Restricted Eigenvalue from Stable Rank with Applications to Sparse Linear Regression

Shiva Prasad Kasiviswanathan[*] and **Mark Rudelson**[†]

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

High-dimensional settings, where the data dimension (d) far exceeds the number of observations (n), are common in many statistical and machine learning applications. Methods based on ℓ_1-relaxation, such as Lasso, are very popular for sparse recovery in these settings. Restricted Eigenvalue (RE) condition is among the weakest, and hence the most general, condition in literature imposed on the Gram matrix that guarantees nice statistical properties for the Lasso estimator. It is hence natural to ask: what families of matrices satisfy the RE condition? Following a line of work in this area (Raskutti et al., 2010; Rudelson and Zhou, 2013; Sivakumar et al., 2015; Oliveira, 2016; Lecué and Mendelson, 2017), we construct a new broad ensemble of dependent random design matrices that have an explicit RE bound. Our construction starts with a fixed (deterministic) matrix $X \in \mathbb{R}^{n\times d}$ satisfying a simple stable rank condition, and we show that a matrix drawn from the distribution $X\Phi^\top\Phi$, where $\Phi \in \mathbb{R}^{m\times d}$ is a subgaussian random matrix, with high probability, satisfies the RE condition. This construction allows incorporating a fixed matrix that has an easily *verifiable* condition into the design process, and allows for generation of *compressed* design matrices that have a lower storage requirement than a standard design matrix. We give two applications of this construction to sparse linear regression problems, including one to a compressed sparse regression setting where the regression algorithm only has access to a compressed representation of a fixed design matrix X.

1. Introduction

A high dimensional setting, where the number of features (d) is much larger than the number of observations (n) appears commonly in statistics and signal processing, for example, in regression, covariance selection on Gaussian graphical models, signal reconstruction, and sparse approximation. Consider a simple setting where we try to recover $\theta^\star$, given $(M, \mathbf{y})$, satisfying the following linear model:

$$\mathbf{y} = M\theta^\star + \mathbf{w}. \tag{1}$$

Here $\mathbf{y} \in \mathbb{R}^n$ is the vector of noisy observations, $M \in \mathbb{R}^{n\times d}$ is the design matrix, and $\mathbf{w} \in \mathbb{R}^n$ is an unknown noise vector. In the setting of $d \gg n$, the model is *unidentifiable* and it is not meaningful to estimate $\theta^\star \in \mathbb{R}^d$. However, many machine learning and statistical applications, exhibit special structure that can lead to an identifiable model. In particular, in many settings, the vector $\theta^\star$ is sparse. Given such a problem, the most direct approach would be to seek an exact sparse minimizer of the least-squares cost, $\|\mathbf{y} - M\theta\|^2$, thereby obtaining an ℓ_0-based estimator. However, since this problem is non-convex, a standard approach is to replace the ℓ_0-constraint with its ℓ_1-norm which is the basis

[*] Amazon AWS AI, Palo Alto, CA, USA. kasivisw@gmail.com.
[†] University of Michigan, Ann Arbor, MI, USA. rudelson@umich.edu. Partially supported by NSF grant, DMS-1464514.

for methods such as Lasso (Tibshirani, 1996) and Dantzig selector (Candes et al., 2007). There is now a well-developed theory of what conditions on the design matrix M are needed for these ℓ_1-based relaxations to succeed. The general idea is that M needs to behave sufficiently nicely in a sense that it satisfies certain *incoherence* conditions. One popular notion of incoherence is *Restricted Isometry Property* (RIP) that states for all k-sparse sets $T \subset \{1, \ldots, d\}$ ($|T| = k$), the matrix M restricted to the columns from T acts as an almost isometry (Candes and Tao, 2005). In the past decade, few variants of the RIP notion for exact and approximate recovery of $\theta^\star$, under the noiseless and noisy setting, have also been proposed (we refer to reader to the books by (Eldar and Kutyniok, 2012; Hastie et al., 2015) for more details).

For the Lasso and Dantzig selector, Bickel et al. (2009) formulated the *restricted eigenvalue* (RE) condition and showed that it is among the weakest,[1] and hence the most general, condition imposed on the Gram matrix that guarantees meaningful recovery. Informally, the RE condition on a matrix M involves lower bounds on $\|M\theta\|$ that hold uniformly over an appropriately defined subset of sparse vectors (see Definition 1 for a formal statement). A natural question is then: for what ensembles of design matrices does the restricted eigenvalue condition hold (say, with high probability)? Standard constructions satisfying the RE condition are based on i.i.d. random matrices, independent draws from a set of uncorrelated basis functions, additive combinatorics, or coding-theoretic techniques (see, e.g., (Mendelson et al., 2008; Adamczak et al., 2011; Rudelson and Vershynin, 2008; Bourgain et al., 2011; Cheraghchi, 2011) and references therein). While these constructions are well-suited for certain compressive sensing tasks, where we have control over the design matrix, it may not be appropriate for statistical inference problems such as sparse linear regression, where the design matrix is not under control of the "experimenter". For example, it is common that the different columns (covariates) of the design matrix are correlated with one other, and in practice ℓ_1-norm methods such as Lasso seem to perform well even in these settings. This has motivated recent work in understanding RE properties for a more realistic class of random design matrices (Raskutti et al., 2010; Rudelson and Zhou, 2013; Sivakumar et al., 2015; Oliveira, 2016; Lecué and Mendelson, 2017). Our paper continues this line of work.

We start with this simple question: can we incorporate a fixed (deterministic) matrix while constructing a family of matrices satisfying the RE condition? In this paper, we answer this question in affirmative by presenting a construction that starts with any deterministic matrix $X \in \mathbb{R}^{n\times d}$, satisfying a very mild easy to check condition, and generates a distribution of matrices centered at X, such that a matrix drawn from this distribution with high probability satisfies the RE condition. More formally, we show that given X, the random matrix $X\Phi^\top\Phi$, where $\Phi \in \mathbb{R}^{m\times d}$ is a i.i.d. subgaussian random matrix, satisfies the RE condition with high probability.[2] All we need is that the *stable rank* of X is not "too small". Stable rank of a matrix X (denoted by $\mathbf{sr}(X)$), defined as the squared ratio of Frobenius and spectral norms of X, is a commonly used robust surrogate to usual matrix rank in linear algebra. We start with an informal statement of our main result which shows, that under some mild conditions on X, with high probability $X\Phi^\top\Phi$ satisfies the restricted eigenvalue property with a parameter value of $\|X\|_F^2/nmk$.

Informal Theorem *[See Theorem 5, Corollary 7] Let X be a fixed $n \times d$ matrix with stable rank greater than m. Let Ψ be an $m \times d$ subgaussian matrix with i.i.d. entries and let $\Phi = \Psi/\sqrt{m}$, then*

1. In particular (Bickel et al., 2009) show that the RE condition is a relaxation of the RIP condition under suitable choices of parameters involved in both of them.
2. We overload notation and use $X\Phi^\top\Phi$ to represent both a random sample and its distribution.

for any k such that $m \gtrsim k^2$, with high probability,

$$\inf_{S\subset[d],|S|=k,\theta\in\mathbb{C}(S)} \frac{\|X\Phi^\top\Phi\theta\|^2}{n\|\theta\|^2} \geq \frac{\|X\|_F^2}{nmk}, \tag{2}$$

where $\mathbb{C}(S)$ is the set of vectors $\theta \in \mathbb{R}^d$ that satisfy the cone constraint, $\mathbb{C}(S) = \{\theta : \|\theta_{S^{\text{co}}}\|_1 \leq 3\|\theta_S\|_1\}$ and $\theta_S, \theta_{S^{\text{co}}}$ represents the subvector of θ confined to coordinates S and $\{1,\ldots,d\} \setminus S$.

The stable rank is independent of the coordinate system, unlike RE which is tied to a concrete coordinate structure. The randomness of Φ is what makes the required condition on X coordinate independent. The above proof is challenging because applying standard concentration tools directly do not give strong enough probability estimates on this quantity for a fixed θ to successfully apply an ε-net argument. To overcome this problem, we develop an orthogonal projection idea that allows us to decouple dependencies and reduce the problem to a state that is amenable to an application of an ε-net argument. Throughout the proof, we rely on the Hanson-Wright inequality and several of its consequences.

Some Key Features of this Construction. We now note some interesting features of the family of random matrices generated by our construction. Firstly, observe that the entries in a matrix $Z = X\Phi^\top\Phi$ are highly correlated with $\mathbb{E}[Z] = X$. Given any matrix X, its stable rank can be computed easily. This is an important advantage while designing RE matrices, as this makes the construction process *verifiable*, i.e., with high probability we can generate a matrix that satisfies an explicit restricted eigenvalue parameter bound. Note that in general, checking whether a matrix satisfies the RE condition is a NP-hard problem (Dobriban and Fan, 2016). To date, the main routes for constructing design matrices with an explicit restricted eigenvalue bound have been via taking i.i.d. random ensembles (under different moment or tail assumptions) or constructions through coding-theoretic techniques (such as expander codes (De Castro, 2014)), both of which generate family of matrices whose assumptions are not always reasonable for machine learning applications. To the best of our knowledge, this is the first construction of a very broad family of (correlated) random matrices that starts with an easy to check condition on the deterministic core. Previous constructions of other such broad family of correlated random designs, such as (Raskutti et al., 2010; Rudelson and Zhou, 2013), require the deterministic matrix to also satisfy some suitable RE condition (more discussion in Section 1.1), thus running into the above mentioned verifiability issues.

An additional salient feature is that the matrix Z can be stored using only $O(m(n+d)) = O(md)$ words of memory as the factorization pair $(X\Phi^\top, \Phi)$. This means that compared to a standard $n \times d$ design matrix which needs $O(nd)$ words of memory (with n generally being much greater than m), the design matrices coming out of this construction have a "compressed" representation. This property is useful when working with large design matrices in presence of memory constraints.

Applications to Sparse Linear Regression. We will give two applications of this result in sparse linear regression. Consider the linear regression model in (1). A popular approach for solving a (traditional) sparse linear regression problem is the Lasso technique of ℓ_1-penalized regression. Lasso minimizes the usual mean squared error loss penalized with (a multiple of) the ℓ_1-norm of θ. The consistency properties of the Lasso estimator under various measurements of performance (such as prediction error, parameter error, support recovery) are now well-understood, see e.g., (Bickel et al., 2009; Wainwright, 2009). We consider the following Lasso problem, defined on the pair $(Z, \mathbf{y})$, where $Z = X\Phi^\top\Phi$.

$$\theta^{\text{comp}} \in \text{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{y} - Z\theta\|^2 + \lambda\|\theta\|_1 = \text{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}(y_i - \langle\Phi\mathbf{x}_i, \Phi\theta\rangle)^2 + \lambda\|\theta\|_1.$$

For brevity, in the following discussion, we make some simplifying assumptions and omit dependence on all but key variables. The ith row in X ($\mathbf{x}_i^\top$) represent the covariates for the ith observation and $\mathbf{y} = (y_1, \ldots, y_n)$. The results stated below all are high probability bounds.

(1) Parameter bound with design matrix Z. Our first application is for the linear model $\mathbf{y} = Z\theta^\star + \mathbf{w}$. In this setting, we have a random design matrix. Here, the RE result on Z leads to a parameter error bound: $\|\theta^{\text{comp}} - \theta^\star\| = O(\sqrt{m}k^{3/2}/\|X\|_F)$, assuming $\mathbf{w}$ is an i.i.d. subgaussian noise vector (see Proposition 12). While this result follows from a simple instantiation of the standard Lasso analysis framework, the result shows that there exists a new broad class of random design matrices for which Lasso succeeds in getting a consistent estimate of $\theta^\star$ (when $\|X\|_F = \omega(\sqrt{m}k^{3/2})$). A similar analysis can also be carried for the Dantzig selector based on the results of (Bickel et al., 2009) (omitted here).

(2) Sparse linear regression with compressed features. Our second application is a variant of sparse linear regression. We start with a linear model $\mathbf{y} = X\theta^\star + \mathbf{w}$, where X is a fixed matrix and $\mathbf{w}$ is an i.i.d. subgaussian noise vector (so in this case, we have fixed design X). However, we assume that the regression algorithm has access to only $(\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$, which is the compressed representation of the original covariate-response pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.[3] Random projections are a class of extremely popular technique for dimensionality reduction (compression), where the original high-dimensional data is projected onto a lower-dimensional subspace using some appropriately chosen random matrix Φ. A motivating scenario for this setting is as follows (see the illustration in Figure 1). Consider a distributed data setting, where n devices each generating its own covariate-response pair is communicating to a central server (cloud). If d is large then communicating $\mathbf{x}_i \in \mathbb{R}^d$ is communication expensive. A natural scheme here is that the server chooses and announces a single random projection matrix Φ, and every input point $\mathbf{x}_i$ can be compressed and sent as $\Phi\mathbf{x}_i$ to the server. Such a scheme can be applied *locally* (i.e., on each $\mathbf{x}_i$ independent of the other), and reduces the overall communication by a factor of d/m.[4] Now the goal of the server is to solve the regression problem for the original linear model ($\mathbf{y} = X\theta^\star + \mathbf{w}$) but from the available compressed representation. Firstly, since the $\mathbf{x}_i$'s are unavailable, it is *a priori* unclear how sparse linear regression performs in this setting. Secondly, just with a stable rank condition on X a parameter error bound on $\theta^\star$, that requires a stronger RE like assumption (Raskutti et al., 2011), is ruled out. In this fixed design setting, we investigate (in-sample) prediction error bounds, and show that θ^{comp} (which can be estimated from the compressed representation) satisfies $\|X\theta^{\text{comp}} - X\theta^\star\|^2/n = O(\|X\|_F^2 k^{3/2}/nm)$ (see Proposition 13). In this case, our use of the RE slightly differs from the standard use of RE in Lasso analysis. We first bound $\|Z\theta^{\text{comp}} - Z\theta^\star\|$ using the Lasso analysis framework, and then use the RE bound on Z to relate that to a bound on $\|X\theta^{\text{comp}} - X\theta^\star\|$.

3. Note that given $\Phi\mathbf{x}_i$ it is not possible to accurately infer $\mathbf{x}_i$ without some strong (sparsity-like) assumptions on $\mathbf{x}_i$. More discussion on this is provided in Section 3.2.

4. We ignore the cost of communicating Φ to devices, which can be achieved using various techniques such as *one-to-all broadcasting*. In a practical implementation, Φ will be generated by a pseudorandom generator initialized by some seed, so by just communicating the seed we can regenerate Φ at each device. Also, with some small degradation in the parameters, the same Φ can be used in a situation where we have to repeatedly solve different sparse linear regression problem instances.

