Random Projections for Large-Scale Regression

Gian-Andrea Thanei, Christina Heinze, and Nicolai Meinshausen

Abstract Fitting linear regression models can be computationally very expensive in large-scale data analysis tasks if the sample size and the number of variables are very large. Random projections are extensively used as a dimension reduction tool in machine learning and statistics. We discuss the applications of random projections in linear regression problems, developed to decrease computational costs, and give an overview of the theoretical guarantees of the generalization error. It can be shown that the combination of random projections with least squares regression leads to similar recovery as ridge regression and principal component regression. We also discuss possible improvements when averaging over multiple random projections, an approach that lends itself easily to parallel implementation.

1 Introduction

Assume we are given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (n samples of a p-dimensional random variable) and a response vector $\mathbf{Y} \in \mathbb{R}^n$. We assume a linear model for the data where $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ for some regression coefficient $\beta \in \mathbb{R}^p$ and ε i.i.d. mean-zero noise. Fitting a regression model by standard least squares or ridge regression requires $\mathcal{O}(np^2)$ or $\mathcal{O}(p^3)$ flops. In the situation of large-scale (n, p very large) or high dimensional ($p \gg n$) data these algorithms are not applicable without having to pay a huge computational price.

Using a random projection, the data can be "compressed" either row- or column-wise. Row-wise compression was proposed and discussed in [7, 15, 19]. These approaches replace the least-squares estimator

$$\underset{\gamma \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\gamma\|_2^2 \qquad \text{with the estimator} \qquad \underset{\gamma \in \mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{\psi}\mathbf{Y} - \boldsymbol{\psi}\mathbf{X}\gamma\|_2^2, \tag{1}$$

where the matrix $\boldsymbol{\psi} \in \mathbb{R}^{m \times n}$ ($m \ll n$) is a random projection matrix and has, for example, i.i.d. $\mathcal{N}(0, 1)$ entries. Other possibilities for the choice of $\boldsymbol{\psi}$ are

G.-A. Thanei • C. Heinze • N. Meinshausen (✉)
ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
e-mail: thanei@stat.math.ethz.ch; heinze@stat.math.ethz.ch; meinshausen@stat.math.ethz.ch

S.E. Ahmed (ed.), *Big and Complex Data Analysis*, Contributions to Statistics,
DOI 10.1007/978-3-319-41573-4_3

discussed below. The high dimensional setting and ℓ_1-penalized regression are considered in [19], where it is shown that a sparse linear model can be recovered from the projected data under certain conditions. The optimization problem is still p-dimensional, however, and computationally expensive if the number of variables is very large.

Column-wise compression addresses this later issue by reducing the problem to a d-dimensional optimization with $d \ll p$ by replacing the least-squares estimator

$$\underset{\gamma \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\gamma\|_2^2 \qquad \text{with the estimator} \qquad \boldsymbol{\phi} \underset{\gamma \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\phi}\gamma\|_2^2, \tag{2}$$

where the random projection matrix is now $\boldsymbol{\phi} \in \mathbb{R}^{p \times d}$ (with $d \ll p$). By right multiplication to the data matrix $\mathbf{X}$ we transform the data matrix to $\mathbf{X}\boldsymbol{\phi}$ and thereby reduce the number of variables from p to d and thus reducing computational complexity. The Johnson–Lindenstrauss Lemma [5, 8, 9] guarantees that the distance between two transformed sample points is approximately preserved in the column-wise compression.

Random projections have also been considered under the aspect of preserving privacy [3]. By pre-multiplication with a random projection matrix as in (1) no observation in the resulting matrix can be identified with one of the original data points. Similarly, post-multiplication as in (2) produces new variables that do not reveal the realized values of the original variables.

In many applications the random projection used in practice falls under the class of Fast Johnson–Lindenstrauss Transforms (FJLT) [2]. One instance of such a fast projection is the Subsampled Randomized Hadamard Transform (SRHT) [17]. Due to its recursive definition, the matrix–vector product has a complexity of $\mathcal{O}(p \log(p))$, reducing the cost of the projection to $\mathcal{O}(np \log(p))$. Other proposals that lead to speedups compared to a Gaussian random projection matrix include random sign or sparse random projection matrices [1]. Notably, if the data matrix is sparse, using a sparse random projection can exploit sparse matrix operations. Depending on the number of non-zero elements in $\mathbf{X}$, one might prefer using a sparse random projection over an FJLT that cannot exploit sparsity in the data. Importantly, using $\mathbf{X}\boldsymbol{\phi}$ instead of $\mathbf{X}$ in our regression algorithm of choice can be disadvantageous if $\mathbf{X}$ is extremely sparse and d cannot be chosen to be much smaller than p. (The projection dimension d can be chosen by cross validation.) As the multiplication by $\boldsymbol{\phi}$ "densifies" the design matrix used in the learning algorithm the potential computational benefit of sparse data is not preserved.

For OLS and row-wise compression as in (1), where n is very large and $p < m < n$, the SRHT (and similar FJLTs) can be understood as a subsampling algorithm. It preconditions the design matrix by rotating the observations to a basis where all points have approximately uniform leverage [7]. This justifies uniform subsampling in the projected space which is applied subsequent to the rotation in order to reduce the computational costs of the OLS estimation. Related ideas can be found in the way columns and rows of $\mathbf{X}$ are sampled in a CUR-matrix decomposition [12]. While the approach in [7] focuses on the concept of leverage, McWilliams et al.

[15] propose an alternative scheme that allows for outliers in the data and makes use of the concept of influence [4]. Here, random projections are used to approximate the influence of each observation which is then used in the subsampling scheme to determine which observations to include in the subsample.

