Lower Bounds on Randomly Preconditioned Lasso via Robust Sparse Designs

Jonathan A. Kelner[*] MIT

Frederic Koehler[†] Stanford

Raghu Meka[‡] UCLA

Dhruv Rohatgi[§] MIT

Abstract

Sparse linear regression with ill-conditioned Gaussian random covariates is widely believed to exhibit a statistical/computational gap, but there is surprisingly little formal evidence for this belief. Recent work has shown that, for certain covariance matrices, the broad class of Preconditioned Lasso programs provably cannot succeed on polylogarithmically sparse signals with a sublinear number of samples. However, this lower bound only holds against deterministic preconditioners, and in many contexts randomization is crucial to the success of preconditioners. We prove a stronger lower bound that rules out randomized preconditioners. For an appropriate covariance matrix, we construct a single signal distribution on which any invertibly-preconditioned Lasso program fails with high probability, unless it receives a linear number of samples. Surprisingly, at the heart of our lower bound is a new *robustness* result in compressed sensing. In particular, we study recovering a sparse signal when a few measurements can be erased adversarially. To our knowledge, this natural question has not been studied before for *sparse* measurements. We surprisingly show that standard sparse Bernoulli measurements are almost-optimally robust to adversarial erasures: if b measurements are erased, then all but $O(b)$ of the coordinates of the signal are identifiable.

1 Introduction

Random-design sparse linear regression (SLR) is a fundamental problem in high-dimensional statistics and learning theory. The simplest formulation of this problem is the following: given independent covariates $X_1, \dots, X_m$ drawn from an n-variable Gaussian distribution with zero mean and positive-definite covariance Σ, and responses $y_i = \langle X_i, w^* \rangle$ for some unknown k-sparse signal w^* (i.e. with at most k nonzero entries), the goal is to recover w^*. While there are more complex models (e.g. with noise or non-Gaussian distributions), we are proving *lower bounds*, so the simplified model only makes our results stronger.

Information-theoretically, for *any* positive-definite covariance Σ, it's possible to recover w^* exactly from only $O(k \log n)$ samples (X_i, y_i). (Again, this is a special case of more general results for the noisy setting.) However, the algorithm which achieves this sample complexity is computationally inefficient — it has time complexity $O(n^k)$. Significant effort has gone into *polynomial-time*

[*] kelner@mit.edu. This work was supported in part by NSF Large CCF-1565235, NSF Medium CCF-1955217, and NSF TRIPODS 1740751.

[†] fkoehler@stanford.edu. This work was supported in part by NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, A. Moitra's ONR Young Investigator Award, and E. Mossel's Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826.

[‡] raghum@cs.ucla.edu. This work was supported in part by NSF CAREER Award CCF-1553605 and NSF Small CCF-2007682.

[§] drohatgi@mit.edu. This work was supported in part by an Akamai Presidential Fellowship and a U.S. DoD NDSEG Fellowship.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

sample-efficient algorithms for sparse linear regression, under as weak conditions on the covariates as possible [22, 16, 44, 15, 13, 21, 20, 14, 18, 12, 45, 53, 48, 38, 47, 11, 46]. Essentially all of these algorithms are variations on one of two fundamental methods: ℓ_1 regularization (most notably, the Lasso program [44]), and greedy variable selection (most notably, Orthogonal Matching Pursuit [45]). Moreover, both of these methods seem to hit the same fundamental barrier: they require that Σ is in some sense *well-conditioned*. The most basic assumption enabling efficient algorithms is a bound on the condition number $\lambda_{\text{max}}(\Sigma)/\lambda_{\text{min}}(\Sigma)$, and this assumption has been weakened (to e.g. the Restricted Isometry Property [15], the Compatibility Condition [48], or Weak Submodularity [18]) but not eliminated. This barrier, and the lack of algorithmic techniques which evade it, suggests that sparse linear regression with general positive-definite Σ may exhibit a computational/statistical gap: that is, sample-efficient signal recovery may be computationally hard.

Preconditioned Lasso. Recent work [32] provided the first computationally-efficient and sample-efficient algorithm for a broad class of *ill-conditioned* covariance matrices Σ (specifically, covariance matrices of Gaussian Graphical Models with low treewidth such as Gaussian time-series data). The algorithm was a generalization of Lasso via an initial *preconditioning* step; while preconditioning has been enormously impactful for efficiently solving linear systems [39], it has only recently been applied to sparse linear regression. The algorithmic results of [32] raise an obvious question: *can preconditioning close the statistical/computational gap for ill-conditioned sparse linear regression? Or is the gap inherent?*

The answer is unclear. The basic obstacle is that no "candidate" hard instances of random-design sparse linear regression are known: nearly all examples on which Lasso provably fails can be trivially fixed by preconditioning [53, 23, 17, 46, 31]. The exception is the lower bound in [32], which provides a limited negative answer to the first question: there is a covariance matrix Σ such that any *deterministic* preconditioner fails (see Section 1.1 for a concrete statement). They showed that for each preconditioner, there exists a specific bad signal on which it fails, where the signal chosen depends heavily on the preconditioner.

This suggests the tantalizing possibility that any ill-conditioned SLR problem can be solved by applying *randomized* preconditioning. This way, the signal can no longer be chosen adversarially based on the preconditioner. Indeed, numerical linear algebra is full of situations where randomized maps are useful precisely for this reason, including the JL embedding and other dimension reduction methods, algorithms based on random metric embeddings, and the famous Nystrom method for low rank approximation and preconditioning (see e.g. [56, 51, 42]). Also, it was observed in [31, 32] that randomized preconditioners (or, roughly equivalently, trying a class of preconditioners) *can* be used to solve SLR problems in various cases where we do not currently know of a fixed deterministic preconditioner that works. Given this context, we need to ask the following question: can *randomized* preconditioning close the statistical/computational gap?

Our result. We give a near-complete negative answer to this revised question: for a carefully-chosen Σ and distribution over the unknown signal w^*, any randomly-preconditioned Lasso program with invertible preconditioners fails (see Section 2 for a more concrete statement).

Broader context on statistical/computational gaps. Random-design sparse linear regression is very interesting among high-dimensional statistics problems, in that there is essentially no evidence for its conjectured statistical/computational gap; while SLR with worst-case covariates (i.e. not drawn from a distribution) has been proven computationally hard under standard worst-case complexity-theoretic assumptions [37, 57, 30, 28], no such hardness is known for random-design SLR.[5] In contrast, problems as diverse as sparse PCA [7], average-case RIP certification [55], planted dense subgraph [10], robust sparse mean estimation [8], and negative-spike sparse PCA [8] have provably statistical/computational gaps under (variants of) the Planted Clique conjecture.

Why isn't sparse linear regression on this list? One possibility is that there is no statistical/computational gap; in this case, our results imply that new algorithmic techniques beyond preconditioning are needed, and our hard instances provide a natural testing ground for new techniques.

[5] This is when Σ is invertible. If Σ is not invertible the task would be learning the concept/regression function, i.e. outputting *any* $\hat{w}$ so that $(\hat{w} - w)^T \Sigma(\hat{w} - w) = 0$. If required to output a k-sparse predictor and Σ is not invertible, the problem is NP hard with infinitely many samples [29] as it's equivalent to finding a sparse solution of linear equations.

But another potential explanation is the current dearth of conjecturally-hard instances of random-design sparse linear regression. Reductions from Planted Clique (and other techniques for proving average-case lower bounds) either explicitly or implicitly involve constructing distributions over the unknown concept class which encapsulate (to some degree) the hardness of the original problem. For instance, sparse PCA is conjecturally hard even when the planted sparse direction is drawn uniformly at random from k-sparse vectors with nonzero entries $\pm 1/\sqrt{k}$ [36, 9]. Similarly, non-reduction based lower bounds like Statistical Query lower bounds[6] are typically derived against the uniform distribution over a finite set of concepts [27]. Thus, it is problematic that for sparse linear regression, no such distribution has even been hypothesized: i.e., a family $(\Sigma_n, \mathcal{D}_n)_{n\in\mathbb{N}}$ where Σ_n describes an $n \times n$ covariance matrix, and $\mathcal{D}_n$ describes a distribution over $k(n)$-sparse n-dimensional signals, specifying a model which seems to encapsulate some of the difficulty of random-design SLR. While it would certainly be necessary that the covariance matrices Σ_n are ill-conditioned, there are also natural families of ill-conditioned covariate distributions which lead to tractable instances (e.g. [32, 31]), so such a property is by no means sufficient.

Our main result provides a family of covariance matrices and signal distributions which is hard at least for all preconditioned Lasso algorithms (this was not accomplished by [32], because they constructed a different signal for each preconditioner). Thus, our work can be viewed as a necessary first step towards average-case computational hardness of ill-conditioned sparse linear regression.

1.1 The Preconditioned Lasso

Perhaps the canonical method to solve sparse linear regression problems is by solving a convex program known as the Lasso [44]. In our noiseless setting, it reduces to the basis pursuit program

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathbb{R}^n : Xw=y} \|w\|_1 .$$

This program is well-studied, and it's known to succeed with high probability with $O(k \log n)$ samples when Σ is well-conditioned [54], but it also is known to fail if Σ is ill-conditioned [23, 17, 31]. In the context of ill-conditioned SLR, the class of programs known as the Preconditioned Lasso [32] is significantly more powerful. Informally, these algorithms apply some change-of-basis to "condition" the covariates before solving the basis pursuit program. Formally:

Definition 1. For an invertible $n \times n$ matrix S (which we think of as an arbitrary function of Σ that however cannot depend on the samples), the *S-preconditioned Lasso* on data (X_i, y_i) applies the transformation $X_i \mapsto S^{-1}X_i$, solves the basis pursuit to get an estimate $\hat{v}$, and returns $\hat{w} := S^T\hat{v}$. This corresponds to solving the convex program

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathbb{R}^n : Xw=y} \left\|S^T w\right\|_1 . \tag{1}$$

Preconditioned Lasso obviously generalizes the Lasso, and it works for SLR tasks where applying the Lasso directly will fail badly. For example, whenever the covariates come from autoregressive time series data (e.g. random walk or $AR(1)$) the sparse linear regression problem can be solved nearly optimally via Preconditioned Lasso, even though the covariates are likely to be heavily correlated. More generally, any covariance matrix with low-treewidth dependency structure induces an SLR model that is tractable via Preconditioned Lasso, even if the matrix is arbitrarily ill-conditioned [32]. Due to this generality, examples that are provably hard against the Preconditioned Lasso are much more difficult to obtain, and much more interesting. Concretely, prior to the present work, the only known hardness result against the Preconditioned Lasso was the following statement:[7]

Theorem 1.1 (Informal theorem statement from [32])**.** *For any $n > 0$, there is a positive-definite covariance matrix $\Sigma : n \times n$ such that for any preconditioner S, there exists some $\operatorname{polylog}(n)$-sparse signal w^* which S-preconditioned Lasso with probability $1 - o(1)$ fails to recover, when given $o(n)$ independent samples $X_i \sim N(0, \Sigma)$ and $y_i = \langle X_i, w^* \rangle$.*

[6] Relatedly, for realizable regression problems there is a general computationally inefficient algorithm which makes a smaller number of SQ queries [49].

[7] To be more precise, Theorem 1.1 and the definition of Preconditioned Lasso actually apply to all rectangular $n \times s$ preconditioners, not just invertible preconditioners (note that Program 1 is still defined, though it no longer corresponds to a change-of-basis). However, most algorithms applying Preconditioned Lasso in the literature use invertible preconditioners [31, 32]. Moreover, restricting to invertible preconditioners does not improve the failure probability achievable by the techniques in [32].

This provides a converse to the algorithmic results of [32], which show that for certain covariance matrices, there is a preconditioner that works for all signals. However, the limitation of Theorem 1.1 is that the hard signal depends on the preconditioner: unpacking the proof of Theorem 1.1 only gives a signal distribution on which any preconditioner fails with probability $\Omega(1/n)$. Thus, the lower bound is extremely weak if we allow for *randomization* in the choice of preconditioner, which is how preconditioners are often applied in other contexts.

2 Main Results

In this paper, we eliminate that limitation and provide a *single* signal distribution which is provably impossible to precondition. That is, we construct a covariance matrix and a sparse signal distribution under which Preconditioned Lasso with any invertible, randomized change-of-basis must fail *with high probability*, unless a linear number of samples are given.

Theorem 2.1 (Theorem E.3 of the Appendix)**.** *Let $n > 0$. There is a positive-definite covariance matrix $\Sigma : n \times n$ and a distribution $\mathcal{D}$ over $\operatorname{polylog}(n)$-sparse signals with the following property: for any invertible randomized preconditioner S, if we draw $w^* \sim \mathcal{D}$, then S-preconditioned Lasso fails to recover w^* with probability at least $1 - o(1)$, when given $o(n)$ independent samples $X_i \sim N(0, \Sigma)$ and $y_i = \langle X_i, w^* \rangle$.*

In fact, the full version of Theorem E.3 is even stronger: it shows that even if we fix a family of $\operatorname{poly}(n)$ different preconditioners, they will all fail on a problem instance sampled from our distribution with probability $1 - o(1)$. This allows us to rule out an even larger class of algorithms: for example, the algorithm used in [31] for solving jointly walk-summable SLR instances adaptively selects one out of n possible (invertible) preconditioners before running the Preconditioned Lasso, and our lower bound shows that this strategy and variants are provably defeated by our new construction.

Additionally, we can extend our result to show hardness against rectangular $n \times s$ preconditioners. For technical reasons we only achieve a failure probability of $1/2 - o(1)$, and require a bound on the preconditioner size. Nonetheless, at this failure probability and with poly-logarithmically sparse signals, we can rule out all polynomially-sized preconditioners.

Theorem 2.2 (Theorem E.10 of the Appendix)**.** *Let $n > 0$. There is a positive-definite covariance matrix $\Sigma : n \times n$ and a distribution $\mathcal{D}$ over k-sparse signals, for any $k \geq \log^{12}(n)$, with the following property: for any randomized preconditioner S with at most $\exp(k/\log^{10}(n))$ columns, if we draw $w^* \sim \mathcal{D}$, then S-preconditioned Lasso fails to recover w^* with probability at least $1/2 - o(1)$, when given $o(n)$ independent samples $X_i \sim N(0, \Sigma)$ and $y_i = \langle X_i, w^* \rangle$.*

In both theorems, the probability is over the choice of preconditioner, the signal distribution, and the random samples. Surprisingly, the key technique used to prove Theorem 2.1 and 2.2 is a *positive* identifiability result in compressed sensing. We show that random sparse compressive design matrices are robust to adversarial erasures. To prove our lower bounds, we will fix one such matrix M, and we will define the covariance matrix $\Sigma = (M^T M + \epsilon I)^{-1}$ for small $\epsilon > 0$; the robustness of M will then imply important properties of Σ. *Note that the final Gaussian SLR design matrix (with independent rows drawn from $N(0, \Sigma)$) will be quite different from the erasure-robust design matrix M.* But first, we describe the application of such matrices M to compressed sensing, as we believe it is of independent interest and a main contribution of this work.

2.1 Key Technique: Erasure-Robust Sparse Designs

In this section we describe the key technique that enables our main results, an erasure-robust sparse design, and provide an independent motivation for this design from compressed sensing. At the end, we provide intuition for how it connects back to hardness against Preconditioned Lasso.

There is a vast literature on sparse linear regression and compressed sensing. Many deterministic conditions and stochastic models for the measurement matrix (also known as design matrix or covariate matrix) have been demonstrated to imply that sparse signals can be recovered either information-theoretically or algorithmically [54]. In noisy settings, the goal is usually either approximate recovery under an ℓ_p norm or prediction error (e.g. [14, 53]); exact support recovery with some assumptions about the signal-to-noise ratio (e.g. [53]); or approximate support recovery under distributional assumptions about the signal (e.g. [40]). However, in noiseless settings, the goal is invariably exact recovery. This is obviously ideal. But in situations where the measurements are not entirely controlled by the algorithm designer, exact recovery could be impossible. A natural goal is then to try to recover *part* of the signal. To our knowledge, this notion of partial recovery of sparse signals (i.e. due to shortcomings of the measurement matrix rather than due to noise) has received essentially no attention; see Section 3 for a discussion of related notions.

Part of the reason may be that it's not obvious what models for a compressive measurement matrix exhibit the behavior that some but not all of the coordinates of a sparse signal are identifiable, besides artificial examples where e.g. unconstrained variables are added to the system. Such examples do not answer the question of whether partial recovery is possible under fundamentally weaker modelling assumptions than total recovery.

Erasure-robustness. Our key technical contribution is a proof that partial recovery is possible in a natural *semi-random* model. Specifically, we show that random sparse compressive measurement matrices M are "erasure-robust", by which we mean that if an adversary *erases* a small fraction of the measurements $\langle M_i, x^* \rangle$, replacing them with question-marks, then most of the coordinates of the sparse signal vector x^* are still information-theoretically identifiable. Moreover, the identifiability result is stable under additive measurement noise with polynomially small norm.

