4

Error Analysis of Triangular Linear System

The vast majority of the occurrences of condition numbers in the study of linear systems of equations involve the normwise condition number $\kappa(A)$. Almost invariably, the use of $\kappa(A)$ is enough to provide a satisfying explanation of the phenomena observed in practice.

The case of triangular systems of linear equations provides, in contrast, an example where $\kappa(A)$ turns out to be inadequate. Practitioners observed since long that triangular systems of equations are generally solved to high accuracy in spite of being, in general, ill-conditioned. Thus, for instance, J.H. Wilkinson in [22, 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."

A first goal in this chapter is to give a precise meaning to the feeling that triangular matrices are, in general, ill-conditioned. We prove that, if $L \in \mathbb{R}^{n\times n}$ is a lower triangular matrix whose entries are independent standard Gaussian random variables (i.e., L is drawn from $N(0, \mathrm{I}_{\frac{n(n+1)}{2}})$) then $\mathbb{E}(\log_\beta \kappa(L)) = \Omega(n)$. Corollary 2.4 then yields an expected loss of precision satisfying

$$\mathbb{E}(\mathsf{LoP}(L^{-1}b)) = \mathcal{O}(n).$$

Were the loss of precision in the solution of triangular systems conform to this bound we would not be able to accurately find these solutions. The reason why we actually do find them can be briefly stated. The error analysis of triangular systems reveals that we may use a componentwise condition number $\mathsf{Cw}(L, b)$ instead of the normwise. The second goal of this chapter is to prove that, when L is drawn from $N(0, \mathrm{I}_{\frac{n(n+1)}{2}})$ and $b \in \mathbb{R}^n$ is drawn from $N(0, \mathrm{I}_n)$ then we have $\mathbb{E}(\log \mathsf{Cw}(L, b)) = \mathcal{O}(\log n)$. This bound, together with some backward error analysis, yields bounds for $\mathbb{E}(\mathsf{LoP}(L^{-1}b))$ which are much smaller than the one above, as well as closer to the loss of precision observed in practice.

4.1 Random Triangular Matrices are Ill-conditioned

The main result of this section states that random lower-triangular matrices are ill-conditioned with respect to the normwise (classical) condition number.

Theorem 4.1. *Let $L = (\ell_{ij}) \in \mathbb{R}^{n\times n}$ be a random lower-triangular matrix with independent standard Gaussian random entries ℓ_{ij} for $i \geq j$. Then we have*

$$\mathbb{E}(\ln \kappa(L)) \geq (\ln 2)n - \ln n - 1.$$

As a warm up, we first show a related result —with very simple proof— that already indicates that on average, $\kappa(L)$ grows exponentially in n. For this we focus on unit lower-triangular matrices L, that is, we additionally assume that $\ell_{ii} = 1$.

Proposition 4.2. *Let $L = (\ell_{ij})$ denote a random unit lower-triangular matrix with $\ell_{ii} = 1$ and with independent standard Gaussian random entries ℓ_{ij} for $i > j$. Then we have*

$$\mathbb{E}(\|L^{-1}\|_F^2) = 2^n - 1.$$

In particular, $\mathbb{E}(\|L\|_F^2 \|L^{-1}\|_F^2) \geq n(2^n - 1)$, hence $\mathbb{E}(\kappa(L)^2)$ grows exponentially in n.

Proof. The first column $(s_1, \ldots, s_n)$ of L^{-1} is characterized by $s_1 = 1$ and the recursive relation

$$s_i = -\sum_{j=1}^{i-1} \ell_{ij} s_j \quad \text{for } i = 2, \ldots, n.$$

This implies that s_i is a function of the first i rows of L. Hence the random variable s_i is independent of the entries of L in the rows with index larger than i. By squaring we obtain for $i \geq 2$

$$s_i^2 = \sum_{\substack{j\neq k \\ j,k<i}} \ell_{ij}\ell_{ik} s_j s_k + \sum_{j<i} \ell_{ij}^2 s_j^2.$$

By the preceding observation, $s_j s_k$ is independent of $\ell_{ij}\ell_{ik}$ for $j, k < i$. If additionally $j \neq k$, we get

$$\mathbb{E}(\ell_{ij}\ell_{ik} s_j s_k) = \mathbb{E}(\ell_{ij}\ell_{ik})\,\mathbb{E}(s_j s_k) = \mathbb{E}(\ell_{ij})\,\mathbb{E}(\ell_{ik})\,\mathbb{E}(s_j s_k) = 0$$

as ℓ_{ij} and ℓ_{ik} are independent and centered. So the expectations of the mixed terms vanish and we obtain, using $\mathbb{E}(\ell_{ij}^2) = 1$ (by Proposition 3.10), that

$$\mathbb{E}(s_i^2) = \sum_{j=1}^{i-1} \mathbb{E}(s_j^2) \quad \text{for } i \geq 2.$$

Solving this recursion with $\mathbb{E}(s_1^2) = 1$ yields

$$\mathbb{E}(s_i^2) = 2^{i-2} \quad \text{for } i \geq 2.$$

Therefore, the first column v_1 of L^{-1} satisfies

$$\mathbb{E}(\|v_1\|^2) = \mathbb{E}\Big(\sum_{i=1}^{n} s_i^2\Big) = 2^{n-1}.$$

By an analogous argument one shows that

$$\mathbb{E}(\|v_k\|^2) = 2^{n-k}$$

for the kth column v_k of L^{-1}. Altogether, we obtain

$$\mathbb{E}(\|L^{-1}\|_F^2) = \mathbb{E}\Big(\sum_{k=1}^{n} \|v_k\|^2\Big) = \sum_{k=1}^{n} \mathbb{E}(\|v_k\|^2) = 2^n - 1.$$

