A New Look at Compressed Ordinary Least Squares

Ata Kabán
School of Computer Science
The University of Birmingham
Edgbaston, B15 2TT
Birmingham, UK
A.Kaban@cs.bham.ac.uk

***Abstract*—The prospect of carrying out data mining on cheaply compressed versions of high dimensional massive data sets holds tremendous potential and promise. However, our understanding of the performance guarantees available from such computationally inexpensive dimensionality reduction methods for data mining and machine learning tasks is currently lagging behind the requirements. In this paper we take a new look at randomly projected ordinary least squares regression, and give improved bounds on its expected excess risk. Our bounds are derived from first principles and use elementary techniques.**

I. Introduction

At the confluence of dealing with high dimensional massive data sets and the recent advances in random projections, compressed sensing and related areas, the prospect of carrying out data mining on compressive versions of the data holds tremendous potential. 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 sort of conditions are required and what sort of guarantees we can get from these techniques specifically for data mining and machine learning tasks.

Bounds on compressive regression and classification have been derived primarily using results from Johnson-Lindenstrauss embeddings and those of Compressed Sensing as building blocks [7], [9]. It is not clear at all if the conditions needed in those results are actually needed 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 really need their strong conditions for guarantees on tasks such as regression and classification of randomly projected data?

Some of our recent research has answered this question for linear classification [3]. It turned out that good classification is possible from randomly projected data with neither the requirement to preserve all inter-point distances in the reduced space, nor to recover the original data points. Intuitively, indeed in linear classification not all distances are important, and not all details of the data matter – we only need to preserve the class structure.

But how about linear regression? In regression the targets are real-valued, and the square loss is of interest. It is not so clear at first sight whether it would still be possible to get away with less than preserving all the geometry of the training set. Prior work on bounding the excess risk of compressive ordinary least squares regression [7], [9] has certainly built on that premise. However, in this paper we show that improved bounds can be obtained on a more direct route. We give improved bounds on randomly projected ordinary linear least squares (OLS) regression that are derived from first principles and use elementary techniques.

II. 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 $x_n v = E[y_n]$, under the linear model assumption:

$$y_n = x_n v + \gamma_n, n = 1, ..., N \tag{1}$$

where γ_n is i.i.d. Gaussian noise as $\mathcal{N}(0, \sigma^2)$.

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. That is, the covariance of the inputs is known and needs not be estimated. This is the simplest setting and it is suitable for studying dimensionality reduction techniques [5], which is our purpose. It should be noted that the finite 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 [5], 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:

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

where Y denotes the column vector of $y_n, n = 1, ..., N$ and X is the $N \times d$ matrix with rows $x_n, n = 1, ..., N$. The expectations are with respect to Y throughout, unless indicated otherwise. Denote by w the true minimiser of the square loss.

$$w = \arg\min_u L(u) \tag{3}$$

The excess risk of an estimator v is defined as:

$$R(v) = L(v) - L(w) \tag{4}$$

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

$$\hat{L}(v) = \frac{1}{N}\|Y - Xv\|^2 \tag{5}$$

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

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

We will make use of the following known result (Proposition 1 in [5]) about the expected excess risk of the OLS estimator in the fixed design setting:

Lemma 1 [5] Let $\sigma^2 = \text{Var}(y_i)$ $(i = 1, ..., N)$, $\Sigma = X^T X/N$ fixed and invertible, w the optimal OLS as above, and $\hat{w}$ the OLS estimator. Then the expected risk $\text{E}[R(\hat{w})]$ equals:

$$\text{E}[L(\hat{w})] - L(w) = \sigma^2 \frac{d}{N} \tag{7}$$

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

III. Compressive OLS revisited

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. Moreover, if $d > N$ 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, $\hat{L}_{ridge}(v) = \hat{L}(v) + \lambda\|v\|^2$, which then yields a regularised covariance $\Sigma + \lambda I_d$ to work with, and 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.

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 [7], [9]. In [7], the Johnson-Lindenstrauss lemma 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 [9], the Johnson-Lindenstrauss lemma (JLL) based argument is replaced by that of compressed sensing (CS), which ensures the same approximate global preservation of pairwise distances via the restricted isometry (RIP) property of the random matrix use for the linear compression. From RIP we have this independently of N, for $k \in \mathcal{O}(s\log(d))$ – but only provided that the input points have a sparse representation with at most s nonzero entries each. The impressive application in [**?**] to music similarity prediction from a million-dimensional data sets [9] clearly outperformed OLS in the original data space.

The JLL-based approach in [7] 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?

