'Compressed Least Squares Regression revisited': Appendix

A Preparations

We here recall some notation from the paper and introduce additional notation as far its is required in the proof sections following below.

Let us recall that for $X \in \mathbb{R}^{n \times d}$ with $\mathrm{rank}(X) = d \wedge n$, the singular value decomposition of X is given by

$$X = U \Sigma V^\top$$

with $U \in \mathbb{R}^{n \times d \wedge n}$, $U^\top U = I$, $\Sigma \in \mathbb{R}^{d \wedge n \times d \wedge n}$, $V \in \mathbb{R}^{d \times d \wedge n}$, $V^\top V = I$.

For $r \in \{1, \dots, d \wedge n\}$, consider

$$U = [U_r \;\; U_{r+}], \qquad \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & \Sigma_{r+} \end{bmatrix}, \qquad V = [V_r \;\; V_{r+}],$$

where U_r and $V_r \in \mathbb{R}^{d \times r}$ contain the top r left respectively right singular vectors, and Σ_r contains the corresponding singular values. The remaining singular vectors respectively singular values are contained in U_{r+}, V_{r+} and Σ_{r+}.

We also define

$$\mathcal{T}_r(X) = U_r \Sigma_r V_r^\top,$$

the best rank-r approximation of X with respect to the Frobenius norm. We write $\Delta_r = X - \mathcal{T}_r(X)$ for the 'residual'.

In general, $\mathcal{T}_r(M)$ wil be used to denote the best rank-r approximation of a matrix M.

Further, P_M denotes the orthogonal projection on the subspace spanned by the columns of M, and we write M^- for the Moore-Penrose pseudoinverse of a matrix M. The i-th column of M is denoted by $M_{:,i}$.

B Proof of Theorem 1

Condition (C1)
Let $r \in [d \wedge n]$ and let $\mathcal{V}_r \subset \mathbb{R}^d$ denote the column space of V_r. For some $\delta \in (0, 1)$, it then holds that $(1 - \delta)\|v\|_2^2 \leq \|R^\top v\|_2^2 \leq (1 + \delta)\|v\|_2^2$ for all $v \in \mathcal{V}_r$.

Condition (C2)
For $\varepsilon \in (0, 1)$, $R^\top$ is an $\varepsilon/\sqrt{r}$-Johnson-Lindenstrauss transform w.r.t. a fixed set of vectors $\mathcal{S} \subset \mathbb{R}^d$ of cardinality $2n \cdot r$, i.e. it holds that $(1 - \varepsilon/\sqrt{r})\|v\|_2^2 \leq \|v\|_2^2 \leq (1 + \varepsilon/\sqrt{r})\|v\|_2^2$ for all $v \in \mathcal{S}$.

Theorem 1. *Let $R^\top$ satisfy condition (C1) and (C2). We then have*

$$\|(I - P_{X_R})X\|_F^2 \leq \left(1 + \frac{\varepsilon^2}{(1-\delta)^2}\right) \|\Delta_r\|_F^2.$$

Proof. The idea of the proof is taken from the proof of Theorem 14 in [6]. It can be partitioned into three basic steps.

Step 1

We start by observing that

$$\|X - P_{X_R}X\|_F^2 \leq \|X - \mathcal{T}_r(P_{X_R}X)\|_F^2, \tag{1}$$

which holds as $P_{X_R}X$ is the best approximation of X by the column space of X_R: we have

$$P_{X_R}X = X_R B^*$$

where $B^* \in \mathbb{R}^{k\times d}$ is given by

$$\min_{B\in\mathbb{R}^{k\times d}} \|X - X_R B\|_F^2. \tag{2}$$

On the other hand, $\mathcal{T}_r(P_{X_R}X) = X_R B^* M_r$, where $M_r \in \mathbb{R}^{d\times r}$ contains the top r right singular values of $X_R B^*$. Since $B^* M_r$ is a feasible solution for the minimization problem (2), we conclude (1). The right hand side of (1) can in turn be bounded as follows:

$$\|X - \mathcal{T}_r(P_{X_R}X)\|_F^2 \leq \|X - \Pi X\|_F^2, \tag{3}$$

where Π is the orthogonal projection on the subspace spanned by the columns of $P_{X_R}\mathcal{T}_r(X)$, i.e.