Finally, we note that $\|L\|_F^2 \geq n$ since $\ell_{ii} = 1$. Hence,

$$\mathbb{E}(\|L\|_F^2 \|L^{-1}\|_F^2) \geq n\, \mathbb{E}(\|L^{-1}\|_F^2) \geq n(2^n - 1).$$

The last assertion follows from $\kappa(L) \geq \frac{1}{n}\|L\|_F \|L^{-1}\|_F$. □

We turn now to the general situation. Consider a lower-triangular matrix $L = (\ell_{ij}) \in \mathbb{R}^{n\times n}$ that is invertible, i.e., $\det L = \ell_{11} \cdots \ell_{nn} \neq 0$. We denote by $t_1, \ldots, t_n$ the entries of the first column of L^{-1}. These entries can be recursively computed as follows

$$\begin{aligned}
t_1 &= \ell_{11}^{-1} \\
t_2 &= \ell_{22}^{-1}\, \ell_{21} t_1 \\
t_3 &= \ell_{33}^{-1}\, (\ell_{31} t_1 + \ell_{32} t_2) \\
&\vdots \\
t_n &= \ell_{nn}^{-1}\, (\ell_{n1} t_1 + \cdots + \ell_{n,n-1} t_{n-1})
\end{aligned}$$

We suppose that the ℓ_{ij} are independent standard Gaussian random variables. The next lemma provides a recursive formula for the joint probability density function f_k of $(t_1, \ldots, t_k)$. We introduce the notation $T_k := \sqrt{t_1^2 + \cdots t_k^2}$.

Lemma 4.3. *The joint probability density function $f_k(t_1, \ldots, t_k)$ satisfies the following recurrence*

$$f_1 = \frac{1}{\sqrt{2\pi} t_1^2}\, e^{-\frac{1}{2t_1^2}}, \quad f_k = \frac{1}{\pi}\, \frac{T_{k-1}}{T_k^2}\, f_{k-1} \quad \textit{for } k > 1.$$

Proof. We have $t_1 = 1/x$ where $x = \ell_{11}$ is standard Gaussian with density $\varphi(x) = (2\pi)^{-1/2} e^{-\frac{1}{2}x^2}$. Therefore, by Proposition 4.5 (with $n = 1$, $\psi(x) = 1/x$, and $\rho_X = \varphi$), the density ρ of the random variable t_1 satisfies

$$\rho(t_1) \;=\; \left|\frac{dt_1}{dx}\right|^{-1} \varphi(x) \;=\; x^2\varphi(x) \;=\; \frac{1}{\sqrt{2\pi}t_1^2}\, e^{-\frac{1}{2t_1^2}}$$

as claimed.

To obtain the recursive expression for f_k, we consider the random variable

$$\tau_k := \ell_{k1}t_1 + \cdots + \ell_{k,k-1}t_k.$$

For fixed values of $t_1, \ldots, t_{k-1}$, the conditional distribution of τ_k is Gaussian with mean 0 and variance T_{k-1}^2. Therefore, the joint probability density of $(t_1, \ldots, t_{k-1}, \tau_k)$ is given by

$$f_{k-1} \cdot \frac{1}{\sqrt{2\pi}\, T_{k-1}}\, e^{-\frac{\tau_k^2}{2T_{k-1}^2}}$$

The variable t_k is obtained as $t_k = \tau_k/\ell$ where $\ell = \ell_{kk}$ is an independent standard Gaussian random variable. Note that the joint probability density of $(t_1, \ldots, t_{k-1}, \tau_k, \ell)$ is given by

$$f_{k-1} \cdot \frac{1}{\sqrt{2\pi}\, T_{k-1}}\, e^{-\frac{\tau_k^2}{2T_{k-1}^2}}\; \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\ell^2}{2}}.$$

We make now the change of variables $(t_1, \ldots, t_{k-1}, \tau_k, \ell) \overset{\Psi}{\mapsto} (t_1, \ldots, t_{k-1}, t_k, \ell)$ which satisfies $\det D\Psi(t_1, \ldots, t_{k-1}, t_k, \ell) = \ell^{-1}$. Proposition 4.5 implies that the density g of $(t_1, \ldots, t_{k-1}, t_k, \ell)$ equals

$$g = f_{k-1} \cdot \frac{1}{\sqrt{2\pi}\, T_{k-1}}\, e^{-\frac{\ell^2 t_k^2}{2T_{k-1}^2}}\; \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\ell^2}{2}} \cdot |\ell|.$$

A straightforward calculation, making the change of variables $b = \ell^2/2$, shows that

$$\begin{aligned} f_k(t_1, \ldots, t_k) \;&=\; \int_{-\infty}^{\infty} g(t_1, \ldots, t_k, \ell)\, d\ell \;=\; \frac{f_{k-1}}{2\pi\, T_{k-1}}\, 2\int_0^{\infty} e^{-\frac{\ell^2}{2}\left(\frac{t_k^2}{T_{k-1}^2}+1\right)}\, \ell\, d\ell \\ &=\; \frac{f_{k-1}}{\pi T_{k-1}}\, \frac{1}{\frac{t_k^2}{T_{k-1}^2}+1} \;=\; \frac{f_{k-1}}{\pi T_{k-1}}\, \frac{T_{k-1}^2}{T_k^2} \;=\; \frac{f_{k-1}}{\pi}\, \frac{T_{k-1}}{T_k^2}, \end{aligned}$$

which proves the desired recursion. □

The recursive description of the joint probability density functions f_k in Lemma 4.3 yields the following recursion for $\mathbb{E}(\ln T_k^2)$.

