Componentwise condition numbers of random sparse matrices

Cheung, Dennis; Cucker, Felipe

Published in:
SIAM Journal on Matrix Analysis and Applications

Published: 01/01/2009

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1137/080729463

Publication details:
Cheung, D., & Cucker, F. (2009). Componentwise condition numbers of random sparse matrices. *SIAM Journal on Matrix Analysis and Applications*, *31*(2), 721-731. https://doi.org/10.1137/080729463

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 721–731

COMPONENTWISE CONDITION NUMBERS OF RANDOM SPARSE MATRICES*

DENNIS CHEUNG† AND FELIPE CUCKER‡

Abstract. We prove an $\mathcal{O}(\log n)$ bound for the expected value of the logarithm of the componentwise (and, a fortiori, the mixed) condition number of a random sparse $n \times n$ matrix. As a consequence, small bounds on the average loss of accuracy for triangular linear systems follow.

Key words. sparse matrices, triangular systems, average loss of precision

AMS subject classifications. Primary, 65F35; Secondary, 65G50

DOI. 10.1137/080729463

1. Introduction.

Triangular systems of linear equations provide one of the few examples in numerical linear algebra where a gap occurs between stability analysis and everyday practice. One could summarize this gap as follows:

> *Triangular systems of equations are generally solved to high accuracy in spite of being, in general, ill-conditioned.*

This state of affairs had been already noted by Wilkinson in [9, p. 105]: "In practice one almost invariably finds that if L is ill-conditioned, so that $\|L\|\|L^{-1}\| \gg 1$, then the computed solution of $Lx = b$ (or the computed inverse) is far more accurate than [what forward stability analysis] would suggest."

An explanation for this gap is suggested by Higham [4] who notes that the backward error analysis given by Wilkinson for the solution of triangular systems yields (small) *componentwise* bounds on the perturbation matrix (see section 6 item (1) below). Higham then uses this fact to deduce small forward error bounds for particular subclasses of triangular systems and to numerically investigate the accuracy of other particular such systems. In doing so, Higham makes use of the mixed condition number introduced by Skeel [5][1]. This condition number has a natural role in analyzing accuracy of triangular systems since bounds for it, together with the backward analysis of Wilkinson mentioned above, yield forward analysis bounds for the computed solution of the system. Furthermore, the restriction to componentwise perturbations—both in the backward error analysis and in the mixed condition number—forces perturbations to preserve the triangular structure of the data matrices.

A further step in explaining the gap, somehow orthogonal to the work of Higham, was given by Viswanath and Trefethen in [8] where a precise meaning to the expression "triangular systems are, in general, ill-conditioned" was given. Indeed, if L_n denotes a random triangular $n \times n$ matrix (whose entries are independent standard Gaussian

*Received by the editors July 7, 2008; accepted for publication (in revised form) by R.-C. Li March 19, 2009; published electronically June 17, 2009.
http://www.siam.org/journals/simax/31-2/72946.html

†United International College, Tang Jia Wan, Zhuhai, Guandong Province, People's Republic of China (dennisc@uic.edu.hk).

‡Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (macucker@cityu.edu.hk). This author's research was partially supported by GRF grant CityU 100808.

[1]Skeel called it "componentwise." In this paper, however, following the notation introduced in [3], we will use this word for the condition numbers measuring both data perturbation and computed errors in a componentwise fashion.

random variables) and $\kappa_n = \|L_n\|\|L_n^{-1}\|$ is its condition number (which is a positive random variable), then it is shown in [8] that

$$\sqrt[n]{\kappa_n} \to 2 \quad \text{almost surely}$$

as $n \to \infty$. A straightforward consequence of this result is that the expected value of $\log \kappa_n$ satisfies $\mathbf{E}(\log \kappa_n) = \Omega(n)$.

The goal of this paper is to close the gap by giving a precise meaning, in the sense of [8], to the other half of the statement above, namely, to the expression "triangular systems are generally solved to high accuracy." More precisely, we consider the mixed condition numbers $\boldsymbol{m}^\dagger(L_n)$ and $\boldsymbol{m}(L_n, b_n)$ for the problems of matrix inversion and linear equation solving, respectively, for a random triangular L_n as above and a random $b_n \in \mathbb{R}^n$. Then we show that

$$\mathbf{E}(\log \boldsymbol{m}^\dagger(L_n)),\ \mathbf{E}(\log \boldsymbol{m}(L_n, b_n)) = \mathcal{O}(\log n). \tag{1}$$

From the bound on $\mathbf{E}(\log \boldsymbol{m}(L_n, b_n))$ it follows that the average loss of precision in the solution of random triangular systems is small. From that on $\mathbf{E}(\log \boldsymbol{m}^\dagger(L_n))$, the one for matrix inversion is small as well. One can therefore replace the summary above by the following:

> *Triangular systems of equations are generally solved to high accuracy because their backward error analysis yields small componentwise perturbations and triangular matrices are, in general, well conditioned for these perturbations.*

The results showing (1), Theorems 2 and 3 below, are proved in the more general context of sparse matrices (which, in this paper, are matrices with a fixed pattern of zeros[2]) and componentwise condition numbers (which ensure high relative accuracy *in each component* of the computed solution A^{-1} or x). Besides triangular matrices, these results apply to other classes of sparse matrices such as, for instance, tridiagonal matrices. In the process of proving them, we found it useful to estimate as well the average mixed condition for the computation of the determinant.