$$\Pi = P_{P_{X_R}\mathcal{T}_r(X)}. \tag{4}$$

To see that (3) holds, consider the following optimization problem:

$$\min_{\mathrm{rank}(B)\leq r} \|X - X_R B\|_F^2.$$

Then any minimizer B^* of the above problem satisfies $X_R B^* = \mathcal{T}_r(P_{X_R}X)$ (see Propostion 1 and Lemma 14 in [2]). Noting that $\Pi = X_R M$ for some matrix $M \in \mathbb{R}^{k\times d}$ with with $\mathrm{rank}(M) \leq r$ (as $\mathcal{T}_r(X)$ has rank no more than r), M is feasible for the above optimization problem, and we conclude (3).

Step 2

In the second step, we decompose $\|X - \Pi X\|_F^2$ into one parts: one 'easy part' and one more delicate part that requires sophisticated analysis.

$$
\begin{aligned}
\|X - \Pi X\|_F^2 &= \|U\Sigma V^\top - \Pi U\Sigma V^\top\|_F^2 \\
&= \|U\Sigma - \Pi U\Sigma\|_F^2 \\
&= \|U_r\Sigma_r - \Pi U_r\Sigma_r\|_F^2 + \|U_{r+}\Sigma_{r+} - \Pi U_{r+}\Sigma_{r+}\|_F^2 \\
&= \|U_r\Sigma_r - \Pi U_r\Sigma_r\|_F^2 + \|(I-\Pi)U_{r+}\Sigma_{r+}\|_F^2 \\
&\le \|U_r\Sigma_r - \Pi U_r\Sigma_r\|_F^2 + \|U_{r+}\Sigma_{r+}\|_F^2 \\
&= \underbrace{\|U_r\Sigma_r - \Pi U_r\Sigma_r\|_F^2}_{\text{part requiring special treatment}} + \underbrace{\|X - \mathcal{T}_r(X)\|_F^2}_{\text{part that we need (up to constant)}} \qquad (5)
\end{aligned}
$$

where the inequality follows from the fact that $I - \Pi$ is an orthogonal projection.

Step 3

It remains to bound

$$\|U_r\Sigma_r - \Pi U_r\Sigma_r\|_F^2 = \|\mathcal{T}_r(X) - \Pi\mathcal{T}_r(X)\|_F^2.$$

Write $C^* = X_R^-$ and $\widetilde{C} = (\mathcal{T}_r(X)R)^-$. Note that for any matrix M of appropriate dimension, we have

$$\|M - P_{X_R}M\|_F^2 = \min_{C\in\mathbb{R}^{k\times n}} \|M - X_R C\|_F^2 = \|M - X_R C^*\|_F^2 \le \|M - X_R\widetilde{C}\|_F^2. \quad (6)$$

Moreover, observe that according to the defintion of Π in (4)

$$\Pi\mathcal{T}_r(X) = P_{P_{X_R}\mathcal{T}_r(X)}\mathcal{T}_r(X) = P_{X_R}\mathcal{T}_r(X). \quad (7)$$

Using (6) and (7), we obtain that

$$
\begin{aligned}
\|\mathcal{T}_r(X) - \Pi\mathcal{T}_r(X)\|_F^2 &= \|\mathcal{T}_r(X) - X_R(X_R)^-\mathcal{T}_r(X)\|_F^2 \\
&\le \|\mathcal{T}_r(X) - X_R\{\mathcal{T}_r(X)R\}^-\mathcal{T}_r(X)\|_F^2 \\
&= \|\mathcal{T}_r(X)^\top - \mathcal{T}_r(X)^\top\{R^\top\mathcal{T}_r(X)\}^- R^\top X^\top\|_F^2 \qquad (8)
\end{aligned}
$$

Define

$$
\begin{aligned}
b_i &= (X^\top)_{:,i} \in \mathbb{R}^d, \quad i \in [n], &(9)\\
A &= \mathcal{T}_r(X)^\top \in \mathbb{R}^{d\times n}, &(10)
\end{aligned}
$$

and consider the least squares problems

$$\min_{\lambda_i} \|b_i - A\lambda_i\|_2^2$$

with minimizer λ_i^*, $i = 1, \ldots, n$, and the corresponding *sketched regression problems* with sketching matrix $R^\top$:

$$\min_{\lambda_i} \|R^\top b_i - R^\top A\lambda_i\|_2^2.$$

with minimizer $\widetilde{\lambda}_i$, $i = 1, \ldots, n$.