Adversarial erasures have been studied in compressing sensing before, and of course have also been long and extensively studied in coding theory (see e.g. [35, 33, 24]). For random dense compressive matrices, it's known that deleting a small fraction of the measurements essentially does nothing; the sparse signal is still totally recoverable with high probability [19, 52, 34]. But for random sparse matrices (which in the absence of erasures do also enable total sparse recovery [5]), no such robustness has been proven, because it's not true: the adversary may simply delete all measurements interacting with a particular coordinate, rendering that coordinate unidentifiable. Given this, partial recovery is the best that can be hoped for. Our result implies that it is also attainable, at least information-theoretically. Moreover, we achieve a nearly-tight bound on the number of unidentifiable coordinates. Here is the informal statement:

Theorem 2.3 (Theorem C.14 of the Appendix)**.** *Let $n, m > 0$ satisfy $n > m > \Theta(\log^2 n)$, and let M be an $m \times n$ matrix with independent Bernoulli-p entries for $p = \Theta(\log^2 n)/m$. With high probability, the following holds. For any set of "deleted" equations $B \subseteq [m]$ of size $|B| \leq O(m/\operatorname{polylog}(n))$, there is a set C (the "unidentifiable coordinates") of size $|C| \leq 2|B|$ such that*

$$\|x_{C^c}\|_2 \leq \operatorname{poly}(n) \cdot \|M_{B^c} x\|_\infty$$

for any $O(m/\operatorname{polylog}(n))$-sparse vector $x \in \mathbb{R}^n$.

To make apparent the implication for compressed sensing, we state the following corollary:

Corollary 2.4. *Let $M \in \{0,1\}^{m \times n}$ be a fixed, known matrix for which the guarantee of Theorem 2.3 holds. Let $x^* \in \mathbb{R}^n$ be an unknown k-sparse vector with $k \leq O(m/\operatorname{polylog}(n))$. Define measurements $y = Mx^* + \eta$ for some noise vector η with $\|\eta\|_\infty \leq \delta$, but suppose that y undergoes adversarial erasures. That is, we only observe $\overline{y} \in \mathbb{R}^m$ where*

$$\overline{y}_i = \begin{cases} ? & \text{if } i \in B \\ y_i & \text{if } i \in B^c \end{cases},$$

*for an adversarially-chosen set $B \subseteq [m]$. Then we can information-theoretically identify an **estimate** $\hat{x} \in \mathbb{R}^n$ and a **region of uncertainty** $C \subseteq [n]$ such that $|\hat{x}_i - x_i| \leq \delta \cdot \operatorname{poly}(n)$ for all $i \notin C$. Moreover, the region of uncertainty is a function of B (not depending on x or y) and is bounded as $|C| \leq 2|B|$.*

In particular, the estimator is simply $\hat{x} = \operatorname{argmin}_{x : \|M_{B^c} x - \overline{y}_{B^c}\|_\infty \leq \delta} \|x\|_0$. The region of uncertainty C is any set C satisfying the guarantee of Theorem 2.3 with respect to B; it can be (inefficiently) computed from B via brute-force computation of singular values of submatrices of M. Then, since $\hat{x} - x^*$ is $2k$ sparse, the guarantee of Theorem 2.3 gives that

$$\begin{aligned} \|(\hat{x} - x^*)_{C^c}\|_\infty &\leq \operatorname{poly}(n) \cdot \|M_{B^c}(\hat{x} - x^*)\|_\infty \\ &\leq \operatorname{poly}(n) \cdot \|(M_{B^c}\hat{x} - \overline{y}_{B^c}) + (\overline{y}_{B^c} - M_{B^c} x^*)\|_\infty \leq \operatorname{poly}(n) \cdot \delta \end{aligned}$$

as claimed in the corollary.

Sparsity of the measurement matrix (and not just the signal) is well-studied in compressed sensing and has various practical applications. For example, in scientific experiments it is often the case that linearity of the response with respect to the covariates is a modelling assumption that's only reasonable for a sparse covariate vector [26]. Our result implies that even with a sparse measurement matrix, adversarial erasures (due to e.g. experimental error) are not disastrous. (Note that in the above theorem, each row of the measurement matrix is roughly $(n \log^2 n)/m$ sparse, which up to logarithmic factors cannot be improved, even without erasures, since the measurement matrix must have $\Omega(n)$ nonzero entries).

Open question. Our results show that partial sparse recovery in this semi-random model is possible. However, finding a computationally efficient algorithm is an interesting open problem.

Connection to lower bounds against Preconditioned Lasso. It may seem rather mysterious that construction of an erasure-robust, "good" design matrix is the key ingredient in a distributional hard example for a family of sparse recovery algorithms. The technical reasons for this connection are deferred to the overview, but here we try to give some high-level intuition. First, for intuition, we restrict our focus to sparse preconditioners, because dense preconditioners (morally) do not preserve the sparsity of the unknown signal and therefore should not work. Now, if Σ is very ill-conditioned, the preconditioner essentially needs to "fix" Σ by reweighting the different eigenspaces.

If we are allowed to construct the signal based on the preconditioner, then the preconditioner is forced to be a good approximation for Σ everywhere, with no bad directions. But from any good measurement matrix, using the fact that it has dense kernel, we can construct a Σ so that no sparse preconditioner can approximate Σ everywhere. This is the approach taken in [32] for Theorem 1.1.

In our case, we need to construct the signal distribution without knowing the preconditioner, so the preconditioner is only forced to be a good approximation for Σ in *most* directions. Ruling out sparse preconditioners then corresponds to a measurement matrix which has a density property even if some of the rows are ignored (specifically, erasure-robustness). Sparsity of the measurement matrix is needed so that the rows are valid sparse signals, and compressivity is needed so that Σ is ill-conditioned in many directions.

3 Related Work

Hard Examples for SLR. As we mentioned earlier, SLR with worst-case covariates is known to be computationally hard under worst-case complexity-theoretic assumptions [37, 57, 30, 28]. However, for the random-design covariate model, there is no known reduction-based hardness, even under average-case or cryptographic assumptions. We enumerate known restricted hardness results:

First, there is a large literature on when the Lasso and Basis Pursuit programs fail at sparse recovery, even for random designs [53, 23, 17, 46, 31]. While these results do (technically speaking) prove lower bounds against classes of algorithms, these classes are quite small; the constructed examples are only proven to be hard for the Lasso and/or Basis Pursuit (at best, these programs have one meta-parameter). In fact, as has been previously observed [58, 32], all of the "hard" examples provided in the above works can be fixed by a simple change-of-basis.

Second, there is a more general lower bound against the class of convex programs solving least-squares regression with coordinate-separable regularizers [58]. While this is a fairly broad class of algorithms (incomparable with the Preconditioned Lasso), the result has two limitations. One is that the constructed hard signals depend on the regularizer, so there is no single signal distribution that is hard for the entire class. The other limitation is that, like in the previous works on hardness against the Lasso, the hard example in [58] can be made easy for the Lasso by a simple change-of-basis.

Third, in [32], motivated by the latter limitation, covariance matrices are constructed such that for any change-of-basis, there is a sparse signal (in the original basis) which causes the "preconditioned" basis pursuit program to fail. However, as we have previously noted, this result is still limited by the strong dependence of the signal on the preconditioner: it does not even rule out the possibility that there are always *two* preconditioners so that every signal can be recovered by one of them.

Fourth, for *isotropic* random-design SLR (i.e. when $\Sigma = I$), there has been work on identifying the precise sample complexity of sparse recovery. In particular, there appears to be a constant-factor gap between the sample complexities of algorithmic recovery and information-theoretic recovery.

Evidence has been given for this gap via the Overlap Gap Property [25], which implies the failure of a restricted class of "stable" algorithms. However, this problem seems fundamentally different from the problem we consider, where the hardness arises from the ill-conditioning of the covariates, and the sample complexity gap is conjecturally exponential rather than a constant.

Partial sparse recovery. There are several other works in compressed sensing that use the terminology of "partial" recovery. To our knowledge, these works all consider different settings from ours; we explain the differences. First, in [3], partial sparse recovery refers to totally recovering a signal that is only partially sparse (where the signal space is divided into two sets of coordinates, and it's known that the signal is sparse on the first set). Second, in [43], the goal is indeed to recover only part of the support of the signal. However, their model is the Gaussian Sequence Model (i.e. where the measurement matrix is the identity), where it is obvious that partial recovery is possible, because there is no compression. Third, as discussed previously, one common goal in noisy models is partial support recovery (see e.g. [40]). There, the goal is to estimate the support with few false positives and false negatives, and the reason for error is simply that some coordinates of the signal may be very small and therefore indistinguishable from noise. In contrast, partial identifiability occurs in our setting even without noise, due to a weaker model for the measurement matrix. Moreover, proving partial support recovery in the setting of [40] requires make strong probabilistic assumptions about the signal, e.g. that the support is a uniform sparse set. In contrast, our results prove conditions under which a measurement matrix enables partial recovery of arbitrary sparse signals.

4 Technical Overview

We start with a sketch of the proof of Theorem 1.1 from [32], which only achieves a failure probability of $O(1/n)$, and which formally motivates the need for erasure-robust sparse designs. We then sketch the proof of our main technical result that random sparse designs are erasure-robust. Finally, we discuss how this result leads to stronger lower bounds against Preconditioned Lasso.

4.1 Lower Bounds via Sparse Designs.

The hard covariance matrix constructed in [32] to prove Theorem 1.1 is defined as $\tilde{\Sigma} = \tilde{\Theta}^{-1}$ where $\tilde{\Theta} = \Theta + \epsilon I$ and $\Theta = M^T M$, for a rectangular matrix M. Note that for small $\epsilon > 0$, this covariance is very ill-conditioned, so long as M has non-trivial kernel. To prove that all Preconditioned Lasso algorithms with m samples fail to recover k-sparse signals, these three properties are used:

1. The rows of M are k-sparse,
2. $\dim \ker M \geq 2m$,
3. $\ker M$ is bounded away from all $(n/k)\log(n)$-sparse vectors.

The first property is self-explanatory. One way to achieve the second property is if M has at most $n - 2m$ rows. And the third property, in compressed sensing, is essentially what a design matrix needs to satisfy to information-theoretically enable $(n/k)\log(n)$-sparse recovery. Thus, to show that $\Omega(n)$ samples are needed to recover $\operatorname{polylog}(n)$-sparse signals, M must be a sparse, compressive matrix which (as a design matrix) enables the recovery of $n/\operatorname{polylog}(n)$-sparse signals.

How do these properties imply that for every preconditioner S, there is a bad k-sparse signal? By the first property, the rows of M are valid signals. For each row M_i, if it is not a bad signal for S-preconditioned Lasso, then it can be shown to induce a certain constraint on S: namely, that every column of S either has small magnitude or is nearly orthogonal to M_i. So if none of the rows of M_i are bad signals, then every column of S either has small magnitude or lies near $\ker M$, in which case by the third property it must be $(n/k)\log(n)$-dense. Roughly speaking, this structure can be used together with the second property to show that a k-sparse signal with uniformly random support causes the Preconditioned Lasso to fail.

A hard signal distribution? The above proof shows that for any preconditioner, either it fails (with high probability) on a random k-sparse signal, or there *exists* some row of M on which it fails. If we want a signal distribution that is uniformly hard, it's therefore natural to equiprobably pick either (a) a random row of M, or (b) a random k-sparse signal. But then the above proof only implies that for this signal distribution, for any preconditioner, the Preconditioned Lasso fails with probability $\Omega(1/n)$. Moreover, it's not clear whether the failure probability can be improved under just the above assumptions: consider the case that for some preconditioner, just a few rows of M are bad signals. Then the likelihood that one of these rows is chosen as the signal is only $O(1/n)$. Moreover, the columns of S are now only forced to be orthogonal to most rows of M, not all. As a result, the columns may have large magnitude and yet fail to be dense, because for sparse matrices like M, it's possible to adversarially delete a few rows so that the kernel of the remaining rows contains sparse vectors. This is an obstacle to proving that such a preconditioner must fail on a sparse signal with uniformly random support.

To circumvent this obstacle, we need to show that a preconditioner S which has columns orthogonal to most rows of M, but not all, still has useful structure. As we suggested earlier, this can be done by reasoning about sparse compressive matrices under adversarial deletions.

4.2 Erasure-robustness

We will return to the lower bound problem in the next section of the overview, but for now focus on the core technical result about partial recovery with adversarial erasures. Based on the discussion after Theorem 2.3, we need to solve the following problem.

Let M be an $m \times n$ sparse random Bernoulli matrix with parameter $p = \Theta(\log n)/m$. We want to show that with high probability, M supports erasure-robust partial sparse recovery: that is, for any set $B \subseteq [m]$ of "bad equations", there is a small set $C \subseteq [n]$ such that if $x \in \mathbb{R}^n$ is τ-sparse, then

$$\|x_{C^c}\|_2 \leq \operatorname{poly}(n) \cdot \|M_{B^c}x\|_\infty .$$

Erasure-robustness: the exact case. For simplicity, in this proof sketch we start by considering the *exact* case, where $M_{B^c}x = 0$, and we want to show that either $|\operatorname{supp}(x)| \geq \tau$ or $\operatorname{supp}(x) \subseteq C$. Without erasures (i.e. $B = \emptyset$), this property follows for $C = \emptyset$ by the fact that the adjacency graph of M is with high probability a unique-neighbor expander.[8] Concretely, because the graph is a $(1-\epsilon)d$ expander for a small constant $\epsilon > 0$, any set $S \subseteq [n]$ of size at most $\tau := O(m/\log(n))$ has at least $(1 - O(\epsilon))d|S|$ unique neighbors in $[m]$. Moreover, if $j \in [m]$ is a unique neighbor of $\operatorname{supp}(x)$ for some vector $x \in \mathbb{R}^n$, then $M_j x \neq 0$. Thus, if $Mx = 0$ then $\operatorname{supp}(x)$ must have no unique neighbors, so either $|\operatorname{supp}(x)| \geq \tau$ or $x = 0$.

However, this argument breaks down in the presence of adversarial erasures. All that can be said is that if $M_{B^c}x = 0$ then $\operatorname{supp}(x)$ must have no unique neighbors in B^c. By the unique neighbor lower bound, it does follow that either $|\operatorname{supp}(x)| \geq \tau$ or $|\operatorname{supp}(x)| \leq O(|B|/d)$ — this can be thought of as a kind of *density amplification* result for $\ker M_{B^c}$, since it eliminates the possibility of any vector in the kernel having an intermediate density. Unfortunately, this does not directly imply erasure-robustness, because we need a single set C that contains the supports of all sparse vectors in $\ker M$, not a different C for each x. (For example, if we allow $C = \operatorname{supp}(x)$ then the result is not very interesting.) Moreover, it's not clear that anything useful can be said about the vertex set $\operatorname{supp}(x)$: certainly many vertices in $\operatorname{supp}(x)$ must be adjacent to "bad" equations, but it's conceivable that other vertices could be farther away. Pictorially, one possible case (of many) is that B could be chosen as the set of "boundary" equations of a ball subgraph; then $\ker M_{B^c}$ certainly contains a vector supported on the ball, which is not actually contained in the neighborhood of B.

Given the above obstacles, one approach is to show that although $\operatorname{supp}(x)$ may not be contained in the neighborhood of B, it must be contained in a distance-r ball around B, for some small but super-constant r. The argument is that if there is a vertex of $\operatorname{supp}(x)$ which is distance greater than r from B, then by iteratively growing neighborhoods of the vertex until B is reached, the support must have size at least d^r, and a contradiction is reached if $d^r > |B|/d$, because then B cannot contain all unique neighbors of $\operatorname{supp}(x)$. Unfortunately, the constructed set C (the distance-r ball around B) then has size $|B| \cdot (d^2)^r \approx |B|^3$, since the distance metric is that two coordinates are adjacent if they share an equation. This is much larger than the desired bound ($O(|B|)$) and in particular, too large to use in our ultimate lower bound application.

In summary, to get the linear bound claimed in Theorem 2.3, we need a different argument. The key idea is to exploit *linearity*. We want to show that the union U of supports of all τ-sparse vectors in

[8] We note that this initial part of the argument (the case without erasures) is quite reminiscent of arguments used in the analysis of LDPC codes (see e.g. [41]).

$\ker M_{B^c}$ has small size. We've seen that for any fixed x, there is a *density amplification* result: if $M_{B^c}x = 0$ and x is $|B|/d$-dense, then x must be τ-dense. So take vectors $x^{(1)}, \dots, x^{(n)} \in \ker M_{B^c}$ which are τ-sparse (and therefore $|B|/d$-sparse) and which cover U. Now observe that since $x^{(1)}$ and $x^{(2)}$ are $O(|B|/d)$-sparse, any linear combination $c_1 x^{(1)} + c_2 x^{(2)}$ must be $2|B|/d$-sparse. But $c_1 x^{(1)} + c_2 x^{(2)} \in \ker M_{B^c}$ by linearity. So if $2|B|/d < \tau$, then by the (contrapositive of the) density amplification result, we in fact know that the sum is $|B|/d$ sparse! Inductively, it follows that any linear combination $c_1 x^{(1)} + \dots + c_n x^{(n)}$ is $|B|/d$-sparse. But for generic $c_1, \dots, c_n$, we have

$$\operatorname{supp}(c_1 x^{(1)} + \dots + c_n x^{(n)}) = \bigcup_{i=1}^{n} \operatorname{supp}(x^{(i)}) = U.$$

This shows that in fact we can find a set C of size $O(|B|/d)$ satisfying the desired property.

Erasure robustness: the general case. Note that the above argument was when $M_{B^c}x = 0$. The proof for the general case, when $M_{B^c}x$ is small but not nonzero, uses the same insight with several complications. First, we need a quantitative density amplification lemma which states that if $M_{B^c}x$ is small and x has more than $|B|$ coordinates with magnitude exceeding some threshold δ, then we can trade off density for magnitude, i.e. find τ coordinates with magnitude exceeding $\delta/\operatorname{poly}(n)$. To prove this without losing a superpolynomial factor on the threshold, we actually need the graph to satisfy a stronger property than just expansion: we also need that for any two disjoint sparse sets $S, T \subseteq [n]$, the intersection of their neighborhoods has size only $O(\sqrt{d} \cdot \max(|S|, |T|))$. Note that expansion would only give a bound of $O(\epsilon d \max(|S|, |T|))$. Nonetheless, it can be proven that the random sparse adjacency matrix of M satisfies the desired stronger property with high probability.

