d]} \left\|X^{(j)}\right\|_2 \le \sqrt{N}$ and $\kappa_k(X) < \infty$, and $\delta \in (0,1)$, a collection of $\Theta(k/\delta)$ random $(k-1)$-subsets defines, with probability at least 0.9, a good preconditioner with an exceptional set of size $O(k^2/\delta)$.*

The formal definitions and proofs appear in Sections 2 to 4.

The previous lower bound of Chandrasekaran, Meka, and Stavropoulos shows that $\Omega(k^2/\delta)$ exceptional coordinates are necessary for the good-preconditioner abstraction, even for correlated Gaussian designs [CMS26]. Thus our construction is optimal up to constants within this framework.

1.2 Intuition behind theorem 1.2

We now describe the main idea. As mentioned in the introduction, we want to find a change-of-basis matrix and a small exceptional set I, such that after the change of basis, the coefficient vector has bounded mass outside I.

Concretely, let $B \in \mathbb{R}^{d\times d}$ be invertible, let $I \subseteq [d]$, and define $Z = XB^\top$. A good preconditioner bounds $\|Z^{(j)}\|_2 \le \sqrt{N}$, has a small exceptional set I, and satisfies the following property for most random supports S:

$$\forall w \text{ supported on } S: \quad \left\|(B^{-\top}w)_{[d]\setminus I}\right\|_\infty \le \frac{1}{\sqrt{N}} \|Xw\|_2 . \tag{3}$$

Our construction is quite simple: we sample $R = \Theta(k/\delta)$ independent sets $T_1, \dots, T_R \in \binom{[d]}{k-1}$ and let $I = \bigcup_{r=1}^R T_r$. Our exceptional set I is composed of a collection of "comparison sets" $T_1, \dots, T_R$, which we will use to compare against a random support.

For a column index $j \in [d]$, define

$$\Delta_j := \min_{r\in[R]} \operatorname{dist}\left(X^{(j)}, \operatorname{colspan}(X_{T_r})\right)$$

and

$$r(j) := \arg\min_{r\in[R]} \operatorname{dist}\left(X^{(j)}, \operatorname{colspan}(X_{T_r})\right),$$

i.e., $T_{r(j)}$ is the comparison set that minimizes the residual. We construct Z as follows: for exceptional $j \in I$, we define $Z^{(j)} = X^{(j)}$. And for non-exceptional $j \notin I$, we define the new j-th column to be the normalized residual

$$Z^{(j)} := \sqrt{N}\frac{X^{(j)} - \operatorname{Proj}_{\operatorname{colspan}(X_{T_{r(j)}})} X^{(j)}}{\Delta_j}.$$

This construction corresponds to a change-of-basis matrix $B^{-\top}$ whose j-th row is $(\Delta_j/\sqrt{N})e_j^\top$ for every $j \notin I$. Therefore,

$$\forall j \notin I: \quad (B^{-\top}w)_j = (\Delta_j/\sqrt{N})w_j. \tag{4}$$

In this way, the magnitude of the non-exceptional coordinate is determined by Δ_j, the distance of the column to the closest comparison set.

Comparing distance with a random support. Now suppose we sample a random support $S \in \binom{[d]}{k}$. This is equivalent to sampling $j \in [d]$ uniformly, sampling $T_0 \in \binom{[d]\setminus\{j\}}{k-1}$ uniformly, and set $S = T_0 \cup \{j\}$. Conditioning on $j \notin I$, we make the crucial observation that

$$T_0, T_1, \ldots, T_R$$

are *exchangeable* uniform subsets of $[d] \setminus \{j\}$. Therefore, focusing on a column $X^{(j)}$, the chance that T_0 has a smaller j-residual than every comparison set is at most $1/(R+1)$. Thus, a union bound over k coordinates of S shows that with high probability over the sampling of $T_1, \cdots, T_R$, for at least a $1-\delta$ fraction of supports S,

$$\forall j \in S \setminus I: \quad \Delta_j \le \operatorname{dist}\left(X^{(j)}, \operatorname{colspan}\left(X_{S\setminus\{j\}}\right)\right). \tag{5}$$

In other words, $X^{(j)}$ is closer to one of the comparison spans than to the span induced by the random support.