It is straightforward to show that

$$A\lambda_i^* = (\mathcal{T}_r(X)^\top)_{:,i}, \quad i \in [n].$$

Observe that for the sketched regressions problems, an optimal set of coefficients is given by

$$\widetilde{\lambda}_i = \{R^\top \mathcal{T}_r(X)\}^- R^\top (X^\top)_{:,i}, \quad i \in [n],$$

so that

$$A\widetilde{\lambda}_i = \mathcal{T}_r(X)^\top \{R^\top \mathcal{T}_r(X)\}^- R^\top (X^\top)_{:,i}, \quad i \in [n].$$

Identifying terms, we see that the right hand side in (8) can be written as

$$\begin{aligned}
&\|\mathcal{T}_r(X)^\top - \mathcal{T}_r(X)^\top \{R^\top \mathcal{T}_r(X)\}^- R^\top X^\top\|_F^2 \\
&= \sum_{i=1}^n \|(\mathcal{T}_r(X)^\top)_{,i} - \mathcal{T}_r(X)^\top \{R^\top \mathcal{T}_r(X)\}^- R^\top (X^\top)_{:,i}\|_2^2 \\
&= \sum_{i=1}^n \|A(\lambda_i^* - \widetilde{\lambda}_i)\|_2^2 \\
&= \sum_{i=1}^n \|\beta_i\|_2^2, \quad \beta_i = A(\lambda_i^* - \widetilde{\lambda}_i), \ i \in [n].
\end{aligned} \tag{11}$$

Consider the residuals

$$w_i = b_i - A\lambda_i^* = (X^\top)_{:,i} - (\mathcal{T}_r(X)^\top)_{:,i}. \tag{12}$$

By analyzing the structure of (general) sketched regression problems, it can be shown that

$$V_r^\top R R^\top V_r \beta_i = V_r^\top R R^\top w_i. \tag{13}$$

The analysis leading to property (13) will be given at the end of this proof. In the sequel, we use this property in combination with conditions (C1) and (C2) to deduce the final result. We will first derive a lower bound on the l.h.s. of (13) with the help of (C1), and then we derive an upper bound on the r.h.s. by means of (C2). Combining both, we obtain an upper bound on $\sum_{i=1}^n \|\beta_i\|_2^2$ and in turn on the quantity $\|\mathcal{T}_r(X) - \Pi \mathcal{T}_r(X)\|_F^2$ that we eventually need to bound.

Note that $V_r \beta_i \subset \mathcal{V}_r$, and $\|V_r \beta_i\|_2^2 = \|\beta_i\|_2^2$. By (C1), it holds that

$$\|R^\top V_r \beta_i\|_2^2 \geq (1 - \delta)\|\beta_i\|_2^2, \ i = 1, \ldots, n.$$

Or equivalently, $\lambda_{\min}(\Gamma) \geq 1 - \delta$, where $\Gamma = V_r^\top R R^\top V_r$, and $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. Now observe that

$$\begin{aligned}
\|V_r^\top R R^\top V_r \beta_i\|_2^2 &= \beta_i^\top \Gamma^2 \beta_i \\
&\geq \lambda_{\min}(\Gamma^2)\|\beta_i\|_2^2 \\
&\geq (1 - \delta)^2 \|\beta_i\|_2^2
\end{aligned} \tag{14}$$

Next, observe that $V_{:,j}^\top w_i = 0$, $j = 1, \dots, r$, $i = 1, \dots, n$ as follows immediately from the definition of the $\{w_i\}_{i=1}^n$ in (12). We now use (C2) for the following set of vectors:

$$\mathcal{S} = \{V_{:,j} + \widetilde{w}_i,\ V_{:,j} - \widetilde{w}_i,\ i = 1, \dots, n,\ j = 1, \dots, r\},$$

where $\widetilde{w}_i = w_i / \|w_i\|_2$, $i = 1, \dots, n$. Note that $|\mathcal{S}| = 2rn$. In the next step, we will establish that the inner products between $V_{:,j}^\top w_i$, are preserved up to an additive term of $\varepsilon' \|w_i\|_2$, $j = 1, \dots, r$, $i = 1, \dots, n$, where $\varepsilon' = \varepsilon / \sqrt{r}$ according to (C2).