2. Preliminaries. Condition numbers measure the worst-case magnification of a small data perturbation in the computed outcome. As originally introduced by Turing [7], they were *normwise* in the sense that both the data perturbation and the outcome's error are measured using norms (in the space of data and outcomes, respectively). In contrast, *mixed* condition numbers measure data perturbation componentwise, and *componentwise* condition numbers measure both data perturbation and the outcome's error in this way.

To define these condition numbers the following form of "distance" function will be useful. For points $u, v \in \mathbb{R}^p$ we define $\frac{u}{v} = (w_1, \ldots, w_p)$ with

$$w_i = \begin{cases} u_i/v_i & \text{if } v_i \neq 0, \\ 0 & \text{if } u_i = v_i = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Then we define

$$d(u, v) = \left\| \frac{u - v}{v} \right\|_\infty .$$

[2]The word "sparse" is also used to denote matrices with pa large number of zeros, not necessarily in fixed positions.

Note that if $d(u,v) < \infty$,

$$d(u,v) = \min\{\nu \geq 0 \mid |u_i - v_i| \leq \nu |v_i| \text{ for } i = 1, \ldots, p\}.$$

For $\delta > 0$ and $a \in \mathbb{R}^p$ we denote $\mathcal{S}(a,\delta) = \{x \in \mathbb{R}^p \mid d(x,a) = \delta\}$.

DEFINITION 1. *Let $\mathcal{D} \subseteq \mathbb{R}^p$ and $F : \mathcal{D} \to \mathbb{R}^q$ be a continuous mapping. Let $a \in \mathcal{D}$ such that $F(a) \neq 0$.*

(i) *The* mixed condition number *of F at a (with respect to a norm $\| \ \|_q$ on $\mathbb{R}^q$) is defined by*

$$\boldsymbol{m}(F,a) = \lim_{\delta \to 0} \sup_{x \in \mathcal{S}(a,\delta)} \frac{\|F(x) - F(a)\|_q}{\|F(a)\|_q} \frac{1}{d(x,a)}.$$

(ii) *Suppose $F(a) = (f_1(a), \ldots, f_q(a))$ is such that $f_j(a) \neq 0$ for $j = 1, \ldots, q$. Then the* componentwise condition number *of F at a is*

$$\boldsymbol{c}(F,a) = \lim_{\delta \to 0} \sup_{x \in \mathcal{S}(a,\delta)} \frac{d(F(x), F(a))}{d(x,a)}.$$

Remark 1. We can extend the definition of componentwise condition numbers to the case where $f_i(a) = 0$ for some $i \in [n]$. Define, for $i \in [n]$ such that $f_i(a) \neq 0$,

$$\boldsymbol{c}_i(F,a) = \lim_{\delta \to 0} \sup_{x \in \mathcal{S}(a,\delta)} \frac{|f_i(x) - f_i(a)|}{\delta |f_i(a)|},$$

and for $i \in [n]$ with $f_i(a) = 0$ take $\boldsymbol{c}_i(F,a) = 0$ if

$$\lim_{\delta \to 0} \sup_{x \in \mathcal{S}(a,\delta)} \frac{|f_i(x) - f_i(a)|}{\delta} = 0$$

and $\boldsymbol{c}_i(F,a) = \infty$ otherwise. Then we have

$$\boldsymbol{c}(F,a) = \max_{i \in [n]} \boldsymbol{c}_i(F,a).$$

PROPOSITION 1. *For all $a \in \mathcal{D}$ and any monotonic norm in $\mathbb{R}^q$, $\boldsymbol{m}(F,a) \leq \boldsymbol{c}(F,a)$.*

Proof. For all $x \in \mathcal{S}(a,\delta)$ and all $i \leq q$, $|F(x)_i - F(a)_i| \leq d(F(x),F(a))|F(a)_i|$. Since $\| \ \|$ is monotonic (cf. [1]), this implies $\|F(x) - F(a)\| \leq d(F(x),F(a))\|F(a)\|$ and hence the statement. ∎

In what follows, for $n \in \mathbb{N}$, we denote the set $\{1, \ldots, n\}$ by $[n]$ and write, as usual, $[n]^2 = [n] \times [n]$.

DEFINITION 2. *We denote by $\mathscr{M}$ the set of $n \times n$ real matrices and by Σ its subset of singular matrices. Also, for a subset $S \subseteq [n]^2$ we denote*

$$\mathscr{M}_S = \{A \in \mathscr{M} \mid \text{ if } (i,j) \notin S, \text{ then } a_{ij} = 0\}.$$

We denote by $\mathscr{R}_S$ the space of random $n \times n$ matrices obtained by setting $a_{ij} = 0$ if $(i,j) \notin S$ and drawing all other entries independently from the standard Gaussian $N(0,1)$. As above, if $S = [n]^2$, we write simply $\mathscr{R}$.

In the rest of this paper, for nonsingular matrices A and A', we denote their inverses by Γ and Γ', respectively. Also, we denote by $A_{(ij)}$ the submatrix of A obtained by removing from A its ith row and its jth column. Denoting by γ_{ij} the (i,j)th entry of Γ we have, by Cramer's rule, $\gamma_{ij} = \frac{\det(A_{(ij)})}{\det(A)}$.