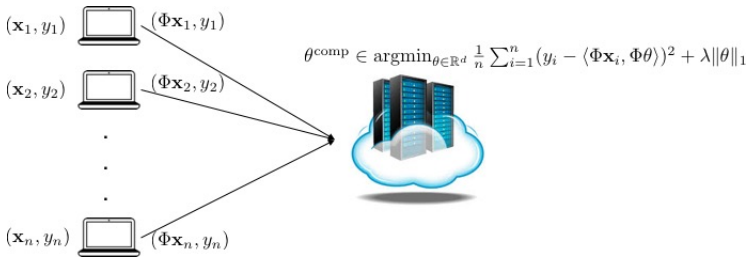

Figure 1: A distributed data setting, where n devices generating $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sending a compressed representation $(\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$ to the cloud server, which then computes the regression parameter.

1.1. Related Work

Restricted Eigenvalue Bound. Matrices that satisfy the restricted isometry (or restricted eigenvalue) property have many interesting applications in high-dimensional statistics and compressed sensing. However, there is no known way to generate them deterministically for a large range of sparsity levels k (some of best constructions here include (Bourgain et al., 2011; Bandeira et al., 2017, 2016)), and additionally as discussed above these constructions lead to family of matrices that are not always reasonable for applications such as linear regression. Interestingly though, it is easy to generate large matrices satisfying the RIP property (and therefore RE) for a wide range of sparsity levels through i.i.d. random design. In statistics and machine learning, one common assumption is that the design matrix is generated randomly by a mechanism which is not under control of the experimenter, and these matrices generally have dependent entries. One may enquire whether such random ensembles will typically satisfy restricted eigenvalue properties. This problem was first addressed for Gaussian ensembles by (Raskutti et al., 2010) and then for subgaussian and bounded-coordinate ensembles by (Rudelson and Zhou, 2013). In particular, these results have established RE bounds for random matrices with subgaussian rows and non-trivial covariance structure, as well as random matrices with independent rows and uniformly bounded entries. Recent papers (Sivakumar et al., 2015; Oliveira, 2016; Lecué and Mendelson, 2017) have developed variants of these bounds under different moment or tail assumptions. The closest relation to our work is the result by (Rudelson and Zhou, 2013), who showed that for a deterministic matrix X satisfying the RE condition, the matrix ΦX satisfies the RE condition too (with a weaker RE parameter), where the rows of Φ are *isotropic* random vectors. Note that, unlike the result in (Rudelson and Zhou, 2013), we have a simple polynomial time checkable stable rank condition on our deterministic matrix X.

Applications to Sparse Linear Regression. Lasso, is the most widely studied scheme for sparse linear regression. There has been a large and rapidly growing body of literature for Lasso and its variants which include theoretical explorations of its behavior and computationally efficient procedures for solving it. We refer the reader to the recent book by (Hastie et al., 2015) for a detailed survey about developments here. For applications of our RE bound to sparse linear regression, we draw on this rich literature studying theoretical properties of Lasso.

Zhou et al. (2009) considered sparse linear regression in a setting where the covariate matrix X is pre-multiplied by a Gaussian random projection matrix to generate a reduced set of new datapoints in d-dimensions. They provide a convergence analysis of the Lasso estimator built from this compressed dataset. This setting is however different from ours, as we consider reducing the dimensionality of

each covariate vector. In a high-dimensional setting, with $d \gg n$, reducing the dimensionality seems intuitively the more desirable way of achieving compression.

A recent area of research is that of distributed sparse linear regression where the dataset is assumed to the distributed across multiple machines. (Lee et al., 2015) showed that if the data is "not too" distributed, and for the random design case, average of individual Lasso estimators properly debiased converges to $\theta^\star$ at almost the same rate as the centralized Lasso estimator. We are not aware of a direct connection between this work and our setting.

There is also a long line of work in using Johnson-Lindenstrauss style transforms with linear regression where the goal is to achieve computational efficiency (see the survey by Woodruff (2014)). Additionally, with random projections and (non-sparse) linear regression, there is a line of work, known as *compressed linear regression* (Maillard and Munos, 2009; Fard et al., 2012; Kabán, 2014), where the goal is to output some $\hat{\vartheta}$ in the projected space (and not in the original $\mathbb{R}^d$ space) that minimizes the expected excess prediction risk. Note that in general, lifting $\hat{\vartheta}$ to the original space is not possible, as $\hat{\vartheta}$ may not lie close to a projection of a sparse vector in $\mathbb{R}^d$. Since our goal is to output θ^{comp} in the original space and our focus is on sparse linear regression, the results from these compressed linear regression papers are also not directly relevant for us.

1.2. Preliminaries

Notation. We denote $[n] = \{1, \ldots, n\}$. For a set $S \subseteq [d]$, S^{co} denotes its complement set. Vectors are in column-wise fashion, denoted by boldface letters. For a vector $\mathbf{v}$, $\mathbf{v}^\top$ denotes its transpose, $\|\mathbf{v}\|_p$ it's ℓ_p-norm, and $\text{supp}(\mathbf{v})$ its support. We use $\mathbf{e}_j \in \mathbb{R}^d$ to denote the standard basis vector with jth entry set to 1. For a matrix M, $\|M\|$ denotes its spectral norm which equals its largest singular value, and $\|M\|_{\text{F}}$ its Frobenius norm. $\mathbb{I}_d$ represents the $d \times d$ identity matrix. For a vector $\mathbf{x}$ and set of indices S, let $\mathbf{x}_S$ be the vector formed by the entries in $\mathbf{x}$ whose indices are in S, and similarly, X_S is the matrix formed by columns of X whose indices are in S. The d-dimensional unit ball in ℓ_p-norm centered at origin is denoted by B_p^d. The Euclidean sphere in $\mathbb{R}^d$ centered at origin is denoted by $\mathbb{S}^{d-1}$. We call a vector $\mathbf{a} \in \mathbb{R}^d$, *$k$-sparse*, if it has at most k non-zero entries. Denote by Σ_k the set of all vectors $\mathbf{a} \in B_2^d$ with support size at most k: $\Sigma_k = \{\mathbf{a} \in B_2^d \,:\, |\text{supp}(\mathbf{a})| \leq k\}$.

Throughout this paper, we assume covariate-response pairs come from some domain $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. In Appendix A, we also review a few additional concepts related to sparse linear regression, ε-nets, and subgaussian random variables.

RE and Sparse Linear Regression. In the following, we consider the linear model: $\mathbf{y} = M\theta^\star + \mathbf{w}$. For a set $S \subset [d]$, let us define a cone set $\mathbb{C}(S)$ as:

$$\mathbb{C}(S) = \{\theta \in \mathbb{R}^d \,:\, \|\theta_{S^{\text{co}}}\|_1 \leq 3\|\theta_S\|_1\}.$$

Restricted eigenvalue is a mild condition on the covariate matrix that is sufficient for estimating $\theta^\star$ in a noisy linear model setup.[5]

Definition 1 (Restricted Eigenvalue (Bickel et al., 2009)) *A matrix $M \in \mathbb{R}^{n \times d}$ satisfies the restricted eigenvalue (RE) condition with parameter ξ if,*

$$\inf_{S \subset [d], |S| = k, \theta \in \mathbb{C}(S)} \frac{\|M\theta\|^2}{n} \geq \xi \|\theta\|^2.$$

5. Given that we observe only a noisy version of the product $M\theta^\star$, it is then difficult to distinguish $\theta^\star$ from other sparse vectors. Thus, it is natural to impose a RE condition if the goal is to produce an estimate $\tilde{\theta}$ such that $\|\theta^\star - \tilde{\theta}\|$ is small.

Restricted eigenvalue is in fact a special case of a general property of loss functions, known as the *restricted strong convexity*, which imposes a type of strong convexity condition for some subset of vectors (Negahban et al., 2012). We now state a well-known result in sparse linear regression that provides a bound on the Lasso error, based on the linear observation model $\mathbf{y} = M\theta^\star + \mathbf{w}$.

Theorem 2 ((Bickel et al., 2009)) *Let $\mathbf{y} = M\theta^\star + \mathbf{w}$ for a noise vector $\mathbf{w} \in \mathbb{R}^n$ and $\theta^\star$ is k-sparse. Let $\lambda \geq 2\|M^\top \mathbf{w}\|_\infty / n$. Suppose M satisfies the restricted eigenvalue condition with parameter $\xi > 0$, then any optimal minimizer, $\tilde{\theta} \in \mathrm{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n}\|\mathbf{y} - M\theta\|^2 + \lambda\|\theta\|_1$, satisfies: $\|\tilde{\theta} - \theta^\star\| \leq 3\sqrt{k}\lambda/\xi$.*

Remark 3 *[A Note on Assumptions] While the above RE condition is common for analyzing the ℓ_2-error of the Lasso estimator,* stronger *conditions are used for achieving the* stronger *guarantee of* consistent support selection *(Wainwright, 2009). These include* mutual incoherence *and* minimum eigenvalue *conditions on M, and* minimum signal value *condition on $\theta^\star$. These conditions are known to be highly restrictive in practice (Tibshirani and Wasserman, 2015) and are not studied in this paper.*

2. Restricted Eigenvalue from Stable Rank

The main result of this paper is to show that the RE condition holds with high probability for systems of random design matrices of a general nature. In particular, we consider design matrices of the form: $Z = X\Phi^\top\Phi$, where X is a fixed matrix and $\Phi \in \mathbb{R}^{m\times d}$ is a subgaussian random matrix. Note that the entries of Z are highly correlated. This construction provides a neat way of incorporating a fixed matrix X into the design of a RE matrix, and also has the advantage that storing Z (as the factorization pair $(X\Phi^\top, \Phi)$) takes only $O(md)$ words of space, compared to storing a standard design matrix which typically requires $O(nd)$ words of space. In the next section, we will discuss few applications of this result to sparse linear regression problems.

We start with the definition of stable rank (denoted by $\mathbf{sr}()$) of a matrix X.

$$\mathbf{sr}(X) = \|X\|_{\mathrm{F}}^2 / \|X\|^2 .$$

Stable rank cannot exceed the usual rank. The stable rank is a more robust notion than the usual rank because it is largely unaffected by tiny singular values. In Appendix D we provide a detailed comparison between these stable rank and restricted eigenvalue. Unsurprisingly, the picture that emerges is that stable rank is in fact a *less restrictive* condition.[6] Throughout this section, $C, C_1, c, c_1, \ldots$ denote positive constants which may depend on the subgaussian norm of the entries of the involved matrices.

We will work with a slightly modified (and a more general) definition of restricted eigenvalue that we state here.

Definition 4 *Let M be an $n \times d$ matrix, and let $k < d,\ \alpha > 0$. Define*

$$\mathrm{RE}(M, k, \alpha) = \inf \frac{\|M\mathbf{z}\|}{\|\mathbf{z}_J\|},$$

where $\mathbf{z}_J$ is the coordinate projection of $\mathbf{z}$ to $\mathbb{R}^J$, and the infimum is taken over all sets $J \subset [d],\ |J| = k$ and all $\mathbf{z} \in \mathbb{R}^m \setminus \{0\}$ satisfying

$$\|\mathbf{z}_{J^{\mathrm{co}}}\|_1 \leq \alpha \|\mathbf{z}_J\|_1 .$$

6. In that a RE bound implies a non-trivial stable rank bound, whereas the other direction does not always hold.

Note that $\alpha = 3$ in Definition 1. Also given $\text{RE}(M, k, \alpha)$, we can get a lower bound on ξ in Definition 1 as $\xi \geq \text{RE}(M, k, 3)^2/k$. Our primary result in this section establishes a lower bound on $\text{RE}(X\Phi^\top\Phi, k, \alpha)$. The proof assumes a stable rank condition on X that we define below. The randomness of Φ makes the required condition on X coordinate independent, unlike the RE condition which is tied to a concrete coordinate structure in $\mathbb{R}^d$.

Theorem 5 *Let $m, n, d \in \mathbb{N},\ m \leq n \leq d$, and let X be a fixed $n \times d$ matrix satisfying*

$$\textbf{Stable Rank Condition}: \quad 2 \leq m \leq \mathbf{sr}(X)/2.$$

Let $\Psi = (\Psi_{ij})$ be an $m \times d$ random matrix with independent entries such that $\mathbb{E}[\Psi_{ij}] = 0$, $\mathbb{E}[\Psi_{ij}^2] = 1$, and $\|\Psi_{ij}\|_{\psi_2}$ is bounded. Let $\Phi = \Psi/\sqrt{m}$. Let $\beta \in (0, 1)$. Then for any $k \in \mathbb{N},\ \alpha > 0$ such that

$$1 \leq \alpha\sqrt{k} \leq \sqrt{\frac{cm}{k \log d + \log(2/\beta)}}$$

the matrix $X\Phi^\top\Phi$ satisfies:

$$\text{RE}(X\Phi^\top\Phi, k, \alpha) \geq \frac{\|X\|_{\text{F}}}{32\sqrt{m}}, \quad \textit{with probability at least } 1 - \beta.$$

Remark 6 *An intuitive explanation why stable rank is the correct notion here is as follows. Firstly, if $\text{rank}(X\Phi^\top\Phi) \leq \text{rank}(X) \leq k$, then RE does not hold for any $X\Phi^\top\Phi$. And it should be the stable rank, because adding an infinitesimally small noise does not change anything. The fact that we have the condition $m \succeq k^2$ and not $m \succeq k$, as this observation would suggest, is due to the model we considered, namely to the multiplication by $\Phi^\top\Phi$.*

Corollary 7 *Let X and Ψ be matrices satisfying the conditions in Theorem 5 with*

$$1 \leq 3\sqrt{k} \leq \sqrt{\frac{cm}{k \log d + \log(2/\beta)}}.$$

Let $\Phi = \Psi/\sqrt{m}$. Then the matrix $X\Phi^\top\Phi$ satisfies:

$$\inf_{S \subset [d], |S| = k, \theta \in \mathbb{C}(S)} \frac{\|X\Phi^\top\Phi\theta\|^2}{n} \geq \frac{\|X\|_{\text{F}}^2 \|\theta\|^2}{1024\, nmk}, \quad \textit{with probability at least } 1 - \beta.$$

We start with a high-level description of the proof idea. In the next section, we describe the key technical ingredients of the proof. The complete proof is presented in Appendix B. Let $\mathbf{e}_j$ denote the standard basis vector with jth entry set to 1.

