NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/nla.2045

Triangular truncation and its extremal matrices

Weiqi Zhou*,†

Jacobs University, Bremen

SUMMARY

The triangular truncation operator is a linear transformation that maps a given matrix to its strictly lower triangular part. The operator norm (with respect to the matrix spectral norm) of the triangular truncation is known to have logarithmic dependence on the dimension, and such dependence is usually illustrated by a specific Toeplitz matrix. However, the precise value of this operator norm as well as on which matrices can it be attained is still unclear. In this article, we describe a simple way of constructing matrices whose strictly lower triangular part has logarithmically larger spectral norm. The construction also leads to a sharp estimate that is very close to the actual operator norm of the triangular truncation. This research is directly motivated by our studies on the convergence theory of the Kaczmarz type method (or equivalently, the Gauß-Seidel type method), the corresponding application of which is also included. Copyright © 2016 John Wiley & Sons, Ltd.

Received 20 October 2015; Revised 13 January 2016; Accepted 5 February 2016

KEY WORDS: triangular truncation; hadamard multiplier; volterra operator; SOR method; kaczmarz method

1. INTRODUCTION

Let $\mathcal{M}_n$ be the space of $n \times n$ $(2 \leqslant n < \infty)$ complex matrices equipped with the spectral norm $\|\cdot\|$, define L to be the $n \times n$ matrix with 1 in its strictly lower triangular part and 0 elsewhere:

$$L_{jk} := \begin{cases} 1 & j > k \\ 0 & j \leqslant k \end{cases}.$$

then the triangular truncation $\mathcal{L}$ on $\mathcal{M}_n$ is defined as

$$\mathcal{L}(A) := L \circ A, \quad A \in \mathcal{M}_n,$$

where $\circ$ stands for the Hadamard product.

Hereafter A always denotes an arbitrary element in $\mathcal{M}_n$. We shall also use $\|\cdot\|$ for the operator norm (with respect to the matrix spectral norm) on $\mathcal{M}_n$ and the Euclidean norm on the underlying vector space $\mathbb{C}^n$, these situations should be clear from the context without raising any confusion.

It is known (see [1, 2]) that

$$\lim_{n\to\infty} \frac{\|\mathcal{L}\|}{\ln n} = \frac{1}{\pi}.$$

In fact, an upper bound of form $\|\mathcal{L}_n\| \leqslant c \ln n$ for some uniform constant c independent of n (hereafter the notation c will always be used for various uniform and absolute constants independent

*Correspondence to: Weiqi Zhou, Jacobs University Bremen, Research I, Campus Ring 1, 28759, Bremen, Germany.
†E-mail: Weiqi.Zhou@outlook.com

of the context) can be established in several ways, for example, see [1, 3–7] and [2, Chapter 4]. In particular, [1] gives the following upper bound which seems to be the tightest one at the moment:

$$\|\mathcal{L}\| \leqslant 1 + \frac{1}{\pi}(1 + \ln n). \tag{1}$$

On the other hand, a lower bound of form $\|\mathcal{L}\| \geqslant c \ln n$ is usually illustrated by the Toeplitz matrices $B \in \mathcal{M}_n$ defined as

$$B_{jk} = \begin{cases} \frac{1}{j-k} & j \neq k \\ 0 & j = k \end{cases},$$

i.e.,

$$B = \begin{pmatrix} 0 & -1 & -\frac{1}{2} & -\frac{1}{3} & \dots \\ 1 & 0 & -1 & -\frac{1}{2} & \dots \\ \frac{1}{2} & 1 & 0 & -1 & \dots \\ \frac{1}{3} & \frac{1}{2} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}. \tag{2}$$

By the standard Toeplitz theory (see [8, Chapter 1]), $\|B\|$ (for all n) are uniformly bounded by the $L^\infty((0, 2\pi))$ norm of its corresponding Toeplitz symbol, which can be found via the Fourier series. In this case here, the Toeplitz symbol is

$$\sum_{k=1}^{\infty} \frac{1}{k} e^{ik\theta} - \sum_{k=1}^{\infty} \frac{1}{k} e^{-ik\theta} = -\ln\left(\frac{1 - e^{i\theta}}{1 - e^{-i\theta}}\right) = -\ln(-\cos\theta - i\sin\theta) = \pi - \theta,$$

which is bounded on the torus, while

$$\frac{\|\mathcal{L}(B)\|}{\|B\|} \geqslant \frac{4}{5\pi} \ln n, \tag{3}$$

can be shown (see [2, Chapter 4]) by applying B to the vector $(0, 1, 1, 1, \dots, 1)$. Slightly different computations are available in [3, 7, 9–11] as well.

The Toeplitz theory also relates the Riesz projection of Fourier series to the triangular truncation operator $\mathcal{L}$ on the Toeplitz class. It is a classical theorem of Riesz that the norm of the Riesz projection is bounded by cp on L^p for $1 < p < \infty$ (see [12, Chapter 4.20]), this boundedness carries over to the Schatten-p class (see [13, Chapter 11.10, p1137] or [14, Chapter 3.6, p118]).

Given a fixed $R \in (0, \ln n/\pi)$, let us call a matrix A has the **R-truncation property** if

$$\|\mathcal{L}(A)\| \geqslant R\|A\|.$$

For example, the Toeplitz matrix in Equation (2) has the $(4 \ln n)/(5\pi)$-truncation property.