Lemma 4.4. *We have* $\mathbb{E}(\ln T_k^2) = \mathbb{E}(\ln T_{k-1}^2) + 2\ln 2$ *for* $k > 1$.

Proof. By Lemma 4.3 we have, omitting the arguments t_i to avoid cluttering the notation,

$$\mathbb{E}(\ln T_k^2) = \int_{\mathbb{R}^k} f_k \ln T_k^2 \, dt_1 \cdots dt_k = \int_{\mathbb{R}^{k-1}} \frac{f_{k-1} T_{k-1}}{\pi} \int_{\mathbb{R}} \frac{\ln T_k^2}{T_k^2} \, dt_k \, dt_1 \cdots dt_{k-1}.$$

We fix $t_1, \ldots, t_{k-1}$ and rewrite the inner integral by making the change of variable $y = t_k / T_{k-1}$. Hence $T_k^2 = T_{k-1}^2(1+y^2)$ and we get

$$\frac{1}{\pi} \int_{\mathbb{R}} \frac{\ln T_k^2}{T_k^2} \, dt_k = \frac{1}{T_{k-1}} \frac{1}{\pi} \int_{\mathbb{R}} \frac{\ln T_{k-1}^2 + \ln(1+y^2)}{1+y^2} \, dy.$$

The function $y \mapsto 1/(\pi(1+y^2))$ is a probability density on $\mathbb{R}$ and a straightforward calculation shows that

$$\frac{1}{\pi} \int_{\mathbb{R}} \frac{\ln(1+y^2)}{1+y^2} \, dy = 2\ln 2.$$

Hence we obtain for the inner integral

$$\frac{1}{\pi} \int_{\mathbb{R}} \frac{\ln T_k^2}{T_k^2} \, dt_k = \frac{1}{T_{k-1}} (\ln T_{k-1}^2 + 2\ln 2).$$

Plugging in this expression above we obtain the stated recursion

$$\mathbb{E}(\ln T_k^2) = \mathbb{E}(\ln T_{k-1}^2) + 2\ln 2. \qquad \square$$

Proof of Theorem 4.1. Using the expression for the density function f_1 provided by Lemma 4.3 we obtain, using software for symbolic integration,

$$\mathbb{E}(\ln T_1^2) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{1}{t_1^2} \, e^{-\frac{1}{2t_1^2}} \ln t_1^2 \, dt_1 = \ln 2 + \gamma$$

where $\gamma \approx 0.577$ denotes the Euler-Mascheroni constant. Combining this with the recursive expression of Lemma 4.4 we get

$$\mathbb{E}(\ln T_n^2) = (2\ln 2)(n-1) + \ln 2 + \gamma \geq (2\ln 2)n - 0.12.$$

Recalling that T_n equals the Euclidean norm of the first column of L^{-1}, this implies

$$\mathbb{E}(\ln \|L^{-1}\|_F) \geq \mathbb{E}(\ln T_n) \geq (\ln 2)n - 0.06.$$

It is known that $\mathbb{E}(\ln \chi_m^2) \geq 0$ for a chi-square distributed random variable χ_m^2 with m degrees of freedom if $m > 1$. Since $\|L\|_F^2$ is chi-square distributed with $n(n+1)/2$ degrees of freedom, this implies $\mathbb{E}(\ln \|L\|_F) \geq 0$ if $n > 1$. Therefore

$$\mathbb{E}\big(\ln(\|L\|_F \|L^{-1}\|_F)\big) \geq \mathbb{E}(\ln T_n) \geq (\ln 2)n - 0.06.$$

Using that $\|L\|\|L^{-1}\| \geq \frac{1}{n}\|L\|_F \|L^{-1}\|_F$, the assertion follows. $\square$

Proposition 4.5. *Let $\psi\colon U \to V$ be a diffeomorphism of open subsets U, V of $\mathbb{R}^n$. Suppose that X is a random vector taking values in U and having the density ρ_X. Then the random variable $Y := \psi(X)$ has the density*

$$\rho_Y(y) = \rho_X(x) \cdot |\det D\psi(x)|^{-1}$$

where $x = \psi^{-1}(y)$.

Proof. `This is an easy consequence of the transformation theorem of integrals.` □

4.2 Backward Analysis of Triangular Linear Systems

Let $L = (\ell_{ij}) \in \mathbb{R}^{n\times n}$ be a non-singular lower triangular matrix and $b \in \mathbb{R}^n$. We are interested in solving the system $Lx = b$. Algorithmically, this is very simple, the components $x_1, \ldots, x_n$ of the solution x are sequentially obtained by back substitution as follows:

algorithm BS
input (L, b)
$x_1 := b_1/\ell_{11}$
for $i = 2 \ldots n$ do
  compute $w := \ell_{i1}x_1 + \cdots + \ell_{i,i-1}x_{i-1}$
  compute $x_i := \frac{b_i - w}{\ell_{ii}}$

It is straightforward to obtain a backward error analysis from the results we proved in Chapter 1. We use notations introduced in Section 1.3.

