Solving Block Low-Rank Linear Systems by LU Factorization is Numerically Stable

Higham, Nicholas J. and Mary, Theo

2019

MIMS EPrint: **2019.15**

ISSN 1749-9097

SOLVING BLOCK LOW-RANK LINEAR SYSTEMS BY LU FACTORIZATION IS NUMERICALLY STABLE[*]

NICHOLAS J. HIGHAM[†] AND THEO MARY[‡]

Abstract. Block low-rank (BLR) matrices possess a blockwise low-rank property that can be exploited to reduce the complexity of numerical linear algebra algorithms. The impact of these low-rank approximations on the numerical stability of the algorithms in floating-point arithmetic has not previously been analyzed. We present rounding error analysis for the solution of a linear system by LU factorization of BLR matrices. Assuming that a stable pivoting scheme is used, we prove backward stability: the relative backward error is bounded by a modest constant times ε, where the low-rank threshold ε is the parameter controlling the accuracy of the blockwise low-rank approximations. In addition to this key result, our analysis offers three new insights into the numerical behavior of BLR algorithms. First, we compare the use of a global or local low-rank threshold and find that a global one should be preferred. Second, we show that performing intermediate recompressions during the factorization can significantly reduce its cost without compromising numerical stability. Third, we consider different BLR factorization variants and determine the compress–factor–update (CFU) variant to be the best. Tests on a wide range of matrices from various real-life applications show that the predictions from the analysis are realized in practice.

Key words. Block low-rank matrices, rounding error analysis, floating-point arithmetic, numerical linear algebra.

AMS subject classifications. 65G50, 65Fxx

1. Introduction.

In many applications requiring the solution of a linear system

$$Ax = v, \tag{1.1}$$

the matrix A has been shown to have a blockwise low-rank property: most of its off-diagonal blocks are of low numerical rank and can therefore be well approximated by low-rank (LR) matrices.

Several formats have been proposed to exploit this property, differing in how the matrix is partitioned into blocks. In this article, we focus on the block low-rank (BLR) format [3], which is based on a flat, non-hierarchical partitioning allowing it to reduce both the theoretical complexity [4] and the practical time and memory costs [5] of key numerical linear algebra computations such as solving (1.1) by LU factorization.

Even though the BLR format has been extensively studied and widely used in numerous applications [1,2,3,4,5,7,12,23,25,27,28,30,31,33], little is known about its numerical behavior in floating-point arithmetic. Indeed, no rounding error analysis has been published for BLR matrix algorithms. The difficulty of such an analysis lies in the fact that, compared with classical algorithms without low-rank approximations, there are two kinds of errors to analyze: the floating-point errors (which depend on the unit roundoff u) and the low-rank truncation errors (which depend on the low-rank threshold $\varepsilon > u$). Yet performing such an analysis is crucial to better understand the effect of BLR approximations on the accuracy and stability of these algorithms and, in particular, to shed light on the following four open problems.

- It has been experimentally observed that the solution to BLR linear systems generally yields a backward error of the same order as the low-rank threshold

[*]Version of September 4, 2019.

[†]Department of Mathematics, The University of Manchester, Manchester, M13 9PL, UK (nick.higham@manchester.ac.uk, http://www.maths.manchester.ac.uk/~higham)

[‡]Department of Mathematics, The University of Manchester, Manchester M13 9PL, UK (theo.mary@manchester.ac.uk)

ε. This is an important and valuable property that has, however, never been formally proved. The dependence of the backward error on the unit roundoff should also be investigated.

- In contrast to hierarchical matrices, the number of block-rows and block-columns in BLR matrices usually grows with the matrix size and may thus become very large. It is therefore important to determine how the error grows as the matrix size increases and whether it depends on the number of blocks.
- The low-rank approximation $\widetilde{A}_{ij}$ to a block A_{ij} is computed such that $\|A_{ij} - \widetilde{A}_{ij}\| \le \varepsilon\beta_{ij}$, where β_{ij} is a scalar that can be freely chosen and whose impact on the numerical behavior of the algorithms is currently not well understood. In particular, local ($\beta_{ij} = \|A_{ij}\|$) and global ($\beta_{ij} = \|A\|$) low-rank thresholds have both been proposed in the literature and should be compared.
- Several BLR LU factorization algorithms can be distinguished, depending on when the compression is performed. These algorithms have been compared in terms of asymptotic complexity, performance, and storage requirements [25]. However, it is not currently known how they compare in terms of numerical stability.

In this article, we develop and present rounding error analyses of various BLR matrix algorithms to answer these questions. We begin in section 2 with the preliminary material necessary for the analysis. Section 3 analyzes several kernels involving LR matrices, such as LR matrix–vector and matrix–matrix products. Then, section 4 builds upon these results to analyze several key BLR matrix algorithms, such as BLR matrix–vector products, BLR LU factorizations, and the solution to BLR linear systems. Throughout the article, numerical experiments are interlaced with theoretical results to illustrate them. We provide additional experiments on a wide range of matrices coming from various real-life applications in section 5. We gather our conclusions in section 6.

2. Technical background and experimental setting.

2.1. Low-rank (LR) and block low-rank (BLR) matrices. Let $A \in \mathbb{R}^{b\times b}$ have the SVD $U\Sigma V^T$, where $\Sigma = \mathrm{diag}(\sigma_i)$ with $\sigma_1 \ge \cdots \ge \sigma_b \ge 0$. Given a target rank $k \le b$, the quantity $\|A - \widetilde{A}\|$ for any rank-k matrix $\widetilde{A}$ is known to be minimized for any unitarily invariant norm by the truncated SVD

$$\widetilde{A} = U_{:,1:k}\Sigma_{1:k,1:k}V_{:,1:k}^T. \tag{2.1}$$

If the singular values of A decay rapidly, $\|A - \widetilde{A}\|$ can be small even for small $k \ll b$, where the meaning of "small" is specified below. In this case, $\widetilde{A}$ is referred to as a low-rank (LR) matrix, and the cost of storing and computing on $\widetilde{A}$ can be greatly reduced. While $\widetilde{A}$ can directly be represented by the truncated SVD (2.1), in this article we use the alternative form $\widetilde{A} = XY^T$, where $X = U_{:,1:k}$ and $Y = V_{:,1:k}\Sigma_{1:k,1:k}^T$; matrix X can thus be assumed to have orthonormal columns.