In this article we would like to present two more classes of matrices that possess the R-truncation property for some R close to $\ln n/\pi$, to be introduced as Example 1 and Example 2 in the next two sections. Estimating these examples also improves both bounds in Equations (1) and (3) to

$$\frac{1}{2n} \sum_{k=1}^{n-1} \csc \frac{(2k-1)\pi}{4n-2} \leqslant \|\mathcal{L}\| \leqslant \frac{1}{2} + \frac{1}{2n} \sum_{k=1}^{n} \left| \csc \frac{(2k+1)\pi}{2n} \right|,$$

or equivalently:

$$\frac{1}{\pi}(\ln n + \gamma) - O\left(\frac{\ln n}{n}\right) \leqslant \|\mathcal{L}\| \leqslant \frac{1}{\pi}\left(\ln n + \frac{1}{4}\right) + 1,$$

where γ is the Euler constant and $O(\cdot)$ is the big O notation. Based on our examples, one can further construct more matrices that have the R-truncation property for R of scale $O(\ln n)$, in particular, they need not be Toeplitz matrices.

DOI: 10.1002/nla

As triangular matrices naturally enters the iteration formula of the Gauß-Seidel type method (see [15] and related references), such results have a direct application to the convergence theory of the Gauß-Seidel type method: the error reduction rate may suffer from a logarithmic deterioration. The corresponding connections and the related numerical results are included in the last section.

2. AN IMPROVED UPPER BOUND

We introduce the matrix $T \in \mathcal{M}_n$ as

$$T_{jk} := \operatorname{sgn}(k-j)i = \begin{cases} 0 & j=k \\ -i & j>k \\ i & j<k \end{cases},$$

where sgn denotes the sign function, i.e.,

$$T = \begin{pmatrix} 0 & i & i & i & \dots \\ -i & 0 & i & i & \dots \\ -i & -i & 0 & i & \dots \\ -i & -i & -i & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

Define the transformation $\mathcal{T}$ to be

$$\mathcal{T}(A) = T \circ A,$$

and let $\mathcal{D}$ be the main diagonal projection on $\mathcal{M}_n$, i.e.,

$$\mathcal{D}(A)_{jk} = \begin{cases} A_{jk} & j=k \\ 0 & j \neq k \end{cases}.$$

$\mathcal{D}$ and $\mathcal{T}$ are obviously related to $\mathcal{L}$ through the Cartesian decomposition of $\mathcal{L}$:

$$\mathcal{L}(A) = \Re(\mathcal{L})(A) + i\Im(\mathcal{L})(A) = \frac{1}{2}(A - \mathcal{D}(A)) + \frac{i}{2}\mathcal{T}(A).$$

Therefore, $\mathcal{L}$ can be written as

$$\mathcal{L} = \frac{1}{2}(\mathcal{I} + i\mathcal{T}) - \frac{1}{2}\mathcal{D},$$

where $\mathcal{I}$ is the identity operator on $\mathcal{M}_n$.

The notion of $\mathcal{T}$ was also used in [6, 10, 14] without an explicit name. Since $\mathcal{T}$ maps the real part of $\mathcal{L}$ to its imaginary part, we may call it the (harmonic) conjugate transform. In particular, applying $\mathcal{T}$ to matrices in the Toeplitz class with vanishing main diagonal (see [10]) is equivalent to applying the discrete Hilbert transform to their corresponding Toeplitz symbol.

As pointed out in [6, 10], $\|\mathcal{T}\|$ provides a very good approximation of $\|\mathcal{L}\|$:

Lemma 1

$$\frac{1}{2}\|\mathcal{T}\| \leqslant \|\mathcal{L}\| \leqslant \frac{1}{2} + \frac{1}{2}\|\mathcal{I} + i\mathcal{T}\|.$$

We omit the proof since it is almost trivial by using the triangular inequality after noticing that

$$\|\mathcal{D}(A)\| = \max_k |(A\xi_k, \xi_k)| \leqslant \|A\|,$$

where ξ_k is the k-th canonical basis (column) vector in $\mathbb{C}^n$. The notation ξ_k will be reused later.

Write tr for the trace and $*$ for the adjoint operation. Also denote $\|\cdot\|_p$ as the Schatten-p norm. For finite matrices, $\|\cdot\|_p$ is just the ℓ^p norm of the singular values. In particular, $\|\cdot\|$ coincides with $\|\cdot\|_\infty$ on the finite dimensional spaces $\mathcal{M}_n$. Consequently, if we write the dual space of $\mathcal{M}_n$ as $\mathcal{M}'_n$ (with element $A' \in \mathcal{M}'_n$), then the following lemma is a direct corollary of the duality on the Schatten-p class, a proof of which can be found in [2, Theorem 1.12]:

DOI: 10.1002/nla

Lemma 2
$\mathcal{M}_n'$ is the space of $n \times n$ complex matrices equipped with $\| \cdot \|_1$, the dual pairing is given by

$$\langle A, A' \rangle = \operatorname{tr}(A(A')^*).$$

It is also easy to verify that

Lemma 3
$\mathcal{T}^*$ acts on $\mathcal{M}_n'$ by

$$\mathcal{T}^*(A') = -T \circ A'.$$

The next two lemmas have already been stated in [6] under a more general framework. To have a clear description of our construction, we repeat them here but using different techniques.