Idea of the Proof of Theorem 5. We now explain the idea behind the proof of the above theorem. Take any $J \subset [d],\ |J| = k$ and any $\mathbf{y} \in S^{d-1}$ with $\text{supp}(\mathbf{y}) \subseteq J$. We wish to show that with overwhelming probability, any $\mathbf{x} \in \mathbb{R}^d$ with $\text{supp}(\mathbf{x}) \subseteq J^{\text{co}}$ and $\|\mathbf{x}\|_1 \leq \alpha \|\mathbf{y}\|_1 \leq \alpha\sqrt{k}$ satisfies

$$\left\| X\Psi^\top\Psi(\mathbf{y} + \mathbf{x}) \right\| \geq r$$

for some $r > 0$. If the probability estimate is strong enough, we would be able to run an ε-net argument over all such $\mathbf{y}$ and take the union bound over all J showing that $\text{RE}(X\Psi^\top\Psi, k, \alpha) \geq r/2$.

The condition above requires checking infinitely many $\mathbf{x}$. To make the problem tractable, let us introduce an orthogonal projection $Q : \mathbb{R}^n \to \mathbb{R}^n$ which we discuss more about later. Assume that $QX\Psi^\top\Psi\mathbf{y} \neq 0$, and let $\mathbf{u}$ be the unit vector in the direction of $QX\Psi^\top\Psi\mathbf{y} \neq 0$. Then

$$\begin{aligned}\left\|X\Psi^\top\Psi(\mathbf{y}+\mathbf{x})\right\| &\geq \left\|QX\Psi^\top\Psi(\mathbf{y}+\mathbf{x})\right\| \geq \mathbf{u}^\top QX\Psi^\top\Psi(\mathbf{y}+\mathbf{x}) \\ &= \left\|QX\Psi^\top\Psi\mathbf{y}\right\| + \mathbf{u}^\top QX\Psi^\top\Psi\mathbf{x}\end{aligned}$$

The quantity above is affine in $\mathbf{x}$, so it is minimized at one of the extreme points of the set $\{\mathbf{x} \in \mathbb{R}^d : \operatorname{supp}(\mathbf{x}) \subseteq J^{\text{co}},\ \|\mathbf{x}\|_1 \leq \alpha\sqrt{k}\}$, i.e., at a vector $\pm\alpha\sqrt{k}\mathbf{e}_j, j \in J^{\text{co}}$. This observation allows us to pass from an infinite set of $\mathbf{x}$'s to a finite set.

Next, we have to establish the concentration bounds on $\left\|QX\Psi^\top\Psi\mathbf{y}\right\|$ and $\mathbf{u}^\top QX\Psi^\top\Psi\mathbf{e}_j$. Notice that $\Psi\mathbf{y}$ and $\Psi\mathbf{e}_j$ are independent centered (mean 0) subgaussian vectors with the unit variance of the coordinates. If these vectors were independent of the random matrix $\Psi^\top$ as well, we would have used the Hanson-Wright inequality to derive the necessary concentration. However, this is obviously not the case. At this moment, the projection Q comes to the rescue. The idea is to carefully construct the projection to take care of the dependencies.

2.1. Proof of Theorem 5: Technical Ingredients

In this section, we describe some of the key technical ingredients behind the proof of Theorem 5. The complete proof is provided in Appendix B. Throughout the proof, we use the Hanson-Wright inequality and its corollaries to get probabilistic estimates for norms of certain matrix products (that we discuss in Appendix B.1).

Bounds for a Fixed Vector. Our first goal will be to investigate a special case of Theorem 5. In particular, we investigate the RE condition in Definition 4 when restricted to vectors of the kind $\mathbf{z} = \mathbf{e}_j + \mathbf{x}$ for a fixed j where $j \notin \operatorname{supp}(\mathbf{x})$ (Proposition 10). The proof is based on two technical lemmas that use careful conditioning arguments along with the probabilistic inequalities that can be derived from Hanson-Wright inequality. We use $\operatorname{conv}()$ to denote the convex hull of a set of vectors. In the following discussion, B and R are fixed matrices and G is a random matrix.

The following lemma bounds the small ball probability of $BG^\top\mathbf{g}$, for a fixed matrix B, random matrix G, and a random vector $\mathbf{g}$.

Lemma 8 *Let B be a fixed $n \times d$ matrix, let $G = (G_{ij})$ be an $m \times d$ random matrix with independent entries and let $\mathbf{g} = (g_1, \ldots, g_m) \in \mathbb{R}^m$ be a random vector with independent entries that satisfy: $\mathbb{E}[G_{ij}] = \mathbb{E}[g_j] = 0$, $\mathbb{E}[G_{ij}^2] = \mathbb{E}[g_j^2] = 1$, and $\|G_{ij}\|_{\psi_2}, \|g_j\|_{\psi_2}$ are bounded. Then*

$$\Pr\left[\left\|BG^\top\mathbf{g}\right\| < \frac{1}{4}\sqrt{m}\,\|B\|_{\text{F}}\right] \leq 8\Big(\exp\big(-c\,\mathbf{sr}(B)\big) + \exp(-cm)\Big).$$

The following lemma provides a large deviation bound for a certain product form.

Lemma 9 *Let B be a fixed $n \times d$ matrix, let $G = (G_{ij})$ be an $m \times d$ random matrix with independent entries and let $\mathbf{g}_1 = (g_{1_1}, \ldots, g_{1_m}) \in \mathbb{R}^m$ and $\mathbf{g}_2 = (g_{2_1}, \ldots, g_{2_m}) \in \mathbb{R}^m$ be random vectors with independent entries that satisfy: $\mathbb{E}[G_{ij}] = \mathbb{E}[g_{l_j}] = 0$, $\mathbb{E}[G_{ij}^2] = \mathbb{E}[g_{l_j}^2] = 1$, and $\|G_{ij}\|_{\psi_2}, \|g_{l_j}\|_{\psi_2}$ are all bounded for $l \in \{1, 2\}$. Assume that $m \leq \mathbf{sr}(B)$. Then for any $t \in \left[0, m\,\|B\|_{\text{F}}^2\right]$,*

$$\Pr\left[|\mathbf{g}_1^\top GB^\top BG^\top\mathbf{g}_2| \geq t\right] \leq 10\exp\left(-c\frac{t^2}{m\,\|B\|_{\text{F}}^4}\right).$$

Using Lemmas 8 and 9, we are ready to prove the following proposition. The main idea here is to introduce an orthogonal projection matrix which lets us decouple various dependencies that appear across various quantities.

Proposition 10 *Let R be a fixed $n \times d$ matrix, and let $G = (G_{i,j})$ be an $m \times d$ random matrix with independent entries that satisfy: $\mathbb{E}[G_{ij}] = 0$, $\mathbb{E}[G_{ij}^2] = 1$, and $\|G_{ij}\|_{\psi_2}$ is bounded. Assume that*

$$2 \leq m \leq \mathbf{sr}(R)/2.$$

Then for any $s \geq 1$,

$$\Pr\left[\exists \mathbf{x} \in s \cdot \mathrm{conv}(\pm\mathbf{e}_2, \ldots, \pm\mathbf{e}_d),\ \left\|RG^\top G(\mathbf{e}_1 + \mathbf{x})\right\| \leq \frac{1}{8}\sqrt{m}\,\|R\|_{\mathrm{F}}\right] \leq 2d\exp\left(-c\frac{m}{s^2}\right).$$

Finishing the Proof of Theorem 5 using a Net Argument. The next theorem is the main technical step in proving Theorem 5. Invoking this theorem with appropriate parameters gives the proof of Theorem 5. The proof of the following theorem is based on generating an orthogonal matrix to reduce the general case to the special case discussed in Proposition 10, and then employing an ε-net argument.

Theorem 11 *Let X be a fixed $n \times d$ matrix satisfying,*

$$2 \leq m \leq \mathbf{sr}(X)/2.$$

Let $\Psi = (\Psi_{ij})$ be an $m \times d$ random matrix with independent entries such that $\mathbb{E}[\Psi_{ij}] = 0$, $\mathbb{E}[\Psi_{ij}^2] = 1$, and $\|\Psi_{ij}\|_{\psi_2}$ is bounded. Let $\beta \in (0,1)$, and let $k \in \mathbb{N}$. Then for any s such that

$$1 \leq s \leq \sqrt{\frac{cm}{k\log d + \log(2/\beta)}},$$

$$\Pr[\exists I \subset [d] \text{ with } |I| = k, \exists \mathbf{y} \in \mathbb{S}^{d-1} \text{ with } \mathrm{supp}(\mathbf{y}) \subseteq I, \exists \mathbf{x} \in s \cdot \mathrm{conv}(\pm\mathbf{e}_i,\ i \notin I),$$
$$\left\|X\Psi^\top\Psi(\mathbf{y} + \mathbf{x})\right\| \leq \frac{1}{32}\sqrt{m}\,\|X\|_{\mathrm{F}}] \leq \beta.$$

Note that the condition $s \geq 1$ in the formulation of the theorem implicitly sets a lower bound on β and an upper bound on k. We now have all the ingredients to complete the proof of Theorem 5.

Proof [Proof of Theorem 5] Assume that the complement of the event described in Theorem 11 occurs. Namely, assume that

$$\forall I \subset [d] \text{ with } |I| = k, \forall \mathbf{y} \in \mathbb{S}^{d-1} \text{ with } \mathrm{supp}(\mathbf{y}) \subseteq I, \forall \mathbf{x} \in s \cdot \mathrm{conv}(\pm\mathbf{e}_i,\ i \notin I)$$
$$\left\|X\Psi^\top\Psi(\mathbf{y} + \mathbf{x})\right\| \geq \frac{1}{32}\sqrt{m}\,\|X\|_{\mathrm{F}}.$$

If s satisfies the condition of this theorem, then the event above occurs with probability at least $1 - \beta$. Pick any $I \subset [d]$ $|I| = k$ and any $\mathbf{z} \in \mathbb{R}^d \setminus \{0\}$ with

$$\|\mathbf{z}_{I^{\mathrm{co}}}\|_1 \leq \alpha\,\|\mathbf{z}_I\|_1.$$

Without loss of generality, we may assume that $\mathbf{y} = \mathbf{z}_I \in \mathbb{S}^{d-1}$. Then, $\|\mathbf{y}\|_1 \leq \sqrt{k}$, and so $\|\mathbf{z}_{I^{\mathrm{co}}}\|_1 \leq \alpha\sqrt{k}$. Theorem 5 now follows from Theorem 11 applied with $s = \alpha\sqrt{k}$ and by plugging $\Phi = \Psi/\sqrt{m}$. ∎

3. Applications to Sparse Linear Regression

We now discuss some applications of our RE bound to the setting of sparse linear regression. We consider two different problems: (a) first one involves a standard regression setting with $Z = X\Phi^\top\Phi$ acting as a design matrix, and the goal is to estimate the sparse $\theta^\star$ from a noisy linear model of observations (b) second one is a variant of sparse linear regression, where the algorithm has access not to the individual covariates, but rather only to a randomly projected version of them, and the goal is to minimize (in-sample) prediction error. Missing details from this section are collected in Appendix C.

3.1. Application 1: Bounding the ℓ_2-error with Random Design $Z = X\Phi^\top\Phi$

Consider the linear model $\mathbf{y} = Z\theta^\star + \mathbf{w}$, where $\mathbf{w}$ is an i.i.d. subgaussian noise. The following proposition establishes a ℓ_2-error bound on estimating $\theta^\star$, using the standard Lasso analysis framework from Theorem 2. This result shows that ℓ_1-relaxations succeed in estimating $\theta^\star$ even for certain dependent design matrices, partially justifying an observation commonly noticed in practice of Lasso succeeding even when the entries of the design matrix has dependencies. We work with a Lasso formulation defined on the pair $(Z, \mathbf{y})$;

$$\theta^{\text{comp}} \in \text{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{y} - Z\theta\|^2 + \lambda\|\theta\|_1 = \text{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}(y_i - \langle\Phi\mathbf{x}_i, \Phi\theta\rangle)^2 + \lambda\|\theta\|_1. \quad (3)$$

The following proposition states the convergence bound of θ^{comp} to $\theta^\star$ under this linear model. The probability in this case is over both the noise realization $\mathbf{w}$ and the randomness in Φ.

Proposition 12 *Let X be a deterministic matrix and Φ be a random matrix satisfying the conditions of Theorem 5. Consider the linear model $\mathbf{y} = X\Phi^\top\Phi\theta^\star + \mathbf{w}$ where the entries of the noise vector $\mathbf{w} = (w_1, \ldots, w_n)$ are independent centered subgaussians with $\|w_i\|_{\psi_2} \le \sigma$. Let $K > 0$ be any constant, and let $dm^{-K} \le \beta < 1$. Then $\theta^{\text{comp}} \in \text{argmin}_{\theta\in\mathbb{R}^d_2} \|\mathbf{y} - X\Phi^\top\Phi\theta\|^2/n + \lambda\|\theta_1\|$ with $\lambda = \Theta(\sigma\|X\|_F/n\sqrt{m})$, satisfies with probability at least $1-\beta$: $\|\theta^{\text{comp}} - \theta^\star\| = O\left(\sigma\sqrt{m}k^{3/2}/\|X\|_F\right)$.*

3.2. Application 2: Sparse Linear Regression with Compressed Features

In this section, we use the results from Section 2 on a sparse linear regression in a model where the regression algorithm only gets access to a compressed representation of the $\mathbf{x}_i$'s in the form of $\Phi\mathbf{x}_i$'s, and not to $\mathbf{x}_i$'s.[7] As discussed in Section 1, these compressed representations of the $\mathbf{x}_i$'s are easier to communicate in a distributed data setting and also reduces the storage requirements as we work with the compressed data. Consider the linear model $\mathbf{y} = X\theta^\star + \mathbf{w}$, where X is some deterministic matrix and $\mathbf{w}$ is subgaussian noise. Note that this is a fixed design setting (unlike the application in Section 3.1).

Since the linear model is $\mathbf{y} = X\theta^\star + \mathbf{w}$, and we only assume a rather weak stable rank assumption on X, getting an error bound on $\theta^\star$ is ruled out because as shown by (Raskutti et al., 2011) a condition closely related to restricted eigenvalue is needed for any parameter recovery method.[8] Therefore,

7. Throughout this section, we will assume that Φ is known to the algorithm.
8. One simple illustration of why a stable rank condition on X is not enough for parameter recovery, is that $\mathbf{sr}(X) \ge m$ (for some m) does not rule $X\theta^\star = 0$, which means $\mathbf{y} = \mathbf{w}$ implying $\mathbf{y}$ provides no information about $\theta^\star$, making any recovery of $\theta^\star$ impossible.

in this section, we measure the performance in terms of minimizing *mean-squared (in-sample) prediction error*. Given $(\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$, the goal is to output $\theta \in \mathbb{R}^d$ that has a relatively low prediction error $\|X\theta - X\theta^\star\|^2/n$. In a matrix-vector form, $(\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$ can be represented as $(X\Phi^\top, \mathbf{y})$. Now in a traditional sparse linear regression setting (with access to $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$) this minimization can be performed without any assumptions on the design matrix X (with faster convergence bounds possible under the RE assumption) (Bickel et al., 2009). However, under the compressed setup, *a priori* it is unclear whether this seemingly simpler problem can even be solved consistently.