A block low-rank (BLR) representation $\widetilde{A}$ of a dense matrix A has the block $p \times p$ form

$$\widetilde{A} = \begin{bmatrix} A_{11} & \widetilde{A}_{12} & \cdots & \widetilde{A}_{1p} \\ \widetilde{A}_{21} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \widetilde{A}_{p1} & \cdots & \cdots & A_{pp} \end{bmatrix}, \tag{2.2}$$

where off-diagonal blocks A_{ij} of size $m_i \times n_j$ are approximated by LR matrices $\widetilde{A}_{ij} = X_{ij} Y_{ij}^T$ of rank k_{ij}, where $X_{ij} \in \mathbb{R}^{m_i \times k_{ij}}$ and $Y_{ij} \in \mathbb{R}^{n_j \times k_{ij}}$.

We assume that the ranks k_{ij} are chosen as

$$k_{ij} = \min \left\{ \ell_{ij} : \|A_{ij} - \widetilde{A}_{ij}\| \le \varepsilon \beta_{ij},\ \operatorname{rank}\big(\widetilde{A}_{ij}\big) = \ell_{ij} \right\}, \tag{2.3}$$

where $\varepsilon > 0$ is referred to as the *low-rank threshold* and controls the accuracy of the approximations $\widetilde{A}_{ij} \approx A_{ij}$, and where we have either $\beta_{ij} = \|A_{ij}\|$ or $\beta_{ij} = \|A\|$. The parameters β_{ij} therefore control whether the blocks are approximated relative to their own norm $\|A_{ij}\|$ or the norm of the global matrix $\|A\|$. We refer to the low-rank threshold as *local* in the former case and as *global* in the latter case.

With a local threshold, k_{ij} corresponds to the usual definition of numerical rank. Importantly, with a global threshold, blocks that do not have rapidly decaying singular values may still be approximated by LR matrices if they are of small norm compared to $\|A\|$. Indeed, even though such blocks have high numerical rank relative to their own norm, their contribution to the global computation can be considered to be negligible compared with other blocks of larger norm. In the most extreme cases, some blocks may be of norm smaller than ε: these blocks can then approximated by *zero-rank* matrices, that is, they may be dropped entirely. Exploiting this fact can drastically improve the compression, sometimes even leading to an improved asymptotic complexity [4].

Throughout this article, we denote by r the largest of the ranks k_{ij}, and we assume for simplicity that all blocks are of the same dimensions $b \times b$, and so $n = pb$.

Note that in general, low-rank approximations only hold in a *normwise* sense: that is, given a matrix A and a LR matrix $\widetilde{A}$ satisfying $\|\widetilde{A} - A\| \le \varepsilon \|A\|$, the componentwise inequality $|\widetilde{A} - A| \le \varepsilon |A|$ does not hold. For this reason we perform a normwise error analysis.

Throughout this article, the unsubscripted norm $\|\cdot\|$ denotes the Frobenius norm

$$\|A\| = \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2},$$

which we use for all our error analysis. We choose to work with the Frobenius norm for three of its useful properties. First, it is submultiplicative (also called consistent): $\|AB\| \le \|A\|\|B\|$. Second, it is invariant under multiplication on the left by a matrix with orthonormal columns X: $\|XA\| = \|A\|$ (note that X is not necessarily unitary: $X^T X = I$ must hold but $XX^T = I$ need not). Finally, unlike for the spectral norm, it is easy to switch between blockwise and global bounds using the relation

$$\|A\| = \left(\sum_{i,j} \|A_{ij}\|^2 \right)^{1/2}$$

for any block partitioning of A. We will use all these properties of the Frobenius norm without comment. More specific examples of why the Frobenius norm is the best choice for our analysis are given throughout sections 3 and 4.

2.2. Floating-point arithmetic and rounding error analysis. Throughout the article, we use the standard model of floating-point arithmetic [19, sec. 2.2]

$$\mathrm{fl}(x \text{ op } y) = (x \text{ op } y)(1 + \delta), \quad |\delta| \le u, \quad \text{op} \in \{+, -, \times, /\}. \tag{2.4}$$

We also define

$$\gamma_k = \frac{ku}{1-ku}$$

for $ku < 1$. We will use without comment the relations $j\gamma_k \le \gamma_{jk}$ and $\gamma_j + \gamma_k \le \gamma_{j+k}$ [19, Lem. 3.3], which hold for any $j, k \ge 1$ (including non-integer j, k, which we will sometimes use).

We recall normwise error bounds for some basic matrix computations.

Lemma 2.1 (Error bounds for matrix–vector and matrix–matrix products). *Let $A \in \mathbb{R}^{a\times b}$, $v \in \mathbb{R}^b$, and $w = Av$. The computed $\widehat{w}$ satisfies*

$$\widehat{w} = (A + \Delta A)v, \quad \|\Delta A\| \le \gamma_b \|A\|. \tag{2.5}$$

Let $B \in \mathbb{R}^{b\times c}$ and let $C = AB$. The computed $\widehat{C}$ satisfies

$$\widehat{C} = AB + \Delta C, \quad \|\Delta C\| \le \gamma_b \|A\|\|B\|. \tag{2.6}$$

Proof. See [19, p. 71]. □

Lemma 2.2 (Backward error bound for triangular systems). *Let $T \in \mathbb{R}^{b\times b}$ be nonsingular and triangular and let $v \in \mathbb{R}^b$. The computed solution $\widehat{x}$ to the triangular system $Tx = v$ satisfies*

$$(T + \Delta T)\widehat{x} = v, \quad \|\Delta T\| \le \gamma_b \|T\|. \tag{2.7}$$

The computed solution $\widehat{X}$ to the multiple right-hand side triangular system $TX = V$, where $V \in \mathbb{R}^{b\times c}$, satisfies

$$T\widehat{X} = B + \Delta B, \quad \|\Delta B\| \le \gamma_b \|T\|\|\widehat{X}\|. \tag{2.8}$$

Proof. See [19, Thm. 8.5]. □

Lemma 2.3 (Backward error bound for LU factorization). *If the LU factorization of $A \in \mathbb{R}^{b\times b}$ runs to completion then the computed LU factors $\widehat{L}$ and $\widehat{U}$ satisfy*

$$\widehat{L}\widehat{U} = A + \Delta A, \quad \|\Delta A\| \le \gamma_b \|\widehat{L}\|\|\widehat{U}\|. \tag{2.9}$$

Proof. See [19, Thm. 9.3]. □

Finally, we make the assumption that rounding errors can be ignored in the computation of the LR approximation $\widetilde{A}_{ij}$ of any block A_{ij} via its truncated SVD.