Lemma 4
Let

$$\zeta := e^{\pi i/n}, \quad \omega := e^{2\pi i/n},$$

then the k-th ($k = 1, 2, \ldots, n$) unit column eigenvector of T is

$$v_k = \frac{1}{\sqrt{n}} \left(1, \zeta^{-1}\omega^{-k}, \left(\zeta^{-1}\omega^{-k}\right)^2, \ldots, (\zeta^{-1}\omega^{-k})^{n-1} \right),$$

and the corresponding eigenvalue is

$$\tau_k = -i \sum_{j=1}^{n-1} \left(\zeta\omega^k\right)^j = \cot \frac{(2k+1)\pi}{2n}.$$

Proof
Notice that T is a skew circulant matrix, it can hence be made into a circulant matrix (which is diagonalizable by the Fourier matrix) after being conjugated with

$$D_\zeta := \operatorname{diag}\left(1,\ \zeta,\ \zeta^2,\ \ldots,\ \zeta^{n-1}\right). \tag{4}$$

The result follows by invoking the standard formula for the eigenstructure of circulant matrices, detailed computations are accessible in either [6] or [16]. □

The notations ζ and ω in the above proof will be followed hereafter. Knowing the explicit eigenstructure of T leads to the conclusion that

Lemma 5

$$\|\mathcal{T}\| = \frac{1}{n} \|T\|_1.$$

Proof
Let $x, y \in \mathbb{C}^n$ be arbitrary unit column vectors, and introduce the notion P_{xy} for the rank 1 norm 1 operator

$$P_{xy} := xy^*, \tag{5}$$

One may check that the eigendecomposition of T can be written as

$$T = \sum_{k=1}^{n} \tau_k P_{v_k v_k} = \sum_{k=1}^{n} \tau_k D_{v_k} E D_{v_k}^*, \tag{6}$$

DOI: 10.1002/nla

where E is the all one matrix (i.e., the identity element with respect to the Hadamard product) and D_{v_k} is the diagonal matrix with its j-th diagonal entry being the j-th entry of v_k. Since the Hadamard product commutes with diagonal scaling, we then have

$$\mathcal{T}(P_{xy}) = T \circ P_{xy} = \sum_{k=1}^{n} \tau_k D_{v_k} P_{xy} D_{v_k}^*,$$

Lemma 4 shows that $\sqrt{n} D_{v_k}$ is unitary, hence

$$\|\mathcal{T}(P_{xy})\|_1 = \left\| \sum_{k=1}^{n} \tau_k D_{v_k} P_{xy} D_{v_k}^* \right\|_1 \leqslant \sum_{k=1}^{n} |\tau_k| \|D_{v_k} P_{xy} D_{v_k}^*\|_1 = \frac{1}{n} \|P_{xy}\|_1 \sum_{k=1}^{n} |\tau_k| = \frac{1}{n} \|T\|_1.$$

By the duality in Lemma 2 and 3, together with the Hölder inequality we get

$$\|\mathcal{T}(A)\| = \sup_{x,y} |y^* \mathcal{T}(A) x| = \sup_{x,y} |(A \mathcal{T}^*(P_{xy}))| \leqslant \sup_{x,y} \|A\| \|\mathcal{T}(P_{xy})\|_1 \leqslant \frac{1}{n} \|A\| \|T\|_1.$$

Dividing by $\|A\|$ at both sides and taking supreme over A we obtain

$$\|\mathcal{T}\| \leqslant \frac{1}{n} \|T\|_1.$$

That

$$\|\mathcal{T}\| \geqslant \frac{1}{n} \|T\|_1,$$

is illustrated by Example 1 below. □

Example 1
Define the matrix S through the eigendecompostion:

$$S := \sum_{k=1}^{n} \mathrm{sgn}(\tau_k) P_{v_k v_k} = \sum_{k=1}^{n} \mathrm{sgn}(\tau_k) D_{v_k} E D_{v_k}^*, \tag{7}$$

i.e., S has the same set of eigenvectors as T, but with its eigenvalues being the image of the spectrum of T under the sgn mapping. Obviously

$$\|S\| = 1.$$

Let

$$\xi = \frac{1}{\sqrt{n}} (1, 1, 1, \ldots),$$

and $P_{\xi\xi}$ as defined in Equation (5), then

$$\|\mathcal{T}(S)\| \geqslant |\xi^* \mathcal{T}(S) \xi| = |\mathrm{tr}(S \mathcal{T}^*(P_{\xi\xi}))| = \left| -\frac{1}{n} \mathrm{tr}(S(T \circ E)) \right| = \frac{1}{n} |\mathrm{tr}(ST)|.$$

Since S by definition has the same set of eigenvectors as T, the above can be further computed as

$$\frac{1}{n} |(ST)| = \frac{1}{n} \left| \mathrm{tr} \left(\sum_{k=1}^{n} \tau_k \mathrm{sgn}(\tau_k) P_{v_k v_k} \right) \right| = \frac{1}{n} \sum_{k=1}^{n} |\tau_k| = \frac{1}{n} \|T\|_1.$$

Therefore

$$\|\mathcal{T}\| \geqslant \frac{\|\mathcal{T}(S)\|}{\|S\|} = \|\mathcal{T}(S)\| = \frac{1}{n} \|T\|_1.$$

The notation ξ will also be reused in the sequel.

Recall Lemma 1, we can conclude that the family of matrices in Example 1 is very close to those matrices on which $\|\mathcal{L}\|$ shall be attained.