3. Determinant computation. We consider here the problem of computing the determinant of a matrix A and its componentwise condition number $\boldsymbol{c}_{\mathsf{det}}(A)$. The main result of this section is the following.

THEOREM 1. *For $S \subseteq [n]^2$ and $t \geq 2|S|$ we have*

$$\operatorname*{Prob}_{A \in \mathscr{R}_S} \{\boldsymbol{c}_{\mathsf{det}}(A) \geq t\} \leq |S|^2 \frac{1}{t}.$$

Here $|S|$ denotes the cardinality of S.

Average loss of precision (in a base b) is measured by the expected value of the logarithm (in that base) of the condition number. We may use Theorem 1 to obtain one such result for the computation of the determinant. To avoid problems caused by this condition number being less than 1 we consider the function $\log_+(x)$ defined to be $\log(x)$ if $x \geq 1$ and 0 otherwise.

COROLLARY 1. *For a base $b > 1$ and a set $S \subseteq [n]^2$ with $|S| \geq 2$, we have $\mathbf{E}(\log_+ \boldsymbol{c}_{\mathsf{det}}(A)) \leq 2 \log |S| + \frac{1}{\ln b}$, where $\mathbf{E}$ denotes expectation over $A \in \mathscr{R}_S$.*

Note that the restriction $|S| \geq 2$ is without loss of generality. The case $|S| = 0$ reduces $\mathscr{M}_S$ to the zero matrix, and the case $|S| = 1$ allows only for one nonzero entry, a singular situation whenever $n \geq 2$. The proof of Corollary 1 follows from Theorem 1 and the following result by taking $Z = \boldsymbol{c}_{\mathsf{det}}(A)$ and $t_0 = |S|^2$ and noting that $t_0 \geq 2|S|$ since $|S| \geq 2$.

PROPOSITION 2. *Let $t_0 > 0$ and $Z \geq 1$ be a random variable satisfying that $\mathsf{Prob}\{Z \geq t\} \leq t_0 t^{-1}$ for all $t \geq t_0$. Then $\mathbf{E}(\log Z) \leq \log t_0 + \frac{1}{\ln b}$, where $b > 1$ is the base of the logarithm.*

Proof. We have $\mathsf{Prob}\{\log Z \geq t\} \leq t_0 b^{-t} = b^{-(t - \log t_0)}$ for all $t > \log t_0$. Therefore,

$$\mathbf{E}(\log Z) = \int_0^\infty \mathsf{Prob}\{\log Z \geq s\} ds \leq \log t_0 + \int_{\log t_0}^\infty b^{-(t - \log t_0)} dt = \log t_0 + \frac{1}{\ln b}. \qquad \square$$

Towards the proof of Theorem 1 we first obtain explicit expressions for $\boldsymbol{c}_{\mathsf{det}}(A)$. We begin by noting that taking $F : \mathscr{M} \to \mathbb{R}$ to be $F(A) = \det(A)$ in Definition 1 we obtain, for $A \in \mathscr{M} \setminus \Sigma$,

$$\boldsymbol{c}_{\mathsf{det}}(A) = \lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A, \delta)} \frac{|\det(A') - \det(A)|}{\delta |\det(A)|}.$$

Also, for $A \in \Sigma$, we have $\boldsymbol{c}_{\mathsf{det}}(A) = 0$ if

$$\lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A, \delta)} \frac{|\det(A')|}{\delta} = 0$$

and $\boldsymbol{c}_{\mathsf{det}}(A) = \infty$ otherwise. Note that $\boldsymbol{c}_{\mathsf{det}}(0) = 0$.

LEMMA 1. *For $A \in \mathscr{M} \setminus \Sigma$,*

$$\boldsymbol{c}_{\mathsf{det}}(A) = \sum_{i,j \in [n]} |a_{ij} \gamma_{ji}|.$$

Proof. Let $A \in \mathscr{M}$. For any $i \in [n]$, expanding by the ith row,

$$\det(A) = \sum_{j \in [n]} (-1)^{i+j} a_{ij} \det(A_{(ij)}).$$

Hence, for all $i, j \in [n]$,

$$\frac{\partial \det(A)}{\partial a_{ij}} = (-1)^{i+j} \det(A_{(ij)}).$$

Using Taylor's expansion and these equalities we obtain

$$\det(A') = \det(A) + \sum_{i,j\in[n]} (-1)^{i+j}(a'_{ij} - a_{ij}) \det(A_{(ij)}) + \mathcal{O}\left(\|A' - A\|^2\right).$$

Here the norm in $\|A' - A\|$ is not relevant since all norms in $\mathscr{M}$ are equivalent. By choosing a monotonic norm we have that if $A' \in \mathcal{S}(A, \delta)$, then $\|A' - A\| = \mathcal{O}(\delta)$. It follows that, for $A \notin \Sigma$,

$$\begin{aligned}
\boldsymbol{c}_{\mathsf{det}}(A) &= \lim_{\delta\to 0} \sup_{A'\in\mathcal{S}(A,\delta)} \frac{|\det(A') - \det(A)|}{\delta|\det(A)|} \\
&= \lim_{\delta\to 0} \sup_{A'\in\mathcal{S}(A,\delta)} \sum_{i,j\in[n]} \frac{|(a'_{ij} - a_{ij})\det(A_{(ij)})|}{\delta|\det(A)|} \\
&= \lim_{\delta\to 0} \sup \left(\sum_{i,j\in[n]} \frac{|(a'_{ij} - a_{ij})\det(A_{(ij)})|}{\delta|\det(A)|} : \frac{|a'_{ij} - a_{ij}|}{|a_{ij}|} \le \delta \right).
\end{aligned}$$