Assumption 2.4 (Error bound for the truncated SVD computation). *Given a block $A_{ij} \in \mathbb{R}^{b\times b}$ and two positive parameters ε (the low-rank threshold) and β_{ij}, the LR block $\widetilde{A}_{ij}$ computed via the truncated SVD $\widetilde{A}_{ij} = \widehat{X}_{:,1:k_{ij}} \widehat{\Sigma}_{1:k_{ij},1:k_{ij}} \widehat{Y}^T_{:,1:k_{ij}}$ satisfies*

$$\widetilde{A}_{ij} = A_{ij} + \Delta A_{ij}, \quad \|\Delta A_{ij}\| \le \varepsilon\beta_{ij}.$$

We recall that $\beta_{ij} = \|A_{ij}\|$ or $\beta_{ij} = \|A\|$ controls whether we use a local or global threshold, as explained in the previous section.

Note that Assumption 2.4 is only satisfied if the unit roundoff u is safely smaller than the low-rank threshold ε, which we assume to be the case throughout the analysis.

2.3. Experimental setting. All numerical experiments reported in this article have been performed with MATLAB R2018b. Unless otherwise specified, we use IEEE double precision floating-point arithmetic. In some experiments we also use IEEE single and half precisions. The use of half precision has been simulated as described in [21]. We have made all our codes used for the experiments available online[1].

One of the most common application domains where block low-rank matrices arise is the solution of discretized partial differential equations. These matrices are then sparse. In our analysis we however consider dense matrices, which serve as building blocks for sparse direct solvers. Therefore, in our experiments, we use a set of dense matrices obtained from the Schur complements of sparse matrices: these correspond to the root separator in the context of a nested dissection [14] solver.

In section 4, we illustrate our analysis by interlacing it with experiments on matrices coming from a Poisson problem

$$-\Delta u = f, \tag{2.10}$$

discretized with a 7-point finite-difference scheme on a 3D domain of dimensions $k \times k \times k$. This leads to a dense $k^2 \times k^2$ matrix. We test variable sizes (from $n = 32^2$ to $n = 128^2$) to explore how the error behaves as $n = k^2$ increases. In section 5, we complement these tests with some experiments on matrices from the SuiteSparse collection [13] coming from various real-life applications.

Throughout sections 4.2 and 4.3 we present experiments on the BLR LU factorization. Instead of measuring the backward error for LU factorization $\|A - \widetilde{L}\widetilde{U}\|/\|A\|$, which is expensive to compute, we solve a linear system $Ax = v$ by forward and backward substitutions with the BLR LU factors (as described in section 4.4), where x is the vector of all ones. We then measure the backward error

$$\frac{\|A\widehat{x} - v\|}{\|A\|\|\widehat{x}\| + \|v\|} \tag{2.11}$$

of the computed $\widehat{x}$, which is much cheaper to compute.

For all experiments, the block size is set to $b = 256$ unless otherwise specified.

3. Rounding error analysis of LR matrix algorithms.

3.1. LR matrix times vector or full matrix. We begin by analyzing the product of an LR matrix $\widetilde{A}$ with a vector v, for which we can establish a backward error bound. We then generalize the analysis to the product of $\widetilde{A}$ with a full matrix B, for which only a forward error bound can be derived. This key kernel is heavily used in the BLR matrix algorithms considered in the subsequent sections, and its analysis gives some insight into the behavior of LR matrix computations.

LEMMA 3.1 (LR matrix times vector). *Let $A \in \mathbb{R}^{b\times b}$, $X \in \mathbb{R}^{b\times r}$, $Y \in \mathbb{R}^{b\times r}$, and $v \in \mathbb{R}^b$, where X has orthonormal columns and $\widetilde{A} = XY^T$ is an LR approximation of A satisfying $\|A - \widetilde{A}\| \le \varepsilon\beta$ for some $\beta > 0$. If the matrix–vector product Av is computed as $z = X(Y^T v)$, the computed $\widehat{z}$ satisfies*

$$\widehat{z} = (\widetilde{A} + \Delta\widetilde{A})v, \quad \|\Delta\widetilde{A}\| \le \gamma_c\|\widetilde{A}\| + O(u^2), \tag{3.1}$$

[1] https://gitlab.com/theo.andreas.mary/BLRstability

where $c = b + r^{3/2}$, and therefore

$$\widehat{z} = (A + \Delta A)v, \quad \|\Delta A\| \le \gamma_c\|A\| + \varepsilon(1+\gamma_c)\beta + O(u^2) \qquad (3.2)$$
$$= \gamma_c\|A\| + \varepsilon\beta + O(u\varepsilon).$$

Proof. Let $w = Y^T v$; the computed $\widehat{w}$ satisfies

$$\widehat{w} = (Y + \Delta Y)^T v, \quad \|\Delta Y\| \le \gamma_b\|Y\| = \gamma_b\|\widetilde{A}\|. \qquad (3.3)$$

Let $z = X\widehat{w}$; the computed $\widehat{z}$ satisfies

$$\widehat{z} = (X + \Delta X)\widehat{w}, \quad \|\Delta X\| \le \gamma_r\|X\| = \gamma_r\sqrt{r}. \qquad (3.4)$$

Combining (3.3) and (3.4), we obtain $\widehat{z} = (X + \Delta X)(Y + \Delta Y)^T v$, which yields (3.1). The bound (3.2) is obtained by replacing $\widetilde{A}$ by $A + E$, where $\|E\| \le \varepsilon\beta$. ∎

Note that for the particular choice $\beta = \|A\|$, the bound (3.2) simplifies to

$$\widehat{z} = (A + \Delta A)v, \quad \|\Delta A\| \le \big(\varepsilon + \gamma_c + \varepsilon\gamma_c\big)\|A\| + O(u^2).$$

This is a backward error bound, from which the forward error bound

$$\|\widehat{z} - Av\| \le \big(\varepsilon + \gamma_c + \varepsilon\gamma_c\big)\|A\|\|v\| + O(u^2)$$

trivially follows.

Before commenting on its significance, we immediately generalize this result to the case where v is a full matrix rather than a vector, in which case only a forward error bound can be obtained.

LEMMA 3.2 (LR matrix times full matrix). *Let $A \in \mathbb{R}^{b\times b}$ and $V \in \mathbb{R}^{b\times m}$, and let $\widetilde{A} = XY^T$ be defined as in Lemma 3.1. If the product AV is computed as $Z = X(Y^T V)$, the computed $\widehat{Z}$ satisfies*

$$\|\widehat{Z} - \widetilde{A}V\| \le \gamma_c\|\widetilde{A}\|\|V\| + O(u^2) \qquad (3.5)$$

and therefore

$$\|\widehat{Z} - AV\| \le \gamma_c\|A\|\|V\| + \varepsilon(1+\gamma_c)\beta\|V\| + O(u^2), \qquad (3.6)$$
$$= \gamma_c\|A\|\|V\| + \varepsilon\beta\|V\| + O(u\varepsilon),$$

with $c = b + r^{3/2}$.