Proposition 4.6. *There is a round-off implementation of algorithm* BS *which, with input $L \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, computes the solution x of $Lx = b$. If $\epsilon_{\mathsf{mach}}(\lceil \log_2 n \rceil + 1) < 1$ then the computed value $\mathsf{fl}(x)$ satisfies $(L + E)\mathsf{fl}(x) = b$ with $|e_{ij}| \le \gamma_{\lceil \log_2 i \rceil + 1}|\ell_{ij}|$.*

Proof. By induction on n. If $n = 1$ then,

$$\mathsf{fl}(x_1) = \frac{b_1}{\ell_{11}}(1 + \theta_1) = \frac{b_1}{(1 + \theta_1)\ell_{11}}$$

and the statement follows since $|\theta_1| \le \gamma_1$.

Now assume $n > 1$ and let $\bar{x} = (x_1, \ldots, x_{n-1})$, $\bar{b} = (b_1, \ldots, b_{n-1})$, and $\bar{L} \in \mathbb{R}^{(n-1)\times(n-1)}$ be the matrix obtained by removing the nth row and the nth column of L. Then, $\bar{L}$ is lower triangular, non-singular, and $\bar{L}\bar{x} = \bar{b}$. By induction hypothesis the point $\overline{\mathsf{fl}(x)} = (\mathsf{fl}(x_1), \ldots, \mathsf{fl}(x_{n-1}))$ computed at the first $(n-2)$ iterations of BS satisfies $(\bar{L} + \bar{E})\overline{\mathsf{fl}(x)} = \bar{b}$ with $|\bar{e}_{ij}| \le \gamma_{\lceil \log_2 i \rceil + 1}|\ell_{ij}|$.

We now use Proposition 1.4 to perform the $(n-1)$th iteration (which computes x_n) with $A = (\ell_{n1}, \ldots, \ell_{n,n-1}) \in \mathbb{R}^{1\times(n-1)}$. By this proposition, we

compute the product $A\overline{\mathsf{fl}(x)} = \ell_{n1}\mathsf{fl}(x_1) + \cdots + \ell_{n,n-1}\mathsf{fl}(x_{n-1})$ and obtain $\mathsf{fl}(w)$ satisfying

$$\mathsf{fl}(w) = (\ell_{n1} + e_{n1})\mathsf{fl}(x_1) + \cdots + (\ell_{n,n-1} + e_{n,n-1})\mathsf{fl}(x_{n-1})$$

with $|e_{nj}| \leq \gamma_{\lceil \log_2(n-1)\rceil+1}|\ell_{nj}|$ for $j \leq n-1$. We then compute x_n and we obtain

$$\begin{aligned}\mathsf{fl}(x_n) &= \mathsf{fl}\Big(\frac{b_n - \mathsf{fl}(w)}{\ell_{nn}}\Big) = \Big(\frac{(b_n - \mathsf{fl}(w))(1+\theta_1)}{\ell_{nn}}\Big)(1+\theta_1) \\ &= \frac{b_n - (\ell_{n1} + e_{n1})\mathsf{fl}(x_1) + \cdots + (\ell_{n,n-1} + e_{n,n-1})\mathsf{fl}(x_{n-1})}{\ell_{nn}(1+\theta_2)}\end{aligned}$$

and the result follows by taking $e_{nn} = \ell_{nn}\theta_2$ and E the matrix obtained by putting $\bar{E}$ in its upper-left $(n-1)\times(n-1)$ corner, appending $(e_{n1}, \ldots, e_{nn})$ as the nth row, and filling the remaining of the nth column with zeros. □

4.3 Componentwise Condition of Random Sparse Matrices

Proposition 4.6 justifies to measure relative errors componentwise and, as a consequence, to use componentwise condition numbers in the error analysis. The goal of this section is to give a (classical) probabilistic analysis for these condition numbers.

We will work in the more general context of sparse matrices (which, in this section, are matrices with a fixed pattern of zeros[1]). Therefore, the following results apply not only to triangular matrices but to other classes of sparse matrices such as, for instance, tridiagonal matrices. Also, in the process of proving our main result we will estimate as well the average componentwise condition for the computation of the determinant and matrix inversion.

4.3.1 Componentwise condition numbers revisited

Recall, for a function $\varphi : \mathcal{D} \subseteq \mathbb{R}^m \to \mathbb{R}^q$ and a point $a \in \mathcal{D}$ with $a_i \neq 0$ and $\varphi_j(a) \neq 0$ for all $i \leq m$ and $j \leq q$, we defined in (1.1) the componentwise condition number

$$\mathsf{Cw}^{\varphi}(a) = \lim_{\delta\to 0}\sup_{\mathsf{RelError}(a)\leq\delta}\frac{\mathsf{RelError}(\varphi(a))}{\mathsf{RelError}(a)}$$

where both $\mathsf{RelError}(a)$ and $\mathsf{RelError}(\varphi(a))$ are measured componentwise and we make the convention that $\frac{0}{0} = 1$. That is,

$$\mathsf{RelError}(a) = \max_{i\leq m}\frac{|\tilde{a}_i - a_i|}{|a_i|}$$

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

and similarly for $\varphi(a)$. In this case, we saw in Section 1.2 that we have $\mathsf{Cw}^{\varphi}(a) = \max_{j\le q} \mathsf{Cw}^{\varphi}_j(a)$ where, for $j \le q$,

$$\mathsf{Cw}^{\varphi}_j(a) = \lim_{\delta\to 0} \sup_{\mathsf{RelError}(a)\le\delta} \frac{\mathsf{RelError}(\varphi(a)_j)}{\mathsf{RelError}(a)}.$$

We want to extend these definitions to the general case where any of the components of a or $\varphi_j(a)$ may be zero.

