Estimating the 2-norm of the columns of a matrix

Stefano Ferioli

Randomized Matrix Computations - fall semester 2024-2025

In this project, we will derive a random estimator to estimate the (squared) 2-norm of each column of a matrix $A \in \mathbb{R}^{m\times n}$, which is a useful quantity in numerical linear algebra. Let $A = [a_1, \dots, a_n] \in \mathbb{R}^{m\times n}$, our goal is to estimate the vector

$$a = \left[\|a_1\|_2^2, \|a_2\|_2^2, \dots, \|a_n\|_2^2\right]^\top \in \mathbb{R}^n,$$

where A can only be accessed through matrix-vector products with A and $A^\top$.

Let $B = A^\top A \in \mathbb{R}^{n\times n}$ and $\omega \in \mathbb{R}^n$ be a Rademacher (random ± 1) vector, we show that

$$a = \bar{b} \quad \text{where} \quad \bar{b} := \mathbb{E}[\omega \odot B\omega],$$

where $\odot$ denotes the Hadamard product.

By matrix multiplication,

$$B_{ij} = \sum_{k=1}^{m} A_{ki}A_{kj}.$$

Then,

$$(\omega \odot B\omega)_i = \omega_i \cdot (B\omega)_i = \omega_i \sum_{j=1}^{n} B_{ij}\omega_j = \sum_{j=1}^{n}\sum_{k=1}^{m} A_{ki}A_{kj}\omega_i\omega_j, \tag{1}$$

Taking the expectation, we have

$$\mathbb{E}[(\omega \odot B\omega)_i] = \mathbb{E}\Big[\sum_{j=1}^{n}\sum_{k=1}^{m} A_{ki}A_{kj}\omega_i\omega_j\Big] = \sum_{j=1}^{n}\sum_{k=1}^{m} A_{ki}A_{kj}\mathbb{E}[\omega_i\omega_j] = \sum_{j=1}^{n}\sum_{k=1}^{m} A_{ki}A_{kj}\delta_{ij} = \sum_{k=1}^{m} A_{ki}^2 = \|a_i\|_2^2.$$

Now we have a column (squared) 2-norm estimator by taking

$$a \approx \frac{1}{\ell}\sum_{i=1}^{\ell} \omega_i \odot B\omega_i := \bar{b}_\ell, \tag{2}$$

where ω_i are i.i.d. Rademacher vectors.

Define $e_i = (\omega_i \odot B\omega_i) - a$, we show that

$$\mathbb{E}[\|\bar{b}_\ell - a\|_2^2] = \mathbb{E}\Big[\Big\|\frac{1}{\ell}\sum_{i=1}^{\ell} e_i\Big\|_2^2\Big] = \frac{1}{\ell}\left(\|B\|_F^2 - \|a\|_2^2\right).$$

First, we notice that

$$\mathbb{E}[e_i] = \mathbb{E}[(\omega_i \odot B\omega_i) - a] = \mathbb{E}[\omega_i \odot B\omega_i] - a = a - a = 0$$

and that $e_i \perp\!\!\!\perp e_j$ for $i \neq j$, since the Rademacher vectors ω_i are pairwise independent. In addition, we introduce the notation $\omega_{ji} = (\omega_i)_j$, that is the j-th entry of the vector ω_i. Note that $\omega_{ji} \perp\!\!\!\perp \omega_{kp}$ for $(i, j) \neq (k, p)$.

Then, for the first equality we have

$$\mathbb{E}[\|\bar{b}_\ell - a\|_2^2] = \mathbb{E}\Big[\Big\|\frac{1}{\ell}\sum_{i=1}^{\ell}(\omega_i \odot B\omega_i) - a\Big\|_2^2\Big] = \mathbb{E}\Big[\Big\|\frac{1}{\ell}\sum_{i=1}^{\ell}(\omega_i \odot B\omega_i) - \frac{1}{\ell}\sum_{i=1}^{\ell} a\Big\|_2^2\Big]$$
$$= \mathbb{E}\Big[\Big\|\frac{1}{\ell}\sum_{i=1}^{\ell}((\omega_i \odot B\omega_i) - a)\Big\|_2^2\Big] = \mathbb{E}\Big[\Big\|\frac{1}{\ell}\sum_{i=1}^{\ell} e_i\Big\|_2^2\Big].$$

