

# Compressed Sparse Linear Regression

Shiva Kasiviswanathan\*

Mark Rudelson†

## Abstract

High-dimensional sparse linear regression is a basic problem in machine learning and statistics. Consider a linear model  $\mathbf{y} = X\theta^* + \mathbf{w}$ , where  $\mathbf{y} \in \mathbb{R}^n$  is the vector of observations,  $X \in \mathbb{R}^{n \times d}$  is the covariate matrix with  $i$ th row representing the covariates for the  $i$ th observation, and  $\mathbf{w} \in \mathbb{R}^n$  is an unknown noise vector. In many applications, the linear regression model is high-dimensional in nature, meaning that the number of observations  $n$  may be substantially smaller than the number of covariates  $d$ . In these cases, it is common to assume that  $\theta^*$  is sparse, and the goal in sparse linear regression is to estimate this sparse  $\theta^*$ , given  $(X, \mathbf{y})$ .

In this paper, we study a variant of the traditional sparse linear regression problem where each of the  $n$  covariate vectors in  $\mathbb{R}^d$  are individually projected by a random linear transformation to  $\mathbb{R}^m$  with  $m \ll d$ . Such transformations are commonly applied in practice for computational savings in resources such as storage space, transmission bandwidth, and processing time. Our main result shows that one can estimate  $\theta^*$  with a low  $\ell_2$ -error, even with access to only these projected covariate vectors, under some mild assumptions on the problem instance. Our approach is based on solving a variant of the popular Lasso optimization problem. While the conditions (such as the restricted eigenvalue condition on  $X$ ) for success of a Lasso formulation in estimating  $\theta^*$  are well-understood, we investigate conditions under which this variant of Lasso estimates  $\theta^*$ . The main technical ingredient of our result, a bound on the restricted eigenvalue on certain projections of a deterministic matrix satisfying a stable rank condition, could be of interest beyond sparse regression.

---

\*Amazon Machine Learning, Palo Alto, CA, USA. [kasivisw@gmail.com](mailto:kasivisw@gmail.com). Part of the work done while the author was at Samsung Research America.

†University of Michigan, Ann Arbor, MI, USA. [rudelson@umich.edu](mailto:rudelson@umich.edu). Partially supported by NSF grant, DMS-1464514.

# 1 Introduction

Problems in high-dimensional statistical inference have attracted a great deal of attention in recent years. Many fields in modern science and engineering such as computational biology, medical imaging, and natural language processing regularly involve collecting datasets in which the dimension of the data exceeds the sample size. In this paper, we consider a prototypical problem in high-dimensional statistics, *sparse linear regression*.

Consider a linear model:  $\mathbf{y} = X\theta^* + \mathbf{w}$ , where  $\mathbf{y} = (y_1, \dots, y_n)$  is the vector of responses,  $X \in \mathbb{R}^{n \times d}$  is the covariate matrix (in which  $i$ th row  $\mathbf{x}_i^\top$  represents the covariates (features) for the  $i$ th observation), and  $\mathbf{w}$  is an unknown  $n$ -dimensional noise vector. The goal of linear regression, given  $(X, \mathbf{y})$ , is to estimate the vector  $\theta^*$ , known as the regression vector. If the linear regression model is high-dimensional, which means that the number of observations  $n$  is substantially smaller than the number of covariates  $d$ , the model is *unidentifiable* and it is not meaningful to estimate  $\theta^* \in \mathbb{R}^d$ . However, many machine learning and statistics applications, exhibit special structure that can lead to an identifiable model. In particular, in many settings, the vector  $\theta^*$  is sparse, which leads to a sparse linear regression problem. Given such a problem, the most direct approach would be to seek an exact sparse minimizer of the least-squares cost,  $\|\mathbf{y} - X\theta\|^2$ , thereby obtaining an  $\ell_0$ -based estimator. However, since this problem is non-convex, a standard approach is to replace the  $\ell_0$ -constraint with its  $\ell_1$ -norm, in either a constrained or penalized form, which leads to the ‘‘Lasso’’ (least absolute shrinkage and selection operator) formulation [18]. A detailed background on sparse linear regression is presented in Appendix A.

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. Random projection techniques, such as the Johnson-Lindenstrauss transform, are attractive for machine learning applications for several reasons: (i) they lead to substantial reduction in resources such as computation time, storage space, and transmission bandwidth, (ii) they are *oblivious* to the data set, meaning that the method does not require any prior knowledge of the data set as input, (iii) in a distributed data setting, they can be carried out *locally* by each party, independent of others, (iv) they are easy to implement and computationally inexpensive, and (v) they come with rigorous theoretical guarantees.

In this paper, we initiate the study of sparse linear regression in the compressed feature setting. A celebrated result in sparse linear regression is that, under a variety of mild assumptions on the instance, the  $\ell_2$ -error of a Lasso estimate decays roughly at the rate  $\sqrt{k \log d/n}$ , where  $k$  is the sparsity level of  $\theta^*$  [21, 1, 11]. We ask: *can we achieve a small  $\ell_2$ -error bound, under some mild assumptions, when we have access to only to the compressed representation of the data?* In this paper, we answer this question in affirmative by establishing both the sufficient conditions and the corresponding achievable error bound in this setting.

**Our Model.** Compressed sampling has been studied in the context of machine learning applications from two points of view. One idea is to use random projections to compress the dataset by combining input vectors using random projections [17, 23, 24]. This does not reduce the dimensionality of the data but rather generates a set of fewer datapoints (reduces  $n$ ). Another idea is to project each input vector into a lower dimensional space (thereby reducing  $d$ ), and then perform the learning with those compressed features. In the context of sparse linear regression, this would mean to estimate  $\theta^*$  given  $(\Phi \mathbf{x}_1, y_1), \dots, (\Phi \mathbf{x}_n, y_n)$ , where  $\Phi \in \mathbb{R}^{m \times d}$  is a random projection matrix with  $m \ll d$ . For sparse linear regression (when  $d \gg n$ ), this form of feature compression has multiple advantages over compressing the number of observations. For example, consider a setting where we care about the cost of communicating the data to the server (e.g., remote devices

communicating to the cloud). If  $d$  is large then communicating  $\mathbf{x}_i \in \mathbb{R}^d$  is costly. A natural scheme here is that the server chooses and announces a single random projection matrix  $\Phi$ , and every input point  $\mathbf{x}_i$  can be compressed and sent as  $\Phi\mathbf{x}_i$  to the server.<sup>1</sup> Such a scheme can be applied locally (i.e., on each  $\mathbf{x}_i$  independent of the other), something that is not possible if the aim is to compress the number of observations. Additionally, for a fixed  $m$ , reducing the dimensionality leads to more storage space savings than the reducing the number of observations, as storing  $n$  compressed features takes  $\approx O(mn)$  space whereas storing the reduced observations takes  $\approx O(md)$  space and  $d \gg n$ . In fact, in a high-dimensional setting, reducing the dimensionality seems intuitively the desirable way of achieving compression.

## 1.1 Our Contributions

We consider algorithms for linear regression that seek a sparse vector of regression coefficients. Our main result shows that, under a set of mild assumptions on the problem instance, we can estimate  $\theta^*$  even with access to only the compressed features. To put our results into context, we start with some background discussion about sparse linear regression using Lasso.

**Error Analysis of Lasso.** In a traditional sparse linear regression problem, given  $(X, \mathbf{y})$  that satisfies a linear system  $\mathbf{y} = X\theta^* + \mathbf{w}$  where  $\theta^*$  is  $k$ -sparse (i.e., has at most  $k$  non-zero entries) in  $\mathbb{R}^d$  and  $\mathbf{w} \in \mathbb{R}^n$  is the noise vector, the goal is to estimate  $\theta^*$ . Typically,  $\theta^*$  is  $k$ -sparse for  $k \ll d$ . Throughout this paper, our main focus will be on the standard Gaussian model for sparse linear regression, in which the entries of the noise vector  $\mathbf{w}$  are i.i.d. subgaussian and the matrix  $X$  is a deterministic matrix. For the purposes of this section, we make some simplifying assumptions and omit dependence on all but key variables.

