New Bounds on Compressive Linear Least Squares Regression

Ata Kabán
School of Computer Science, The University of Birmingham, Edgbaston, B15 2TT

Abstract

In this paper we provide a new analysis of compressive least squares regression that removes a spurious $\log N$ factor from previous bounds, where N is the number of training points. Our new bound has a clear interpretation and reveals meaningful structural properties of the linear regression problem that makes it solvable effectively in a small dimensional random subspace. In addition, the main part of our analysis does not require the compressive matrix to have the Johnson-Lindenstrauss property, or the RIP property. Instead, we only require its entries to be drawn i.i.d. from a 0-mean symmetric distribution with finite first four moments.

1 Introduction

The requirement of dealing with high dimensional massive data sets has motivated a lot of interest in applying recent advances in random projections, compressed sensing and related areas. Such techniques allow us to gain computationally feasible statistical learning methods at the expense of a controlled loss in performance. Indeed, the prospect of carrying out data mining on compressive versions of the data holds much potential and promise. However, in order to be able to make full and informed use of these computationally cheap methods of dimensionality reduction, we need to develop a better understanding of what conditions are required and what guarantees we can get from these techniques specifically for data mining and machine learning tasks.

Much of the theoretical guarantees on compressive regression [12, 14] and classification [1, 4] have been derived by employing results from Johnson-Lindenstrauss embeddings [6] and from Compressed Sensing [5] as building blocks. At present it is not at all clear if the conditions that were needed for those results are still required for guarantees on learning from compressive data. The Johnson-Lindenstrauss embedding of a data set requires conditions that ensure that the Euclidean distances between all pairs of points are preserved within a small distortion after projection. Compressed Sensing requires conditions that ensure that the high dimensional data can be recovered from just a few of its random projections. Do we need the same (strong) conditions to ensure guarantees on statistical learning tasks carried out in a random subspace? Some recent research on compressed linear classification [7, 11, 8] found that less is sufficient. This is perhaps not that surprising – intuitively, indeed in classification not all distances are important, and not all details of the data matter.

But how about linear regression? In regression the targets are real-valued, and one might feel that a global preservation of the geometry of the training set might be needed. Prior work on bounding the excess risk of compressive ordinary least squares regression [12, 14] has certainly built on that premise. However, in this paper we show that improved bounds may be obtained on a more direct route, from first principles. Our improved bounds on randomly projected ordinary linear least squares (OLS) regression remove a spurious logarithmic factor, and show that the bias term of compressive OLS does not depend on the training set size at all. In developing this result we also obtain a generalisation of a random matrix theory result by [13], which may be of independent interest.

1.1 Preliminaries

We consider ordinary linear least squares regression in the fixed design setting. Given a set of N input-target pairs $S = \{(x_1, y_1), ..., (x_N, y_N)\}$ where $x_n \in \mathbb{R}^d, y_n \in \mathbb{R}, n = 1, ..., N$, the goal is to learn an estimator v so that $x_n^T v$ approximates $x_n^T w$, under the linear model assumption:

$$y_n = x_n^T w + \xi_n, n = 1, ..., N \qquad (1)$$

Appearing in Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

where ξ_n is i.i.d. 0-mean noise with variance γ.

The fixed design setting means that the inputs (covariates) $x_n, n = 1, ..., N$ are non-random, with only the targets (responses) $y_1, ..., y_n$ being treated as random variables. This is the simplest setting and it is suitable for studying dimensionality reduction techniques [10], which is our purpose. It should be noted that the fixed design setting does not address out-of-sample prediction because with fixed inputs the estimated regression vector is an unbiased estimate of the minimiser of the regression objective whereas with random inputs it would be not. However, as noted in [10], by conditioning on the inputs, many results extend from the fixed design setting to the random design setting under some more elaborated conditions.

The square loss of an estimator v is defined as:

$$L(v) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}[(y_n - x_n^T v)^2] = \frac{1}{N}\mathrm{E}[\|Y - X^T v\|^2] \quad (2)$$

where Y denotes the column vector of $y_n, n = 1, ..., N$ and X is the $d \times N$ matrix with columns $x_n, n = 1, ..., N$. The expectations are with respect to Y throughout, unless indicated otherwise.

Denote by w the true minimiser of the square loss:

$$w = \arg\min_u L(u) \quad (3)$$

The excess risk of an estimator v is defined as:

$$R(v) = L(v) - L(w) \quad (4)$$

The empirical square loss of an estimator v is the following:

$$\hat{L}(v) = \frac{1}{N}\|Y - X^T v\|^2 \quad (5)$$

The ordinary least square (OLS) estimator is the minimiser of the empirical square loss:

$$\hat{w} = \arg\min_u \hat{L}(u) \quad (6)$$

We will make use of the following well known result about the expected excess risk of the OLS estimator in the fixed design setting ([9], Thm 11.1.):

Lemma 1 *Let $\gamma = Var(y_i)$ $(i = 1, ..., N)$, $\Sigma = XX^T/N$ fixed and invertible, w the optimal OLS as above, and $\hat{w}$ the OLS estimator. Then the expected risk $E[R(\hat{w})]$ equals:*

$$E[L(\hat{w})] - L(w) = \gamma\frac{d}{N} \quad (7)$$

where the expectation is w.r.t. $\hat{w}$ that is a function of the random vector Y.

1.2 Prior work and motivation

From Lemma 1 it is obvious that the expected excess risk of OLS grows linearly with d. Hence when d is large and N is small compared to d then OLS becomes poor. When $d > N$, then Σ is not invertible and OLS is not applicable at all. A common approach to overcome these problems is to use a ridge regression estimator instead, which is obtained by minimising a regularised version of the loss, $\hat{L}_{ridge}(v) = \hat{L}(v) + \lambda\|v\|^2$, yielding a regularised covariance $\Sigma + \lambda I_d$ to work with. This $d \times d$ matrix needs to be inverted to obtain the estimator v. An alternative is to apply some dimensionality reduction prior to OLS. Taking the reduced dimension to be less than that of the span of the training input points we ensure that the compressed OLS will have a unique solution.