For the second inequality,

$$\mathbb{E}\Big[\Big\|\frac{1}{\ell}\sum_{i=1}^{\ell} e_i\Big\|_2^2\Big] = \mathbb{E}\Big[\frac{1}{\ell^2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} e_i^\top e_j\Big] = \frac{1}{\ell^2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\mathbb{E}[e_i^\top e_j] = \frac{1}{\ell^2}\sum_{i=1}^{\ell}\mathbb{E}[e_i^\top e_i] + \frac{1}{\ell^2}\sum_{i=1}^{\ell}\sum_{\substack{j=1\\ j\neq i}}^{\ell}\mathbb{E}[e_i^\top e_j]$$
$$= \frac{1}{\ell^2}\sum_{i=1}^{\ell}\mathbb{E}[e_i^\top e_i] + \frac{1}{\ell^2}\sum_{i=1}^{\ell}\sum_{\substack{j=1\\ j\neq i}}^{\ell}\mathbb{E}[e_i]^\top\mathbb{E}[e_j] = \frac{1}{\ell^2}\sum_{i=1}^{\ell}\mathbb{E}[e_i^\top e_i] = \frac{1}{\ell}\mathbb{E}[e_1^\top e_1]$$
$$= \frac{1}{\ell}\mathbb{E}[((\omega_1 \odot B\omega_1) - a)^\top((\omega_1 \odot B\omega_1) - a)] = \frac{1}{\ell}\left(\mathbb{E}[(\omega_1 \odot B\omega_1)^\top(\omega_1 \odot B\omega_1)] - 2\mathbb{E}[(\omega_1 \odot B\omega_1)]^\top a + a^\top a\right)$$
$$= \frac{1}{\ell}\left(\mathbb{E}[(\omega_1 \odot B\omega_1)^\top(\omega_1 \odot B\omega_1)] - 2a^\top a + a^\top a\right) = \frac{1}{\ell}\left(\mathbb{E}[\|\omega_1 \odot B\omega_1\|_2^2] - \|a\|_2^2\right)$$

Finally, we show $\mathbb{E}[\|\omega_1 \odot B\omega_1\|_2^2] = \|B\|_F^2$. Using (1), we have

$$\mathbb{E}[\|\omega_1 \odot B\omega_1\|_2^2] = \mathbb{E}\Big[\sum_{i=1}^{n}(\omega_1 \odot B\omega_1)_i^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{n}\Big(\sum_{j=1}^{n} B_{ij}\omega_{i1}\omega_{j1}\Big)^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} B_{ij}B_{ik}\omega_{i1}^2\omega_{j1}\omega_{k1}\Big]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} B_{ij}B_{ik}\mathbb{E}[\omega_{j1}\omega_{k1}] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} B_{ij}B_{ik}\delta_{jk} = \sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}^2 = \|B\|_F^2.$$

Let $e = \sum_{i=1}^{\ell} e_i$, the idea we use to provide an error estimate is to bound the k-th moment of $\|e\|_2^2$ by using a scalar random variable.

Using Jensen's inequality, we show that for a fixed integer $k > 0$,

$$\mathbb{E}[\|e\|_2^{2k}] \leq \mathbb{E}[\|\hat{e} - \tilde{e}\|_2^{2k}] = \mathbb{E}\Big[\Big\|\sum_{i=1}^{\ell}(\hat{e}_i - \tilde{e}_i)\Big\|_2^{2k}\Big], \tag{3}$$

where $\hat{e}$ and $\tilde{e}$ are i.i.d. copies of e and $\hat{e}_i$ and $\tilde{e}_i$ are i.i.d. copies of e_i.

First, we prove the inequality.

$$\mathbb{E}[\|e\|_2^{2k}] = \mathbb{E}_{\hat{e}}[\|\hat{e}\|_2^{2k}] = \mathbb{E}_{\hat{e}}[\|\hat{e} - \mathbb{E}_{\tilde{e}}[\tilde{e}]\|_2^{2k}] \leq \mathbb{E}_{\hat{e}}[\mathbb{E}_{\tilde{e}}[\|\hat{e} - \tilde{e}]\|_2^{2k}] = \mathbb{E}[\|\hat{e} - \tilde{e}\|_2^{2k}],$$

where we can use Jensen's inequality since, for fixed $\hat{e}$, the function $x \mapsto \|\hat{e} - x\|_2^{2k}$ is convex.

