S©MR

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ
Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru
ISSN 1813-3304

Vol. 21, No. 2, pp. A70–A81 (2024)
https://doi.org/10.33048/semi.2024.21.A04

УДК 519.61
MSC 65F22

REGULARIZED CHOLESKY DECOMPOSITION METHOD FOR FINITE BIT WIDTH COMPUTING

Z. ZHANG, **V. LYASHEV**

Dedicated to 85th birthday of academician Vladimir G. Romanov

Abstract: This research focuses on the Cholesky decomposition of symmetric positive definite matrices. While the Cholesky decomposition is known for its computational efficiency and numerical robustness, it may encounter decomposition failures when applied to ill-conditioned matrices with large condition numbers. To address these computational challenges, this paper proposes an improved probabilistic rounding error analysis method. This method can more accurately estimate the rounding errors and thereby guide the selection of the optimal diagonal loading value. The main contribution of this research is the determination of a diagonal loading value applicable to all positive definite matrices, ensuring the successful completion of Cholesky decomposition. In addition, taking into account the binary representation of numbers in computers, the diagonal loading value is converted to exponential form, allowing multiplication to be replaced by the floating-point bitwise operations. This approach is both practical and efficient, effectively solving the challenges posed by ill-conditioned matrices and limited computational precision.

Keywords: cholesky decomposition, diagonal loading, regularization, numerical robustness, low bit-width computations, probabilistic rounding error analysis.

Zhang, Z., Lyashev, V., Regularized Cholesky decomposition method for finite bit width computing.

Received November, 21, 2024, Published December, 31, 2024.

1 Introduction

Wiener filter-based algorithms are widely used in modern digital signal processing, antenna combining techniques, receivers [1], MMSE channel estimation algorithms [2], etc. A pivotal aspect of these algorithms is their reliance on the accurate inversion of the covariance matrix, a process recognized as a classic problem in digital signal processing. This is particularly crucial when dealing with tasks like multidimensional parameter estimation of a linear system. In such a context, the weight matrix $\mathbf{W}$ based on Wiener filtering is fundamentally dependent on the inverse of the covariance matrix. Since the covariance matrix is a symmetric positive definite matrix, using the Cholesky decomposition combined with the forward and backward substitution method for triangular matrices to find the inverse is a method with the lowest computational complexity and the highest numerical stability.

Cholesky decomposition is favored for its computational efficiency and numerical robustness, but it encounters challenges when dealing with ill-conditioned matrices with very large condition numbers [3]. The main issue is that if the condition number is too large, the computational process becomes unstable, which can lead to negative numbers or zeros on the diagonal of the Cholesky factor during the Cholesky decomposition process, thus causing the decomposition to fail. This instability is particularly evident when computational precision is limited, such as when using 16-bit floating-point arithmetic and 32-bit floating-point arithmetic. Several algorithms have been proposed to overcome these computational challenges, among which the improved Cholesky decomposition [4] stands out. Although, the diagonal loading method is becoming increasingly popular in practical applications through its relatively easy implementation, determining the optimal loading value for this method remains a significant research challenge. Rounding errors are a critical issue when considering low-bit-width computations. Traditional deterministic rounding error analysis [5] often overestimates rounding error and provides limited guidance for practical implementations.

In this paper, we introduce probabilistic rounding error analysis [6] as the theoretical basis. This approach offers a more accurate estimate of the rounding error, allowing for a more accurate assessment of algorithmic stability and reliability, and provides a theoretical rationale for the selection of diagonal loading values. The main contribution of this paper is to obtain a diagonal loading value which ensures that any positive definite matrix can successfully complete the Cholesky decomposition.

Throughout this paper we use the following notations to denote scalars, vectors and matrices respectively: $a, \mathbf{b}, \mathbf{C}$. The condition number of the matrix $\mathbf{A}$ is referred as $k_2(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2$. As an element-by-element comparison between the two matrices goes the expression $|\mathbf{A}| < |\mathbf{B}|$ signifies that the absolute value of each element in the matrix $\mathbf{A}$ is less than the absolute value of the corresponding element in the matrix $\mathbf{B}$.