A computationally attractive dimensionality reduction technique that is also amenable to analysis is random projections. The tandem of random projections plus OLS was indeed put forth in [12, 14]. In [12], the Johnson-Lindenstrauss lemma (JLL) guarantees are used to ensure that the mismatch between the predictions of the compressive-OLS and those of the data-space OLS are close enough on all training points. This requires of the order $k \in \mathcal{O}(\epsilon^{-2}\log(N))$ dimensions for the reduced space, where ϵ controls the allowed distortion in the pairwise Euclidean distances after projection. In [14], the JLL based argument is replaced by that of compressed sensing (CS) [5], which ensures the same approximate global preservation of pairwise distances between sparse vectors via the restricted isometry (RIP) property of the random matrix used for the linear compression. However this works only provided that the input points have a sparse representation with at most s nonzero entries each. Then, from RIP we have the guarantee of pairwise distance preservation independently of N, for $k \in \mathcal{O}(s\log(d))$. The impressive application in [14] to music similarity prediction from a million-dimensional data sets [14] clearly outperformed OLS in the original data space.

The JLL-based approach in [12] is intended to be a worst-case analysis: The required k ensures that all dot-products are approximately preserved so of course the mismatch of the in-sample predictions in the two spaces is controlled. But do we really need to ensure that all dot-products are approximately preserved in order to achieve this? Put in a slightly different way, in what conditions would the linear regression problem be solvable in a smaller dimensional random subspace than the one required by the JLL-based analysis?

One might then attempt to speculate based on the CS-based argument in [14] whether sparsity of the inputs is perhaps a fortuitous structure that makes the regression problem easier? Note however that the requirement of a subspace of dimension $k \in \mathcal{O}(s\log(d))$ in [14], along with the requirement of sparsity of the input data, are conditions that are just simply inherited from the Compressed Sensing literature, and are in fact sufficient to recover each data point exactly from their random projections. This again leaves the question open, as to whether a linear regression task would still need the same?

A natural conjecture that we will make more formal shortly is that there should be a more direct and problem-specific characterisation of what makes a linear regression problem solvable in a small dimensional random subspace. It was in fact already noted in [14] that the cases where compressive OLS was experimentally observed to be particularly effective are not predicted by the currently existing theory. The remainder of this paper aims to elucidate this open issue.

2 Main Result

Let k be the dimension of a randomly oriented subspace that we project our input points to, which we take as $k < N, k < d$. Let R be the $k \times d$ random projection matrix with entries drawn i.i.d. from a zero mean distribution with variance $1/k$ and finite fourth moment. We are interested in the expected excess risk of OLS that receives only a k-dimensional randomly projected version of the training set $S_R = \{(Rx_1, y_1), ..., (Rx_N, y_N)\}$, where $x_n \in \mathbb{R}^d, Rx_n \in \mathbb{R}^k, y_n \in \mathbb{R}, n = 1, ..., N$.

We will use notations analogous to those defined in the previous section for OLS. To indicate that we now operate in the random subspace defined by R, we use the subscript R. From S_R, we seek to learn an estimator $\hat{w}_R$ so that $x_n^T R^T \hat{w}_R$ approximates $x_n^T w$. For a given R, the square loss of an estimator v_R is:

$$L_R(v_R) = \frac{1}{N}\mathrm{E}_{Y|R}[\|Y - X^T R^T v_R\|^2] \tag{8}$$

The optimal OLS achievable in the random subspace defined by R is:

$$w_R = \arg\min_{u_R} L(u_R) \tag{9}$$

The empirical square loss of an estimator v_R is:

$$\hat{L}_R(v_R) = \frac{1}{N}\|Y - X^T R^T v_R\|^2 \tag{10}$$

and the OLS estimate in the randomly projected space is

$$\hat{w}_R = \arg\min_{u_R} \hat{L}_R(u_R) \tag{11}$$

Finally, our quantities of interest in this section is:

$$\mathrm{E}_{Y,R}[L_R(\hat{w}_R)] - L(w) \tag{12}$$

We will prove the following upperbound on the expected excess risk, where the expectation is over both Y and R.

Theorem 1 ***(Expected excess risk bound on compressive OLS)*** *Let $\gamma = Var(y_i)$, and $\Sigma = XX^T/N$ fixed. Let w be the optimal OLS in $\mathbb{R}^d$, and $\hat{w}_R \in \mathbb{R}^k$, $k < d, k < N$ the OLS estimator in the random projection space $\mathbb{R}^k$ defined by the $k \times d$ random matrix R with entries drawn i.i.d. from a zero-mean symmetric distribution with variance $1/k$ and excess kurtosis $\frac{E[R_{ij}^4]}{E[R_{ij}^2]^2} - 3 = \kappa$. Then,*

$$E_{R,Y}[L_R(\hat{w}_R)] - L(w) \leqslant \gamma\frac{k}{N} + \frac{1}{k} \cdot \|w\|^2_{\Sigma+(1+\kappa)Tr(\Sigma)I_d} \tag{13}$$

where $\|u\|_M = \sqrt{u^T M u}$ stands for the Mahalanobis norm, and I_d is the d-dimensional identity matrix.

The first term on the r.h.s. is the variance of the estimator. Of course this is greatly reduced in comparison with the data space, where it was $\gamma \cdot d/N$. The second term is the expected bias of the estimator, which is the price for the reduced variance.