Moreover, since

$$\frac{\pi}{n}\|T\|_1 = \frac{\pi}{n}\sum_{k=1}^{n}\left|\cot\frac{(2k+1)\pi}{2n}\right|, \tag{8}$$

is a Riemann sum that approximates the integral $\int |\cot\theta| d\theta$, we may apply elementary calculus to obtain a marginal improvement over the previous estimate in Equation (1):

Lemma 6

$$\|\mathcal{L}\| \leqslant 1 + \frac{1}{\pi}\left(\frac{1}{4} + \ln n\right).$$

Proof

Rewrite the upper bound in Lemma 1 as

$$\begin{aligned}\|\mathcal{L}\| &\leqslant \frac{1}{2} + \frac{1}{2}\|(\mathcal{I} + i\mathcal{T})\| \\ &= \frac{1}{2} + \frac{1}{2}\left|1 + i\frac{1}{n}\sum_{k=1}^{n}\right|\cot\frac{(2k+1)\pi}{2n}|| \\ &\leqslant \frac{1}{2} + \frac{1}{2n}\sum_{k=1}^{n}\left|\csc\frac{(2k+1)\pi}{2n}\right| \\ &= \frac{1}{2} + \frac{1}{n}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor}\csc\frac{(2k+1)\pi}{2n}.\end{aligned}$$

We can bound the summation by

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor}\left|\csc\frac{(2k+1)\pi}{2n}\right| &= \frac{1}{2n}\csc\frac{\pi}{2n} + \frac{1}{\pi}\left(\frac{\pi}{2n}\csc\frac{\pi}{2n} + \frac{\pi}{n}\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}\csc\frac{(2k+1)\pi}{2n}\right) \\ &\leqslant \frac{1}{2} + \frac{1}{\pi}\int_{\frac{\pi}{2n}}^{\frac{\pi}{2}}\csc\theta d\theta \\ &= \frac{1}{2} + \frac{1}{\pi}\ln\cot\frac{\pi}{4n} \\ &\leqslant \frac{1}{2} + \frac{1}{\pi}\left(\frac{1}{4} + \ln n\right),\end{aligned}$$

where in both inequalities we used the fact that $n \geqslant 2$. In particular, the last inequality holds since

$$\ln\cot\frac{\pi}{4n} - \ln n = \ln\frac{\cos\frac{\pi}{4n}}{n\sin\frac{\pi}{4n}} \to \ln\frac{4}{\pi} \approx 0.24 < \frac{1}{4},$$

where the convergence is monotonic from below as $n \to \infty$.

Combining the above computations together establishes this lemma. □

3. AN IMPROVED LOWER BOUND

The derivation of the lower bound is very similar to what we have shown in the last section, first one may verify that

Lemma 7
$\mathcal{L}^*$ acts on $A' \in \mathcal{M}_n'$ by

$$\mathcal{L}^*(A') = L^* \circ A'.$$

Then we have:

Example 2
Let the singular value decomposition of L be

$$L = U\Sigma V,$$

and define

$$\tilde{S} := UV,$$

clearly

$$\|\mathcal{L}(\tilde{S})\| \geqslant |\xi^*\mathcal{L}(\tilde{S})\xi| = |\mathrm{tr}(\tilde{S}\mathcal{L}^*(P_{\xi\xi}))| = \left|\frac{1}{n}\mathrm{tr}(\tilde{S}(L^* \circ E))\right| = \frac{1}{n}|\mathrm{tr}(\tilde{S}L^*)| = \frac{1}{n}\mathrm{tr}(U\Sigma U^*) = \frac{1}{n}\|L\|_1.$$

Since $\tilde{S}$ is unitary, this implies that

$$\|\mathcal{L}\| \geqslant \frac{\|\mathcal{L}(\tilde{S})\|}{\|\tilde{S}\|} = \frac{1}{n}\|L\|_1. \tag{9}$$

$\|L\|_1$ can also be computed explicitly:

Lemma 8
The k-th singular value of L is:

$$\sigma_k = \begin{cases} 0 & k = 0 \\ \frac{1}{2}\csc\frac{2k\pi-\pi}{4n-2} & k = 1, 2, \ldots, n-1 \end{cases}.$$

Proof
Straight forward computation shows

$$(LL^*)_{jk} = \min(j,k) - 1,$$

i.e.,

$$LL^* = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 1 & \ldots & 1 & 1 \\ 0 & 1 & 2 & \ldots & 2 & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 2 & \ldots & n-2 & n-2 \\ 0 & 1 & 2 & \ldots & n-2 & n-1 \end{pmatrix}.$$

Hence 0 is an eigenvalue, while the other eigenvalues are the eigenvalues of the following principal submatrix (which we denote as K) obtained by removing the first row and the first column of LL^*, i.e.,

$$K := \begin{pmatrix} 1 & 1 & \ldots & 1 & 1 \\ 1 & 2 & \ldots & 2 & 2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2 & \ldots & n-2 & n-2 \\ 1 & 2 & \ldots & n-2 & n-1 \end{pmatrix}.$$

Observe that

$$K^{-1} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

i.e., K^{-1} differs from the tridiagonal matrix $\{-1, 2, -1\}$ by 1 at the bottom right corner.