Using random projections column-wise as in (2) as a dimensionality reduction technique in conjunction with (ℓ_2 penalized) regression has been considered in [10, 11, 13]. The main advantage of these algorithms is the computational speedup while preserving predictive accuracy. Typically, a variance reduction is traded off against an increase in bias. In general, one disadvantage of reducing the dimensionality of the data is that the coefficients in the projected space are not interpretable in terms of the original variables. Naively, one could reverse the random projection operation by projecting the coefficients estimated in the projected space back into the original space as in (2). For prediction purposes this operation is irrelevant, but it can be shown that this estimator does not approximate the optimal solution in the original p-dimensional coefficient space well [18]. As a remedy, Zhang et al. [18] propose to find the dual solution in the projected space to recover the optimal solution in the original space. The proposed algorithm approximates the solution to the original problem accurately if the design matrix is low-rank or can be sufficiently well approximated by a low-rank matrix.

Lastly, random projections have been used as an auxiliary tool. As an example, the goal of McWilliams et al. [16] is to distribute ridge regression across variables with an algorithm called LOCO. The design matrix is split across variables and the variables are distributed over processing units (workers). Random projections are used to preserve the dependencies between all variables in that each worker uses a randomly projected version of the variables residing on the other workers in addition to the set of variables assigned to itself. It then solves a ridge regression using this local design matrix. The solution is the concatenation of the coefficients found from each worker and the solution vector lies in the original space so that the coefficients are interpretable. Empirically, this scheme achieves large speedups while retaining good predictive accuracy. Using some of the ideas and results outlined in the current manuscript, one can show that the difference between the full solution and the coefficients returned by LOCO is bounded.

Clearly, row- and column-wise compression can also be applied simultaneously or column-wise compression can be used together with subsampling of the data instead of row-wise compression. In the remaining sections, we will focus on the column-wise compression as it poses more difficult challenges in terms of statistical performance guarantees. While row-wise compression just reduces the effective sample size and can be expected to work in general settings as long as the compressed dimension $m < n$ is not too small [19], column-wise compression can only work well if certain conditions on the data are satisfied and we will give an overview of these results. If not mentioned otherwise, we will refer with compressed regression and random projections to the column-wise compression.

The structure of the manuscript is as follows: We will give an overview of bounds on the estimation accuracy in the following Sect. 2, including both known results and new contributions in the form of tighter bounds. In Sect. 3 we will discuss the

possibility and properties of variance-reducing averaging schemes, where estimators based on different realized random projections are aggregated. Finally, Sect. 4 concludes the manuscript with a short discussion.

2 Theoretical Results

We will discuss in the following the properties of the column-wise compressed estimator as in (2), which is defined as

$$\hat{\beta}_d^{\boldsymbol{\phi}} \;=\; \boldsymbol{\phi} \operatorname*{argmin}_{\gamma \in \mathbb{R}^d} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\phi}\gamma\|_2^2, \tag{3}$$

where we assume that ϕ has i.i.d. $\mathcal{N}(0, 1/d)$ entries. This estimator will be referred to as the compressed least-squares estimator (CLSE) in the following. We will focus on the unpenalized form as in (3) but note that similar results also apply to estimators that put an additional penalty on the coefficients β or γ. Due to the isotropy of the random projection, a ridge-type penalty as in [11, 16] is perhaps a natural choice. An interesting summary of the bounds on random projections is, on the other hand, that the random projection as in (3) already acts as a regularization and the theoretical properties of (3) are very much related to the properties of a ridge-type estimator of the coefficient vector in the absence of random projections.

We will restrict discussion of the properties mostly to the mean-squared error (MSE)

$$\mathbb{E}_{\boldsymbol{\phi}}\big[\mathbb{E}_{\varepsilon}(\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2)\big]. \tag{4}$$

First results on compressed least squares have been given in [13] in a random design setting. It was shown that the bias of the estimator (3) is of order $\mathcal{O}(\log(n)/d)$. This proof used a modified version of the Johnson–Lindenstrauss Lemma. A recent result [10] shows that the $\log(n)$-term is not necessary for fixed design settings where $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ for some $\beta \in \mathbb{R}^p$ and ε is i.i.d. noise, centred $\mathbb{E}_{\varepsilon}[\varepsilon] = 0$ and with the variance $\mathbb{E}_{\varepsilon}[\varepsilon\varepsilon'] = \sigma^2 I_{n\times n}$. We will work with this setting in the following.

The following result of [10] gives a bound on the MSE for fixed design.

Theorem 1 ([10]) *Assume fixed design and* $\operatorname{Rank}(\mathbf{X}) \geq d$. *Then*

$$\mathbb{E}_{\boldsymbol{\phi}}\big[\mathbb{E}_{\varepsilon}(\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2)\big] \;\leq\; \sigma^2 d + \frac{\|\mathbf{X}\beta\|_2^2}{d} + \operatorname{trace}(\mathbf{X}'\mathbf{X})\frac{\|\beta\|_2^2}{d}. \tag{5}$$

Proof See Appendix.

Compared with [13], the result removes an unnecessary $\mathcal{O}(\log(n))$ term and demonstrates the $\mathcal{O}(1/d)$ behaviour of the bias. The result also illustrates the tradeoffs when choosing a suitable dimension d for the projection. Increasing d

will lead to a $1/d$ reduction in the bias terms but lead to a linear increase in the estimation error (which is proportional to the dimension in which the least-squares estimation is performed). An optimal bound can only be achieved with a value of d that depends on the unknown signal and in practice one would typically use cross validation to make the choice of the dimension of the projection.

One issue with the bound in Theorem 1 is that the bound on the bias term in the noiseless case ($Y = \mathbf{X}\beta$)

$$\mathbb{E}_{\boldsymbol{\phi}}\left[\mathbb{E}_{\varepsilon}(\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2)\right] \leq \frac{\|\mathbf{X}\beta\|_2^2}{d} + \operatorname{trace}(\mathbf{X}'\mathbf{X})\frac{\|\beta\|_2^2}{d} \tag{6}$$

is usually weaker than the trivial bound (by setting $\hat{\beta}_d^{\boldsymbol{\phi}} = 0$) of

$$\mathbb{E}_{\boldsymbol{\phi}}\left[\mathbb{E}_{\varepsilon}(\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2)\right] \leq \|\mathbf{X}\beta\|_2^2 \tag{7}$$

for most values of $d < p$. By improving the bound, it is also possible to point out the similarities between ridge regression and compressed least squares.