We see from eq. (13) that the variance term is smallest when k is small – this term represents the expected excess risk of $\hat{w}_R$ with respect to the best achievable in the reduced space i.e. w_R. In turn, the second term, the bias, is smallest when k is large – this term makes the relation back to the original data space – and we now see clearly from the form of this term that a norm of the best OLS in the data space i.e. $\|w\|^2_{\Sigma+(1+\kappa)Tr(\Sigma)I_d}$ is the quantity that governs the compressibility of the working space. More specifically, if the linear regression problem in the original space has its best OLS regressor w with a small (Mahalanobis) norm then we can effectively work with a small k. On the other hand if the best w has a large norm then compressing to a small k will no longer guarantee a low excess risk for the problem at hand. In practice of course w is unknown, but it is theoretically pleasing to have a characterisation of problem compressibility in terms of the specific problem structure.

Before starting the proof, let us point out that the main difference from the bound obtained in [12] is that our bias term in the above eq. (13) is independent of N. The proof technique in [12] brings a spurious factor $\mathcal{O}(\log N)$ into this term, which leaves the interpretation of the overall bound unclear. This spurious factor comes from the union bound after N applications of JLL for dot-products.

A second difference is of course that our result in Theorem 1 is stated in expectation whereas those in the mentioned previous works are given with high probability (w.r.t. the random choice of R). As it turns out, bounding the excess risk in expectation allows us much more generality in the choice of the distribution of the entries of R, as well as a tighter bounding. However, high probability bounds of the same order also hold when the entries of R are subgaussian, these are given in Sec. 4.

3 Proof of Theorem 1

We start by applying Lemma 1 in $\mathbb{R}^k$, which yields:

$$\mathrm{E}_{Y|R}[L_R(\hat{w}_R)] - L_R(w_R) = \gamma\frac{k}{N} \tag{14}$$

This is the variance of the compressive estimator, which is much reduced. The price to pay is that the expected risk in eq. 14 is w.r.t the best achievable in $\mathbb{R}^k$ rather than in $\mathbb{R}^d$, so next we need to bound $L_R(w_R)$ with an expression that contains $L(w)$. By definition we have the inequality:

$$\begin{aligned}
L_R(w_R) &\leqslant L_R(Rw) && (15)\\
&= \frac{1}{N}\mathrm{E}_{Y|R}[\|Y - X^T R^T R w\|^2] && (16)\\
&= \frac{1}{N}\mathrm{E}_{Y|R}[\|Y - X^T w\|^2]... &&\\
&+ \frac{1}{N}\|X^T w - X^T R^T R w\|^2 && (17)\\
&= L(w) + \frac{1}{N}\|X^T w - X^T R^T R w\|^2 &&
\end{aligned}$$

where the decomposition in eq. (17) can be verified by elementary algebra. Hence, so far we have:

$$\mathrm{E}_{Y|R}[L_R(\hat{w}_R)] - L(w) \leqslant \gamma\frac{k}{N} + \frac{1}{N}\|X^T w - X^T R^T R w\|^2 \tag{18}$$

Remains to bound the last term in eq. (18), namely:

$$\frac{1}{N}\|X^T w - X^T R^T R w\|^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n^T w - x_n^T R^T R w)^2 \tag{19}$$

From this point, our analysis will deviate from previous techniques. Since this term contains dot products of randomly projected vectors, the approach in previous works [12, 14] was to use the approximate preservation of dot products under random projections, and require the conditions that are needed for all N dot products to be approximately preserved. As already mentioned, unfortunately this needs either k to grow as $\mathcal{O}(\log N)$ – cf. the JLL-based approach – or it needs sparsity to be imposed on the data points – cf. the compressed sensing based approach. We do not wish to impose sparsity on the data inputs because that would lose generality. Observe the expression in eq. 19 is a sum of dependent random variables where all terms depend on the same R. Although for this very reason we cannot expect a concentration bound to decay with N, there is no reason for it to increase with N either.

We will simply compute $\frac{1}{N}\mathrm{E}_R[\|X^T w - X^T R^T R w\|^2]$. Expanding, we get:

$$\begin{aligned}
&\mathrm{E}_R[w^T\frac{XX^T}{N}w + w^T R^T R\frac{XX^T}{N}R^T R w - 2w^T\frac{XX^T}{N}R^T R w]\\
&\quad = w^T\Sigma w + w^T\mathrm{E}_R[R^T R\Sigma R^T R]w - 2w^T\Sigma\mathrm{E}_R[R^T R]w
\end{aligned}$$

Observe that $\mathrm{E}_R[R^T R] = I_d$ since we had the entries of R have mean zero and variance $1/k$. Hence, after cancellations we get:

$$-w^T\Sigma w + w^T\mathrm{E}_R[R^T R\Sigma R^T R]w \tag{20}$$

A matrix expectation of the form that appears in eq. (20) has been studied in random matrix theory for handling singular sample covariances in [13] for the special case when R is a Haar distributed, i.e. it has orthonormal rows. Here we will obtain a generalisation for the case when the random matrix R has entries drawn i.i.d. from a 0-mean symmetric distribution with finite first four moments. The following lemma computes the expectation $\mathrm{E}_R[R^T R\Sigma R^T R]$ for such R, which turns out to have a closed form, and may be of independent interest.

Lemma 2 *Let R be a $k \times d$ random matrix, $k < d$, with entries drawn i.i.d. from a symmetric distribution with 0-mean and finite first four moments. Let Σ be a $d \times d$ fixed positive semi-definite matrix with eigenvalues $\lambda_1, ..., \lambda_d$. Then,*
$E[R^T R\Sigma R^T R] = ...$

$$k\cdot E[R_{ij}^2]^2\left[(k+1)\Sigma + Tr(\Sigma)I_d + \left(\frac{E[R_{ij}^4]}{E[R_{ij}^2]^2} - 3\right)\sum_{i=1}^{d}\lambda_i A_i\right] \tag{21}$$

where A_i are $d \times d$ diagonal matrices with their j-th diagonal elements being $\sum_{a=1}^{d} U_{ai}^2 U_{aj}^2$, and U_{ai} is the a-th entry of the i-th eigenvector of Σ.

The proof of this lemma is given in Subsection 3.1.