A popular approach for solving a (traditional) sparse linear regression problem is the Lasso technique of  $\ell_1$ -penalized regression. Lasso minimizes the usual mean squared error loss penalized with (a multiple of) the  $\ell_1$ -norm of  $\theta$ :

$$\theta^{\text{Lasso}} \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \|\mathbf{y} - X\theta\|^2 + \lambda \|\theta\|_1. \quad (1)$$

The consistency properties of the Lasso are now well-understood under a variety of assumptions on the instance [21, 7]. One of weakest known sufficient condition for the convergence of the Lasso estimator ( $\theta^{\text{Lasso}}$ ) to  $\theta^*$  is the *restricted eigenvalue* (RE) condition due to Bickel *et al.* [1].<sup>2</sup> Informally, the RE condition on  $X$  lower bounds the quadratic form defined by  $X$  over a subset of sparse vectors (formally defined in Definition 1). If  $X$  satisfies the RE condition then it can be shown that with an appropriate choice of the regularization parameter  $\lambda$ ,  $\theta^{\text{Lasso}}$  satisfies the error bound:  $\|\theta^{\text{Lasso}} - \theta^*\| = O(\sqrt{k \log d/n})$ , with high probability over  $\mathbf{w}$ . The above error decay rate is known to be minimax optimal, meaning that it cannot be substantially improved upon by any estimator [12].

**Our Results and Techniques.** In a compressed sparse linear regression setting, the goal is still to estimate  $\theta^*$  where  $\mathbf{y} = X\theta^* + \mathbf{w}$ , but however we now have to do with just the compressed representation of  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  (i.e.,  $(\Phi\mathbf{x}_1, y_1), \dots, (\Phi\mathbf{x}_n, y_n)$ ), where  $\Phi \in \mathbb{R}^{m \times d}$  is a random projection matrix.<sup>3</sup>

Since the  $\mathbf{x}_i$ 's are not provided, directly applying an approach like Lasso is ruled out. Also since the original  $\mathbf{x}_i$ 's are not available, it is *a priori* unclear whether a good reconstruction of  $\theta^*$  is even possible. The aim of this paper is to resolve this question. For this, we consider a natural extension to the Lasso

<sup>1</sup>Note that communicating  $\Phi$  can be very efficient, e.g., by sending a seed to a pseudorandom generator.

<sup>2</sup>The RE condition is less severe than the *Restricted Isometry Property* (RIP) and other related conditions that can also be used for similar analyses [1].

<sup>3</sup>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.

formulation (1) that is based on using the projected covariate vectors:

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

Our goal then is to show that the  $\ell_2$ -error between  $\theta^{\text{comp}}$  and  $\theta^*$  is small under some reasonable assumptions on the instance.

Our main result (Theorem 3.4) shows if the *stable rank* of the Gramian matrix ( $X^\top X$ ) of  $X$  exceeds  $m$ , then  $\theta^{\text{comp}}$  satisfies the error bound:

$$\|\theta^{\text{comp}} - \theta^*\| = O\left(\frac{k^{5/2} \log^{3/2}(d)}{\|X\|_F} + \frac{k^{7/2} \log d}{\sqrt{d}}\right),$$

with high probability over  $\Phi, \mathbf{w}$ . Ignoring polylog factors, note that the second term  $k^{7/2}/\sqrt{d}$  is much smaller than  $k^{7/2}/\sqrt{n}$  as  $d \gg n$ . Also, as we discuss in Section 3.2, for many interesting families of covariate matrices,  $\|X\|_F = \Omega(\sqrt{nd})$ . Therefore, in these cases, the error in estimation decays at a rate much greater than  $k^{7/2}/\sqrt{n}$ .

Let us now talk about the stable rank condition. Stable rank of a matrix  $M$  ( $\operatorname{sr}(M)$ ), defined as the squared ratio of Frobenius and spectral norms of  $M$ , is a commonly used robust surrogate to usual matrix rank in linear algebra.<sup>4</sup> In our case, we rely on a stable rank condition on  $X^\top X$ . We compare various conditions in more detail in Appendix C. The picture that emerges is roughly as follows: (i) a stable rank condition on  $X$  ( $\operatorname{sr}(X)$ ) is less restrictive than a RE condition on  $X$ , and (ii) in many interesting settings of  $X$ ,  $\operatorname{sr}(X^\top X) \approx \operatorname{sr}(X)$ .

Our analysis follows the framework used in the traditional Lasso error analysis. For the purposes of the analysis, we consider a modified linear model:  $\mathbf{y} = X \Phi^\top \Phi \theta^* + \tilde{\mathbf{w}}$ . The matrix of interest now becomes  $X \Phi^\top \Phi$ , which we show satisfies a RE bound under the above stable rank condition on  $X^\top X$ . To establish a RE bound, we need a lower bound on  $\|\Phi^\top \Phi \theta\|$  on all unit vectors  $\theta$  from a certain sparse set. The proof is challenging because applying standard concentration tools directly do not give strong enough probability estimates on this quantity for a fixed  $\theta$  to successfully apply an  $\varepsilon$ -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  $\varepsilon$ -net argument. Throughout the proof, we rely on the Hanson-Wright inequality and several of its consequences. With a RE bound on  $X \Phi^\top \Phi$ , we investigate the setting of the regularization parameter  $\lambda$  that leads to a small  $\ell_2$ -error between  $\theta^{\text{comp}}$  and  $\theta^*$ .

Our results also trivially hold for the traditional sparse linear regression, as given  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ , the algorithm can pick  $\Phi$  and generate the input  $(\Phi \mathbf{x}_1, y_1), \dots, (\Phi \mathbf{x}_n, y_n)$  before using (2). While as discussed above this results in a weaker  $\ell_2$ -error bound than using the Lasso directly on  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ , nevertheless it does provide a result for the traditional sparse linear regression problem that operates under slightly different assumptions on  $X$ . Further exploring this connection is an interesting research direction.

## 1.2 Related Work

**Lasso and Sparse Regression.** Sparsity is the most widely studied structure of data that also provides attractive statistical properties and computational advantages. There is an extensive literature on the topic of sparse machine learning which have explored the close connections between it and areas such as compressed sensing, high-dimensional geometry, convex optimization, etc. (we refer the reader to books by Eldar *et al.*

<sup>4</sup>For every matrix  $M$ ,  $\operatorname{sr}(M^\top M) \leq \operatorname{sr}(M) \leq \operatorname{rank}(M)$ .

[3] and Rish *et al.* [13] for a detailed treatment). 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.* [7] for a detailed survey about developments here. In this paper, we draw on the rich literature studying theoretical properties of Lasso for sparse linear regression.

A recent area of research is that of distributed (communication efficient) sparse linear regression, where the dataset is assumed to be distributed across multiple machines (see, e.g., [9] and references therein). We do not know of a direct connection between these works and our setting.

Zhou *et al.* [24] considered sparse linear regression in a setting where the covariate matrix  $X$  is pre-multiplied by a Gaussian random projection matrix to generate  $m$  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, which as we discussed earlier has advantages in the context of sparse linear regression.

**Compression on the Feature Space (Compressed Learning).** Our problem setting is related to the framework of *compressed learning* [3], where the goal is to “learn” directly from the compressed features. Compressed learning algorithms have been developed for variety of common machine learning tasks such as ordinary least squares [10, 4, 8], classification [2], sparse subspace clustering [22], and robust PCA [5]. To the best of our knowledge ours is the first work dealing with the problem of sparse linear regression given only the projected data.

**Speeding up Regression using Random Projections.** There is a long line of work in using Johnson-Lindenstrauss style transforms for speeding up linear regression and its variants. For linear regression, the general idea is to consider the problem  $\min_{\theta} \|Ry - RX\theta\|^2$  instead of the original least-squares problem, where  $R$  is some appropriate choice of random matrix. Recent work in this space, have used *structured random projections*, such as those based on randomized Hadamard transform or Fourier transform, to generate a subsampled matrix, which is then used for estimating the regression coefficient  $\theta$  (we refer the reader to the survey by Woodruff [23] for more details). An open question here is to extend the results in this paper to  $\Phi$ 's that come from structured random projections as it could lead to better computational efficiency.

## 2 Preliminaries

**Notation.** We denote  $[n] = \{1, \dots, n\}$ . For a set  $S \subseteq [d]$ ,  $S^{\text{co}}$  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,  $\|\mathbf{v}\|_p$  its  $\ell_p$ -norm, and  $\text{supp}(\mathbf{v})$  its support. We use  $\mathbf{e}_j \in \mathbb{R}^d$  to denote the standard basis vector with  $j$ th entry set to 1. For a matrix  $M$ ,  $\|M\|$  denotes its spectral norm which equals its largest singular value, and  $\|M\|_F$  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  $\ell_p$ -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  $\Sigma_k$  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 B, we also review a few additional concepts related to  $\varepsilon$ -nets, subgaussian random variables, and randomized dimensionality reduction techniques.