The case of $a_i = 0$ is easily dealt with. We say that $\mathsf{RelError}(a) \le \delta$ for a perturbation $\tilde{a}$ when $|\tilde{a}_i - a_i| \le \delta|a_i|$ for all $i = 1, \dots, m$. The case of $\varphi_j(a) = 0$ for some $j \in [q]$ is dealt with the following definition.

Definition 4.7. For $j \le q$ such that $\varphi_j(a) \neq 0$, we let

$$\mathsf{Cw}_j(\varphi, a) := \lim_{\delta\to 0} \sup_{\mathsf{RelError}(a)\le\delta} \frac{|\varphi_j(\tilde{a}) - \varphi_j(a)|}{\mathsf{RelError}(a)|\varphi_j(a)|}$$

and for $j \le q$ with $\varphi_j(a) = 0$ we take $\mathsf{Cw}^{\varphi}_j(a) := 1$ if

$$\lim_{\delta\to 0} \sup_{\mathsf{RelError}(a)\le\delta} \frac{|\varphi_j(\tilde{a}) - \varphi_j(a)|}{\mathsf{RelError}(a)} = 0$$

and $\mathsf{Cw}^{\varphi}_j(a) = \infty$ otherwise. Then, we define

$$\mathsf{Cw}^{\varphi}(a) := \max_{j\le q} \mathsf{Cw}^{\varphi}_j(a).$$

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

Definition 4.8. 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

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

and $|S|$ for its cardinality. We denote by $\mathcal{R}_S$ 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 $\mathcal{R}$.

Remark 4.9. Note that by the definition of relative error $\mathscr{M}_S$ is closed under perturbations: if $A \in \mathscr{M}_S$ and $\tilde{A}$ is a perturbation with $\mathsf{RelError}(A) \le \delta$, then $\tilde{A} \in \mathscr{M}_S$.

In the rest of this chapter, for non-singular matrices $A, \tilde{A}$, we denote their inverses by $\Gamma, \tilde{\Gamma}$, respectively. Also, we denote by $A_{(ij)}$ the sub-matrix 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, $\gamma_{ij} = (-1)^{i+j} \frac{\det(A_{(ji)})}{\det(A)}$.

4.3.2 Determinant computation

We consider here the problem of computing the determinant of a matrix A and its componentwise condition number $\mathsf{Cw}^{\mathsf{det}}(A)$ which is defined by taking $\varphi : \mathscr{M} \to \mathbb{R}$ to be $\varphi(A) = \det(A)$ in Definition 4.7. Our main result for $\mathsf{Cw}^{\mathsf{det}}(A)$ is the following.

Theorem 4.10. *For $S \subseteq [n]^2$ and $t \geq 2|S|$ we have*

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

We may use Theorem 4.10 to estimate the average componentwise condition number for the computation of the determinant.

Corollary 4.11. *For a base $\beta \geq 2$ and a set $S \subseteq [n]^2$ with $|S| \geq 2$, we have $\mathbb{E}(\log_\beta \mathsf{Cw}^{\mathsf{det}}(A)) \leq 2 \log_\beta |S| + \log_\beta e$ where $\mathbb{E}$ denotes expectation over $A \in \mathcal{R}_S$.*

Proof. Use Propositions 3.18 and 4.12 below together with Theorem 4.10 taking $Z = \mathsf{Cw}^{\mathsf{det}}(A)$, $\alpha = 1$ and $t_0 = K = |S|^2$ (note that $|S|^2 \geq 2|S|$ since $|S| \geq 2$). □

Proposition 4.12. *For all $A \in \mathscr{M} \setminus \Sigma$ we have $\mathsf{Cw}^{\mathsf{det}}(A) \geq 1$.*

Proof. For each $\delta > 0$ consider $\tilde{A} \in \mathscr{M}$ with rows $a_1(1+\delta), a_2, \ldots, a_n$ where $a_1, \ldots, a_n$ are the rows of A. Then $\mathsf{RelError}(A) = \delta$ and

$$\mathsf{RelError}(\det A) = \frac{|\det(\tilde{A}) - \det(A)|}{|\det(A)|} = \delta. \qquad \square$$

Lemma 4.13. *For $A \in \mathscr{M} \setminus \Sigma$,*

$$\mathsf{Cw}^{\mathsf{det}}(A) = \sum_{i,j \in [n]} \left| \frac{a_{ij} \det(A_{(ij)})}{\det(A)} \right|.$$

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 $i, j \in [n]$,

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

Let $\delta > 0$ and $\tilde{A}$ such that $\mathsf{RelError}(A) \leq \delta$. Then, $\|\tilde{A} - A\| \leq \mathsf{RelError}(A)\|A\|$. Using Taylor's expansion and the equalities above we obtain