2 Problem statement

The Cholesky decomposition, named after the French military officer and mathematician Andre-Louis Cholesky (1875–1918), is an efficient method commonly used to compute the inverse of symmetric positive definite matrices. Aforementioned type of factorization is backward stable. However, if the computational precision is limited or the condition number of the matrix $\mathbf{A}$ (where $\mathbf{A}$ is a symmetric positive definite matrix) is too large, the decomposition fails due to rounding errors. A more intuitive explanation can be found in the implementations of the Cholesky decomposition presented in [8]. Algorithm 1 is the most common floating-point implementation of Cholesky decomposition, which is based on gaxpy computations and suitable for deployment on vector processors.

If this algorithm is applied to an ill-conditioned matrix, due to the accumulation of rounding errors, the computer may attempt to take the square root of a negative number $\mathbf{v}[j]$ while processing line 8, causing to the the decomposition to fail. Alternatively, it may result in $\mathbf{L}[j,j] = 0$, in which case $\mathbf{v}[j:n]$ will be divided by zero on the next iteration. Even for positive definite matrices the algorithm may fail due to these issues when using floating-point arithmetic.

Algorithm 1 Cholesky Decomposition [8]

Input: Symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$

1: $n \leftarrow \text{length}(\mathbf{A})$
2: $\mathbf{L} \leftarrow \mathbf{0}_{n\times n}$ {Initialize $\mathbf{L}$ as an $n \times n$ zero matrix}
3: **for** $j = 1$ to n **do**
4: **for** $i = 1$ to j **do**
5: $\mathbf{L}[j,i] \leftarrow \mathbf{A}[i,j] - \mathbf{L}[i,1:i] \times \mathbf{L}[j,1:i]^{\mathrm{T}}$
6: $\mathbf{L}[j,i] \leftarrow \mathbf{L}[j,i]/\mathbf{L}[i,i]$
7: **end for**
8: $\mathbf{L}[j,j] \leftarrow \sqrt{\mathbf{A}[j,j] - \mathbf{L}[j,1:j]\mathbf{L}[j,1:j]^{\mathrm{T}}}$
9: **end for**

Output: Lower triangular matrix $\mathbf{L} \in \mathbb{R}^{n\times n}$ such that $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$

Wilkinson [9] conducted a comprehensive analysis of the computational conditions for Cholesky decomposition. He pointed out that the Cholesky decomposition can be guaranteed to be complete if $q_n u k_2(\mathbf{A}) \leq 1$, where q_n is a small constant, and u is the unit rounding error (which depends on machine precision). To ensure the completion of the factorization, the paper [10] proposes a small offset parameter s to the matrix $\mathbf{A}$, keeping the condition number of $\mathbf{A}$ within an acceptable range. However, to obtain the offset parameter s, it is necessary to first compute the Frobenius norm of the $\mathbf{A}$ matrix, which adds unnecessary complexity. Modified Cholesky decomposition [4] is also a solution, where the authors compute diagonal

loading values during the Cholesky decomposition process to minimize perturbations. However, this method is more complex and disrupts the computational pipeline of the Cholesky decomposition, leading to increased complexity in hardware implementation.

We suggest a method similar to that of Fukaya [10], which involves adding a small offset δ to the diagonal of the matrix $\mathbf{A}$ as follows:

$$\hat{\mathbf{A}} = \mathbf{A} + \delta\mathbf{D}, \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with elements identical to those on the diagonal of the matrix $\mathbf{A}$.

This method can protect the smallest eigenvalue of $\mathbf{A}$ from falling below a certain threshold, thus ensuring that the condition number of the matrix does not become too large. The δ here is an exponent with a base of 2, and its exponent is a negative integer. Computing $\delta\mathbf{D}$ does not require any multiplication operations, just a shift in floating-point arithmetic. Therefore, this method hardly adds any complexity, requiring only n addition operations to ensure completion of the Cholesky decomposition.