We may notice at this point that from the regression estimation point of view it seems counter-intuitive that the sample size detrimentally affects a quantity that features in the performance of the estimator. One might then attempt to speculate based on the CS-based argument in [9] 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 [9], along with the requirement of sparsity of the input data, are conditions that are just simply inherited from the CS literature, where these conditions 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 [9] that the cases where rp-OLS was experimentally observed to be particularly effective are not predicted by the currently existing theory. The remainder of this paper aims to fill this gap.

Let k be the dimension of a randomly oriented subspace that we project our input points to, and let R be the $k \times d$ random projection matrix with entries drawn i.i.d. from a zero mean Gaussian with variance $1/k$, i.e. $N(0, 1/k)$. We are interested in the expected excess risk of OLS that receives only a k-dimensional randomly projected version of the training set, i.e. it receives $S_R = \{(Rx_1, y_1), ..., (Rx_N, y_N)\}$ where $x_n \in \mathbb{R}^d, Rx_n \in \mathbb{R}^k, y_n \in \mathbb{R}, n = 1, ..., N$. Thus, we seek to bound, with high probability w.r.t. the random draw of R, the difference between the expected square loss of the k-dimensional OLS estimate obtained from S_R and the square loss of the optimal d-dimensional OLS, w.

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 $x_n R^T \hat{w}_R$ approximates $E[y_n]$. The square loss of an estimator v_R is:

$$L_R(v_R) = \frac{1}{N}\text{E}[\|Y - XR^T v_R\|^2] \tag{8}$$

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

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

The empirical square loss of an estimator v_R is:

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

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 quantity of interest is:

$$\text{E}[L_R(\hat{w}_R)] - L(w) \tag{12}$$

where the expectation is again w.r.t. Y. We seek a bound on this quantity that would hold w.h.p. with respect to the random draw of R.

We will prove the following result.

Theorem 2. Let $\sigma^2 = \text{Var}(y_i)$, and $\Sigma = X^T X/N$ fixed. Let w be the optimal OLS in $\mathbb{R}^d$, and $\hat{w}_R$ 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 $\mathcal{N}(0, 1/k)$. Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\mathrm{E}[L_R(\hat{w}_R)] - L(w) \leqslant \sigma^2 \frac{k}{N} + \frac{1}{\delta} \cdot \frac{1}{k} \cdot \|w\|^2_{\Sigma + Tr(\Sigma) I_d} \quad (13)$$

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

The first term is the variance of the estimator. Of course this is greatly reduced in comparison with the data space where it was $\sigma^2 d/N$. As it will become clear from the proof, the variance term is deterministic. The second term is a high probability bound on the bias of the estimator. This bias is the price for the reduced variance.

Before starting the proof, let us point out that the main difference from the bound obtained in [7] is that our bias term in the above eq. (13) is independent of N. The proof technique in [7] 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. As we shall see during the proof, this difference not only tightens their bound, but also yields a clear interpretation where the bias term becomes revealing of which characteristics of the linear regression problem determine its compressibility.

We can already see from eq. (13) that the first term is smallest when k is small – this is the variance term and 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 see clearly from the form of this term that a norm of the best OLS in the data space i.e. $\|w\|^2_{\Sigma + Tr(\Sigma) I_d}$ is the quantity that governs to what extent we can compress 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 in unknown, but it is theoretically pleasing to have a characterisation of problem compressibility in terms of the specific problem structure, i.e. a notion of norm of the best OLS w. In addition, the finding that this term does not grow with N is of both practical and theoretical relevance.

IV. Proof of Theorem 2

We start by applying Lemma 1 in $\mathbb{R}^k$. For any full row rank R this yields the following:

$$\mathrm{E}[L_R(\hat{w}_R)] - L_R(w_R) = \sigma^2 \frac{k}{N} \quad (14)$$

Note that the Gaussian random matrix R has full row rank a.s.

The l.h.s. of eq.(14) is much smaller than what we had in the full space in Lemma 1, since we reduced the dimensionality. The price to pay is that this expected risk is w.r.t the best achievable in $\mathbb{R}^k$ rather than in $\mathbb{R}^d$, so next we bound $L_R(w_R)$ with an expression that contains $L(w)$. By definition we have the inequality:

$$\begin{aligned} L_R(w_R) &\leqslant L_R(Rw) && (15)\\ &= \frac{1}{N}\mathrm{E}[\|Y - XR^T Rw\|^2] && (16)\\ &= \frac{1}{N}\mathrm{E}[\|Y - Xw\|^2] ... \\ &+ \frac{1}{N}\|Xw - XR^T Rw\|^2 && (17)\\ &= L(w) + \frac{1}{N}\|Xw - XR^T Rw\|^2 && (18) \end{aligned}$$

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

$$\mathrm{E}[L_R(\hat{w}_R)] - L(w) = \sigma^2 \frac{k}{N} + \frac{1}{N}\|Xw - XR^T Rw\|^2 \quad (19)$$

Remains to bound the last term in eq. (19), namely:

$$\frac{1}{N}\|Xw - XR^T Rw\|^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n w - x_n R^T Rw)^2 \quad (20)$$

This is the point where we will deviate from previous techniques. Since this term contains dot products of randomly projected vectors, the approach in previous works [7], [9] 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 either needs 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.