Second, the iterative addition procedure in the noiseless case requires a modification for the noisy case; each addition causes the quantitative threshold to decay, and after n additions it would decay by a factor superpolynomial in n. Instead, we add the vectors $x^{(1)}, \dots, x^{(n)}$ recursively according to a d-ary tree. This tree has depth only $\log_d n$, which allows the decay to be controlled to only a $\operatorname{poly}(n)$ factor, proving Theorem 2.3.

4.3 Stronger lower bound via erasure-robustness

We now return to the problem of proving hardness against Preconditioned Lasso. Theorem 2.3 can be used to show that for an appropriately chosen M, if the number of rows of M that are bad signals for S-preconditioned Lasso is at most $n/\operatorname{polylog}(n)$, then there is a set $C \subseteq [n]$ of size $n/\operatorname{polylog}(n)$ such that each column of S is either $n/\operatorname{polylog}(n)$-dense, or has small magnitude on coordinates outside the set C. This is precisely the structure lemma we need for preconditioners that succeed on most rows of M: it crucially allows for a nearly-linear number of rows of M that are bad signals, although in exchange there is a set C of sublinear size where we cannot control the columns of S (the corresponding structure lemma in [32] could not tolerate any bad rows). We also show that in this situation, the number of dense columns of S must be $\Omega(n)$.

With these results, we can prove our lower bounds. First, to prove our lower bound against invertible preconditioners (Theorem 2.1), we define a distribution over $\operatorname{polylog}(n)$-sparse signals by taking w^* to be the sum of $\operatorname{polylog}(n)$ random rows of M, plus an infinitesimal uniformly random $\operatorname{polylog}(n)$-sparse vector. Under certain conditions, if at least one of the rows in the sum is a bad signal, then the sum must also be a bad signal. With this amplification (at the cost of a $\operatorname{polylog}(n)$ factor in sparsity), any invertible preconditioner must fail with probability $1 - o(1)$: either there are $\Omega(n/\operatorname{polylog}(n))$ bad rows of M, in which case the sum of the chosen rows is a bad signal with high probability, and the infinitesimal perturbation does not affect the program failure. Or, S has many dense columns, in which case $S^T w^*$ is dense due to the perturbation. By a dimension-counting argument (which crucially uses invertibility of S), this implies that there exists a feasible direction of improvement for the program objective.

Extending the lower bound to rectangular preconditioners is more involved and involves generalizations of techniques from [32]. The factor of $1/2$ in Theorem 2.2 arises because we are not able to construct a single signal distribution that causes failure of both "incompatible" preconditioners (i.e. for which more than $n/\operatorname{polylog}(n)$ rows of M are bad signals) and "compatible" preconditioners (for which at most $n/\operatorname{polylog}(n)$ rows of M are bad signals, so the structure lemma applies) with high probability. Instead, we take a mixture of the two cases' hard distributions: either a sum of rows of M, or a uniformly random sparse vector. Due to the lack of invertibility of the preconditioners, the second case is no longer a simple dimension-counting argument. In [32], the proof crucially relies on a "projection lemma" which states that if $\dim \ker M \geq 2m$, then any fixed direction is unlikely to align with the span of the covariates. Since our structure lemma has no control over the preconditioner columns in the subspace indexed by the set C, we prove a generalized projection lemma which states that alignment is unlikely even on C^c. This yields Theorem 2.2.

References

[1] Noga Alon and Joel H Spencer. *The probabilistic method.* John Wiley & Sons, 2004.

[2] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[3] Afonso S Bandeira, Katya Scheinberg, and Luis Nunes Vicente. On partial sparse recovery. *arXiv preprint arXiv:1304.2809*, 2013.

[4] Anirban Basak and Mark Rudelson. Sharp transition of the invertibility of the adjacency matrices of sparse random graphs. *Probability Theory and Related Fields*, pages 1–76, 2021.

[5] Radu Berinde, Anna C Gilbert, Piotr Indyk, Howard Karloff, and Martin J Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 798–805. IEEE, 2008.

[6] Radu Berinde and Piotr Indyk. Sparse recovery using sparse random matrices. *preprint*, 2008.

[7] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on learning theory*, pages 1046–1066. PMLR, 2013.

[8] Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

[9] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference On Learning Theory*, pages 48–166. PMLR, 2018.

[10] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Universality of computational lower bounds for submatrix detection. In *Conference on Learning Theory*, pages 417–468. PMLR, 2019.

[11] T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory*, 57(7):4680–4688, 2011.

[12] Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of statistics*, 35(6):2313–2351, 2007.

[13] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[14] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[15] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[16] Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

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

[18] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[19] Mark A Davenport, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. A simple proof that random matrices are democratic. *arXiv preprint arXiv:0911.0736*, 2009.

[20] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[21] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

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

[23] Rina Foygel and Nathan Srebro. Fast rate and optimistic rate for l1-regularized regression. Technical report, Toyota Technological Institute. arXiv: 1108.037 v1, 2011.

[24] Matthew Franklin, Ran Gelles, Rafail Ostrovsky, and Leonard J Schulman. Optimal coding for streaming authentication and interactive communication. *IEEE Transactions on Information Theory*, 61(1):133–145, 2014.

[25] David Gamarnik and Ilias Zadik. High dimensional regression with binary coefficients. estimating squared error and a phase transtition. In *Conference on Learning Theory*, pages 948–953. PMLR, 2017.

[26] Anna Gilbert and Piotr Indyk. Sparse recovery using sparse matrices. *Proceedings of the IEEE*, 98(6):937–947, 2010.

[27] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020.

[28] Aparna Gupte and Vinod Vaikuntanathan. The fine-grained hardness of sparse linear regression. *arXiv preprint arXiv:2106.03131*, 2021.

[29] Aparna Ajit Gupte and Kerri Lu. Fine-grained complexity of sparse linear regression. 2020.

[30] Sariel Har-Peled, Piotr Indyk, and Sepideh Mahabadi. Approximate sparse linear regression. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[31] Jonathan Kelner, Frederic Koehler, Raghu Meka, and Ankur Moitra. Learning some popular gaussian graphical models without condition number bounds. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2020.

[32] Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. *62nd Annual IEEE Symposium on Foundations of Computer Science*, 2021.

[33] Michael Langberg. Private codes or succinct random codes that are (almost) perfect. In *FOCS*, volume 4, pages 325–334, 2004.

[34] Ran Lu. On the strong restricted isometry property of bernoulli random matrices. *Journal of Approximation Theory*, 245:1–22, 2019.

[35] Michael G Luby, Michael Mitzenmacher, Mohammad Amin Shokrollahi, and Daniel A Spielman. Efficient erasure correcting codes. *IEEE Transactions on Information Theory*, 47(2):569–584, 2001.

[36] Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse pca. *Advances in Neural Information Processing Systems*, 28, 2015.

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

[38] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[39] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[40] Jonathan Scarlett and Volkan Cevher. Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Transactions on Information Theory*, 63(1):593–620, 2016.

[41] Michael Sipser and Daniel A Spielman. Expander codes. *IEEE transactions on Information Theory*, 42(6):1710–1722, 1996.

[42] Daniel A Spielman and Jaeoh Woo. A note on preconditioning by low-stretch spanning trees. *arXiv preprint arXiv:0903.2816*, 2009.

[43] Ali Tajer and H Vincent Poor. Hypothesis testing for partial sparse recovery. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 901–908. IEEE, 2012.

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

[45] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[46] Sara Van De Geer. On tight bounds for the lasso. *Journal of Machine Learning Research*, 19:46, 2018.

[47] Sara van de Geer, Johannes Lederer, et al. The lasso, correlated design, and improved oracle inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 303–316. Institute of Mathematical Statistics, 2013.

[48] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[49] Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. In *Conference on Learning Theory*, pages 3115–3117. PMLR, 2019.

[50] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[51] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[52] Vladislav Voroninski and Zhiqiang Xu. A strong restricted isometry property, with an application to phaseless compressed sensing. *Applied and computational harmonic analysis*, 40(2):386–395, 2016.

[53] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[54] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[55] Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of rip certification. *Advances in Neural Information Processing Systems*, 29, 2016.

[56] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

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

[58] Yuchen Zhang, Martin J Wainwright, Michael I Jordan, et al. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 2
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplemental material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Organization

In Appendix B, we discuss notation and collect important definitions, e.g. of erasure-robustness. In Appendix C, we prove that random sparse compressive matrices are erasure-robust (Theorem 2.3). In Appendix D, we prove the key structure lemma that connects erasure-robustness with lower bounds against Preconditioned Lasso. In Appendix E, we then use this lemma together with our result about erasure-robustness to construct example distributions that are provably hard against Preconditioned Lasso algorithms (Theorem 2.1 and Theorem 2.2).

B Preliminaries

For vectors x, y we denote the inner product as $\langle x, y\rangle = x^T y$. For a matrix M, we let $\operatorname{span}(M)$ and $\ker(M)$ denote the row space and null space of M respectively. For a matrix $M \in \mathbb{R}^{n\times p}$, a vector $v \in \mathbb{R}^n$, and a subset $U \subseteq [n]$, we say that M_U is the $|U| \times p$ matrix consisting of the rows of M indexed by U; similarly, v_U consists of the entries of v indexed by U; and the complement of U in $[n]$ is $U^c = \bar{U} = [n] \setminus U$. We use the standard notation for vector norms that $\|v\|_p = \left(\sum_{i=1}^n |v_i|^p\right)^{1/p}$.

B.1 Preconditioners and the Preconditioned Lasso

Definition 2. For $n, s \in \mathbb{N}$, a *preconditioner* is a matrix $S \in \mathbb{R}^{n\times s}$ with $\ker(S^T) = \{0\}$. The S*-preconditioned-Lasso* on samples (X, y) is the convex program

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathbb{R}^n : Xw=y} \left\|S^T w\right\|_1 .$$

In our lower bounds, we will say that the S-preconditioned Lasso *fails* if the true signal vector w^* is not contained in the set of optimal solutions to the program (i.e. some other vector achieves strictly smaller objective value). As a result, some restriction on $\ker(S^T)$ is necessary, to rule out programs with multiple (and therefore infinitely many) co-optimal solutions. In [32], it is only assumed that S is not identically zero; this suffices for their purposes because they are allowed to pick the true signal depending on S. However, we want an algorithm-independent distribution over signals. If S^T is nonzero in only a few directions, then it's obviously impossible to cause the S-preconditioned Lasso to fail (in our strong sense) with non-trivial probability without knowing S. Nonetheless such a program is clearly not useful.

More to the point, any matrix S where $\ker(S^T)$ is non-trivial can be perturbed infinitesimally (possibly by adding columns) so that $\ker(S^T)$ is trivial, but so that if the original program *uniquely* recovered the true signal, then the new program still does so.

B.2 Supports, erasure-robustness, and quantitative density

Definition 3. For $x \in \mathbb{R}^n$ and $\delta > 0$, define the δ-support of x to be

$$\operatorname{supp}_\delta(x) := \{i \in [n] : |x_i| \geq \delta\}.$$

We sometimes refer to $\operatorname{supp}_\delta(x)$ as a *quantitative support* of x with threshold δ. The support of x is $\operatorname{supp}(x) := \|x\|_0 = \{i \in [n] : |x_i| > 0\}$.

Definition 4. Let $M \in \mathbb{R}^{m\times n}$ be a matrix. We say that M is (b, b', η, τ)*-erasure-robust* if for any set $B \subseteq [m]$ of size $|B| \leq b$, there is a set $C \subseteq [n]$ of size $|C| \leq b'$ with the following property: for every $x \in \mathbb{R}^n$, either:

- $|\operatorname{supp}(x)| = \|x\|_0 \geq \tau$, or
- $\|x_{C^c}\|_2 \leq \eta \|M_{B^c} x\|_\infty$.

If M satisfies this property, we say that it tolerates b erasures and sparsity level τ, with only b' unidentifiable coordinates.

Definition 5. For any subspace $V \subseteq \mathbb{R}^n$ and vector $x \in \mathbb{R}^n$, the (Euclidean) distance from x to V is

$$\operatorname{dist}(x, V) := \inf_{v\in V} \|x - v\|_2 = \|\operatorname{Proj}_{V^\perp} x\|_2 .$$

Definition 6. Let $V \subseteq \mathbb{R}^n$ be a subspace. We say that V is (δ, η, τ)-*quantitatively dense* if for any set $C \subseteq [n]$ of size $|C| \leq \tau$, for any $x \in \mathbb{R}^n$ with $\operatorname{dist}(x, V) \leq \delta \|x\|_2$, it holds that $\|x_{C^c}\|_2 \geq \eta \|x\|_2$.

B.3 Random Matrix Theory

We will use the following standard bound on the singular values of Gaussian random matrices.

Theorem B.1 (e.g. Corollary 5.35 in [50])**.** *Let $n, N \in \mathbb{N}$. Let $A \in \mathbb{R}^{N \times n}$ be a random matrix with entries i.i.d. $N(0,1)$. Then for any $t > 0$, it holds with probability at least $1 - 2\exp(-t^2/2)$ that*

$$\sqrt{N} - \sqrt{n} - t \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq \sqrt{N} + \sqrt{n} + t.$$

Theorem B.2 (See e.g. Exercise 4.7.3 of [51])**.** *Suppose $X_1, \dots, X_m \sim N(0, \Sigma)$ with $\Sigma : n \times n$ a positive definite matrix, $t > 0$ and $m = \Omega(n + t^2)$. Let $\hat{\Sigma} = \frac{1}{m}\sum_i X_i X_i^T$. Then with probability at least $1 - 2\exp(-t^2/2)$,*

$$(1-\epsilon)\Sigma \preceq \hat{\Sigma} \preceq (1+\epsilon)\Sigma$$

with $\epsilon = O(\sqrt{n/m} + \sqrt{t^2/m})$.

C Erasure-Robustness via Expanders

In this section, we construct a sparse and compressive measurement matrix M which satisfies erasure-robustness with near-linear erasure tolerance and sparsity level, and so that the number of unidentifiable coordinates is only a constant multiple of the number of erasures. In fact, we'll show that a sparse random Bernoulli matrix has the desired property with high probability, proving Theorem 2.3 and also enabling our construction of a hard distribution family for Preconditioned Lasso.

Recall that the erasure-robustness property is defined as follows: if B is an adversarially chosen subset of indices of rows of M, then M_{B^c} is still a good measurement matrix, meaning that any sparse signal x can still be approximately recovered from noisy measurements $M_{B^c}x + \xi$, except for the entries of x in a small set of input coordinates $C = C(B)$ that is *independent of* x. In our construction, we will prove that $|C(B)| = O(|B|)$, which is nearly optimal because M is sparse, and erasing all equations involving one particular coordinate of the signal clearly makes it impossible to recover that coordinate.

Notation. Let $n, m \in \mathbb{N}$. Let $M \in \{0,1\}^{m \times n}$ be a binary matrix. The matrix M defines a bipartite graph between a set of "equations" $[m]$ and a set of "coordinates" $[n]$. For $S \subseteq [n]$ and $E \subseteq [m]$, we can define neighborhoods

$$N(S) = \{i \in [m] : \exists j \in S : M_{ij} = 1\}$$

and

$$N'(E) = \{j \in [n] : \exists i \in E : M_{ij} = 1\}.$$

For $S \subseteq [n]$, further define the "unique neighborhood" of S to be

$$U(S) = \{i \in [m] : \|M_{iS}\|_0 = 1\}.$$

C.1 Deterministic Conditions for Erasure-Robustness

We start by proving that erasure-robustness holds whenever M satisfies certain deterministic conditions. Throughout this section, we make the following assumption encapsulating the needed properties: approximate regularity, vertex expansion, and a bounded intersection property.

Assumption 1. For some $d, k \in \mathbb{N}$ and $\epsilon > 0$, suppose that M satisfies the following deterministic conditions:

- **(Degree bounds)** For all $i \in [m]$ and $j \in [n]$,

$$|N(j)| \leq (1+\epsilon)d$$

and

$$|N'(i)| \leq (1+\epsilon)(n/m)d.$$

- **(Expansion)** For all $S \subseteq [n]$ with $|S| \leq k$,

$$|N(S)| \geq (1-\epsilon)d|S|.$$

- **(Bounded intersection)** For all disjoint $S, T \subseteq [n]$ with $|S|, |T| \leq k$,

$$|N(S) \cap N(T)| \leq \frac{\sqrt{d}}{8} \max(|S|, |T|).$$

Note that the first two conditions imply a weaker form of the bounded intersection property, with $2\epsilon d$ instead of $\sqrt{d}/8$. However, we need the stronger bound to prove our result.

An immediate corollary of the first two conditions is a unique-neighbor lower bound.

Corollary C.1 (Folklore). *For all $S \subseteq [n]$ with $|S| \leq k$,*

$$|U(S)| \geq (1-3\epsilon)d|S|.$$

Proof. By the degree bound assumption, we have $\sum_{j \in S} |N(j)| \leq (1+\epsilon)d|S|$. Together with the expansion assumption, it follows that $\sum_{j\in S} |N(j)| - |N(S)| \leq 2\epsilon d|S|$. Every unique neighbor counts once in both terms, and every non-unique neighbor counts at least twice in the first term but only once in the second term. So there are at most $2\epsilon d|S|$ non-unique neighbors, which means that $|U(S)| \geq |N(S)| - 2\epsilon d|S| \geq (1-3\epsilon)d|S|$. □

Density amplification. For exact solutions $x \in \mathbb{R}^n$ to $M_{B^c}x = 0$, it's easy to see that the unique neighbor lower bound implies a density amplification statement: if $|\operatorname{supp}(x)| > |B|/((1-3\epsilon)d)$ and $|\operatorname{supp}(x)| \leq k$, then by Corollary C.1, $U(\operatorname{supp}(x)) \cap B^c$ must be nonempty, which contradicts $M_{B^c}x = 0$. Thus, $|\operatorname{supp}(x)| > |B|/((1-3\epsilon)d)$ implies $|\operatorname{supp}(x)| > k$.