A first idea given only $\Phi\mathbf{x}_i$'s will be to: (a) for all i, construct $\hat{\mathbf{x}}_i$, an approximation to $\mathbf{x}_i$ from $\Phi\mathbf{x}_i$, (b) use the Lasso formulation on the resulting $(\hat{\mathbf{x}}_i, y_i)$'s. This idea, however, is problematic because good reconstruction of $\mathbf{x}_i$'s from $\Phi\mathbf{x}_i$'s will require some strong (sparsity-like) assumptions on the structure of the $\mathbf{x}_i$'s, that is generally untrue. Another idea will be to construct an estimator $\hat{\vartheta} \in \mathbb{R}^m$ in the projected space say by minimizing the squared loss between $\mathbf{y}$ and $X\Phi^\top\vartheta$ (over $\vartheta \in \mathbb{R}^m$). This minimization would correspond to a different linear model: $\mathbf{y} = X\Phi^\top\vartheta^\star + \hat{\mathbf{w}}$. Since the true linear model is $\mathbf{y} = X\theta^\star + \mathbf{w}$, this would mean the new noise vector $\hat{\mathbf{w}} = (X\theta^\star - X\Phi^\top\vartheta^\star) + \mathbf{w}$ is no longer i.i.d. subgaussian. For bounding the prediction error (which in this case means bounding the norm of the difference between $X\Phi^\top\hat{\vartheta}$ and $X\theta^\star$) this could be problematic. Additionally, given $\hat{\vartheta}$, lifting it to $\mathbb{R}^d$ is problematic as $\hat{\vartheta}$ may not be close to a projection of a sparse vector in $\mathbb{R}^d$. We overcome these hurdles by working with a Lasso formulation defined on the pair $(Z = X\Phi^\top\Phi, \mathbf{y})$ as in (3). Again define: $\theta^{\text{comp}} \in \text{argmin}_{\theta\in\mathbb{R}^d} \|\mathbf{y} - X\Phi^\top\Phi\theta\|^2/n + \lambda\|\theta\|_1$. Our basic idea is to establish a bound on $\|Z\theta^{\text{comp}} - Z\theta^\star\|$, with the error vector $(\theta^{\text{comp}} - \theta^\star)$ satisfying the cone set condition, and then using the RE bound on Z to relate $\|Z\theta^{\text{comp}} - Z\theta^\star\|$ and $\|X\theta^{\text{comp}} - X\theta^\star\|$. We get the following result in this compressed feature setting.

Proposition 13 [9] *Let X be a deterministic matrix and Φ be a Gaussian random matrix satisfying the conditions of Theorem 5. Consider the linear model $\mathbf{y} = X\theta^\star + \mathbf{w}$ where the entries of the noise vector $\mathbf{w} = (w_1, \ldots, w_n)$ are independent centered subgaussians with $\|w_i\|_{\psi_2} \le \sigma$ and $\theta^\star \in S^{d-1}$. Let $K > 0$ be any constant, and let $dm^{-K} \le \beta < 1$. Then $\theta^{\text{comp}} \in$ argmin$_{\theta\in B_2^d} \|\mathbf{y} - X\Phi^\top\Phi\theta\|^2/n + \lambda\|\theta_1\|$ with $\lambda = \Theta(\sigma\|X\|_{\text{F}} \log m/n\sqrt{m} + \|X\|_{\text{F}}^2/nm)$, satisfies with probability at least $1 - \beta$:*

$$\frac{1}{n}\|X\theta^{\text{comp}} - X\theta^\star\|^2 = O\left(\frac{\sigma\|X\|_{\text{F}}k^{3/2}\log m}{n\sqrt{m}} + \frac{\|X\|_{\text{F}}^2 k^{3/2}}{nm}\right).$$

Remark 14 *For small σ, the dominant term in the error bound in Proposition 13 is the $\|X\|_{\text{F}}^2 k^{3/2}/nm$ term. If we set, $m = \mathbf{sr}(X)/2$, then $\|X\|_{\text{F}}^2 k^{3/2}/nm = 2\|X\|^2 k^{3/2}/n$, and therefore in this case we get a consistent prediction if $\|X\| = o(\sqrt{n}/k^{3/4})$.*

References

Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

9. We assume that the algorithm has an estimate of $\|X\|_F$ (a good upper bound suffices). This is easy achievable in the distributed data setting described in Section 1 (Figure 1), as each device i in addition to $(\Phi\mathbf{x}_i, y_i)$ can also communicate $\|\mathbf{x}_i\|$ to the server.

Afonso S Bandeira, Matthew Fickus, Dustin G Mixon, and Joel Moreira. Derandomizing restricted isometries via the legendre symbol. *Constructive Approximation*, 43(3):409–424, 2016.

Afonso S Bandeira, Dustin G Mixon, and Joel Moreira. A conditional construction of restricted isometries. *International Mathematics Research Notices*, 2017(2):372–381, 2017.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, Denka Kutzarova, et al. Explicit constructions of rip matrices and related problems. *Duke Mathematical Journal*, 159(1):145–185, 2011.

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

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

Mahdi Cheraghchi. Coding-theoretic methods for sparse recovery. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 909–916. IEEE, 2011.

Yohann De Castro. Optimal designs for lasso and dantzig selector using expander codes. *IEEE Transactions on Information Theory*, 60(11):7293–7299, 2014.

Edgar Dobriban and Jianqing Fan. Regularity properties for sparse regression. *Communications in mathematics and statistics*, 4(1):1–19, 2016.

Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

Mahdi Milani Fard, Yuri Grinberg, Joelle Pineau, and Doina Precup. Compressed least-squares regression on sparse spaces. In *AAAI*, 2012.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

Ata Kabán. New bounds on compressive linear least squares regression. In *The 17-th International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, volume 33, pages 448–456, 2014.

Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society*, 19(3):881–904, 2017.

Jason D Lee, Yuekai Sun, Qiang Liu, and Jonathan E Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.

Odalric Maillard and Rémi Munos. Compressed least-squares regression. In *NIPS*, 2009.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4), 2012.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3-4):1175–1194, 2016.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *JMLR*, 11:2241–2259, 2010.

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

Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013.

Vidyashankar Sivakumar, Arindam Banerjee, and Pradeep K Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. In *Advances in neural information processing systems*, pages 2206–2214, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Ryan Tibshirani and Larry Wasserman. Sparsity and the lasso, http://www.stat.cmu.edu/~larry/=sml/sparsity.pdf, 2015.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5), 2009.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *FnT-TCS*, 10(1–2):1–157, 2014. URL http://dx.doi.org/10.1561/0400000060.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Compressed and privacy-sensitive sparse regression. *Information Theory, IEEE Transactions on*, 55(2):846–866, 2009.

Appendix A. Additional Preliminaries

Background on Sparse Linear Regression. If the linear model $\mathbf{y} = M\theta^\star + \mathbf{w}$, where $M \in \mathbb{R}^{n\times d}$ is high-dimensional in nature, meaning that the number of observations n is substantially smaller than d, then it is easy to see that without further constraints on $\theta^\star$, the statistical model $\mathbf{y} = M\theta^\star + \mathbf{w}$ is not *identifiable*. This is because (even when $\mathbf{w} = 0$), there are many vectors $\theta^\star$ that are consistent with the observations $\mathbf{y}$ and M. This identifiability concern may be eliminated by imposing some type of sparsity assumption on the regression vector $\theta^\star$. Typically, $\theta^\star$ is k-sparse for $k \ll d$. Disregarding computational cost, the most direct approach to estimating a k-sparse θ in the linear regression model would be solving a quadratic optimization problem with an ℓ_0-constraint:

$$\theta^{\text{sparse}} \in \text{argmin}_{\theta\in\Sigma_k} \frac{1}{n}\|\mathbf{y} - M\theta\|^2. \tag{4}$$

Lasso Regression. Since (4) leads to a non-convex problem, a natural alternative is obtained by replacing the ℓ_0-constraint with its tightest convex relaxation, the ℓ_1-norm. This leads to the popular Lasso regression, defined as,

$$\text{Lasso Regression (penalized form):} \quad \theta^{\text{Lasso}} \in \text{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y} - M\theta\|^2. + \lambda\|\theta\|_1,$$

for some choice $\lambda > 0$.

The consistency properties of Lasso are well-understood. Under a variety of mild assumptions on the instance, the Lasso estimator (θ^{Lasso}) is known to converge to the sparse $\theta^\star$ in the ℓ_2-norm. Under stronger assumptions (such as mutual incoherence, minimum eigenvalue, and minimum signal condition) on the instance, it is also known that θ^{Lasso} will have the same support as $\theta^\star$. We refer the reader to the recent book (Hastie et al., 2015) for a detailed survey of developments in this area.

Background on ε-Nets. Consider a subset T of $\mathbb{R}^d$, and let $\varepsilon > 0$. A ε-net of T is a subset $\mathcal{N} \subseteq T$ such that for every $\mathbf{x} \in T$, there exists a $\mathbf{y} \in \mathcal{N}$ such that $\|\mathbf{x} - \mathbf{y}\| \leq \varepsilon$.

Proposition 15 (Volumetric Estimate) *Let T be a subset of B_2^d and let $\varepsilon > 0$. Then there exists an ε-net N of T of cardinality at most $(1 + 2/\varepsilon)^d$. For any $\varepsilon \leq 1$, this can be simplified as $(1 + 2/\varepsilon)^d \leq (3/\varepsilon)^d$.*

Background on Subgaussian Random Variables. Subgaussian random variables are a wide class of random variables, which contains in particular the standard normal, Bernoulli, and all bounded random variables.

Definition 16 (Subgaussian Random Variable) *We call a random variable $x \in \mathbb{R}$ subgaussian if there exists a constant $C > 0$ if $\Pr[|x| > t] \leq 2\exp(-t^2/C^2)$ for all $t > 0$.*

Definition 17 (Norm of a Subgaussian Random Variable) *The ψ_2-norm of a subgaussian random variable $x \in \mathbb{R}$, denoted by $\|x\|_{\psi_2}$ is: $\|x\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}[\exp(|x|^2/t^2)] \leq 2\right\}$.*

Note that the ψ_2 condition on a scalar random variable x is equivalent to the subgaussian tail decay of x.

Appendix B. Complete Proof of Theorem 5

In this section, we provide the complete proof of our RE bound for the random matrix $X\Phi^\top\Phi$. In Section B.1, we use the Hanson-Wright theorem and its corollaries to get probabilistic estimates for norms of certain matrix products. In Section B.2, we prove Theorem 5 for a fixed vector of a special form. We finish the proof in Section B.3.

B.1. Hanson-Wright Preliminaries

We start by establishing probability estimates for the spectral and Frobenius norms for certain matrix products. The results in this section form the basic building blocks that are used throughout the proof. An important tool used here is the Hanson-Wright inequality and its several consequences. Hanson-Wright inequality establishes the concentration of a quadratic form of independent centered subgaussian random variables. An original (slightly weaker) version of this inequality was first proved in (Hanson and Wright, 1971).

Theorem 18 (Hanson-Wright Inequality (Rudelson and Vershynin, 2013)) *Let $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ be a random vector with independent components x_i which satisfy $\mathbb{E}[x_i] = 0$ and $\|x_i\|_{\psi_2}$ is bounded. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\Pr\left[\left|\mathbf{x}^\top A\mathbf{x} - \mathbb{E}[\mathbf{x}^\top A\mathbf{x}]\right| > t\right] \leq 2\exp\Big(-c\min\Big(\frac{t^2}{\|A\|_{\mathrm{F}}^2}, \frac{t}{\|A\|}\Big)\Big).$$

Besides the theorem itself, we need several corollaries.

Corollary 19 (Spectral Norm of the Product) *Let B be a fixed $n \times d$ matrix, and let $G = (G_{ij})$ be an $m \times d$ random matrix with independent entries that satisfy: $\mathbb{E}[G_{ij}] = 0$, $\mathbb{E}[G_{ij}^2] = 1$, and $\|G_{ij}\|_{\psi_2}$ is bounded. Then for any $s, t \geq 1$,*

$$\Pr\left[\left\|BG^\top\right\| > C(s\,\|B\|_{\mathrm{F}} + t\sqrt{m}\,\|B\|)\right] \leq 2\exp(-s^2\,\mathbf{sr}(B) - t^2 m)$$

and

$$\Pr\left[\left\|BG^\top\right\| < \frac{1}{2}\,\|B\|_{\mathrm{F}}\right] \leq 2\exp(-c\,\mathbf{sr}(B)).$$

Corollary 19 can be found in (Rudelson and Vershynin, 2013). Assuming that $m \leq \mathbf{sr}(B)$, we can rewrite the above inequalities as

$$\Pr\left[\frac{1}{2}\,\|B\|_{\mathrm{F}} < \left\|BG^\top\right\| < C\,\|B\|_{\mathrm{F}}\right] \geq 1 - 2\exp(-c\,\mathbf{sr}(B)). \tag{5}$$

Applying this corollary in the case $m = 1$, we obtain a small ball probability estimate for the image of a subgaussian vector. The small ball probability bounds the probability $\|B\mathbf{g}\|$ is small for a fixed matrix B and a subgaussian vector $\mathbf{g}$.

Corollary 20 (Concentration for the Norm of a Vector) *Let B be a fixed $n \times d$ matrix, and let $\mathbf{g} = (g_1, \dots, g_d) \in \mathbb{R}^d$ be a random vector with independent entries that satisfy $\mathbb{E}[g_j] = 0$, $\mathbb{E}[g_j^2] = 1$, and $\|g_j\|_{\psi_2}$ is bounded. Then*

$$\Pr\left[\frac{1}{2}\,\|B\|_{\mathrm{F}} < \|B\mathbf{g}\| < C\,\|B\|_{\mathrm{F}}\right] \geq 1 - 2\exp(-c\,\mathbf{sr}(B)).$$

Using this inequality, we can easily derive a small ball probability estimate for the Frobenius norm.