The second equality follows from the fact that we can choose A' such that the terms $(a'_{ij} - a_{ij})(-1)^{i+j}\det(A_{(ij)})$ have the same sign for all $i, j \in [n]$. Actually, the supremum above is attained by taking $a'_{ij} = a_{ij}(1 \pm \delta)$ where we take the plus sign if $(-1)^{i+j}\det(A_{(ij)}) \ge 0$ and the minus sign otherwise. Therefore,

$$\boldsymbol{c}_{\mathsf{det}}(A) = \sum_{i,j\in[n]} \left| \frac{a_{ij}\det(A_{ij})}{\det(A)} \right| = \sum_{i,j\in[n]} |a_{ij}\gamma_{ji}|. \qquad \square$$

We denote by $N(0, \mathrm{Id}_n)$ the n-dimensional Gaussian distribution with mean zero and covariance matrix the identity $n \times n$ matrix. A vector x has this distribution when its entries are independently drawn from $N(0, 1)$.

LEMMA 2. *Let p and q be two fixed vectors in $\mathbb{R}^n$ such that $\|p\| \le \|q\|$. If $x \sim N(0, \mathrm{Id}_n)$, then, for all $t \ge 2$,*

$$\mathsf{Prob}\left\{ \left| \frac{x^{\mathsf{T}}p}{x^{\mathsf{T}}q} \right| \ge t \right\} \le \frac{1}{t}.$$

Proof. Let $\nu = \|q\|$. By the orthogonal invariance of $N(0, \mathrm{Id}_n)$ we may assume $q = (\nu, 0, \ldots, 0)$. Also, by appropriately scaling, we may assume that $\nu = 1$. Note that then $\|p\| \le 1$. We therefore have

$$\begin{aligned}
\mathsf{Prob}\left\{ \left| \frac{x^{\mathsf{T}}p}{x^{\mathsf{T}}q} \right| \ge t \right\} &= \mathsf{Prob}\left\{ \left| p_1 + \sum_{i\in\{2,\ldots,n\}} \frac{x_i p_i}{x_1} \right| \ge t \right\} \\
&= \mathsf{Prob}\left\{ \left| p_1 + \frac{1}{x_1}\alpha Z \right| \ge t \right\} \qquad (2) \\
&= \mathsf{Prob}\left\{ \frac{Z}{x_1} \ge \frac{t - p_1}{\alpha} \right\} + \mathsf{Prob}\left\{ \frac{Z}{x_1} \le \frac{-t - p_1}{\alpha} \right\},
\end{aligned}$$

where $Z = N(0, 1)$ is independent of x_1 and $\alpha = \sqrt{p_2^2 + \cdots + p_n^2} \le 1$. Here we used that a sum of independent centered Gaussians is a centered Gaussian whose variance is the sum of the terms' variances. Note that in case $\alpha = 0$ the statement is trivially true.

Since x_1 and Z are independent, both with distribution $N(0,1)$, the angle $\theta = \arctan(Z/x_1)$ is uniformly distributed in $[-\pi/2, \pi/2]$ and we have, for $\gamma \in [0,\infty)$,

$$\begin{aligned}\mathsf{Prob}\left\{\frac{Z}{x_1} \geq \gamma\right\} &= \mathsf{Prob}\{\theta \geq \arctan\gamma\} = \frac{1}{\pi}\left(\frac{\pi}{2} - \arctan\gamma\right)\\ &= \frac{1}{\pi}\int_\gamma^\infty \frac{1}{1+t^2}dt \;\leq\; \frac{1}{\pi}\int_\gamma^\infty \frac{1}{t^2}dt \;=\; \frac{1}{\pi\gamma}.\end{aligned}$$

Similarly, for $\sigma \in (-\infty, 0]$,

$$\begin{aligned}\mathsf{Prob}\left\{\frac{Z}{x_1} \leq \sigma\right\} &= 1 - \mathsf{Prob}\{\theta \geq \arctan\sigma\} = 1 - \frac{1}{\pi}\left(\frac{\pi}{2} - \arctan\sigma\right)\\ &= \frac{1}{\pi}\left(\frac{\pi}{2} - \arctan(-\sigma)\right) \;\leq\; \frac{1}{\pi(-\sigma)}.\end{aligned}$$

Using these bounds in (2) with $\gamma = \frac{t-p_1}{\alpha}$ and $\sigma = \frac{-t-p_1}{\alpha}$ we obtain

$$\mathsf{Prob}\left\{\left|\frac{x^\mathsf{T}p}{x^\mathsf{T}q}\right| \geq t\right\} \leq \frac{1}{\pi}\left(\frac{\alpha}{t-p_1} + \frac{\alpha}{t+p_1}\right) = \frac{\alpha}{\pi}\frac{2t}{t^2-p_1^2} \leq \frac{2}{\pi}\frac{t}{t^2-1} \leq \frac{1}{t}$$

the last inequality since $t > 2$. ∎

LEMMA 3. *Let $S \subseteq [n]^2$ be such that $\mathscr{M}_S \subseteq \Sigma$. Then, for all $A \in \mathscr{M}_S$, $\boldsymbol{c}_{\mathsf{det}}(A) = 0$.*