To prove the equality, we simply substitute $\hat{e}$ and $\tilde{e}$ with $\sum_{i=1}^{\ell}\hat{e}_i$ and $\sum_{i=1}^{\ell}\tilde{e}_i$, respectively.

$$\mathbb{E}[\|\hat{e} - \tilde{e}\|_2^{2k}] = \mathbb{E}\Big[\Big\|\sum_{i=1}^{\ell}\hat{e}_i - \sum_{i=1}^{\ell}\tilde{e}_i\Big\|_2^{2k}\Big] = \mathbb{E}\Big[\Big\|\sum_{i=1}^{\ell}(\hat{e}_i - \tilde{e}_i)\Big\|_2^{2k}\Big].$$

Now, we denote a scalar random variable $W_i = r_i\|\hat{e}_i - \tilde{e}_i\|_2$, where r_i is another independent Rademacher random variable. We show that

$$\mathbb{E}\Big[\Big\|\sum_{i=1}^{\ell}(\hat{e}_i - \tilde{e}_i)\Big\|_2^{2k}\Big] \leq \mathbb{E}\Big[\Big(\sum_{i=1}^{\ell} W_i\Big)^{2k}\Big]. \tag{4}$$

From this it will follow that

$$\mathbb{E}[e^{\lambda\|e\|_2^2}] \leq \mathbb{E}[e^{\lambda(\sum_{i=1}^{\ell} W_i)^2}]. \tag{5}$$

Since the distribution of $(\hat{e}_i - \tilde{e}_i)$ is symmetric, we have that $(\hat{e}_i - \tilde{e}_i) \overset{d}{=} r_i(\hat{e}_i - \tilde{e}_i)$. We have

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\sum_{i=1}^{\ell}(\hat{e}_i - \tilde{e}_i)\Big\|_2^{2k}\Big] = \mathbb{E}\Big[\Big\|\sum_{i=1}^{\ell} r_i(\hat{e}_i - \tilde{e}_i)\Big\|_2^{2k}\Big] &= \mathbb{E}\Big[\Big(\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} r_i r_j (\hat{e}_i - \tilde{e}_i)^\top(\hat{e}_j - \tilde{e}_j)\Big)^k\Big] \\
&= \mathbb{E}\Big[\sum_{i\in[\ell]^k}\sum_{j\in[\ell]^k}\prod_{t=1}^{k}(r_{i_t} r_{j_t}(\hat{e}_{i_t} - \tilde{e}_{i_t})^\top(\hat{e}_{j_t} - \tilde{e}_{j_t}))\Big] \\
&= \sum_{i\in[\ell]^k}\sum_{j\in[\ell]^k}\mathbb{E}\Big[\prod_{t=1}^{k}(r_{i_t} r_{j_t})\Big]\mathbb{E}\Big[\prod_{t=1}^{k}((\hat{e}_{i_t} - \tilde{e}_{i_t})^\top(\hat{e}_{j_t} - \tilde{e}_{j_t}))\Big]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{i=1}^{\ell} W_i\Big)^{2k}\Big] = \mathbb{E}\Big[\Big(\sum_{i=1}^{\ell} r_i\|\hat{e}_i - \tilde{e}_i\|_2\Big)^{2k}\Big] &= \mathbb{E}\Big[\Big(\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} r_i r_j \|\hat{e}_i - \tilde{e}_i\|_2\|\hat{e}_j - \tilde{e}_j\|_2\Big)^k\Big] \\
&= \mathbb{E}\Big[\sum_{i\in[\ell]^k}\sum_{j\in[\ell]^k}\prod_{t=1}^{k}(r_{i_t} r_{j_t}\|\hat{e}_{i_t} - \tilde{e}_{i_t}\|_2\|\hat{e}_{j_t} - \tilde{e}_{j_t}\|_2)\Big] \\
&= \sum_{i\in[\ell]^k}\sum_{j\in[\ell]^k}\mathbb{E}\Big[\prod_{t=1}^{k}(r_{i_t} r_{j_t})\Big]\mathbb{E}\Big[\prod_{t=1}^{k}(\|\hat{e}_{i_t} - \tilde{e}_{i_t}\|_2\|\hat{e}_{j_t} - \tilde{e}_{j_t}\|_2)\Big].
\end{aligned}$$

Now, we perform a term-by-term comparison. Fix $i \in [\ell]^k$ and $j \in [\ell]^k$. If

$$\mathbb{E}\Big[\prod_{t=1}^{k}(r_{i_t} r_{j_t})\Big] = 0,$$

we are done. Otherwise, necessarily, we have

$$\mathbb{E}\Big[\prod_{t=1}^{k}(r_{i_t} r_{j_t})\Big] = 1$$

and, using Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}\Big[\prod_{t=1}^{k}((\hat{e}_{i_t} - \tilde{e}_{i_t})^\top(\hat{e}_{j_t} - \tilde{e}_{j_t}))\Big] \leq \mathbb{E}\Big[\prod_{t=1}^{k}(\|\hat{e}_{i_t} - \tilde{e}_{i_t}\|_2\|\hat{e}_{j_t} - \tilde{e}_{j_t}\|_2)\Big].$$