Corollary 21 (Frobenius Norm of the Product) *Let B be a fixed $n \times d$ matrix, and let $G = (G_{ij})$ be an $m \times d$ random matrix with independent entries that satisfy: $\mathbb{E}[G_{ij}] = 0$, $\mathbb{E}[G_{ij}^2] = 1$, and $\|G_{ij}\|_{\psi_2}$ is bounded. Then*

$$\Pr\left[\frac{1}{2}\sqrt{m}\,\|B\|_{\mathrm{F}} < \left\|BG^{\top}\right\|_{\mathrm{F}} < C\sqrt{m}\,\|B\|_{\mathrm{F}}\right] \geq 1 - 2\exp(-c\,\mathbf{sr}(B)).$$

Proof Denote the rows of G by $\gamma_1, \ldots, \gamma_m$. Then,

$$\left\|BG^{\top}\right\|_{\mathrm{F}} = \left(\sum_{j=1}^{m} \|B\gamma_j\|^2\right)^{1/2}.$$

The right-hand side can be interpreted as the Euclidean norm of the image of the vector $\tilde{\gamma} \in \mathbb{R}^{dm}$ obtained by concatenation of the vectors $\gamma_1, \ldots, \gamma_m$ under the $nm \times dm$ block-diagonal matrix $\tilde{B} = \mathrm{diag}(B, \ldots, B)$. The result follows from the Corollary 20, since $\left\|\tilde{B}\right\|_{\mathrm{F}}^2 = m\,\|B\|_{\mathrm{F}}^2$ implying $\left\|\tilde{B}\right\|_{\mathrm{F}} = \sqrt{m}\,\|B\|_{\mathrm{F}}$. ∎

We will need a similar estimate for the Frobenius norm of the triple product of the form $GHG^{\top}$, where H is a positive semidefinite matrix. Let $\mathrm{tr}()$ denote the trace of a matrix.

Corollary 22 (Frobenius norm of the Triple Product) *Let H be a fixed $d \times d$ symmetric positive semidefinite matrix, and let $G = (G_{ij})$ be an $m \times d$ random matrix with independent entries that satisfy: $\mathbb{E}[G_{ij}] = 0$, $\mathbb{E}[G_{ij}^2] = 1$, and $\|G_{ij}\|_{\psi_2}$ is bounded. If $m \leq \mathrm{tr}(H)/\,\|H\|$, then*

$$\Pr\left[\left\|GHG^{\top}\right\|_{\mathrm{F}} \geq C\sqrt{m} \cdot \mathrm{tr}(H)\right] \leq 4\exp\left(-c\frac{\mathrm{tr}(H)}{\|H\|}\right).$$

Proof Let $H^{1/2}$ be the square root of the matrix H. Since $\mathrm{tr}(H) = \left\|H^{1/2}\right\|_{\mathrm{F}}^2$, the assumption of the corollary reads $m \leq \mathbf{sr}(H^{1/2})$. By Corollary 20,

$$\Pr\left[\left\|H^{1/2}G^{\top}\right\| \geq C\left\|H^{1/2}\right\|_{\mathrm{F}}\right] \leq 2\exp(-c\,\mathbf{sr}(H^{1/2})).$$

Similarly, Corollary 21 implies

$$\Pr[\left\|H^{1/2}G^{\top}\right\|_{\mathrm{F}} \geq C\sqrt{m}\left\|H^{1/2}\right\|_{\mathrm{F}}] \leq 2\exp(-c\,\mathbf{sr}(H^{1/2})).$$

As $\left\|GHG^{\top}\right\|_{\mathrm{F}} \leq \left\|H^{1/2}G^{\top}\right\|_{\mathrm{F}} \cdot \left\|H^{1/2}G^{\top}\right\|$, we have

$$\Pr\left[\left\|GHG^{\top}\right\|_{\mathrm{F}} \geq C\sqrt{m} \cdot \left\|H^{1/2}\right\|_{\mathrm{F}}^2\right] \leq 4\exp(-c\,\mathbf{sr}(H^{1/2})) = 4\exp\left(-c\frac{\mathrm{tr}(H)}{\|H\|}\right),$$

which completes the proof. ∎

B.2. Bounds for a Fixed Vector

In this section, our goal will be to investigate a special case of Theorem 5. In particular, we investigate the RE condition in Definition 4 when restricted to vectors of the kind $\mathbf{z} = \mathbf{e}_j + \mathbf{x}$ for a fixed j where $j \notin \text{supp}(\mathbf{x})$ (Proposition 25). The proof is based on two technical lemmas that use careful conditioning arguments along with the probabilistic inequalities established in the previous section. We use $\text{conv}()$ and $\text{span}()$ to denote the convex hull and span of a set of vectors. We use $\text{Ker}()$ to denote the kernel of a matrix.

The following lemma bounds the small ball probability of $BG^\top \mathbf{g}$, for a fixed matrix B, random matrix G, and a random vector $\mathbf{g}$.

Lemma 23 (Restated Lemma 8) *Let B be a fixed $n \times d$ matrix, let $G = (G_{ij})$ be an $m \times d$ random matrix with independent entries and let $\mathbf{g} = (g_1, \ldots, g_m) \in \mathbb{R}^m$ be a random vector with independent entries that satisfy: $\mathbb{E}[G_{ij}] = \mathbb{E}[g_j] = 0$, $\mathbb{E}[G_{ij}^2] = \mathbb{E}[g_j^2] = 1$, and $\|G_{ij}\|_{\psi_2}, \|g_j\|_{\psi_2}$ are bounded. Then*

$$\Pr\left[\left\|BG^\top \mathbf{g}\right\| < \frac{1}{4}\sqrt{m}\,\|B\|_{\text{F}}\right] \leq 8\Big(\exp\big(-c\,\mathbf{sr}(B)\big) + \exp(-cm)\Big).$$

Proof Conditioning on G and applying Corollary 20, we obtain

$$\Pr\left[\left\|BG^\top \mathbf{g}\right\| \leq \frac{1}{2}\left\|BG^\top\right\|_{\text{F}} \mid G\right] \leq 2\exp(-c\,\mathbf{sr}(BG^\top)).$$

Define the events Ω_{F} and Ω_{op} as in Corollary 22:

$$\Omega_{\text{F}} = \left\{G : \left\|BG^\top\right\|_{\text{F}} \geq \frac{1}{2}\sqrt{m}\,\|B\|_{\text{F}}\right\}$$
$$\Omega_{op} = \left\{G : \left\|BG^\top\right\| \leq C(\|B\|_{\text{F}} + \sqrt{m}\,\|B\|)\right\}$$

Let $\Omega_{\text{F}}^{\text{co}}$ and Ω_{op}^{co} denote the complement of these events respectively. Then by Corollaries 21 and 19,

$$\begin{aligned}
&\Pr\left[\left\|BG^\top \mathbf{g}\right\| \leq \frac{1}{4}\sqrt{m}\,\|B\|_{\text{F}}\right] \\
&\leq \Pr\left[\left\|BG^\top \mathbf{g}\right\| \leq \frac{1}{2}\left\|BG^\top\right\|_{\text{F}} \mid G \in \Omega_{\text{F}} \cap \Omega_{op}\right] + \Pr\left[\Omega_{\text{F}}^{\text{co}}\right] + \Pr\left[\Omega_{op}^{\text{co}}\right] \\
&\leq 2\exp\left(-c\frac{m\,\|B\|_{\text{F}}^2}{\|B\|_{\text{F}}^2 + m\,\|B\|^2}\right) + 4\exp(-c\,\mathbf{sr}(B)) \\
&\leq 8\Big(\exp\big(-c\,\mathbf{sr}(B)\big) + \exp(-cm)\Big).
\end{aligned}$$

■

The following lemma provides a large deviation bound for a certain product form.

Lemma 24 (Restated Lemma 9) *Let B be a fixed $n \times d$ matrix, let $G = (G_{ij})$ be an $m \times d$ random matrix with independent entries and let $\mathbf{g}_1 = (g_{1_1}, \ldots, g_{1_m}) \in \mathbb{R}^m$ and $\mathbf{g}_2 = (g_{2_1}, \ldots, g_{2_m}) \in \mathbb{R}^m$ be random vectors with independent entries that satisfy: $\mathbb{E}[G_{ij}] = \mathbb{E}[g_{l_j}] = 0$, $\mathbb{E}[G_{ij}^2] = \mathbb{E}[g_{l_j}^2] = 1$,*

and $\|G_{ij}\|_{\psi_2}, \|g_{l_j}\|_{\psi_2}$ *are all bounded for* $l \in \{1,2\}$. *Assume that* $m \leq \mathbf{sr}(B)$. *Then for any* $t \in \left[0, m\|B\|_{\mathrm{F}}^2\right]$,

$$\Pr\left[|\mathbf{g}_1^\top G B^\top B G^\top \mathbf{g}_2| \geq t\right] \leq 10\exp\left(-c\frac{t^2}{m\|B\|_{\mathrm{F}}^4}\right).$$

Proof Define the vector $\mathbf{g} \in \mathbb{R}^{2m}$ and the $2m \times 2m$ matrix Γ by

$$\mathbf{g} = \begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{pmatrix}, \qquad \Gamma = \begin{pmatrix}0 & GB^\top BG^\top\\ GB^\top BG^\top & 0\end{pmatrix}.$$

Condition on G. By Theorem 18, for any $t \geq 0$,

$$\Pr\left[|\mathbf{g}^\top \Gamma \mathbf{g}| > t\right] \leq 2\exp\Big[-c\min\Big(\frac{t^2}{\|\Gamma\|_{\mathrm{F}}^2}, \frac{t}{\|\Gamma\|}\Big)\Big].$$

Note that $\|\Gamma\| = \left\|GB^\top BG^\top\right\| = \left\|BG^\top\right\|^2$. Let Ω_F and Ω_{op} be the events defined by

$$\Omega_F = \left\{G:\ \left\|GB^\top BG^\top\right\|_{\mathrm{F}} \leq C\left(m\left\|B^\top B\right\|_{\mathrm{F}} + \sqrt{m}\cdot\mathrm{tr}(B^\top B)\right)\right\},$$
$$\Omega_{op} = \left\{G:\ \tfrac{1}{4}\|B\|_{\mathrm{F}}^2 \leq \left\|GB^\top BG^\top\right\| \leq C\|B\|_{\mathrm{F}}^2\right\}.$$

Again, let $\Omega_{\mathrm{F}}^{\mathrm{co}}$ and $\Omega_{op}^{\mathrm{co}}$ denote the complement events. For any $G \in \Omega_{\mathrm{F}}$,

$$\|\Gamma\|_{\mathrm{F}}^2 \leq Cm\cdot\mathrm{tr}(B^\top B)^2 = C'm\|B\|_{\mathrm{F}}^4.$$

Notice that

$$\frac{\mathrm{tr}(B^\top B)}{\|B^\top B\|} = \mathbf{sr}(B).$$

Finally, combining this with Corollary 22, and (5), we obtain

$$\begin{aligned}\Pr\left[|g_1^\top GB^\top BG^\top g_2| \geq t\right] &\leq 2\exp\Big[-c\min\Big(\frac{t^2}{m\|B\|_{\mathrm{F}}^4}, \frac{t}{\|B\|_{\mathrm{F}}^2}\Big)\Big] + \Pr\left[\Omega_{\mathrm{F}}^{\mathrm{co}}\right] + \Pr\left[\Omega_{op}^{\mathrm{co}}\right]\\ &\leq 4\exp\left(-c\frac{t^2}{m\|B\|_{\mathrm{F}}^4}\right) + 6\exp(-c\,\mathbf{sr}(B))\end{aligned}$$

for any $t \in \left[0, m\|B\|_{\mathrm{F}}^2\right]$. Since $m \leq \mathbf{sr}(B)$, the first term in the right-hand side dominates the second one, and the proof is complete. ■

Using Lemmas 23 and 24, we are ready to prove the following proposition. The main idea here is to introduce an orthogonal projection matrix which lets us decouple various dependencies that appear across various quantities.

Proposition 25 (Restated Proposition 10) *Let* R *be a fixed* $n \times d$ *matrix, and let* $G = (G_{i,j})$ *be an* $m \times d$ *random matrix with independent entries that satisfy:* $\mathbb{E}[G_{ij}] = 0$, $\mathbb{E}[G_{ij}^2] = 1$, *and* $\|G_{ij}\|_{\psi_2}$ *is bounded. Assume that*

$$2 \leq m \leq \mathbf{sr}(R)/2.$$

Then for any $s \geq 1$,

$$\Pr\left[\exists \mathbf{x} \in s\cdot\mathrm{conv}(\pm\mathbf{e}_2, \ldots, \pm\mathbf{e}_d),\ \left\|RG^\top G(\mathbf{e}_1 + \mathbf{x})\right\| \leq \frac{1}{8}\sqrt{m}\|R\|_{\mathrm{F}}\right] \leq 2d\exp\left(-c\frac{m}{s^2}\right).$$

Proof Let P_1 be the orthogonal projection in $\mathbb{R}^n$ with $\mathrm{Ker}(P_1) = \mathrm{span}(R\mathbf{e}_1)$, where span() denote the span. Assume that $P_1 R G^\top G\mathbf{e}_1 \neq 0$ and set

$$\mathbf{u} = \frac{P_1 R G^\top G\mathbf{e}_1}{\|P_1 R G^\top G\mathbf{e}_1\|}.$$

Then

$$\left\|RG^\top G(\mathbf{e}_1+\mathbf{x})\right\| \geq \left\|P_1 RG^\top G(\mathbf{e}_1+\mathbf{x})\right\| \geq \left\|P_1 RG^\top G\mathbf{e}_1\right\| - \mathbf{u}^\top P_1 RG^\top G\mathbf{x}. \tag{6}$$

The minimal value of this expression over $\mathbf{x} \in s\cdot \mathrm{conv}(\pm\mathbf{e}_2,\dots,\pm\mathbf{e}_d)$ is attained at the extreme points of this set. Consider $\mathbf{x} = s\mathbf{e}_2$ since all other extreme points are treated the same way. Since $\mathbf{sr}(R) > 4$ and by the interlacing, we have

$$\|P_1 R\|_{\mathrm{F}}^2 \geq \|R\|_{\mathrm{F}}^2 - \|R\|^2 \geq \|R\|_{\mathrm{F}}^2/2$$

and so, $\mathbf{sr}(P_1 R) \geq (1/2)\,\mathbf{sr}(R)$ (as $\|P_1 R\| = \|R\|$).

Denote by $\mathbf{g}_1$ and $\mathbf{g}_2$ the first and the second columns of G. We have introduced P_1 to ensure that that the matrix $P_1 RG^\top$ is independent of $\mathbf{g}_1$. This allows us to replace the vector $\mathbf{g}_1$ by its copy independent of G. Hence, by Lemma 23,

$$\begin{aligned}\Pr\left[\left\|P_1 RG^\top G\mathbf{e}_1\right\| < \frac{1}{4}\sqrt{m}\,\|R\|_{\mathrm{F}}\right] &= \Pr\left[\left\|P_1 RG^\top \mathbf{g}_1\right\| < \frac{1}{4}\sqrt{m}\,\|R\|_{\mathrm{F}}\right] \\ &\leq 8\Big(\exp\big(-c\,\mathbf{sr}(R)\big) + \exp(-cm)\Big) \leq 2\exp(-c'm),\end{aligned} \tag{7}$$

where we used that $m \leq \mathbf{sr}(R)$.