Proof. Since $\mathscr{M}_S \subseteq \Sigma$ and $A \in \mathscr{M}_S$, we have $\mathcal{S}(A,\delta) \subseteq \Sigma$ for all $\delta > 0$. The result now follows. ∎

LEMMA 4. *Let $S \subset [n]^2$ such that $\mathscr{M}_S \not\subseteq \Sigma$. Then*

$$\operatorname*{Prob}_{A\in\mathscr{R}_S}(A \text{ is singular}) = 0.$$

Proof. The set of singular matrices in $\mathscr{M}_S$ is the zero set of the restriction of the determinant to $\mathscr{M}_S$. This restriction is a polynomial in $\mathbb{R}^{|S|}$ whose zero set, if different from $\mathbb{R}^{|S|}$, has dimension smaller than $|S|$. ∎

Proof of Theorem 1. Case (i): $\mathscr{M}_S \subseteq \Sigma$. In this case, the desired inequality is trivial by Lemma 3.

Case (ii): $\mathscr{M}_S \not\subseteq \Sigma$. By Lemma 4, with probability 1, A is nonsingular. So, by Lemma 1,

$$\begin{aligned}\mathsf{Prob}\{\boldsymbol{c}_{\mathsf{det}}(A) \geq t\} &= \mathsf{Prob}\left\{\sum_{i,j\in[n]} |a_{ij}\gamma_{ji}| \geq t\right\}\\ &= \mathsf{Prob}\left\{\sum_{(i,j)\in S}\left|\frac{a_{ij}\det(A_{(ij)})}{\det(A)}\right| \geq t\right\}. \qquad (3)\end{aligned}$$

Assume $(1,1) \in S$, and let $x = a_1$ be the first column of A. Also, let $I = \{i \in [n] \mid (i,1) \in S\}$ and x_I be the vector obtained by removing entries x_i with $i \notin I$. Then

$$x_I \sim N(0, \mathrm{Id}_{|I|}). \qquad (4)$$

For $i \in [n]$ write $q_i = (-1)^{i+1}\det(A_{(i1)})$. Let $q = (q_1,\dots,q_n)^\mathsf{T}$ and q_I be the vector obtained by removing entries q_i with $i \notin I$. Clearly, q_I is independent of x_I. Using

this notation, the expansion by the first column yields

$$\det(A) = \sum_{i\in[n]} (-1)^{i+1} a_{i1} \det(A_{(i1)}) = x_I^{\mathsf{T}} q_I.$$

In addition, $a_{11}\det(A_{(11)}) = x_I^{\mathsf{T}}(q_1 e_1)$, where e_1 is the vector with the first entry equal to 1 and all others equal to 0. Hence,

$$\frac{a_{11}\det(A_{(11)})}{\det(A)} = \frac{x_I^{\mathsf{T}}(q_1 e_1)}{x_I^{\mathsf{T}} q_I}.$$

Let φ be the density of $q_1 = \det(A_{(11)})$ and ψ be that of the random vector q_I. Then, for $z \geq 2$,

$$\begin{aligned}
&\mathsf{Prob}\left\{\left|\frac{a_{11}\det(A_{(11)})}{\det(A)}\right| \geq z\right\} = \mathsf{Prob}\left\{\left|\frac{x_I^{\mathsf{T}}(q_1 e_1)}{x_I^{\mathsf{T}} q_I}\right| \geq z\right\} \\
&= \int_{u\in\mathbb{R}} \int_{v\in\mathbb{R}^{|I|}} \mathsf{Prob}\left\{\left|\frac{x_I^{\mathsf{T}}(ue_1)}{x_I^{\mathsf{T}} v}\right| \geq z \middle| q_1 = u, q_I = v\right\} \varphi(u)\psi(v) \\
&\leq \int_{u\in\mathbb{R}} \int_{v\in\mathbb{R}^{|I|}} \frac{1}{z}\,\varphi(u)\psi(v) \;=\; \frac{1}{z}.
\end{aligned}$$

Here the inequality follows since x_I is independent of q_1 and q_I, and therefore we can use (4) and Lemma 2 (with $p = (q_1 e_1)$ and $q = q_I$).

The same bound can be proven for all $(i,j) \in S$. Using these bounds with $z = \frac{t}{|S|}$ and (3) we obtain

$$\mathsf{Prob}\{\boldsymbol{c}_{\mathsf{det}}(A) \geq t\} \leq \sum_{(i,j)\in S} \mathsf{Prob}\left\{\left|\frac{a_{ij}\det(A_{(ij)})}{\det(A)}\right| \geq \frac{t}{|S|}\right\} \leq \frac{|S|^2}{t}. \qquad \square$$

4. Matrix inversion. We now focus on the problem of inverting a matrix A and its componentwise condition number $\boldsymbol{c}^{\dagger}(A)$. Our main results in this section are the following.

THEOREM 2. *Let $S \subset [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then, for all $t \geq 4|S|$,*

$$\operatorname*{\mathsf{Prob}}_{A\in\mathscr{R}_S}\{\boldsymbol{c}^{\dagger}(A) \geq t\} \leq 4|S|^2 n^2 \frac{1}{t}.$$

Using Proposition 2 we prove the following corollary.

