Regularized Cholesky decomposition method for finite bit width computing

Zhibin Zhang[1,*] and Vladimir Lyashev[1,†]

[1]*Moscow Institute of Physics and Technology, Moscow*

This research focuses on the Cholesky decomposition of symmetric positive definite matrices. While Cholesky decomposition is known for its computational efficiency and numerical robustness, it may encounter decomposition failures when performing with 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 rounding errors, thereby guiding 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. Additionally, considering the binary representation of computers, the diagonal loading value is converted into an exponential form, allowing multiplication to be replaced by bitwise operations of floating-point numbers. This approach is both practical and efficient, effectively solving the challenges posed by ill-conditioned matrices and limited computational precision.

I. 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 which is 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 contexts, the weight matrix $\mathbf{W}$ based on Wiener filtering de-

*Electronic address: zhibin@phystech.edu

†Electronic address: lyashev.va@mipt.ru

pends fundamentally on the inverse of the covariance matrix. Since the covariance matrix is a symmetric positive definite matrix, using Cholesky decomposition combined with the back-substitution method for triangular matrices to find the inverse is a method with the lowest computational complexity and highest numerical stability.

Cholesky decomposition is favored for its computational efficiency and numerical robustness, but it encounters challenges when performing ill-conditioned matrices that have very large condition numbers. The main issue is that when 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. To address these computational challenges, various algorithms have been proposed, among which the improved Cholesky decomposition [3] stands out. However, due to simple implementation, the diagonal loading method is becoming increasingly popular in practical applications. Determining the optimal loading value for this method remains a significant research challenge. In the context of low-bit-width computations, rounding errors present a pivotal challenge. Traditional determinate rounding error analysis [4] often provide overestimations of rounding error, offering limited guidance for practical implementations.

In this paper, we introduce probabilistic rounding error analysis [5] as the theoretical foundation. This approach offers a more precise rounding error estimation, allowing for a more accurate assessment of algorithmic stability and reliability, and providing a theoretical basis for the selection of diagonal loading values. The main contribution of this paper is getting a diagonal loading value that ensures any positive definite matrix can successfully completion of Cholesky decomposition.

In the context of this paper, the notation, $a, \mathbf{b}, \mathbf{C}$ denote scalars, vectors, and matrices, respectively. The $k_2(\mathbf{A}) = \|A\|_2\|A^{-1}\|_2$ is condition number of $\mathbf{A}$ matrix. The expression $|\mathbf{A}| < |\mathbf{B}|$ signifies that the absolute value of every element in matrix $\mathbf{A}$ is smaller than the absolute value of the corresponding element in matrix $\mathbf{B}$. This is an element-by-element comparison between the two matrices.

II. STATEMENT OF THE PROBLEM

Cholesky decomposition, named the French military officer and mathematician André-Louis Cholesky (1875–1918), is a commonly used method for solving the inverse of symmetric positive definite matrices. The Cholesky decomposition is a backward stable algorithm. However, when the computational precision is limited or the condition number of matrix A (where 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 implementation algorithms of Cholesky decomposition [6]. Algorithm 1 is the most common floating-point implementation algorithm for Cholesky decomposition, which is based on gaxpy computations and is convenient for deployment on vector processors.

If this algorithm is applied to an ill-conditioned matrix, due to the accumulation of rounding errors, line 9 might attempt to take the square root of a negative number $v[j]$, leading to the failure of the decomposition. Alternatively, it might result in $L[j, j] = 0$, in which case $v[j : n]$ will be divided by zero in the next iteration. Even for positive definite matrices, in floating-point arithmetic, the algorithm may fail due to these issues.

Algorithm 1 Gaxpy-based Cholesky Decomposition [6]

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

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

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

Wilkinson [7] conducted a comprehensive analysis of the computational conditions for Cholesky decomposition. Wilkinson pointed out that Cholesky decomposition can be guar-

anteed to complete when $q_n u k_2(\mathbf{A}) \leq 1$, where q_n is a small constant, and u is the unit roundoff error (dependent on machine precision). To ensure the completion of the decomposition, work [8] suggests 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 [3] 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 Cholesky decomposition, leading to increased complexity in hardware implementation.

We have proposed a method similar to that of Fukaya [8], 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 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, only a shift in floating-point arithmetic. Therefore, this method hardly adds any complexity, requiring only n addition operations to ensure the execution of Cholesky decomposition.

III. DIAGONAL LOADING FOR CHOLESKY DECOMPOSITION

In this section, we will analyze the rounding errors in Cholesky decomposition based on the probabilistic roundoff error model. We will further investigate the relationship between the error in Cholesky decomposition and the diagonal elements of matrix $\mathbf{A}$. Subsequently, building on these two theorems, we will derive a formula to calculate the optimal diagonal loading value, δ.