The improvement in the bound rests on a small modification in the original proof in [10]. The idea is to bound the bias term of (4) by optimizing over the upper bound given in the foregoing theorem. Specifically, one can use the inequality

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\beta\|_2^2]] \\ \leq \min_{\hat{\beta}\in\mathbb{R}^p} \mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'\hat{\beta}\|_2^2]],\end{aligned}$$

instead of

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\beta\|_2^2]] \\ \leq \mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'\beta\|_2^2]].\end{aligned}$$

To simplify the exposition we will from now on always assume we have rotated the design matrix to an orthogonal design so that the Gram matrix is diagonal:

$$\Sigma = \mathbf{X}'\mathbf{X} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p). \tag{8}$$

This can always be achieved for any design matrix and is thus not a restriction. It implies, however, that the optimal regression coefficients β are expressed in the basis in which the Gram matrix is orthogonal, this is the basis of principal components. This will turn out to be the natural choice for random projections and allows for easier interpretation of the results.

Furthermore note that in Theorem 1 we have the assumption $\operatorname{Rank}(\mathbf{X}) \geq d$, which tells us that we can apply the CLSE in the high dimensional setting $p \gg n$ as long as we choose d small enough (smaller than $\operatorname{Rank}(\mathbf{X})$, which is usually equal to n) in order to have uniqueness.

With the foregoing discussion on how to improve the bound in Theorem 1 we get the following theorem:

Theorem 2 *Assume* $\text{Rank}(\mathbf{X}) \geq d$*, then the MSE (4) can be bounded above by*

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2]] \leq \sigma^2 d + \sum_{i=1}^{p} \beta_i^2 \lambda_i w_i \tag{9}$$

where

$$w_i = \frac{(1 + 1/d)\lambda_i^2 + (1 + 2/d)\lambda_i \,\text{trace}(\Sigma) + \text{trace}(\Sigma)^2/d}{(d + 2 + 1/d)\lambda_i^2 + 2(1 + 1/d)\lambda_i \,\text{trace}(\Sigma) + \text{trace}(\Sigma)^2/d}. \tag{10}$$

Proof See Appendix.

The w_i are shrinkage factors. By defining the proportion of the total variance observed in the direction of the ith principal component as

$$\alpha_i = \frac{\lambda_i}{\text{trace}(\Sigma)}, \tag{11}$$

we can rewrite the shrinkage factors in the foregoing theorem as

$$w_i = \frac{(1 + 1/d)\alpha_i^2 + (1 + 2/d)\alpha_i + 1/d}{(d + 2 + 1/d)\alpha_i^2 + 2(1 + 1/d)\alpha_i + 1/d}. \tag{12}$$

Analyzing this term shows that the shrinkage is stronger in directions of high variance compared to directions of low variance. To explain this relation in a bit more detail we compare it to ridge regression. The MSE of ridge regression with penalty term $\lambda\|\beta\|_2^2$ is given by

$$\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\beta^{\text{Ridge}}\|_2^2] = \sigma^2 \sum_{i=1}^{p} \Big(\frac{\lambda_i}{\lambda_i + \lambda}\Big)^2 + \sum_{i=1}^{p} \beta_i^2 \lambda_i \Big(\frac{\lambda}{\lambda + \lambda_i}\Big)^2. \tag{13}$$

Imagine that the signal lives on the space spanned by the first q principal directions, that is $\beta_i = 0$ for $i > q$. The best MSE we could then achieve is $\sigma^2 q$ by running a regression on the first q first principal directions. For random projections, we can see that we can indeed reduce the bias term to nearly zero by forcing $w_i \approx 0$ for $i = 1, \ldots, q$. This requires $d \gg q$ as the bias factors will then vanish like $1/d$. Ridge regression, on the other hand, requires that the penalty λ is smaller than the qth largest eigenvalue λ_q (to reduce the bias on the first q directions) but large enough to render the variance factor $\lambda_i/(\lambda_i + \lambda)$ very small for $i > q$. The tradeoff in choosing the penalty λ in ridge regression and choosing the dimension d for random projections is thus very similar. The number of directions for which the eigenvalue λ_i is larger than the penalty λ in ridge corresponds to the effective dimension and

will yield the same variance bound as in random projections. The analogy between the MSE bounds (9) for random projections and (13) for ridge regression illustrates thus a close relationship between compressed least squares and ridge regression or principal component regression, similar to Dhillon et al. [6].

Instead of an upper bound for the MSE of CLSE as in [10, 13], we will in the following try to derive explicit expressions for the MSE, following the ideas in [10, 14] and we give a closed form MSE in the case of orthonormal predictors. The derivation will make use of the following notation:

Definition 1 Let $\boldsymbol{\phi} \in \mathbb{R}^{p\times d}$ be a random projection. We define the following matrices:

$$\boldsymbol{\phi}_d^{\mathbf{X}} = \boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}' \in \mathbb{R}^{p\times p} \quad \text{and} \quad T_d^{\boldsymbol{\phi}} = \mathbb{E}_{\boldsymbol{\phi}}[\boldsymbol{\phi}_d^{\mathbf{X}}] = \mathbb{E}_{\boldsymbol{\phi}}[\boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'] \in \mathbb{R}^{p\times p}.$$

The next Lemma [14] summarizes the main properties of $\boldsymbol{\phi}_d^{\mathbf{X}}$ and $T_d^{\boldsymbol{\phi}}$.

Lemma 1 *Let $\boldsymbol{\phi} \in \mathbb{R}^{p\times d}$ be a random projection. Then*

(*i*) $(\boldsymbol{\phi}_d^{\mathbf{X}})' = \boldsymbol{\phi}_d^{\mathbf{X}}$ *(symmetric),*
(*ii*) $\boldsymbol{\phi}_d^{\mathbf{X}}\mathbf{X}'\mathbf{X}\boldsymbol{\phi}_d^{\mathbf{X}} = \boldsymbol{\phi}_d^{\mathbf{X}}$ *(projection),*
(*iii*) *if $\Sigma = \mathbf{X}'\mathbf{X}$ is diagonal $\Rightarrow T_d^{\boldsymbol{\phi}}$ is diagonal.*

Proof See Marzetta et al. [14].