COROLLARY 2. *Let $S \subset [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then*

$$\mathbf{E}(\log_+(\boldsymbol{c}^{\dagger}(A))) \leq 2\log n + 2\log|S| + \log 4 + \frac{1}{\ln b},$$

where $\mathbf{E}$ *denotes expectation over* $A \in \mathscr{R}_S$.

Remark 2. Note that, for all monotonic norm on $\mathscr{M}_S$, the bound above also holds for $\boldsymbol{m}(A)$ by Proposition 1. This is in contrast with the lower bound linear in n for the expected value of the logarithm of normwise condition numbers which follows from [8].

Definition 1 (and Remark 1) yields expressions for the (mixed and componentwise) condition numbers of a matrix A by taking $\mathcal{D} = \mathscr{M} \setminus \Sigma$ and $F : \mathscr{M} \setminus \Sigma \to \mathscr{M}$ given by $F(A) = A^{-1}$. For $k, \ell \in [n]$ such that $\gamma_{k\ell} \neq 0$, we let

$$\boldsymbol{c}^{\dagger}_{k\ell}(A) = \lim_{\delta\to 0} \sup_{A'\in\mathcal{S}(A,\delta)} \frac{|\gamma'_{k\ell} - \gamma_{k\ell}|}{\delta|\gamma_{k\ell}|},$$

and for $k, \ell \in [n]$ such that $\gamma_{k\ell} = 0$, we let $\boldsymbol{c}^\dagger_{k\ell}(A) = 0$ if

$$\lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A,\delta)} \frac{|\gamma'_{k\ell} - \gamma_{k\ell}|}{\delta} = 0$$

and $\boldsymbol{c}^\dagger_{k\ell}(A) = \infty$ otherwise. Then

$$\boldsymbol{c}^\dagger(A) = \max_{k,\ell \in [n]} \boldsymbol{c}^\dagger_{k\ell}(A).$$

Similarly, for a norm $\| \ \|$ on $\mathscr{M}$,

$$\boldsymbol{m}(A) = \lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A,\delta)} \frac{\|\Gamma' - \Gamma\|}{\delta \|\Gamma\|}.$$

LEMMA 5. *For $A \in \mathscr{M} \setminus \Sigma$ and $k, \ell \in [n]$,*

$$\boldsymbol{c}^\dagger_{k\ell}(A) \leq \boldsymbol{c}_{\mathsf{det}}(A) + \boldsymbol{c}_{\mathsf{det}}(A_{(\ell k)}).$$

Proof. We divide the proof into cases. Case (i): $\gamma_{k\ell} \neq 0$.

Let $\delta > 0$ be sufficiently small so that $\mathcal{S}(A,\delta) \cap \Sigma = \emptyset$ and, for all $A' \in \mathcal{S}(A,\delta)$, $\left|\frac{\det(A') - \det(A)}{\det(A)}\right| < 1$. Let $A' \in \mathcal{S}(A,\delta)$.

Since $\gamma_{k\ell} = \frac{\det(A_{(\ell k)})}{\det(A)}$,

$$\begin{aligned}
\frac{\gamma'_{k\ell} - \gamma_{k\ell}}{\gamma_{k\ell}} &= \frac{\det(A)}{\det(A_{(\ell k)})} \left(\frac{\det(A'_{(\ell k)})}{\det(A')} - \frac{\det(A_{(\ell k)})}{\det(A)} \right) \\
&= \frac{\det(A)}{\det(A_{(\ell k)})} \frac{\det(A'_{(\ell k)})}{\det(A')} - 1 \\
&= \frac{1 + \frac{\det(A'_{(\ell k)}) - \det(A_{(\ell k)})}{\det(A_{(\ell k)})}}{1 + \frac{\det(A') - \det(A)}{\det(A)}} - 1 \\
&= \frac{\frac{\det(A'_{(\ell k)}) - \det(A_{(\ell k)})}{\det(A_{(\ell k)})} - \frac{\det(A') - \det(A)}{\det(A)}}{1 + \frac{\det(A') - \det(A)}{\det(A)}}.
\end{aligned}$$

Using that $\left|\frac{\det(A') - \det(A)}{\det(A)}\right| < 1$,

$$\left|\frac{\gamma'_{k\ell} - \gamma_{k\ell}}{\gamma_{k\ell}}\right| \leq \frac{\left|\frac{\det(A'_{(lk)}) - \det(A_{(\ell k)})}{\det(A_{(\ell k)})}\right| + \left|\frac{\det(A') - \det(A)}{\det(A)}\right|}{1 - \left|\frac{\det(A') - \det(A)}{\det(A)}\right|}$$

and therefore

$$\begin{aligned}
&\sup_{A' \in \mathcal{S}(A,\delta)} \left|\frac{\gamma'_{k\ell} - \gamma_{k\ell}}{\delta \gamma_{k\ell}}\right| \\
&\qquad \leq \frac{\sup_{A' \in \mathcal{S}(A,\delta)} \left|\frac{\det(A'_{(\ell k)}) - \det(A_{(\ell k)})}{\delta \det(A_{(\ell k)})}\right| + \sup_{A' \in \mathcal{S}(A,\delta)} \left|\frac{\det(A') - \det(A)}{\delta \det(A)}\right|}{1 - \sup_{A' \in \mathcal{S}(A,\delta)} \left|\frac{\det(A') - \det(A)}{\det(A)}\right|}.
\end{aligned}$$

Taking limits for $\delta \to 0$ on both sides we get

$$\boldsymbol{c}^{\dagger}_{k\ell}(A) \leq \boldsymbol{c}_{\mathsf{det}}(A) + \boldsymbol{c}_{\mathsf{det}}(A_{(\ell k)}).$$

Case (ii): $\gamma_{k\ell} = 0$ and

$$\lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A,\delta)} \frac{|\gamma'_{k\ell}|}{\delta} = 0.$$

In this case, $\boldsymbol{c}^{\dagger}_{k\ell}(A) = 0$ and the statement holds.