Now that we have proved (4), we combine it with (3) to prove (5). The derivation proceeds as follows.

$$\begin{aligned}
\mathbb{E}[e^{\lambda\|e\|_2^2}] &= \mathbb{E}\Big[\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\|e\|_2^{2k}\Big] && \text{Expand the exponential using its Taylor series} \\
&= \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[\|e\|_2^{2k}] && \text{Interchange expectation and summation} \\
&\leq \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}\Big[\Big\|\sum_{i=1}^{\ell}(\hat{e}_i - \tilde{e}_i)\Big\|_2^{2k}\Big] && \text{Use (3)} \\
&\leq \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}\Big[\Big(\sum_{i=1}^{\ell} W_i\Big)^{2k}\Big] && \text{Use (4)} \\
&= \mathbb{E}\Big[\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\Big(\sum_{i=1}^{\ell} W_i\Big)^{2k}\Big] && \text{Interchange expectation and summation} \\
&= \mathbb{E}[e^{\lambda(\sum_{i=1}^{\ell} W_i)^2}] && \text{Recognize and recombine the Taylor series of an exponential}
\end{aligned}$$

Note that, in general, it is not possible to interchange the expectation with an infinite sum. Here, this operation is possible since the terms of the series are non-negative. This is a consequence of the Monotone Convergence Theorem. Given $(X_n)_{n\in\mathbb{N}}$ a sequence of non-negative random variables, we have

$$\mathbb{E}\Big[\sum_{k=0}^{\infty} X_k\Big] = \mathbb{E}\Big[\lim_{N\to\infty}\sum_{k=0}^{N} X_k\Big] \overset{\text{MCT}}{=} \lim_{N\to\infty}\mathbb{E}\Big[\sum_{k=0}^{N} X_k\Big] = \lim_{N\to\infty}\sum_{k=0}^{N}\mathbb{E}\Big[X_k\Big] = \sum_{k=0}^{\infty}\mathbb{E}\Big[X_k\Big].$$

We continue showing that

$$\mathbb{E}[e^{\lambda|W_i|^2}] \le \mathbb{E}[e^{4\lambda\|e_i\|_2^2}] \le e^{\tilde{c}\lambda\,\mathrm{trace}(D)},$$

for $D := (B-\mathrm{diag}(a))^\top(B-\mathrm{diag}(a))$ and for some constant $\tilde{c}$ with $\mathrm{diag}(a)$ denoting the diagonal matrix whose diagonal entries are the entries of a.

To prove the first inequality, we have

$$|W_i|^2 = |r_i\|\hat{e}_i - \tilde{e}_i\|_2|^2 = |r_i|^2\|\hat{e}_i - \tilde{e}_i\|_2^2 = \|\hat{e}_i - \tilde{e}_i\|_2^2 \overset{\star}{\le} 2\|\hat{e}_i\|_2^2 + 2\|\tilde{e}_i\|_2^2,$$

where $\overset{\star}{\le}$ holds since $-\hat{e}_i^\top\tilde{e}_i \le \|\hat{e}_i\|_2^2 + \|\tilde{e}_i\|_2^2$ and $\|\hat{e}_i - \tilde{e}_i\|_2^2 = \|\hat{e}_i\|_2^2 - \hat{e}_i^\top\tilde{e}_i + \|\tilde{e}_i\|_2^2 \le 2\|\hat{e}_i\|_2^2 + 2\|\tilde{e}_i\|_2^2$.

Since $x \mapsto e^{\lambda x}$ is non-decreasing for $\lambda \ge 0$, we have

$$e^{\lambda|W_i|^2} \le e^{\lambda(2\|\hat{e}_i\|_2^2 + 2\|\tilde{e}_i\|_2^2)}$$

and

$$\mathbb{E}[e^{\lambda|W_i|^2}] \le \mathbb{E}[e^{\lambda(2\|\hat{e}_i\|_2^2 + 2\|\tilde{e}_i\|_2^2)}] \overset{\perp\!\!\!\perp}{=} \mathbb{E}[e^{2\lambda\|\hat{e}_i\|_2^2}]\mathbb{E}[e^{2\lambda\|\tilde{e}_i\|_2^2}] = \mathbb{E}[e^{2\lambda\|e_i\|_2^2}]^2 \le \mathbb{E}[e^{4\lambda\|e_i\|_2^2}], \tag{6}$$

by Jensen's inequality, since $x \mapsto x^2$ is convex.