The important point of this lemma is that when we assume orthogonal design then $T_d^{\boldsymbol{\phi}}$ is diagonal. We will denote this by

$$T_d^{\boldsymbol{\phi}} = \operatorname{diag}(1/\eta_1, \ldots, 1/\eta_p),$$

where the terms η_i are well defined but without an explicit representation.

A quick calculation reveals the following theorem:

Theorem 3 *Assume* $\operatorname{Rank}(\mathbf{X}) \geq d$*, then the MSE (4) equals*

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2]] = \sigma^2 d + \sum_{i=1}^{p} \beta_i^2 \lambda_i \Big(1 - \frac{\lambda_i}{\eta_i}\Big). \tag{14}$$

Furthermore we have

$$\sum_{i=1}^{p} \frac{\lambda_i}{\eta_i} = d. \tag{15}$$

Proof See Appendix.

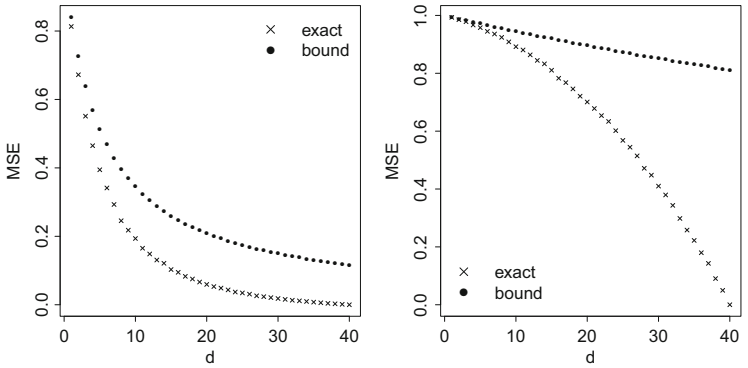

Fig. 1 Numerical simulations of the bounds in Theorems 2 and 3. *Left*: the exact factor $(1-\lambda_1/\eta_1)$ in the MSE is plotted versus the bound w_1 as a function of the projection dimension d. *Right*: the exact factor $(1-\lambda_p/\eta_p)$ in the MSE and the upper bound w_p. Note that the upper bound works especially well for small values of d and for the larger eigenvalues and is always below the trivial bound 1

By comparing coefficients in Theorems 2 and 3, we obtain the following corollary, which is illustrated in Fig 1:

Corollary 1 *Assume* $\text{Rank}(\mathbf{X}) \geq d$, *then*

$$\forall i \in \{1, \ldots, p\} : \quad 1 - \frac{\lambda_i}{\eta_i} \leq w_i \tag{16}$$

As already mentioned in general we cannot give a closed form expression for the terms η_i in general. However, for some special cases (26) can help us to get to an exact form of the MSE of CLSE. If we assume orthonormal design ($\Sigma = CI_{p\times p}$), then we have that λ_i/η_i is a constant for all i and and thus, by (26), we have $\eta_i = Cp/d$. This gives

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2]] = \sigma^2 d + C\sum_{i=1}^{p} \beta_i^2 \Big(1 - \frac{d}{p}\Big), \tag{17}$$

and thus we end up with a closed form MSE for this special case.

Providing the exact mean-squared errors allows us to quantify the conservativeness of the upper bounds. The upper bound has been shown to give a good approximation for small dimensions d of the projection and for the signal contained in the larger eigenvalues.

3 Averaged Compressed Least Squares

We have so far looked only into compressed least-squares estimator with one single random projection. An issue in practice of the compressed least-squares estimator is its variance due to the random projection as an additional source of randomness. This variance can be reduced by averaging multiple compressed least-squares estimates coming from different random projections. In this section we will show some properties of the averaged compressed least-squares estimator (ACLSE) and discuss its advantage over the CLSE.

Definition 2 (ACLSE) Let $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\} \in \mathbb{R}^{p\times d}$ be independent random projections, and let $\hat{\beta}_d^{\boldsymbol{\phi}_i}$ for all $i \in \{1, \ldots, K\}$ be the respective compressed least-squares estimators. We define the averaged compressed least-squares estimator (ACLSE) as

$$\hat{\beta}_d^K := \frac{1}{K}\sum_{i=1}^{K} \hat{\beta}_d^{\boldsymbol{\phi}_i}. \tag{18}$$

One major advantage of this estimator is that it can be calculated in parallel with the minimal number of two communications, one to send the data and one to receive the result. This means that the asymptotic computational cost of $\hat{\beta}_d^K$ is equal to the cost of $\hat{\beta}_d^{\boldsymbol{\phi}}$ if calculations are done on K different processors. To investigate the MSE of $\hat{\beta}_d^K$, we restrict ourselves for simplicity to the limit case

$$\hat{\beta}_d = \lim_{K\to\infty} \hat{\beta}_d^K \tag{19}$$

and instead only investigate $\hat{\beta}_d$. The reasoning being that for large enough values of K (say $K > 100$) the behaviour of $\hat{\beta}_d$ is very similar to $\hat{\beta}_d^K$. The exact form of the MSE in terms of the η_i's is given in [10]. Here we build on these results and give an explicit upper bound for the MSE.

Theorem 4 *Assume* $\mathrm{Rank}(\mathbf{X}) \geq d$. *Define*

$$\tau = \sum_{i=1}^{p} \left(\frac{\lambda_i}{\eta_i}\right)^2.$$

The MSE of $\hat{\beta}_d$ *can be bounded from above by*

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d\|_2^2]] \leq \sigma^2\tau + \sum_{i=1}^{p} \beta_i^2 \lambda_i w_i^2,$$

where the w_i's are given (as in Theorem 1) by

$$w_i = \frac{(1+1/d)\lambda_i^2 + (1+2/d)\lambda_i \operatorname{trace}(\Sigma) + \operatorname{trace}(\Sigma)^2/d}{(d+2+1/d)\lambda_i^2 + 2(1+1/d)\lambda_i \operatorname{trace}(\Sigma) + \operatorname{trace}(\Sigma)^2/d}.$$

and

$$\tau \in [d^2/p, d].$$

Proof See Appendix.