Distance to a bounded mass. Fix a support satisfying Equation (5) and a vector w supported on S. For every coordinate $j \in [d]$, we can decompose

$$Xw = w_j X^{(j)} + X_{S\setminus\{j\}} w_{S\setminus\{j\}};$$

taking the distance to $\operatorname{colspan}\left(X_{S\setminus\{j\}}\right)$ gives

$$\|Xw\|_2 \ge |w_j| \cdot \operatorname{dist}\left(X^{(j)}, \operatorname{colspan}\left(X_{S\setminus\{j\}}\right)\right).$$

Combining this with Equations (4) and (5) gives

$$\left|(B^{-\top} w)_j\right| = (\Delta_j/\sqrt{N})\,|w_j| \le \frac{1}{\sqrt{N}}\,\|Xw\|_2 .$$

This proves Equation (3) simultaneously for every vector w supported on any support S satisfying Equation (5).

Comparison with [CMS26]. We note that our use of random $(k-1)$-subsets is directly motivated by the improvement step of [CMS26]. In their construction, a random set $T \in \binom{[d]}{k-1}$ is sampled, and $\operatorname{colspan}(X_T)$ is used in a convex program to identify coordinate-wise updates for indices $j \notin T$. Their analysis uses the viewpoint $S = T_0 \cup \{j\}$ in an averaging argument to show that, whenever the current preconditioner is not sufficiently good, a random set T has nonnegligible probability of exposing many coordinates that can be improved. The algorithm repeatedly samples T until it finds one that exposes sufficiently many such coordinates; it then adds this T to the exceptional set and performs the corresponding row updates. The improvement procedure is subsequently repeated over successively smaller coefficient scales.

Here, we instead sample the entire collection $T_1, \ldots, T_R$ in advance, let each coordinate choose its closest comparison span, and compare $T_0 = S \setminus \{j\}$ directly with the closet comparison set using the exchangeability argument. This one-shot comparison eliminates the adaptive selection and repeated improvement rounds, and yields a construction with no quantitative dependence on the condition number.

2 Prelimiaries

Notations. For any integer $n \in \mathbb{Z}_{>0}$, we use $[n]$ to denote $\{1, \ldots, n\}$. For $T \subseteq [d]$, X_T is the submatrix of X consisting of columns indexed by T, and w_T is the corresponding subvector. We use $X^{(j)}$ for the j-th column of X and B_j for the j-th row of a matrix B. For $I \subseteq [d]$, we write $\overline{I} = [d] \setminus I$. We use $S \sim \binom{[d]}{k}$ for a uniformly random k-subset of $[d]$. The support of a vector w is denoted by $\mathrm{supp}(w)$. For $w \in \mathbb{R}^d$ and a fixed design matrix X, we define

$$\|w\|_X := \frac{1}{\sqrt{N}} \|Xw\|_2 .$$

A random vector $\xi \in \mathbb{R}^N$ is σ-sub-gaussian if, for every $v \in \mathbb{R}^N$ with $\|v\|_2 = 1$ and every $t \geq 0$,

$$\Pr\left[|\langle v, \xi \rangle| > t\right] \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

2.1 Good preconditioners

We now give the formal definition of good preconditioners. The following definition is an almost equivalent rewriting of the original definition of good preconditioners in [CMS26]. We will describe how this implies definition 3.1 in [CMS26] when we use such a preconditioner for regression in Section 4.

Definition 2.1 (Good preconditioner)**.** Let $X \in \mathbb{R}^{N \times d}$ satisfy $\max_{j \in [d]} \left\|X^{(j)}\right\|_2 \leq \sqrt{N}$ and $\kappa_k(X) < \infty$. A pair (B, I), where $B \in \mathbb{R}^{d \times d}$ is an invertible conditioning matrix and $I \subseteq [d]$ is an exceptional set-, is an (ℓ, k, δ)-good preconditioner for X if the following conditions hold.