Proof. The result follows from the columnwise bounds

$$\|\widehat{z}_j - \widetilde{A}v_j\| \le \gamma_c\|\widetilde{A}\|\|v_j\| + O(u^2), \quad j = 1\colon m,$$

that are obtained from (3.1). ∎

For $\beta = \|A\|$, bound (3.6) simplifies to

$$\|\widehat{Z} - AV\| \le \big(\varepsilon + \gamma_c + \varepsilon\gamma_c\big)\|A\|\|V\| + O(u^2). \qquad (3.7)$$

The bounds (3.1) and (3.5) generalize the classical bounds (2.5) and (2.6) to the case where $\widetilde{A}$ is an LR matrix rather than a full one. The bounds (3.2) and (3.6) have a more informative form, as they measure not only the effect of floating-point errors

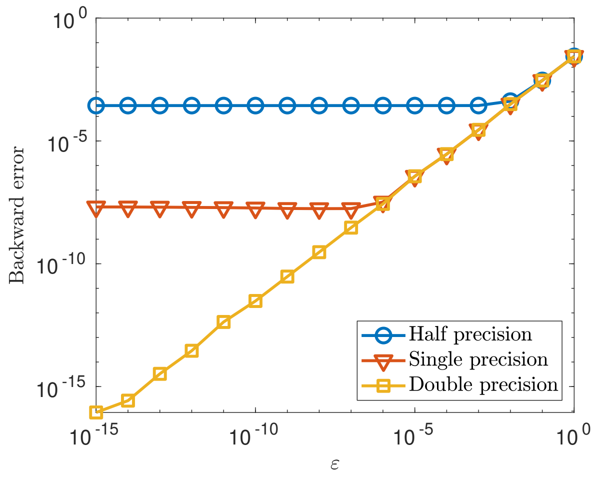


Fig. 3.1: Backward error (3.8) for computing $z = \widetilde{A}v$, with $\beta = \|A\|$, $b = 1000$, $A =$ `gallery('randsvd',b,1e16,3)`, and $v =$ `rand(b,1)`, for varying low-rank thresholds ε and floating-point precisions.

but also that of the low-rank truncation errors. They consist of three terms: the term ε, corresponding to the low-rank truncation errors; the term γ_c, corresponding to the floating-point errors; and their product $\varepsilon\gamma_c$, which reflects the fact that the two types of error accumulate, although this $O(u\varepsilon)$ term can be considered to be of lower order and will not always be explicitly tracked in the rest of the analysis below.

Since $u \ll \varepsilon$, the main conclusion of Lemma 3.1 is that the bound (3.2) is dominated by the low-rank truncation error term $\varepsilon\beta$ and is almost independent of both the unit roundoff and the constants b and r. This is a very positive result, since for $\beta = \|A\|$ we have $\|\varDelta A\| \lesssim \varepsilon\|A\|$, meaning that the computation is backward stable with respect to the precision ε. Another crucial consequence of the fact that $u \ll \varepsilon$ is that we can and should use the lowest floating-point precision such that u remains safely smaller than ε. This is illustrated by Figure 3.1, which shows that the backward error is not affected by the use of single rather than double precision arithmetic when $\varepsilon \gg 10^{-8}$. Similarly, half precision arithmetic can be used with no impact on the error when $\varepsilon \gg 10^{-4}$. In this experiment, the backward error is computed by the formula

$$\frac{\|\widehat{z} - Av\|}{\|A\|\|v\|}, \tag{3.8}$$

which is a consequence of the Rigal–Gaches theorem [19, Thm. 7.1], [29].

It is important to note that Lemma 3.1, as well as all other results in this paper, provides only a normwise error bound. This is unavoidable because low-rank approximations do not hold componentwise, as already mentioned. Interestingly, this means that algorithms that sacrifice componentwise stability for speed, such as the 3M algorithm to multiply complex matrices with only three real multiplications [18], or fast matrix algorithms such as Strassen's algorithm [32], are much more attractive when used on LR matrices, since only normwise stability can be expected anyway. Another possible drawback of Strassen's algorithm is that the constants in the error bound are much larger: for example, for matrix multiplication, the constant b in (2.6) increases

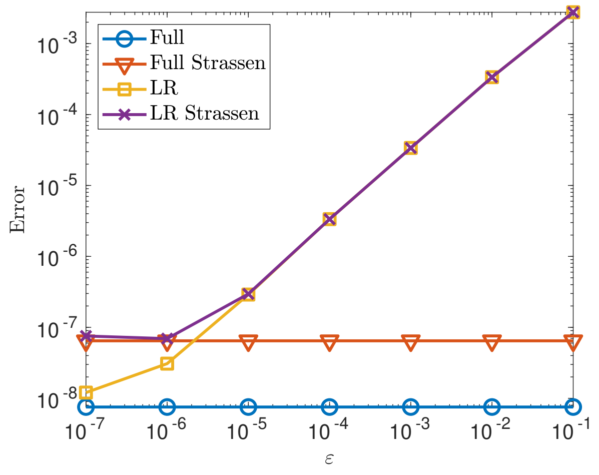


Fig. 3.2: Forward error (3.9) for computing $C = AV$ and $C = \widetilde{A}V$ in single precision, with $\beta = \|A\|$, $b = 1024$, $A =$ `gallery('randsvd',b,1e16,3)`, and $V =$ `rand(b,128)`, for varying low-rank thresholds ε, with multiplications performed classically and by Strassen's algorithm.

to $O(b^{\log_2 12}) \approx O(b^{3.6})$ [17], [19, sec. 23.2.2]. This larger constant is insignificant for LR matrices because the low-rank errors dominate the floating-point ones, and so the constant c in (3.2) has essentially no impact on the overall error. This is illustrated in Figure 3.2, where we compare the error[2]

$$\frac{\|\widehat{Z} - AV\|}{\|A\|\|V\|} \tag{3.9}$$

for computing the product of two full matrices, $Z = AV$, with that for computing the product of an LR matrix with a full one, $Z = \widetilde{A}V$, for multiplication performed classically and by Strassen's algorithm. While the use of Strassen's algorithm leads to an error larger by about an order of magnitude for the product $Z = AV$, the error for the product $Z = \widetilde{A}V$ is unaffected as long as the threshold ε is large enough to hide the larger constant of Strassen's algorithm.