3 Diagonal loading for the Cholesky decomposition

This section analyzes the rounding errors in the Cholesky decomposition based on the probabilistic rounding error model. Subsequently, as we investigate the relationship between the error in Cholesky decomposition and the diagonal elements of matrix $\mathbf{A}$, we derive a formula to calculate the optimal diagonal loading value δ, based on the theoretical foundation available to the date.

3.1. Error analysis of Cholesky decomposition. In numerical linear algebra, traditional rounding error analysis provides deterministic backward error bounds that depend on $\gamma_n = nu/(1-nu)$, where n is the size of the matrix, and u is the unit rounding error. This type of rounding error analysis offers an important framework for understanding and assessing the accumulation of errors in computations.

However, as for low-precision computations, these deterministic backward error bounds cannot provide useful information. Higham [6] has developed a new probabilistic rounding error analysis method. It uses concentration inequalities [11] and makes probabilistic estimates of rounding errors. The research has shown that the inner product error is approximately $\sqrt{n}u$, which aligns with simulation results.

A pioneering contribution by Higham [6] was the introduction of two critical expressions in the field of probabilistic rounding error analysis. The first expression, formulated as

$$\widetilde{\gamma}_n(\lambda) = \exp\left(\lambda\sqrt{n}u + \frac{nu^2}{1-u}\right) \leq \lambda\sqrt{n}u + O(u^2) \approx \lambda\sqrt{n}u \tag{2}$$

is used to determine the probabilistic bounds of backward rounding errors. The second expression, formulated as

$$Q(\lambda, n) = 1 - 2n \exp\left(-\frac{\lambda^2(1-u)^2}{2}\right) \tag{3}$$

quantifies the probability with that rounding errors will exceed a specified threshold in computations. In these expressions, λ acts as a tuning parameter, often referred as a relaxation constant, which adjusts the probability bounds.

Furthermore, Higham extended this probabilistic rounding error framework to analyze the backward error in Cholesky decomposition, demonstrating its applicability in a broader range of computational contexts.

Theorem 1 (Cholesky decomposition error analysis)**.** *If Cholesky decomposition applied to the symmetric positive definite matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *runs to completion then the computed factor* $\tilde{\mathbf{L}}$ *satisfies*

$$\mathbf{A} + \Delta\mathbf{A} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^{\mathrm{T}}, \qquad |\Delta\mathbf{A}| \leq \widetilde{\gamma}_{n+1}(\lambda)|\tilde{\mathbf{L}}||\tilde{\mathbf{L}}|^{\mathrm{T}}, \tag{4}$$

where $\Delta\mathbf{A}$ *is a perturbation matrix of* $\mathbf{A}$*, with probability at least* $Q(\lambda, n^3/6 + n^2/2 + n/3)$.

Proof. See Theorem 3.8 in [6]. □

Aforementioned probability model may not be perfect in some cases, as it may yield negative results, which is unreasonable in probability theory. Nonetheless, for sufficiently large values of λ, $Q(\lambda, n)$ remains within the range of $[0, 1]$, making the model effective in these cases.

Таблица 1. Values of $Q(\lambda, n)$ in (3) for half precision and single precision arithmetic with $n = 32$.

λ	half	single
4.5	0.5157	0.5205
5	0.9548	0.9554
5.5	0.9967	0.9968
6	0.9998	0.9998

As the value of λ increases, the $Q(\lambda, n)$ rapidly approaches 1, as shown in the Table 1. This indicates that the higher the relaxation constant λ, the lower the probability with that the error in the computation will exceed the threshold.