**Background on Lasso for Sparse Linear Regression.** Here we describe necessary background on how

Lasso provides an estimate of sparse regression vector (we refer the reader to the book by Hastie *et al.* [7] for a detailed treatment on this topic).

A dominant goal<sup>5</sup> in this line of work has been to establish conditions on the instance under which the  $\ell_2$ -error on estimating  $\theta^*$  is well-controlled. For aiding this discussion, we would need few additional definitions. We also assume access to the original  $(\mathbf{x}_i, y_i)$ 's.

For a set  $S \subset [d]$ , let us define a cone set  $\mathbb{C}(S)$  as:

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

Restricted eigenvalue is a mild condition on the covariate matrix that is sufficient for estimating  $\theta^*$  in a noisy linear model setup.<sup>6</sup>

**Definition 1** (Restricted Eigenvalue [1]). *A matrix  $X \in \mathbb{R}^{n \times d}$  satisfies the restricted eigenvalue (RE) condition with parameter  $\xi$  if,*

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

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 [11].

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  $\mathbf{y} = X\theta^* + \mathbf{w}$ .

**Theorem 2.1** ([1, 11, 7]). *Let  $\mathbf{y} = X\theta^* + \mathbf{w}$  for a noise vector  $\mathbf{w} \in \mathbb{R}^n$  and  $\theta^*$  is  $k$ -sparse. Let  $\lambda_n \geq 2\|X^\top \mathbf{w}\|_\infty/n$ . Suppose  $X$  satisfies the restricted eigenvalue condition with parameter  $\xi > 0$ , then any optimal minimizer,  $\hat{\theta} \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n}\|\mathbf{y} - X\theta\|^2 + \lambda_n\|\theta\|_1$ , satisfies:  $\|\hat{\theta} - \theta^*\| \leq 3\sqrt{k}\lambda_n/\xi$ .*

**Remark 2.2.** [A Note on Assumptions] *While the above RE condition is common for analyzing the  $\ell_2$ -error of the Lasso estimator [11], stronger conditions are used for achieving the stronger guarantee of consistent support selection [21, 7]. These include mutual incoherence and minimum eigenvalue conditions on  $X$ , and minimum signal value condition on  $\theta^*$ . These conditions are known to be highly restrictive [19].*

### 3 Sparse Linear Regression with Compressed Features

In this section, we consider the problem of sparse linear regression in a model where the algorithm only gets access to  $\Phi \mathbf{x}_i$ 's and  $\Phi$ , and not  $\mathbf{x}_i$ 's. 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 (1) on  $(\hat{\mathbf{x}}_i, y_i)$ 's. This idea, however, is problematic because good reconstruction of  $\mathbf{x}_i$ 's from  $\Phi \mathbf{x}_i$ 's will require (sparsity-like) assumptions on the structure of the  $\mathbf{x}_i$ 's. Additionally, sparse linear regression analyses (such as for Lasso) require certain assumptions (such as RE) about the instance, which may not be satisfied by  $\hat{\mathbf{x}}_i$ 's, even if the original  $\mathbf{x}_i$ 's satisfy these assumptions.

Our idea for tackling the compressed sparse linear regression problem is based on using a variant of the Lasso formulation. Let  $\Phi$  be an  $m \times d$  random matrix with independent subgaussian entries. If the algorithm

<sup>5</sup>Other goals considered in the literature include establishing conditions for recovery of the support set of the unknown regression vector [7]. More on this in Remark 2.2.

<sup>6</sup>Given that we observe only a noisy version of the product  $X\theta^*$ , it is then difficult to distinguish  $\theta^*$  from other sparse vectors. Thus, it is natural to impose an RE condition if the goal is to produce an estimate  $\hat{\theta}$  such that  $\|\theta^* - \hat{\theta}\|$  is small.

has only access to  $(\Phi \mathbf{x}_1, y_1), \dots, (\Phi \mathbf{x}_n, y_n)$  and  $\Phi$ , a natural extension to Lasso is:

$$\theta^{\text{comp}} \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (y_i - \langle \Phi \mathbf{x}_i, \Phi \theta \rangle)^2 + \lambda_n \|\theta\|_1 \equiv \operatorname{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \|\mathbf{y} - X \Phi^\top \Phi \theta\|^2 + \lambda_n \|\theta\|_1$$

Our goal is to establish a bound on the  $\ell_2$ -error between  $\theta^{\text{comp}}$  and  $\theta^*$  (Theorem 3.4). For this, we consider a modified linear model:  $\mathbf{y} = X \Phi^\top \Phi \theta^* + \tilde{\mathbf{w}}$  (note that the true linear model is  $\mathbf{y} = X \theta^* + \mathbf{w}$ ). In the following, we establish the conditions needed for invoking Theorem 2.1 on this modified linear model. The matrix of interest is now  $X \Phi^\top \Phi$ . We start off by establishing a RE bound on this matrix (Section 3.1). In Section 3.2, we investigate the setting of the regularization parameter  $\lambda_n$ . Putting these pieces together in the framework of Theorem 2.1 bounds  $\|\theta^{\text{comp}} - \theta^*\|$ .

### 3.1 Restricted Eigenvalue Condition on $X \Phi^\top X$

In this section, we show how a stable rank condition on  $X^\top X$  translates into a RE bound on the matrix  $X \Phi^\top \Phi$ . We start with the definition of stable rank (denoted by  $\operatorname{sr}(\cdot)$ ) of a matrix  $X$ .

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

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. Also since,

$$\|X^\top X\|_{\text{F}} \leq \|X\|_{\text{F}} \cdot \|X\| \implies \operatorname{sr}(X^\top X) \leq \operatorname{sr}(X).$$

Throughout this section,  $C, C_1, c, c_1, \dots$  denote positive constants which may depend on the subgaussian norm of the entries of the involved matrices.

For the proof, it will be convenient to work with a slightly modified (and a more general) definition of restricted eigenvalue that we state here.

**Definition 2.** Let  $V$  be an  $N \times M$  matrix, and let  $k < M$ ,  $\alpha > 0$ . Define

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

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

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

Note that  $\alpha = 3$  in Definition 1. Also given  $\operatorname{RE}(V, k, \alpha)$ , we can get a lower bound on  $\xi$  in Definition 1 as  $\xi \geq \operatorname{RE}(V, k, 3)^2/k$ .

Our primary result in this section is the following theorem which establishes a lower bound on  $\operatorname{RE}(X \Psi^\top \Psi, k, \alpha)$ . The proof assumes a stable rank condition on  $X^\top X$  that we define below. In Appendix C, we provide a detailed discussion about how the stable rank condition is practically reasonable and compares with the RE condition.

**Theorem 3.1.** Let  $m, n, d \in \mathbb{N}$ ,  $m \leq n \leq d$ , and let  $X$  be a fixed  $n \times d$  matrix satisfying

$$\text{Stable Rank Condition : } 4 \leq m \leq \operatorname{sr}(X^\top X)/4.$$

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  $p \in (0, 1)$ . Then for any  $k \in \mathbb{N}$ ,  $\alpha > 0$  such that

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

the matrix  $X\Psi^\top\Psi$  satisfies

$$\text{RE}(X\Psi^\top\Psi, k, \alpha) \geq \frac{1}{32}\sqrt{m}\|X\|_F$$

with probability at least  $1 - p$ .

**Corollary 3.2.** Let  $X$  and  $\Psi$  be matrices satisfying the conditions in Theorem 3.1 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:

$$\inf_{S \subset [d], |S|=k, \theta \in \mathbb{C}(S)} \frac{\|X\Phi^\top\Phi\theta\|^2}{n} \geq \frac{\|X\|_F^2 \|\theta\|^2}{1024 nmk},$$

with probability at least  $1 - \beta$ .

The complete proof of the above theorem is presented in Section 4. Here we provide a high-level description of the proof idea.

**Idea of the Proof of Theorem 3.1.** 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  $\varepsilon$ -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  $Q : \mathbb{R}^n \rightarrow \mathbb{R}^n$  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  $\{\mathbf{x} \in \mathbb{R}^d : \text{supp}(\mathbf{x}) \subseteq J^{\text{co}}, \|\mathbf{x}\|_1 \leq \alpha\sqrt{k}\}$ , 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.



### 3.2 Bounding the $\ell_2$ -error

In this section, we bound the  $\ell_2$ -error between  $\theta^{\text{comp}}$  and  $\theta^*$ . We do so by using the RE bound established in Corollary 3.2 and some additional simple conditions needed for our analysis. We start with the definition of a *well-behaved* instance that precisely state these additional conditions.