The estimate of the inner product is a little more complicated. Let P_2 be the orthogonal projection with $\mathrm{Ker}(P_2) = \mathrm{span}(R\mathbf{e}_1, P_1 R\mathbf{e}_2)$. Then we can write

$$\begin{aligned}P_1 RG^\top G\mathbf{e}_1 &= P_2 RG^\top \mathbf{g}_1 + P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_1\\ P_1 RG^\top G\mathbf{e}_2 &= P_2 RG^\top \mathbf{g}_2 + P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_2\end{aligned}$$

and therefore,

$$(P_1 RG^\top G\mathbf{e}_1)^\top P_1 RG^\top G\mathbf{e}_2 = (P_2 RG^\top \mathbf{g}_1)^\top P_2 RG^\top \mathbf{g}_2 + (P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_1)^\top P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_2.$$

Note that $P_2 RG^\top$ is independent of $\mathbf{g}_1$ and $\mathbf{g}_2$. Similarly to (10), we have

$$\|P_2 R\|_{\mathrm{F}}^2 \geq \|R\|_{\mathrm{F}}^2 - 2\,\|R\|^2 \geq \|R\|_{\mathrm{F}}^2/2$$

and so, $\mathbf{sr}(P_2 R) \geq (1/2)\,\mathbf{sr}(R) \geq m$. This allows us to use Lemma 24 to estimate

$$\Pr\left[|\mathbf{g}_1^\top G(P_2 R)^\top P_2 RG^\top \mathbf{g}_2| \geq t\right] \leq 8\exp\left(-c\frac{t^2}{m\,\|P_2 R\|_{\mathrm{F}}^4}\right) \tag{8}$$

for any $t \in [0, m\,\|P_2 R\|_{\mathrm{F}}^2]$.

The estimate for the last term is straightforward as $P_1 R\mathbf{e}_2$ is deterministic. Since

$$\forall s \geq 0 \qquad \Pr\left[|\mathbf{g}_2^\top \mathbf{g}_1| > Cs\right] \leq 2\exp\left(-c\frac{s^2}{m}\right) + \exp(-m),$$

and

$$\Pr\left[|\mathbf{g}_2^\top \mathbf{g}_2| > Cm\right] \leq \exp(-m),$$

we obtain

$$\Pr\left[|(P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_1)^\top P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_2| \geq sm\left\|P_1 R\mathbf{e}_2\right\|^2\right] \leq 2\exp\left(-c\frac{s^2}{m}\right) + \exp(-m)$$

or

$$\Pr\left[|(P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_1)^\top P_1 R\mathbf{e}_2\mathbf{g}_2^\top \mathbf{g}_2| \geq t\right] \leq 2\exp\left(-c\frac{t^2}{m^3\left\|P_1 R\mathbf{e}_2\right\|^4}\right) + \exp(-m) \tag{9}$$

for all $t \geq 0$. Combining (8) and (9), we conclude that

$$\begin{aligned}\Pr\left[|(P_1 RG^\top G\mathbf{e}_1)^\top P_1 RG^\top G\mathbf{e}_2| > t\right] &\leq 2\exp\left(-c\frac{t^2}{m\left\|R\right\|_{\mathrm{F}}^4}\right) + 2\exp\left(-c\frac{t^2}{m^3\left\|P_1 R\mathbf{e}_2\right\|^4}\right) + \exp(-cm)\\ &\leq 4\exp\left(-c\frac{t^2}{m\left\|R\right\|_{\mathrm{F}}^4}\right) + \exp(-cm)\end{aligned}$$

for any $t \in [0, m\left\|P_2 R\right\|_{\mathrm{F}}^2]$. Here we used the inequality

$$m\left\|P_1 R\mathbf{e}_2\right\|^2 \leq m\left\|R\right\|^2 \leq \left\|R\right\|_{\mathrm{F}}^2,$$

where the last one follows from the assumption $m \leq \mathbf{sr}(R)$. Taking into account the result from (7), we see that

$$\Pr\left[|\mathbf{u}^\top P_1 RG^\top G\mathbf{e}_2| > \tau\right] \leq 2\exp\left(-c\frac{\tau^2}{\left\|R\right\|_{\mathrm{F}}^2}\right) + \exp(-cm),$$

for all $\tau \in [0, \frac{1}{8}\sqrt{m}\left\|R\right\|_{\mathrm{F}}]$. After taking the union bound, we show that

$$\Pr\left[\exists j \geq 2,\ |\mathbf{u}^\top P_1 RG^\top G\mathbf{e}_j| > \tau\right] \leq 2d\left(\exp\left(-c\frac{\tau^2}{\left\|R\right\|_{\mathrm{F}}^2}\right) + \exp(-cm)\right). \tag{10}$$

Recall (6). Setting $\tau = \frac{1}{8s}\sqrt{m}\left\|R\right\|_{\mathrm{F}}$ with $s \geq 1$, and using together (7) and (10), we conclude that

$$\Pr\left[\exists \mathbf{x} \in s\cdot\mathrm{conv}(\pm\mathbf{e}_2, \ldots, \pm\mathbf{e}_d),\ \left\|RG^\top G(\mathbf{e}_1 + \mathbf{x})\right\| \leq \frac{1}{8}\sqrt{m}\left\|R\right\|_{\mathrm{F}}\right] \leq 2d\exp\left(-c\frac{m}{s^2}\right),$$

as the second term in the right-hand side gets absorbed in the first one. The proof of the proposition is complete. ■

B.3. Finishing the Proof of Theorem 5: Net Argument

The next theorem is the main technical step in proving Theorem 5. Invoking this theorem with appropriate parameters (that we explain later in this section) gives the proof of Theorem 5. The proof of the following theorem is based on generating an orthogonal matrix to reduce the general case to the special case discussed in Proposition 25, and then employing an ε-net argument.

Theorem 26 (Restated Theorem 11) *Let X be a fixed $n \times d$ matrix satisfying,*

$$2 \le m \le \mathbf{sr}(X)/2.$$

Let $\Psi = (\Psi_{ij})$ be an $m \times d$ random matrix with independent entries such that $\mathbb{E}[\Psi_{ij}] = 0$, $\mathbb{E}[\Psi_{ij}^2] = 1$, and $\|\Psi_{ij}\|_{\psi_2}$ is bounded. Let $\beta \in (0,1)$, and let $k \in \mathbb{N}$. Then for any s such that

$$1 \le s \le \sqrt{\frac{cm}{k \log d + \log(2/\beta)}},$$

$$\Pr[\exists I \subset [d] \text{ with } |I| = k, \exists \mathbf{y} \in \mathbb{S}^{d-1} \text{ with } \operatorname{supp}(\mathbf{y}) \subseteq I, \exists \mathbf{x} \in s \cdot \operatorname{conv}(\pm \mathbf{e}_i,\ i \notin I), \left\| X\Psi^\top \Psi(\mathbf{y}+\mathbf{x}) \right\| \le \frac{1}{32}\sqrt{m}\,\|X\|_{\mathrm{F}}] \le \beta.$$

Note that the condition $s \ge 1$ in the formulation of the theorem implicitly sets a lower bound on β and an upper bound on k.

Proof Fix the set I with $|I| = k$. For instance, consider $I = [k] \subset [d]$. Fix also a point $\mathbf{y} \in \mathbb{S}^{k-1}$. Define the subspace $E \subset \mathbb{R}^d$ as

$$E = \operatorname{span}(\mathbf{y}, \mathbf{e}_j,\ j > k).$$

Note that the vectors $\mathbf{y}$ and $\mathbf{e}_j,\ j > k$ form an orthonormal basis of E. Let $P_E : \mathbb{R}^d \to E$ be matrix of the orthogonal projection onto E with respect to this basis and the standard basis in $\mathbb{R}^d$. Then $P_E^\top$ is the matrix of the embedding of E into $\mathbb{R}^d$.

Let $Q : \mathbb{R}^n \to \mathbb{R}^n$ be the orthogonal projection with $\operatorname{Ker}(Q) = XE^\perp$, where $E^\perp$ represents the orthogonal complement of E. Then for any $\mathbf{z} \in E$,

$$\left\| X\Psi^\top \Psi \mathbf{z} \right\| \ge \left\| QX\Psi^\top \Psi \mathbf{z} \right\|. \tag{11}$$

We can represent the restriction of the linear operator $QX\Psi^\top\Psi$ to E as the following composition of linear operators:

$$E \xrightarrow{P_E^\top} \mathbb{R}^d \xrightarrow{\Psi} \mathbb{R}^m \xrightarrow{\Psi^\top} \mathbb{R}^d \xrightarrow{P_E} E \xrightarrow{P_E^\top} \mathbb{R}^d \xrightarrow{X} \mathbb{R}^n \xrightarrow{Q} \mathbb{R}^n.$$

Since $\|\mathbf{y}\| = 1$ and $\operatorname{supp}(\mathbf{y}) \subseteq [k]$, the $m \times (d-k+1)$ matrix $G = \Psi P_E^\top$ in the basis $\{\mathbf{y}, \mathbf{e}_j,\ j > k\}$ has centered subgaussian entries of unit variance. Denote $R = QXP_E^\top$. Then by the interlacing

$$\|X\|_{\mathrm{F}}^2 \ge \|R\|_{\mathrm{F}}^2 \ge \|X\|_{\mathrm{F}}^2 - 2k\,\|X\|^2 \ge \frac{1}{2}\,\|X\|_{\mathrm{F}}^2,$$

since by the assumptions on k and X, $k \le m/8 \le \mathbf{sr}(X)/16$. This implies

$$\mathbf{sr}(R) \ge \frac{1}{2}\,\mathbf{sr}(X) \ge m.$$

Applying Proposition 25 to the matrices G, R, with $\mathbf{y}$ playing the role of $\mathbf{e}_1$, and taking into account (11), we obtain

$$\Pr\left[\exists \mathbf{x} \in s \cdot \mathrm{conv}(\pm \mathbf{e}_j \; j > k),\; \left\|X\Psi^\top \Psi(\mathbf{y}+\mathbf{x})\right\| \le \frac{1}{16}\sqrt{m}\,\|X\|_{\mathrm{F}}\right] \le 2d\exp\left(-c\frac{m}{s^2}\right)$$

for any $s \ge 1$.

In the rest of the proof, we employ the net argument. Since Ψ is a subgaussian random matrix,

$$\begin{aligned}\left\|X\Psi^\top\Psi\right\| &\le \left\|X\Psi^\top\right\| \cdot \|\Psi\| \le C'(\|X\|_{\mathrm{F}} + \sqrt{m}\,\|X\|) \cdot C''(\sqrt{d}+\sqrt{m}) \\ &\le C\sqrt{d}\,\|X\|_{\mathrm{F}}\end{aligned}$$

with probability at least $1-\exp(-m)$, where we used Corollary 19. Let $\varepsilon > 0$ be a number to be chosen later, and (by Proposition 15) let $\mathcal{N} \subset \mathbb{S}^{k-1}$ be an ε-net of cardinality

$$|\mathcal{N}| \le \left(\frac{3}{\varepsilon}\right)^k.$$

Assume that for any $\mathbf{y} \in \mathcal{N}$, and for any $\mathbf{x} \in s \cdot \mathrm{conv}(\pm\mathbf{e}_j \; j > k)$,

$$\left\|X\Psi^\top\Psi(\mathbf{y}+\mathbf{x})\right\| \ge \frac{1}{16}\sqrt{m}\,\|X\|_{\mathrm{F}}.$$

Assume also that $\left\|X\Psi^\top\Psi\right\| \le C\sqrt{d}\,\|X\|_{\mathrm{F}}$. Let $\mathbf{z} \in \mathbb{S}^{k-1}$, and chose $\mathbf{y} \in \mathcal{N}$ such that $\|\mathbf{z}-\mathbf{y}\| < \varepsilon$. Then setting $\varepsilon = c\sqrt{m/d}$ for an appropriately small constant $c > 0$, we obtain

$$\left\|X\Psi^\top\Psi(\mathbf{z}+\mathbf{x})\right\| \ge \left\|X\Psi^\top\Psi(\mathbf{y}+\mathbf{x})\right\| - \left\|X\Psi^\top\Psi\right\| \cdot \|\mathbf{z}-\mathbf{y}\| \ge \frac{1}{32}\sqrt{m}\,\|X\|_{\mathrm{F}}.$$

Thus,

$$\begin{aligned}&\Pr\left[\exists \mathbf{y} \in \mathbb{S}^{k-1},\; \exists \mathbf{x} \in s \cdot \mathrm{conv}(\pm\mathbf{e}_i,\; i > k),\; \left\|X\Psi^\top\Psi(\mathbf{y}+\mathbf{x})\right\| \le \frac{1}{32}\sqrt{m}\,\|\Psi\|_{\mathrm{F}}\right] \\ &\le |\mathcal{N}| \cdot 2d\exp\left(-c\frac{m}{s^2}\right) + \exp(-m) \\ &\le 2\exp\left(-c\frac{m}{s^2} + k\log\left(\frac{C\sqrt{d}}{\sqrt{m}}\right)\right).\end{aligned}$$

It remains to take the union bound over all possible supports of $\mathbf{y}$. It yields,

$$\begin{aligned}&\Pr\Big[\exists I \subset [d] \text{ with } |I| = k, \exists \mathbf{y} \in \mathbb{S}^{d-1} \text{ with } \mathrm{supp}(\mathbf{y}) \subseteq I, \exists \mathbf{x} \in s \cdot \mathrm{conv}(\pm\mathbf{e}_i,\; i \notin I), \\ &\qquad \left\|X\Psi^\top\Psi(\mathbf{y}+\mathbf{x})\right\| \le \frac{1}{32}\sqrt{m}\,\|\Psi\|_{\mathrm{F}}\Big] \\ &\le \binom{d}{k} \cdot 2\exp\left(-c\frac{m}{s^2} + k\log\left(\frac{C\sqrt{d}}{\sqrt{m}}\right)\right) \\ &\le 2\exp\left(-c\frac{m}{s^2} + \frac{k}{2}\log\left(\frac{Cd^2}{mk}\right)\right).\end{aligned}$$

The last quantity is smaller than β provided that[10]

$$1 \leq s \leq \sqrt{\frac{cm}{k \log d + \log(2/\beta)}}.$$

This completes the proof of the theorem. ■

Using the above result, we can finish the proof of Theorem 5 as shown in Section 2.