Returning to the proof of Theorem 1 we employ Lemma 2 with $\mathrm{E}[R_{ij}^2] = 1/k$, and note also that by the Cauchy-Schwartz inequality, the diagonal elements of A_i are no greater than one. Hence, the last term in eq (21) is upper-bounded as $\kappa\sum_{i=1}^{d}\lambda_i A_i \preccurlyeq \kappa \cdot Tr(\Sigma)\cdot I_d$, and so we get that eq. (20) is upper bounded by the following:

$$\begin{aligned}
&\leqslant \quad -w^T\Sigma w + \left(1+\frac{1}{k}\right) w^T\Sigma w + \frac{1}{k} w^T w \cdot Tr(\Sigma)... \\
&+ \quad \frac{1}{k} w^T w \cdot \kappa \cdot Tr(\Sigma) \qquad (22)\\
&= \quad \frac{1}{k} w^T(\Sigma + (1+\kappa)Tr(\Sigma)I_d)w \qquad (23)\\
&= \quad \frac{1}{k} \cdot \|w\|^2_{\Sigma+(1+\kappa)Tr(\Sigma)I_d} \qquad (24)
\end{aligned}$$

Summarising this main step of the proof, we obtained w.p. at least $1-\delta$ that:

$$\frac{1}{N}\mathrm{E}[\|X^Tw - X^TR^TRw\|^2] \leqslant \frac{1}{k}\|w\|^2_{\Sigma+(1+\kappa)Tr(\Sigma)I_d} \qquad (25)$$

Finally, plugging eq. (25) back into eq. (18) we obtain the statement of Theorem 2. ■

3.1 Proof of Lemma 2

Take the SVD of $\Sigma = U\Lambda U$, where U is the $d \times d$ matrix of eigenvectors, $UU^T = U^TU = I$, and Λ is the diagonal matrix of eigenvalues.

Denote $P := RU$, and observe that P has the following properties:

$$\begin{aligned}
&\mathrm{E}[P_{ij}] = 0 && (26)\\
&P_{ij} \text{ and } P_{i'j'} \text{ are independent } \forall i \neq i' && (27)\\
&\mathrm{Cov}(P_{ij}, P_{ij'}) = 0, \forall j \neq j' && (28)\\
&-P_{ij} \sim P_{ij} && (29)
\end{aligned}$$

where the last eq. says that P_{ij} and $-P_{ij}$ have the same distribution. In other words, P has 0-mean symmetrically distributed entries, independent rows, and dependent but uncorrelated columns. To see property (28), note first that for $j \neq j'$, both P_{ij} and $P_{ij'}$ are functions of the i-th row of R, so they are dependent. However, their covariance evaluates to zero:

$$\begin{aligned}
\mathrm{Cov}(P_{ij}, P_{ij'}) &= \mathrm{Cov}(\sum_{\ell=1}^d R_{i\ell}U_{\ell j}, \sum_{\ell=1}^d R_{i\ell}U_{\ell j'})\\
&= \sum_{\ell=1}^d \sum_{\ell'=1}^d U_{\ell j}U_{\ell' j'}\mathrm{Cov}(R_{i\ell}, R_{i\ell'})\\
&= \sum_{\ell=1}^d U_{\ell j}U_{\ell j'}\mathrm{Var}(R_{i\ell})\\
&= \mathrm{Var}(R_{i\ell})U_j^TU_{j'} = 0
\end{aligned}$$

since the entries of R are i.i.d., and the columns of U are orthogonal.

Now, rewrite:

$$\begin{aligned}
\mathrm{E}[R^TR\Sigma R^TR] &= \mathrm{E}[R^TRU\Lambda U^TR^TR]\\
&= U\mathrm{E}[U^TR^TRU\Lambda U^TR^TRU]U^T\\
&= U\mathrm{E}[P^TP\Lambda P^TP]U^T \qquad (30)
\end{aligned}$$

so it is sufficient to work with the matrix expectation $\mathrm{E}[P^TP\Lambda P^TP]$ where Λ is diagonal with nonnegative elements.

We will need the individual elements of this matrix, so we start by rewriting: $\mathrm{E}[P^TP\Lambda P^TP] = ...$

$$\sum_{i=1}^{\rho} \lambda_i \begin{bmatrix} \mathrm{E}[(P_1^TP_i)^2] & \dots & \mathrm{E}[(P_1^TP_i)(P_i^TP_d)] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[(P_d^TP_i)(P_i^TP_1)] & \dots & \mathrm{E}[(P_d^TP_i)^2] \end{bmatrix} \qquad (31)$$

where ρ stands for the rank of Σ, λ_i are the diagonal elements of Λ, and P_i is the i-th column of P.

Step 1: We show that $\mathrm{E}[P^TP\Lambda P^TP]$ is diagonal. By implication, the expectation on the l.h.s. of the statement of Lemma 2, i.e. $\mathrm{E}[P^TP\Sigma P^TP]$, has the same eigenvectors as Σ.

Indeed, the off-diagonal entries of the matrix in the i-th summand have the following form:

$$\begin{aligned}
\mathrm{E}[(P_j^TP_i)(P_i^TP_\ell)] &= \mathrm{E}[(\sum_{m=1}^k P_{mi}P_{mj})(\sum_{m'=1}^k P_{m'i}P_{m'\ell})]\\
&= \sum_{m=1}^k \sum_{m'=1}^k \mathrm{E}[P_{mi}P_{mj}P_{m'i}P_{m'\ell}]
\end{aligned}$$

with $j \neq \ell$. Now it is a straightforward matter to go though all the different cases, and verify that the properties in eqs. (26)-(29) imply that this expectation evaluates to zero.