Instead, we will simply use Markov inequality. Our reasoning is as follows. Firstly, we do not wish to impose sparsity on the data because then our results will not be applicable when the data is not known to have a sparse representation. Secondly, the expression to bound is a sum of dependent (positive valued) 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.

By Markov inequality we have for any $\delta > 0$ that the following holds w.p. at least $1 - \delta$:

$$\frac{1}{N}\|Xw - XR^T Rw\|^2 \leqslant \frac{1}{\delta}\frac{1}{N}\mathrm{E}_R[\|Xw - XR^T Rw\|^2]$$

Expanding, we get:

$$\frac{1}{\delta}\mathrm{E}_R[w\frac{X^T X}{N}w + w^T R^T R\frac{X^T X}{N}R^T Rw - 2w^T\frac{X^T X}{N}R^T Rw]$$

$$= \frac{1}{\delta}\left(w\frac{X^T X}{N}w + w^T\mathrm{E}_R[R^T R\frac{X^T X}{N}R^T R]w - 2w^T\frac{X^T X}{N}\mathrm{E}_R[R^T R]w\right) \quad (21)$$

Observe that $\mathrm{E}_R[R^TR] = I_d$ since we had the entries of R i.i.d. from $\mathcal{N}(0, 1/k)$. Hence, after cancellations we get:

$$\frac{1}{\delta}\left(-w\Sigma w + w^T \mathrm{E}_R[R^TR\Sigma R^TR]w\right) \tag{22}$$

The following lemma computes the expectation $\mathrm{E}_R[R^TR\Sigma R^TR]$, which turns out to have a closed form.

Lemma 3 Let R be a $k \times d$ random matrix, $k < d$, with entries drawn i.i.d. from $\mathcal{N}(0, \omega^2)$, and Σ a $d \times d$ fixed positive semi-definite matrix. Then,

$$\mathrm{E}_R[R^TR\Sigma R^TR] = \omega^4 k((k+1)\Sigma + Tr(\Sigma)I_d) \tag{23}$$

Proof [of Lemma 3]. Take the SVD decomposition $\Sigma = U\Lambda U^T$, where $UU^T = I_d$, and Λ is diagonal. Then we can rewrite:

$$\begin{aligned} \mathrm{E}[R^TR\Sigma R^TR] &= \mathrm{E}[R^TRU\Lambda U^TR^TR] &(24)\\ &= U\mathrm{E}[U^TR^TRU\Lambda U^TR^TRU]U^T &(25)\end{aligned}$$

Note the Gaussian distribution is rotation-invariant, so RU has the same distribution as R. Therefore we can absorb U into R and have the r.h.s. of eq.(25) further equals to:

$$U\mathrm{E}[R^TR\Lambda^TR^TR]U^T \tag{26}$$

Therefore it is enough to compute $\mathrm{E}[R^TR\Lambda R^TR]$ with Λ being diagonal. Denoting by ρ the rank of Σ, this can be further rewritten as the following:

$$\mathrm{E}_R[R^TR\Lambda R^TR] = \sum_{i=1}^{\rho}\lambda_i \begin{bmatrix} \mathrm{E}[(r_1^Tr_i)^2] & \dots & \mathrm{E}[(r_1^Tr_i)(r_i^Tr_d)] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[(r_d^Tr_i)(r_i^Tr_1)] & \dots & \mathrm{E}[(r_d^Tr_i)^2] \end{bmatrix} \tag{27}$$

We will first compute the diagonal elements of a generic term of the above sum. These have the form $\mathrm{E}[(r_j^Tr_i)^2]$. We need to take separately the case when $j = i$ and when $j \neq i$.