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

It follows that, for $A \notin \Sigma$,

$$\begin{aligned}
\mathsf{Cw}^{\mathsf{det}}(A) &= \lim_{\delta\to 0} \sup_{\mathsf{RelError}(A)\le\delta} \frac{|\det(\tilde{A}) - \det(A)|}{\mathsf{RelError}(A)|\det(A)|} \\
&= \lim_{\delta\to 0} \sup_{\mathsf{RelError}(A)\le\delta} \frac{\left|\sum_{i,j\in[n]}(-1)^{i+j}(\tilde{a}_{ij} - a_{ij})\det(A_{(ij)})\right|}{\mathsf{RelError}(A)|\det(A)|} \\
&= \lim_{\delta\to 0} \sup_{\mathsf{RelError}(A)\le\delta} \sum_{i,j\in[n]} \frac{|(\tilde{a}_{ij} - a_{ij})\det(A_{(ij)})|}{\mathsf{RelError}(A)|\det(A)|}.
\end{aligned}$$

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

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

Lemma 4.14. *Let p, q be two fixed vectors in $\mathbb{R}^n$ such that $\|p\| \le \|q\|$. If $x \sim N(0, \mathrm{I}_n)$ then, for all $t \ge 2$,*

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

Proof. Let $\nu = \|q\|$. By the rotational invariance of $N(0, \mathrm{I}_n)$ 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

$$\begin{aligned}
\mathsf{Prob}\left\{\left|\frac{x^{\mathrm{T}}p}{x^{\mathrm{T}}q}\right| \ge t\right\} &= \mathsf{Prob}\left\{\left|p_1 + \sum_{i\in\{2,\ldots,n\}} \frac{x_i p_i}{x_1}\right| \ge t\right\} \\
&= \mathsf{Prob}\left\{\left|p_1 + \frac{1}{x_1}\alpha Z\right| \ge t\right\} \qquad (4.1) \\
&= \mathsf{Prob}\left\{\frac{Z}{x_1} \ge \frac{t - p_1}{\alpha}\right\} + \mathsf{Prob}\left\{\frac{Z}{x_1} \le \frac{-t - p_1}{\alpha}\right\}
\end{aligned}$$

where $Z = N(0,1)$ is independent of x_1 and $\alpha = \sqrt{p_2^2 + \ldots + p_n^2} \le 1$. Here we used that a sum of independent centered Gaussians is a centered Gaussian whose variance is the sum of the terms' variances (cf. §3.1.1). Note that in case $\alpha = 0$ the statement of the lemma is trivially true.

The random variable x_1 and Z are independent $N(0,1)$. It therefore follows from Proposition 3.10 that the angle $\theta = \arctan(Z/x_1)$ is uniformly distributed

in $[-\pi/2, \pi/2]$. Hence, for $\gamma \in [0, \infty)$,

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

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

$$\mathsf{Prob}\left\{\frac{Z}{x_1} \leq \sigma\right\} \;\leq\; \frac{1}{\pi(-\sigma)}.$$

Using these bounds in (4.1) with $\gamma = \frac{t-p_1}{\alpha}$ and $\sigma = \frac{-t-p_1}{\alpha}$ we obtain

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

the last since $t \geq 2$. □

Lemma 4.15. *Let $S \subseteq [n]^2$. Then,*

1. *if $\mathscr{M}_S \subseteq \Sigma$ then, for all $A \in \mathscr{M}_S$, $\mathsf{Cw}^{\mathsf{det}}(A) = 1$,*
2. *if $\mathscr{M}_S \not\subseteq \Sigma$, then $\underset{A\in\mathcal{R}_S}{\mathsf{Prob}}(A \text{ is singular}) = 0$.*

Proof. Since $\mathscr{M}_S \subseteq \Sigma$ and $A \in \mathscr{M}_S$ we have, for all $\delta > 0$, that if $\tilde{A}$ is such that $\mathsf{RelError}(A) \leq \delta$ then $\tilde{A} \in \Sigma$. Part (1) now follows.

For part (2), we note that 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 4.10. Case (i): $\mathscr{M}_S \subseteq \Sigma$. In this case, the desired inequality is trivial by Lemma 4.15(1).

Case (ii): $\mathscr{M}_S \not\subseteq \Sigma$. By Lemma 4.15(2), with probability 1, A is non-singular. So, by Lemma 4.13,

$$\begin{aligned}\mathsf{Prob}\{\mathsf{Cw}^{\mathsf{det}}(A) \geq t\} &= \mathsf{Prob}\left\{\sum_{(i,j)\in S}\left|\frac{a_{ij}\det(A_{(ij)})}{\det(A)}\right| \geq t\right\} \\ &\leq \sum_{(i,j)\in S}\mathsf{Prob}\left\{\left|\frac{a_{ij}\det(A_{(ij)})}{\det(A)}\right| \geq \frac{t}{|S|}\right\}. \qquad (4.2)\end{aligned}$$

It is therefore enough to prove that, for all $(i,j) \in S$ and all $z > 0$,

$$\mathsf{Prob}\left\{\left|\frac{a_{ij}\det(A_{(ij)})}{\det(A)}\right| \geq z\right\} \leq \frac{1}{z}. \qquad (4.3)$$

Without loss of generality, take $(i,j) = (1,1)$. 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,

$$x_I \sim N(0, \mathrm{I}_{|I|}). \tag{4.4}$$

For $i \in [n]$ write $q_i = (-1)^{i+1} \det(A_{(i1)})$. Let $q = (q_1, \ldots, q_n)$ 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