For approximate solutions, i.e. when $\|M_{B^c}x\|_\infty \leq \delta$, this argument does not quite work: if some equation $i \in B^c$ has a unique neighbor in $\operatorname{supp}(x)$, this is not necessarily a contradiction because the value of x at that coordinate may be within the error tolerance δ. Thus, we can only hope for amplification if we have a lower bound on $|\operatorname{supp}_\gamma(x)|$, the number of coordinates of x exceeding some threshold $\gamma \gg \delta$ in magnitude. However, there is now another problem: even if some equation has a unique neighbor in $\operatorname{supp}_\gamma(x)$, it may have many neighbors where x has magnitude just slightly less than γ, so that these coordinates together cancel out the large coordinate. If this happens, then that equation must neighbor both $\operatorname{supp}_\gamma(x)$ and $\operatorname{supp}_{\gamma/(2d)}(x) \setminus \operatorname{supp}_\gamma(x)$. Naively, we can use expansion to bound the number of such equations by $2\epsilon d \cdot |\operatorname{supp}_{\gamma/(2d)}(x)|$. However, combined with the unique neighbor lower bound, this only shows that $|\operatorname{supp}_{\gamma/(2d)}(x)| \geq \Omega(1/\epsilon) \cdot |\operatorname{supp}_\gamma(x)| - |B|$. That is, density is amplified by a constant factor, but the threshold decays by a factor of $2d = \Omega(\log n)$.

This tradeoff is not good enough for our purposes (tolerating inverse-polynomial measurement error). The following lemma uses the bounded intersection property to prove a better tradeoff:

Lemma C.2 (One-step quantitative density amplification). *Suppose that $\epsilon \leq 1/12$ and $d \geq 2$. Let $x \in \mathbb{R}^n$ and $\delta > 0$. If $\|M_{B^c}x\|_\infty \leq \delta$, then*

$$|\operatorname{supp}_\delta x| \geq \min\left(6\sqrt{d}|\operatorname{supp}_{2d\delta} x| - |B|, k\right).$$

Proof. Let $S = \operatorname{supp}_{2d\delta} x$ and let $T = \operatorname{supp}_\delta x \setminus S$. If $|\operatorname{supp}_\delta x| > k$ then the claim holds, so suppose that $|\operatorname{supp}_\delta x| \leq k$. On the one hand, by Assumption 1, we have that

$$|N(S) \cap N(T)| \leq \frac{\sqrt{d}}{8} \max(|S|, |T|) \leq \frac{\sqrt{d}}{8} |\operatorname{supp}_\delta x|.$$

On the other hand, we claim that $U(S) \backslash B \subseteq N(S) \cap N(T)$. Clearly $U(S) \backslash B \subseteq N(S)$, so it remains to show inclusion in $N(T)$. Suppose that this is false. Then pick some $i \in U(S) \setminus (B \cup N(T))$.

Since $i \in U(S)$, there is a unique $j \in S$ such that $M_{ij} = 1$. Now

$$\begin{aligned}|M_i x| &\geq |x_j| - \sum_{j' \neq j} |M_{ij'} x_{j'}| \\ &\geq 2d\delta - \sum_{j' \neq j : M_{ij'} = 1} \delta \\ &\geq (1-\epsilon)d\delta.\end{aligned}$$

The first inequality is by the triangle inequality. The second inequality uses $j \in S$ and the definition of S; it also uses that any $j' \neq j$ with $M_{ij'} = 1$ satisfies $j' \notin T$ (since $i \notin N(T)$) and $j' \notin S$ (since j is the unique neighbor of i in S): thus $j' \notin \operatorname{supp}_\delta x$, so $|x_{j'}| \leq \delta$. The last inequality is by the degree bound on i.

But since $\epsilon < 1 - 1/d$, the resulting inequality $|M_i x| \geq (1-\epsilon)d\delta$ contradicts the lemma assumption that $\|M_{B^c} x\|_\infty \leq \delta$, since $i \notin B$. So in fact the initial claim that

$$U(S) \setminus B \subseteq N(S) \cap N(T)$$

was true. As a result,

$$|\operatorname{supp}_\delta x| \geq \frac{8|N(S) \cap N(T)|}{\sqrt{d}} \geq \frac{8|U(S) \setminus B|}{\sqrt{d}} \geq 6\sqrt{d}|S| - |B|$$

where the final inequality uses Corollary C.1 and the assumption that $\epsilon \leq 1/12$. □

Recursively applying Lemma C.2 gives our quantitative analogue of the noiseless density amplification: if $\|M_{B^c} x\|_\infty$ is small (i.e. x approximately satisfies the "good" equations of M), then if x has at least $|B|$ not-too-small entries, it must in fact have at least k not-too-small entries. Moreover, the threshold defining "not-too-small" only decays by a polynomial factor.

Corollary C.3 (Quantitative density amplification)**.** *Suppose that $\epsilon \leq 1/12$ and $d \geq 16$. Let $f(n,d) = 2dn^2$. Let $x \in \mathbb{R}^n$ and $\delta > 0$. If $\|M_{B^c} x\|_\infty \leq \delta$, and $|\operatorname{supp}_{f(n,d)\delta}(x)| \geq |B|$, then $|\operatorname{supp}_\delta(x)| \geq k$.*

Proof. For any $\delta' \geq \delta$ with $|\operatorname{supp}_{2d\delta'}(x)| \geq |B|$, Lemma C.2 implies that

$$|\operatorname{supp}_{\delta'}(x)| \geq \min\left(\sqrt{2d}|\operatorname{supp}_{2d\delta'}(x)|, k\right).$$

Let $t \in \mathbb{N}$ be maximal subject to $|\operatorname{supp}_{(2d)^t\delta}(x)| \geq |B|$. Applying the inequality to all $\delta' \in \{\delta, 2d\delta, \dots, (2d)^{t-1}\delta\}$ gives that

$$|\operatorname{supp}_\delta(x)| \geq \min((\sqrt{2d})^t, k).$$

But since $|\operatorname{supp}_{f(n,d)\delta}(x)| \geq |B|$, we know that $(2d)^{t+1} > f(n,d) = 2dn^2$, so $(\sqrt{2d})^t \geq n$. Thus, $|\operatorname{supp}_\delta(x)| \geq k$. □

Bounding the union of supports. We've shown that if a single vector x has quantitative support larger than $|B|$ at some threshold, then we can amplify the density by decreasing the threshold (so long as x is an "approximate solution", i.e. $\|M_{B^c} x\|_\infty$ is sufficiently small). Ultimately, we want to bound the union of the quantitative supports over all k-sparse vectors for which $M_{B^c} x$ is small. The key step is the following lemma, which shows that adding quantitatively sparse approximate solutions preserves quantitative sparsity (although we have to decrease the threshold by a factor polynomial in the number of terms in the sum).

The lemma is proved recursively: we divide the terms into $O(\sqrt{d})$ groups, apply the inductive hypothesis to bound the supports of the group sums, and then add up the $O(\sqrt{d})$ sums. When there are only $O(\sqrt{d})$ terms, the proof is via Lemma C.2: if the sum has quantitative support larger than $|B|$, then by amplification it must have quantitative support at least $O(\sqrt{d}|B|)$. But by the triangle inequality, adding up r terms can only increase the quantitative support size by a factor of r. For an appropriate choice of constants, this yields a contradiction, proving the desired claim.

Lemma C.4. *Suppose that $\epsilon \leq 1/12$ and $k \geq 4\sqrt{d}|B|$ and $d \geq 16$. Let $\delta > 0$ be a threshold. Let $r \leq (\sqrt{2d})^t$ for some $t \in \mathbb{N}$. Let $x_1, \dots, x_r \in \mathbb{R}^n$ satisfy $\|M_{B^c} x_i\|_\infty \leq \delta/\sqrt{2d}$ and $|\operatorname{supp}_\delta x_i| \leq |B|$ for all $i \in [r]$. Then*

$$\left|\operatorname{supp}_{(2d)^{3t/2}\delta} \sum_{i=1}^{r} x_i\right| \leq |B|.$$

Proof. We induct on t. If $t = 0$ then $r \leq 1$, and the claim is trivial. Suppose $t = 1$. For the sake of contradiction, assume that

$$\left|\operatorname{supp}_{(2d)^{3/2}\delta} \sum_{i=1}^{r} x_i\right| > |B|.$$

We know that

$$\left\|M_{B^c} \sum_{i=1}^{r} x_i\right\|_\infty \leq \delta r/\sqrt{2d} \leq \delta.$$

So by Lemma C.2,

$$\left|\operatorname{supp}_{\sqrt{2d}\delta} \sum_{i=1}^{r} x_i\right| \geq \min(6\sqrt{d}|B| - |B|, k) \geq 4\sqrt{d}|B|.$$

But

$$\operatorname{supp}_{\sqrt{2d}\delta} \sum_{i=1}^{r} x_i \subseteq \bigcup_{i=1}^{r} \operatorname{supp}_\delta x_i$$

by the triangle inequality and since $r \leq \sqrt{2d}$: if some coordinate $j \in [n]$ is contained in none of the sets $\operatorname{supp}_\delta x_i$, then $|x_{ij}| < \delta$ for all $i \in [r]$, so $|\sum_{i=1}^r x_{ij}| < r\delta \leq \sqrt{2d}\delta$. Now, the right-hand-side set in the above inclusion has size at most $|B|r \leq \sqrt{2d}|B|$, which contradicts the previous bound we obtained on the size of the left-hand-side set. This proves the lemma statement for $t = 1$.

Now suppose $t \geq 2$. Suppose that the lemma holds for all $r \leq (\sqrt{2d})^{t-1}$. Pick some $x_1, \dots, x_r \in \mathbb{R}^n$ with $r \leq (\sqrt{2d})^t$, satisfying $\|M_{B^c} x_i\|_\infty \leq \delta/\sqrt{2d}$ and $|\operatorname{supp}_\delta x_i| \leq |B|$ for all $i \in [r]$. Define vectors $y_1, \dots, y_{\sqrt{2d}} \in \mathbb{R}^n$ by $y_b = \sum_{i=1+(b-1)(\sqrt{2d})^{t-1}}^{\min(b(\sqrt{2d})^{t-1}, r)} x_i$. By the inductive hypothesis applied to each of these smaller sums, we have for all $b \in [\sqrt{2d}]$ that

$$|\operatorname{supp}_{(2d)^{3(t-1)/2}\delta} y_b| \leq |B|.$$

Moreover for each $b \in [\sqrt{2d}]$,

$$\|M_{B^c} y_b\|_\infty \leq \sum_{i=1+(b-1)(\sqrt{2d})^{t-1}}^{\min(b(\sqrt{2d})^{t-1}, r)} \|M_{B^c} x_i\|_\infty \leq (\sqrt{2d})^{t-1}\delta/\sqrt{2d} \leq (2d)^{3(t-1)/2}\delta/\sqrt{2d}.$$

By the inductive hypothesis applied to $y_1, \dots, y_{\sqrt{2d}}$ with threshold $(2d)^{3(t-1)/2}\delta$, we have that

$$\left|\operatorname{supp}_{(2d)^{3t/2}\delta} \sum_{b=1}^{\sqrt{2d}} y_b\right| \leq |B|$$

as desired. □

The previous lemma lets us bound the quantitative support of a sum of vectors. To show that this bounds the union of the supports, we also need the following lemma, which is proved by the probabilistic method.

Lemma C.5. *Let $\delta > 0$. Let $x_1, \dots, x_r \in \mathbb{R}^n$ and let $V = \bigcup_{i=1}^r \operatorname{supp}_\delta x_i$. Then there is some $\sigma \in \{-1, 1\}^r$ such that*

$$\left|\operatorname{supp}_\delta \sum_{i=1}^{n} \sigma_i x_i\right| \geq \frac{|V|}{2}.$$

Proof. Suppose we choose σ uniformly at random. For any $j \in V$, let $a \in [r]$ be such that $|x_{aj}| \geq \delta$. For any σ, if $|\sum \sigma_i x_{ij}| < \delta$, then if we flip σ_a to obtain a sign vector σ', it's necessary that $|\sum \sigma'_i x_{ij}| \geq \delta$. Thus, conditioned on σ_a it holds that $|\sum_i \sigma_i x_{ij}| \geq \delta$ with probability at least $1/2$, so unconditionally it holds with probability at least $1/2$. Thus,

$$\mathbb{E}_{\sigma \sim \{-1,1\}^r} \left| \operatorname{supp}_\delta \sum_{i=1}^r \sigma_i x_i \right| \geq \frac{|V|}{2}.$$

So there must exist at least one $\sigma \in \{-1,1\}^r$ satisfying the desired inequality. □

We can now bound the union of the quantitative supports of sparse approximate solutions to $M_{B^c} x = 0$, by picking n vectors whose quantitative supports cover the union, and adding them up with the signs suggested by Lemma C.5.

Lemma C.6. *Suppose that $\epsilon \leq 1/12$ and $|B| \leq k/(4\sqrt{d})$ and $d \geq 16$. Let $\delta > 0$. Define*

$$S = \{x \in \mathbb{R}^n : \|M_{B^c} x\|_\infty \leq \delta/\sqrt{2d} \wedge |\operatorname{supp}_\delta(x)| \leq |B|\}$$

and define

$$C = \bigcup_{x \in S} \operatorname{supp}_{g(n,d)\delta}(x)$$

where $g(n,d) = (2nd)^3$. Then $|C| \leq 2|B|$.

Proof. Since $|C| \leq n$, we can find $x_1, \dots, x_n \in \mathbb{R}^n$ such that $C = \bigcup_{i=1}^n \operatorname{supp}_{g(n,d)\delta}(x_i)$, and $\|M_{B^c} x_i\|_\infty \leq \delta/\sqrt{2d}$ and $|\operatorname{supp}_\delta(x)| \leq |B|$ for all $i \in [n]$. By Lemma C.5, there are some $\sigma_1, \dots, \sigma_n \in \{-1,1\}$ such that

$$\left| \operatorname{supp}_{g(n,d)\delta} \sum_{i=1}^n \sigma_i x_i \right| \geq \frac{|C|}{2}.$$

Let t be minimal so that $n \leq (\sqrt{2d})^t$, and note that $(\sqrt{2d})^t \leq 2nd$, so $(2d)^{3t/2} \leq (2nd)^3 = g(n,d)$. By Lemma C.4, which is applicable since $\|M_{B^c} \sigma_i x_i\|_\infty \leq \delta/\sqrt{2d}$ and $|\operatorname{supp}_\delta \sigma_i x_i| \leq |B|$ for all $i \in [n]$, we have that

$$\left| \operatorname{supp}_{g(n,d)\delta} \sum_{i=1}^n \sigma_i x_i \right| \leq \left| \operatorname{supp}_{(2d)^{3t/2}\delta} \sum_{i=1}^n \sigma_i x_i \right| \leq |B|.$$

It follows that $|C| \leq 2|B|$. □

From the above lemma, erasure-robustness of M is essentially immediate; it only remains to observe that although we took the union of all approximate solutions with quantitative supports of size at most $|B|$, this is equivalent to taking the union of all approximate solutions with quantitative supports of size less than k (for a slightly different threshold) due to density amplification.

Corollary C.7 (Erasure-robustness under Assumption 1)**.** *Suppose that $\epsilon \leq 1/12$ and $n, d \geq D_0$, for some universal constant D_0. For any $b \leq k/d$, the matrix M is $(b, 2b, d^4 n^6, k)$-erasure-robust (Definition 4). To restate: let $B \subseteq [m]$ satisfy $|B| \leq k/d$. Then there is a set $C \subseteq [n]$ with $|C| \leq 2|B|$, such that for any $x \in \mathbb{R}^n$, either:*

- $\|x_{C^c}\|_2 \leq d^4 n^6 \|M_{B^c} x\|_\infty$, *or*
- $|\operatorname{supp}(x)| \geq k$.

Proof. Pick arbitrary $\delta > 0$ and define S and C as in Lemma C.6, which guarantees that $|C| \leq 2|B|$. For any $x \in \mathbb{R}^n$, pick some $c \in \mathbb{R}^+$ such that $\|M_{B^c} cx\|_\infty \leq \delta/f(n,d)$, where $f(n,d) = 2dn^2$. If $|\operatorname{supp}_\delta(cx)| > |B|$, then by Corollary C.3, $|\operatorname{supp}_{\delta/f(n,d)}(cx)| \geq k$. As a result, $|\operatorname{supp}(x)| = |\operatorname{supp}(cx)| \geq k$. Otherwise, we have $|\operatorname{supp}_\delta(cx)| \leq |B|$. But we also know that $\|M_{B^c} cx\|_\infty \leq \delta/f(n,d) \leq \delta/d$. Thus, $cx \in S$. As a result, by definition of C, we get $\operatorname{supp}_{g(n,d)\delta}(cx) \subseteq C$, where $g(n,d) = (2nd)^3$. Therefore $\|cx_{C^c}\|_\infty \leq g(n,d)\delta$, so $\|cx_{C^c}\|_2 \leq \sqrt{n} g(n,d)\delta$.