Case $j \neq i \neq \ell, m \neq m'$:

$$\begin{aligned}
\mathrm{E}[P_{mi}P_{mj}P_{m'i}P_{m'\ell}] &= \mathrm{E}[P_{mi}P_{mj}]\mathrm{E}[P_{m'i}P_{m'\ell}]\\
&= 0 \quad \{\text{by eq. (28)}\}
\end{aligned}$$

Case $j \neq i \neq \ell, m = m'$:

$$\begin{aligned}
\mathrm{E}[P_{mi}P_{mj}P_{m'i}P_{m'\ell}] &= \mathrm{E}[P_{mi}^2P_{mj}P_{m'\ell}]\\
&= -\mathrm{E}[P_{mi}^2P_{mj}P_{m'\ell}] \quad \{\text{by eq.(29)}\}\\
&= 0
\end{aligned}$$

Case $j = i \neq \ell, m \neq m'$:

$$\begin{aligned}
\mathrm{E}[P_{mi}P_{mj}P_{m'i}P_{m'\ell}] &= \mathrm{E}[P_{mi}^2P_{m'i}P_{m'\ell}]\\
&= \mathrm{E}[P_{mi}^2]\mathrm{E}[P_{m'j}P_{m'\ell}] \ \{\text{by (27)}\}\\
&= 0 \quad \{\text{by eq.(28)}\}
\end{aligned}$$

Case $j = i \neq \ell, m = m'$:

$$\begin{aligned} \mathrm{E}[P_{mi}P_{mj}P_{m'i}P_{m'\ell}] &= \mathrm{E}[P_{mi}^3 P_{m\ell}] \\ &= -\mathrm{E}[P_{mi}^3 P_{m\ell}] \quad \{\text{by eq.(29)}\} \\ &= 0 \end{aligned}$$

and the remaining cases $j \neq i = \ell, m \neq m'$, and $j \neq i = \ell, m = m'$ are equivalent to the previous two, by symmetry.

Hence indeed, all the off-diagonal entries of all the summands in eq. (31) are zero, so the matrix $\mathrm{E}[P^TP\Lambda P^TP]$ is diagonal.

Step 2: Now we compute the diagonal entries of $\mathrm{E}[P^TP\Lambda P^TP]$. There are two cases to consider:

Case $i = j$:

$$\begin{aligned} \mathrm{E}[(P_i^TP_i)^2] &= \mathrm{E}\left[\left(\sum_{\ell=1}^k P_{\ell i}^2\right)^2\right] = \sum_{\ell=1}^k \sum_{\ell'=1}^k \mathrm{E}[P_{\ell i}^2 P_{\ell' i}^2] \\ &= \sum_{\ell=1}^k \sum_{\ell'=1,\ell'\neq\ell}^k \mathrm{E}[P_{\ell i}^2]\mathrm{E}[P_{\ell' i}^2] + \sum_{\ell=1}^k \mathrm{E}[P_{\ell i}^4] \end{aligned}$$

by eq. (27). Taking aside the term $\mathrm{E}[P_{\ell i}^2]$, we have:

$$\begin{aligned} \mathrm{E}[P_{\ell i}^2] &= \mathrm{E}[(R_\ell \cdot U_i)^2] = \mathrm{E}\left[\left(\sum_{a=1}^d R_{\ell a}U_{ai}\right)^2\right] \\ &= \sum_{a=1}^d \sum_{a'=1}^d U_{ai}U_{a'i}\mathrm{E}[R_{\ell a}R_{\ell a'}] \\ &= \sum_{a=1}^d U_{ai}^2 \mathrm{E}[R_{\ell a}^2] \quad \{\text{entries of } R \text{ are i.i.d.}\} \\ &= \sigma^2\|U_i\|^2 = \sigma^2 \quad \{\text{as } \|U_i\| = 1\} \end{aligned}$$

Replacing,

$$\mathrm{E}[(P_i^TP_i)^2] = (k^2-k)\sigma^4 + k\mathrm{E}[P_{\ell i}^4] \tag{32}$$

and remains to compute the expectation in the last term: $\mathrm{E}[P_{\ell i}^4] = ...$

$$\begin{aligned} &= \mathrm{E}\left[\left(\sum_{a=1}^d R_{\ell a}U_{ai}\right)^4\right] \\ &= \sum_{a=1}^d \sum_{a'=1}^d \sum_{b=1}^d \sum_{b'=1}^d U_{ai}U_{a'i}U_{bi}U_{b'i}\mathrm{E}[R_{\ell a}R_{\ell a'}R_{\ell b}R_{\ell b'}] \\ &= \sum_{a=1}^d U_{ai}^4 \mathrm{E}[R_{\ell a}^4] + 3\sum_{a=1}^d \sum_{a'=1,a'\neq a}^d U_{ai}^2U_{a'i}^2\mathrm{E}[R_{\ell a}^2R_{\ell a'}^2] \end{aligned}$$

by eq. (29). Now denote $\mu_4 = \mathrm{E}[R_{\ell a}^4], \forall \ell = 1, ..., k, a = 1, ..., d$, and $\sigma^2 = \mathrm{E}[R_{\ell a}^2]$, and rearrange, this further equals:

$$\begin{aligned} &= \mu_4 \sum_{a=1}^d U_{ai}^4 + 3\sigma^4 \sum_{a=1}^d \sum_{a'=1,a'\neq a}^d U_{ai}^2U_{a'i}^2 \\ &= (\mu_4 - 3\sigma^4)\sum_{a=1}^d U_{ai}^4 + 3\sigma^4 \sum_{a=1}^d \sum_{a'=1}^d U_{ai}^2U_{a'i}^2 \\ &= (\mu_4 - 3\sigma^4)\sum_{a=1}^d U_{ai}^4 + 3\sigma^4\|U_i\|^4 \\ &= (\mu_4 - 3\sigma^4)\sum_{a=1}^d U_{ai}^4 + 3\sigma^4 \quad \{\text{since } \|U_i\| = 1\} \end{aligned}$$