To prove the second inequality, we first relate $\|e_i\|_2^2$ to the hutchinson estimator applied to D. We have

$$\|e_i\|_2^2 = \|(\omega_i \odot B\omega_i) - a\|_2^2 = \sum_{k=1}^{n}((\omega_i \odot B\omega_i)_k - a_k)^2 = \sum_{k=1}^{n}\Big(\sum_{j=1}^{n} B_{kj}\omega_{ki}\omega_{ji} - a_k\Big)^2$$
$$= \sum_{k=1}^{n}\Big(\sum_{j=1}^{n} B_{kj}\omega_{ki}\omega_{ji} - \omega_{ki}^2 a_k\Big)^2 = \sum_{k=1}^{n}\Big(\sum_{j=1}^{n}(B_{kj} - a_k\delta_{kj})\omega_{ki}\omega_{ji}\Big)^2 = \sum_{k=1}^{n}\Big(\sum_{j=1}^{n}(B_{kj} - a_k\delta_{kj})\omega_{ji}\Big)^2$$
$$= \sum_{j=1}^{n}\sum_{p=1}^{n}\omega_{ji}\omega_{pi}\sum_{k=1}^{n}(B_{kj} - a_k\delta_{kj})(B_{kp} - a_k\delta_{kp})$$

where $\sum_{k=1}^{n}(B_{kj} - a_k\delta_{kj})(B_{kp} - a_k\delta_{kp})$ is the (j,p)-th entry of $(B-\mathrm{diag}(a))^\top(B-\mathrm{diag}(a)) = D$.

Follows that $\|e_i\|_2^2 = \omega_i^\top D\omega_i$.

Then, we prove that $\mathbb{E}[e^{4\lambda\omega_i^\top D\omega_i}] \le e^{\tilde{c}\lambda\,\mathrm{trace}(D)}$ for some constant $\tilde{c}$. To do this, we introduce a lemma that we use without proof.

Lemma: Let $\mathcal{Z}(B) = \omega^\top B\omega - \mathrm{trace}(B)$ be the error of Hutchinson's trace estimator with one Rademacher random vector ω. For absolute constants c, C, we have

$$\mathbb{E}[e^{\lambda\mathcal{Z}(B)}] \le e^{C\lambda^2\|B\|_F^2} \quad \text{for all} \quad |\lambda| \le \frac{c}{\|B\|_2}. \tag{7}$$

Using the lemma we have

$$\mathbb{E}[e^{4\lambda\omega_i^\top D\omega_i}] = e^{4\lambda\,\mathrm{trace}(D)}\mathbb{E}[e^{4\lambda\mathcal{Z}(D)}] \le e^{4\lambda\,\mathrm{trace}(D)}e^{16C\lambda^2\|D\|_F^2} \quad \text{for all} \quad 0 \le 4\lambda \le \frac{c}{\|D\|_2} \tag{8}$$

for absolute constants c, C.

We work now on the bound over λ. Recall that, since D is symmetric positive semi-definite matrix, given the eigenvalues of D, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, we have $\|D\|_F^2 = \sum_{i=1}^n \lambda_i^2$, $\|D\|_2 = \lambda_1$, and $\operatorname{trace}(D) = \sum_{i=1}^n \lambda_i$. Then we have

$$
\begin{aligned}
0 \leq 4\lambda &\leq \frac{c}{\|D\|_2}, \\
0 \leq 4\lambda\|D\|_F^2 &\leq c\frac{\|D\|_F^2}{\|D\|_2}, \\
0 \leq 4\lambda\|D\|_F^2 &\leq c\frac{\sum_{i=1}^n \lambda_i^2}{\lambda_1}, \\
0 \leq 4\lambda\|D\|_F^2 &\leq c\sum_{i=1}^n \frac{\lambda_i}{\lambda_1}\lambda_i \leq c\sum_{i=1}^n \lambda_i = c\operatorname{trace}(D),
\end{aligned}
$$

since $\lambda_i/\lambda_1 \leq 1$ for all $i \in [n]$.