(i) For $Z := XB^\top$,

$$\max_{j \in [d]} \left\|Z^{(j)}\right\|_2 \leq \sqrt{N}.$$

(ii) The exceptional set has size $|I| \leq \ell$, and

$$\mathrm{supp}\big((B^{-1})_j\big) \subseteq I \cup \{j\} \qquad \text{for every } j \in [d].$$

(iii) With probability at least $1 - \delta$ over $S \sim \binom{[d]}{k}$, the following holds simultaneously for every $w \in \mathbb{R}^d$ supported on S:

$$\left\|(B^{-\top} w)_{\overline{I}}\right\|_\infty \leq \|w\|_X . \tag{6}$$

3 The one-shot good preconditioner

We now restate our construction and prove theorem 1.2 (formally stated as theorem 3.1).

For $j \in [d]$ and $T \subseteq [d] \setminus \{j\}$, define

$$\Delta_j(T) := \mathrm{dist}\left(X^{(j)}, \mathrm{colspan}(X_T)\right) = \min_{a \in \mathbb{R}^T} \left\|X^{(j)} - X_T a\right\|_2 .$$

We note that under the assumption $\kappa_k(X) < \infty$, for every T with $|T| \leq k - 1$ and every $j \notin T$, one has $\Delta_j(T) > 0$. Geometrically, $X^{(j)}$ has a nonzero component outside the span of C_T. This rules

Algorithm 1 RandomPreconditioner(X, k, δ)

Set $R \leftarrow \lceil 10k/\delta \rceil$.
Sample independently $T_1, \dots, T_R \sim \binom{[d]}{k-1}$.
Set $I \leftarrow \bigcup_{r=1}^{R} T_r$.
for $j \in I$ **do**
    Set $B_j \leftarrow e_j^\top$.
for $j \in [d] \setminus I$ **do**
    Choose $r(j) \in \arg\min_{r \in [R]} \Delta_j(T_r)$.
    Let $a^{(j)} \in \mathbb{R}^d$ be supported on $T_{r(j)}$ and satisfy

$$X a^{(j)} = \mathrm{Proj}_{\mathrm{colspan}(X_{T_{r(j)}})} X^{(j)}.$$

    Set

$$\rho_j \leftarrow \frac{\Delta_j(T_{r(j)})}{\sqrt{N}}, \qquad B_j \leftarrow \frac{e_j^\top - (a^{(j)})^\top}{\rho_j}.$$

return (B, I).

out a basic obstruction of controlling $\|w\|_\infty$ for k-sparse w by the prediction norm $\|w\|_X$: otherwise there would be a nonzero k-sparse null vector, which could be scaled arbitrarily without changing its prediction norm. In particular, all residual normalizations used below are well-defined.

The construction can be interpreted as follows. The sampled sets form a collection of comparison subspaces. Their union is the set of coordinates on which the final regression algorithm is allowed to place arbitrary coefficients, so we retain corresponding columns as it. Every other column is residualized against its closet subspace and normalized to have length $\sqrt{N}$.

Theorem 3.1. *Let $X \in \mathbb{R}^{N \times d}$ satisfy $\max_{j \in [d]} \left\|X^{(j)}\right\|_2 \le \sqrt{N}$ and $\kappa_k(X) < \infty$, and let $2 \le k \le d$ and $\delta \in (0, 1)$. With probability at least 0.9, algorithm 1 outputs an (ℓ, k, δ)-good preconditioner for X with $\ell = O(\frac{k^2}{\delta})$.*

Moreover, algorithm 1 can be implemented using $O(\frac{Ndk^2}{\delta})$ real-arithmetic operations.

We divide the proof into the structural properties of B ((i) and (ii) of definition 2.1), and the control over the non-exceptional mass ((iii) of definition 2.1).

3.1 Structural properties (i) and (ii)

Lemma 3.2 (Positivity of the residuals)**.** *For every $j \notin I$, one has $\rho_j > 0$.*

Proof. For every $r \in [R]$, $j \notin T_r$ because $j \notin I$. If $\Delta_j(T_r) = 0$, then $X^{(j)} \in \mathrm{colspan}(X_{T_r})$, giving a nonzero vector in the nullspace of X supported on $T_r \cup \{j\}$, a set of size k. This contradicts the assumption $\kappa_k(X) < \infty$.