If $\left\| M_{B^c} x \right\|_\infty > 0$, then we can choose $c = \delta / (f(n,d) \left\| M_{B^c} x \right\|_\infty)$, in which case $\left\| x_{C^c} \right\|_2 \le \sqrt{n} f(n,d) g(n,d) \left\| M_{B^c} x \right\|_\infty \le d^4 n^6$ for sufficiently large n, d. And if $\left\| M_{B^c} x \right\|_\infty = 0$, then we can choose c arbitrarily large, so $\left\| x_{C^c} \right\|_2 = 0$. In either case, it holds that $\left\| x_{C^c} \right\|_2 \le d^4 n^6 \left\| M_{B^c} x \right\|_\infty$. □

All that remains to prove Theorem 2.3 is showing that a random sparse binary matrix satisfies Assumption 1 with high probability. However, for the application to lower bounds against Preconditioned Lasso, we also must recall the following result (essentially due to [6] and slightly generalized in [32]), which states that under the degree bound and expansion assumptions, any vector in the kernel of M must be quantitatively dense.

Lemma C.8 (Lemma 8.7 in [32])**.** *Let $x \in \mathbb{R}^n$ be such that $Mx = 0$. Let $S \subseteq [n]$ with $|S| \le k$. If $\epsilon \le 1/17$, then*

$$\left\| x_S \right\|_1 \le \frac{\left\| x \right\|_1}{3}.$$

C.2 Random unbalanced bipartite graph

In this section we show that a random sparse binary matrix satisfies Assumption 1 with high probability. Let $n, d, m \in \mathbb{N}$ with $d \le m \le n$, and let $p = d/m$. Define a random matrix $M \in \mathbb{R}^{m \times n}$ with independent entries $M_{ij} \sim \text{Ber}(p)$.

The following result is folklore (see e.g. [1]):

Lemma C.9 (Expansion of a random bipartite graph)**.** *Let $\epsilon \in (0,1)$. Suppose that $p \ge 32\epsilon^{-2}(\log n)/m$. It holds with probability at least $1 - 2/n$ that for all $S \subseteq [n]$ with $|S| \le \epsilon/(2p)$,*

$$|N(S)| \ge d(1-\epsilon)|S|.$$

Proof. For $1 \le l \le \epsilon/(2p)$ let q_l be the probability that there exists some $S \subseteq [n]$ with $|S| = l$ and $|N(S)| < d(1-\epsilon)|S|$. To bound this probability, fix $S \subseteq [n]$ with $|S| = l$. For any $y \in [m]$, we have

$$\Pr[y \in N(S)] = 1 - (1-p)^l \ge 1 - e^{-pl} \ge pl - (pl)^2 \ge (1 - \epsilon/2)pl$$

so long as $0 \le pl \le \epsilon/2$. Thus, by the Chernoff bound,

$$\Pr[|N(S)| < (1-\epsilon)plm] \le \exp(-(\epsilon/2)^2(1-\epsilon/2)plm/2) \le \exp(-\epsilon^2 plm/16).$$

By the union bound over sets of size l, if $\epsilon^2 pm/16 \ge 2\log n$, we have that

$$q_l \le \binom{n}{l} \exp(-\epsilon^2 plm/16) \le \exp(l \log n - \epsilon^2 plm/16) \le \exp(-l \log n).$$

Finally, by a union bound over $1 \le l \le \epsilon/(2p)$, the lemma holds with probability at least

$$1 - \sum_{l=1}^{\epsilon/(2p)} q_l \ge 1 - \sum_{l=1}^{\epsilon/(2p)} n^{-l} \ge 1 - \frac{2}{n}$$

as claimed. □

We'll also need the following simple result:

Lemma C.10 (Degree bounds)**.** *Let $\epsilon \in (0,1)$. Suppose that $p \ge 6\epsilon^{-2}(\log n)/m$. It holds with probability at least $1 - 1/n$ that*

$$|N(x)| \le d(1+\epsilon)$$

for all $x \in [n]$. Similarly, it holds with probability at least $1 - 1/n$ that $|N'(y)| \le (n/m)d(1+\epsilon)$ for all $y \in [m]$.

Proof. Fix $x \in [n]$. By the Chernoff bound,

$$\Pr[|N(x)| > (1+\epsilon)pm] \le \exp(-\epsilon^2 pm/3).$$

Since $\epsilon^2 pm/3 \ge 2\log n$, this bound is at most $1/n^2$. Union bounding over $x \in [n]$ completes the proof of the first claim.

Similarly, fix $y \in [m]$. By the Chernoff bound,

$$\Pr[|\{x : y \in N(x)\}| > (1+\epsilon)pn] \leq \exp(-\epsilon^2 pn/3).$$

Since $\epsilon^2 pn/3 \geq 2\log n$, this bound is at most $1/n^2$, and union bounding over $y \in [m]$ completes the proof. □

To prove the last condition of Assumption 1, we need the following version of the Chernoff-Hoeffding bound:

Lemma C.11 (Chernoff-Hoeffding)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with values in $\{0,1\}$. Let $\mu = \mathbb{E}\sum X_i$. Then for any $t > 0$,*

$$\Pr\left[\sum_{i=1}^n X_i > t\right] \leq \exp(-t\log(t/(\mu e))).$$

Now, the proof of the bounded intersection property is analogous to the proof of expansion.

Lemma C.12. *Suppose that $d \geq \log^2 n$ and n is sufficiently large. It holds with probability at least $1 - 2/m^2$ that for all disjoint $S, T \subseteq [n]$ with $|S|, |T| \leq m/d^3$,*

$$|N(S) \cap N(T)| \leq \frac{\sqrt{d}}{8}\max(|S|, |T|).$$

Proof. Fix $1 \leq l \leq m^3/(n^2 d^7)$. Let q_l be the probability that some sets S, T of size exactly l violate the inequality. Fix disjoint $S, T \subseteq [n]$ with $|S|, |T| = l$. For any $i \in [m]$, we have that

$$\Pr[i \in N(S)] = 1 - (1-p)^l \leq pl.$$

Thus, since S and T are disjoint,

$$\Pr[i \in N(S) \cap N(T)] \leq p^2 l^2.$$

Let $D = \sqrt{d}/8 \geq 3$. By the Chernoff-Hoeffding bound (Lemma C.11), we get

$$\Pr[|N(S) \cap N(T)| > Dl] \leq \exp(-Dl\log(Dl/(p^2 l^2 me))) \leq \left(\frac{D}{p^2 lme}\right)^{-Dl}.$$

By the union bound over sets S, T of size l,

$$q_l \leq \left(\frac{D}{p^2 lme}\right)^{-Dl}\left(\frac{en}{l}\right)^{2l} \leq (p^{2D} m^D l^{D-2} n^2)^l = \left(\frac{d^{2D} l^{D-2} n^2}{m^D}\right)^l.$$

So long as $l \leq m/d^3$ and $d \geq \log^2 n$, we have $d^{2D} l^{D-2} n^2/m^D \leq n^2/(m^2 d^{D-6}) \leq 1/m^2$ for large n. As a result, summing over l, we have

$$\sum_{l=1}^{m/d^3} q_l \leq \sum_{l=1}^{m/d^3} (1/m^2)^l \leq 2/m^2$$

as claimed. □

Together, the above three lemmas immediately imply that the random matrix M satisfies Assumption 1 with high probability.

Theorem C.13. *Let n be an even positive integer that is sufficiently large. Let $m, d \in \mathbb{N}$ with $d \leq m \leq n$, and let $\epsilon \in (0,1)$. If $d \geq 32\epsilon^{-2}\log n$ and $d \geq \log^2 n$, then with probability at least $1 - 4/n - 2/m^2$, the random binary matrix $M \in \mathbb{R}^{m \times n}$ with i.i.d. entries $M_{ij} \sim Ber(d/m)$ satisfies Assumption 1 with sparsity parameter $k = m/d^3$ and error parameter ϵ.*

Putting together Theorem C.13 and Corollary C.7 immediately gives the following.

Theorem C.14. *There are constants N, C with the following property. Let $n, m \in \mathbb{N}$ with $n \geq N$. Let $p \in (0,1)$ satisfy $p \geq C(\log^2 n)/m$. Let M be the $m \times n$ random matrix with independent entries $M_{ij} \sim Ber(p)$. Then with probability $1 - O(1/n) - O(1/m^2)$, it holds that for all $b \leq m/d^4$, the matrix M is $(b, 2b, d^4 n^6, m/d^3)$-erasure-robust.*

C.3 Properties of Final Construction

We can now prove the following existence result, which collects all the important properties of the matrix M which we will use for our lower bound against Preconditioned Lasso: sparsity, quantitative density of the kernel, erasure-robustness, and eigenvalue bounds.

Theorem C.15. *Let $n \in \mathbb{N}$ be an even number larger than some absolute constant n_0. There is a density parameter $\tau = \Omega(n/\log^6 n)$, $\eta = O(n^6 \log^8 n)$, $b = \Omega(n/\log^8 n)$, and a matrix $M \in \mathbb{R}^{n/2 \times n}$ with the following properties:*

1. *The rows of M are $O(\log^2 n)$-sparse*
2. *For any $x \in \mathbb{R}^n$ with $\operatorname{dist}(x, \ker M) \le \|x\|_2 / (12\sqrt{n})$, and any $S \subseteq [n]$ with $|S| \le \tau$, it holds that $\|x_{S^c}\|_2 \ge \|x\|_2 / (2\sqrt{n})$.*
3. *For any $B \subseteq [n/2]$ with $|B| \le b$, there is a set $C \subseteq [n]$ with $|C| \le 2|B|$, such that for any $x \in \mathbb{R}^n$, either*
   - *$\|x_{C^c}\|_2 \le \eta \|M_{B^c} x\|_\infty$, or*
   - *$|\operatorname{supp}(x)| \ge \tau$.*
4. *$\|M^T M\|_F \le O(n \log n)$*
5. *The smallest nonzero eigenvalue λ of Θ satisfies $\lambda \ge \Omega(n^{-5/2})$.*

Proof. Let $\epsilon = 1/17$ and $d = \log^2 n$. Let $M \in \mathbb{R}^{n/2 \times n}$ be the random binary matrix with independent entries $M_{ij} \sim \mathrm{Ber}(2d/n)$. By Theorem C.13, M satisfies Assumption 1 with $k = n/\log^6 n$ and error ϵ, with probability at least $1 - 12/n$. Let $\tau = k$. Claim (1) follows from the degree bound condition.

To prove claim (2), let $x \in \mathbb{R}^n$ with $\operatorname{dist}(x, \ker M) \le \|x\|_2 / (12\sqrt{n})$, and let $C \subseteq [n]$ with $|C| \le k$. Let $y = \mathrm{Proj}_{\ker M}\, x$. Then $My = 0$, so $\|y_C\|_1 \le \|y\|_1 / 3$ by Lemma C.8. Therefore

$$\begin{aligned} \|x_{C^c}\|_2 &\ge \|y_{C^c}\|_2 - \frac{1}{12\sqrt{n}} \|x\|_2 \\ &\ge \frac{1}{\sqrt{n}} \|y_{C^c}\|_1 - \frac{1}{12\sqrt{n}} \|x\|_2 \\ &\ge \frac{2}{3\sqrt{n}} \|y\|_1 - \frac{1}{12\sqrt{n}} \|x\|_2 \\ &\ge \frac{2}{3\sqrt{n}} \|y\|_2 - \frac{1}{12\sqrt{n}} \|x\|_2 \\ &\ge \frac{1}{2\sqrt{n}} \|x\|_2 \end{aligned}$$

as desired.

Claim (3) follows from Corollary C.7 with $b = k/d$. Claim (4) follows from the degree bound assumption, and claim (5) holds with probability at least $1/2$ by results from random matrix theory (Theorem 1.1 in [4]) , and the observation that $\lambda = \sigma^2$, where σ is the smallest singular value of M^T.

Thus, all claims hold with probability at least $1/2 - 12/n > 0$, so in particular the desired matrix M exists. □

D Structure Lemma under Erasure-Robustness

In this section we show how erasure-robust sparse designs M can be used to construct a covariance matrix Σ so that preconditioners which are "compatible" with most rows of M (i.e. succeed at recovery with non-trivial probability when the covariates are drawn from $N(0, \Sigma)$ and the signal is the row of M) satisfy a useful structure lemma. Concretely, the covariance matrix we use to fool the preconditioned Lasso is given by

$$\tilde{\Sigma} := \tilde{\Theta}^{-1}, \qquad \tilde{\Theta} := \Theta + \epsilon I_n, \qquad \Theta = M^T M$$

where $\epsilon > 0$ is polynomially small, so that $\tilde{\Sigma}$ still has polynomial condition number. Ultimately, we will instantiate M as the matrix constructed in the previous section, but for now we state our results in generality. We will show that if M is erasure-robust and a preconditioner is compatible with most rows of M, then the columns of the preconditioner are either dense or have small magnitude outside a small set of coordinates. Moreover, we'll show that the number of dense columns must be linear.

D.1 Failure on incompatible signals

To start with, we recall the weak compatibility coefficients $\alpha^{(1)}$ and $\beta^{(1)}$ from [32], which for any fixed covariance matrix and preconditioner, provide a simple necessary condition for the success of the (Preconditioned) Lasso. (We remind the reader that for fixed w^*, there also exists a sharp characterization of its failure/success due to Gordon's theorem (see e.g. [2]), though the exact characterization is not as easy to work with directly.)

Definition 7 (Weak S-Preconditioned Compatibility Condition [32])**.** We say that

$$\alpha^{(1)}_{\Sigma,S,k} = \inf_{w \in B_0(k) \setminus \{0\}} \frac{\langle w, \Sigma w \rangle}{\left\| S^T w \right\|_1^2}$$

where we let $B_0(k)$ denotes the set of k-sparse vectors and

$$\beta^{(1)}_{\Sigma,S,m,k} = \sup\{\beta \in \mathbb{R} : \dim W_{\Sigma,S,\beta} \geq 2m\}$$

where

$$W_{\Sigma,S,\beta} = \left\{ w : \langle w, \Sigma w \rangle \geq \beta \left\| S^T w \right\|_1^2 \right\},$$

and $\dim W_{\Sigma,S,\beta}$ is defined as the largest dimension of any subspace contained in $W_{\Sigma,S,\beta}$.

Remark 1. Since we only deal with invertible covariance matrices Σ in this paper, it always holds that $\alpha^{(1)}_{\Sigma,S,k} > 0$. Moreover, so long as S is not identically zero, $\alpha^{(1)}_{\Sigma,S,k}$ is finite. We also assume that S^T has trivial kernel, which implies that $\beta^{(1)}_{\Sigma,S,m,k}$ is finite. This assumption is essentially without loss of generality; see Section B.1 for discussion.

In [32], it shown that if the ratio $\beta^{(1)}_{\Sigma,S,m,k}/\alpha^{(1}_{\Sigma,S,m,k}$ exceeds a certain constant, then there is a k-sparse signal such that the S-preconditioned Lasso with m samples fails with high probability. However, this signal depends on S (specifically, it's the signal for which $\alpha^{(1)}_{\Sigma,S,k}$ achieves the infimum). Because we ultimately need to construct a signal distribution independent of S, we need a slightly more general statement which provides a condition under which a given signal causes S-preconditioned Lasso to fail. Such a statement is in fact implicit in [32]:

Theorem D.1 ([32])**.** *Let $\Sigma \in \mathbb{R}^{n \times n}$ be positive-definite and let $S \in \mathbb{R}^{n \times s}$. Let $m, k \in \mathbb{N}$. If $w^* \in \mathbb{R}^n$ is a k-sparse signal with*

$$(w^*)^T \Sigma w^* < \frac{\beta^{(1)}_{\Sigma,S,m,k}}{18} \left\| S^T w^* \right\|_1^2,$$

then the S-preconditioned Lasso exactly recovers w^* with probability at most $\exp(-\Omega(m))$, from m samples with independent covariates $X_1, \ldots, X_m \sim N(0, \Sigma)$ and noiseless responses $Y_i = \langle w^*, X_i \rangle$.

Proof. The proof of this result is implicit in (Theorem 6.5, [32]) and we include it for the reader's convenience. For convenience of notation let $\beta = \beta^{(1)}_{\Sigma,S,m,k}$ and $\Theta = \Sigma^{-1}$. We want to show that with high probability, the S-preconditioned Lasso (1) fails to recover w^*, i.e.

$$w^* \notin \operatorname*{argmin}_{w : Xw = Xw^*} \left\| S^T w \right\|_1$$

where X has rows $X_1, \dots, X_m \sim N(0, \Sigma)$. We know that

$$(w^*)^T \Sigma w^* = \alpha \left\|S^T w^*\right\|_1^2$$

for some $\alpha < \beta/18$.

By definition of β, there is a subspace $U \subseteq \mathbb{R}^n$ of dimension $2m$ such that $w^T \Sigma w \geq \beta \left\|S^T w\right\|_1^2$ for all $w \in U$. Let $v_1, \dots, v_{2m} \in U$ form an orthonormal basis for U, and let $V \in \mathbb{R}^{n \times 2m}$ be the matrix with columns $v_1, \dots, v_{2m}$.