Using this in (8) and using (6), we obtain

$$
\mathbb{E}[e^{\lambda|W_i|^2}] \leq \mathbb{E}[e^{4\lambda\|e_i\|_2^2}] = \mathbb{E}[e^{4\lambda\omega_i^\top D\omega_i}] \leq e^{4\lambda\operatorname{trace}(D)}e^{16C\lambda^2\|D\|_F^2} \leq e^{4\lambda\operatorname{trace}(D)}e^{4Cc\lambda\operatorname{trace}(D)} = e^{\tilde{c}\lambda\operatorname{trace}(D)}
$$
$$
\text{for all} \quad 0 \leq 4\lambda \leq \frac{c}{\|D\|_2},
$$

with $\tilde{c} = 4 + 4Cc$. This is equivalent to

$$
\mathbb{E}[e^{\lambda|W_i|^2}] \leq e^{\tilde{c}\lambda\operatorname{trace}(D)} \quad \text{for all} \quad 0 \leq \lambda \leq \frac{c}{4\|D\|_2}.
$$

Since $\|D\|_2 \leq \operatorname{trace}(D)$ and $c/(4\operatorname{trace}(D)) \leq c/(4\|D\|_2)$, this implies

$$
\mathbb{E}[e^{\lambda|W_i|^2}] \leq e^{\tilde{c}\lambda\operatorname{trace}(D)} \quad \text{for all} \quad 0 \leq \lambda \leq \frac{c}{4\operatorname{trace}(D)}.
$$

Now, let $\tilde{C} = \max\{4 + 4Cc, 4/c\}$, then

$$
\mathbb{E}[e^{\lambda|W_i|^2}] \leq e^{\tilde{C}\lambda\operatorname{trace}(D)} \quad \text{for all} \quad 0 \leq \lambda \leq \frac{1}{\tilde{C}\operatorname{trace}(D)}.
$$

Then, $|W_i|^2$ satisfies 2.7.1.c in [1] and it is thus sub-exponential with $K_3 = \tilde{C}\operatorname{trace}(D)$. Then, using Lemma 2.7.6 from [1], $|W_i|$ is sub-Gaussian with parameter $\sqrt{\tilde{C}}\sqrt{\operatorname{trace}(D)}$. Since, $W_i \leq |W_i|$, also W_i is sub-Gaussian with parameter $\sqrt{\tilde{C}}\sqrt{\operatorname{trace}(D)}$. Using the result from L2S25 or Proposition 2.6.1 from [1], $\sum_{i=1}^\ell W_i$ is sub-Gaussian with parameter $C_\star\sqrt{\ell}\sqrt{\operatorname{trace}(D)}$ as it is the sum of independent sub-Gaussian random variables. Using Lemma 2.7.6 from [1], $(\sum_{i=1}^\ell W_i)^2$ is sub-exponential with parameter $C_\star^2\ell\operatorname{trace}(D)$. Then, by (5), also $\|e\|_2^2$ is sub-exponential with parameter $C_\star^2\ell\operatorname{trace}(D)$. Using Lemma 2.7.6 from [1], $\|e\|_2$ is sub-Gaussian with parameter $C_\star\sqrt{\ell}\sqrt{\operatorname{trace}(D)}$. Using $\sqrt{\operatorname{trace}(D)} = \|B - \operatorname{diag}(a)\|_F$ and Proposition 2.5.2.i from [1], we obtain

$$
\begin{aligned}
\mathbb{P}(\|a - \bar{b}_\ell\|_2 \geq t) &= \mathbb{P}\Big(\Big\|\frac{1}{\ell}e\Big\|_2 \geq t\Big) = \mathbb{P}(\|e\|_2 \geq \ell t) \\
&\leq 2\exp\Big(-\frac{\ell^2t^2}{C_\star^2\ell\|B - \operatorname{diag}(a)\|_F^2}\Big) = 2\exp\Big(-\frac{\ell t^2}{C_\star^2\|B - \operatorname{diag}(a)\|_F^2}\Big).
\end{aligned}
$$

We solve

$$
2\exp\Big(-\frac{\ell t^2}{C_\star^2\|B - \operatorname{diag}(a)\|_F^2}\Big) \leq \delta,
$$

for $0 < \delta \leq 1$, and we obtain

$$
t \geq C_\star\sqrt{\frac{\log(2/\delta)}{\ell}}\|B - \operatorname{diag}(a)\|_F,
$$

so that

$$\mathbb{P}\Big(\|a - \bar{b}_\ell\|_2 \geq C_\star \sqrt{\frac{\log(2/\delta)}{\ell}} \|B - \mathrm{diag}(a)\|_F\Big) \leq \delta. \tag{9}$$

Now, we implement estimator (2) in MATLAB. We plot the error against ℓ and we investigate the tightness of the probability bound (9). We test the algorithm on a matrix $A = U\Sigma V^\top$ where $U, V \in \mathbb{R}^{1000\times 1000}$ are independent random orthogonal matrices and $\Sigma_{ij} = \delta_{ij}/i$ for $(i,j) \in [1000]^2$.