□

Lemma 3.3 (Structural properties)**.** *Let (B, I) be the output of algorithm 1 and let $Z = XB^\top$. Then the following hold deterministically:*

(i) $\left\|Z^{(j)}\right\|_2 \le \sqrt{N}$ *for every $j \in [d]$.*

(ii) *B is invertible and*

$$\operatorname{supp}\big((B^{-1})_j\big) \subseteq I \cup \{j\} \qquad \text{for every } j \in [d].$$

Proof. If $j \in I$, then $B_j = e_j^\top$ and hence $Z^{(j)} = X^{(j)}$, whose norm is at most $\sqrt{N}$. If $j \notin I$, then

$$Z^{(j)} = XB_j^\top = \frac{X^{(j)} - Xa^{(j)}}{\rho_j}.$$

The numerator is the residual of $X^{(j)}$ from $\operatorname{colspan}(X_{T_{r(j)}})$, and has norm $\rho_j\sqrt{N}$. Thus $\left\|Z^{(j)}\right\|_2 = \sqrt{N}$. This proves part (i).

For (ii), we start with permuting the coordinates so that coordinates in I comes first. Because $a^{(j)}$ is supported on $T_{r(j)} \subseteq I$, the matrix B has block form

$$B = \begin{pmatrix} \mathrm{I}_{|I|} & 0 \\ A & D \end{pmatrix},$$

where D is diagonal and $D_{jj} = 1/\rho_j$ for $j \notin I$, and A denotes the matrix recording the coefficients used to subtract the projections onto the comparison spans. By lemma 3.2, D is invertible. Therefore B is invertible and

$$B^{-1} = \begin{pmatrix} \mathrm{I}_{|I|} & 0 \\ -D^{-1}A & D^{-1} \end{pmatrix}. \tag{7}$$

This proves part (ii). □

Since I is a union of R sets of size $k-1$, the size bound $|I| = O(\frac{k^2}{\delta})$ holds deterministically. It remains to bound the mass of non-exceptional coordinates for most random supports.

3.2 Bounding the non-exceptional mass (iii)

We first state the elementary symmetry fact used in the proof.

Claim 3.4. *Let $D_0, D_1, \ldots, D_R$ be exchangeable real-valued random variables. Then*

$$\Pr\left[D_0 < \min_{r\in[R]} D_r\right] \le \frac{1}{R+1}.$$

Proof. For each $i \in \{0, \ldots, R\}$, let E_i be the event that D_i is strictly smaller than every other variable. Exchangeability implies that the events E_i have the same probability. They are also disjoint, thus the claim follows. □

Fix a realization of the samples $T_1, \ldots, T_R$. We say a support $S \in \binom{[d]}{k}$ is *covered* if

$$\min_{r\in[R]} \Delta_j(T_r) \le \Delta_j(S \setminus \{j\}) \qquad \text{for every } j \in S \setminus I.$$

Lemma 3.5 (Most supports are covered)**.** *With probability at least* 0.9 *over the sampled sets $T_1, \ldots, T_R$,*

$$\Pr_{S\sim\binom{[d]}{k}} [S \text{ is covered}] \ge 1 - \delta. \tag{8}$$

Proof. As outlined in Section 1.2, we sample $S \sim \binom{[d]}{k}$ and then sample j uniformly from S. This is equivalent to sample j uniformly from $[d]$, sample

$$T_0 \sim \binom{[d] \setminus \{j\}}{k-1},$$

and set $S = T_0 \cup \{j\}$.