We construct $v \in \mathbb{R}^n$ to satisfy $Xw^* = Xv$ and $\left\|S^T v\right\|_1 < \left\|S^T w^*\right\|_1$ as follows. Let $\Gamma = V^T \Sigma V \in \mathbb{R}^{2m \times 2m}$. The columns of V have no linear dependencies, and Σ is symmetric positive-definite, so Γ is symmetric positive-definite. Thus, there is an invertible matrix $N \in \mathbb{R}^{2m \times 2m}$ such $\Gamma = N^T N$. Define

$$c = N^{-1}(XVN^{-1})^\dagger X w^* \in \mathbb{R}^{2m}$$

and define $v = Vc \in \mathbb{R}^n$. By construction we have $v \in U$, so

$$v^T \Sigma v \geq \beta \left\|S^T v\right\|_1^2. \tag{2}$$

Second, note that

$$\mathbb{E}[(XVN^{-1})^T (XVN^{-1})] = m(N^{-1})^T V^T \Sigma V N^{-1} = m(N^{-1})^T \Gamma N^{-1} = mI_{2m}.$$

Moreover, the rows of XVN^{-1} are independent and Gaussian. So in fact XVN^{-1} has i.i.d. $N(0,1)$ entries. Thus, with probability $1 - \exp(-\Omega(m))$, we have $\sigma_{\text{min}}((XVN^{-1})^T) \geq \sqrt{m}/3$ since the dimensions of $(XVN^{-1})^T$ are $2m \times m$ (by Theorem B.1). Hence, $\sigma_{\text{max}}((XVN^{-1})^\dagger) \leq 3/\sqrt{m}$. We can conclude that

$$v^T \Sigma v = c^T N^T N c = (w^*)^T X^T (XVN^{-1})^{\dagger T} (XVN^{-1})^\dagger X w^* \leq (9/m)(w^*)^T X^T X w^*. \tag{3}$$

We can now check that $\left\|S^T v\right\|_1 < \left\|S^T w\right\|_1$. Indeed, by Theorem B.2, $(w^*)^T X^T X w^* \leq 2m(w^*)^T \Sigma w^*$ with probability at least $1 - \exp(-\Omega(m))$, so

$$\begin{aligned}
\left\|S^T v\right\|_1 &\leq \sqrt{\frac{1}{\beta} v^T \Sigma v} \\
&\leq \sqrt{\frac{9}{m\beta}(w^*)^T X^T X w^*} \\
&\leq \sqrt{\frac{18}{\beta}(w^*)^T \Sigma w^*} \\
&\leq \sqrt{\frac{18\alpha}{\beta}} \left\|S^T w^*\right\|_1
\end{aligned}$$

which produces the desired inequality as long as $\beta/\alpha > 18$.

Finally, since XVN^{-1} is rank-m with probability 1, we have $(XVN^{-1})(XVN^{-1})^\dagger = I_m$, and thus

$$Xv = XVN^{-1}(XVN^{-1})^\dagger X w^* = Xw^* \tag{4}$$

as desired. □

D.2 Structure Lemma for compatible preconditioners

We now consider the case where the preconditioner does a good job of preconditioning the rows M_i, so that Lasso can succeed with non-trivial probability for most such signals. Then very informally, one would think this forces $\Sigma \approx SS^T$, which by the construction of Σ will force S to be dense, and hence the preconditioner will fail on basis vectors.

The following Lemma starts to make this intuition correct and precise. It shows that if the preconditioner is compatible with most of the rows of M, then the "sparse part" of the preconditioner S is not too large. The "sparse part" is actually given not just by removing the dense rows of S^T but also by removing a small number b' of columns of S^T: this is inevitable because the assumption that the preconditioner is compatible with *most* of the rows of M cannot imply something about *all* of the columns of S^T, and this is where erasure-robustness is crucially used.

Lemma D.2 (Structure Lemma)**.** *Let $M \in \mathbb{R}^{n-r\times n}$, and define $\Theta = M^TM$. Let λ be the smallest nonzero eigenvalue of Θ. Let $\epsilon > 0$, and let $\tilde{\Theta} = \Theta + \epsilon I$. Let $k, m, \alpha, \tau, b, b', \eta > 0$. Suppose that M satisfies (b, b', η, τ)-erasure-robustness (Definition 4).*

Let $S \in \mathbb{R}^{n\times s}$. Suppose that

$$\Pr_{i\in[n-r]}\left[M_i^T\tilde{\Theta}^{-1}M_i < \alpha\left\|S^TM_i\right\|_1^2\right] \le \frac{b}{n-r}$$

and define $\gamma = \beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}/\alpha$. Let

$$D := \{i \in [s] : \|(S^T)_i\|_0 \ge \tau\}$$

be the set of τ-dense columns of S (rows of S^T). There exists a subset of row indices $C \subseteq [n]$ with $|C| \le b'$ such that the submatrix $S_{C^cD^c}$ satisfies

$$\sum_{j\in D^c}\|S_{C^cj}\|_2 \le \frac{n^{3/2}\eta\,\|M\|_F}{\sqrt{\lambda\alpha}}.$$

Proof. Let $B \subseteq [n-r]$ be the set of $i \in [n-r]$ such that $M_i^T\tilde{\Theta}^{-1}M_i < \alpha\left\|S^TM_i\right\|_1^2$. By assumption, $|B| \le b$; let $C := C_B \subseteq [n]$ be the set guaranteed by erasure-robustness, which indeed satisfies $|C| \le b'$. For any $i \notin B$, we have

$$\left\|S^TM_i\right\|_1^2 \le \frac{1}{\alpha}M_i^T\tilde{\Theta}^{-1}M_i \le \frac{\|M_i\|_2^2}{\lambda\alpha},$$

where the last inequality is because $M_i \in \operatorname{span}\Theta$. Expanding the ℓ_1 norm as a sum, and summing $\left\|S^TM_i\right\|_1$ over $i \notin B$, we have

$$\sum_{j\in[s]}\sum_{i\notin B}|\langle(S^T)_j, M_i\rangle| \le \frac{n\sum_{i\notin B}\|M_i\|_2}{\sqrt{\lambda\alpha}} \le \frac{n^{3/2}\,\|M\|_F}{\sqrt{\lambda\alpha}}.$$

Thus,

$$\sum_{j\in[s]}\left\|M_{B^c}(S^T)_j\right\|_\infty \le \sum_{j\in[s]}\left\|M_{B^c}(S^T)_j\right\|_1 \le \frac{n^{3/2}\,\|M\|_F}{\sqrt{\lambda\alpha}}.$$

As specified in the theorem statement, let $D \subseteq [s]$ be the set of $i \in [s]$ such that $\left\|(S^T)_i\right\|_0 \ge \tau$. Then for any $i \in D^c$, we have that $\left\|(S^T)_{iC^c}\right\|_2 \le \eta\left\|M_{B^c}(S^T)_i\right\|_\infty$. Thus,

$$\sum_{j\in D^c}\left\|(S^T)_{jC^c}\right\|_2 \le \eta\sum_{j\in D^c}\left\|M_{B^c}(S^T)_j\right\|_\infty \le \frac{n^{3/2}\eta\,\|M\|_F}{\sqrt{\lambda\alpha}}$$

as claimed. □

As a consequence of the previous lemma, we can show that if the compatibility ratio is small then S must have many dense columns (S^T has many dense rows).

Lemma D.3. *In the setting of Lemma D.2, suppose that $1/\sqrt{\beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}\epsilon} > n^{3/2}\eta\,\|M\|_F/\sqrt{\lambda\alpha}$. Then $|D| > n - r - b' - 2m$.*

Proof. Suppose for contradiction that $|D| \le n-r-b'-2m$. Pick some $\beta' > \beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}$ sufficiently small that $1/\sqrt{\beta'\epsilon} > n^{3/2}\eta\,\|M\|_F/\sqrt{\lambda\alpha}$. Define

$$W = \ker(S_D^T) \cap \ker(\Theta) \cap \operatorname{span}\{e_i : i \in C^c\} \subseteq \mathbb{R}^n.$$

Then

$$\dim(W) \ge n - |D| - r - |C| \ge 2m.$$

For any $w \in W$, we have

$$w^T \tilde{\Theta}^{-1} w \geq \epsilon^{-1} \left\| \mathrm{Proj}_{\ker \Theta} w \right\|_2^2 = \epsilon^{-1} \left\| w \right\|_2^2$$

since $w \in \ker \Theta$. On the other hand,

$$\begin{aligned} \left\| S^T w \right\|_1 &= \sum_{j \in D^c} |\langle (S^T)_j, w \rangle| && (w \in \ker(S_D^T)) \\ &= \sum_{j \in D^c} |\langle (S^T)_{jC^c}, w_{C^c} \rangle| && (\mathrm{supp}(w) \subseteq C^c) \\ &\leq \sum_{j \in D^c} \left\| (S^T)_{jC^c} \right\|_2 \left\| w \right\|_2 && \text{(Cauchy-Schwarz)} \\ &\leq \frac{n^{3/2} \eta \left\| M \right\|_F}{\sqrt{\lambda \alpha}} \left\| w \right\|_2 . && \text{(Lemma D.2)} \end{aligned}$$

As a consequence, by choice of β', we have $w^T \tilde{\Theta}^{-1} w \geq \beta' \left\| S^T w \right\|_1^2$. Therefore $W \subseteq W_{\tilde{\Theta}^{-1}, S, \beta'}$, contradicting the definition of $\beta_{\tilde{\Theta}^{-1}, S, k, m}$. □

E Failure of the Preconditioned Lasso

In this section, we prove Theorems 2.1 and 2.2. We start by proving our lower bound against invertible preconditioners. We then prove a key projection lemma and use it to prove the lower bound against rectangular preconditioners.

E.1 Invertible Preconditioners

To construct a signal distribution which fails invertible preconditioners with probability $1 - o(1)$, we'll need to amplify the failure probability by adding together multiple signals. The following lemma formalizes why this works in certain cases: a random combination of vectors where at least one of them has large ℓ_1 norm is very unlikely to have small ℓ_1 norm.

Lemma E.1. *Let $\mathcal{D}$ be a continuous distribution on $\mathbb{R}$ with density upper bounded by $1/2$. Let $v_1, \dots, v_t \in \mathbb{R}^n$ and let $Z_1, \dots, Z_t \sim \mathcal{D}$ be independent random variables. Then*

$$\Pr\left[\left\| \sum_{i=1}^{t} Z_i v_i \right\|_1 < \delta \max_{i \in [t]} \left\| v_i \right\|_1 \right] \leq \delta.$$

Proof. Without loss of generality assume that $\left\| v_t \right\|_1 = \max_{i \in [t]} \left\| v_i \right\|_1$. Condition on $Z_1, \dots, Z_{t-1}$ and define

$$f(z) = \left\| \sum_{i=1}^{t-1} Z_i v_i + z v_t \right\|_1 .$$

For any z, z' with $f(z) \leq \delta \left\| v_t \right\|_1$ and $f(z') \leq \delta \left\| v_t \right\|_1$ we have by the triangle inequality that $\left\| z v_t - z' v_t \right\|_1 \leq 2\delta \left\| v_t \right\|_1$, so $|z - z'| \leq 2\delta$. Since the density of $\mathcal{D}$ is upper bounded by $1/2$, it follows that

$$\Pr[f(Z_t) < \delta \left\| v_t \right\|_1] \leq \frac{2\delta}{2} = \delta$$

which proves the claim. □

We know show that if M is a sparse compressive erasure-robust design matrix (e.g. as we constructed previously), then the covariance matrix $\Sigma = (M^T M + \epsilon I)^{-1}$ together with an appropriate signal distribution describes a hard distribution family for (invertibly) Preconditioned Lasso.

Theorem E.2. *Let $n, r \in \mathbb{N}$. Let $M \in \mathbb{R}^{n-r \times n}$ and $\epsilon > 0$. Define $\Theta = M^T M$ and $\tilde{\Theta} = \Theta + \epsilon I$. Let λ be the smallest nonzero eigenvalue of Θ. Let $k, m, \alpha, \tau, \eta, b, b', t > 0$ and suppose that M satisfies (b, b', η, τ)-erasure-robustness and has k-sparse rows.*

Suppose $k > 2(n/\tau)\log(n)$ and $n - r - b' \geq 3m$. There is a distribution $\mathcal{D}$ on $k(t+1)$-sparse signals in $\mathbb{R}^n$ with the following property.

Let $S \in \mathbb{R}^{n\times n}$ be invertible. Suppose that

$$9n^{105/2}\eta \left\|M\right\|_F \sqrt{\epsilon/\lambda} < 1. \tag{5}$$

Then with probability at least $1 - \exp(-tb/(n-r)) - 1/n^{50}$ over true signals $w^* \sim \mathcal{D}$, it holds that S-preconditioned Lasso fails with probability $1 - \exp(-\Omega(m))$ over independent samples $X_1, \dots, X_m \sim N(0, \tilde{\Theta}^{-1})$: that is, w^* is not a unique minimizer of $\left\|S^T w\right\|_1$ subject to $Xw = Xw^*$.

Proof. Let $\mathcal{D}$ be the signal distribution where we draw independent and uniformly random indices $R_1, \dots, R_t \in [n-r]$ as well as independent $Z_0, Z_1, \dots, Z_t \sim \text{Unif}([-1,1])$, and draw $\tilde{w}$ with uniformly random k-sparse support and entries $\text{Unif}([-1,1])$ on that support, and set the signal to be

$$w^* = Z_0\sqrt{\epsilon}\tilde{w} + \sum_{i=1}^t \frac{Z_i M_{R_i}}{\sqrt{M_{R_i}^T \tilde{\Theta}^{-1} M_{R_i}}}.$$

Pick any invertible $S \in \mathbb{R}^{n\times n}$. Define $\alpha = \beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}/(72n^{102})$. We distinguish two cases.

Case I: incompatible preconditioner. On the one hand, suppose that

$$\Pr_{i\in[n-r]}\left[M_i\tilde{\Theta}^{-1}M_i < \alpha \left\|S^T M_i\right\|_1^2\right] \geq \frac{b}{n-r}.$$

Then with probability at least

$$\left(1 - (1 - b/(n-r))^t\right) \geq \left(1 - e^{-tb/(n-r)}\right)$$

over the row indices $R_1, \dots, R_t$, there is some R_i with

$$\left\|S^T M_{R_i}\right\|_1^2 > \frac{1}{\alpha} M_{R_i}^T \tilde{\Theta}^{-1} M_{R_i}.$$

Under this event, we have

$$\max_{i\in[t]} \frac{\left\|S^T M_{R_i}\right\|_1}{\sqrt{M_{R_i}^T \tilde{\Theta}^{-1} M_{R_i}}} > \frac{1}{\sqrt{\alpha}},$$

so by Lemma E.1, it holds with probability at least $1 - 1/n^{50}$ over $Z_1, \dots, Z_t$ that $\left\|S^T w^*\right\|_1 \geq \frac{1}{n^{50}\sqrt{\alpha}}$. But by the triangle inequality, we have $\sqrt{(w^*)^T \tilde{\Theta}^{-1} w^*} \leq t + \sqrt{k} \leq 2n$ (since $\tilde{w}^T\tilde{\Theta}^{-1}\tilde{w} \leq \epsilon^{-1}\left\|\tilde{w}\right\|_2^2$). Thus,

$$\Pr_{w^*\sim\mathcal{D}}\left[\left\|S^T w^*\right\|_1^2 > \frac{1}{4n^{102}\alpha}(w^*)^T\tilde{\Theta}^{-1}w^*\right] \geq \left(1 - e^{-tb/(n-r)}\right)\left(1 - \frac{1}{n^{50}}\right).$$

Moreover, for such w^*, by choice of α and by Theorem D.1, the S-preconditioned Lasso recovers w^* with probability at most $\exp(-\Omega(m))$, from m independent samples $X_1, \dots, X_m \sim N(0, \tilde{\Theta}^{-1})$.

Case II: compatible preconditioner. On the other hand, suppose that

$$\Pr_{i\in[n-r]}\left[M_i\tilde{\Theta}^{-1}M_i \geq \alpha \left\|S^T M_i\right\|_1^2\right] \geq \frac{b}{n-r}.$$

By choice of α and the theorem assumptions, we know that $1/\sqrt{\beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}\epsilon} > n^{3/2}\eta\left\|M\right\|_F/\sqrt{\lambda\alpha}$. So we can apply Lemma D.2 and Lemma D.3: there is a set $D \subseteq [s]$ satisfying $|D| > n - r - b' - 2m$, and $\left\|(S^T)_j\right\|_0 \geq \tau$ for all $j \in D$. Let $U = \text{supp}(S^T w^*)$. Then since the support of $\tilde{w}$ is uniformly random of size $k \geq 2(n/\tau)\log(n)$, we know that $U \supseteq \text{supp}(S^T\tilde{w}) \supseteq D$ with probability at least $1 - 1/n$. Let $V = \{d \in \mathbb{R}^n : \text{supp}(S^T d) \subseteq U\}$ and let $z = \text{sign}(S^T w^*)$. We claim that there is some $d \in V$ with $Xd = 0$ but $\langle d, Sz\rangle \neq 0$. Indeed, since S^T is invertible, it suffices to show that there is a vector $f \in \mathbb{R}^n$ supported on U, such that $\langle S^{-1}X_i, f\rangle = 0$ for all

$i \in [m]$, but $\langle z, f\rangle \neq 0$. This holds because z is with probability 1 outside the span of the vectors $\{(S^{-1}X_i)_U : i \in [m]\}$, and the number of degrees of freedom is $|U| \geq |D| > n - r - b' - 2m \geq m$. We conclude that the desired direction of improvement d exists, so w^* is not a minimizer of the S-preconditioned Lasso.

□

Remark 2. The term $1/n^{50}$ in the probability of failure in Theorem E.2 and Theorem E.3 can be replaced by $1/n^{\ell}$ for any particular ℓ if we modify the left hand side of (5) accordingly. All that happens if we pick a larger ℓ is that to satisfy (5), we need to pick a correspondingly (polynomially) smaller ϵ which means that the covariance matrix of the data, $\Sigma = \tilde{\Theta}^{-1}$, becomes more ill-conditioned. We stated the results with $\ell = 50$ only to simplify the statements.

Finally, we instantiate the above theorem with the parameters of the design matrix M we constructed in Theorem C.15.