**Definition 3** (Well-behaved Instance). *An instance  $(X, \mathbf{y}, \theta^*)$ , where  $X \in \mathbb{R}^{n \times d}$  and  $\mathbf{y} \in \mathbb{R}^n$ , and  $\theta \in \mathbb{R}^d$ , is  $(k, \sigma)$ -well behaved if there exists a  $\mathbf{w} \in \mathbb{R}^d$  such that  $\mathbf{y} = X\theta^* + \mathbf{w}$  and:*

1. *Bounded estimator vector:  $\theta^* \in \Sigma_k$  (i.e.,  $\theta^*$  is  $k$ -sparse and  $\|\theta^*\| \leq 1$ ).*<sup>7</sup>
2. *Noise condition: The entries of the noise vector  $\mathbf{w} = (w_1, \dots, w_n)$  are independent centered subgaussians with  $\|w_i\|_{\psi_2} \leq \sigma$  (Definition 5).*

Note that these above assumptions are typical in the analysis of Lasso and related approaches to sparse linear regression (see, e.g., Hastie *et al.* [7]).

We now assume that  $(X, \mathbf{y}, \theta^*)$  is  $(k, \sigma)$ -well behaved. Again consider the modified linear model:  $\mathbf{y} = X\Phi^\top \Phi \theta^* + \tilde{\mathbf{w}}$ . To establish the necessary bound on  $\lambda_n$  for Theorem 2.1, we bound  $\|(X\Phi^\top \Phi)^\top \tilde{\mathbf{w}}\|_\infty / n$ . The proof of the following proposition is presented in Appendix D.

**Proposition 3.3.** *Let  $(X, \mathbf{y}, \theta^*)$  be  $(k, \sigma)$ -well behaved. 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  $m = \Theta(k \log(d/\beta))$  and  $\Phi = \Psi/\sqrt{m}$ . Then with probability at least  $1 - \beta$ ,*

$$\frac{\|(X\Phi^\top \Phi)^\top \tilde{\mathbf{w}}\|_\infty}{n} = O\left(\frac{\sigma \|X\|_F \log(d/\beta)}{n\sqrt{m}} + \frac{\|X\|_F^2}{n\sqrt{d}}\right).$$

Our main result now follows by invoking the Theorem 2.1 on the modified linear model  $\mathbf{y} = X\Phi^\top \Phi \theta^* + \tilde{\mathbf{w}}$  with the results from Corollary 3.2 and Proposition 3.3.

**Theorem 3.4** (Main Theorem). *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  $(X, \mathbf{y}, \theta^*)$  be  $(k, \sigma)$ -well behaved. Let  $m = \Theta(k^2 \log(d/\beta))$  for  $0 \leq \beta \leq 1/2$ . If  $X$  satisfies the stable rank condition:  $\text{sr}(X^\top X) \geq 4m$ , then any optimal minimizer,*

$$\theta^{\text{comp}} \in \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - \langle \Phi \mathbf{x}_i, \Phi \theta \rangle)^2 + \lambda_n \|\theta\|_1, \text{ with } \lambda_n = \Theta\left(\frac{\sigma \|X\|_F \log(d/\beta)}{n\sqrt{m}} + \frac{\|X\|_F^2}{n\sqrt{d}}\right),$$

with probability at least  $1 - \beta$  satisfies:

$$\|\theta^{\text{comp}} - \theta^*\| = O\left(\frac{k^{3/2} \sqrt{m} \sigma \log(d/\beta)}{\|X\|_F} + \frac{k^{3/2} m}{\sqrt{d}}\right) = O\left(\frac{k^{5/2} \sigma \log^{3/2}(d/\beta)}{\|X\|_F} + \frac{k^{7/2} \log(d/\beta)}{\sqrt{d}}\right).$$

**Discussion about Theorem 3.4.** In the first term of the error bound, note that  $\|X\|_F$  is a function of both  $n$  and  $d$ . As a point of comparison, for a very broad class of random matrices  $X$ , including ones with significant dependencies between the entries, with high probability,  $\|X\|_F = \Omega(\sqrt{nd})$  [16]. In general, if  $X$  satisfies the RE condition (Definition 1) with parameter  $\xi$ , then  $\|X\|_F \geq \sqrt{\xi nd}$  as:

$$\|X\|_F = \left(\sum_{j=1}^d \|X \mathbf{e}_j\|^2\right)^{1/2} \geq \sqrt{\xi nd}.$$

<sup>7</sup>To simplify presentation, we assume  $\|\theta^*\| \leq 1$ , but our results directly extend to any bound on  $\|\theta^*\|$ .

The second term in the error bound is independent of  $n$ , but since  $d \gg n$ , it implies that  $k^{7/2}\sqrt{d} \ll k^{7/2}\sqrt{n}$ . Therefore, when  $\|X\|_F = \Omega(\sqrt{nd})$ , the estimation error decays at a rate much greater than  $k^{7/2}/\sqrt{n}$ . In other words, the estimator  $\theta^{\text{comp}}$  is *consistent* when  $n = \omega(k^7)$ .

We suspect that the dependence on the sparsity factor in bound of Theorem 3.4 could possibly be reduced with a tighter analysis of Theorem 3.1.

## 4 Restricted Eigenvalue Bound on $X\Phi^\top\Phi$ : Proof of Theorem 3.1

In this section, we present the complete proof of Theorem 3.1. In Section 4.1, we use the Hanson-Wright theorem and its corollaries to get probabilistic estimates for norms of certain matrix products. In Section 4.2, we prove Theorem 3.1 for a fixed vector of a special form. We finish the proof in Section 4.3.

### 4.1 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 [6].

**Theorem 4.1** (Hanson-Wright Inequality [15]). *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  $t \geq 0$ ,*

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

Besides the theorem itself, we need several corollaries.

**Corollary 4.2** (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  $s, t \geq 1$ ,*

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

and

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

Corollary 4.2 can be found in [15]. Assuming that  $m \leq \mathbf{sr}(B)$ , we can rewrite the above inequalities as

$$\Pr \left[ \frac{1}{2} \|B\|_F < \left\| BG^\top \right\| < C \|B\|_F \right] \geq 1 - 2 \exp(-c \mathbf{sr}(B)). \quad (3)$$

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 4.3** (Small ball). *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*

$$\Pr \left[ \|B\mathbf{g}\| \leq \frac{1}{2} \|B\|_{\text{F}} \right] \leq 2 \exp(-c \mathbf{sr}(B)).$$

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

**Corollary 4.4** (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*

$$\Pr \left[ \|BG^{\top}\|_{\text{F}} \leq \frac{1}{2} \sqrt{m} \|B\|_{\text{F}} \right] \leq 2 \exp(-cm \mathbf{sr}(B)).$$

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

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

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  $\gamma_1, \dots, \gamma_m$  under the  $nm \times dm$  block-diagonal matrix  $\tilde{B} = \text{diag}(B, \dots, B)$ .

The result follows from the Corollary 4.3, since  $\|\tilde{B}\|_{\text{F}}^2 = m \|B\|_{\text{F}}^2$  implying  $\|\tilde{B}\|_{\text{F}} = \sqrt{m} \|B\|_{\text{F}}$ .  $\square$

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  $\text{tr}()$  denote the trace of a matrix.

**Corollary 4.5** (Frobenius norm of the Triple Product). *Let  $m > 2$ . 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. Then*

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

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

$$\|GHG^{\top}\|_{\text{F}}^2 = \sum_{\substack{i,j \in [m] \\ i \neq j}} (\gamma_i^{\top} H \gamma_j)^2 + \sum_{j=1}^m (\gamma_j^{\top} H \gamma_j)^2.$$

Fix  $j \in [m]$  and denote by  $G_j$  the  $(m-1) \times d$  matrix obtained from  $G$  by removing the  $j$ th row. Define

$$Y_j = \sum_{i \in [m] \setminus \{j\}} (\gamma_i^{\top} H \gamma_j)^2 = \|G_j H \gamma_j\|^2.$$

Conditioning on  $G_j$  and using Corollary 4.2 with  $m = 1$ , we obtain

$$\Pr \left[ Y_j \geq C \|G_j H \gamma_j\|_{\text{F}}^2 \mid G_j \right] \leq 2 \exp(-c \mathbf{sr}(G_j H)).$$