Even though the constant $c = b + r^{3/2}$ has no significant impact on the overall error, further insight can nevertheless be gained by analyzing its form. First, note that c heavily depends on the choice of norm used for the analysis: the proof of Lemma 3.1 repeatedly uses the fact that the Frobenius norm is both submultiplicative and unitarily invariant, which helps us to obtain a relative small constant. With the maximum norm $\|A\|_M = \max_{i,j} |a_{ij}|$ (for example) which satisfies neither of these two properties, bound (3.5) holds with a much larger constant $c = b^{3/2}r(b + r)$. Second, it may seem surprising that γ_c is larger than γ_b, the constant in the error bound (2.5) for the classical matrix–vector product, since computing $\widetilde{A}v$ rather than Av reduces the number of flops from $2b^2$ to $4br$. However, not all flops are equal, as is clearly illustrated by the case of inner and outer products x^Ty and xy^T of two vectors $x, y \in \mathbb{R}^n$, which require $O(n)$ and $O(n^2)$ flops, but yield error bounds proportional

[2]Note that the quantity (3.9) is not a backward error; it can be interpreted as a combination of the columnwise backward errors (3.8) for each column v_i of V.

to γ_n and γ_1, respectively. In the case of Lemma 3.1, b rounding errors combine in the first product $w = Y^T v$, but r additional errors combine in the second product Xw. The extra $\sqrt{r}$ factor comes from the term $\|X\|$, and could thus be avoided by using the 2-norm rather than the Frobenius norm; however we need the Frobenius norm when working on BLR matrices, as will become clear in section 4.

We also note the importance of assuming that X has orthonormal columns, as it allows for replacing $\|Y\|$ by $\|\widetilde{A}\|$ in (3.3). Without that assumption, the bound would be proportional to $\|X\|\|Y\|$ instead of $\|\widetilde{A}\|$, reflecting the fact that the computation $X(Y^T v)$ would be subject to possibly severe cancellation when $\|X\|\|Y\| \gg \|\widetilde{A}\|$. Note that assuming the columns of Y, rather than X, to be orthonormal would lead to the same bound. We also mention that using the alternative form $\widetilde{A} = X\Sigma Y^T$, with $\Sigma \in \mathbb{R}^{r\times r}$ and where both X and Y have orthonormal columns would not significantly change the bound: we simply would need to account for the extra product with Σ, slightly increasing the constant c.

3.2. LR matrix times LR matrix. Next we analyze the product of two LR matrices $\widetilde{A} = X_A Y_A^T$ and $\widetilde{B} = Y_B X_B^T$, where X_A and X_B have orthonormal columns. Note that we consider $\widetilde{B}$ of the form $Y_B X_B^T$ rather than $X_B Y_B^T$, for reasons which will be made clear below.

LEMMA 3.3 (LR matrix times LR matrix). *Let $A, B \in \mathbb{R}^{b\times b}$ and*

$$\widetilde{A} = \underbrace{X_A}_{b\times r}\underbrace{Y_A^T}_{r\times b}, \quad \widetilde{B} = \underbrace{Y_B}_{b\times r}\underbrace{X_B^T}_{r\times b},$$

where X_A and X_B have orthonormal columns and

$$\|A - \widetilde{A}\| \le \varepsilon\beta_A, \quad \|B - \widetilde{B}\| \le \varepsilon\beta_B.$$

If the product $C = \widetilde{A}\widetilde{B}$ is computed as either $C = (X_A(Y_A^T Y_B))X_B^T$ or $C = X_A((Y_A^T Y_B)X_B^T)$, where the parentheses indicate the order in which the intermediate products are performed, then the computed $\widehat{C}$ satisfies

$$\|\widehat{C} - \widetilde{A}\widetilde{B}\| \le \gamma_c\|\widetilde{A}\|\|\widetilde{B}\| + O(u^2), \tag{3.10}$$

where $c = b + 2r^{3/2}$, and therefore

$$\|\widehat{C} - AB\| \le \gamma_c\|A\|\|B\| + \varepsilon(1+\gamma_c)\big(\beta_A\|B\| + \|A\|\beta_B + \varepsilon\beta_A\beta_B\big) + O(u^2). \tag{3.11}$$

Proof. We consider the case where the product is computed as $(\widetilde{A}Y_B)X_B^T = (X_A(Y_A^T Y_B))X_B^T$, the other case being analogous. Let $W = \widetilde{A}Y_B$; by Lemma 3.2, the computed W satisfies

$$\widehat{W} = \widetilde{A}Y_B + \Delta W, \quad \|\Delta W\| \le \gamma_{b+r^{3/2}}\|\widetilde{A}\|\|Y_B\| + O(u^2) = \gamma_{b+r^{3/2}}\|\widetilde{A}\|\|\widetilde{B}\| + O(u^2).$$

Let $C = \widehat{W}X_B^T$; the computed $\widehat{C}$ satisfies

$$\begin{aligned}\widehat{C} &= \widehat{W}X_B^T + \Delta C, \quad \|\Delta C\| \le \gamma_r\|\widehat{W}\|\|X_B\| \le \gamma_{r^{3/2}}\|\widetilde{A}\|\|\widetilde{B}\| + O(u^2),\\ &= \widetilde{A}\widetilde{B} + \Delta W X_B^T + \Delta C = \widetilde{A}\widetilde{B} + F.\end{aligned}$$

Bounding $\|F\| \le \gamma_c\|\widetilde{A}\|\|\widetilde{B}\|$, with $c = b + 2r^{3/2}$, proves (3.10). We then replace $\widetilde{A}$ by $A + E_A$ and $\widetilde{B}$ by $B + E_B$ to obtain

$$\begin{aligned}\widehat{C} = AB + F + G, \quad &\|F\| \le \gamma_c\|A\|\|B\| + \gamma_c\varepsilon\big(\beta_A\|B\| + \|A\|\beta_B + \varepsilon\beta_A\beta_B\big) + O(u^2),\\ &\|G\| = \|E_A B + A E_B + E_A E_B\| \le \varepsilon\big(\beta_A\|B\| + \|A\|\beta_B + \varepsilon\beta_A\beta_B\big),\end{aligned}$$

which yields (3.11). ∎

In the case $\beta_A = \|A\|$, $\beta_B = \|B\|$, the bound (3.11) simplifies to

$$\|\widehat{C} - AB\| \le \big(2\varepsilon + \varepsilon^2 + \gamma_c(1+\varepsilon)^2\big)\|A\|\|B\| + O(u^2), \tag{3.12}$$

which is similar to bound (3.7) from Lemma 3.2 for an LR matrix times a full matrix.