Writing β_k as the k-th eigenvalue of K^{-1} and $(u_1, u_2, \ldots, u_{n-1})$ the corresponding eigenvector, we arrive at the following recurrence relation and boundary conditions:

$$\begin{cases} u_{j-1} + (\beta_k - 2)u_j + u_{j+1} = 0 \\ u_0 = 0 \\ u_{n-1} = u_n \end{cases},$$

which has solution

$$u_j = c_1 r_1^j + c_2 r_2^j,$$

where r_1, r_2 are the roots of the associated characteristic equation

$$x^2 + (\beta_k - 2)x + 1 = 0.$$

Using the boundary conditions

$$\begin{cases} u_0 = 0 \\ u_{n-1} = u_n \end{cases},$$

together with the following relation from the characteristic equation

$$r_1 r_2 = 1,$$

one easily gets that

$$\begin{cases} c_1 + c_2 = 0 \\ (r_1)^{2n-1} + 1 = 0 \end{cases},$$

and therefore

$$\beta_k = 2 - \frac{u_2}{u_1} = 2 - \frac{r_1^2 - \frac{1}{r_1^2}}{r_1 - \frac{1}{r_1}} = 2 - 2\cos(\arg(r_1)) = 4\sin^2 \frac{(2k-1)\pi}{4n-2}.$$

Since σ_k (for $k = 1, 2, \ldots, n-1$) is the square root of the k-th eigenvalue of LL^*, which is the reciprocal of the k-th eigenvalue of K^{-1}, this completes the proof. □

The above method also applies to L^*L and thus one can derive explicit form for the singular vectors and compute $\tilde{S}$ as well, these elementary computations are left to the interested readers.

We may estimate the growth of $\|L\|_1/n$ and conclude that

Lemma 9

$$\|\mathcal{L}\| \geqslant \frac{1}{n}\|L\|_1 \geqslant \frac{1}{\pi}(\ln n + \gamma) - O\left(\frac{\ln n}{n}\right).$$

DOI: 10.1002/nla

Proof
The first inequality has already been shown in Equation (9). For the second one, we apply Lemma 8 to get

$$\begin{aligned}
\frac{1}{n}\|L\|_1 &= \frac{1}{2n}\sum_{k=1}^{n-1}\frac{1}{\sin\frac{2k\pi-\pi}{4n-2}} \\
&\geqslant \frac{1}{2n}\sum_{k=1}^{n-1}\frac{4n-2}{2k\pi-\pi} \\
&= \frac{1}{\pi}\left(2-\frac{1}{n}\right)\sum_{k=1}^{n-1}\frac{1}{2k-1} \\
&\geqslant \frac{2}{\pi}\sum_{k=1}^{n}\frac{1}{2k}-\frac{1}{\pi n}\sum_{k=1}^{n-1}\frac{1}{2k-1} \\
&\geqslant \frac{1}{\pi}(\ln n+\gamma)-O\left(\frac{\ln n}{n}\right).
\end{aligned}$$

□

4. A SHARP ESTIMATE OF $\|\mathcal{L}\|$

Combining the results in the previous sections, we are able to pin down $\|\mathcal{L}\|$ into a narrow range whose upper and lower bound can be computed explicitly:

Theorem 1
There holds

$$\frac{1}{2n}\sum_{k=1}^{n-1}\csc\frac{(2k-1)\pi}{4n-2} \leqslant \|\mathcal{L}\| \leqslant \frac{1}{2}+\frac{1}{2n}\sum_{k=1}^{n}\left|\csc\frac{(2k+1)\pi}{2n}\right|, \tag{10}$$

in particular,

$$\frac{1}{\pi}(\ln n+\gamma)-O\left(\frac{\ln n}{n}\right) \leqslant \|\mathcal{L}\| \leqslant \frac{1}{\pi}\left(\ln n+\frac{1}{4}\right)+1.$$

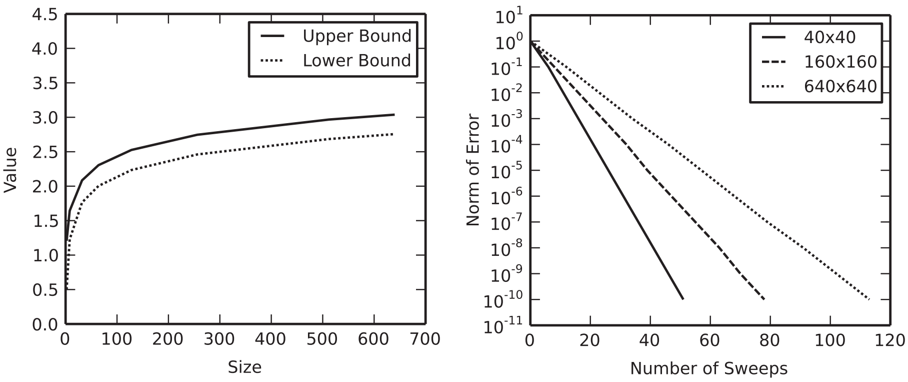
Simple numerical computation shows that the difference between upper and lower bounds in Equation (10) starts off at around 0.71 when $n = 2$ and quickly stabilizes to around 0.28 after $n > 500$ (see Figure 1(a)). Example 1 and Example 2 are among those matrices that illustrate the above theorem.

Obviously one can shift the spectrum of the matrices in Example 1 and 2 a bit to obtain other matrices with similar truncation property. Moreover, since the Hadamard product commutes with diagonal scaling, the R-truncation properties of S and $\tilde{S}$ remain invariant if we conjugate them by diagonal unitary matrices. For instance, conjugating the matrix S in Example 1 with the matrix D_ξ defined in Equation (4) yields a circulant matrix which has the same R-truncation property as S, i.e.,

$$\frac{\|\mathcal{L}(D_\xi S D_\xi^*)\|}{\|D_\xi S D_\xi^*\|} = \frac{\|\mathcal{L}(S)\|}{\|S\|}.$$

In general, left or right multiplying S and $\tilde{S}$ by a diagonal matrix changes R by at most the spectral conditioning of the multiplied diagonal matrix, while a bounded additive perturbation also has bounded impact on the truncation. Therefore it is easy to construct various matrices with the $O(\ln n)$-truncation property based on these prototypes.