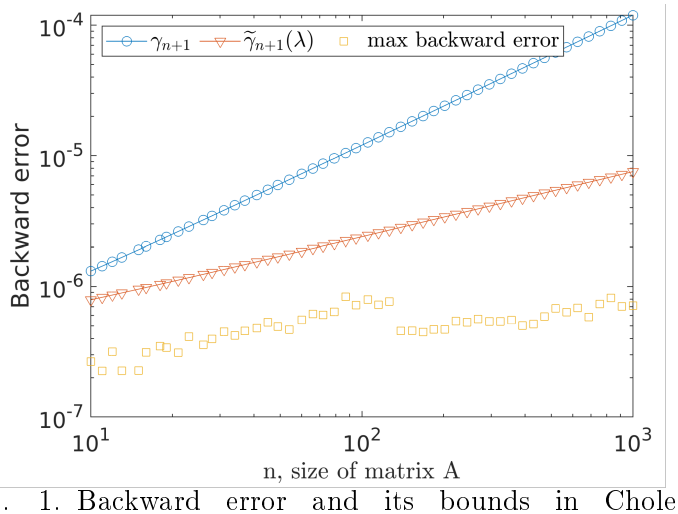

FIG. 1. Backward error and its bounds in Cholesky decomposition in single precision. Here, $N_{test} = 100$ and $\lambda = 2$.

We generate matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$ randomly from the normal distribution using the 'randn' function in Matlab. Subsequently, we perform matrix multiplication to obtain $\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$. Although the condition number of these matrices could be any value greater than zero, it is bounded by the Marchenko-Pastur law [7].

For each matrix size n, the Cholesky decomposition is applied to $\mathbf{A}$, yielding the lower triangular matrix $\tilde{\mathbf{L}}$. We conduct N_{test} numerical experiments to calculate the backward rounding error as follows:

$$\varepsilon = \frac{|\Delta\mathbf{A}|}{|\tilde{\mathbf{L}}||\tilde{\mathbf{L}}|^{\mathrm{T}}}. \tag{5}$$

Afterwards, we estimate the maximum backward error observed across these experiments.

As can be seen from Figure 1, the probability $Q(\lambda, f(n))$ is actually quite conservative. In practical simulations, smaller values of λ are already sufficient to meet the error bound requirements. This indicates that although the probabilistic error analysis method is better than the deterministic error analysis method, it is still quite discreet. Therefore, this approach ensures the reliability of the error bounds even in the worst-case scenario.

Theorem 2 (Diagonal-Dependent Stability of Cholesky Decomposition)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. If the Cholesky decomposition is applied to $\mathbf{A}$ and reaches completion, then the computed factor $\tilde{\mathbf{L}}$ satisfies the following condition:*

$$\mathbf{A} + \Delta\mathbf{A} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^{\mathrm{T}}, \quad |\Delta\mathbf{A}| \leq (1 - \widetilde{\gamma}_{n+1})^{-1}\widetilde{\gamma}_{n+1}\mathbf{d}\mathbf{d}^{\mathrm{T}}, \tag{6}$$

where $\Delta\mathbf{A}$ is a perturbation matrix of $\mathbf{A}$, d_i is the square root of the diagonal elements of $\mathbf{A}$, with probability at least $Q(\lambda, n^3/6 + n^2/2 + n/3)$.

Proof. Theorem 1 states that with probability at least $Q(\lambda, n^3/6 + n^2/2 + n/3)$, the bound of $|\Delta\mathbf{A}|$ is given by $\widetilde{\gamma}_{n+1}(\lambda)|\tilde{\mathbf{L}}||\tilde{\mathbf{L}}^{\mathrm{T}}|$. Let $\tilde{\mathbf{l}}_i$ denote the i-th row of $\tilde{\mathbf{L}}$.