To apply the previous inequality, we have to bound  $\|G_j H\|_{\mathbb{F}}$  and  $\|G_j H\|$ . Let  $\Omega_{\mathbb{F}}$  be the event  $\|G_j H\|_{\mathbb{F}} \geq \frac{1}{2}\sqrt{m-1}\|H\|_{\mathbb{F}}$ . By Corollary 4.4,

$$\Pr [\Omega_{\mathbb{F}}^{\text{co}}] \leq 2 \exp(-c(m-1) \mathbf{sr}(H)).$$

Also, let  $\Omega_{op}$  be the event  $\|G_j H\| \leq C(\sqrt{m-1}\|H\| + \|H\|_{\mathbb{F}})$  and  $\Omega_{op}^{\text{co}}$  be the complement event. Then by Corollary 4.2,

$$\Pr [\Omega_{op}^{\text{co}}] \leq 2 \exp(-\mathbf{sr}(H) - (m-1)).$$

If both  $\Omega_{\mathbb{F}}$  and  $\Omega_{op}$  occur, then

$$\mathbf{sr}(G_j H) \geq c \frac{(m-1) \|H\|_{\mathbb{F}}^2}{(m-1) \|H\|^2 + \|H\|_{\mathbb{F}}^2} \geq c' \min(\mathbf{sr}(H), m),$$

where we used  $m > 2$  to replace  $m-1$  by  $m$ . Therefore,

$$\begin{aligned} \Pr [Y_j \geq Cm \|H\|_{\mathbb{F}}^2] &\leq \Pr [Y_j \geq Cm \|H\|_{\mathbb{F}}^2 \mid \Omega_{\mathbb{F}} \cap \Omega_{op}] + \Pr [\Omega_{\mathbb{F}}^{\text{co}}] + \Pr [\Omega_{op}^{\text{co}}] \\ &\leq 2(\exp(-c' \min(\mathbf{sr}(H), m)) + \exp(-cm \mathbf{sr}(H)) + \exp(-\mathbf{sr}(H) - (m-1))) \\ &\leq 2(\exp(-c \mathbf{sr}(H)) + \exp(-m)). \end{aligned}$$

Consider the diagonal terms now. For  $j \in [m]$ , set  $Z_j = \gamma_i^\top H \gamma_j$ . Then  $\mathbb{E}[Z_j] = \text{tr}(H)$ . Since  $\|H\|_{\mathbb{F}}^2 \leq \|H\| \cdot \text{tr}(H)$ , we can apply Theorem 4.1 to get

$$\Pr [Z_j > 2 \text{tr}(H)] \leq 2 \exp\left(-c \frac{\text{tr}(H)}{\|H\|}\right) \leq 2 \exp(-c \mathbf{sr}(H)).$$

Thus,

$$\begin{aligned} \Pr \left[ \left\| \left\| GHG^\top \right\|_{\mathbb{F}}^2 \geq Cm^2 \|H\|_{\mathbb{F}}^2 + 2m \cdot \text{tr}(H)^2 \right\| \right] &\leq \sum_{j=1}^m \Pr [Y_j \geq Cm \|H\|_{\mathbb{F}}^2] + \sum_{j=1}^m \Pr [Z_j \geq 2 \text{tr}(H)] \\ &\leq 4m (\exp(-c \mathbf{sr}(H)) + \exp(-m)), \end{aligned}$$

as claimed.  $\square$

## 4.2 Bounds for a Fixed Vector

In this section, our goal will be to investigate a special case of Theorem 3.1. In particular, we investigate the RE condition in Definition 2 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 4.8). 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 4.6.** *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  $\mathbf{g} = (g_1, \dots, g_m) \in \mathbb{R}^m$  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*

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

*Proof.* Conditioning on  $G$  and applying Corollary 4.3, we obtain

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

Define the events  $\Omega_{\mathbb{F}}$  and  $\Omega_{op}$  as in Corollary 4.5:

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

Let  $\Omega_{\mathbb{F}}^{\text{co}}$  and  $\Omega_{op}^{\text{co}}$  denote the complement of these events respectively. Then by Corollaries 4.4 and 4.2,

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

□

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

**Lemma 4.7.** *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  $\mathbf{g}_1 = (g_{11}, \dots, g_{1m}) \in \mathbb{R}^m$  and  $\mathbf{g}_2 = (g_{21}, \dots, g_{2m}) \in \mathbb{R}^m$  be random vectors with independent entries that satisfy:  $\mathbb{E}[G_{ij}] = \mathbb{E}[g_{lj}] = 0$ ,  $\mathbb{E}[G_{ij}^2] = \mathbb{E}[g_{lj}^2] = 1$ , and  $\|G_{ij}\|_{\psi_2}, \|g_{lj}\|_{\psi_2}$  are all bounded for  $l \in \{1, 2\}$ . Assume that  $m \leq \mathbf{sr}(B^\top B)$ . Then for any  $t \in [0, m \|B\|_{\mathbb{F}}^2]$ ,*

$$\Pr \left[ |\mathbf{g}_1^\top G B^\top B G^\top \mathbf{g}_2| \geq t \right] \leq 8 \exp \left( -c \frac{t^2}{m \|B\|_{\mathbb{F}}^4} \right).$$

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

$$\mathbf{g} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 0 & G B^\top B G^\top \\ G B^\top B G^\top & 0 \end{pmatrix}.$$

Condition on  $G$ . By Theorem 4.1, for any  $t \geq 0$ ,

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

Note that  $\|\Gamma\| = \|G B^\top B G^\top\| = \|B G^\top\|^2$ . Let  $\Omega_{HS}$  and  $\Omega_{op}$  be the events defined by

$$\begin{aligned} \Omega_{HS} &= \left\{ G : \left\| G B^\top B G^\top \right\|_{\mathbb{F}} \leq C \left( m \left\| B^\top B \right\|_{\mathbb{F}} + \sqrt{m} \cdot \text{tr}(B^\top B) \right) \right\} \\ \Omega_{op} &= \left\{ G : \frac{1}{4} \|B\|_{\mathbb{F}}^2 \leq \left\| G B^\top B G^\top \right\| \leq C \|B\|_{\mathbb{F}}^2 \right\} \end{aligned}$$

Again, let  $\Omega_{\mathbb{F}}^{\text{co}}$  and  $\Omega_{op}^{\text{co}}$  denote the complement events. Since  $\|B^\top B\|_{\mathbb{F}} \leq \|B\|_{\mathbb{F}} \cdot \|B\|$ , for any  $G \in \Omega_{\mathbb{F}} \cap \Omega_{op}$ ,

$$\|\Gamma\|_{\mathbb{F}}^2 \leq C \left( m^2 \|B^\top B\|_{\mathbb{F}}^2 + m \cdot \text{tr}(B^\top B)^2 \right) \leq C' m \|B\|_{\mathbb{F}}^4,$$

where we used the assumption  $m \leq \text{sr}(B)$  in the last inequality.

Finally, combining this with Corollary 4.5, and (3), we obtain

$$\begin{aligned} \Pr \left[ |\mathbf{g}_1^\top G B^\top B G^\top \mathbf{g}_2| \geq t \right] &\leq 2 \exp \left[ -c \min \left( \frac{t^2}{m \|B\|_{\mathbb{F}}^4}, \frac{t}{\|B\|_{\mathbb{F}}^2} \right) \right] + \Pr [\Omega_{\mathbb{F}}^{\text{co}}] + \Pr [\Omega_{op}^{\text{co}}] \\ &\leq 4 \exp \left( -c \frac{t^2}{m \|B\|_{\mathbb{F}}^4} \right) + 2 \exp(-c \text{sr}(B)) + 4m \left( \exp(-c \text{sr}(B^\top B)) + \exp(-m) \right) \end{aligned}$$

for any  $t \in [0, m \|B\|_{\mathbb{F}}^2]$ . Since for any such  $t$ , the first term in the right-hand side dominates the other three, the proof is complete.  $\square$

Using Lemmas 4.6 and 4.7, 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 4.8.** *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*

$$4 \leq m \leq \text{sr}(R^\top R).$$

Then for any  $s \geq 1$ ,

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

*Proof.* Let  $P_1$  be the orthogonal projection in  $\mathbb{R}^n$  with  $\text{Ker}(P_1) = \text{span}(R\mathbf{e}_1)$ , where  $\text{span}()$  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

$$\left\| R G^\top G (\mathbf{e}_1 + \mathbf{x}) \right\| \geq \left\| P_1 R G^\top G (\mathbf{e}_1 + \mathbf{x}) \right\| \geq \left\| P_1 R G^\top G \mathbf{e}_1 \right\| - \mathbf{u}^\top P_1 R G^\top G \mathbf{x}. \quad (4)$$

The minimal value of this expression over  $\mathbf{x} \in s \cdot \text{conv}(\pm \mathbf{e}_2, \dots, \pm \mathbf{e}_d)$  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  $\text{sr}(R) > 4$  and by the interlacing, we have