In the context of the BLR matrix LU factorization, computing products of LR matrices asymptotically represents the dominant cost. This has generated interest in strategies seeking to reduce this cost. For instance, in [4] it is proposed that the middle product $M = Y_A^T Y_B$ should be recompressed, that is, we should compute an LR approximation $\widetilde{M} = X_M Y_M^T \approx M$. Indeed, matrix M often has a lower numerical rank than A and B, because even though $\sigma_{\min}(\widetilde{A})$ and $\sigma_{\min}(\widetilde{B})$ are both larger than ε, $\sigma_{\min}(M) = \sigma_{\min}(\widetilde{A}\widetilde{B})$ can potentially be as small as ε^2. Note that this is only true when $\widetilde{A} = X_A Y_A^T$ and $\widetilde{B} = Y_B X_B^T$, that is, when the matrices with orthonormal columns X_A and X_B are on the outside of the product $\widetilde{A}\widetilde{B}$.

The cost of computing the product $C = \widetilde{A}\widetilde{B}$ can thus be reduced by replacing M by an LR matrix. The following lemma bounds the additional errors introduced in doing so.

Lemma 3.4 (LR matrix times LR matrix with intermediate recompression). *Let $A, B, \widetilde{A}, \widetilde{B}$ be defined as in Lemma 3.3 and let $M = Y_A^T Y_B$. If the product $C = \widetilde{A}\widetilde{B}$ is computed as $(X_A X_M)(Y_M^T X_B^T)$, where $X_M, Y_M \in \mathbb{R}^{r\times r}$, X_M has orthonormal columns, and $\widetilde{M} = X_M Y_M^T$ satisfies $\|M - \widetilde{M}\| \le \varepsilon\beta_M$, then the computed $\widehat{C}$ satisfies*

$$\|\widehat{C} - \widetilde{A}\widetilde{B}\| \le \varepsilon(1+\gamma_c)\beta_M + \gamma_c\|\widetilde{A}\|\|\widetilde{B}\| + O(u^2), \tag{3.13}$$

where $c = b + r^2 + 2r^{3/2}$, and therefore

$$\|\widehat{C} - AB\| \le \gamma_c\|A\|\|B\| + \varepsilon(1+\gamma_c)\big(\beta_M + \beta_A\|B\| + \|A\|\beta_B + \varepsilon\beta_A\beta_B\big) + O(u^2). \tag{3.14}$$

Proof. In order to do the compression, we first compute $M = Y_A^T Y_B$, obtaining $\widehat{M}$ satisfying

$$\widehat{M} = M + \Delta M, \quad \|\Delta M\| \le \gamma_b\|Y_A\|\|Y_B\| = \gamma_b\|\widetilde{A}\|\|\widetilde{B}\|.$$

Then, we compute an LR approximation to $\widehat{M}$ satisfying $\widetilde{M} = \widehat{M} + E_M$, $\|E_M\| \le \varepsilon\beta_M$. Let $W = X_A X_M$ and $Z = Y_M^T X_B^T$; the computed $\widehat{W}$ and $\widehat{Z}$ satisfy

$$\widehat{W} = X_A X_M + \Delta W, \quad \|\Delta W\| \le \gamma_r\|X_A\|\|X_M\| \le \gamma_{r^2}, \tag{3.15}$$

$$\widehat{Z} = Y_M^T X_B^T + \Delta Z, \quad \|\Delta Z\| \le \gamma_r\|X_B\|\|Y_M\| \le \gamma_{r^{3/2}}\|\widetilde{M}\|. \tag{3.16}$$

Finally, let $C = \widehat{W}\widehat{Z}$; the computed $\widehat{C}$ satisfies

$$\begin{aligned}
\widehat{C} &= \widehat{W}\widehat{Z} + F, \quad \|F\| \le \gamma_r\|\widehat{W}\|\|\widehat{Z}\| \le \gamma_{r^{3/2}}\|\widetilde{M}\| + O(u^2),\\
&= X_A X_M Y_M^T X_B^T + \Delta W Y_M^T X_B^T + X_A X_M \Delta Z + \Delta W \Delta Z + F\\
&= X_A\big(\widehat{M} + E_M\big)X_B^T + \Delta W Y_M^T X_B^T + X_A X_M \Delta Z + \Delta W \Delta Z + F\\
&= \widetilde{A}\widetilde{B} + X_A\big(\Delta M + E_M\big)X_B^T + \Delta W Y_M^T X_B^T + X_A X_M \Delta Z + \Delta W \Delta Z + F\\
&= \widetilde{A}\widetilde{B} + \Delta C, \quad \|\Delta C\| \le \varepsilon(1+\gamma_{r^2+2r^{3/2}})\beta_M + \gamma_c\|\widetilde{A}\|\|\widetilde{B}\| + O(u^2),
\end{aligned}$$

where $c = b + r^2 + 2r^{3/2}$. We obtain the slightly weaker bound (3.13) by bounding $\gamma_{r^2+2r^{3/2}}$ by γ_c. Replacing $\widetilde{A}$ by $A+E_A$ and $\widetilde{B}$ by $B+E_B$ yields (3.14) and concludes the proof. ∎

The introduction of an intermediate low-rank approximation has two consequences. First, the constant c is slightly larger in Lemma 3.4 than in Lemma 3.3, a consequence of computing one more product than previously (four instead of three products). Second, and more importantly, a new low-rank truncation term $\varepsilon\beta_M$ is introduced and is dominant in the bound (3.13) on $\|\widehat{C} - \widetilde{A}\widetilde{B}\|$. However, in the overall bound (3.14) on $\|\widehat{C} - AB\|$, $\varepsilon\beta_M$ is just one more term that adds to $\varepsilon\big(\beta_A\|B\| + \|A\|\beta_A\big)$. For instance, in the case $\beta_A = \|A\|$, $\beta_B = \|B\|$, and $\beta_M = \|A\|\|B\|$, the bound (3.14) simplifies to

$$\|\widehat{C} - AB\| \le \big(3\varepsilon + \varepsilon^2 + \gamma_c(1 + 3\varepsilon + \varepsilon^2)\big)\|A\|\|B\| + O(u^2),$$

which is roughly $3\varepsilon\|A\|\|B\|$, compared with roughly $2\varepsilon\|A\|\|B\|$ in the bound (3.12).

3.3. Triangular system with LR right-hand side. We now consider the solution of a triangular system where the right-hand side is an LR matrix $\widetilde{B} = YX^T$, which will be useful to analyze the CFU factorization in section 4.3. Note that we consider $\widetilde{B}$ of the form YX^T rather than XY^T, where X has orthonormal columns, because this is the form that arises in the CFU factorization.