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

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

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

Let ρ be the density of the random vector q_I. Then, for $z \geq 2$,

$$\begin{aligned} & \mathsf{Prob}\left\{\left|\frac{a_{11}\det(A_{(11)})}{\det(A)}\right| \geq z\right\} = \mathsf{Prob}\left\{\left|\frac{x_I^{\mathrm{T}}(q_1 e_1)}{x_I^{\mathrm{T}} q_I}\right| \geq z\right\} \\ = & \int_{u \in \mathbb{R}^{|I|}} \mathsf{Prob}\left\{\left|\frac{x_I^{\mathrm{T}}(u_1 e_1)}{x_I^{\mathrm{T}} u}\right| \geq z \;\middle|\; q_I = u\right\} \rho(u) du \\ \leq & \int_{u \in \mathbb{R}^{|I|}} \frac{1}{z}\, \rho(u)\, du \;=\; \frac{1}{z}. \end{aligned}$$

Here the inequality follows since x_I is independent of q_1 and q_I and therefore we can use (4.4) and Lemma 4.14 (with $p = q_1 e_1$ and $q = q_I$). This proves (4.3) and hence the lemma. □

4.3.3 Matrix inversion

We now focus on the problem of inverting a matrix A and its componentwise condition number $\mathsf{Cw}^{\dagger}(A)$ obtained from Definition 4.7 by taking $\mathcal{D} = \mathscr{M} \setminus \Sigma$ and $\varphi : \mathscr{M} \setminus \Sigma \to \mathscr{M}$ given by $\varphi(A) = A^{-1}$. Our main results for $\mathsf{Cw}^{\dagger}(A)$ are the following two.

Theorem 4.16. *Let $S \subseteq [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then, for all $t \geq 4|S|$,*

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

Using Propositions 3.18 and 4.18 below we obtain the following corollary.

Corollary 4.17. *Let $S \subseteq [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then,*

$$\mathbb{E}(\log_\beta(\mathsf{Cw}^\dagger(A))) \leq 2\log_\beta n + 2\log_\beta |S| + \log_\beta 4e$$

where $\mathbb{E}$ denotes expectation over $A \in \mathcal{R}_S$. □

Proposition 4.18. *For all $A \in \mathscr{M} \setminus \Sigma$ we have $\mathsf{Cw}^\dagger(A) \geq 1$.*

Proof. Take $\tilde{A} = (1-\delta)A$ so that $\tilde{A}^{-1} = \frac{1}{1-\delta}A^{-1} = A^{-1} + \delta A^{-1} + o(\delta)$. Now reason as in Proposition 4.12. □

Lemma 4.19. *For $A \in \mathscr{M} \setminus \Sigma$ and $k, \ell \in [n]$,*

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

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

Let $\delta > 0$ be sufficiently small so that if $\mathsf{RelError}(A) \leq \delta$ then $\tilde{A} \notin \Sigma$ and $\left|\frac{\det(\tilde{A})-\det(A)}{\det(A)}\right| < 1$. Let $\tilde{A}$ be such that $\mathsf{RelError}(A) \leq \delta$.

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

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

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

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

and therefore

$$\begin{aligned}
&\sup_{\mathsf{RelError}(A)\leq\delta}\left|\frac{\tilde{\gamma}_{k\ell} - \gamma_{k\ell}}{\mathsf{RelError}(A)\gamma_{k\ell}}\right| \\
&\quad\leq \frac{\sup_{\mathsf{RelError}(A)\leq\delta}\left|\frac{\det(\tilde{A}_{(\ell k)})-\det(A_{(\ell k)})}{\mathsf{RelError}(A)\det(A_{(\ell k)})}\right| + \sup_{\mathsf{RelError}(A)\leq\delta}\left|\frac{\det(\tilde{A})-\det(A)}{\mathsf{RelError}(A)\det(A)}\right|}{1 - \sup_{\mathsf{RelError}(A)\leq\delta}\left|\frac{\det(\tilde{A})-\det(A)}{\det(A)}\right|}.
\end{aligned}$$

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

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

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

$$\lim_{\delta \to 0} \sup_{\mathsf{RelError}(A) \leq \delta} \frac{|\tilde{\gamma}_{k\ell}|}{\mathsf{RelError}(A)} = 0.$$

In this case, $\mathsf{Cw}^{\dagger}_{k\ell}(A) = 1$ and the statement holds.

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

$$0 \neq \lim_{\delta \to 0} \sup_{\mathsf{RelError}(A) \leq \delta} \frac{|\tilde{\gamma}_{k\ell}|}{\mathsf{RelError}(A)} = \lim_{\delta \to 0} \sup_{\mathsf{RelError}(A) \leq \delta} \frac{|\det(\tilde{A}_{\ell k})|}{\mathsf{RelError}(A)|\det(\tilde{A})|}.$$

In this case $\mathsf{Cw}^{\mathsf{det}}(A_{\ell k}) = \infty$ and the statement holds as well. □

Proof of Theorem 4.16. By definition of $\mathsf{Cw}^{\dagger}(A)$,

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

By Lemma 4.15(2), A is non-singular with probability 1. So, since $\frac{t}{2} \geq 2|S|$ by hypothesis, we can apply Lemma 4.19 to obtain

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

the last inequality by applying Theorem 4.10 to A and $A_{(k\ell)}$. The statement now follows. □

4.3.4 Linear equations solving

We finally deal with the problem of solving linear systems of equations. That is, we consider a matrix $A \in \mathscr{M}_S$ and a vector $b \in \mathbb{R}^n$ and we want to solve $Ax = b$. We denote by $\mathsf{Cw}(A, b)$ the corresponding componentwise condition number which, again, is obtained from Definition 4.7 by taking $\mathcal{D} = (\mathscr{M} \setminus \Sigma) \times \mathbb{R}^n$ and $\varphi : (\mathscr{M} \setminus \Sigma) \times \mathbb{R}^n \to \mathbb{R}^n$ given by $\varphi(A, b) = A^{-1}b$.