We condition on $j \notin I$, which is the intersection of the independent events $j \notin T_r$ for $r \in [R]$. Under this conditioning, the sets $T_1, \ldots, T_R$ remain independent and each is uniform in $\binom{[d]\setminus\{J\}}{k-1}$. They are also independent of T_0. Therefore

$$T_0, T_1, \ldots, T_R$$

are exchangeable. Hence by claim 3.4,

$$\Pr\left[\min_{r\in[R]} \Delta_j(T_r) > \Delta_j(T_0) \,\middle|\, j \notin I\right] \le \frac{1}{R+1}. \tag{9}$$

If a support S is not covered, then at least one $j \in S \setminus I$ witnesses the bad event

$$\mathcal{E} := \left\{j \in S \setminus I \text{ and } \min_{r\in[R]} \Delta_j(T_r) > \Delta_j(T_0)\right\}.$$

Since j is uniform in S,

$$\mathbf{1}[S \text{ is not covered}] \le k \Pr_j\left[\mathcal{E} \mid S, \{T_1, \cdots, T_R\}\right].$$

where $\mathbf{1}[\cdot]$ denotes the indicator random variable. Taking expectation and applying Equation (9), we get that

$$\mathbb{E}_{\{T_1,\cdots,T_R\}} \Pr_S[S \text{ is not covered}] \le \frac{k}{R+1} \le \frac{\delta}{10}.$$

By Markov's inequality, with probability at least 0.9 over $\{T_1, \cdots, T_R\}$,

$$\Pr_S[S \text{ is not covered}] \le \delta,$$

which is exactly Equation (8). □

The viewpoint $S = T_0 \cup \{j\}$ follows [CMS26, Lemmas 3.5, 3.6, and 4.9]; here it is use with a exchangeability argument to compare T_0 directly with an independent collection of comparison sets.

The property of being covered can be translated into a bound connecting the coordinate's magnitude with the distance to the comparison span.

Lemma 3.6 (Bounding the the non-exceptional mass)**.** *Let $S \in \binom{[d]}{k}$ be covered. Then, for every $w \in \mathbb{R}^d$ supported on S and every $j \in S \setminus I$,*

$$\rho_j \left|w_j\right| \le \|w\|_X .$$

Proof. Fix $j \in S \setminus I$. We decompose

$$Xw = w_j X^{(j)} + X_{S\setminus\{j\}} w_{S\setminus\{j\}}.$$

The second term lies in the subspace $\mathrm{colspan}(X_{S\setminus\{j\}})$. Taking the distance of both sides to $\mathrm{colspan}(X_{S\setminus\{j\}})$ gives

$$\|Xw\|_2 \geq \left|w_j\right| \Delta_j(S \setminus \{j\}). \tag{10}$$

Since S is covered,

$$\rho_j \sqrt{N} = \min_{r\in[R]} \Delta_j(T_r) \leq \Delta_j(S \setminus \{j\}).$$

Combining this inequality with Equation (10) and dividing by $\sqrt{N}$ proves the lemma. □

Proof of theorem 3.1. By lemma 3.3, conditions (i) and (ii) of definition 2.1 hold deterministically, and $|I| = O(\frac{k^2}{\delta})$.

By lemma 3.5, with probability at least 0.9 over the sampling, a $1-\delta$ fraction of supports are covered. Fix such a covered support S. Let w be supported on S. We can observe from Equation (7) that for $j \notin I$, column j of B^{-1} is zero except at its diagonal entry, which equals ρ_j. Thus

$$(B^{-\top}w)_j = (B^{-1}e_j)^\top w = \rho_j w_j.$$

For $j \notin S \cup I$, this gives $(B^{-\top}w)_j = 0$. For $j \in S \setminus I$, combining this with lemma 3.6 gives

$$\left|(B^{-\top}w)_j\right| = \rho_j \left|w_j\right| \leq \|w\|_X .$$

Hence

$$\left\|(B^{-\top}w)_{\bar{I}}\right\|_\infty \leq \|w\|_X .$$

This is condition (iii) of definition 2.1.