Appendix C. Missing Details from Section 3

We provide missing details from Section 3. For brevity, we will say that the event which holds with probability at least $1 - O(m^{-K})$ occurs with a *large* probability.

Proposition 27 (Restated Proposition 12) *Let X be a deterministic matrix and Φ be a random matrix satisfying the conditions of Theorem 5. Consider the linear model $\mathbf{y} = X\Phi^\top \Phi\theta^\star + \mathbf{w}$ where the entries of the noise vector $\mathbf{w} = (w_1, \ldots, w_n)$ are independent centered subgaussians with $\|w_i\|_{\psi_2} \leq \sigma$. Let $K > 0$ be any constant, and let $dm^{-K} \leq \beta < 1$. Then $\theta^{\text{comp}} \in \operatorname{argmin}_{\theta \in \mathbb{R}^d_2} \|\mathbf{y} - X\Phi^\top \Phi\theta\|^2/n + \lambda\|\theta_1\|$ with $\lambda = \Theta(\sigma \|X\|_F / n\sqrt{m})$, satisfies with probability at least $1 - \beta$:*

$$\|\theta^{\text{comp}} - \theta^\star\| = O\left(\frac{\sigma\sqrt{m}k^{3/2}}{\|X\|_F}\right).$$

Proof We use the framework of Theorem 2 to bound $\|\theta^{\text{comp}} - \theta^\star\|$. The matrix of interest is now $X\Phi^\top\Phi$. Our first aim will be to bound $\left\|\Phi^\top\Phi X^\top \mathbf{w}\right\|_\infty$ that is used to set λ in Theorem 2. Take any $\mathbf{u} \in S^{d-1}$. Then conditioning on Φ, with a large probability,

$$|\mathbf{u}^\top\Phi^\top\Phi X^\top\mathbf{w}| \leq C\sigma \left\|\mathbf{u}^\top\Phi^\top\Phi X^\top\right\|_2.$$

Applying the Hanson-Wright inequality, we can show that with a large probability with respect to Φ,

$$\left\|\mathbf{u}^\top\Phi^\top\Phi X^\top\right\|_2 \leq C\frac{\|X\|_F}{\sqrt{m}}.$$

The estimate for $\left\|\Phi^\top\Phi X^\top\mathbf{w}\right\|_\infty$ follows by combining two previous inequalities and using the union bound for $\mathbf{u} = \mathbf{e}_j,\ j \in [d]$ as before. We get that probability at least $1 - O(dm^{-K})$ ($\geq 1 - O(\beta)$),

$$\left\|\Phi^\top\Phi X^\top\mathbf{w}\right\|_\infty = O\left(\sigma\frac{\|X\|_F}{\sqrt{m}}\right).$$

Plugging this bound into Theorem 2 along with the RE bound from Corollary 7 gives the claimed result. ■

10. Here we ignored smaller order terms assuming $d^2 \gg mk$. If this does not hold, one can obtain a slightly better estimate.

C.1. Missing Details from Section 3.2

In this section, we provide missing details from Section 3.2 where we considered sparse linear regression with compressed features. Following the same notation. We start with the following inequality:

$$\frac{1}{n}\|\mathbf{y} - Z\theta^{\text{comp}}\|^2 + \lambda\|\theta^{\text{comp}}\|_1 \leq \frac{1}{n}\|\mathbf{y} - Z\theta^\star\|^2 + \lambda\|\theta^\star\|_1.$$

Rearranging this gives

$$\frac{1}{n}(\|Z\theta^{\text{comp}}\|^2 - \|Z\theta^\star\|^2) \leq \frac{2}{n}\mathbf{y}^\top(Z\theta^{\text{comp}} - Z\theta^\star) + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1)$$

and plugging in $\tilde{\mathbf{w}} = \mathbf{y} - Z\theta^\star$ (i.e., $\tilde{\mathbf{w}} = (X - Z)\theta^\star + \mathbf{w}$),

$$\frac{1}{n}\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 \leq \frac{2}{n}\tilde{\mathbf{w}}^\top(Z\theta^{\text{comp}} - Z\theta^\star) + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1).$$

Rearranging the terms, we get,

$$\frac{1}{n}\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 \leq \frac{2}{n}\langle Z^\top\tilde{\mathbf{w}}, \theta^{\text{comp}} - \theta^\star\rangle + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1).$$

Adding and subtracting $2\langle\mathbb{E}[Z^\top\tilde{\mathbf{w}}], \theta^{\text{comp}} - \theta^\star\rangle/n$ on the right-hand side gives,

$$\frac{1}{n}\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 \leq \frac{2}{n}\langle Z^\top\tilde{\mathbf{w}} - \mathbb{E}[Z^\top\tilde{\mathbf{w}}], \theta^{\text{comp}} - \theta^\star\rangle + \frac{2}{n}\langle\mathbb{E}[Z^\top\tilde{\mathbf{w}}], \theta^{\text{comp}} - \theta^\star\rangle + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1).$$

By applying Hölder's inequality,

$$\frac{1}{n}\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 \leq \frac{2}{n}\|Z^\top\tilde{\mathbf{w}} - \mathbb{E}[Z^\top\tilde{\mathbf{w}}]\|_\infty\|\theta^{\text{comp}} - \theta^\star\|_1 + \frac{2}{n}\|\,\mathbb{E}[Z^\top\tilde{\mathbf{w}}]\|\|\theta^\star - \theta^{\text{comp}}\| + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1). \tag{12}$$

The following lemma establishes a bound on $\|Z^\top\tilde{\mathbf{w}} - \mathbb{E}[Z^\top\tilde{\mathbf{w}}]\|_\infty$. For simplicity, we focus on Gaussian random matrices Φ, the extension of the lemma to a more general class of subgaussian random matrices is possible, but omitted here. Also w.l.o.g. we assume that $\theta^\star \in S^{d-1}$. The proof involves careful analysis of the projections of $Z^\top(X - Z)\theta^\star$ onto $\theta^\star$ and a vector in its orthogonal direction.

Lemma 28 *Let X be an $n \times d$ matrix. Let Ψ be an $m \times d$ standard Gaussian matrix with independent entries, and let $\Phi = \Psi/\sqrt{m}$. Let $Z = X\Phi^\top\Phi$. Let $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ be a vector with independent centered coordinates having $\|w_j\|_{\psi_2} \leq \sigma$. Let $\theta^\star \in S^{d-1}$.*

$$\tilde{\mathbf{w}} = (X - Z)\theta^\star + \mathbf{w}.$$

Assume that $\mathbf{sr}(X) \geq m$. Then

$$\mathbb{E}[Z^\top\tilde{\mathbf{w}}] = -\frac{\|X\|_F^2}{m}\theta^\star \cdot (1 + \delta),$$

where δ depends on $\theta^\star$ and $|\delta| \leq 1/m$. and for any $K > 0$, with probability at least $1 - O(dm^{-K})$,

$$\left\|Z^\top\tilde{\mathbf{w}} - \mathbb{E}[Z^\top\tilde{\mathbf{w}}]\right\|_\infty \leq C\log m\left(\frac{\|X\|_F^2}{m^{3/2}} + \sigma\frac{\|X\|_F}{\sqrt{m}}\right)$$

with a constant C depending on K.

Proof Assume first that $\theta^\star = \mathbf{e}_1$. We will remove this assumption later. Set

$$\mathbf{z} := \Phi^\top \Phi X^\top (X - X\Phi^\top \Phi)\mathbf{e}_1.$$

To estimate $\mathbf{z}$ we consider projections of $\mathbf{z}$ on $\mathbf{e}_1$ and on a vector $\mathbf{v} \in S^{d-1}$ orthogonal to $\mathbf{e}_1$ separately. For brevity, we will say that the event which holds with probability at least $1 - O(m^{-K})$ occurs with a *large* probability.

Step 1 *We will show that*

$$\mathbb{E}[\mathbf{e}_1^\top \mathbf{z}] = -\frac{\|X\|_F^2}{m} \cdot (1+\delta), \tag{13}$$

where $|\delta| \le 1/m$ *and with a large probability,*

$$|\mathbf{e}_1^\top \mathbf{z} - \mathbb{E}[\mathbf{e}_1^\top \mathbf{z}]| \le C \log m \frac{\|X\|_F^2}{m^{3/2}}.$$

We will start with estimating $\mathbf{e}_1^\top \mathbf{z}$. To this end, denote $\Psi = (\upsilon_1\ G)$ and $X = (\mathbf{f}_1\ Y)$ separating the first column in each matrix. Then

$$\begin{aligned}\mathbf{e}_1^\top \mathbf{z} &= \frac{1}{m}\upsilon_1^\top (\upsilon_1\ G) X^\top \mathbf{f}_1 \cdot \left(1 - \frac{1}{m}\upsilon_1^\top \upsilon_1\right) - \frac{1}{m^2}\upsilon_1^\top (\upsilon_1\ G) X^\top Y G^\top \upsilon_1 \\ &=: A + B.\end{aligned}$$

Let us estimate A first. By Gaussian concentration, with a large probability, for a constant (independent of the parameters) C,

$$|1 - \frac{1}{m}\upsilon_1^\top \upsilon_1| \le C\frac{1}{\sqrt{m}}.$$

Hence,

$$\left|\frac{1}{m}(\upsilon_1^\top \upsilon_1)\mathbf{f}_1^\top \mathbf{f}_1 \left(\frac{1}{m} - \upsilon_1^\top \upsilon_1\right)\right| \le C\frac{1}{\sqrt{m}}\|\mathbf{f}_1\|_2^2 \le C\frac{1}{m^{3/2}}\|X\|_F^2,$$

where we used $\mathbf{sr}(X) \ge m$ in the last inequality. Also, conditioning on G, we also have that with large probability

$$\left|\frac{1}{m}\upsilon_1^\top G Y^\top \mathbf{f}_1 \left(1 - \frac{1}{m}\upsilon_1^\top \upsilon_1\right)\right| \le C m^{-3/2}\sqrt{\log m}\left\|GY^\top \mathbf{f}_1\right\|_2.$$

By Corollary 21,

$$\Pr\left[\left\|GY^\top \mathbf{f}_1\right\|_2 \ge C\sqrt{m}\left\|Y^\top \mathbf{f}_1\right\|_2\right] \le \exp(-m),$$

so with a large probability,

$$\left|\frac{1}{m}\upsilon_1^\top G Y^\top \mathbf{f}_1 \cdot \left(1 - \frac{1}{m}\upsilon_1^\top \upsilon_1\right)\right| \le C m^{-1}\sqrt{\log m}\left\|Y^\top \mathbf{f}_1\right\|_2 \le C m^{-3/2}\sqrt{\log m}\,\|X\|_F^2.$$

Summarizing, we proved that with a large probability,

$$|A| \le C m^{-3/2}\sqrt{\log m}\,\|X\|_F^2.$$

We move to estimating B. Denote

$$B = -\frac{1}{m^2}(v_1^\top v_1)\mathbf{f}_1^\top Y G^\top v_1 - \frac{1}{m^2} v_1^\top G Y^\top Y G^\top v_1 =: B_1 + B_2.$$

Then $\mathbb{E}[B_1] = 0$ and

$$\begin{aligned}\mathbb{E}[B_2] &= -\frac{1}{m^2}\,\mathbb{E}[v_1^\top G Y^\top Y G^\top v_1] = -\frac{1}{m^2}\,\mathrm{tr}(\mathbb{E}[G Y^\top Y G^\top]) \\ &= -\frac{1}{m}\,\mathrm{tr}(Y^\top Y) = -\frac{\|Y\|_F^2}{m}.\end{aligned}$$

Note that by assumption on $\mathbf{sr}(X)$,

$$\|X\|_F^2 - \|Y\|_F^2 = \|\mathbf{f}_1\|_2^2 \le \frac{1}{m}\,\|X\|_F^2\,,$$

which yields (13) with the required bound on δ.

Now, let us bound the deviation of B from its expectation. Arguing as above, we conclude that with a large probability,

$$|B_1| \le Cm^{-1/2}\sqrt{\log m}\,\left\|\mathbf{f}_1^\top Y\right\|_2 \le Cm^{-3/2}\sqrt{\log m}\,\|X\|_F^2\,.$$

Also, conditioning on G, with a large probability,

$$\left|v_1^\top G Y^\top Y G^\top v_1 - \mathbb{E}[v_1^\top G Y^\top Y G^\top v_1 \mid G]\right| \le C\sqrt{\log m}\,\left\|G Y^\top Y G^\top\right\|_F\,. \tag{14}$$

Using the Hanson-Wright inequality as in Corollary 22, we conclude that with a large probability,

$$\left\|G Y^\top Y G^\top\right\|_F \le Cm\sqrt{\log m}\,\left\|Y^\top Y\right\|_F + C\sqrt{m}\,\mathrm{tr}(Y^\top Y) \le C\sqrt{m}\sqrt{\log m}\,\|X\|_F^2 \tag{15}$$

since $m\left\|Y^\top Y\right\|_F \le m\,\|Y\|\cdot\|Y\|_F \le \sqrt{m}\,\|X\|_F^2 \le \sqrt{m}\,\|X\|_F^2$. The measure concentration with respect to the Gaussian matrix G yields that with a large probability

$$\left|\mathbb{E}[v_1^\top G Y^\top Y G^\top v_1 \mid G] - \mathbb{E}[v_1^\top G Y^\top Y G^\top v_1]\right| \le Cm\left\|Y^\top Y\right\|_F \le C\sqrt{m}\,\|X\|_F^2\,. \tag{16}$$

Combining (14), (15), and (16) shows that with a large probability,

$$|B_2 - \mathbb{E}[B_2]| \le Cm^{-3/2}\log m\,\|X\|_F^2\,.$$

This together with the bounds on A and B_1 obtained above completes the proof of Step 1.

Step 2 *Let* $\mathbf{v} \in S^{d-1}$, $\mathbf{v}^\top \mathbf{e}_1 = 0$. *Then* $\mathbb{E}[\mathbf{v}^\top \mathbf{z}] = 0$ *and with a large probability,*

$$|\mathbf{v}^\top \mathbf{z}| \le Cm^{-3/2}\log m\,\|X\|_F^2\,.$$

The equality $\mathbb{E}[\mathbf{v}^\top \mathbf{z}] = 0$ follows from independence of $\Phi\mathbf{e}_1$ and $\Phi\mathbf{v}$ (recall that we assumed that the matrix Ψ is Gaussian). To prove the concentration, we can use rotation invariance of the Gaussian distribution. More precisely, the matrix Φ is distributed like ΦU, where U is an orthogonal matrix such that $U\mathbf{e}_1 = \mathbf{e}_1$ and $U\mathbf{v} = \mathbf{e}_2$. Note that replacing X by $XU^\top$ does not change the Hilbert-Schmidt norm. Using these observations, we can reduce the case of a general $\mathbf{v} \perp \mathbf{e}_1$ to $\mathbf{v} = \mathbf{e}_2$.