Comparing averaging to the case where we only have one single estimator we see that there are two differences: First the variance due to the model noise ε turns into $\sigma^2\tau$ with $\tau \in [d^2/p, d]$, thus $\tau \leq d$. Second the shrinkage factors w_i in the bias are now squared, which in total means that the MSE of $\hat{\beta}_d$ is always smaller or equal to the MSE of a single estimator $\hat{\beta}_d^{\boldsymbol{\phi}}$.

We investigate the behaviour of τ as a function of d in three different situations (Fig. 2). We first look at two extreme cases of covariance matrices for which the respective upper and lower bounds $[d^2/p, d]$ for τ are achieved. For the lower bound, let $\Sigma = I_{p\times p}$ be orthonormal. Then $\lambda_i/\eta_i = c$ for all i, as above. From

$$\sum_{i=1}^{p} \lambda_i/\eta_i = d$$

we get $\lambda_i/\eta_i = d/p$. This leads to

$$\tau = \sum_{i=1}^{p} (\lambda_i/\eta_i)^2 = p\frac{d^2}{p^2} = \frac{d^2}{p},$$

which reproduces the lower bound.

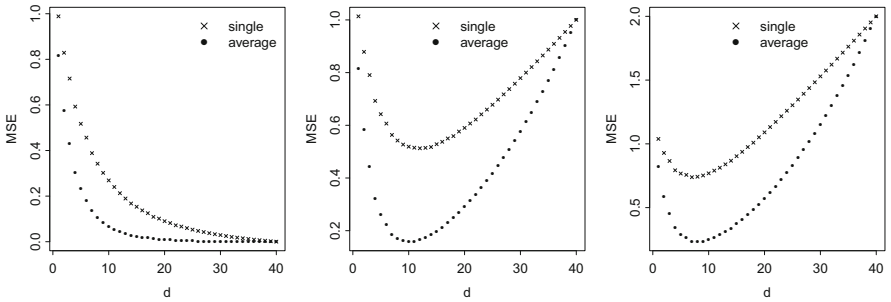

Fig. 2 MSE of averaged compressed least squares (*circle*) versus the MSE of the single estimator (*cross*) with covariance matrix $\Sigma_{i,i} = 1/i$. On the *left* with $\sigma^2 = 0$ (only bias), in the *middle* $\sigma^2 = 1/40$ and on the *right* $\sigma^2 = 1/20$. One can clearly see the quadratic improvement in terms of MSE as predicted by Theorem 4

We will not be able to reproduce the upper bound exactly for all $d \leq p$. But we can show that for any d there exists a covariance matrix Σ, such that the upper bound is reached. The idea is to consider a covariance matrix that has equal variance in the first d direction and almost zero in the remaining $p - d$. Define the diagonal covariance matrix

$$\Sigma_{i,j} = \begin{cases} 1, & \text{if } i = j \text{ and } i \leq d \\ \epsilon, & \text{if } i = j \text{ and } i > d \\ 0, & \text{if } i \neq j \end{cases}. \tag{20}$$

We show $\lim_{\epsilon \to 0} \tau = d$. For this decompose Φ into two matrices $\Phi_d \in \mathbb{R}^{d \times d}$ and $\Phi_r \in \mathbb{R}^{(p-d) \times d}$:

$$\Phi = \begin{pmatrix} \Phi_d \\ \Phi_r \end{pmatrix}.$$

The same way we define β_d, β_r, $\mathbf{X}_d$ and $\mathbf{X}_r$. Now we bound the approximation error of $\hat{\beta}_d^{\Phi}$ to extract information about λ_i / η_i. Assume a squared data matrix ($n = p$) $\mathbf{X} = \sqrt{\Sigma}$, then

$$\begin{aligned} \mathbb{E}_{\Phi}[\operatorname*{argmin}_{\gamma \in \mathbb{R}^d} \|\mathbf{X}\beta - \mathbf{X}\Phi\gamma\|_2^2] &\leq \mathbb{E}_{\Phi}[\|\mathbf{X}\beta - \mathbf{X}\Phi\Phi_d^{-1}\beta_d\|_2^2] \\ &= \mathbb{E}_{\Phi}[\|\mathbf{X}_r\beta_r - \mathbf{X}_r\Phi_r\Phi_d^{-1}\beta_d\|_2^2] \\ &= \epsilon\mathbb{E}_{\Phi}[\|\beta_r - \Phi_r\Phi_d^{-1}\beta_d\|_2^2] \\ &\leq \epsilon(2\|\beta_r\|_2^2 + 2\|\beta_d\|_2^2\mathbb{E}_{\Phi}[\|\Phi_r\|_2^2]\mathbb{E}_{\Phi}[\|\Phi_d^{-1}\|_2^2]) \\ &\leq \epsilon C, \end{aligned}$$

where C is independent of ϵ and bounded since the expectation of the smallest and largest singular values of a random projection is bounded. This means that the approximation error decreases to zero as we let $\epsilon \to 0$. Applying this to the closed form for the MSE of $\hat{\beta}_d^{\Phi}$ we have that

$$\sum_{i=1}^{p} \beta_i^2 \lambda_i \Big(1 - \frac{\lambda_i}{\eta_i}\Big) \leq \sum_{i=1}^{d} \beta_i^2 \Big(1 - \frac{\lambda_i}{\eta_i}\Big) + \epsilon \sum_{i=d+1}^{p} \beta_i^2 \Big(1 - \frac{\lambda_i}{\eta_i}\Big)$$

has to go to zero as $\epsilon \to 0$, which in turn implies

$$\lim_{\epsilon \to 0} \sum_{i=1}^{d} \beta_i^2 \Big(1 - \frac{\lambda_i}{\eta_i}\Big) = 0,$$