Recall that for arbitrary x, y, it holds that $\langle x, y \rangle = \frac{1}{4}\left(\|x + y\|_2^2 - \|x - y\|_2^2\right)$.

In virtue of the fact that $R^\top$ is a Johnson-Lindenstrauss transform for $\mathcal{S}$, we therefore have

$$\begin{aligned}
4\left\langle R^\top V_{:,j}, R^\top \widetilde{w}_i \right\rangle &= \|R^\top V_{:,j} + R^\top \widetilde{w}_i\|_2^2 - \|R^\top V_{:,j} - R^\top \widetilde{w}_i\|_2^2 \\
&\geq (1 - \varepsilon')\|V_{:,j} + \widetilde{w}_i\|_2^2 - (1 + \varepsilon')\|V_{:,j} - \widetilde{w}_i\|_2^2 \\
&= 4\left\langle V_{:,j}, \widetilde{w}_i \right\rangle - 2\varepsilon' \left(\|V_{:,j}\|_2^2 + \|\widetilde{w}_i\|_2^2\right) \\
&= 4\left\langle V_{:,j}, \widetilde{w}_i \right\rangle - 4\varepsilon'.
\end{aligned}$$

It follows that $\left\langle R^\top V_{:,j}, R^\top \widetilde{w}_i \right\rangle \geq \langle V_{:,j}, \widetilde{w}_i \rangle - \varepsilon'$ and in turn also $\left\langle R^\top V_{:,j}, R^\top w_i \right\rangle \geq \langle V_{:,j}, w_i \rangle - \varepsilon' \|w_i\|_2$ by homogeneity.

Regarding the upper bound, we argue analogously:

$$\begin{aligned}
4\left\langle R^\top V_{:,j}, R^\top \widetilde{w}_i \right\rangle &= \|R^\top V_{:,j} + R^\top \widetilde{w}_i\|_2^2 - \|R^\top V_{:,j} - R^\top \widetilde{w}_i\|_2^2 \\
&\leq (1 + \varepsilon')\|V_{:,j} + \widetilde{w}_i\|_2^2 - (1 - \varepsilon')\|V_{:,j} - \widetilde{w}_i\|_2^2 \\
&= 4\left\langle V_{:,j}, \widetilde{w}_i \right\rangle + 2\varepsilon' \left(\|V_{:,j}\|_2^2 + \|\widetilde{w}_i\|_2^2\right) \\
&= 4\left\langle V_{:,j}, \widetilde{w}_i \right\rangle + 4\varepsilon'.
\end{aligned}$$

and thus $\left\langle R^\top V_{:,j}, R^\top \widetilde{w}_i \right\rangle \leq \langle V_{:,j}, \widetilde{w}_i \rangle + \varepsilon'$ and in turn $\left\langle R^\top V_{:,j}, R^\top w_i \right\rangle \leq \langle V_{:,j}, w_i \rangle + \varepsilon' \|w_i\|_2$.

We now use this result as follows (recall that $\langle V_{:,j}, w_i \rangle = 0$, $j \in [r]$, $i \in [n]$):

$$\begin{aligned}
\sum_{i=1}^n \|V_r^\top R R^\top w_i\|_2^2 &= \sum_{i=1}^n \sum_{j=1}^r \left\langle R^\top V_{:,j}, R^\top w_i \right\rangle^2 \\
&\leq \sum_{i=1}^n \sum_{j=1}^r (\varepsilon')^2 \|w_i\|_2^2 \\
&= r(\varepsilon')^2 \sum_{i=1}^n \|w_i\|_2^2 \\
&= \varepsilon^2 \|X - \mathcal{T}_r(X)\|_F^2
\end{aligned} \tag{15}$$

where the last line is immediate from the definition of the $\{w_i\}_{i=1}^n$ in (12). Combining (5), (8), (11), (14), (15), the assertion of the theorem follows.