In the last case, we separate the first two columns of the matrices Ψ and X as we did in Step 1:

$$\Psi = (\upsilon_1\ \upsilon_2\ \Lambda) \quad \text{and} \quad X = (\mathbf{f}_1\ \mathbf{f}_2\ P).$$

Then the inner product $\mathbf{e}_2^\top \mathbf{z}$ can be decomposed into a sum of 9 terms containing different combinations of independent random variables υ_1, υ_2, and Λ. The absolute value of each of these terms does not exceed $Cm^{-3/2}\log m\,\|X\|_F^2$ with a large probability. These estimates closely follow the argument of Step 1, so we omit the details. Combining these nine estimates completes the proof of Step 2.

Let us summarize what we proved. We have shown that in the case $\theta^\star = \mathbf{e}_1$,

$$\mathbb{E}[\mathbf{z}] = -\frac{\|X\|_F^2}{m} \cdot (1+\delta)\mathbf{e}_1 = -\frac{\|X\|_F^2}{m} \cdot (1+\delta)\theta^\star$$

and for any $\mathbf{u} \in S^{d-1}$,

$$|\mathbf{u}^\top(\mathbf{z} - \mathbb{E}[\mathbf{z}])| \le Cm^{-3/2}\log m\,\|X\|_F^2\,.$$

The last inequality follows by decomposing $\mathbf{u}$ into its projection on $\mathbf{e}_1$ and the orthogonal component and applying Steps 1 and 2 respectively to these components.

Now, we can use the invariance of the Gaussian matrix under multiplication by an orthogonal one to remove the assumption that $\theta^\star = \mathbf{e}_1$ in two last inequalities. To derive the bound for $\|\mathbf{z} - \mathbb{E}[z]\|_\infty$, we apply the last inequality with $\mathbf{u} = \mathbf{e}_j$ and take the union bound over $j \in [d]$.

Finally, it remains to handle the term $\left\|\Phi^\top \Phi X^\top \mathbf{w}\right\|_\infty$ which we do as in Proposition 12. As before, take any $\mathbf{u} \in S^{d-1}$. Then conditioning on Φ, with a large probability,

$$|\mathbf{u}^\top \Phi^\top \Phi X^\top \mathbf{w}| \le C\sigma \left\|\mathbf{u}^\top \Phi^\top \Phi X^\top\right\|_2.$$

Applying the Hanson-Wright inequality, we show that with a large probability with respect to Φ,

$$\left\|\mathbf{u}^\top \Phi^\top \Phi X^\top\right\|_2 \le C\frac{\|X\|_F}{\sqrt{m}}.$$

The estimate for $\left\|\Phi^\top \Phi X^\top \mathbf{w}\right\|_\infty$ follows by combining two previous inequalities and using the union bound for $\mathbf{u} = \mathbf{e}_j,\ j \in [d]$ as before.

This completes the proof of the lemma. ■

Remark 29 *The same argument shows that for any $K > 0$, with probability at least $1 - O(d^{-K})$,*

$$\left\|\Phi^\top \Phi X^\top \tilde{\mathbf{w}} - \mathbb{E}[\Phi^\top \Phi X^\top \tilde{\mathbf{w}}]\right\|_\infty \le C\log d\left(\frac{\|X\|_F^2}{m^{3/2}} + \sigma\frac{\|X\|_F}{\sqrt{m}}\right)$$

with a constant C depending on K.

The following proposition follows by using Lemma 28 in (12) and by setting λ appropriately (as guided by Theorem 2) to ensure that $\theta^{\text{comp}} - \theta^\star \in \mathbb{C}(S_\star)$ where $S_\star = \text{supp}(\theta^\star)$. The RE bound lets us relate $\|Z\theta^{\text{comp}} - Z\theta^\star\|$ and $\|X\theta^{\text{comp}} - X\theta^\star\|$.

Proposition 30 (Restated Proposition 13) *Let X be a deterministic matrix and Φ be a Gaussian random matrix satisfying the conditions of Theorem 5. Consider the linear model $\mathbf{y} = X\theta^\star + \mathbf{w}$ where the entries of the noise vector $\mathbf{w} = (w_1, \ldots, w_n)$ are independent centered subgaussians with $\|w_i\|_{\psi_2} \le \sigma$ and $\theta^\star \in S^{d-1}$. Let $K > 0$ be any constant, and let $dm^{-K} \le \beta < 1$. Then $\theta^{\text{comp}} \in \text{argmin}_{\theta \in B_2^d} \|\mathbf{y} - X\Phi^\top \Phi\theta\|^2/n + \lambda\|\theta_1\|$ with $\lambda = \Theta(\sigma\|X\|_{\text{F}} \log m/n\sqrt{m} + \|X\|_{\text{F}}^2/nm)$, satisfies with probability at least $1 - \beta$:*

$$\frac{1}{n}\|X\theta^{\text{comp}} - X\theta^\star\|^2 = O\left(\frac{\sigma\|X\|_{\text{F}} k^{3/2} \log m}{n\sqrt{m}} + \frac{\|X\|_{\text{F}}^2 k^{3/2}}{nm}\right).$$

Proof We start by restating (12).

$$\frac{1}{n}\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 \le \frac{2}{n}\|Z^\top \tilde{\mathbf{w}} - \mathbb{E}[Z^\top \tilde{\mathbf{w}}]\|_\infty \|\theta^{\text{comp}} - \theta^\star\|_1 + \frac{2}{n}\|\mathbb{E}[Z^\top \tilde{\mathbf{w}}]\| \|\theta^\star - \theta^{\text{comp}}\| + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1).$$

Applying Lemma 28 in the above equation, gives that with probability at least $1 - O(dm^{-K})$,

$$\frac{1}{n}\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 \le \frac{C \log m}{n}\left(\frac{\|X\|_F^2}{m^{3/2}} + \sigma\frac{\|X\|_F}{\sqrt{m}}\right)\|\theta^{\text{comp}} - \theta^\star\|_1 + \frac{4\|X\|_F^2}{nm}\|\theta^{\text{comp}} - \theta^\star\|_1 + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1). \quad (17)$$

For the remainder of this proof, we condition on (17) holding true. Let $S_\star = \text{supp}(\theta^\star)$. We first argue that $\hat{\theta} := \theta^{\text{comp}} - \theta^\star$ is such that $\hat{\theta} \in \mathbb{C}(S_\star)$. We start by observing that:

$$\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1 = \|\theta^\star\|_1 - \|\theta^\star + \hat{\theta}\|_1 = \|\theta^\star\|_1 - \|\theta^\star_{S_\star} + \hat{\theta}_{S_\star}\|_1 - \|\hat{\theta}_{S_\star^{\text{co}}}\|_1 \le \|\hat{\theta}_{S_\star}\|_1 - \|\hat{\theta}_{S_\star^{\text{co}}}\|_1. \quad (18)$$

Set

$$\lambda \ge \frac{2C \log m}{n}\left(\frac{\|X\|_F^2}{m^{3/2}} + \sigma\frac{\|X\|_F}{\sqrt{m}}\right) + \frac{8\|X\|_F^2}{nm}.$$

Using this value of λ, we can observe that,

$$\frac{1}{n}\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 \le \frac{\lambda}{2}\|\hat{\theta}\|_1 + \lambda(\|\theta^\star\|_1 - \|\theta^{\text{comp}}\|_1).$$

From (18) and by noting $\|Z\theta^{\text{comp}} - Z\theta^\star\|^2 > 0$,

$$0 \le \frac{\lambda}{2}\|\hat{\theta}\|_1 + \lambda(\|\hat{\theta}_{S_\star}\|_1 - \|\hat{\theta}_{S_\star^{\text{co}}}\|_1),$$

implying $\|\hat{\theta}_{S_\star^{\text{co}}}\|_1 \le 3\|\hat{\theta}_{S_\star}\|_1$, i.e., $\hat{\theta} \in \mathbb{C}(S_\star)$. We can now simplify (17) as,

$$\frac{1}{n}\|Z\hat{\theta}\|^2 = O\left(\frac{\|X\|_{\text{F}}^2 \log m}{nm^{3/2}} + \frac{\sigma\|X\|_{\text{F}} \log m}{n\sqrt{m}} + \frac{\|X\|_{\text{F}}^2}{nm}\right)\|\hat{\theta}\|_1.$$

Now,

$$\|\hat{\theta}\|_1 = \|\hat{\theta}_{S_\star}\|_1 + \|\hat{\theta}_{S_\star^{\text{co}}}\|_1 \leq \hat{\theta}_{S_\star}\|_1 + 3\|\hat{\theta}_{S_\star}\|_1 \leq 4\sqrt{k}\|\hat{\theta}_{S_\star}\| \leq 4\sqrt{k}\|\hat{\theta}\| \leq 8\sqrt{k},$$

as $\|\hat{\theta}\| = \|\theta^{\text{comp}} - \theta^\star\| \leq \|\theta^{\text{comp}}\| + \|\theta^\star\| \leq 2$.[11] Plugging this in the above inequality,

$$\begin{aligned}\frac{1}{n}\|Z\hat{\theta}\|^2 &= O\left(\frac{\|X\|_{\text{F}}^2\sqrt{k}\log m}{nm^{3/2}} + \frac{\sigma\|X\|_{\text{F}}\sqrt{k}\log m}{n\sqrt{m}} + \frac{\|X\|_{\text{F}}^2\sqrt{k}}{nm}\right) \\ &= O\left(\frac{\sigma\|X\|_{\text{F}}\sqrt{k}\log m}{n\sqrt{m}} + \frac{\|X\|_{\text{F}}^2\sqrt{k}}{nm}\right).\end{aligned}$$

Now by our stable rank assumption on X, $\|X\hat{\theta}\| = O((\|X\|_{\text{F}}/\sqrt{m})\|\hat{\theta}\|)$. Under the conditions of Corollary 7, with probability at least $1-\beta$, $\|Z\hat{\theta}\|^2 = \Omega((\|X\|_{\text{F}}^2/mk)\|\hat{\theta}\|^2)$. Putting these two together gives that, $\|X\hat{\theta}\|^2 = O(k\|Z\hat{\theta}\|^2)$.

Using this in the above bound on $\|Z\hat{\theta}\|^2 = \|Z(\theta^{\text{comp}} - \theta^\star)\|^2$ gives that with probability at least $1-\beta$, under the conditioning on (17):

$$\frac{1}{n}\|X\theta^{\text{comp}} - X\theta^\star\|^2 = O\left(\frac{\sigma\|X\|_{\text{F}}k^{3/2}\log m}{n\sqrt{m}} + \frac{\|X\|_{\text{F}}^2k^{3/2}}{nm}\right).$$

Finally, we can remove the conditioning on (17). To simplify the result, assume $\beta > dm^{-K}$. ■

Appendix D. Stable Rank vs. Restricted Eigenvalue Condition

In this section, we investigate how stable rank relates to the restricted eigenvalue (RE) condition that is commonly used in the analysis of Lasso. The picture that emerges is the following: stable rank is a *less restrictive* condition to impose on a design matrix than RE. We show this by establishing that a RE bound on a matrix implies a non-trivial[12] stable rank for that matrix, whereas other direction does not always hold.

We first look at the case, when we have a stable rank condition on X. The RE condition (and of course, RIP) governs the behavior of the matrix on *all* coordinate subspaces of a small dimension. In this sense, a bound on the stable rank on X is much more relaxed. We now provide a simple pedagogical example to illustrate this fact. We rely on the fact that if $X\mathbf{e}_j = 0$ for even one $j \in d$, then no RE condition holds. Consider, for example the $d \times n$ matrix

$$X = \begin{pmatrix}\mathbb{I}_{2m} & 0 \\ 0 & 0\end{pmatrix},$$

where $\mathbb{I}_{2m}$ is the identity $2m \times 2m$ matrix. Then, $\mathbf{sr}(X) = 2m$, while the RE condition does not hold for X. This simple example illustrates that there exist families of matrices for which a stable rank condition (as required in Theorem 5) holds, but a RE condition is not satisfied.

11. Since $\theta^\star \in S^{d-1}$, it suffices to define $\theta^{\text{comp}} \in \text{argmin}_{\theta \in B_2^d} \|\mathbf{y} - Z\theta\|^2/n + \lambda\|\theta_1\|$, implying that $\|\theta^{\text{comp}}\| \leq 1$.

12. A direct numerical extension is not possible as stable rank is invariant to matrix scaling, whereas RE is not.

To make the comparison in the other direction, we need an additional normalization of X, as $\mathbf{sr}(X)$ is invariant under scaling, and $\mathrm{RE}(X,k,\alpha)$ is degree 1 homogenous (in that scaling each element in X by a factor c changes $\mathrm{RE}(X,k,\alpha)$ by c). Assume that $\mathrm{RE}(X,k,\alpha) \geq r$ and define

$$\|X\|_{(k)} = \max_{\substack{J\subset[d]\\|J|=k}} \|X_J\| \leq R.$$

An upper bound on $\|X\|_{(k)}$ is usually applied together with a lower bound on $\mathrm{RE}(X,k,\alpha) \geq r$ in derivation of the vector reconstruction conditions (see, e.g. (Rudelson and Zhou, 2013)). These assumptions yield that

$$\|X\|_{\mathrm{F}} = \left(\sum_{j=1}^{d} \|X\mathbf{e}_j\|^2\right)^{1/2} \geq r\sqrt{d}.$$

Also, assume for simplicity that $d = kL$ and decompose $[d] = \bigcup_{l=1}^{L} J_l$, where $J_l \subset [d]$ are consecutive sets of k coordinates. Let $\mathbf{y} \in \mathbb{S}^{d-1}$. Then

$$\|X\mathbf{y}\| \leq \sum_{l=1}^{L} \|X_{J_l}\| \cdot \|\mathbf{y}_{J_l}\| \leq \left(\sum_{l=1}^{L} \|X_{J_l}\|^2\right)^{1/2} \left(\sum_{l=1}^{L} \|\mathbf{y}_{J_l}\|^2\right)^{1/2} \leq R\sqrt{L} = R\sqrt{\frac{d}{k}}.$$

Therefore, $\|X\| \leq R\sqrt{\frac{d}{k}}$ and so

$$\mathbf{sr}(X) \geq \left(\frac{r}{R}\right)^2 k.$$

This shows that a RE bound on X implies a non-trivial stable rank bound on X.

Putting both these directions together implies that while a RE bound always translates into stable rank bound, the other direction does not always hold.