DOI: 10.1002/nla



(a)

(b)

Figure 1. (a) The Estimate of $\|\mathcal{L}\|$, (b) The Logarithmic Deterioration.

5. APPLICATION TO THE GAUß-SEIDEL TYPE METHODS

Consider a consistent linear system

$$By = b,$$

where $B \in \mathcal{M}_n$ is Hermitian positive definite. Eigenvalues of B are denoted as

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n > 0,$$

and the spectral condition number κ of B is as usual defined to be

$$\kappa := \frac{\lambda_1}{\lambda_n}.$$

Let L, U (we apologize for reusing L here for a different meaning) respectively be the strictly lower and upper triangular parts of B, let D be the main diagonal part of B, and denote I as the identity matrix. Then one step (which corresponds to a full sweep of the entire system) of the Gauß-Seidel iteration reads

$$y^{(k+1)} = (D + L)^{-1} \left(b - Uy^k\right). \tag{11}$$

Let $\tilde{y}$ be the solution of the system, one may thus replace b with $B\tilde{y}$ and rewrites the above formula as

$$y^{(k+1)} - \tilde{y} = -(D + L)^{-1}U \left(y^k - \tilde{y}\right) = \left(I - (D + L)^{-1}B\right) \left(y^k - \tilde{y}\right).$$

Therefore the error reduction matrix has form

$$Q := I - (D + L)^{-1}B.$$

To estimate the error reduction rate, we need to switch to the energy norm $\|\cdot\|_B$:

$$\|y\|_B^2 := (By, y),$$

which is a well defined norm since B is Hermitian positive definite. $\|y\|_B$ is equivalent to the Euclidean norm $\|y\|$ as

$$\sqrt{\lambda_n}\|y\| \leqslant \|y\|_B \leqslant \sqrt{\lambda_1}\|y\|.$$

Proposition 1

$$\|Qy\|_B^2 \leqslant \left(1 - \frac{c}{\kappa^2 \ln^2 n}\right) \|y\|_B^2, \quad y \in \mathbb{C}^n. \tag{12}$$

Proof
Observe that the smallest singular value of D admits a lower bound

$$\sigma_{\min}(D) = \min_k |D_{kk}| = \min_k |(B\xi_k, \xi_k)| \geqslant \lambda_n,$$

while similarly for its largest singular value we have

$$\sigma_{\max}(D) = \max_k |D_{kk}| = \max_k |(B\xi_k, \xi_k)| \leqslant \lambda_1.$$

Straightforward computation shows that

$$\begin{aligned} \|Qy\|_B^2 &= (By, y) - (B(D+U)^{-1}(D+U+D+L-B)(D+L)^{-1}By, y) \\ &= \|y\|_B^2 - (B(D+U)^{-1}D(D+L)^{-1}By, y) \\ &\leqslant \|y\|_B^2 - \frac{(By, y)}{\|B^{-1}\|\|D^{-1}\|\|D+L\|^2} \\ &\leqslant \|y\|_B^2 - \frac{\lambda_n^2}{(1+\|\mathcal{L}\|)^2\lambda_1^2}\|y\|_B^2 \\ &\leqslant \left(1 - \frac{c}{\kappa^2 \ln^2 n}\right)\|y\|_B^2, \end{aligned}$$

where by Theorem 1 the constant c is asymptotically π^2. This completes the proof. □

The cases of successive over relaxation method (i.e., SOR, which adds a relaxation parameter in each step) and symmetric successive over relaxation method (i.e., SSOR, which sweeps the rows in the reversed order after a SOR sweep) are rather similar, to keep our notations simple we will leave out the details here, some discussions can be found in [15] and [17].

Of particular interest is the Kaczmarz method. For a linear system

$$A_{n\times m}x = b,$$

Each step of the Kaczmarz method is an orthogonal projection from the current solution onto the affine plane defined by the j-th (with j runs over the row indices in a circulant order) component equation

$$a_j x = b_j,$$

i.e.,

$$x^{(k+1)} = (I - \mathcal{P}_j)x^{(k)} + b_j \frac{a_j^*}{\|a_j\|^2}, \quad j = (k \pmod{n}) + 1,$$

where $\mathcal{P}_j$ is the orthogonal projection onto the span of a_j^*.

Therefore the corresponding error reduction matrix Q_{kacz} for a full sweep of the linear system is

$$Q_{kacz} = (I - \mathcal{P}_n)(I - \mathcal{P}_{n-1})\dots(I - \mathcal{P}_1). \tag{13}$$

Let $\tilde{A}$ be the row normalized version of A, i.e., $\tilde{A}$ is obtained by left multiplying A with

$$\operatorname{diag}\left(\frac{1}{\|a_1\|}, \frac{1}{\|a_2\|}, \dots, \frac{1}{\|a_n\|}\right).$$