Then, we have

$$\|\tilde{\mathbf{l}}_i\|_2^2 = \tilde{\mathbf{l}}_i\tilde{\mathbf{l}}_i^{\mathrm{T}} = a_{ii} + \Delta a_{ii} \leq a_{ii} + \widetilde{\gamma}_{n+1}|\tilde{\mathbf{l}}_i||\tilde{\mathbf{l}}_i|^{\mathrm{T}}, \tag{7}$$

which implies that $\|\tilde{\mathbf{l}}_i\|_2^2 \leq (1-\widetilde{\gamma}_{n+1})^{-1}a_{ii}$. Applying the Cauchy-Schwarz inequality, we obtain

$$|\tilde{\mathbf{l}}_j\tilde{\mathbf{l}}_i^{\mathrm{T}}| \leq \|\tilde{\mathbf{l}}_i\|_2\|\tilde{\mathbf{l}}_j\|_2 \leq (1-\widetilde{\gamma}_{n+1})^{-1}(a_{ii}a_{jj})^{1/2}, \tag{8}$$

leading to

$$|\tilde{\mathbf{L}}||\tilde{\mathbf{L}}|^{\mathrm{T}} \leq (1-\widetilde{\gamma}_{n+1})^{-1}\mathbf{d}\mathbf{d}^{\mathrm{T}}, \tag{9}$$

which provides the necessary bound for $\Delta\mathbf{A}$. □

This theorem was originally outlined and proved by Demmel [12]. However, since he used the traditional deterministic error analysis method, the rounding error bounds given are relatively broad. In contrast, applying the probabilistic rounding error analysis method can yield more compact error bounds. The probabilistic method takes into account the statistical characteristics of rounding errors, offering more precise and realistic error bounds.

In the following subsection, we will apply this theorem to derive the expression for calculating the optimal diagonal loading value. By analyzing the size of the matrix $\mathbf{A}$ and the rounding errors, we will be able to determine an optimal value for diagonal loading.

3.2. Optimal diagonal loading value.

Theorem 3 (Regularization For Cholesky Decomposition Completion)**.** *Let* $\mathbf{A}$ *be* $\mathbb{R}^{n\times n}$ *symmetric and positive definite. If the diagonal loading values satisfy the condition*

$$\delta > n\widetilde{\gamma}_{n+1}/(1-\widetilde{\gamma}_{n+1}), \tag{10}$$

then the Cholesky decomposition applied to $\hat{\mathbf{A}} = \mathbf{A} + \delta\mathbf{D}$*, where* $\mathbf{D}$ *is the diagonal matrix composed of the diagonal elements of* $\mathbf{A}$*, can be completed (excluding underflow and overflow issues), with probability at least* $Q(\lambda, n^3/6 + n^2/2 + n/3)$.

Proof. Assume that the algorithm has successfully completed $k-1$ stages, yielding a nonsingular $\tilde{\mathbf{L}}_{k-1}$. At the k-th step, we consider the partitioned matrix:

$$\hat{\mathbf{A}}_k = \begin{bmatrix} \hat{\mathbf{A}}_{k-1} & \mathbf{a} \\ \mathbf{a}^{\mathrm{T}} & b \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{L}}_{k-1} & 0 \\ \mathbf{l}^{\mathrm{T}} & \sqrt{b-\mathbf{l}^{\mathrm{T}}\mathbf{l}} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{L}}_{k-1}^{\mathrm{T}} & \mathbf{l} \\ 0 & \sqrt{b-\mathbf{l}^{\mathrm{T}}\mathbf{l}} \end{bmatrix} = \tilde{\mathbf{L}}_k\tilde{\mathbf{L}}_k^{\mathrm{T}}, \tag{11}$$

where $\hat{\mathbf{A}}_k$, $\hat{\mathbf{A}}_{k-1}$ are the leading principal submatrices of $\hat{\mathbf{A}}$ and $\mathbf{l} \in \mathbb{R}^{k-1}$ is the k-th row of $\mathbf{L}$ up to column k-1.

During the calculation, we may encounter the case where $b - \mathbf{l}^{\mathrm{T}}\mathbf{l} < 0$. This step would result in an attempt to compute the square root of a

negative number, yielding an imaginary value and causing the Cholesky decomposition to fail.