A possible implementation of estimator (2) is the following.

```
function a_est = estimator(A, ell)
    B = A' * A;
    w = rademacher(size(A,2), ell);
    a_est = mean(w .* (B*w),2);
end

function w = rademacher(m, n)
    w = 2 * (rand(m, n) > 0.5) - 1;
end

function A = generate_matrix(n)
    Sigma = diag((1:n).^(-1));
    U = orth(randn(n));
    V = orth(randn(n));
    A = U * Sigma * V';
end
```

We plot the error $\|a - \bar{b}_\ell\|_2$ for different values of ℓ in logarithmic scale. We see in Figure 1 that the error asymptotically behaves like $1/\sqrt{\ell}$.

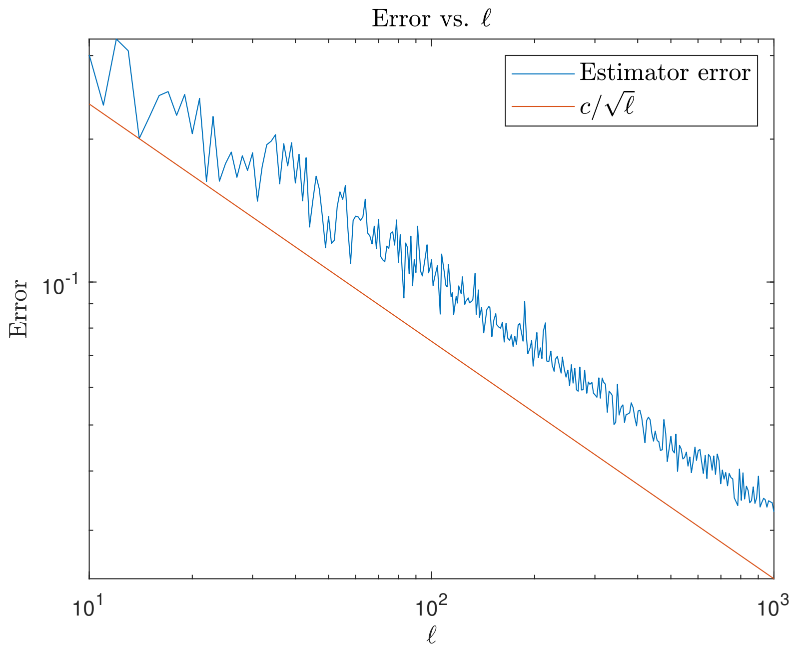
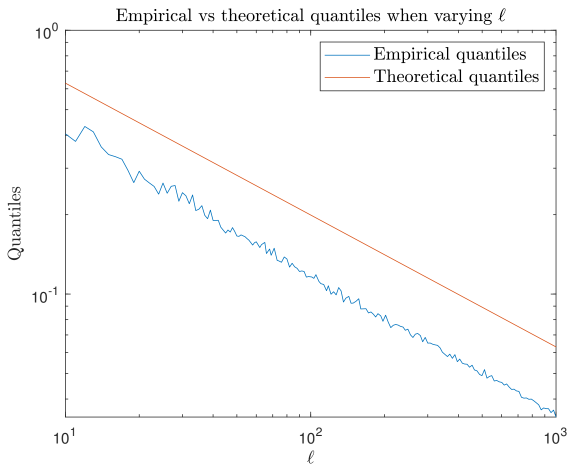
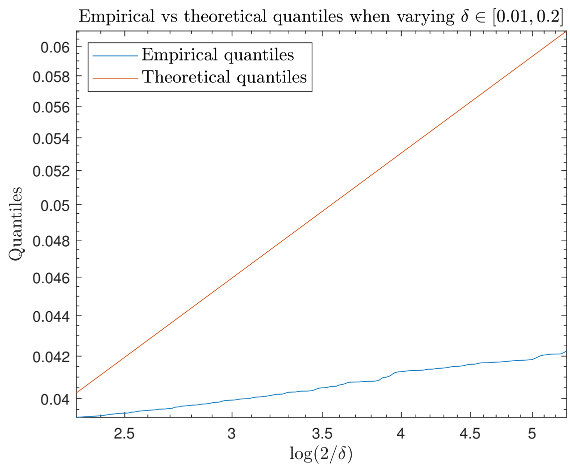


Figure 1

We plot the theoretical and the empirical quantiles when varying ℓ and δ. In the first case (Figure 2a), we plot the quantiles against ℓ in logarithmic scale. We can see here that the bound (9) is tight for ℓ as the two lines are parallel. In the second case (Figure 2b), to make the comparison easier, we transform the x axis with $\delta \mapsto \log(2/\delta)$ and we plot the quantiles against the transformed δ in logarithmic scale. In this case, the bound appears not to be tight as the two lines have different slopes.