Theorem E.3. *Let $n \in \mathbb{N}$ be sufficiently large. There is a matrix $\Sigma \in \mathbb{R}^{n\times n}$ with condition number* $\operatorname{poly}(n)$ *such that the following holds. Let $k \in \mathbb{N}$ with $k \geq \log^8 n$. There is a distribution $\mathcal{D}$ over $O(k \log^9 n)$-sparse signals such that for any positive integer $m \leq n/7$ and any invertible preconditioner $S \in \mathbb{R}^{n\times n}$, with probability at least $1 - O(1/n^{50})$ over $w^* \sim \mathcal{D}$, the S-preconditioned Lasso recovers w^* with probability at most* $\exp(-\Omega(m))$ *from m independent samples $X_1, \dots, X_m \sim N(0, \Sigma)$ with noiseless responses $Y_i = \langle X_i, w^*\rangle$.*

Proof. Let Θ be the matrix guaranteed by Theorem C.15. We check the conditions of Theorem E.2. First, $\dim\ker(\Theta) \geq n/2$. Second, $\lambda = \Omega(n^{-5/2})$. We can take $b = n/(\log^8 n)$, $b' = 2n/(\log^8 n)$, $\eta = n^6 \log^8 n$, and $\tau = \Omega(n/\log^7 n)$. We have $\|M\|_F \leq O(\sqrt{n \log n})$. Thus, we can take $\epsilon = \Omega(n^{-111})$. We know that the rows of M are k-sparse. Let $t = 2\log^9 n$.

Applying Theorem E.2, there is a distribution $\mathcal{D}$ over $O(k \log^9 n)$-sparse signals such that for any invertible preconditioner $S \in \mathbb{R}^{n\times n}$, with probability at least $1 - O(1/n^{50})$ over $w^* \sim \mathcal{D}$, the S-preconditioned Lasso recovers w^* uniquely with probability $\exp(-\Omega(m))$ from m samples, so long as $m \leq n/7$ (so that $n - r - b' \geq 3m$). □

E.2 Projection Lemma

To extend our lower bound to rectangular preconditioners, we need Lemma E.6, a projection lemma generalizing an analogous result from [32]. To prove it, we recall two lemmas which are essentially taken from [32]; the second of these is the original projection lemma.

First, recall that our covariance matrix has the form $\tilde{\Sigma} = \tilde{\Theta}^{-1}$, where $\tilde{\Theta} = \Theta + \epsilon I$ for some PSD matrix Θ. The following lemma establishes that if ϵ is sufficiently small relative to the smallest nonzero eigenvalue of Θ, then the row span of the design matrix X is nearly orthogonal to all but the top eigenspace of the covariance $\tilde{\Sigma} = \left(\tilde{\Theta}\right)^{-1}$ (i.e. the kernel of Θ). In other words, by taking ϵ small enough the top eigenspace dominates, as expected.

Lemma E.4 ([32])**.** *Let $\Theta \in \mathbb{R}^{n\times n}$ be a PSD matrix with minimum nonzero eigenvalue λ. Let $\epsilon, m > 0$ and let $\tilde{\Theta} = \Theta + \epsilon I$. Let $X_1, \dots, X_m \sim N(0, \tilde{\Theta}^{-1})$. If $r := \dim\ker\Theta > 2m$, then with probability at least $1 - \exp(-\Omega(m))$ it holds that for all $a \in \mathbb{R}^m$,*

$$\left\|\operatorname{Proj}_{\operatorname{span}\Theta} X^T a\right\|_2 \leq C\sqrt{\frac{n\epsilon}{\lambda}}\left\|X^T a\right\|_2$$

where C is an absolute constant, and where $X : m \times n$ is the matrix with rows $X_1, \dots, X_m$.

Proof. The proof of this result is essentially contained in [32], and we repeat it to make this paper self-contained. The statement of the lemma is basis-independent (e.g. does not depend on sparsity of Θ or a), so we can assume without loss of generality that Θ is diagonal. Then $\tilde{\Theta}^{-1}$ is diagonal, and we can choose a basis ordering such that the first $r = \dim\ker\Theta$ diagonal entries are each ϵ^{-1}.

Next, note that $(X^T)_{[r]}$ is a $r \times m$ matrix with i.i.d. $N(0, \epsilon^{-1})$ entries, so $\sigma_{\min}((X^T)_{[r]}) \geq c\epsilon^{-1/2}\sqrt{r}$ with probability at least $1 - \exp(-\Omega(m))$, for some constant $c > 0$ (by Theorem B.1). On the other hand, since the entries of $\tilde{\Theta}^{-1}_{[r]^c,[r]^c}$ are bounded by $1/\lambda$, we also have $\sigma_{\max}((X^T)_{[r]^c}) \le C\sqrt{n/\lambda}$ with probability at least $1 - \exp(-\Omega(n))$, for some constant C. This means that for any $u \in \mathbb{R}^m$,

$$\left\|(X^T u)_{[r]^c}\right\|_2 \le C\sqrt{\frac{n}{\lambda}}\left\|u\right\|_2 \le \frac{C\sqrt{n\epsilon}}{c\sqrt{\lambda r}}\left\|(X^T u)_{[r]}\right\|_2 \le \frac{C}{c}\sqrt{\frac{n\epsilon}{\lambda}}\left\|X^T u\right\|_2.$$

But $(X^T u)_{[r]^c}$ is precisely the projection of $X^T u$ onto $\operatorname{span}\Theta$. So this proves the lemma. □

The next lemma establishes that if the top eigenspace of the covariance $\tilde{\Sigma} = \left(\tilde{\Theta}\right)^{-1}$ has a dimension significantly larger than the number of samples, then the row span of the design matrix X is unlikely to align with any particular direction v (i.e. the projection of v onto the null space has large norm). Informally, this is because the worst-case v to consider would be a vector in the top eigenspace (since a lot of the energy of the samples is in this space), and a random lower-dimensional subspace of the top eigenspace (corresponding to the samples) is not likely to contain any particular direction.

Lemma E.5 (Lemma 7.4 in [32]). *Let $\Theta \in \mathbb{R}^{n\times n}$ be a PSD matrix with minimum nonzero eigenvalue λ. Let $\epsilon, m > 0$ and let $\tilde{\Theta} = \Theta + \epsilon I$. Let $X_1, \dots, X_m \sim N(0, \tilde{\Theta}^{-1})$. If $\epsilon \le c\lambda/n$ for a sufficiently small absolute constant $c > 0$, and $r := \dim\ker\Theta > 2m$, then for any fixed $v \in \mathbb{R}^n$, we have*

$$\Pr_{X_1,\dots,X_m}[v^T(I-P)v \ge (v^T v)/8] \ge 1 - \frac{4m}{3r} - \exp(-\Omega(m)),$$

where $P = X^T(XX^T)^{-1}X$ is the projection map onto $\operatorname{span}\{X_1,\dots,X_m\}$, and where $X : m \times n$ is the matrix with rows $X_1, \dots, X_m$. As an equivalent statement, it holds with probability at least $1 - (4m)/(3r) - \exp(-\Omega(m))$ that

$$\inf_{a\in\mathbb{R}^m}\left\|v - X^T a\right\|_2 \ge \frac{1}{2\sqrt{2}}\left\|v\right\|_2.$$

Proof. The proof of this result is essentially contained in [32], and we repeat it to make this paper self-contained.

The statement of the lemma is basis-independent (e.g. does not depend on sparsity of Θ or v), so we can assume without loss of generality that Θ is diagonal. Then $\tilde{\Theta}^{-1}$ is diagonal, and we can choose a basis ordering such that the first $r = \dim\ker\Theta$ diagonal entries are each ϵ^{-1}. Let $w = v_{[r]}$ be the first r coordinates of v. For $i \in [m]$ let $Y_i = (X_i)_{[r]}$ be the first r coordinates of X_i. Then $Y_1, \dots, Y_m$ are i.i.d. $N(0, \epsilon^{-1}I_r)$. So if $P_Y = Y^T(YY^T)^{-1}Y$, then P_Y is projection onto an isotropically random dimension-m subspace of $\mathbb{R}^r$. Hence,

$$\mathbb{E}\left\|P_Y w\right\|_2^2 = \frac{m}{r}\left\|w\right\|_2^2.$$

With probability at least $1 - 4m/(3r)$ we have $\left\|P_Y w\right\|_2^2 \le \frac{3}{4}\left\|w\right\|_2^2$. So $\left\|w - P_Y w\right\|_2^2 \ge \left\|w\right\|_2^2/4$. Now $\left\|w - P_Y w\right\|_2$ is the distance from w to the subspace $\operatorname{span}\{Y_1,\dots,Y_m\}$. For any vector in $\operatorname{span}\{X_1,\dots,X_m\}$, its first r coordinates lie in $\operatorname{span}\{Y_1,\dots,Y_m\}$, so the distance to v must be at least $\left\|w - P_Y w\right\|_2$. Thus,

$$\left\|v - Pv\right\|_2^2 \ge \left\|w - P_Y w\right\|_2^2 \ge \frac{1}{4}\left\|w\right\|_2^2. \tag{6}$$

Next, note that $(X^T)_{[r]}$ is a $r\times m$ matrix with i.i.d. $N(0,\epsilon^{-1})$ entries, so $\sigma_{\min}((X^T)_{[r]}) \ge c\epsilon^{-1/2}\sqrt{r}$ with probability at least $1 - \exp(-\Omega(m))$, for some constant $c > 0$ (by Theorem B.1). On the other hand, since the entries of $\tilde{\Theta}^{-1}_{[r]^c,[r]^c}$ are bounded by $1/\lambda$, we also have $\sigma_{\max}((X^T)_{[r]^c}) \le C\sqrt{n/\lambda}$ with probability at least $1 - \exp(-\Omega(n))$, for some constant C. This means that for any $u \in \mathbb{R}^m$,

$$\left\|(X^T u)_{[r]^c}\right\|_2 \le C\sqrt{\frac{n}{\lambda}}\left\|u\right\|_2 \le \frac{C\sqrt{n\epsilon}}{c\sqrt{\lambda r}}\left\|(X^T u)_{[r]}\right\|_2.$$

By assumption, $\epsilon \le (c/4C)^2\lambda r/n$, so that $\left\|(X^T u)_{[r]^c}\right\|_2 \le \left\|(X^T u)_{[r]}\right\|_2/4$. Now Pv lies in the span of $X_1, \dots, X_m$, so there is some $u \in \mathbb{R}^m$ with $Pv = X^T u$. This means that

$$\left\|(Pv)_{[r]^c}\right\|_2 \le \frac{1}{4}\left\|Pv\right\|_2 \le \frac{1}{4}\left\|v\right\|_2.$$

So

$$\|v - Pv\|_2 \geq \left\|(v - Pv)_{[r]^c}\right\|_2 \geq \left\|v_{[r]^c}\right\|_2 - \frac{1}{4}\|v\|_2 .$$

Together with Equation 6, which states that $\|v - Pv\|_2 \geq \frac{1}{2}\left\|v_{[r]}\right\|_2$, we get that $\|v - Pv\|_2^2 \geq \frac{1}{8}\|v\|_2^2$. □

Finally, we extend the previous lemma to show that projection of a fixed direction onto the null space of the covariates has lower bounded norm even when restricted to a set of coordinates P with sparse complement. We need an extra condition, that $\ker(\Theta)$ is quantitatively dense (Definition 6), meaning that there is no approximate solution u to the equation $u_P = 0$ near the top eigenspace of $\tilde{\Sigma} = \tilde{\Theta}^{-1}$. Then we show that under the assumptions of the previous two lemmas, for any particular vector u_P supported in P, the equation $u_P = (X^T a)_P$ is unlikely to have an approximate solutions a. The proof is by contradiction: if the conclusion is false, then there is a positive probability that if we sample two independent design $X, \bar{X}$ that: (1) there are approximate solutions $a, \bar{a}$ to $u_P = (X^T a)_P$ and $u_P = ((\bar{X})^T \bar{a})_P$, (2) by Lemma E.4 both approximate solutions are near the top eigenspace, and (3) by Lemma E.5 these two solutions are not well-aligned; combining (1-3) gives that $(X^T a - (\bar{X})^T \bar{a})_P$ is an approximate solution to $(u_P) = 0$ near the top eigenspace, which is impossible.

Lemma E.6. *There are absolute constants $c, C > 0$ such that the following holds. Let $\Theta \in \mathbb{R}^{n\times n}$ be a PSD matrix with minimum nonzero eigenvalue λ. Let $\epsilon > 0$ and $\delta, \eta, \tau > 0$ and let $\tilde{\Theta} = \Theta + \epsilon I$. Let $m \geq C \log n$. Suppose that $\epsilon \leq c\delta^2\lambda/n$, and $r := \dim\ker\Theta > 2m$, and $\ker(\Theta)$ is (δ, η, τ)-quantitatively dense (Definition 6). Fix $u \in \mathbb{R}^n$ and $P \subseteq [n]$ with $|P| \geq n - \tau$. Then with probability at least $p = 1 - (4m)/(3r) - \exp(-\Omega(m))$ over independent $X_1, \dots, X_m \sim N(0, \tilde{\Theta}^{-1})$, it holds that*

$$\inf_{a\in\mathbb{R}^m} \left\|u_P - (X^T a)_P\right\|_2 \geq \frac{\eta}{32}\|u_P\|_2 .$$

Proof. Suppose that the claim is false. Then using that $r = \dim\ker\Theta \leq n$ we have

$$\Pr_X\left[\exists a \in \mathbb{R}^m : \left\|(u - X^T a)_P\right\|_2 < (\eta/32)\|u_P\|_2\right] > \frac{4m}{3r} \geq \frac{1}{n}.$$

By Lemma E.4, we also have that

$$\Pr_X\left[\forall a \in \mathbb{R}^m : \left\|\mathrm{Proj}_{\mathrm{span}\,\Theta} X^T a\right\|_2 \leq C\sqrt{\frac{n\epsilon}{\lambda}}\left\|X^T a\right\|_2\right] \geq 1 - \exp(-\Omega(m)).$$

Since $m \geq C\log n$, for a sufficiently large constant C the latter probability must exceed $1 - 1/n$. So with positive probability, these two events occur simultaneously. Hence, there exists some deterministic $\bar{X} \in \mathbb{R}^{m\times n}$ such that both events occur. Let $\bar{a}$ be the witness of the first event for $\bar{X}$, and define $v = \bar{X}^T\bar{a}$. Then the following two equations hold:

$$\|(u - v)_P\|_2 < \frac{\eta}{32}\|u_P\|_2 \tag{7}$$

$$\mathrm{dist}(v, \ker\Theta) \leq C\sqrt{\frac{n\epsilon}{\lambda}}\|v\|_2 \tag{8}$$

Now suppose that the claim of Lemma E.4 holds (for the original samples $X_1, \dots, X_m$). Also suppose that the claim of Lemma E.5 holds for vector v (and samples $X_1, \dots, X_m$), i.e. $v^T(I - P)v \geq (v^T v)/8$: or equivalently, the following inequality holds for arbitrary vectors $a \in \mathbb{R}^m$:

$$\left\|v - X^T a\right\|_2 \geq \frac{1}{2\sqrt{2}}\|v\|_2 . \tag{9}$$

By Lemmas E.4 and E.5, we may assume that both claims hold with probability $p := 1 - (4m)/(3r) - \exp(-\Omega(m))$. Fix $a \in \mathbb{R}^m$. Then

$$\begin{aligned}
\mathrm{dist}(X^T a, \ker\Theta) &\leq C\sqrt{\frac{n\epsilon}{\lambda}}\left\|X^T a\right\|_2 && \text{(by claim of Lemma E.4)}\\
&\leq C\sqrt{\frac{n\epsilon}{\lambda}}(\left\|v - X^T a\right\|_2 + \|v\|_2)\\
&\leq 4C\sqrt{\frac{n\epsilon}{\lambda}}\left\|v - X^T a\right\|_2 && \text{(by Equation 9)}
\end{aligned}$$

By Equation 8 and Equation 9,

$$\operatorname{dist}(v, \ker \Theta) \leq 3C\sqrt{\frac{n\epsilon}{\lambda}} \left\|v - X^T a\right\|_2 .$$

Thus, by the triangle inequality,

$$\operatorname{dist}(v - X^T a, \ker \Theta) \leq 7C\sqrt{\frac{n\epsilon}{\lambda}} \left\|v - X^T a\right\|_2 .$$

By assumption, $7C\sqrt{(n\epsilon)/\lambda} \leq \delta$. So because $\ker(\Theta)$ is (δ, η, τ)-quantitatively dense and $|P| \geq n - \tau$, and by Equation 9,

$$\left\|(v - X^T a)_P\right\|_2 \geq \eta \left\|v - X^T a\right\|_2 \geq \frac{\eta}{2\sqrt{2}} \left\|v\right\|_2 .$$

Finally, we convert this into a bound on $(u - X^T a)_P$, repeatedly using the fact that vectors u and v are close on P (Equation 7):

$$\begin{aligned}
\left\|(u - X^T a)_P\right\|_2 &\geq \left\|(v - X^T a)_P\right\|_2 - \left\|(u - v)_P\right\|_2 \\
&\geq \frac{\eta}{2\sqrt{2}} \left\|v\right\|_2 - \frac{\eta}{32} \left\|u_P\right\|_2 \\
&\geq \frac{\eta}{2\sqrt{2}} \left\|v_P\right\|_2 - \frac{\eta}{32} \left\|u_P\right\|_2 \\
&\geq \frac{\eta}{2\sqrt{2}} \left\|u_P\right\|_2 - \frac{\eta}{2\sqrt{2}} \left\|(u - v)_P\right\|_2 - \frac{\eta}{32} \left\|u_P\right\|_2 \\
&\geq \frac{\eta}{8} \left\|u_P\right\|_2
\end{aligned}$$

where the second and last inequalities apply Equation 7. We showed this inequality holds for all $a \in \mathbb{R}^m$ with probability at least p, which is the desired conclusion of the Lemma. This contradicts the initial assumption that the conclusion is false, proving the conclusion unconditionally. □

E.3 Failure of rectangular preconditioners

Lemma E.7. *Let $r := \dim \ker \Theta$. Then for any $S \in \mathbb{R}^{n \times s}$,*

$$\Pr_{i \in [n]} \left[\left\|S^T e_i\right\|_1 \geq \frac{1}{\sqrt{\beta \epsilon n}} \right] > \frac{r - 2m}{n}$$

where $\beta = \beta_{\tilde{\Theta}^{-1}, S, k, m}$.