Denote the j-th row in $\tilde{A}$ as $\tilde{a}_j$, and set

$$B = \tilde{A}\tilde{A}^*,$$

and L again its strictly lower triangular part, then straight forward computation would also reveal that

$$Q_{kacz} = I - \tilde{A}^*(I - L)^{-1}\tilde{A}. \tag{14}$$

Indeed, since L is nilpotent, the right hand side of the above can be written as:

$$I - \tilde{A}^*(I + L)^{-1}\tilde{A} = I - \tilde{A}^*(I + (-L) + (-L)^2 + \ldots + (-L)^{n-1})\tilde{A},$$

while by definition of $\mathcal{P}_j$ we have:

$$\begin{aligned} \mathcal{P}_{j_1}\mathcal{P}_{j_2}\ldots\mathcal{P}_{j_k} &= \tilde{A}^*\xi_{j_1}\xi_{j_1}^*\tilde{A}\tilde{A}^*\xi_{j_2}\xi_{j_2}^*\tilde{A}\ldots\tilde{A}^*\xi_{j_k}\xi_{j_k}^*\tilde{A} \\ &= \tilde{A}^*\xi_{j_1}\left(\tilde{a}_{j_1}\tilde{a}_{j_2}^*\tilde{a}_{j_2}\ldots\tilde{a}_{j_k}^*\right)\xi_{j_k}^*\tilde{A} \\ &= \tilde{A}^*\xi_{j_1}\left(B_{j_1j_2}B_{j_2j_3}\ldots B_{j_{k-1}j_k}\right)\xi_{j_k}^*\tilde{A} \\ &= \tilde{A}^*\xi_{j_1}\left(L_{j_1j_2}L_{j_2j_3}\ldots L_{j_{k-1}j_k}\right)\xi_{j_k}^*\tilde{A}, \end{aligned}$$

where the last equality holds since the ordering of the projections in Equation (13) imposed the condition that:

$$j_1 > j_2 > j_3 > \ldots > j_k. \tag{15}$$

Hence for any fixed k, enumerating $j_1, j_2, \ldots, j_k$ through all possible values establishes

$$(-1)^k \sum_{j_1,j_2,\ldots,j_k} \mathcal{P}_{j_1}\mathcal{P}_{j_2}\ldots\mathcal{P}_{j_k} = (-1)^k \sum_{j_1,j_k} \tilde{A}^*\left(\xi_{j_1}L_{j_1j_k}^{k-1}\xi_{j_k}^*\right)\tilde{A} = (-1)^k\tilde{A}^*L^{k-1}\tilde{A},$$

which leads to the equivalence between Equation (13) and Equation (14).

In fact, applying the Kaczmarz method to $Ax = b$ is the same as applying the Gauß-Seidel method to the normalized system $By = \tilde{b}$ ($\tilde{b}$ is defined analogously as $\tilde{A}$), which should be clear by comparing Q_{kacz} with Q. A different derivation can also be found in [18, p.210].

Therefore the estimate in Proposition 1 carries over to $\|Q_{kacz}\|$ for the Kaczmarz method (or equivalently, the Schwarz algorithm for the domain decomposition method) up to a bounded factor depending on the spectral conditioning of A. This has been stated in a series of papers in [17, 19, 20].

Moreover, since all the iteration steps of the Kaczmarz method take place in range(A^*), choosing $x^{(0)} \in$ range(A^*) guarantees the convergence even if B is only positive semi-definite: one simply replaces λ_n with λ_r (r being the rank of B) in the corresponding estimate in Proposition 1 (see [21, p.125] for the formula of the limit vector from the iteration, and see [17] for the estimate for the error reduction rate).

To show that such a logarithmic upper bound as in Proposition 1 is attainable in practice, we can set

$$B = I + \frac{1}{2}S,$$

where S is the matrix defined in Example 1 and for convenience we assume n is a large even number.

It is clear from the eigenstructure of S that B is Hermitian, and its spectrum consists of only $3/2$ and $1/2$, with each appearing precisely $n/2$ times. Thus B is positive definite with $\kappa = 3$. Moreover, as Lemma 4 shows the symmetry of the entries in the eigenvectors of S, one may verify that $(S\xi_k, \xi_k)$ is 0 for all $k = 1, 2, \ldots, n$, hence the main diagonal of B is just I, and it can be factored into AA^* with the row norm of each row in A normalized to 1.

It is also easy to see from Lemma 4 that the spectrum of T lies symmetrically on the real axis, therefore a similar computation as in Example 1 shows that

$$\|I + L\| \geqslant \|\frac{i}{2}((I + U) - (I + L))\| = \frac{1}{2}\|\mathcal{T}(B)\| \geqslant \frac{1}{2}|(\mathcal{T}(B)\xi, \xi)| = \frac{1}{2n}|\text{tr}(BT)| = \frac{1}{4n}\|T\|_1,$$

where the extra $1/2$ factor in the last inequality comes from the fact that half of the spectrum from T cancels out in the summation $\operatorname{tr}(BT)$.

Now applying a similar technique as in Lemma 8 leads to

$$\frac{1}{n}\|T\|_1 = \frac{1}{n}\sum_{k=1}^{n}\left|\cot\frac{(2k+1)\pi}{2n}\right| \geqslant \frac{2}{\pi}\int_{\frac{\pi}{2n}}^{\frac{\pi}{2}}\cot\theta d\theta = \frac{2}{\pi}\left|\ln\sin\left(\frac{\pi}{2n}\right)\right| \geqslant \frac{1}{\pi}\ln n,$$

where the last inequality holds since $n \geqslant 2$.

Consequently, if we take u to be the eigenvector for the smallest eigenvalue of $(I+U)^{-1}I(I+L)^{-1}$ and set

$$y = \frac{B^{-1}u}{\|B^{-1}u\|},$$

then

$$\|Q^*Q\| \geqslant |(Q^*Qy, y)| \geqslant 1 - \frac{|(I+U)^{-1}I(I+L)^{-1}u, u)|}{\|B^{-1}u\|^2} \geqslant 1 - \frac{c}{\ln^2 n},$$

where the constant c is bounded by $36\pi^2$ and we have absorbed κ into it.