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

and so,  $\text{sr}(P_1 R) \geq (1/2) \text{sr}(R)$  (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  $P_1 R G^\top$  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 4.6,

$$\begin{aligned} \Pr \left[ \left\| P_1 R G^\top G \mathbf{e}_1 \right\| < \frac{1}{4} \sqrt{m} \|R\|_F \right] &= \Pr \left[ \left\| P_1 R G^\top \mathbf{g}_1 \right\| < \frac{1}{4} \sqrt{m} \|R\|_F \right] \\ &\leq 8 \left( \exp(-c \mathbf{sr}(R)) + \exp(-cm) \right) \leq 2 \exp(-c'm), \end{aligned} \quad (5)$$

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

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

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

and therefore,

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

Note that  $P_2 R G^\top$  is independent of  $\mathbf{g}_1$  and  $\mathbf{g}_2$ . This allows us to use Lemma 4.7 to estimate,

$$\Pr \left[ \left| \mathbf{g}_1^\top G (P_2 R)^\top P_2 R G^\top \mathbf{g}_2 \right| \geq t \right] \leq 8 \exp \left( -c \frac{t^2}{m \|P_2 R\|_F^4} \right), \quad (6)$$

for any  $t \in [0, m \|P_2 R\|_F^2]$ .

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

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

and

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

we obtain

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

or

$$\Pr \left[ \left| (P_1 R \mathbf{e}_2 \mathbf{g}_2^\top \mathbf{g}_1)^\top P_1 R \mathbf{e}_2 \mathbf{g}_2^\top \mathbf{g}_2 \right| \geq t \right] \leq 2 \exp \left( -c \frac{t^2}{m^3 \|P_1 R \mathbf{e}_2\|^4} \right) + \exp(-m) \quad (7)$$

for all  $t \geq 0$ . Combining (6) and (7), we conclude that

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

for any  $t \in [0, m \|P_2 R\|_F^2]$ . Here we used the inequality

$$m \|P_1 R \mathbf{e}_2\|^2 \leq m \|R\|^2 \leq \|R\|_F^2,$$

where the last one follows from the assumption  $m \leq \text{sr}(R^\top R) \leq \text{sr}(R)$ . Taking into account the result from (5), we see that

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

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

$$\Pr \left[ \exists j \geq 2, |\mathbf{u}^\top P_1 R G^\top G \mathbf{e}_j| > \tau \right] \leq 2d \left( \exp \left( -c \frac{\tau^2}{\|R\|_F^2} \right) + \exp(-cm) \right). \quad (8)$$

Recall (4). Setting  $\tau = \frac{1}{8s} \sqrt{m} \|R\|_F$  with  $s \geq 1$ , and using together (5) and (8), we conclude that

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

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

### 4.3 Finishing the Proof of Theorem 3.1: Net Argument

The next theorem is the main technical step in proving Theorem 3.1. Invoking this theorem with appropriate parameters (that we explain later in this section) gives the proof of Theorem 3.1. 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 4.8, and then employing an  $\varepsilon$ -net argument.

**Theorem 4.9.** *Let  $X$  be a fixed  $n \times d$  matrix satisfying,*

$$4 \leq m \leq \text{sr}(X^\top X)/4.$$

*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  $p \in (0, 1)$ , and let  $k \in \mathbb{N}$ . Then for any  $s$  such that*

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

$\Pr[\exists I \subset [d]$  with  $|I| = k$ ,  $\exists \mathbf{y} \in \mathbb{S}^{d-1}$  with  $\text{supp}(\mathbf{y}) \subseteq I$ ,  $\exists \mathbf{x} \in s \cdot \text{conv}(\pm \mathbf{e}_i, i \notin I)$ ,

$$\left\| X \Psi^\top \Psi (\mathbf{y} + \mathbf{x}) \right\| \leq \frac{1}{32} \sqrt{m} \|X\|_F] \leq p.$$

Note that the condition  $s \geq 1$  in the formulation of the theorem implicitly sets a lower bound on  $p$  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

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

Note that the vectors  $\mathbf{y}$  and  $\mathbf{e}_j, j > k$  form an orthonormal basis of  $E$ . Let  $P_E : \mathbb{R}^d \rightarrow E$  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  $Q : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be the orthogonal projection with  $\text{Ker}(Q) = XE^\perp$ , where  $E^\perp$  represents the orthogonal complement of  $E$ . Then for any  $\mathbf{z} \in E$ ,

$$\left\| X\Psi^\top\Psi\mathbf{z} \right\| \geq \left\| QX\Psi^\top\Psi\mathbf{z} \right\|. \quad (9)$$

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  $\text{supp}(\mathbf{y}) \subseteq [k]$ , 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  $R = QX P_E^\top$ . Then by the interlacing

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

since by the assumptions on  $k$  and  $X$ ,  $k \leq m/8 \leq \text{sr}(X^\top X)/8 \leq \text{sr}(X)/8$ . Similarly, writing  $\|R^\top R\|_{\text{F}}^2$  in terms of the singular values of  $R$  and using the interlacing, we obtain

$$\left\| X^\top X \right\|_{\text{F}}^2 \geq \left\| R^\top R \right\|_{\text{F}}^2 \geq \frac{1}{4} \left\| X^\top X \right\|_{\text{F}}^2,$$

which implies

$$\text{sr}(R^\top R) \geq \frac{1}{4} \text{sr}(X^\top X) \geq m.$$

Applying Proposition 4.8 to the matrices  $G, R$ , with  $\mathbf{y}$  playing the role of  $\mathbf{e}_1$ , and taking into account (9), we obtain

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

for any  $s \geq 1$ .

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

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

with probability at least  $1 - \exp(-m)$ , where we used Corollary 4.2. Let  $\varepsilon > 0$  be a number to be chosen later, and (by Proposition B.1) let  $\mathcal{N} \subset \mathbb{S}^{k-1}$  be an  $\varepsilon$ -net of cardinality

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

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

$$\left\| X \Psi^\top \Psi(\mathbf{y} + \mathbf{x}) \right\| \geq \frac{1}{16} \sqrt{m} \|X\|_F.$$

Assume also that  $\|X \Psi^\top \Psi\| \leq C\sqrt{d} \|X\|_F$ . 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

$$\left\| X \Psi^\top \Psi(\mathbf{z} + \mathbf{x}) \right\| \geq \left\| X \Psi^\top \Psi(\mathbf{y} + \mathbf{x}) \right\| - \left\| X \Psi^\top \Psi \right\| \cdot \|\mathbf{z} - \mathbf{y}\| \geq \frac{1}{32} \sqrt{m} \|X\|_F.$$

Thus,

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

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

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

The last quantity is smaller than  $p$  provided that<sup>8</sup>

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

This completes the proof of the theorem.  $\square$

We now have all the ingredients to complete the proof of Theorem 3.1.

**Proof of Theorem 3.1.** Assume that the complement of the event described in Theorem 4.9 occurs. Namely, assume that

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

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

$$\|\mathbf{z}_{I^c}\|_1 \leq \alpha \|\mathbf{z}_I\|_1.$$

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  $\|\mathbf{z}_{I^c}\|_1 \leq \alpha\sqrt{k}$ . Theorem 3.1 now follows from Theorem 4.9 applied with  $s = \alpha\sqrt{k}$ .  $\square$

<sup>8</sup>Here we ignored smaller order terms assuming  $d^2 \gg mk$ . If this does not hold, one can obtain a slightly better estimate.

## References

- [1] 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.
- [2] Robert Calderbank, Sina Jafarpour, and Robert Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. *preprint*, 2009.
- [3] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.
- [4] Mahdi Milani Fard, Yuri Grinberg, Joelle Pineau, and Doina Precup. Compressed least-squares regression on sparse spaces. In *AAAI*, 2012.
- [5] Wooseok Ha and Rina Foygel Barber. Robust pca with compressed data. In *NIPS*, pages 1927–1935, 2015.
- [6] 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.
- [7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.
- [8] 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.
- [9] 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.
- [10] Odalric Maillard and Rémi Munos. Compressed least-squares regression. In *NIPS*, 2009.
- [11] 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.
- [12] 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.
- [13] Irina Rish and Genady Grabarnik. *Sparse modeling: theory, algorithms, and applications*. CRC Press, 2014.
- [14] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21, 2007.
- [15] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.
- [16] Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013.
- [17] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152. IEEE, 2006.