A. Error analysis of Cholesky decomposition

In numerical linear algebra, traditional rounding error analysis provides deterministic backward error bounds, depended on $\gamma_n = nu/(1-nu)$, where n is the size of the computation, and u is the unit roundoff error. This type of roundoff error analysis offers an important framework for understanding and assessing the accumulation of errors during computations.

However, in low-precision computations, these deterministic backward error bounds can not provide useful information. Higham [5] developed a new probabilistic rounding error analysis method. This method uses concentration inequalities [9] and makes probabilistic estimation about rounding errors. Research have shown that the inner product error is roughly $\sqrt{n}u$, which aligns with simulation results.

A pioneering contribution by Higham [5] was the introduction of two critical expressions in the realm of probabilistic rounding error analysis. The first expression,

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

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 that rounding errors exceed a specified threshold in computational tasks. In these expressions, λ acts as a tuning parameter, often referred to 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* $\tilde{L}$ *satisfies*

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

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. For the proof, see Theorem 3.8 in Higham (2019) [5]. □

However, this probability model may not be perfect in some cases, as it could 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.

Table I: 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 Table II. This indicates that under a higher relaxation constant λ, the probability of the error exceeding the threshold in the computation process decreases.

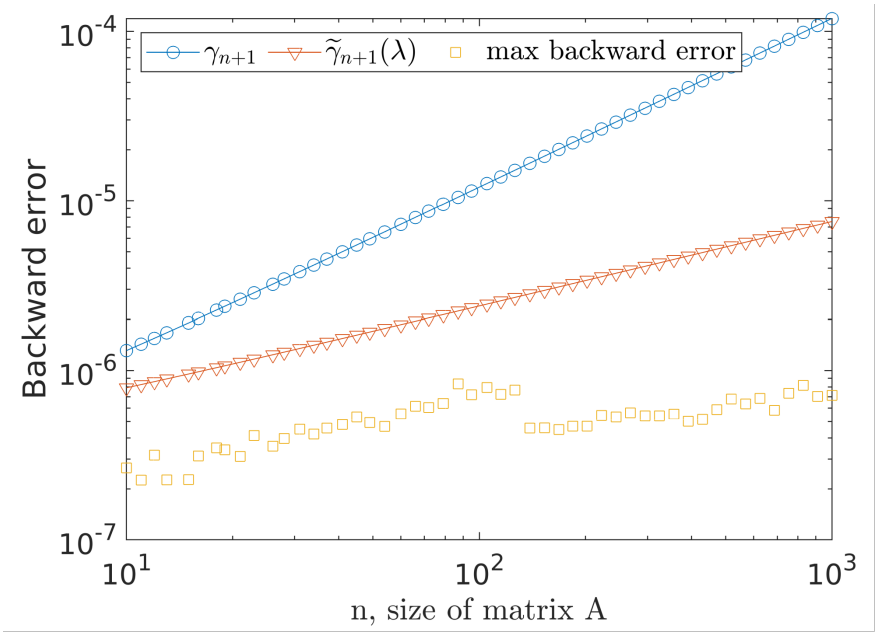

Figure 1: Backward error and its bounds in Cholesky decomposition in single precision. Here, $N_{test} = 100$ and $\lambda = 2$.

As can be seen from Figure 2, 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 batter than the deterministic error analysis method, it is still conservative. Therefor 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 Cholesky decomposition is applied to $\mathbf{A}$ and runs to completion. Then the computed Cholesky factor $\tilde{L}$ satisfies the following condition:*

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

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

Then, we have

$$\|\tilde{l}_i\|_2^2 = \tilde{l}_i\tilde{l}_i^T = a_{ii} + \Delta a_{ii} \leq a_{ii} + \widetilde{\gamma}_{n+1}|\tilde{l}_i||\tilde{l}_i|^T, \tag{5}$$

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

$$|\tilde{l}_j\tilde{l}_i^T| \leq \|\tilde{l}_i\|_2\|\tilde{l}_j\|_2 \leq (1 - \widetilde{\gamma}_{n+1})^{-1}(a_{ii}a_{jj})^{1/2}, \tag{6}$$

leading to

$$|\tilde{L}||\tilde{L}|^T \leq (1 - \widetilde{\gamma}_{n+1})^{-1}dd^T, \tag{7}$$

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

This theorem was originally observed and proven by Demmel [10]. 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 next subsection, we will apply this theorem to derive the expression for calculating the optimal diagonal loading value. By analyzing the size of matrix $\mathbf{A}$ and rounding errors, we can determine an optimal value for diagonal loading.