(a)



(b)

Figure 2

Finally, we report the code we utilized to generate the plots.

```
clc, clear all, close all

rng(42);

n = 1000;

A = generate_matrix(n);
a_real = sum(A .* A, 1)';

nnorm = norm(A'*A - diag(a_real), 'fro');
f = @(delta, ell) sqrt(log(2./delta)./ell) .* nnorm;

%% Plot error vs ell

ell_values = unique(floor(logspace(1,3,400)).');
error = zeros(length(ell_values), 1);
for ell_idx = 1:length(ell_values)
    ell = ell_values(ell_idx);

    a_est = estimator(A, ell);

    error(ell_idx) = norm(a_real - a_est);
end

figure;
plot(ell_values, error, '-', 'DisplayName', 'Estimator error')
hold on
plot(ell_values, 0.75./ell_values.^0.5, '-', 'DisplayName', '$c/\sqrt{\ell}$')

xlabel('$\ell$', 'Interpreter', 'latex', 'FontSize', 12);
ylabel('Error', 'Interpreter', 'latex', 'FontSize', 12);
legend('Location', 'northeast', 'Interpreter', 'latex', 'FontSize', 12);
title('Error vs. $\ell$', 'Interpreter', 'latex', 'FontSize', 12);
set(gca, 'XScale', 'log');
set(gca, 'YScale', 'log');
print('-depsc2','-vector','error_vs_ell')

```

```matlab
%% Plot empirical vs theoretical quantiles when varying ell

ell_values = unique(floor(logspace(1,3,200)).');
delta = 0.05;
times = 30;
sample_quantiles = zeros(length(ell_values), 1);
for ell_idx = 1:length(ell_values)
    ell = ell_values(ell_idx);

    error = zeros(times, 1);
    for times_idx = 1:times
        a_est = estimator(A, ell);
        error(times_idx) = norm(a_real - a_est);
    end

    sample_quantiles(ell_idx) = quantile(error,1-delta);
end

figure;
plot(ell_values, sample_quantiles, '-', 'DisplayName', 'Empirical quantiles')
hold on
plot(ell_values, f(delta, ell_values), '-', 'DisplayName', 'Theoretical quantiles')

xlabel('$\ell$', 'Interpreter', 'latex', 'FontSize', 12);
ylabel('Quantiles', 'Interpreter', 'latex', 'FontSize', 12);
legend('Location', 'northeast', 'Interpreter', 'latex', 'FontSize', 12);
title('Empirical vs theoretical quantiles when varying $\ell$', 'Interpreter', ...
    'latex', 'FontSize', 12);
set(gca, 'XScale', 'log');
set(gca, 'YScale', 'log');
print('-depsc2','-vector','quantiles_ell')

%% Plot empirical vs theoretical quantiles when varying delta
% for efficiency we compute 1000 errors and then we compute the empirical quantiles

ell = 750;
deltas = logspace(log10(0.01),log10(0.2),200);
times = 1000;
errors = zeros(times, 1);
sample_quantiles = zeros(length(deltas), 1);

for times_idx = 1:times
    a_est = estimator(A, ell);
    errors(times_idx) = norm(a_real - a_est);
end

for delta_idx = 1:length(deltas)
    delta = deltas(delta_idx);
    sample_quantiles(delta_idx) = quantile(errors,1-delta);
end

figure;
plot(log(2./deltas), sample_quantiles, '-', 'DisplayName', 'Empirical quantiles')
hold on
plot(log(2./deltas), 0.7*f(deltas, ell), '-', 'DisplayName', 'Theoretical quantiles')

xlabel('$\log(2/\delta)$', 'Interpreter', 'latex', 'FontSize', 12);
ylabel('Quantiles', 'Interpreter', 'latex', 'FontSize', 12);
legend('Location', 'northwest', 'Interpreter', 'latex', 'FontSize', 12);
title('Empirical vs theoretical quantiles when varying $\delta\in[0.01,0.2]$', ...
    'Interpreter', 'latex', 'FontSize', 12);
set(gca, 'XScale', 'log');
set(gca, 'YScale', 'log');
print('-depsc2','-vector','quantiles_delta')
```

References

[1] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press.