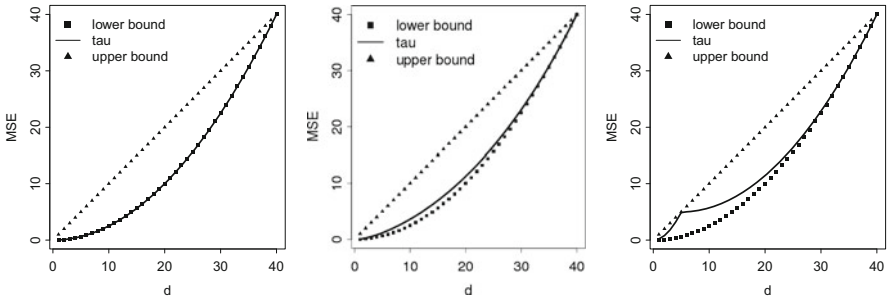

Fig. 3 Simulations of the variance factor τ (*line*) as a function of d for three different covariance matrices and in lower bound (d^2/p) and upper bound (d) (*square*, *triangle*). On the *left* ($\Sigma = I_{p\times p}$) τ as proven reaches the lower bound. In the *middle* ($\Sigma_{i,i} = 1/i$) τ reaches almost the lower bound, indicating that in most practical examples τ will be very close to the lower bound and thus averaging improves MSE substantially compared to the single estimator. On the *right* the extreme case example from (20) with $d = 5$, where τ reaches the upper bound for $d = 5$

and thus $\lim_{\epsilon\to 0} \lambda_i/\eta_i = 1$ for all $i \in \{1, \ldots, d\}$. This finally yields a limit

$$\lim_{\epsilon\to 0} \sum_{i=1}^{p} \frac{\lambda_i^2}{\eta_i^2} = d.$$

This illustrates that the lower bound d^2/p and upper bound d for the variance factor τ can both be attained. Simulations suggest that τ is usually close to the lower bound, where the variance of the estimator is reduced by a factor d/p compared to a single iteration of a compressed least-squares estimator, which is on top of the reduction in the bias error term. This shows, perhaps unsurprisingly, that averaging over random projection estimators improves the mean-squared error in a Rao–Blackwellization sense. We have quantified the improvement. In practice, one would have to decide whether to run multiple versions of a compressed least-squares regression in parallel or run a single random projection with a perhaps larger embedding dimension. The computational effort and statistical error tradeoffs will depend on the implementation but the bounds above will give a good basis for a decision (Fig. 3).

4 Discussion

We discussed some known results about the properties of compressed least-squares estimation and proposed possible tighter bounds and exact results for the mean-squared error. While the exact results do not have an explicit representation, they allow nevertheless to quantify the conservative nature of the upper bounds on the error. Moreover, the shown results allow to show a strong similarity of the

error of compressed least squares, ridge and principal component regression. We also discussed the advantages of a form of Rao–Blackwellization, where multiple compressed least-square estimators are averaged over multiple random projections. The latter averaging procedure also allows to compute the estimator trivially in a distributed way and is thus often better suited for large-scale regression analysis. The averaging methodology also motivates the use of compressed least squares in the high dimensional setting where it performs similar to ridge regression and the use of multiple random projection will reduce the variance and result in a non-random estimator in the limit, which presents a computationally attractive alternative to ridge regression.

Appendix

In this section we give proofs of the statements from the section theoretical results.

Theorem 1 ([10]) *Assume fixed design and* $\mathrm{Rank}(\mathbf{X}) \geq d$*, then the AMSE 4 can be bounded above by*

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2]] \leq \sigma^2 d + \frac{\|\mathbf{X}\beta\|_2^2}{d} + \mathrm{trace}(\mathbf{X}'\mathbf{X})\frac{\|\beta\|_2^2}{d}. \tag{21}$$

Proof (Sketch)

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2]] &= \mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\beta\|_2^2] + \sigma^2 d \\
&\leq \mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'\beta\|_2^2] + \sigma^2 d \\
&= \mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'\beta\|_2^2] + \sigma^2 d.
\end{aligned}$$

Finally a rather lengthy but straightforward calculation leads to

$$\mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'\beta\|_2^2] = \frac{\|\mathbf{X}\beta\|_2^2}{d} + \mathrm{trace}(\mathbf{X}'\mathbf{X})\frac{\|\beta\|_2^2}{d} \tag{22}$$

and thus proving the statement above. □

Theorem 2 *Assume* $\mathrm{Rank}(\mathbf{X}) \geq d$*, then the AMSE (4) can be bounded above by*

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2]] \leq \sigma^2 d + \sum_{i=1}^{p} \beta_i^2 \lambda_i w_i \tag{23}$$

where

$$w_i = \frac{(1+1/d)\lambda_i^2 + (1+2/d)\lambda_i \operatorname{trace}(\Sigma) + \operatorname{trace}(\Sigma)^2/d}{(d+2+1/d)\lambda_i^2 + 2(1+1/d)\lambda_i \operatorname{trace}(\Sigma) + \operatorname{trace}(\Sigma)^2/d}. \tag{24}$$

Proof We have for all $v \in \mathbb{R}^p$

$$\mathbb{E}_{\boldsymbol{\phi}}[\min_{\hat{\gamma}\in\mathbb{R}^d} \|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\hat{\gamma}\|_2^2] \leq \mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'v\|_2^2].$$

Which we can minimize over the whole set $\mathbb{R}^p$:

$$\mathbb{E}_{\boldsymbol{\phi}}[\min_{\hat{\gamma}\in\mathbb{R}^d} \|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\hat{\gamma}\|_2^2] \leq \min_{v\in\mathbb{R}^p} \mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'v\|_2^2].$$