- [18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [19] Ryan Tibshirani and Larry Wasserman. Sparsity and the lasso, <http://www.stat.cmu.edu/~larry/=sml/sparsity.pdf>.
- [20] Roman Vershynin. Introduction to the Non-asymptotic Analysis of Random Matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- [21] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5), 2009.
- [22] Yining Wang, Yu-Xiang Wang, and Aarti Singh. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1422–1431, 2015.
- [23] David P. Woodruff. Sketching as a tool for numerical linear algebra. *FnT-TCS*, 10(1–2):1–157, 2014.
- [24] Shuheng Zhou, John Lafferty, and Larry Wasserman. Compressed and privacy-sensitive sparse regression. *Information Theory, IEEE Transactions on*, 55(2):846–866, 2009.

## A Background on Sparse Linear Regression

If the linear model  $\mathbf{y} = X\theta^* + \mathbf{w}$ , where  $X \in \mathbb{R}^{n \times d}$  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  $\theta^*$ , the statistical model  $\mathbf{y} = X\theta^* + \mathbf{w}$  is not *identifiable*. This is because (even when  $\mathbf{w} = 0$ ), there are many vectors  $\theta^*$  that are consistent with the observations  $\mathbf{y}$  and  $X$ . This identifiability concern may be eliminated by imposing some type of sparsity assumption on the regression vector  $\theta^*$ . Typically,  $\theta^*$  is  $k$ -sparse for  $k \ll d$ . Under this assumption, the goal of *sparse linear regression* is to find a sparse  $\theta$  with few nonzero entries such that  $\langle \mathbf{x}_i, \theta \rangle \approx y_i$  for “most”  $(\mathbf{x}_i, y_i)$  pairs. Disregarding computational cost, the most direct approach to estimating a  $k$ -sparse  $\theta$  in the linear regression model would be solving a quadratic optimization problem with an  $\ell_0$ -constraint:

$$\theta^{\text{sparse}} \in \operatorname{argmin}_{\theta \in \Sigma_k} \frac{1}{n} \sum_{i=1}^n (y_i - \langle \mathbf{x}_i, \theta \rangle)^2. \quad (10)$$

**Lasso Regression.** Since (10) leads to a non-convex problem, a natural alternative is obtained by replacing the  $\ell_0$ -constraint with its tightest convex relaxation, the  $\ell_1$ -norm. This leads to the popular Lasso regression, defined as,

$$\text{Lasso Regression (penalized form): } \theta^{\text{Lasso}} \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (y_i - \langle \mathbf{x}_i, \theta \rangle)^2 + \lambda \|\theta\|_1,$$

for some choice  $\lambda > 0$ .

The consistency properties of Lasso are now well-understood. Under a variety of mild assumptions on the instance, the Lasso estimator ( $\theta^{\text{Lasso}}$ ) is known to converge to the sparse  $\theta^*$  in the  $\ell_2$ -norm. Under stronger assumptions (such as mutual incoherence, minimum eigenvalue, and minimum signal condition) on the instance, it is also known that  $\theta^{\text{Lasso}}$  will have the same support as  $\theta^*$ .

## B Additional Preliminaries

**Background on  $\varepsilon$ -Nets.** Consider a subset  $T$  of  $\mathbb{R}^d$ , and let  $\varepsilon > 0$ . A  $\varepsilon$ -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 B.1** (Volumetric Estimate). *Let  $T$  be a subset of  $B_2^d$  and let  $\varepsilon > 0$ . Then there exists an  $\varepsilon$ -net  $\mathcal{N}$  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 4** (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 5** (Norm of a Subgaussian Random Variable). *The  $\psi_2$ -norm of a subgaussian random variable  $x \in \mathbb{R}$ , denoted by  $\|x\|_{\psi_2}$  is:  $\|x\|_{\psi_2} = \inf \{t > 0 : \mathbb{E}[\exp(|x|^2/t^2)] \leq 2\}$ .*

Note that the  $\psi_2$  condition on a scalar random variable  $x$  is equivalent to the subgaussian tail decay of  $x$ .

**Johnson-Lindenstrauss (JL) Transformations.** In the following few paragraphs, we will review some useful facts about randomized dimension reduction using Johnson-Lindenstrauss transformation. Johnson-Lindenstrauss (JL) transformation is a low-dimensional embedding which preserves, up to a small distortion, pairwise  $\ell_2$ -distances between vectors according to the JL lemma.

**Lemma B.2** (JL Lemma). *For any  $0 < \gamma, \beta < 1/2$  and positive integer  $d$ , there exists a distribution  $\mathcal{D}$  over  $\mathbb{R}^{m \times d}$  for  $m = O(\log(1/\beta)/\gamma^2)$  such that for any  $\mathbf{a} \in \mathbb{R}^d$ ,  $\Pr_{\Phi \sim \mathcal{D}}[\|\Phi \mathbf{a}\|^2 - \|\mathbf{a}\|^2 \geq \gamma \|\mathbf{a}\|^2] \leq \beta$ .*

The original proof of the JL lemma chose  $\Phi$  as a scaled projection onto a random  $m$ -dimensional linear subspace, whereas subsequent works showed that the entries of  $\Phi$  can be i.i.d. subgaussian random variables.

A simple consequence of the JL lemma is that, for any  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ , the inner-product between  $\mathbf{a}$  and  $\mathbf{b}$  is approximately preserved under these transformations, in that, if  $m = \Omega(\log(1/\beta)/\gamma^2)$ ,

$$\Pr[|\langle \Phi \mathbf{a}, \Phi \mathbf{b} \rangle - \langle \mathbf{a}, \mathbf{b} \rangle| \geq \gamma \|\mathbf{a}\| \|\mathbf{b}\|] \leq \beta. \quad (11)$$

Using (11) and a net argument<sup>9</sup> over the set of sparse vectors gives the following standard fact.

**Proposition B.3.** *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  $m = \Theta(k \log(d/\beta)/\gamma^2)$  and  $\Phi = \Psi/\sqrt{m}$ . Then for any fixed set of vectors  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ , with  $n \leq d$ , we have,*

$$\Pr[|\langle \Phi \mathbf{x}_i, \Phi \theta \rangle - \langle \mathbf{x}_i, \theta \rangle| \geq \gamma \|\mathbf{x}_i\| \|\theta\| \text{ for all } i \in [n], \theta \in \Sigma_k] \leq \beta.$$

## C Comparison of the Stable Rank and the Restricted Eigenvalue Condition

In this section, we investigate the relationship between the stable rank condition that we used in Theorem 3.1 and the standard restricted eigenvalue (RE) condition commonly used in the analysis of Lasso [7]. The

<sup>9</sup>Let  $\mathcal{N}$  be an  $\varepsilon$ -net over  $\Sigma_k$  (set of  $k$ -sparse vectors in  $B_2^d$ ). Using Proposition B.1,  $|\mathcal{N}| \leq \binom{d}{k} \cdot \left(\frac{3}{\varepsilon}\right)^k = O\left(\frac{d}{\varepsilon}\right)^k$ .

picture that appears is as follows: (a) the stable rank condition on  $X$  is a less restrictive<sup>10</sup> condition than a RE condition on  $X$ , (b) the stable rank condition on  $X^\top X$  (as we have in Theorem 3.1) appears incomparable with a RE condition on  $X$ , and (c) in most settings of concern for sparse regression,  $\text{sr}(X^\top X)$  approximately equals  $\text{sr}(X)$ .

**Stable Rank on  $X$  vs. RE Condition.** 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

$$X = \begin{pmatrix} \mathbb{I}_k & 0 \\ 0 & 0 \end{pmatrix},$$

where  $\mathbb{I}_k$  is the identity  $k \times k$  matrix. Then,  $\text{sr}(X^\top X) = \text{sr}(X) = k$ , while the RE condition does not hold for  $X$ . This shows that there exist families of matrices for which a non-trivial stable rank condition will hold, but no RE condition is possible.

To make the comparison in the other direction, we need an additional normalization of  $X$ , as  $\text{sr}(X)$  is invariant under scaling, and  $\text{RE}(X, k, \alpha)$  is degree 1 homogenous (in that scaling each element in  $X$  by a factor  $c$  changes  $\text{RE}(X, k, \alpha)$  by  $c$ ). Assume that  $\text{RE}(X, k, \alpha) \geq r$  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  $\text{RE}(X, k, \alpha) \geq r$  in derivation of the vector reconstruction conditions (see, e.g. [16]). These assumptions yield that

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

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

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

This shows that always a RE condition on  $X$  implies a non-trivial stable rank condition on  $X$ . Putting both these directions together, implies that while a RE bound always translates into stable rank bound, the other direction does not hold.