Case $j = i$:

$$\begin{aligned} \mathrm{E}[(r_i^Tr_i)^2] &= \mathrm{E}[(\sum_{j=1}^k r_{ji}^2)^2] = \sum_{j=1}^k\sum_{j'=1}^k \mathrm{E}[r_{ji}^2r_{j'i}^2] &(28)\\ &= \sum_{j=1}^k\sum_{j'=1,j'\neq j}^k \mathrm{E}[r_{ji}^2]\mathrm{E}[r_{j'i}^2] + \sum_{j=1}^k \mathrm{E}[r_{ji}^4] \\ &= (k^2-k)\omega^4 + 3k\omega^4 \\ &= \omega^4(k^2+2k) &(29)\end{aligned}$$

Case $j \neq i$:

$$\begin{aligned} \mathrm{E}[(r_i^Tr_j)^2] &= \mathrm{E}[(\sum_{\ell=1}^k r_{\ell i}r_{\ell j})^2] = \sum_{\ell=1}^k\sum_{\ell'=1}^k \mathrm{E}[r_{\ell i}r_{\ell j}r_{\ell' i}r_{\ell' j}] \\ &= \sum_{\ell=1}^k\sum_{\ell'=1,\ell'\neq\ell}^k \mathrm{E}[r_{\ell i}]\mathrm{E}[r_{\ell j}]\mathrm{E}[r_{\ell' i}]\mathrm{E}[r_{\ell' j}] + ... \\ &+ \sum_{\ell=1}^k \mathrm{E}[r_{\ell i}^2r_{\ell j}^2] = k\cdot\omega^4. &(30)\end{aligned}$$

Next, we compute the off-diagonal elements. These have the form $\mathrm{E}[(r_j^Tr_i)(r_i^Tr_\ell)]$ with $j \neq \ell$.

$$\begin{aligned} \mathrm{E}[(r_j^Tr_i)(r_i^Tr_\ell)] &= \mathrm{E}[(\sum_{m=1}^k r_{mi}r_{mj})(\sum_{m'=1}^k r_{m'i}r_{m'\ell})] \\ &= \sum_{m=1}^k\sum_{m'=1}^k \mathrm{E}[r_{mi}r_{mj}r_{m'i}r_{m'\ell}] \\ &= 0\end{aligned}$$

by the independence of the entries of R and the fact that they have zero mean. Indeed, since $j \neq \ell$, the product inside the expectation will always have at least one independent entry of R on its own.

Hence, we obtained that for diagonal Λ, $\mathrm{E}_R[R^TR\Lambda R^TR]$ is a diagonal matrix.

Putting together,

$$\mathrm{E}[R^TR\Lambda R^TR] = \sum_{i=1}^{\rho}\lambda_i D_i \tag{31}$$

where D_i is a diagonal matrix having its (i,i)-th element equal to $\omega^4(k^2+2k)$ and all other diagonal elements equal to $\omega^4 k$. After some algebra, this may be further rewritten as:

$$\mathrm{E}[R^TR\Lambda R^TR] = \omega^4 k\left(\mathrm{Trace}(\Lambda)I_d + (k+1)\Lambda\right) \tag{32}$$

So by implication, we obtained a regularised version of the sample covariance estimate:

$$\mathrm{E}[R^TR\Sigma R^TR] = \omega^4 k\left(\mathrm{Trace}(\Sigma)I_d + (k+1)\Sigma\right) \tag{33}$$

which concludes the proof of Lemma 3. ■

Returning to the proof of Theorem 2, replacing ω^2 with $1/k$ we get for eq. (22) the following:

$$\begin{aligned} &\frac{1}{\delta}\cdot\left(-w\Sigma w + \left(1+\frac{1}{k}\right)w\Sigma w + \frac{1}{k}w^Tw\cdot Tr(\Sigma)\right) \\ =\ &\frac{1}{\delta}\left(\frac{1}{k}w^T(\Sigma + Tr(\Sigma)I_d)w\right) &(34)\\ =\ &\frac{1}{\delta}\cdot\frac{1}{k}\cdot\|w\|^2_{\Sigma+Tr(\Sigma)I_d} &(35)\end{aligned}$$

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

$$\frac{1}{N}\|Xw - XR^TRw\|^2 \leqslant \frac{1}{\delta}\cdot\frac{1}{k}\|w\|^2_{\Sigma+Tr(\Sigma)I_d} \tag{36}$$

Finally, plugging eq. (36) back into eq. (19) we obtain the statement of Theorem 2. ■

V. Discussion

A. Comparison with previous bounds on RP-OLS and numerical validation

Let us contrast eq. (36) with the JLL-based approach used in previous work [7]. It would have lead to the following:

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

with the same probability $1-\delta$. We see our bound eliminated the $\log(N)$ factor and otherwise it has a very similar form