This last expression we can calculate following the same path as in Theorem 1:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'v\|_2^2] =& \beta'\mathbf{X}'\mathbf{X}\beta - 2\beta'\mathbf{X}'\mathbf{X}\mathbb{E}_{\boldsymbol{\phi}}[\boldsymbol{\phi}\boldsymbol{\phi}']v \\ &+ v'\mathbb{E}_{\boldsymbol{\phi}}[\boldsymbol{\phi}\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}']v \\ =& \beta'\mathbf{X}'\mathbf{X}\beta - 2\beta'\mathbf{X}'\mathbf{X}v \\ &+ (1+1/d)v'\mathbf{X}'\mathbf{X}v + \frac{\operatorname{trace}(\Sigma)}{d}\|v\|_2^2,\end{aligned}$$

where $\Sigma = X'X$. Next we minimize the above expression w.r.t v. For this we take the derivative w.r.t. v and then we zero the whole expression. This yields

$$2\Big(1+\frac{1}{d}\Big)\Sigma v + 2\frac{\operatorname{trace}(\Sigma)}{d} I_{p\times p} v - 2\Sigma\beta = 0.$$

Hence we have

$$v = \Big(\Big(1+\frac{1}{d}\Big)\Sigma + \frac{\operatorname{trace}(\Sigma)}{d} I_{p\times p}\Big)^{-1}\Sigma\beta,$$

which is element wise equal to

$$v_i = \frac{\beta_i\lambda_i}{(1+1/d)\lambda_i + \operatorname{trace}(\Sigma)/d}.$$

Define the notation $s = \operatorname{trace}(\Sigma)$. We now plug this back into the original expression and get

$$\begin{aligned}\min_{v\in\mathbb{R}^p} \mathbb{E}_{\boldsymbol{\phi}}[\|\mathbf{X}\beta - \mathbf{X}\boldsymbol{\phi}\boldsymbol{\phi}'v\|_2^2] =& \beta'\Sigma\beta - 2\beta'\Sigma v \\ &+ (1+1/d)v'\Sigma v + \frac{s}{d}\|v\|_2^2\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{p} \beta_i^2 \lambda_i - 2\beta_i v_i \lambda_i + (1 + 1/d) v_i^2 \lambda_i + s/d v_i^2 \\
&= \sum_{i=1}^{p} \Big(\beta_i^2 \lambda_i - 2\beta_i^2 \lambda_i \frac{\lambda_i}{(1 + 1/d)\lambda_i + s/d} \\
&\quad + \beta_i^2 \lambda_i (1 + 1/d) \frac{\lambda_i^2}{((1 + 1/d)\lambda_i + s/d)^2} \\
&\quad + \beta_i^2 \lambda_i \frac{s}{d} \frac{\lambda_i}{((1 + 1/d)\lambda_i + s/d)^2} \Big) \\
&= \sum_{i=1}^{p} \beta_i^2 \lambda_i w_i,
\end{aligned}
$$

by combining the summands we get for w_i the expression mentioned in the theorem. □

Theorem 3 *Assume* $\mathrm{Rank}(\mathbf{X}) \geq d$*, then the MSE (4) equals*

$$
\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d^{\boldsymbol{\phi}}\|_2^2]] = \sigma^2 d + \sum_{i=1}^{p} \beta_i^2 \lambda_i \Big(1 - \frac{\lambda_i}{\eta_i}\Big). \tag{25}
$$

Furthermore we have

$$
\sum_{i=1}^{p} \frac{\lambda_i}{\eta_i} = d. \tag{26}
$$

Proof Calculating the expectation yields

$$
\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d\|_2^2]] = \beta' \Sigma \beta - 2\beta' \Sigma T_d^{\phi} \Sigma \beta + \mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[Y' \mathbf{X} \boldsymbol{\phi}_d^{\mathbf{X}} \mathbf{X}' Y]].
$$

Going through these terms we get:

$$
\begin{aligned}
\beta' \Sigma \beta &= \sum_{i=1}^{p} \beta_i^2 \lambda_i \\
\beta' \Sigma T_d^{\phi} \Sigma \beta &= \sum_{i=1}^{p} \beta_i^2 \frac{\lambda_i^2}{\eta_i} \\
\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[Y' \mathbf{X} \boldsymbol{\phi}_d^{\mathbf{X}} \mathbf{X}' Y]] &= \beta' \Sigma \mathbb{E}_{\boldsymbol{\phi}}[\boldsymbol{\phi}_d^{\mathbf{X}}] \Sigma \beta + \mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\varepsilon' \mathbf{X} \boldsymbol{\phi}_d^{\mathbf{X}} \mathbf{X}' \varepsilon]].
\end{aligned}
$$

The first term in the last line equals $\sum_{i=1}^{p} \beta_i^2 \lambda_i^2 / \eta_i$. The second can be calculated in two ways, both relying on the shuffling property of the trace operator:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\varepsilon' \mathbf{X} \phi_d^{\mathbf{X}} \mathbf{X}' \varepsilon]] &= \mathbb{E}_{\varepsilon}[\varepsilon' \mathbf{X} T_d^{\mathbf{X}} \mathbf{X}' \varepsilon]] = \sigma^2 \operatorname{trace}(\mathbf{X} T_d^{\mathbf{X}} \mathbf{X}') \\ &= \sigma^2 \operatorname{trace}(\Sigma T_d^{\mathbf{X}}) = \sum_{i=1}^{p} \frac{\lambda_i}{\eta_i}. \\ \mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\varepsilon' \mathbf{X} \phi_d^{\mathbf{X}} \mathbf{X}' \varepsilon]] &= \sigma^2 \mathbb{E}_{\boldsymbol{\phi}}[\operatorname{trace}(\mathbf{X} \phi_d^{\mathbf{X}} \mathbf{X}')] = \sigma^2 \mathbb{E}_{\boldsymbol{\phi}}[\operatorname{trace}(\Sigma \phi_d^{\mathbf{X}})] \\ &= \sigma^2 \mathbb{E}_{\boldsymbol{\phi}}[\operatorname{trace}(I_{d \times d})] = \sigma^2 d.\end{aligned}$$