Lemma 3.5 (Triangular system with LR right-hand side). *Let $T \in \mathbb{R}^{b\times b}$ be a triangular matrix and let $B \in \mathbb{R}^{b\times m}$ such that the LR matrix $\widetilde{B} = YX^T$, where X has orthonormal columns, satisfies $\|B - \widetilde{B}\| \le \varepsilon\beta$ for some $\beta > 0$. If the solution to the triangular system $T\widetilde{Z} = \widetilde{B}$ is obtained as the LR matrix $\widetilde{Z} = \widehat{W}X^T$, where $\widehat{W}$ is the computed solution ot the system $TW = Y$, then $\widetilde{Z}$ satisfies*

$$T\widetilde{Z} = \widetilde{B} + \Delta\widetilde{B}, \quad \|\Delta\widetilde{B}\| \le \gamma_b\|T\|\|\widetilde{Z}\| \tag{3.17}$$

and therefore

$$T\widetilde{Z} = B + \Delta B, \quad \|\Delta B\| \le \gamma_b\|T\|\|\widetilde{Z}\| + \varepsilon\beta. \tag{3.18}$$

Proof. By Lemma 2.2, the computed solution $\widehat{W}$ to the triangular system $TW = Y$ satisfies

$$T\widehat{W} = Y + \Delta Y, \quad \|\Delta Y\| \le \gamma_b\|T\|\|\widehat{W}\|.$$

Defining $\widetilde{Z} = \widehat{W}X^T$, we obtain $T\widetilde{Z} = \widetilde{B} + \Delta\widetilde{B}$, with

$$\|\Delta\widetilde{B}\| = \|\Delta Y\| \le \gamma_b\|T\|\|\widetilde{Z}\|,$$

proving (3.17). Replacing $\widetilde{B}$ by $B + E$, with $\|E\| \le \varepsilon\beta$, yields (3.18) and concludes the proof. ∎

Note that the assumption that X has orthonormal columns is important, as otherwise we could only prove that $\|\Delta\widetilde{B}\| \le \gamma_b\|T\|\|\widehat{W}\|\|X\|$, reflecting the possibility of cancellation if $\|\widetilde{Z}\| \ll \|\widehat{W}\|\|X\|$.

Also note that the solution $\widetilde{Z}$ is given in LR form. If it is needed as a full matrix instead, we must compute the product $\widehat{W}X^T$ and the analysis must therefore be adapted to take into account the errors introduced by this additional computation.

4. Rounding error analysis of BLR matrix algorithms.

We now turn to BLR matrices, building on the results on LR matrices obtained in the previous section. We analyze several key algorithms, starting with the BLR matrix–vector product in section 4.1, then two LU factorization algorithms in sections 4.2 and 4.3, and finally the solution to BLR linear systems in section 4.4.

4.1. BLR matrix–vector product.

We begin with the product of a BLR matrix $\widetilde{A}$ with a vector v, which is a kernel that arises in the iterative solution of linear systems based on matrix–vector products [26].

Computing $\widetilde{A}v$ requires performing p^2 LR matrix–vector products, which have been analyzed in Lemma 3.1. In the following theorem, we assume for simplicity that every block of $\widetilde{A}$ is represented as an LR matrix. In practice, some blocks (in particular the diagonal ones) would be kept as full matrices, but since the error bound (3.2) for LR matrices is larger than the bound (2.5) for full matrices, the bound derived below is still valid.

THEOREM 4.1 (BLR matrix–vector product). *Let $\widetilde{A} \in \mathbb{R}^{n\times n}$ be a $pb \times pb$ BLR matrix partitioned into p^2 LR blocks $\widetilde{A}_{ij}$ satisfying $\|A_{ij} - \widetilde{A}_{ij}\| \le \varepsilon\beta_{ij}$, and let $v \in \mathbb{R}^n$. If the ith block $z_i = z((i-1)b+1\colon ib)$ of the product $z = Av$ is computed as $z_i = \sum_{j=1}^{p} \widetilde{A}_{ij}v_j$, the computed $\widehat{z}$ satisfies*

$$\widehat{z} = (\widetilde{A} + \Delta\widetilde{A})v, \quad \|\Delta\widetilde{A}\| \le \gamma_c\|\widetilde{A}\| + O(u^2) \tag{4.1}$$

and therefore $\widehat{z} = (A + \Delta A)v$, with

$$\|\Delta A\| \le \big(\varepsilon + \gamma_c + \varepsilon\gamma_c\big)\|A\| + O(u^2) \qquad \text{if } \beta_{ij} = \|A_{ij}\|, \tag{4.2}$$

$$\|\Delta A\| \le \big(p\varepsilon + \gamma_c + p\varepsilon\gamma_c\big)\|A\| + O(u^2) \quad \text{if } \beta_{ij} = \|A\|, \tag{4.3}$$

where $c = b + r^{3/2} + p$.

Proof. Consider the computation of the block-row z_i for some $i \le p$. Let $z_i^{(j)} = \widetilde{A}_{ij}v_j$. By Lemma 3.1, the computed $\widehat{z}_i^{(j)}$ satisfies, by the error analysis for recursive summation in [19, Chap. 4],

$$\widehat{z}_i^{(j)} = \big(\widetilde{A}_{ij} + F_{ij}\big)v_j, \quad \|F_{ij}\| \le \gamma_{b+r^{3/2}}\|\widetilde{A}_{ij}\| + O(u^2).$$

Since $z_i = \sum_{j=1}^{p} \widehat{z}_i^{(j)}$, the computed $\widehat{z}_i$ satisfies

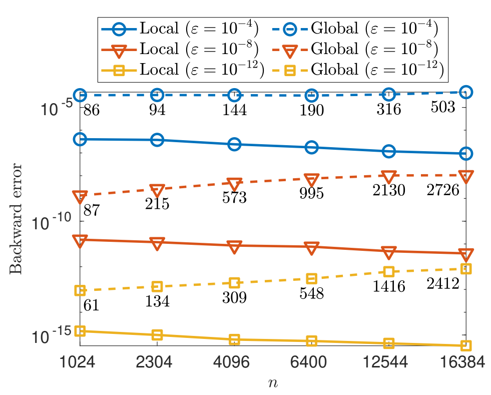
$$\begin{aligned}
\widehat{z}_i &= \sum_{j=1}^{p} \widehat{z}_i^{(j)}(1+\theta_j), \quad |\theta_j| \le \gamma_p,\\
&= \sum_{j=1}^{p} \big(\widetilde{A}_{ij} + \widetilde{A}_{ij}\theta_j + F_{ij}(1+\theta_j)\big)v_j,\\
&= \big(\widetilde{A}_i + \Delta\widetilde{A}_i\big)v,
\end{aligned}$$

where $\widetilde{A}_i$ denotes the ith block-row of $\widetilde{A}$. We therefore have $\widehat{z} = (A + \Delta\widetilde{A})v$, where the (i,j) block of $\Delta\widetilde{A}$ is given by $\Delta\widetilde{A}_{ij} = \widetilde{A}_{ij}\theta_j + F_{ij}(1+\theta_j)$. From the blockwise bounds $\|\Delta\widetilde{A}_{ij}\| \le \gamma_{b+r^{3/2}+p+O(u)}\|\widetilde{A}_{ij}\|$, we obtain