In order to finish the proof, it remains to establish (13) as is done below.

For $A \in \mathbb{R}^{d\times n}$, $b \in \mathbb{R}^d$, consider the least squares problem of the form

$$\min_{\lambda\in\mathbb{R}^n} \|A\lambda - b\|_2^2$$

and the corresponding sketched regression problem with sketching matrix $R^\top$

$$\min_{\lambda} \|R^\top A\lambda - R^\top b\|_2^2,$$

Let λ^* denote a minimizer of the original least squares problem and let $\widetilde{\lambda}$ denote the minimizer of the sketched least squares problem. Furthermore, we write $\mathcal{U}$ for the matrix of left singular vectors of A.

We then have the following properties:

(P1) $A\lambda^* = \mathcal{U}\alpha$,

(P2) $b = A\lambda^* + w$, with w orthogonal to $\mathcal{U}$.

(P3) $A\widetilde{\lambda} - A\lambda^* = \mathcal{U}\beta$,

for certain vectors α and β.

We now decompose the least squares error when using $\widetilde{\lambda}$:

$$\begin{aligned}
\|b - A\widetilde{\lambda}\|_2^2 &= \|b - A\lambda^* + A(\lambda^* - \widetilde{\lambda})\|_2^2 \\
&= \|b - A\lambda^*\|_2^2 + \|A(\lambda^* - \widetilde{\lambda})\|_2^2 \\
&= \|w\|_2^2 + \|\mathcal{U}\beta\|_2^2 \\
&= \|w\|_2^2 + \|\beta\|_2^2
\end{aligned}$$

Bringing the sketching matrix $R^\top$ into play, we have

$$\begin{aligned}
R^\top \mathcal{U}(\alpha + \beta) &= R^\top A\lambda^* + R^\top(A\widetilde{\lambda} - A\lambda^*) \\
&= R^\top A\widetilde{\lambda} \\
&= P_{R^\top A} R^\top b \\
&= P_{R^\top \mathcal{U}} R^\top b.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
P_{R^\top \mathcal{U}} R^\top b &= P_{R^\top \mathcal{U}} R^\top(\mathcal{U}\alpha + w) \\
&= R^\top \mathcal{U}\alpha + P_{R^\top \mathcal{U}} R^\top w.
\end{aligned}$$

Combining the previous displays, we obtain that

$$R^\top \mathcal{U}(\alpha + \beta) = R^\top \mathcal{U}\alpha + P_{R^\top \mathcal{U}} R^\top w$$

and thus

$$R^\top \mathcal{U}\beta = P_{R^\top \mathcal{U}} R^\top w.$$

Multiplying both sides with $\mathcal{U}^\top R$, this implies

$$\begin{aligned}\mathcal{U}^\top R R^\top \mathcal{U} \beta &= \mathcal{U}^\top R P_{R^\top \mathcal{U}} R^\top w \\ &= \mathcal{U}^\top R R^\top w.\end{aligned} \tag{16}$$

Note that (16) has the form as claimed in (13) with V_r playing the role of $\mathcal{U}$: according to (9), this is as it should be since V_r contains the left singular vectors of $\mathcal{T}_r(X)^\top$. The proof is thus complete.

□

C Proof of Proposition 1

Proposition 1. *Let R have entries drawn i.i.d. from a zero-mean sub-Gaussian distribution and variance k^{-1}. If $k = \Omega(\varepsilon^{-2} r \{\log(r) + \log(n)\} + \delta^{-2} \log(\delta^{-1}) r)$, then $R^\top$ satisfies conditions (C1), (C2) with probability at least $1 - \exp(-c \log(\delta^{-1}) r) - \exp(-c' \log(n\, r))$ for absolute constants $c, c' > 0$.*