B. Optimal diagonal loading value

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

$$\delta > n\widetilde{\gamma}_{n+1}/(1-\widetilde{\gamma}_{n+1}) \tag{8}$$

, then 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. Assuming the algorithm has successfully completed $k-1$ stages, yielding a nonsingular $\tilde{L}_{k-1}$. In the k-th step, the following situation might be encountered:

$$\hat{\mathbf{A}}_k = \begin{bmatrix} \hat{\mathbf{A}}_{k-1} & a \\ a & b \end{bmatrix} = \begin{bmatrix} \tilde{L}_{k-1} & 0 \\ l & \sqrt{b-ll} \end{bmatrix} \begin{bmatrix} \tilde{L}_{k-1}^T & l \\ 0 & \sqrt{b-ll} \end{bmatrix} = \tilde{L}_k\tilde{L}_k^T \tag{9}$$

$b - ll < 0$, then performing the square root operation will yield an imaginary number. However, even in the latter case, the error analysis presented in Theorem 2 remains valid. This results in obtaining $\tilde{L}_k$ satisfying

$$\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k = \tilde{L}_k\tilde{L}_k^T, \quad |\Delta\hat{\mathbf{A}}_k| \le (1-\widetilde{\gamma}_{k+1})^{-1}\widetilde{\gamma}_{k+1}\sqrt{d_k}\sqrt{d_k}^T \tag{10}$$

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

$$\begin{aligned} \lambda_{\min}(\mathbf{D}_k^{-\frac{1}{2}}(\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k)\mathbf{D}_k^{-\frac{1}{2}}) &= \lambda_{\min}(\mathbf{D}_k^{-\frac{1}{2}}(\mathbf{A}_k + \delta\mathbf{D}_k + \Delta\hat{\mathbf{A}}_k)\mathbf{D}_k^{-\frac{1}{2}}) \\ &= \lambda_{\min}(\mathbf{H}_k + \delta + \mathbf{D}_k^{-\frac{1}{2}}\Delta\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}) \\ &\ge \lambda_{\min}(\mathbf{H}_k) + \delta - \|\mathbf{D}_k^{-\frac{1}{2}}\Delta\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\|_2 \\ &\ge \lambda_{\min}(\mathbf{H}_k) + \delta - \frac{\widetilde{\gamma}_{k+1}}{1-\widetilde{\gamma}_{k+1}}\|\mathbf{1}_k\|_2 \\ &\ge \lambda_{\min}(\mathbf{H}_k) + \delta - \frac{k\widetilde{\gamma}_{k+1}}{1-\widetilde{\gamma}_{k+1}} > 0. \end{aligned} \tag{11}$$

Here, $\mathbf{H}_k$ 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. Hence $\mathbf{D}_k^{-1}(\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k)\mathbf{D}_k^{-1}$ is positive definite, and therefore so is the congruent matrix $\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k$, showing that $\tilde{L}_k$ must be real and non-singular one.

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. Consequently, the congruent matrix $\hat{\mathbf{A}}_k + \Delta\hat{\mathbf{A}}_k$ is also positive definite. This result is significant as it implies that $\tilde{L}_k$ is necessarily real and non-singular.

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

This theorem naturally links the diagonal loading value to the diagonal elements of the matrix **A**. Higham [5] and Demmel [10] have previously proven similar theorems. This theorem is a diagonal loading variant of their theorems and integrates probabilistic rounding error analysis.

According to equation 8, we have derived a diagonal loading value that depends only on the unit roundoff error precision and the dimensions of the matrix. This value is very small and does not require any additional computation of the norm of matrix **A**. By transforming this value with the logarithmic function $\log_2$, we make it easier to handle and adapt it to the binary representation in computers. Finally, by using the fix function fix, we 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. In this way, all related multiplications can be efficiently performed through bitwise operations of floating-point numbers, thereby enhancing computational efficiency.

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

$$\delta(\lambda, n, u) = \mathrm{fix}\left(\log_2\left(\frac{n\sqrt{n}u\lambda}{1-\sqrt{n}u\lambda}\right)\right) \tag{12}$$

It can be seen that this expression perfectly incorporates both matrix dimensions and rounding errors. We believe 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 2. Its most significant contribution is the demonstration that rounding errors do not increase linearly with the growth in the dimensions of a matrix. Through extensive numerical analysis and simulations, we have set the parameter λ to a value of 2. When $\lambda = 2$, occurrences exceeding the error bounds do not arise.

Furthermore, based on equation 12, we have calculated diagonal loading values for matrices of varying dimensions. These values were compared with diagonal loading values from deterministic rounding error analysis. The comparison reveals that our diagonal loading values are smaller, thereby minimizing bias in the computational results while still ensuring the smooth execution of the Cholesky decomposition.