**Stable Rank on  $X^\top X$  vs. RE Condition.** Doing an exact comparison of the RE condition with  $\text{sr}(X^\top X)$  is trickier, since for a general matrix, there are no relation between  $\text{sr}(X^\top X)$  number and  $\text{sr}(X)$  besides the latter is larger, i.e.,

$$\text{sr}(X^\top X) \leq \text{sr}(X) \leq \text{rank}(X). \quad (12)$$

<sup>10</sup>In that a larger class of matrices satisfy it.

However, for many interesting classes of matrices  $\text{sr}(X^\top X)$  approximately equals  $\text{sr}(X)$ . For example, if  $X$  is an  $n \times d$  random matrix with independent centered subgaussian entries of unit variance, and  $d \geq 2n$ , then with high probability, all three terms in (12) are of the same order. Indeed, let  $s_1(X) \geq \dots \geq s_n(X) \geq 0$  be the singular values of  $X$ . Then,

$$\text{sr}(X^\top X) = \frac{1}{s_1(X)^4} \sum_{j=1}^n s_j(X)^4.$$

It is a standard fact [20] that with high probability the least singular value of  $X$  is  $\Omega(s_1(X))$ . Therefore, in this case, with high probability,  $\text{sr}(X^\top X) = \Theta(\text{sr}(X))$ .

Even in a non-random setting, for matrices  $X$  generally used in sparse reconstruction problems, it is reasonable to assume that  $n$  and  $\text{sr}(X)$  are of the same order, since otherwise, one can obtain a good approximation of  $X^\top X$  by randomly sampling  $O(\text{sr}(X) \log(\text{sr}(X)))$  rows of  $X$  (see, e.g., [14, Theorem 1.1]). Under this additional assumption,  $\text{sr}(X^\top X)$  and  $\text{sr}(X)$  are again comparable. Indeed, define  $\sigma_j = s_j^2(X)/s_1^2(X)$ . If  $n \leq \rho \text{sr}(X)$  for some  $\rho \in \mathbb{R}^+$ , then by the Cauchy-Schwarz inequality,

$$\text{sr}(X^\top X) = \sum_{j=1}^n \sigma_j^2 \geq \frac{1}{n} \left( \sum_{j=1}^n \sigma_j \right)^2 = \frac{1}{n} \text{sr}(X)^2 \geq \frac{1}{\rho} \text{sr}(X).$$

These above examples illustrate that in many common settings of  $X$ ,  $\text{sr}(X^\top X)$  is comparable to  $\text{sr}(X)$ . This along with our previous discussion about the relation between  $\text{sr}(X)$  and RE shows that a stable rank assumption on  $\text{sr}(X^\top X)$  is reasonable and probably practically even less restrictive than a RE assumption on  $X$ .

## D Proof of Proposition 3.3

The following Hoeffding bound will be useful in our analysis.

**Proposition D.1** (Hoeffding Bound). *Suppose that the variables  $x_i$ ,  $i = 1, \dots, n$  are independent, and  $x_i$  has mean  $\mu_i$  and  $\|x_i\|_{\psi_2} \leq \sigma_i$ . Then for all  $t \geq 0$ , we have*

$$\Pr \left[ \sum_{i=1}^n (x_i - \mu_i) \geq t \right] \leq \exp \left( \frac{-t^2}{2 \sum_{i=1}^n \sigma_i^2} \right).$$

**Proposition D.2** (Proposition 3.3 Restated). *Let  $(X, \mathbf{y}, \theta^*)$  be  $(k, \sigma)$ -well behaved. 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  $m = \Theta(k \log(d/\beta))$  and  $\Phi = \Psi/\sqrt{m}$ . Then with probability at least  $1 - \beta$ ,*

$$\frac{\|(X\Phi^\top \Phi)^\top \tilde{\mathbf{w}}\|_\infty}{n} = O \left( \frac{\sigma \|X\|_F \log(d/\beta)}{n\sqrt{m}} + \frac{\|X\|_F^2}{n\sqrt{d}} \right).$$

*Proof.* Let  $\mathbf{w} = (w_1, \dots, w_n)$  and  $\mathbf{y} = (y_1, \dots, y_n)$ . By Definition 3,  $w_i = y_i - \langle \mathbf{x}_i, \theta^* \rangle$ . Let  $\tilde{\mathbf{w}} = \mathbf{y} - X\Phi^\top \Phi \theta^*$ . Therefore, by invoking Proposition B.3 with  $\gamma = O(1)$  provides that with probability at least

$1 - \beta$ ,

$$\begin{aligned}
\frac{\|(X\Phi^\top\Phi)^\top\tilde{\mathbf{w}}\|_\infty}{n} &= \left\| \frac{1}{n} \sum_{i=1}^n \Phi^\top\Phi\mathbf{x}_i (y_i - \langle\Phi\mathbf{x}_i, \Phi\theta^*\rangle) \right\|_\infty \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n \Phi^\top\Phi\mathbf{x}_i (y_i - \langle\mathbf{x}_i, \theta^*\rangle \pm \|\mathbf{x}_i\|\|\theta^*\|) \right\|_\infty \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \Phi^\top\Phi\mathbf{x}_i (y_i - \langle\mathbf{x}_i, \theta^*\rangle \pm \|\mathbf{x}_i\|) \right\|_\infty \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \Phi^\top\Phi\mathbf{x}_i (w_i \pm \|\mathbf{x}_i\|) \right\|_\infty \\
&= \max_{j \in [d]} \left\{ \left| \frac{1}{n} \langle \mathbf{c}_j, \bar{\mathbf{w}} \rangle \right| \right\}
\end{aligned} \tag{13}$$

where  $\mathbf{c}_j = (c_{j1}, \dots, c_{jn})$  is the  $j$ th column in  $X\Phi^\top\Phi$ , and  $\bar{\mathbf{w}} = (\bar{w}_1, \dots, \bar{w}_n)$  with  $\bar{w}_i = w_i \pm \|\mathbf{x}_i\|$ . Note that we used Proposition B.3 for the first inequality.

We now bound the term in the right-hand side of (13). For a fixed  $j$ , using Proposition D.1 on the set of subgaussian variables  $c_{j1}\bar{w}_1, \dots, c_{jn}\bar{w}_n$  gives that

$$\Pr \left[ \sum_{i=1}^n (c_{ji}\bar{w}_i - \|\mathbf{x}_i\|) \geq t \right] \leq \exp \left( \frac{-t^2}{2\sigma^2\|\mathbf{c}_j\|^2} \right).$$

Taking a union bound over  $j \in [d]$ ,

$$\Pr \left[ \max_{j \in [d]} \left\{ \sum_{i=1}^n (c_{ji}\bar{w}_i - \|\mathbf{x}_i\|) \right\} \geq t \right] \leq d \exp \left( \frac{-t^2}{2\sigma^2\|\mathbf{c}_j\|^2} \right). \tag{14}$$

We now investigate the norm of  $\mathbf{c}_j$ . Let  $\phi_j$  be the  $j$ th column in  $\Phi$ . Now, the  $i$ th entry in  $\mathbf{c}_j \in \mathbb{R}^n$ , can be expressed as  $c_{ji} = \langle\Phi\mathbf{x}_i, \phi_j\rangle$  (i.e., the  $(i, j)$ th entry of  $X\Phi^\top\Phi$  equals  $\langle\Phi\mathbf{x}_i, \phi_j\rangle$ ). With the choice of  $m$ , with probability at least  $1 - \beta$ ,  $\|\Phi\mathbf{x}_i\| = O(\|\mathbf{x}_i\|)$  for all  $i \in [n]$ , where the first expression follows from the norm preservation property of JL-style transform (Lemma B.2). Using definition of subgaussian random variables yields that with probability at least  $1 - \beta$ , for each  $j \in [d]$ ,  $\|\mathbf{c}_j\| = O(\|X\|_F \sqrt{\log(d/\beta)/m})$ . Using this bound in (14) and setting  $t = O(\sigma\|X\|_F \log(d/\beta)/\sqrt{m})$ , and proper conditioning, gives that with probability at least  $1 - \beta$ ,

$$\max_{j \in [d]} \{ \langle \mathbf{c}_j, \bar{\mathbf{w}} \rangle \} = O \left( \frac{\sigma\|X\|_F \log(d/\beta)}{\sqrt{m}} + \sum_{i=1}^n \|\mathbf{x}_i\| \right) = O \left( \frac{\sigma\|X\|_F \log(d/\beta)}{\sqrt{m}} + \frac{\|X\|_F^2}{\sqrt{d}} \right).$$

Plugging this bound into (13) gives the claimed result.  $\square$