Case (iii): $\gamma_{k\ell} = 0$ and

$$0 \neq \lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A,\delta)} \frac{|\gamma'_{k\ell}|}{\delta} = \lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A,\delta)} \frac{|\det(A'_{\ell k})|}{\delta |\det(A')|}.$$

In this case,

$$\lim_{\delta \to 0} \sup_{A' \in \mathcal{S}(A,\delta)} \frac{|\det(A'_{\ell k})|}{\delta} \neq 0$$

and therefore $\boldsymbol{c}_{\mathsf{det}}(A_{\ell k}) = \infty$. The statement holds as well. ∎

Proof of Theorem 2. By definition of $\boldsymbol{c}^{\dagger}(A)$,

$$\mathsf{Prob}\{\boldsymbol{c}^{\dagger}(A) \geq t\} = \mathsf{Prob}\left\{\max_{k,\ell \in [n]} \boldsymbol{c}^{\dagger}_{k\ell}(A) \geq t\right\} \leq \sum_{k,\ell \in [n]} \mathsf{Prob}\{\boldsymbol{c}^{\dagger}_{k\ell}(A) \geq t\}.$$

By Lemma 4, with probability 1, A is nonsingular. So, since $\frac{t}{2} \geq 2|S|$ by hypothesis, we can apply Lemma 5 to obtain

$$\begin{aligned}\mathsf{Prob}\{\boldsymbol{c}^{\dagger}_{k\ell}(A) \geq t\} &\leq \mathsf{Prob}\left\{\boldsymbol{c}_{\mathsf{det}}(A) \geq \frac{t}{2}\right\} + \mathsf{Prob}\left\{\boldsymbol{c}_{\mathsf{det}}(A_{(k\ell)}) \geq \frac{t}{2}\right\} \\ &\leq 4|S|^2 \frac{1}{t}\end{aligned}$$

the last inequality by Theorem 1. The statement now follows. ∎

5. Linear equations solving. We finally deal with the problem of solving linear systems of equations.

THEOREM 3. *Let $S \subset [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then, for all $t \geq 4(|S| + n)$,*

$$\mathsf{Prob}\{\boldsymbol{c}(A,b) \geq t\} \leq 10|S|^2 n \frac{1}{t},$$

where Prob *denotes probability over $(A,b) \in \mathscr{R}_S \times N(0, \mathrm{Id}_n)$.*

We may use Proposition 2 a last time.

COROLLARY 3. *Let $S \subset [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then*

$$\mathbf{E}(\log_+(\boldsymbol{c}(A,b))) \leq \log n + 2\log|S| + \log 10 + \frac{1}{\ln b}.$$

Definition 1 (and Remark 1) yields again expressions for the (mixed and componentwise) condition numbers of a pair (A,b) by taking $\mathcal{D} = (\mathscr{M} \setminus \Sigma) \times \mathbb{R}^n$ and $F : (\mathscr{M} \setminus \Sigma) \times \mathbb{R}^n \to \mathbb{R}^n$ given by $F(A,b) = A^{-1}b$.

For $A \in \mathscr{M} \setminus \Sigma$ and $b \in \mathbb{R}^n$ we denote $x = A^{-1}b$. For $k \in [n]$ such that $x_k \neq 0$ we let

$$\boldsymbol{c}_k(A,b) = \lim_{\delta \to 0} \sup_{(A',b') \in \mathcal{S}((A,b),\delta)} \frac{|x'_k - x_k|}{\delta |x_k|}.$$

For $k \in [n]$ such that $x_k = 0$ we let $\boldsymbol{c}_k(A,b) = 0$ if

$$\lim_{\delta \to 0} \sup_{(A',b') \in \mathcal{S}((A,b),\delta)} \frac{|x'_k - x_k|}{\delta} = 0$$

and $\boldsymbol{c}_k(A,b) = \infty$ otherwise. Then

$$\boldsymbol{c}(A,b) = \max_{k \in [n]} \boldsymbol{c}_k(A,b).$$

Similarly, for a norm $\| \ \|$ in $\mathbb{R}^n$,

$$\boldsymbol{m}(A,b) = \lim_{\delta \to 0} \sup_{(A',b') \in \mathcal{S}((A,b),\delta)} \frac{\|x' - x\|}{\delta \|x\|}.$$

In what follows let R_k be the matrix obtained by replacing the kth column of A by b.