Proof. Suppose not. Then $\left\|S^T e_i\right\| < 1/\sqrt{\beta \epsilon n}$ for at least $n + 2m - r$ choices of $i \in [n]$. Pick some $\beta' > \beta$ which is sufficiently close to β that $I = \{i \in [n] : \left\|S^T e_i\right\| < 1/\sqrt{\beta' \epsilon n}\}$ also satisfies $|I| \geq n + 2m - r$. Let $V = \operatorname{span}\{e_i : i \in I\}$. Then $\dim V = |I| \geq n + 2m - r$. Define $W = V \cap \ker \Theta$. Then $\dim W \geq \dim V - (n - r) \geq 2m$. Moreover, for any $w \in W$, we have

$$\begin{aligned}
w^T \tilde{\Theta}^{-1} w &\geq \epsilon^{-1} \left\|\operatorname{Proj}_{\ker \Theta} w\right\|_2^2 \\
&= \epsilon^{-1} \left\|w\right\|_2^2
\end{aligned}$$

whereas

$$\left\|S^T w\right\|_1 \leq \sum_{i \in I} |w_i| \cdot \left\|S^T e_i\right\|_1 \leq \frac{1}{\sqrt{\beta' \epsilon n}} \left\|w\right\|_1 \leq \frac{\left\|w\right\|_2}{\sqrt{\beta' \epsilon}} .$$

As a consequence, $w^T \tilde{\Theta}^{-1} w \geq \beta' \left\|S^T w\right\|_1^2$. Therefore $W \subseteq W_{\tilde{\Theta}^{-1}, S, \beta'}$, contradicting the definition of $\beta_{\tilde{\Theta}^{-1}, S, k, m}$. □

Lemma E.8. *There are constants $c, c_m > 0$ so that the following holds. Let $M \in \mathbb{R}^{n - r \times n}$ and $\epsilon > 0$. Define $\Theta = M^T M$ and $\tilde{\Theta} = \Theta + \epsilon I$. Let λ be the smallest nonzero eigenvalue of Θ. Let $k, m, s, \alpha, \tau, \eta, b, b' > 0$ and suppose that M satisfies (b, b', η, τ)-erasure-robustness, and $\ker(M)$ is $(1/(12\sqrt{n}), 1/2\sqrt{n}, b')$-quantitatively-dense. Suppose $r > 2m$, $m \geq c_m \log n$, $k > (n/\tau) \log(sn)$*

and $\epsilon < c\lambda/n^2$. Let $\mathcal{D}_k$ be the distribution of k-sparse signals with uniformly random support in $\mathbb{R}^n$, and Gaussian entries on the support. For any preconditioner $S \in \mathbb{R}^{n\times s}$ satisfying

$$\Pr_{i\in[n-r]}\left[M_i^T\tilde{\Theta}^{-1}M_i < \alpha\left\|S^TM_i\right\|_1^2\right] \le \frac{b}{n-r} \tag{10}$$

and

$$128n^{7/2}k\eta\left\|M\right\|_F\sqrt{\frac{\epsilon\beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}}{\alpha\lambda}} \le 1, \tag{11}$$

we have the following: with probability at least

$$1 - 2/n - kb'/n - \exp(-k(r-2m)/n)$$

over true signals drawn from $\mathcal{D}_k$, it holds that S-preconditioned Lasso fails with probability at least $1-(4m)/(3r)-\exp(-\Omega(m))$ over independent samples $X_1,\dots,X_m \sim N(0,\tilde{\Theta}^{-1})$.

Proof. Let $S : n \times s$ be an arbitrary preconditioning matrix satisfying (10) and (11). By (10), we can apply Lemma D.2 to define sets C, D as functions of S, with the following properties:

- For every $i \in D$, the column $v = (S^T)_i \in \mathbb{R}^n$ satisfies $\left\|v\right\|_0 \ge \tau$,
- The submatrix $S_{C^cD^c}$ satisfies
$$\sum_{j\in D^c}\left\|S_{C^cj}\right\|_2 \le \frac{n^{3/2}\eta\left\|M\right\|_F}{\sqrt{\lambda\alpha}}$$
- $|C| \le b'$.

Draw $w^* \sim \mathcal{D}_k$. Let $K = \operatorname{supp}(w^*)$ and let $U = \operatorname{supp}(S^Tw^*)$. Recall that for every $i \in D$, the ith row of S^T (the vector $(S^T)_i \in \mathbb{R}^n$) has at least τ nonzero entries. Therefore, for a particular choice of $i \in D$, the probability that $(S^Tw^*)_i = 0$ is at most the probability that a uniformly random set of k elements from $[n]$ misses all τ elements of the support of $(S^T)_i$, which is at most

$$(1-\tau/n)(1-\tau/(n-1))\cdots(1-\tau/(n-k)) \le e^{-k\tau/n} \le 1/(sn).$$

Hence by the union bound over all $i \in D$, recalling that $D \subset [s]$, we have that $D \subseteq U$ with probability at least $1-1/n$.

Similarly, because $|C| \le b'$, with probability at least $1-kb'/n$ it holds that $K \subseteq C^c$. By Lemma E.7, with probability at least $1-(1-(r-2m)/n)^k \ge 1-e^{-k(r-2m)/n}$ it holds that $\left\|S_{j^*}\right\|_1 \ge 1/\sqrt{\beta\epsilon n}$ for some $j^* \in K$. Now conditioned on K, observe that $(w^*)_K$ has independent $N(0,1)$ entries. So by Lemma E.1,

$$\Pr[\left\|S^Tw^*\right\|_1 \ge \left\|S_{j^*}\right\|_1/n] \ge 1-1/n.$$

Thus, it follows that $\left\|S^Tw^*\right\|_1 \ge 1/(n\sqrt{\beta\epsilon n})$ with probability at least $1-1/n-\exp(-k(r-2m)/n)$ over w^*. Moreover $\left\|w^*\right\|_2 \le 2\sqrt{k}$ with probability at least $1-\exp(-\Omega(k))$. Assume for the rest of the proof that all of the above events (on w^*) occur: $D \subseteq U$, $K \subseteq C^c$, $\left\|S^Tw^*\right\|_1 \ge 1/(n\sqrt{\beta\epsilon n})$, and $\left\|w^*\right\|_2 \le 2\sqrt{k}$. We have shown that these together occur with probability at least $1-2/n-kb'/n-\exp(-k(r-2m)/n)$, and in the rest of the proof we show that under these events, S-preconditioned Lasso fails with probability at least $1-(4m)/(3r)-\exp(-\Omega(m))$ over samples $X_1,\dots,X_m$.

Let $z = \operatorname{sign}(S^Tw^*)$. Since $\operatorname{supp}(w^*) = K \subseteq C^c$, we have by Cauchy-Schwarz that

$$\left\|(Sz)_{C^c}\right\|_2 \ge \frac{\langle w^*, Sz\rangle}{\left\|w^*\right\|_2} = \frac{\langle S^Tw^*, z\rangle}{\left\|w^*\right\|_2} = \frac{\left\|S^Tw^*\right\|_1}{\left\|w^*\right\|_2} \ge \frac{1}{2n\sqrt{\beta\epsilon nk}} \tag{12}$$

where the last inequality uses the above bounds on $\left\|S^Tw^*\right\|_1$ and $\left\|w^*\right\|_2$. Define

$$d = \operatorname*{argmin}_{x\in\ker X_{[m],C^c}}\left\|x-(Sz)_{C^c}\right\|_2,$$

implicitly zero-extending d from C^c to $[n]$. Then $Xd = X_{[m],C^c} d_{C^c} = 0$ by construction, and moreover

$$\begin{aligned}\left\|(S^T d)_{U^c}\right\|_1 &\leq \left\|(S^T d)_{D^c}\right\|_1 && (D \subseteq U)\\ &= \sum_{j \in D^c} |\langle (S^T)_{jC^c}, d_{C^c}\rangle| && (\mathrm{supp}(d) \subseteq C^c)\\ &\leq \|d\|_2 \sum_{j \in D^c} \|S_{C^c j}\|_2 && \text{(Cauchy-Schwarz)}\\ &\leq \|d\|_2 \cdot \frac{n^{3/2}\eta \|M\|_F}{\sqrt{\lambda\alpha}}.\end{aligned}$$

On the other hand,

$$\|d\|_2^2 = \langle S^T d, z\rangle = \left\|\mathrm{Proj}_{\ker X_{[m],C^c}} (Sz)_{C^c}\right\|_2^2.$$

We now apply Lemma E.6 to vector Sz and set C^c, using that $\ker(M)$ is $(1/(12\sqrt{n}), 1/(2\sqrt{n}), b')$-quantitatively dense in conjunction with the bounds $\epsilon \leq c\lambda/n^2$, $m \geq c_m \log n$, $r > 2m$, and $|C| \leq b'$. By this lemma and by (12), with probability at least $1 - (4m)/(3r) - \exp(-\Omega(m))$, we can lower bound the norm of the projection to get

$$\|d\|_2 \geq \frac{1}{64\sqrt{n}} \|(Sz)_{C^c}\|_2 \geq \frac{1}{128n^2\sqrt{\beta\epsilon k}}.$$

As a result, so long as $n^{3/2}\eta \|M\|_F / \sqrt{\lambda\alpha} < 1/(128n^2\sqrt{\beta\epsilon k})$, which holds by (11), we have that $\left\|(S^T d)_{U^c}\right\|_1 < \langle S^T d, z\rangle$ so the preconditioned Lasso fails on w^*. □

Theorem E.9. *There are constants $c, c_m > 0$ so that the following holds. Let $M \in \mathbb{R}^{n-r\times n}$, and define $\Theta = M^T M$. Let $k, k', m, \tau, b, b', \eta, t > 0$. Let λ be the smallest non-zero eigenvalue of Θ. Suppose that M satisfies (b, b', η, τ)-erasure-robustness, and that $\ker(M)$ is $(1/(12\sqrt{n}), 1/(2\sqrt{n}), b')$-quantitatively-dense. Let $\epsilon > 0$ and define $\tilde{\Theta} = \Theta + \epsilon I$. Suppose that $544n^{15/2}k\eta \|M\|_F \sqrt{\epsilon/\lambda} \leq 1$, that $r > 2m$, that $m \geq c_m \log n$, and that $\epsilon < c\lambda/n^2$. And suppose that the rows of M are k'-sparse. Then there is a distribution $\mathcal{D}$ over $\max(k't, k)$-sparse signals such that for any preconditioner $S \in \mathbb{R}^{n\times s}$ with $(n/\tau)\log(sn) < k$, with probability at least*

$$\frac{1}{2}\min\left(1 - e^{-bt/(n-r)} - \frac{1}{n}, 1 - \frac{2}{n} - \frac{kb'}{n} - e^{-k(r-2m)/n}\right)$$

over $w^* \sim \mathcal{D}$, the S-preconditioned Lasso fails to recover w^* with probability at least $1 - (4m)/(3r) - \exp(-\Omega(m))$ over the samples $X_1, \ldots, X_m$.

Proof. Let $\mathcal{D}_M$ be the signal distribution where we draw independent and uniformly random indices $R_1, \ldots, R_t \in [n-r]$ as well as independent $Z_1, \ldots, Z_t \sim \mathrm{Unif}([-1, 1])$, and set the signal to be

$$w^* = \sum_{i=1}^{t} \frac{Z_i M_{R_i}}{\sqrt{M_{R_i}^T \tilde{\Theta}^{-1} M_{R_i}}}.$$

We define $\mathcal{D}$ to be the mixture which with probability $1/2$ draws a signal from $\mathcal{D}_M$, and with probability $1/2$ draws from the distribution $\mathcal{D}_k$ defined in Lemma E.8 (note that in the former case, the signal is $k't$-sparse, and in the latter case, it is k-sparse).

Pick any $S \in \mathbb{R}^{n\times s}$ with $(n/\tau)\log(sn) < k$. Define $\alpha = \beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}/(18n^4)$. We distinguish two cases.

Case I: incompatible preconditioner. On the one hand, suppose that

$$\Pr_{i\in[n-r]}\left[M_i \tilde{\Theta}^{-1} M_i < \alpha \left\|S^T M_i\right\|_1^2\right] \geq \frac{b}{n-r}.$$

Then with probability at least

$$\frac{1}{2}\left(1 - (1 - b/(n-r))^t\right) \geq \frac{1}{2}\left(1 - e^{-tb/(n-r)}\right)$$

over the row indices $R_1, \dots, R_t$, there is some R_i with

$$\left\|S^T M_{R_i}\right\|_1^2 > \frac{1}{\alpha} M_{R_i}^T \tilde{\Theta}^{-1} M_{R_i}.$$

In this event, we have

$$\max_{i \in [t]} \frac{\left\|S^T M_{R_i}\right\|_1}{\sqrt{M_{R_i}^T \tilde{\Theta}^{-1} M_{R_i}}} > \frac{1}{\sqrt{\alpha}},$$

so by Lemma E.1, it holds with probability at least $1 - 1/n$ over $Z_1, \dots, Z_t$ that $\left\|S^T w^*\right\|_1 \geq \frac{1}{n\sqrt{\alpha}}$. But by the triangle inequality, we have $\sqrt{(w^*)^T \tilde{\Theta}^{-1} w^*} \leq t \leq n$. Thus,

$$\Pr_{w^* \sim \mathcal{D}} \left[\left\|S^T w^*\right\|_1^2 > \frac{1}{n^4 \alpha}(w^*)^T \tilde{\Theta}^{-1} w^*\right] \geq \frac{1}{2}\left(1 - e^{-tb/(n-r)}\right)\left(1 - \frac{1}{n}\right).$$

Moreover, for such w^*, by choice of α and by Theorem D.1, the S-preconditioned Lasso recovers w^* with probability at most $\exp(-\Omega(m))$, from m independent samples $X_1, \dots, X_m \sim N(0, \tilde{\Theta}^{-1})$.

Case II: compatible preconditioner. On the other hand, suppose that

$$\Pr_{i \in [n-r]} \left[M_i \tilde{\Theta}^{-1} M_i < \alpha \left\|S^T M_i\right\|_1^2\right] \leq \frac{b}{n-r}.$$

This is precisely the condition (10) in Lemma E.8. Moreover, (11) is satisfied because we defined α so that $\beta^{(1)}_{\tilde{\Theta}^{-1},S,k,m}/\alpha = 18n^4$, and because of the bound assumed in the theorem statement. Thus, we can invoke Lemma E.8, which states that for $w^* \sim \mathcal{D}_k$, with probability at least $1 - 2/n - kb'/n - \exp(-k(r - 2m)/n)$ over the signal, the S-preconditioned Lasso fails with the desired probability. Thus, the same holds for $w^* \sim \mathcal{D}$ up to a factor of 2. □

Finally, we instantiate the above theorem with the matrix constructed in Theorem C.15.

Theorem E.10. *Let $n \in \mathbb{N}$ be sufficiently large. There is a matrix $\Sigma \in \mathbb{R}^{n \times n}$ with condition number $\operatorname{poly}(n)$ such that the following holds. Let $k \in \mathbb{N}$ with $k \geq \log^8 n$. There is a distribution $\mathcal{D}$ over $k \log^4 n$-sparse signals such that for any positive integer $m \leq n/5$ and any preconditioner $S \in \mathbb{R}^{n \times s}$ with $s < \exp(ck/\log^6 n - \log n)$, with probability at least $\frac{1}{2} - O(1/\log n)$ over $w^* \sim \mathcal{D}$, the S-preconditioned Lasso recovers w^* with probability at most $8m/3n + \exp(-\Omega(m))$ from m independent samples $X_1, \dots, X_m \sim N(0, \Sigma)$ with noiseless responses $Y_i = \langle X_i, w^* \rangle$.*

Proof. Let Θ be the matrix guaranteed by Theorem C.15. We check the conditions of Theorem E.9. Clearly, $r := \dim\ker(\Theta) \geq n/2$. By claim (5), we have $\lambda = \Omega(n^{-5/2})$. By claims (3) and (2), we can take $b = n/(k \log n)$, $b' = 2n/(k \log n)$, $\eta = n^6 \log^8 n$, and $\tau = \Omega(n/\log^6 n)$. By claim (4), we have $\|M\|_F \leq O(\sqrt{n \log n})$. Thus, we can take $\epsilon = \Omega(n^{-22})$. By claim (1), the rows of M are k'-sparse with $k' = O(\log^2 n)$. Let $t = k \log^2 n$.

So applying Theorem E.9, we get that there is a distribution $\mathcal{D}$ over $k \log^4 n$-sparse signals such that for any preconditioner $S \in \mathbb{R}^{n \times s}$ with $s < \exp(ck/\log^6 n - \log n)$, with probability at least

$$\frac{1}{2} \min\left(1 - \frac{2}{n}, 1 - \frac{2}{n} - \frac{2}{\log n} - e^{-k(n/2 - 2m)/n}\right)$$

over $w^* \sim \mathcal{D}$, the S-preconditioned Lasso recovers w^* with probability at most $8m/(3n) + \exp(-\Omega(m))$ from m samples. So long as $m \leq n/5$, both terms in the minimum are $1 - O(1/\log n)$, yielding the claimed result. □