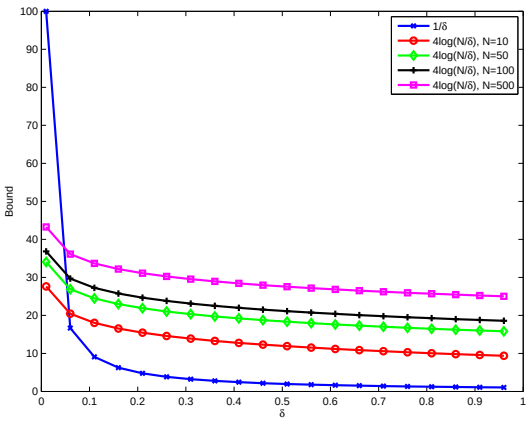

Fig. 1. Illustration of the difference between our bound and the previous JLL-based bound for different values of N. Except for very small values of the confidence parameter δ, our new bound is numerically tighter even when N is small.

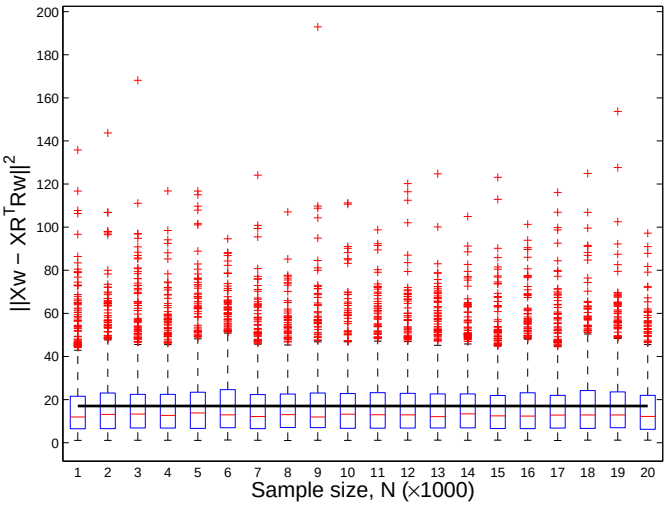

Fig. 2. Empirical verification that the bias term of RP-OLS, $\frac{1}{N}\|Xw - XR^TRw\|^2$, has no dependence on N. The entries of R are drawn i.i.d. from $\mathcal{N}(0, 1/k)$. The straight line is $\mathrm{E}_R[\frac{1}{N}\|Xw - XR^TRw\|^2]$ as computed using Lemma 3.

and behaviour w.r.t the rest of the factors. In particular, $\|w\|^2_{\Sigma+Tr(\Sigma)I_d} \leqslant 2\|w\|^2 Tr(\Sigma)$ by Hölder inequality. It is numerically less tight for extremely small values of δ because it has a $1/\delta$ dependence rather than a $\log(1/\delta)$ dependence which comes from the use of Markov inequality, but despite this it becomes tighter very quickly due to its independence on N, as seen in Figure 1.

Figure 2 provides a numerical verification of the behaviour of the term $\frac{1}{N}\|Xw - XR^TRw\|^2$ when N is varied between 1,000 and 20,000. The figure clearly confirms that this term does not depend on N. Here, for each sample size N tested, the input points were generated i.i.d. from a d-dimensional standard Gaussian and then fixed; w was also randomly generated, normalised to unit norm and then kept fixed. Each boxplot represents the distribution of the values of $\frac{1}{N}\|Xw - XR^TRw\|^2$ across 1000 independent draws of the random matrix R, each with entries drawn i.i.d from $\mathcal{N}(0, 1/k)$. The dimensionality was $d = 50$, and the projection space dimension $k = 3$. The straight line is $\mathrm{E}_R[\frac{1}{N}\|Xw - XR^TRw\|^2]$ as computed using Lemma 3.

B. Other random projection matrices

In the proof of Theorem 2 we made use of a crucial property of i.i.d. Gaussian random matrices, namely rotation invariance. It is of interest to know if other RP matrices would give similar guarantees. The proof technique in [7], via JLL ensures their bound holds for any matrix R that satisfies the JLL property – these are the sub-Gaussian random matrices, i.e. random matrices with entries drawn i.i.d. from a distribution that has moment generating function upper-bounded by that of the Gaussian. The proof technique in [9] goes via the RIP property in Compressed Sensing, which ensures their bound holds for any matrix R that satisfies the Restricted Isometry Property (RIP). This is a wider class of random matrices [2]. Since our Theorem 2 was proved from the first principles for Gaussian R, and did not make use of any of these building blocks, we need to find out if it would hold beyond the choice a Gaussian R. In particular, sparse RP matrices would be of practical interest to further speed up computations.