□

3.3 Implementation and arithmetic complexity

The construction is particularly simple to implement. For each $r \in [R]$, we compute a (reduced) QR factorization $X_{T_r} = Q_r U_r$, where $Q_r \in \mathbb{R}^{N\times(k-1)}$ has orthonormal columns and $U_r \in \mathbb{R}^{(k-1)\times(k-1)}$ is upper triangular. U_r is also invertible because $\kappa_k(X) < \infty$. Then, for every $j \in [d]$, we can compute

$$\Delta_j(T_r)^2 = \left\|X^{(j)}\right\|_2^2 - \left\|(Q_r^\top X)^{(j)}\right\|_2^2.$$

Moreover, let $c^{(j,r)} \in \mathbb{R}^{k-1}$ be the unique solution of

$$U_r c^{(j,r)} = (Q_r^\top X)^{(j)}.$$

Then

$$X_{T_r} c^{(j,r)} = Q_r U_r c^{(j,r)} = Q_r Q_r^\top X^{(j)} = \mathrm{Proj}_{\mathrm{colspan}(X_{T_r})} X^{(j)}.$$

Thus, extending $c^{(j,r)}$ by zero outside T_r gives the coefficient vector representing this projection. In algorithm 1, while iterating over the loop (Line 6), we retain the smallest residual encountered and its corresponding coefficient vector. At termination, these are $\Delta_j(T_{r(j)})$ and $a^{(j)}$, respectively.

For a fixed r, computing the QR factorization, the product $Q_r^\top X$, and solving the d triangular systems $U_r c^{(j,r)} = (Q_r^\top X)^{(j)}$ takes

$$O(Nk^2 + Ndk + dk^2) = O(Ndk)$$

real-arithmetic operations. Since $R = O(k/\delta)$, the total arithmetic complexity is $O\left(\frac{Ndk^2}{\delta}\right)$.

4 Regression from a good preconditioner

We use the regression theorem of Chandrasekaran, Meka, and Stavropoulos [CMS26, Theorems 3.4 and 5.1]. We first verify that our reformulation of the good-preconditioner definition 2.1 implies the corresponding definition 3.1 in [CMS26].

Claim 4.1. *Let (B, I) be an (ℓ, k, δ)-good preconditioner for X as in definition 2.1. Let D be the uniform distribution over the rows of the design matrix X. Then (B, I) is an (ℓ, k, δ)-good preconditioner for D as in definition 3.1 in [CMS26]. In particular, the following holds:*

(a) $\max_{j\in[d]} \mathbb{E}_{x\sim D}[(Bx)_j^2] \le 1$.

(b) *The exceptional set has size $|I| \le \ell$, and*

$$\operatorname{supp}\big((B^{-1})_j\big) \subseteq I \cup \{j\} \qquad \text{for every } j \in [d].$$

(c) *With probability at least $1-\delta$ over $S \sim \binom{[d]}{k}$, for all $w \in \mathbb{R}^d$ with $\operatorname{supp}(w) \subseteq S$ and $\mathbb{E}_{x\sim D}[(w\cdot x)^2] \le 1$, it holds that*

$$\left\|(B^{-\top}w)_{\overline{I}}\right\|_\infty \le O(1).$$

Proof. For (a), let $Z = XB^\top$. Since D is the uniform distribution over the rows of X, for every $j \in [d]$,

$$\mathbb{E}_{x\sim D}\left[(Bx)_j^2\right] = \frac{1}{N}\left\|XB_j^\top\right\|_2^2 = \frac{1}{N}\left\|Z^{(j)}\right\|_2^2.$$

By condition (i) of definition 2.1, $\left\|Z^{(j)}\right\|_2 \le \sqrt{N}$ for every $j \in [d]$, which implies (a).

Condition (b) is exactly condition (ii) in definition 2.1. It remains to verify (c). By condition (iii) of definition 2.1, with probability $1-\delta$ over $S \sim \binom{[d]}{k}$, simultaneously for every $w \in \mathbb{R}^d$ supported on S,

$$\left\|(B^{-\top}w)_{\overline{I}}\right\|_\infty \le \|w\|_X.$$

Fix a support S for which this guarantee holds, and let w additionally satisfy $\mathbb{E}_{x\sim D}\left[(w\cdot x)^2\right] \le 1$. Since D is uniform over the rows of X,

$$\mathbb{E}_{x\sim D}\left[(w\cdot x)^2\right] = \frac{1}{N}\|Xw\|_2^2 = \|w\|_X^2.$$

Hence $\|w\|_X \le 1$, and therefore

$$\left\|(B^{-\top}w)_{\overline{I}}\right\|_\infty \le \|w\|_X \le 1.$$

This proves (c) and completes the proof of the claim. □

We now restate the regression theorem of [CMS26] in our notation. This immediately leads to our main result.