Since the error reduction matrix for the SSOR method is simply Q^*Q (if the relaxation parameter is set to 1), the above suggests that the bound in Proposition 1 is achievable at every step of the iteration if the SSOR method is applied to B with the initial error vector $y^{(0)} - \tilde{y}$ chosen as the eigenvector for the largest eigenvalue of Q^*Q.

Now consider a new matrix

$$G = \begin{pmatrix} A \\ \forall \end{pmatrix},$$

where the set of rows in $\forall$ is the same as in A, but arranged in reversed order. Then obviously applying the Gauß-Seidel method to GG^* is same as applying the Kaczmarz method to G, and also equivalent as applying the above SSOR method (with the relaxation parameter 1) to B.

Therefore the estimate in Proposition 1 is indeed not further improvable beyond a constant factor, the logarithmic deterioration can happen in all Gauß-Seidel type methods, including the SOR and SSOR method, the Kaczmarz method, and the Schwarz method if the matrix B has the R-truncation property with R being logarithmically large (for example, a similar situation also holds for the Toeplitz matrix defined in Equation (2)).

A simple way to circumvent this situation is to adopt a random row ordering in the iteration. That a random row ordering may accelerate the iteration has been observed in various applications (for example see [22, 23]), the recent result in [24] has drawn even more attention in this direction. It is provable that the random row ordering can, in expectation, remove the $\ln n$ factor in Proposition 1, and thus make the iteration only depend on the spectral conditioning of the linear system, details can be found in [15].

As a concluding remark, we include the result from a numerical test as Figure 1(b). It illustrates the error reduction when applying the Gauß-Seidel method to $B = 2I + S$ of growing sizes. In all tests we solve the homogeneous system $By = 0$ with the initial vector set to $B^{-1}\xi/\|B^{-1}\xi\|$, thus the initial errors in all tests are 1, and the graph shows that it takes increasing number of full sweeps (as defined in Equation (11)) to reach the same error level as the size grows, which suggests the existence of a deterioration here.

REFERENCES

1. Angelos JR, Cowen CC, Narayan SK. Triangular truncation and finding the norm of a Hadamard multiplier. *Linear Algebra and Its Applications* 1992; **170**:117–135.
2. Davidson KR. *Nest Algebras*. Longman Scientific and Technical: Harlow, England, 1988.
3. Bhatia R. Pinching, trimming, truncating, and averaging of matrices. *American Mathematical Monthly* 2000; **107**(7):602–608.

DOI: 10.1002/nla

4. Chobanyan S, Levental S, Salehi H. On the constant in Meńshov-Rademacher inequality. *Journal of Inequalities and Applications* 2006; **68969**.
5. Kwapień S, Pełczyński A. The main triangle projection in matrix spaces and its applications. *Studia Mathematica* 1970; **34**(1):43–67.
6. Mathias R. The hadamard operator norm of a circulant and applications. *SIAM Journal on Matrix Analysis and Applications* 1993; **14**(4):1152–1167.
7. Oswald P. On the convergence rate of SOR: A worst case estimate. *Computing* 1994; **52**(3):245–255.
8. Böttcher A, Grudsky SM. *Toeplitz Matrices, Asymptotic Linear Algebra, and Functional analysis*. Springer: Basel, Switzerland, 2000.
9. Choi M-D. Tricks or treats with the Hilbert matrix. *American Mathematical Monthly* 1983; **90**(5):301–312.
10. Gasch J, Gilbert JE. Triangularization of Hankel operators and the bilinear Hilbert transform. *Contemporary Mathematics* 1999; **247**:235–248.
11. Hardy GH, Littlewood JE, Pólya G. *Inequalities*. Cambridge University Press: London, England, 1934.
12. King FW. *Hilbert Transforms*. Cambridge University Press: New York, USA, 2009.
13. Dunford N, Schwartz JT, Bade WG, Bartle RG. *Linear Operators*. Wiley-Interscience: New York, 1971.
14. Gohberg I, Krein MG. *Theory and Applications of Volterra Operators in Hilbert space*. American Mathematical Society, 1970.
15. Oswald P, Zhou W. Random reordering in SOR-type methods. *arXiv preprint* 2015. (1510.04727).
16. Huckle T. *Circulant-skewcirculant matrices as preconditioners for Hermitian Toeplitz systems*, 1990.
17. Oswald P, Zhou W. Convergence analysis for Kaczmarz-type methods in a Hilbert space framework. *Linear Algebra and its Applications* 2015; **478**:131–161.
18. Hackbusch W. *Iterative Solution of Large Sparse Systems of Equations*. Springer: New York, USA, 1994.
19. Griebel M, Oswald P. On the abstract theory of additive and multiplicative Schwarz algorithms. *Numerische Mathematik* 1995; **70**(2):163–180.
20. Griebel M, Oswald P. Greedy and randomized versions of the multiplicative Schwarz method. *Linear Algebra and its Applications* 2012; **437**(7):1596–1610.
21. Galántai A. *Projectors and Projection Methods*. Springer: New York, USA, 2004.
22. Varga R. Orderings of the Successive Overrelaxation Scheme. *Pacific Journal of Mathematics* 1959; **9**(3):925–939.
23. Young D. Iterative methods for solving partial difference equations of elliptic type. *Transactions of the American Mathematical Society* 1954; **76**(1):92–111.
24. Strohmer T, Vershynin R. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications* 2009; **15**(2):262–278.