Table II: 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

IV. NUMERICAL EXPERIMENTS

We consider linear systems $\mathbf{Ax} = \mathbf{b}$ with symmetric positive definite $\mathbf{A}$ to conduct numerical experiments. Firstly, diagonal loading value is applied to matrix $\mathbf{A}$, resulting in $\hat{\mathbf{A}} = \mathbf{A} + \delta\mathbf{D}$. Subsequently, the $\mathbf{L}$ matrix is obtained through Cholesky decomposition, and the estimated value $\hat{\mathbf{x}}$ is then calculated using forward and backward substitution method for triangular matrices. The objective of these experiments is to compare the effects of two types of diagonal loading values on the residuals, achieved by incrementally increasing the condition number of the matrix $\mathbf{A}$.

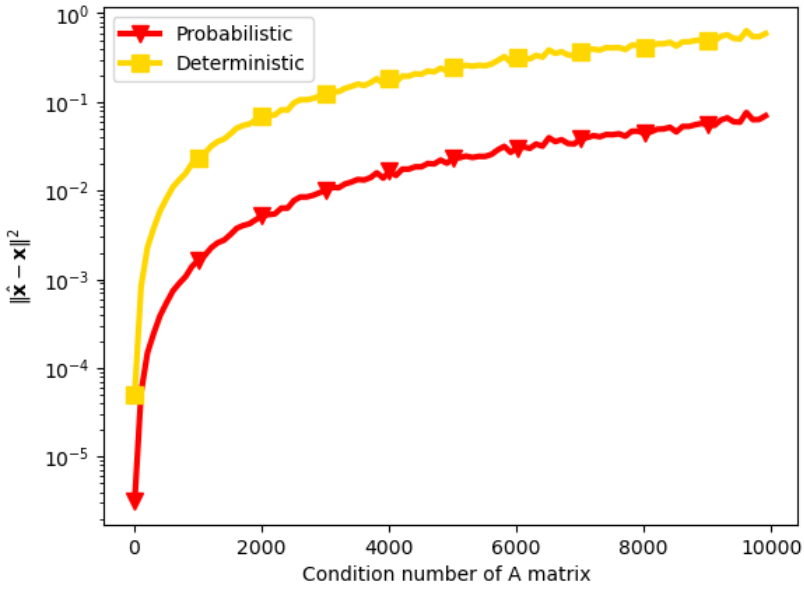

Figure 2: Residual value comparison. Here, $N_{\text{test}} = 100, \lambda = 2$ and dimension of $\mathbf{A}$ is 64.

The results of the numerical experiments demonstrate that the utilization of diagonal loading values determined by probabilistic rounding error analysis leads to a markedly reduced residual in linear system. This residual is smaller than that observed when employing diagonal loading values derived from deterministic rounding error analysis.

V. 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 encountered by Cholesky decomposition when dealing with ill-conditioned matrices, large dimension of matrix and 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 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 efficiently completed through bitwise operations of floating-point numbers, thereby enhancing computational efficiency.

In summary, this paper combines theoretical analysis with practical applications, providing new methods and insights for processing key algorithms in digital signal processing.

VI. REFERENCES

[1] *Z. Bai et al.,* "On the equivalence of MMSE and IRC receiver in MU-MIMO systems," IEEE Commun. Lett., vol. 15, no. 12, pp. 1288–1290, Dec. 2011.

[2] *Savaux, Vincent, and Yves Louët.* "LMMSE channel estimation in OFDM context: a review." IET Signal Processing 11.2 (2017): 123-134.

[3] *Cheng, Sheung Hun, and Nicholas J. Higham.* A modified Cholesky algorithm based on a symmetric indefinite factorization. SIAM Journal on Matrix Analysis and Applications 19.4 (1998): 1097-1110.

[4] *Higham, Nicholas J.* Accuracy and stability of numerical algorithms. Society for industrial and applied mathematics, 2002.

[5] *Higham, Nicholas J., and Theo Mary.* A new approach to probabilistic rounding error analysis. SIAM journal on scientific computing 41.5 (2019): A2815-A2835.

[6] *Golub G H, Van Loan C F.* Matrix computations[M]. JHU press, 2013.

[7] *Wilkinson J H.* A priori error analysis of algebraic processes[C]//Intern. Congress Math. 1968, 19: 629-639.

[8] *Fukaya T, Kannan R, Nakatsukasa Y, Yamamoto Y, Yanagisawa Y.* Performance evaluation of the shifted Cholesky QR algorithm for ill-conditioned matrices.

[9] *W. Hoeffding.* Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30, https://doi.org/10.1080/01621459.1963.10500830.

[10] *James W. Demmel.* On floating point errors in Cholesky. Technical Report CS-89-87, Department of Computer Science, University of Tennessee, Knoxville, TN, USA, October 1989. 6 pp. LAPACK Working Note 14.