Theorem 4.2 (Regression from a good preconditioner, theorem 3.4, 5.1 of [CMS26])**.** *Let $X \in \mathbb{R}^{N\times d}$ satisfy $\max_{j\in[d]} \left\|X^{(j)}\right\|_2 \le \sqrt{N}$, and let (B, I) be an (ℓ, k, δ)-good preconditioner for X. There is a polynomial-time algorithm with the following guarantee.*

For every support S such that definition 2.1 (iii) holds and every w^* supported on S with $\|Xw^*\|_2 \le \sqrt{N}$, given

$$y = Xw^* + \xi,$$

where ξ is a mean-zero σ-sub-gaussian random vector independent of X, the algorithm outputs $\widehat{w} \in \mathbb{R}^d$ such that, with probability at least $1-\gamma$ over the noise,

$$\frac{1}{N}\|X(\widehat{w}-w^*)\|_2^2 \le \frac{C\sigma k}{\sqrt{N}}\left(\sqrt{\ell}+\sqrt{\log(d/\gamma)}\right). \tag{11}$$

Proof of theorem 1.1. This is a direct corollary of theorem 3.1,claim 4.1, and theorem 4.2. □

5 Conclusions

We gave a one-shot, condition-number-independent construction of good preconditioners for sparse linear regression on random supports. The construction is based on a collection of random comparison spans. The resulting exceptional set has size $\Theta(k^2/\delta)$, matching the known lower bound for the framework. As a result, we obtain a sparse linear regression algorithm on random supports, with a sublinear-in-d, condition-number-independent sample complexity and polynomial runtime.

References

[BDT24] Rares-Darius Buhai, Jingqiu Ding, and Stefan Tiegel. Computational-statistical gaps for improper learning in sparse linear regression. In *Proceedings of the 37th Conference on Learning Theory*, pages 752–771, 2024.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[CMS26] Gautam Chandrasekaran, Raghu Meka, and Konstantinos Stavropoulos. Sparse linear regression is easy on random supports. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*, pages 2242–2253, 2026.

[CT05] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[CT07] Emmanuel J. Candès and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[DHL17] Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

[DS89] David L. Donoho and Philip B. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[FKT15] Dean Foster, Howard Karloff, and Justin Thaler. Variable selection is hard. In *Proceedings of the 28th Conference on Learning Theory*, pages 696–709, 2015.

[FS11] Rina Foygel and Nathan Srebro. Fast-rate and optimistic-rate error bounds for ℓ_1-regularized regression. *CoRR*, abs/1108.0373, 2011.

[GV21] Aparna Gupte and Vinod Vaikuntanathan. The fine-grained hardness of sparse linear regression. *CoRR*, abs/2106.03131, 2021.

[GVV24] Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Sparse linear regression and lattice problems. In *Theory of Cryptography Conference*, pages 276–307, 2024.

[KKMM20] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Ankur Moitra. Learning some popular gaussian graphical models without condition number bounds. In *Advances in Neural Information Processing Systems 33*, 2020.

[KKMR22] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. In *Proceedings of the 62nd IEEE Symposium on Foundations of Computer Science*, pages 550–561, 2022.

[KKMR23] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Feature adaptation for sparse linear regression. In *Advances in Neural Information Processing Systems 36*, pages 29951–29995, 2023.

[KKMR24] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Lasso with latents: Efficient estimation, covariate rescaling, and computational-statistical gaps. In *Proceedings of the 37th Conference on Learning Theory*, pages 2840–2886, 2024.

[Nat95] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[RWY11] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[vdGB09] Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[ZWJ14] Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of the 27th Conference on Learning Theory*, pages 921–948, 2014.