However, even in this case, the error analysis presented in Theorem 2 remains valid. The error bounds derived in Theorem 2 are independent of the successful completion of the Cholesky decomposition. They provide a perturbation $\Delta\hat{\mathbf{A}}_k$ such that:

$$\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k = \tilde{\mathbf{L}}_k\tilde{\mathbf{L}}_k^{\mathrm{T}}, \quad |\Delta\hat{\mathbf{A}}_k| \leq (1-\widetilde{\gamma}_{k+1})^{-1}\widetilde{\gamma}_{k+1}\sqrt{\mathbf{d}_k}\sqrt{\mathbf{d}_k}^{\mathrm{T}}, \tag{12}$$

where $\mathbf{d}_k = [a_{11}, \ldots, a_{kk}]^{\mathrm{T}}$. Now, let $\mathbf{D}_k = \mathrm{diag}(\mathbf{d}_k)$, it follows that

$$\begin{aligned}
\lambda_{\min}(\mathbf{D}_k^{-\frac{1}{2}}(\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k)\mathbf{D}_k^{-\frac{1}{2}}) &= \lambda_{\min}(\mathbf{D}_k^{-\frac{1}{2}}(\mathbf{A}_k + \delta\mathbf{D}_k + \Delta\hat{\mathbf{A}}_k)\mathbf{D}_k^{-\frac{1}{2}}) \\
&= \lambda_{\min}(\mathbf{H}_k + \delta + \mathbf{D}_k^{-\frac{1}{2}}\Delta\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}) \\
&\geq \lambda_{\min}(\mathbf{H}_k) + \delta - \|\mathbf{D}_k^{-\frac{1}{2}}\Delta\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\|_2 \\
&\geq \lambda_{\min}(\mathbf{H}_k) + \delta - \frac{\widetilde{\gamma}_{k+1}}{1-\widetilde{\gamma}_{k+1}}\|\mathbf{1}_k\|_2 \\
&\geq \lambda_{\min}(\mathbf{H}_k) + \delta - \frac{k\widetilde{\gamma}_{k+1}}{1-\widetilde{\gamma}_{k+1}} > 0.
\end{aligned} \tag{13}$$

In this context, $\mathbf{H}_k$ is defined as $\mathbf{H}_k = \mathbf{D}_k^{-\frac{1}{2}}\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}$ and $\mathbf{1}_k$ represents a $k \times k$ matrix with all elements equal to 1.

Given this definition, it follows that the matrix $\mathbf{D}_k^{-1}(\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k)\mathbf{D}_k^{-1}$ is positive definite. Thus, the congruent matrix $\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k$ is also positive definite. This result is significant because it implies that $\tilde{\mathbf{L}}_k$ is necessarily real and non-singular.

The theorem is proven based on the principle of induction. □

Theorem 2 naturally links the diagonal loading value to the diagonal elements of the matrix $\mathbf{A}$. Higham [6] and Demmel [12] had similar theorems proven previously. Our diagonal loading variant of these theorems presented above integrates the probabilistic rounding error analysis.

According to equation (10), we have derived a diagonal loading value that depends only on the unit rounding error precision and the size of the matrix. This value is very small and does not require any additional computation of the norm of the matrix $\mathbf{A}$. By transforming this value with the logarithmic function $\log_2$, we make it easier to handle and adapt to the binary representation of numbers in computers. Finally, we use the 'fix' function to adjust the resulting value to an exponent with a base of 2. This adjustment ensures that the final diagonal loading value is a small decimal with a base of 2 and a negative integer exponent. This allows all related multiplications to be performed efficiently using bit-wise floating-point operations, improving computational efficiency.

From this, we have obtained the following equation for calculating the optimal diagonal loading:

$$\delta(\lambda, n, u) = \mathrm{fix}\left(\log_2\left(\frac{n\sqrt{n}u\lambda}{1-\sqrt{n}u\lambda}\right)\right). \tag{14}$$

It can be seen that this expression perfectly incorporates both the matrix size and the rounding errors. We believe that this is a sufficiently simple yet effective diagonal loader.