It turns out that it is possible to derive an analogue of our Lemma 3 without assuming Gaussianity, and without even assuming the existence of a moment generating function, as given in the following lemma. An example of sparse RP is also given below, which gives exactly the same guarantee as the Gaussian one.

Lemma 4 Let R be a $k \times d$ random matrix, $k < d$, with entries drawn i.i.d. from a symmetric distribution having the same first four moments as those of $\mathcal{N}(0, \omega^2)$. Let Σ be a $d \times d$ fixed positive semi-definite matrix. Then,

$$\mathrm{E}_R[R^T R\Sigma R^T R] = \omega^4 k((k+1)\Sigma + Tr(\Sigma)I_d) \qquad (38)$$

Proof sketch The proof is somewhat similar but much more tedious than that of Lemma 3. After SVD decomposing $\Sigma = U\Lambda U^T$, we define $\tilde{R} = RU$, and work with the matrix expectation $\mathrm{E}_R[\tilde{R}^T\tilde{R}\Lambda\tilde{R}^T\tilde{R}]$. Notice that $\tilde{R}$ has 0-mean symmetrically distributed entries, independent rows, and dependent but uncorrelated columns. Using these properties we first show that $\mathrm{E}_R[\tilde{R}^T\tilde{R}\Lambda\tilde{R}^T\tilde{R}]$ is diagonal. Then, by calculating its diagonal entries, we ultimately arrive at the following closed form expression:

$$\begin{aligned}\mathrm{E}_R[\tilde{R}^T\tilde{R}\Sigma\tilde{R}^T\tilde{R}] &= \omega^4 k[(k+1)\Sigma + Tr(\Sigma)I_d \\ &+ (\frac{\mu_4}{\omega^4} - 3)\sum_{i=1}^{d}\lambda_i \mathrm{Diag}_{j=1}^{d}(\sum_{a=1}^{d} u_{ai}^2 u_{aj}^2)]\end{aligned}$$

where u_{ai} is the a-th entry of the i-th eigenvector of Σ, μ_4 is the fourth moment of any entry of R, and $\mathrm{Diag}_{j=1}^{d}(e_j)$ stands for a diagonal matrix with diagonal entries $e_1, e_2, ..., e_d$.

We see that whenever $\mu_4 = 3\omega^4$, the last term cancels – this is the case e.g. for the Gaussian $\mathcal{N}(0, \omega^2)$ – which indeed recovers Lemma 3. Finally, noting that the r.h.s. of eq.(39) only depends on μ_4 and ω^4 concludes the proof. ■

Example. The following sparse random matrix, originally proposed for computational efficiency in [1], satisfies Lemma

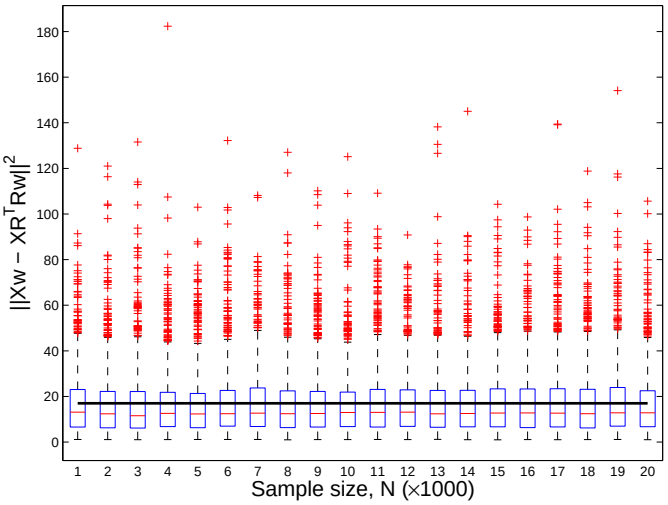

Fig. 3. Empirical verification of the behaviour of the term $\frac{1}{N}\|Xw - XR^TRw\|^2$ when the entries of R are drawn i.i.d. from the sparse random matrix in eq. (39). The straight line is $\mathrm{E}_R[\frac{1}{N}\|Xw - XR^TRw\|^2]$ as given by Lemma 4.

4. This is,

$$r_{ij} \overset{iid}{\sim} \begin{cases} -\omega\sqrt{3} & \text{w.p.} 1/6 \\ 0, & \text{w.p.} 2/3 \\ \omega\sqrt{3} & \text{w.p.} 1/6 \end{cases} \tag{39}$$

Indeed, we can easily verify that $\mathrm{Var}(r_{ij}) = \omega^2$, and $\mu_4(r_{ij}) = 3\omega^4$. Therefore, our bound on the bias of RP-OLS (and overall risk bound) is exactly the same for this RP matrix as it is for the Gaussian. Figure 3 verifies this empirically, and we see indeed the same behaviour and no dependence on N.