Adding the first version to the expectation from above we get the exact expected mean-squared error. Setting both versions equal we get the equation

$$d = \sum_{i=1}^{p} \frac{\lambda_i}{\eta_i} \quad .$$

□

Theorem 4 *Assume* $\operatorname{Rank}(\mathbf{X}) \geq d$*, then there exists a real number* $\tau \in [d^2/p, d]$ *such that the AMSE of* $\hat{\beta}_d$ *can be bounded from above by*

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d\|_2^2]] \leq \sigma^2 \tau + \sum_{i=1}^{p} \beta_i^2 \lambda_i w_i^2,$$

where the w_i*'s are given as*

$$w_i = \frac{(1 + 1/d)\lambda_i^2 + (1 + 2/d)\lambda_i \operatorname{trace}(\Sigma) + \operatorname{trace}(\Sigma)^2/d}{(d + 2 + 1/d)\lambda_i^2 + 2(1 + 1/d)\lambda_i \operatorname{trace}(\Sigma) + \operatorname{trace}(\Sigma)^2/d}$$

and

$$\tau \in [d^2/p, d].$$

Proof First a simple calculation [10] using the closed form solution gives the following equation:

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbb{E}_{\varepsilon}[\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}_d\|_2^2]] = \sigma^2 \sum_{i=1}^{p} \left(\frac{\lambda_i}{\eta_i}\right)^2 + \sum_{i=1}^{p} \beta_i^2 \lambda_i \left(1 - \frac{\lambda_i}{\eta_i}\right)^2. \tag{27}$$

Now using the corollary from the last section we can bound the second term by the following way:

$$\left(1 - \frac{\lambda_i}{\eta_i}\right)^2 \leq w_i^2. \tag{28}$$

For the first term we write

$$\tau = \sum_{i=1}^{p} \left(\frac{\lambda_i}{\eta_i}\right)^2. \tag{29}$$

Now note that since $\lambda_i/\eta_i \leq 1$ we have

$$\left(\frac{\lambda_i}{\eta_i}\right)^2 \leq \frac{\lambda_i}{\eta_i} \tag{30}$$

and thus we get the upper bound by

$$\sum_{i=1}^{p} \left(\frac{\lambda_i}{\eta_i}\right)^2 \leq \sum_{i=1}^{p} \frac{\lambda_i}{\eta_i} = d. \tag{31}$$

For the lower bound of τ we consider an optimization problem. Denote $t_i = \frac{\lambda_i}{\eta_i}$, then we want to find $t \in \mathbb{R}^p$ such that

$$\sum_{i=1}^{p} t_i^2 \text{ is minimal}$$

under the restrictions that

$$\sum_{i=1}^{p} t_i = d \text{ and } 0 \leq t_i \leq 1.$$

The problem is symmetric in each coordinate and thus $t_i = c$. Plugging this into the linear sum gives $c = d/p$ and we calculate the quadratic term to give the result claimed in the theorem. □

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4), 671–687 (2003)
2. Ailon, N., Chazelle, B.: Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In: Proceedings of the 38th Annual ACM Symposium on Theory of Computing (2006)

3. Blocki, J., Blum, A., Datta, A., and Sheffet, O.: The Johnson-Lindenstrauss transform itself preserves differential privacy. In: 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS), pp. 410–419. IEEE, Washington, DC (2012)
4. Cook, R.D.: Detection of influential observation in linear regression. Technometrics **19**, 15–18 (1977)
5. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algoritm. **22**, 60–65 (2003)
6. Dhillon, P.S., Foster, D.P., Kakade, S.: A risk comparison of ordinary least squares vs ridge regression. J. Mach. Learn. Res. **14**, 1505–1511 (2013)
7. Dhillon, P., Lu, Y., Foster, D.P., Ungar, L.: New subsampling algorithms for fast least squares regression. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 26, pp. 360–368. Curran Associates, Inc. (2013). http://papers.nips.cc/paper/5105-new-subsampling-algorithms-for-fast-least-squares-regression.pdf
8. Indyk, P. and Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the 30th Annual ACM Symposium on Theory of Computing (1998)
9. Johnson, W., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Contemporary Mathematics: Conference on Modern Analysis and Probability (1984)
10. Kabán, A.: A new look at compressed ordinary least squares. In: 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 482–488 (2013). doi:10.1109/ICDMW.2013.152, ISSN:2375-9232
11. Lu, Y., Dhillon, P.S., Foster, D., Ungar, L.: Faster ridge regression via the subsampled randomized hadamard transform. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, pp. 369-377. Curran Associates Inc., Lake Tahoe (2013). http://dl.acm.org/citation.cfm?id=2999611.2999653
12. Mahoney, M.W., Drineas, P.: CUR matrix decompositions for improved data analysis. Proc. Natl. Acad. Sci. **106**(3), 697–702 (2009)
13. Maillard, O.-A., Munos, R.: Compressed least-squares regression. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 22, pp. 1213–1221. Curran Associates, Inc. (2009). http://papers.nips.cc/paper/3698-compressed-least-squares-regression.pdf
14. Marzetta, T., Tucci, G., Simon, S.: A random matrix-theoretic approach to handling singular covariance estimates. IEEE Trans. Inf. Theory **57**(9), 6256–6271 (2011)
15. McWilliams, B., Krummenacher, G., Lučić, M., and Buhmann, J.M.: Fast and robust least squares estimation in corrupted linear models. In: NIPS (2014)
16. McWilliams, B., Heinze, C., Meinshausen, N., Krummenacher, G., Vanchinathan, H.P.: Loco: distributing ridge regression with random projections. arXiv preprint arXiv:1406.3469 (2014)
17. Tropp, J.A.: Improved analysis of the subsampled randomized Hadamard transform. arXiv:1011.1595v4 [math.NA] (2010)
18. Zhang, L., Mahdavi, M., Jin, R., Yang, T., Zhu, S.: Recovering optimal solution by dual random projection. arXiv preprint arXiv:1211.3046 (2012)
19. Zhou, S., Lafferty, J., Wasserman, L.: Compressed and privacy-sensitive sparse regression. IEEE Trans. Inf. Theory. **55**(2), 846-866 (2009). doi:10.1109/TIT.2008.2009605. ISSN:0018-9448