Although the probabilistic error analysis is elegant, it tends to provide overly pessimistic probability lower bounds, as can be seen in Figure 1. Its most significant contribution is the demonstration that rounding errors do not increase linearly with the growth of the size of a matrix. Through extensive numerical analysis and simulations, we have set the parameter λ to a value of 2. When $\lambda = 2$, there are no events that exceed the error bounds.

Furthermore, based on equation (14), we calculated diagonal loading values for matrices varying their size, as shown in Table 2. These values were compared with the diagonal loading values from the deterministic rounding error analysis. The comparison reveals that our diagonal loading values are smaller, thereby minimizing the bias in the computational results while still ensuring the smooth execution of the Cholesky decomposition.

Таблица 2. Diagonal Loading Values: Probabilistic vs. Deterministic Analysis in Single Precision Arithmetic

n	Probabilistic	Deterministic
32	-14	-12
64	-12	-10
128	-11	-8
256	-9	-6
512	-8	-4
1024	-6	-2

4 Numerical experiments

We consider linear systems $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a symmetric positive definite matrix, to conduct numerical experiments. The matrix $\mathbf{A}$ is constructed as $\mathbf{A} = \mathbf{U\Lambda U}^{\mathrm{T}}$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{A}$. The orthogonal matrix $\mathbf{U}$ is obtained by orthogonalizing a randomly generated matrix that follows a normal distribution.

To systematically control the condition number of $\mathbf{A}$, we manipulate its eigenvalues. Specifically, the smallest eigenvalue is set to 1, and the largest eigenvalue is set equal to the desired condition number. The remaining

eigenvalues are spaced linearly between these bounds, with the spacing calculated as:

$$\lambda_i = 1 + \frac{(i-1) \times (\text{cond} - 1)}{n-1},$$

where n is the size of the matrix, and 'cond' denote the target condition number.

After constructing $\mathbf{A}$, we apply diagonal loading, modifying it as $\hat{\mathbf{A}} = \mathbf{A} + \delta\mathbf{D}$, where δ is a scalar and $\mathbf{D}$ is a diagonal matrix. The Cholesky decomposition is then applied to $\hat{\mathbf{A}}$ to obtain the lower triangular matrix $\tilde{\mathbf{L}}$, and the estimated value $\hat{\mathbf{x}}$ is then calculated using the forward and backward substitution method for triangular matrices. The aim of these experiments is to compare the effects of two types of the diagonal loading values on the residuals $\|\hat{\mathbf{x}} - \mathbf{x}\|^2$, achieved by incrementally increasing the condition number of the matrix $\mathbf{A}$.

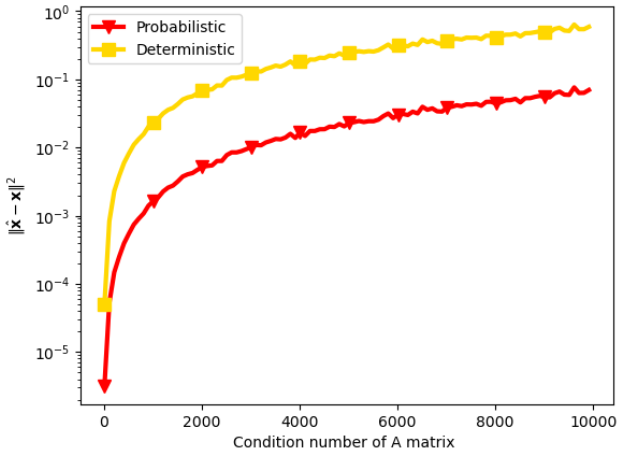

FIG. 2. Residual value comparison. Here, $N_{\text{test}} = 100, \lambda = 2$ and the size of $\mathbf{A}$ is 64.