From Lemma 4 and Theorem 2 we conclude the following more general version of Theorem 2:

Theorem 5 Let $\sigma^2 = \mathrm{Var}(y_i)$, and $\Sigma = X^TX/N$ fixed. Let w be the optimal OLS in $\mathbb{R}^d$, and $\hat{w}_R$ 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 symmetric distribution with mean 0, variance $1/k$, and fourth moment $3/k^2$. Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\mathrm{E}[L_R(\hat{w}_R)] - L(w) \leqslant \sigma^2 \frac{k}{N} + \frac{1}{\delta} \cdot \frac{1}{k} \cdot \|w\|^2_{\Sigma + Tr(\Sigma) I_d} \tag{40}$$

VI. A perspective on combining RP-OLS regressions

In order to reduce the variability w.r.t. the random choice of R of the bias term of RP-OLS, we may consider an ensemble of RP-OLS, each working in a different random subspace. Here we will simply take an averaging ensemble to illustrate this, without the ambition of being the optimal combination scheme. Of course, ensembling will affect both the bias and the variance of the new ensemble-regressor, which we can quantify.

Further to the definitions in Section II, for a linear regression estimator v in $\mathbb{R}^d$ we denote $\bar{v} = \mathrm{E}[v]$. We should note this only evaluates to the optimal w in the fixed design setting if v was the OLS estimator. For other, biased estimators (e.g. ridge regression) it would be different.

The bias-variance decomposition of the expected risk of a linear regression estimator v will be useful:

$$\begin{aligned} \mathrm{E}[R(v)] &= \frac{1}{N}\|Xv - Xw\|^2 = \|v - w\|^2_\Sigma & (41) \\ &= \mathrm{E}[\|v - \bar{v}\|^2_\Sigma] + \|\bar{v} - w\|^2_\Sigma & (42) \end{aligned}$$

where the first term on the r.h.s. is the variance and the second term is the bias.

Now, consider M random projections of the same data X, via $k \times d$ random matrices $R_i, i = 1, ..., M$ with entries i.i.d. from $\mathcal{N}(0, 1/k)$, and estimate the OLS on each. We then combine the predictions that these estimators give on projected points via averaging.

It is easy to see that the OLS estimator on the i-th random subspace has the form:

$$\hat{w}_{R_i} = (R_i \Sigma R_i^T)^{-1} \frac{1}{N} R_i X^T Y \in \mathbb{R}^k \tag{43}$$

Hence its prediction on a projected point $R_i x$ is $x^T R^T \hat{w}_{R_i} = x^T R_i^T (R_i \Sigma R_i^T)^{-1} \frac{1}{N} R_i X^T Y$.

Consequently, the averaged predictions take the form:

$$x^T \frac{1}{M} \sum_{i=1}^{M} R_i^T (R_i \Sigma R_i^T)^{-1} \frac{1}{N} R_i X^T Y \tag{44}$$

In the limit when $M \to \infty$, this converges to:

$$x^T \mathrm{E}_R[R^T (R \Sigma R^T)^{-1} R] \frac{1}{N} X^T Y \tag{45}$$

and we can regard the finite ensemble as the finite approximation of this. Therefore we can analyse the risk of this ensemble in the original data space $\mathbb{R}^d$ by regarding it as the following ensemble-estimator:

$$\hat{w}_{ens} = \mathrm{E}_R[R^T (R \Sigma R^T)^{-1} R] \frac{1}{N} X^T Y \in \mathbb{R}^d \tag{46}$$

The following theorem gives the expected risk of this estimator.

Theorem 6 Let $\sigma^2 = \mathrm{Var}(y_i)$, $\Sigma = X^TX/N$ fixed. Then the expected risk of $\hat{w}_{ens}$ is:

$$\mathrm{E}[L(\hat{w}_{ens})] - L(w) = \sum_{j=1}^{d} \left(\frac{\sigma^2}{N} \left(\frac{\lambda_j}{\eta_j} \right)^2 + \lambda_j \beta_j^2 \left(\frac{\lambda_j}{\eta_j} - 1 \right)^2 \right) \tag{47}$$

where $\lambda_j = \lambda_j(\Sigma)$, $\eta_j = 1/\lambda_j(\mathrm{E}[R^T(R\Sigma R^T)^{-1}R])$, and where $\lambda_j(\cdot)$ denotes the j-th eigenvalue of its argument.