Finally, we plug this back into eq. (32), which gives:

$$\begin{aligned} \mathrm{E}[(P_i^TP_i)^2] &= (k^2-k)\sigma^4 + k(\mu_4 - 3\sigma^4)\sum_{a=1}^d U_{ai}^4 \\ &+ 3\sigma^4 k \\ &= k\sigma^4\left(k - 1 + \left(\frac{\mu_4}{\sigma^4} - 3\right)\sum_{a=1}^d U_{ai}^4 + 3\right) \\ &= k\sigma^4\left(k + 2 + \left(\frac{\mu_4}{\sigma^4} - 3\right)\sum_{a=1}^d U_{ai}^4\right) \end{aligned}$$

Case $i \neq j$:

$$\begin{aligned} &\mathrm{E}[(P_i^TP_j)^2] = ... \\ &= \mathrm{E}\left[\left(\sum_{\ell=1}^k P_{\ell i}P_{\ell j}\right)^2\right] = \sum_{\ell=1}^k \sum_{\ell'=1}^k \mathrm{E}[P_{\ell i}P_{\ell j}P_{\ell' i}P_{\ell' j}] \\ &= \sum_{\ell=1}^k \sum_{\ell'=1,\ell'\neq\ell}^k \mathrm{E}[P_{\ell i}P_{\ell j}]\mathrm{E}[P_{\ell' i}P_{\ell' j}] + \sum_{\ell=1}^k \mathrm{E}[P_{\ell i}^2P_{\ell j}^2] \\ &= \sum_{\ell=1}^k \mathrm{E}[P_{\ell i}^2P_{\ell j}^2] \ \{\text{by eq. (28)}\} \end{aligned} \tag{33}$$

Calculating $\mathrm{E}[P_{\ell i}^2P_{\ell j}^2]$ we get:

$$\begin{aligned} &\mathrm{E}[P_{\ell i}^2P_{\ell j}^2] = ... \\ &= \mathrm{E}\left[\left(\sum_{a=1}^d R_{\ell a}U_{ai}\right)^2\left(\sum_{a'=1}^d R_{\ell a'}U_{a'j}\right)^2\right] \\ &= \sum_{a=1}^d \sum_{a'=1}^d \sum_{b=1}^d \sum_{b'=1}^d U_{ai}U_{a'i}U_{bj}U_{b'j}\mathrm{E}[R_{\ell a}R_{\ell a'}R_{\ell b}R_{\ell b'}] \\ &= \sum_{a=1}^d U_{ai}^2U_{aj}^2\mathrm{E}[R_{\ell a}^4]... \\ &+ \sum_{a=1}^d \sum_{a'=1,a'\neq a}^d U_{ai}^2U_{a'j}^2\mathrm{E}[R_{\ell a}^2]\mathrm{E}[R_{\ell a'}^2]... \\ &+ \sum_{a=1}^d \sum_{a'=1,a'\neq a}^d U_{ai}U_{a'i}U_{aj}U_{a'j}\mathrm{E}[R_{\ell a}^2]\mathrm{E}[R_{\ell a'}^2]... \\ &+ \sum_{a=1}^d \sum_{a'=1,a'\neq a}^d U_{ai}U_{a'i}U_{a'j}U_{aj}\mathrm{E}[R_{\ell a}^2]\mathrm{E}[R_{\ell a'}^2] \end{aligned}$$

since all odd moments are 0. This further equals:

$$
\begin{aligned}
&= \mu_4 \sum_{a=1}^{d} U_{ai}^2 U_{aj}^2 + \sigma^4 \sum_{a=1}^{d} \sum_{a'=1,a'\neq a}^{d} U_{ai}^2 U_{a'j}^2 ... \\
&+ 2\sigma^4 \sum_{a=1}^{d} \sum_{a'=1,a'\neq a}^{d} U_{ai} U_{a'i} U_{aj} U_{a'j} \\
&= (\mu_4 - 3\sigma^4) \sum_{a=1}^{d} U_{ai}^2 U_{aj}^2 + \sigma^4 \|U_i\|^4 + 2\sigma^4 (U_i^T U_j)^2 \\
&= (\mu_4 - 3\sigma^4) \sum_{a=1}^{d} U_{ai}^2 U_{aj}^2 + \sigma^4
\end{aligned}
$$

as $U_i^T U_j = 0$ and $\|U_i\| = 1$.

Replacing this into eq. (33), we have that:

$$
\begin{aligned}
\mathrm{E}[(P_i^T P_j)^2] &= k(\mu_4 - 3\sigma^4) \sum_{a=1}^{d} U_{ai}^2 U_{aj}^2 + k\sigma^4 \\
&= k\sigma^4 \left(\left(\frac{\mu_4}{\sigma^4} - 3 \right) \sum_{a=1}^{d} U_{ai}^2 U_{aj}^2 + 1 \right)
\end{aligned}
$$

Step 3 Putting everything together, after some algebra we arrive at the following closed form expression: $\mathrm{E}[P^T P \Lambda P^T P] = ...$

$$
\sigma^4 k \left[(k+1)\Lambda + \mathrm{Tr}(\Lambda) I_d + \left(\frac{\mu_4}{\sigma^4} - 3 \right) \sum_{i=1}^{d} \lambda_i A_i \right] \quad (34)
$$

where A_i is the following diagonal matrix:

$$
A_i = \begin{bmatrix} \sum_{a=1}^{d} U_{ai}^2 U_{a1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{a=1}^{d} U_{ai}^2 U_{ad}^2 \end{bmatrix}
$$

The diagonal matrix obtained in eq. (34) holds the eigenvalues of the matrix expectation of our interest, $\mathrm{E}[R^T R \Sigma R^T R]$. Assembling this with eq. (30) to bring back the eigenvectors we have the final result that concludes the proof: $\mathrm{E}[R^T R \Sigma R^T R] = ...$

$$
\sigma^4 k \left[(k+1)\Sigma + \mathrm{Tr}(\Sigma) I_d + \left(\frac{\mu_4}{\sigma^4} - 3 \right) \sum_{i=1}^{d} \lambda_i A_i \right] \quad (35)
$$

with the diagonal matrices $A_i = \mathrm{Diag}_{j=1}^{d} \left(\sum_{a=1}^{d} U_{ai}^2 U_{aj}^2 \right)$. ■

4 Discussion

4.1 On the distribution of entries of R

Throughout so far, we did not require the random projection matrix R to have the Johnson-Lindenstrauss property, or to have the Restricted Isometry Property. Our analysis suggests that these properties may be stronger than what is needed to carry out linear regression in a random subspace. In fact the distribution of the entries of R is not even required to have a moment generating function for the existence of the matrix expectation that has been at the heart of our proof, and the moments of the entries of R that are higher than four make no difference to the form of this expectation.