LEMMA 6. *For any nonsingular matrix A and $k \in [n]$,*

$$\boldsymbol{c}_k(A,b) \leq \boldsymbol{c}_{\mathsf{det}}(A) + \boldsymbol{c}_{\mathsf{det}}(R_k).$$

Proof. By Crammer's rule,

$$x_k = \frac{\det(R_k)}{\det(A)}.$$

The rest of this proof is similar to the proof of Lemma 5. ∎

Proof of Theorem 3. The proof follows the lines of that of Theorem 2. First, we get

$$\mathsf{Prob}\{\boldsymbol{c}(A,b) \geq t\} \leq \sum_{k \in [n]} \mathsf{Prob}\{\boldsymbol{c}_k(A,b) \geq t\}.$$

Then we apply Lemma 6 (using that, with probability 1, $A \notin \Sigma$ and that $\frac{t}{2} \geq 2|S|$) to get

$$\begin{aligned} \mathsf{Prob}\{\boldsymbol{c}_k(A,b) \geq t\} &\leq \mathsf{Prob}\left\{\boldsymbol{c}_{\mathsf{det}}(A) \geq \frac{t}{2}\right\} + \mathsf{Prob}\left\{\boldsymbol{c}_{\mathsf{det}}(R_k) \geq \frac{t}{2}\right\} \\ &\leq 2|S|^2 \frac{1}{t} + 2(|S|+n)^2 \frac{1}{t} \leq 10|S|^2 \frac{1}{t}. \end{aligned}$$

For the second inequality we used the fact that $|S| \geq n$. The statement now follows. ∎

6. Additional remarks. (1) To obtain bounds for the average loss of precision of random triangular systems one may combine Theorem 3 with the following result by Wilkinson [9, Chapter 3, section 19] which we quote as given in [4].

THEOREM 4. *Let $T \in \mathbb{R}^{n\times n}$ be a nonsingular triangular matrix, and assume $nu < 0.1$ (here u is the round-off unit). Then the computed solution $\widehat{x}$ to the system $Tx = b$ satisfies*

$$(T + E)\widehat{x} = b,$$

where, for some universal constant c,

$$|e_{ij}| \leq (|i - j| + 2)cu|t_{ij}|.$$

The use of $\boldsymbol{c}(A)$ actually yields average loss of precision with the latter measured componentwise in the computed solution.

(2) The bound in Corollary 2 appears to be worse than what computer simulations suggest $\mathbf{E}(\log \boldsymbol{c}^\dagger(L_n))$ should be. In [2] matrices L_n were generated for various values of n and an experimental mean of $\mathbf{E}(\log \boldsymbol{c}^\dagger(L_n))$ was obtained from these values. A linear regression for these means shows a best fit of $3.065 \log n - 1.1466$. A probable source of (a good part of) the difference between this value and the estimate $6 \log n + \mathcal{O}(1)$ following from Corollary 2 is the broad bound $\mathsf{Prob}\{\max_{k,\ell\in[n]} \boldsymbol{c}^\dagger_{k\ell}(A) \geq t\} \leq \sum_{k,\ell\in[n]} \mathsf{Prob}\{\boldsymbol{c}^\dagger_{k\ell}(A) \geq t\}$ in the proof of Theorem 2. In addition to this, the bound $\mathbf{E}(\boldsymbol{m}^\dagger(L_n)) \leq \mathbf{E}(\boldsymbol{c}^\dagger(L_n))$ following from Proposition 1 may be coarse as well. Numerical experiments in [2] suggest a best fit of $\mathbf{E}(\log \boldsymbol{m}^\dagger(L_n)) \approx 1.5334 \log n - 0.5723$.

(3) Section 6 of [8] discusses stability of Gaussian elimination. Having shown that, almost surely, $\kappa(L_n) \approx 2^n$ the authors reflect on how this behavior can be reconciled with the fact that "Gaussian elimination is overwhelmingly stable." They point out that "the reason appears to be statistical: the matrices A for which $\|L^{-1}\|$ is large occupy an exponentially small proportion of the space of matrices" A, a claim for which experimental evidence is given in [6]. It would thus follow that "the matrices L produced by Gaussian elimination are far from random."

Our results show that, in addition, Gaussian elimination (for linear equation solving) needs much less than producing matrices L in the vanishingly small set of triangular matrices with $\kappa(L)$ small. It is enough to produce matrices L *outside* the vanishingly small set of matrices with $\boldsymbol{c}(L, b)$ large.

Aknowledgments. We are grateful to Ernesto Mordecki for helpful discussions.

REFERENCES

[1] F.L. BAUER, J. STOER, AND C. WITZGALL, *Absolute and monotonic norms*, Numer. Math., 3 (1961), pp. 257–264.
[2] F. CUCKER AND H. DIAO, *Mixed and componentwise condition numbers for rectangular structured matrices*, CALCOLO, 44 (2007), pp. 89–115.
[3] I. GOHBERG AND I. KOLTRACHT, *Mixed, componentwise, and structured condition numbers*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 688–704.
[4] N. HIGHAM, *The accuracy of solutions to triangular systems*, SIAM J. Numer. Anal., 26 (1989), pp. 1252–1265.
[5] R.D. SKEEL, *Scaling for numerical stability in Gaussian elimination*, J. ACM, 26 (1979), pp. 494–526.
[6] L.N. TREFETHEN AND R.S. SCHREIBER, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.
[7] A.M. TURING, *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math., 1 (1948), pp. 287–308.
[8] D. VISWANATH AND L.N. TREFETHEN, *Condition numbers of random triangular matrices*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 564–581.
[9] J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice–Hall, Englewood Cliffs, NJ, 1963.