Proof. We adapt the analysis technique of ridge regression in [5] (Proposition 2), see also [6], in the fixed design setting. To do this, we first observe that $\hat{w}_{ens}$ is rotation-invariant and so it is no loss of generality to assume Σ diagonal. Indeed, writing $\Sigma = U \Lambda U^T$ for the SVD decomposition of Σ, one can easily verify that

$$\hat{w}_{ens} = U \mathrm{E}_R[R^T (R \Lambda R^T)^{-1} R] U^T \frac{1}{N} X^T Y \tag{48}$$

and that the matrix $\mathrm{E}[R^T(R\Lambda R)^{-1}R]$ is diagonal. The proof of this fact is straightforward and can be found in [8](Sec. IV.A.). The diagonal elements of $\mathrm{E}[R^T(R\Lambda R)^{-1}R]$ will be denoted as $1/\eta_j$.

Now, following the steps of the proof of Proposition 2 in [5] we get the following for the variance and bias terms respectively:

$$\mathrm{E}[\|X\hat{w}_{ens} - X\bar{\hat{w}}_{ens}\|^2] = \frac{\sigma^2}{N}\sum_{j=1}^{d}\left(\frac{\lambda_j}{\eta_j}\right)^2 \quad (49)$$

$$\|X\bar{\hat{w}}_{ens} - Xw\|^2 = \sum_{j=1}^{d}\lambda_j w_j^2\left(\frac{\lambda_j}{\eta_j} - 1\right)^2 \quad (50)$$

where $\bar{\hat{w}}_{ens} = \mathrm{E}_Y[\hat{w}_{ens}]$, and the sum of these completes the proof. ■

The exact form of the eigenvalues η_j are given in [8]. Alternatively bounds that are more interpretable are also available [4]. These bounds imply that $\mathrm{E}_R[R^T(R\Sigma R)^{-1}R]^{-1}$ implements a sophisticated regularisation to Σ, which is parametrised by k. In the range space of Σ it behaves as a shrinkage regulariser, and in its null-space it acts as a ridge regulariser. Therefore the averaging combination of RP-OLS is applicable even when OLS in the data space would be not, i.e. when Σ is singular, and it has guarantees of similar flavour as those of ridge regression – with η_j replacing $\lambda_j + \lambda$ (were λ is the regularisation parameter in ridge regression).

VII. Conclusions and outlook

We gave improved bounds on the excess risk of randomly projected OLS regression in the fixed design setting. The new bounds are more interpretable, remove a spurious factor of $\log(N)$ from the h.p. bound on the bias term of RP-OLS, and show that this term does not depend on the sample size at all. Our bound on RP-OLS holds for any random projection matrix that has entries drawn i.i.d. from a symmetric distribution with the first four moments equal to those of the Gaussian $\mathcal{N}(0, 1/k)$. We also briefly considered the possibility of ensembling several RP-OLS regressors, so far with Gaussian random projection matrices, and have seen this has an expected excess risk of the form that resembles that of ridge-regression.

It would be of interest to try to improve the $1/\delta$ term to a logarithmic dependence in δ. Future work also includes a more detailed study of the practical implications of these results, comparison with some recent findings about PCA-OLS and ridge regression [6], looking at other RP-OLS ensembles and other combination schemes along the lines of [10]. Extensions of our results to the random design setting is also subject to future work.

References

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

[2] 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.

[3] R.J. Durrant and A. Kaban. 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.

[4] R.J. Durrant and A. Kabán. Random Projections as Regularizers: Learning a Linear Discriminant Ensemble from Fewer Observations than Dimensions, Tech. Report CSR-12-01, University of Birmingham, School of Computer Science, 2012.

[5] D. Hsu, Sham M. Kakade, Tong Zhang. Random Design Analysis of Ridge Regression, COLT 12, 2012.

[6] Paramveer Dhillon, Dean P. Foster, Sham M. Kakade, Lyle Ungar. A Risk Comparison of Ordinary Least Squares vs Ridge Regression. To appear in JMLR.ArXiv Report, arXiv:1105.0875.

[7] 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.

[8] T.L. Marzetta, G.H. Tucci, and S.H. Simon, A Random Matrix-Theoretic Approach to Handling Singular Covariance Estimates, IEEE Trans. Information Theory 57 (2011), no. 9, 6256-71.

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

[10] Y. Zhang, J.C. Duchi, M. Wainwright. Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates. COLT'12, JMLR: Workshop and Conference Proceedings vol 30 (2013) 1-26.