Theorem 4.20. *Let $S \subseteq [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then, for all $t \geq 4(|S|+n)$,*

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

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

We may use Proposition 3.18 once more (together with Proposition 4.22 below).

Corollary 4.21. *Let $S \subseteq [n]^2$ be such that $\mathscr{M}_S \not\subseteq \Sigma$. Then,*

$$\mathbb{E}(\log_\beta(\mathsf{Cw}(A,b))) \leq \log_\beta n + 2\log_\beta |S| + \log_\beta 10e. \qquad \square$$

Proposition 4.22. *For all $A \in \mathscr{M} \setminus \Sigma$ and all $b \in \mathbb{R}^n$ we have $\mathsf{Cw}(A,b) \geq 1$.*

Proof. Take $\tilde{A} = A$ and $\tilde{b} = (1+\delta)b$ so that $\tilde{A}^{-1}\tilde{b} = (1+\delta)x$ where $x = A^{-1}b$. Now reason as in Proposition 4.12. $\square$

Recall that we denote by $A[k:b]$ the matrix obtained by replacing the kth column of A by b.

Lemma 4.23. *For any non-singular matrix A and $k \in [n]$ we have*

$$\mathsf{Cw}_k(A,b) \leq \mathsf{Cw}^{\mathsf{det}}(A) + \mathsf{Cw}^{\mathsf{det}}(A[k:b]).$$

Proof. By Cramer's rule,

$$x_k = \frac{\det(A[k:b])}{\det(A)}.$$

The rest of this proof is similar to the proof of Lemma 4.19. $\square$

Proof of Theorem 4.20. It follows the lines of that of Theorem 4.16. First, we get

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

Then, we apply Lemma 4.23 and Theorem 4.10 (using that, with probability 1, $A \notin \Sigma$ and that $\frac{t}{2} \geq 2|S|$) to get

$$\begin{aligned}\mathsf{Prob}\{\mathsf{Cw}_k(A,b) \geq t\} &\leq \mathsf{Prob}\left\{\mathsf{Cw}^{\mathsf{det}}(A) \geq \frac{t}{2}\right\} + \mathsf{Prob}\left\{\mathsf{Cw}^{\mathsf{det}}(A[k:b]) \geq \frac{t}{2}\right\} \\ &\leq 2|S|^2\frac{1}{t} + 2(|S|+n)^2\frac{1}{t} \leq 10|S|^2\frac{1}{t}.\end{aligned}$$

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

4.4 Error Bounds for Triangular Linear Systems

We may now use the results in the preceding sections to estimate the expected loss of precision in the solution of a triangular system $Lx = b$.

Theorem 4.24. *Assume we solve $Lx = b$ using algorithm* BS. *Then, for standard Gaussian L and b we have*

$$\mathbb{E}\big(\mathsf{LoP}(L^{-1}b)\big) \leq 5\log_\beta n + \log_\beta(\lceil\log_2 n\rceil + 1) + \log_\beta 10e + o(1).$$

Proof. By Proposition 4.6 and Theorem 1.3 (with $f(\mathsf{dims}(L,b)) = \lceil\log_2 n\rceil + 1$) we have

$$\mathsf{LoP}(L^{-1}b) \leq \log_\beta(\lceil\log_2 n\rceil + 1) + \log_\beta \mathsf{Cw}(L,b) + o(1).$$

Therefore, using Corollary 4.21 with $|S| = \frac{n^2+n}{2}$,

$$\begin{aligned}\mathbb{E}\big(\mathsf{LoP}(L^{-1}b)\big) &\leq \log_\beta(\lceil\log_2 n\rceil + 1) + \mathbb{E}(\log_\beta \mathsf{Cw}(L,b)) + o(1)\\ &\leq \log_\beta(\lceil\log_2 n\rceil + 1) + 5\log_\beta n + + \log_\beta 10e + o(1). \qquad \square\end{aligned}$$

If $\mathsf{fl}(x) = (\mathsf{fl}(x_1), \ldots, \mathsf{fl}(x_n))$ is the solution of $Lx = b$ computed by BS, the number of correct significant figures of its ith component is

$$\left|\log_\beta \frac{|\mathsf{fl}(x_i) - x_i|}{|x_i|}\right|.$$

We can rephrase Theorem 4.24 stating that for standard Gaussian L and b

$$\begin{aligned}&\mathbb{E}\left(\min_{i\leq n}\left|\log_\beta \frac{|\mathsf{fl}(x_i) - x_i|}{|x_i|}\right|\right)\\ &\qquad \geq t - \Big(5\log_\beta n + \log_\beta(\lceil\log_2 n\rceil + 1) + \log_\beta 10e + o(1)\Big)\end{aligned}$$

where $t = |\log_\beta \epsilon_{\mathsf{mach}}|$ is the number of significant figures the machine works with (compare §1.3.2).