We have seen that the matrix expectation involved has a term that depends on the excess kurtosis of the entries of R. For platikurtic distributions the excess kurtosis is negative and so is the last term in eq. (21). Our bound tightens for these distributions. For leptokurtic distributions the excess kurtosis is positive so we may expect that strongly kurtotic distributions are less appropriate to make a good R. This is in remarkable analogy with what is known for Johnson-Lindenstrauss embeddings, where an R with sub-Gaussian entries is known to be a near-isometry whereas an R with heavy-tailed entries is less so.

It may be worth noting that whenever the distribution of the entries of R have zero excess kurtosis then the upper bound on the expected bias term holds with equality since the last term on the r.h.s. of Lemma 2 cancels out. As an obvious example, for a random matrix R with i.i.d. Gaussian entries we have that the excess kurtosis is zero. Unsurprisingly, this makes the random projection step rotationally invariant – that is, the orientation of the training set axes makes no difference. It is interesting to note that there also exist non-Gaussian random matrices that are not rotationally invariant but have entries with zero excess kurtosis. For example, the following sparse random projection matrix that was originally proposed for computational efficiency in [2], also has $\kappa = 0$:

$$
R_{ij} \overset{iid}{\sim} \begin{cases} -\sigma\sqrt{3} & \text{w.p.}1/6 \\ 0, & \text{w.p.}2/3 \\ \sigma\sqrt{3} & \text{w.p.}1/6 \end{cases} \quad (36)
$$

Indeed, one can easily verify that $\mathrm{Var}(R_{ij}) = \sigma^2$, and $\mathrm{E}[R_{ij}^4] = 3\sigma^4$, so $\kappa = 0$. Therefore, our bound on the bias of compressive OLS (and the overall expected risk bound) is exactly the same if we use this sparse R as if we use the Gaussian R.

An example of platicurtic random matrix, also proposed in [2], is to have the entries drawn i.i.d. from {-1,+1} with probability 1/2 each. Then the excess kurtosis is $\kappa = -2$. Both of these random matrices happen to have the Johnson-Lindenstrauss property [2]; however, as pointed out already, the R in our Theorem 1 can be from a much larger class.

4.2 Comparison with previous bounds on RP-OLS, and new bounds with high probability in the case of subgaussian R

Let us contrast eq. (25), i.e. the our bound on the expected bias, with the JLL-based approach used in previous work [12], which yielded to the following:

$$\frac{1}{N}\|X^T w - X^T R^T R w\|^2 \leqslant \|w\|^2 \cdot Tr(\Sigma) \cdot \frac{8}{k}\log\frac{4N}{\delta} \tag{37}$$

with probability $1-\delta$, R must be subgaussian.

We see that our bound eliminated the $\log(N)$ factor and otherwise it has a very similar flavour. In particular, $\|w\|^2_{\Sigma+Tr(\Sigma)I_d} \leqslant 2\|w\|^2 Tr(\Sigma)$ by Hölder inequality.

Of course, eq.(37) is a high probability bound whereas our Theorem 1 is a bound on the expectation of the excess risk. From eq. (13) we can trivially obtain the following h.p. bound by using Markov inequality for the bias term (as the variance term deterministically holds for any choice of R): For any $\delta > 0$, the following holds with probability at least $1-\delta$:

$$\mathrm{E}_{Y|R}[L_R(\hat{w}_R)] - L(w) \leqslant \gamma\frac{k}{N} + \frac{1}{\delta}\cdot\frac{1}{k}\cdot\|w\|^2_{\Sigma+(1+\kappa)Tr(\Sigma)I_d} \tag{38}$$

We have not made an effort to improve the $1/\delta$ dependence under the general conditions of Theorem 1, but in the case of R with i.i.d. subgaussian entries we give the following upper and lower bounds on the bias term – both of these are of the same order as our bound in Theorem 1, and independent of N.

Theorem 2 *Let $X \in \mathbb{R}^{d\times N}, \Sigma = XX^T/N$ fixed, and denote $\rho = rank(\Sigma)$. Let $R \in \mathbb{R}^{k\times d}$ be a random matrix with i.i.d. 0-mean subgaussian entries with variance $1/k$. Then, for any $\delta > 0$, the following upper and lower bounds hold simultaneously w.p. $1-\delta$:*

$$\frac{1}{N}\|X^T w - X^T R^T R w\|^2 \leqslant \ldots$$
$$\left(1+2\sqrt{\frac{2\log(4/\delta)}{k}}\right)\left(\sqrt{\frac{\rho}{k}} + C + \sqrt{\frac{2\log(4/\delta)}{ck}}\right)^2 \|w\|^2\lambda_{\max}(\Sigma)$$
$$-w^T\Sigma w + 4\sqrt{\frac{2\log(4/\delta)}{k}}\|w\|^2\lambda_{\max}(\Sigma)$$
$$\frac{1}{N}\|X^T w - X^T R^T R w\|^2 \geqslant \ldots$$
$$\left(1-2\sqrt{\frac{2\log(4/\delta)}{k}}\right)_+\left(\sqrt{\frac{\rho}{k}} - C - \sqrt{\frac{2\log(4/\delta)}{ck}}\right)_+^2 \|w\|^2\lambda_{\min\neq 0}(\Sigma)$$
$$-w^T\Sigma w - 4\sqrt{\frac{2\log(4/\delta)}{k}}\|w\|^2\lambda_{\min\neq 0}(\Sigma)$$

where C and c are positive constants that only depend on the 'subgaussian norm' of the rows of R; $\lambda_{\max}(\cdot)$ and $\lambda_{\min\neq 0}(\cdot)$ denote the largest and the smallest nonzero eigenvalues respectively, and $(\cdot)_+ = \max(\cdot, 0)$.