Proof. Regarding (C1), it follows from the reasoning in [1] (cf. Lemma 5.1 therein) that for any fixed subspace $\mathcal{X}$ of dimension r in in $\mathbb{R}^d$, $r < k$,

$$(1-\delta)\|x\|_2^2 \leq \|R^\top x\|_2^2 \leq (1+\delta)\|x\|_2^2 \quad \text{for all } x \in \mathcal{X},$$

with probability at least

$$1 - 2(12/\delta)^r \exp(-c_0 \delta^2 k) = 1 - \exp\left(-c_0 \delta^2 k + r \log(12/\delta) + \log(2)\right),$$

for some absolute constant $c_0 > 0$. Hence, for $k = \Omega(\delta^{-2} \log(\delta^{-1}) r)$, (C1) holds with probability at least $1 - \exp(-c \log(\delta^{-1}) r)$.

Turning to (C2), in [5] (see Theorem 3.1 and the proof therein) it is shown that for any fixed $v \in \mathbb{R}^d$ and $\varepsilon' \in (0,1)$

$$\mathbf{P}\left((1-\varepsilon')\|v\|_2^2 \leq \|R^\top v\|_2^2 \leq (1+\varepsilon')\|v\|_2^2\right) \leq 2\exp(-c_0 (\varepsilon')^2 k).$$

It hence follows from the union bound that for any set $\mathcal{S}$ of vectors in $\mathbb{R}^d$, $|S| = 2n \cdot r$,

$$\mathbf{P}\left(\forall v \in \mathcal{S}: \; (1-\varepsilon')\|v\|_2^2 \leq \|R^\top v\|_2^2 \leq (1+\varepsilon')\|v\|_2^2\right) \leq \exp(-c_0 (\varepsilon')^2 k + \log(4nr)).$$

Setting $\varepsilon' = \varepsilon/\sqrt{r}$ for $\varepsilon \in (0,1)$, it follows that for $k = \Omega(\varepsilon^{-2} r \log(nr))$, condition (C2) holds with probability at least $1 - \exp(-c' \log(nr))$. This concludes the proof of the proposition. □

D Proof of Proposition 2

Proposition 2. *Consider a collection of L i.i.d. d-dimensional standard Gaussian random vectors $\{\omega_l\}_{l=1}^L$ independent of R and the estimator*

$$\widehat{\delta}_R^2 = \frac{1}{L} \sum_{l=1}^{L} \|X\omega_l - P_{X_R} X \omega_l\|_2^2.$$

Then, for any $c \in (0,1)$ *and any* $C > 1$*, as long as*

$$L \geq \max\left\{\frac{16}{(1-c)^2}, \frac{144}{(C-1)^2}\right\}$$

it holds that $\mathbf{P}\left(c\delta_R^2 \leq \widehat{\delta}_R^2 \leq C\delta_R^2\right) \geq 0.96$.

Proof. we first verify that $\|X\omega_l - P_{X_R}X\omega_l\|_2^2$ is an unbiased estimator of δ_R^2, $l = 1, \dots, L$. We have

$$\begin{aligned}
\mathbf{E}[\|X\omega_l - P_{X_R}X\omega_l\|_2^2] &= \mathbf{E}[\|(I - P_{X_R})X\omega_l\|_2^2] \\
&= \mathbf{E}[\mathrm{tr}(\omega_l^\top X^\top (I - P_{X_R})X\omega_l)] \\
&= \mathrm{tr}(X^\top (I - P_{X_R})X\, \mathbf{E}[\omega_l\omega_l^\top]) \\
&= \mathrm{tr}(X^\top (I - P_{X_R})X) \\
&= \|X - P_{X_R}X\|_F^2.
\end{aligned}$$

Concentration. We now establish concentration for the estimator $\widehat{\delta}_R^2$ by invoking results in [3, 4]. Let $\boldsymbol{\omega} \in \mathbb{R}^{d \cdot L}$ be the vector one obtains when stacking $\omega_1, \dots, \omega_L$ vertically. Let us also introduce $\Psi = X^\top (I - P_{X_R})X$ and let $\boldsymbol{\Psi} = \frac{1}{L} I_L \otimes \Psi$, where $\otimes$ denotes the Kronecker product. Then $\widehat{\delta}_R^2$ can be re-written in the following way:

$$\begin{aligned}
\boldsymbol{\omega}^\top \boldsymbol{\Psi} \boldsymbol{\omega} &= \boldsymbol{\omega}^\top \frac{1}{L} \begin{bmatrix} \Psi & 0 & \dots & \dots & 0 \\ 0 & \Psi & \dots & \dots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & \Psi \end{bmatrix} \boldsymbol{\omega} \\
&= \frac{1}{L} \sum_{l=1}^{L} \omega_l^\top \Psi \omega_l \\
&= \frac{1}{L} \sum_{l=1}^{L} \omega_l^\top X^\top (I - P_{X_R}) X \omega_l \\
&= \frac{1}{L} \sum_{l=1}^{L} \|(I - P_{X_R}) X \omega_l\|_2^2 \\
&= \widehat{\delta}_R^2.
\end{aligned}$$

In other words, $\widehat{\delta}_R^2$ can be expressed as a quadratic form in a Gaussian random vector of dimension dL and a positive definite matrix. We can thus use the following tail inequalities [3, 4]

$$\begin{aligned}
&\mathbf{P}(\boldsymbol{\omega}^\top \boldsymbol{\Psi} \boldsymbol{\omega} > \mathrm{tr}(\boldsymbol{\Psi}) + 2\sqrt{t\, \mathrm{tr}(\boldsymbol{\Psi}^2)} + 2\|\boldsymbol{\Psi}\|_2 t) \leq \exp(-t), \quad t > 0. \\
&\mathbf{P}(\boldsymbol{\omega}^\top \boldsymbol{\Psi} \boldsymbol{\omega} < \mathrm{tr}(\boldsymbol{\Psi}) - 2\sqrt{t\, \mathrm{tr}(\boldsymbol{\Psi}^2)}) \leq \exp(-t), \quad t > 0.
\end{aligned}$$

This can be re-written using the following relations:

$$\operatorname{tr}(\boldsymbol{\Psi}) = \operatorname{tr}(\Psi) = \mathbf{E}[\widehat{\delta}_R^2] = \delta_R^2, \qquad \sqrt{\operatorname{tr}(\boldsymbol{\Psi}^2)} = \|\boldsymbol{\Psi}\|_F = \frac{\|\Psi\|_F}{\sqrt{L}} \leq \frac{\operatorname{tr}(\Psi)}{\sqrt{L}},$$

$$\|\boldsymbol{\Psi}\|_2 \leq \|\boldsymbol{\Psi}\|_F \leq \operatorname{tr}(\boldsymbol{\Psi}),$$

$$\mathbf{P}\left(\widehat{\delta}_R^2 > \delta_R^2\left(1 + \frac{2(t+\sqrt{t})}{\sqrt{L}}\right)\right) \leq \exp(-t),$$
$$\mathbf{P}\left(\widehat{\delta}_R^2 < \delta_R^2\left(1 - \frac{2\sqrt{t}}{\sqrt{L}}\right)\right) \leq \exp(-t),$$

Setting $t = 4$

$$\mathbf{P}\left(\left(1 - \frac{4}{\sqrt{L}}\right)\delta_R^2 \leq \widehat{\delta}_R^2 \leq \delta_R^2\left(1 + \frac{12}{\sqrt{L}}\right)\right) \geq 1 - 2\exp(-4) \geq 0.96.$$

As a result, for any $0 < c < 1$ and any $C > 1$, as long as

$$L \geq \max\left\{\frac{16}{(1-c)^2}, \frac{144}{(C-1)^2}\right\}$$

it holds that

$$\mathbf{P}\left(c\delta_R^2 \leq \widehat{\delta}_R^2 \leq C\delta_R^2\right) \geq 1 - 2\exp(-4) \geq 0.96.$$

□

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation*, 28:253–263, 2006.

[2] F. Bunea, Y. She, and M. Wegkmap. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39:1282–1309, 2011.

[3] D. Hsu, S. Kakade, and T. Thang. A tail inequality for quadratic forms of sub-Gaussian random vectors. *Electronic Communications in Probability*, 52:1–6, 2012.

[4] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28:1302–1338, 2000.

[5] J. Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Structures and Algorithms*, 33:142–156, 2008.

[6] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science (FOCS)*, pages 143–152, 2006.