In Figure 2 it can be seen that the utilization of the diagonal loading values determined by probabilistic rounding error analysis leads to a significantly reduced residual in the linear system. This residual is smaller than that observed when using the diagonal loading values derived from the deterministic rounding error analysis.

5 Conclusion

This paper delves into the Cholesky decomposition of symmetric positive definite matrices and its widespread application in modern digital signal processing. It analyzes the challenges that the Cholesky decomposition faces when dealing with ill-conditioned matrices and large matrix sizes in finite-precision computing environments.

By introducing probabilistic rounding error analysis, this study successfully identifies a diagonal loading value applicable to any positive definite matrix, ensuring the effective execution of the Cholesky decomposition. This discovery highlights the advantages of probabilistic methods in overcoming the limitations of traditional rounding error analysis and provides a new perspective on error handling in finite-bit-width computations.

The work also explores the transformation of the diagonal loading value into an exponent. This enables all related multiplication operations to be performed efficiently through the floating-point bitwise operations, thereby enhancing computational efficiency.

In summary, this paper combines theoretical analysis with practical applications, providing new methods and gaining insights into the handling of the key algorithms in digital signal processing.

References

[1] Z. Bai et al., *On the equivalence of MMSE and IRC receiver in MU-MIMO systems,* IEEE Commun. Lett., **15**:12 (2011), 1288–1290.

[2] V. Savaux, Y. Louët, *LMMSE channel estimation in OFDM context: a review*, IET Signal Processing, **11**:2 (2017), 123–134.

[3] A. Osinsky, R. Bychkov, M. Trefilov, V. Lyashev, A. Ivanov, *Regularization for Cholesky decomposition in massive MIMO detection*, IEEE Wireless Communications Letters, **12**:9 (2023), 1603–1607.

[4] S.H. Cheng, N.J. Higham, *A modified Cholesky algorithm based on a symmetric indefinite factorization*, SIAM J. Matrix Anal. Appl., **19**:4 (1998), 1097–1110. Zbl 0949.65022

[5] N.J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, Philadelphia, 2002. Zbl 1011.65010

[6] N.J. Higham, T. Mary, *A new approach to probabilistic rounding error analysis*, SIAM J. Sci. Comput., **41**:5 (2019), A2815–A2835. Zbl 7123205

[7] I. Kolesnikov, V. Lyashev, M. Kirichenko, *Fast algorithm for estimating singular values of Hermitian matrix*, in 31st Telecommunications Forum (TELFOR), Belgrade, Serbia, 2023, 1–4,

[8] G.H. Golub, C.F. Van Loan, *Matrix computations*, The Johns Hopkins University Press, Baltimore, 2013.

[9] J.H. Wilkinson, *A priori error analysis of algebraic processes*, in Tr. Mezhdunarod. Kongr. Mat., Moskva 1966, 629–639. Zbl 0197.13301

[10] Fukaya T, Kannan R, Nakatsukasa Y, Yamamoto Y, Yanagisawa Y., *Performance evaluation of the shifted Cholesky QR algorithm for ill-conditioned matrices*, SC18 Supercomputing, Proceedings, Poster No 69.

[11] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Stat. Assoc., **58** (1963), 13–30. Zbl 0127.10602

[12] Z. Bai, J. Demmel, A. McKenney, *On floating point errors in Cholesky*, Technical Report CS-89-87, Department of Computer Science, University of Tennessee, Knoxville, TN, USA, October 1989, LAPACK Working Note 14.

Zhibin Zhang
Moscow Institute of Physics and Technology,
9 Institutskiy Lane
Dolgoprudny City
Moskovskaya oblast
141700, Moscow, Russia
Email address: zhibin@phystech.edu

Vladimir Lyashev
Moscow Institute of Physics and Technology,
9 Institutskiy Lane
Dolgoprudny City
Moskovskaya oblast
141700, Moscow, Russia
Email address: lyashev.va@mipt.ru