Proof of Theorem 2. We prove the first part, and the second part goes analogously. Expanding the l.h.s,

$$w^T\Sigma w + w^T R^T R\Sigma R^T R w - 2w^T\Sigma R^T R w \tag{39}$$

bound the last two terms w.h.p. By Rayleigh quotient,

$$\begin{aligned} w^T R^T R\Sigma R^T R w &\leqslant \|Rw\|^2\lambda_{\max}(R\Sigma R^T) \\ &\leqslant (1+\epsilon_1)\|w\|^2\cdot\left(\sqrt{\frac{\rho}{k}} + C + \frac{\epsilon_2}{\sqrt{k}}\right)^2\lambda_{\max}(\Sigma) \end{aligned} \tag{40}$$

w.p. $1-\exp(-\epsilon_1^2 k/8) - \exp(-c\epsilon_2^2/2)$. In the last line we used one side of the Johnson-Lindenstrauss Lemma (noting that subgaussian random matrices with i.i.d. entries satisfy this), and the known upperbound on the largest singular value of a random matrix with i.i.d. subgaussian entries [15] (Theorem 5.39). In the latter, C and c are positive constants that only depend on the 'subgaussian norm' of the rows of R [15].

To bound the last term of eq.(39) we use one side of the Johnson-Lindenstrauss Lemma for the inner product, and then Hölder inequality to get:

$$\begin{aligned} -2w^T\Sigma R^T R w &\leqslant -2w^T\Sigma w + 2\epsilon\cdot\|w^T\Sigma\|\cdot\|w\| \\ &\leqslant -2w^T\Sigma w + 2\epsilon\cdot\|w\|^2\lambda_{\max}(\Sigma) \end{aligned} \tag{41}$$

w.p. $1-2\exp(-\epsilon^2 k/8)$. Now, put together eqs (40) and (41) by union bound, then equate each exponentially decaying probability to $\delta/4$ and solve for the corresponding ϵ to get the first part of the statement of Theorem 2.

Analogous arguments involving the other side of JLL and the known lower bound on the smallest singular value of subgaussian random matrices [15] (Theorem 5.39) yield the second part of Theorem 2. ■

5 Conclusions

We gave improved bounds on the excess risk of compressive ordinary least squares regression in the fixed design setting. The new bounds remove a spurious factor of $\log(N)$ from the bias term of compressive OLS, and have a clearer interpretation that reveals the structure of the problem that makes a linear regression task solvable effectively in a random subspace of the data space. As the main technical ingredient, we developed a generalisation of a result of [13] for computing a matrix expectation that was required in our proof, and which may be of independent interest, e.g. in contexts that involve dealing with singular covariances matrices. Our upper bound on the expected excess risk holds for any random projection matrix that has entries drawn i.i.d. from a symmetric distribution with finite first four moments. We also obtained high probability bounds of the same order when the random projection matrix is subgaussian. Regarding the latter, it remains an open question whether it would be possible to relax the subgaussianity assumption.

References

[1] R.Arriaga and S. Vempala, An algorithmic theory of learning: Robust concepts and random projection, Machine Learning, Vol. 63 Issue 2, 2006, pp. 161-182.

[2] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. Journal of Computer and System Sciences 66 (2003) 671-687.

[3] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices. Constructive Approximation, December 2008, Volume 28, Issue 3, pp 253-263.

[4] R. Calderbank, S. Jafarpour, R. Schapire. Compressed Learning: Universal Sparse Dimensionality Reduction and Learning in the Measurement Domain. Technical Report, Rice University, 2009.

[5] E.J. Candès and T. Tao. Decoding by Linear Programming. IEEE Transactions on Information Theory, Vol. 51, Issue 12, pp. 42034215, 2005.

[6] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures and Algorithms, Vol. 22, pp. 6065, 2003.

[7] R.J. Durrant and A. Kabán. Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data. Proc. of the 16-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), pp. 1119-1128, 2010.

[8] R.J. Durrant and A. Kabán. Sharp Generalization Error Bounds for Randomly-projected Classifiers. 30th International Conference on Machine Learning (ICML 2013), Journal of Machine Learning Research-Proceedings Track 28(3):693-701, 2013.

[9] L. Györfi, M. Kohler, A. Krzyzak, H. Walk. A Distribution-Free Theory of Nonparametric Regression. Springer, 2014.

[10] D. Hsu, S.M. Kakade, T. Zhang. Random Design Analysis of Ridge Regression, In Proc. of the 25th Conference on Learning Theory (COLT 12), 2012.

[11] A. Kabán and R.J. Durrant. Dimension-Adaptive Bounds on Compressive FLD Classification. In Proc of the 24th International Conference on Algorithmic Learning Theory (ALT 2013), pp. 294-308.

[12] O. Maillard and R. Munos. Compressed least squares regression. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 22, pages 1213-1221, 2009.

[13] T. Marzetta, G. Tucci, S. Simon. A random matrix theoretic approach to handling singular covariance estimates. IEEE Transactions on Information Theory, Vol. 57, Issue 9, 2011.

[14] M. Fard, Y. Grinberg, J. Pineau, and D. Precup, Compressed least-squares regression on sparse spaces, Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[15] R. Vershynin. Introduction to the Non-Asymptotic Analysis of Random Matrices. Compressed sensing, pp. 210-268, Cambridge University Press, 2012.