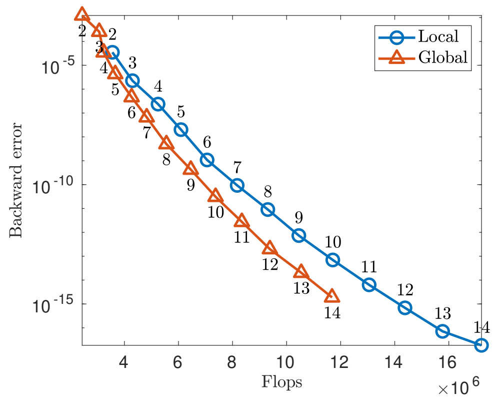
$$\|\Delta\widetilde{A}\| = \left(\sum_{i=1}^{p}\sum_{j=1}^{p}\|\Delta\widetilde{A}_{ij}\|^2\right)^{1/2} \le \gamma_{c+O(u)}\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widetilde{A}_{ij}\|^2\right)^{1/2} = \gamma_c\|\widetilde{A}\| + O(u^2),$$

with $c = b + r^{3/2} + p$, which proves (4.1). All that is left now is to replace $\widetilde{A}$ by $A + E$, where $\|E_{ij}\| \le \varepsilon\beta_{ij}$, so we obtain

$$\widehat{z} = \big(A + E + \Delta\widetilde{A}\big)v = \big(A + \Delta A\big)v, \quad \|\Delta A\| \le \gamma_c\|A\| + (1+\gamma_c)\|E\| + O(u^2).$$



(a) Backward error (3.8) for an increasing n and for three values of ε. The numbers indicate the ratio between the errors with global and local thresholds.



(b) Flops–accuracy tradeoff ($n = 4096$). The backward error (3.8) is plotted on the y-axis as a function of the flops for computing $\widetilde{A}v$ on the x-axis. Each point corresponds to a different ε threshold: the number printed next to each point is equal to $-\log_{10}\varepsilon$.

Fig. 4.1: Comparison of local and global thresholds ε for computing the product $\widetilde{A}v$, where $\widetilde{A}$ is a Poisson BLR matrix (see (2.10)) and where $v = \texttt{rand(n,1)}$.

If $\beta_{ij} = \|A_{ij}\|$, then

$$\|E\| = \left(\sum_{i=1}^{p}\sum_{j=1}^{p}\|E_{ij}\|^2\right)^{1/2} \le \varepsilon\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\|A_{ij}\|^2\right)^{1/2} = \varepsilon\|A\|,$$

which yields (4.2). However, if $\beta_{ij} = \|A\|$, then $\|E\| \le p\varepsilon\|A\|$ which yields (4.3). ∎

Bounds (4.1)–(4.3) are backward error bounds that have obvious analogues bounding the forward error $\|Av - \widehat{z}\|$.

The comments made about Lemma 3.1 on LR matrix–vector products still apply to Theorem 4.1. Most notably, the backward error bounds (4.2) and (4.3) are dominated by the low-rank truncation errors; the floating-point errors have almost no effect when $u \ll \varepsilon$.

The main novelty compared with Lemma 3.2 is that we now have BLR matrices in play, and therefore the distinction between a local or global low-rank threshold ε becomes relevant, that is: should we approximate the blocks A_{ij} relative to their own "local" norm or to the norm of the global matrix? Theorem 4.1 provides part of the answer by determining that the type of threshold mainly affects the low-rank error term, which increases from ε for a local threshold to $p\varepsilon$ for a global threshold. This is confirmed by numerical experiments in Figure 4.1a, which show that the backward error (3.8) is larger for a global threshold than a local one, and that this gap increases with n (and hence with $p = n/b$).

The question is then whether this increased error allows for a better compression and thus pays off by reducing the number of flops for the computation. To answer this question we must assess which type of threshold achieves the best flops–accuracy tradeoff. To do so, we perform the following experiment in Figure 4.1b: taking several values of ε, we plot the error (3.8) as a function of the corresponding number of flops required to compute the product. This experiment shows that a global threshold achieves the best tradeoff, since it is always closer to the bottom left corner of the plot: that is, for the same accuracy as a local threshold, a global threshold performs fewer flops or, equivalently, for the same number of flops as a local threshold, a global threshold delivers a more accurate result. A global threshold is therefore the best choice for this matrix and algorithm. We will show in section 5 this remains the case for LU factorization of a wide range of BLR matrices.

We finally mention that Theorem 4.1 can be easily adapted to the product of a BLR matrix $\widetilde{A}$ with a matrix V (instead of a vector), yielding the following forward error result.

Theorem 4.2 (BLR matrix–matrix product). *Let $\widetilde{A} \in \mathbb{R}^{n\times n}$ be a $pb \times pb$ BLR matrix partitioned into p^2 LR blocks $\widetilde{A}_{ij}$ satisfying $\|A_{ij} - \widetilde{A}_{ij}\| \le \varepsilon\beta_{ij}$, and let $V \in \mathbb{R}^{n\times m}$. If the columns z_j of Z of the product $Z = AV$ are computed as $z_j = \widetilde{A}v_j$ then the computed $\widehat{Z}$ satisfies*

$$\|\widehat{Z} - \widetilde{A}V\| \le \gamma_c\|\widetilde{A}\|\|V\| + O(u^2), \tag{4.4}$$

where $c = b + r^{3/2} + p$, and therefore

$$\|\widehat{Z} - AV\| \le \big(\varepsilon + \gamma_c + \varepsilon\gamma_c\big)\|A\|\|V\| + O(u^2) \qquad \text{if } \beta_{ij} = \|A_{ij}\|, \tag{4.5}$$

$$\|\widehat{Z} - AV\| \le \big(p\varepsilon + \gamma_c + p\varepsilon\gamma_c\big)\|A\|\|V\| + O(u^2) \quad \text{if } \beta_{ij} = \|A\|. \tag{4.6}$$

Proof. The result directly follows from the columnwise bounds, from Theorem 4.1,

$$\|\widehat{z_j} - \widetilde{A}v_j\| \le \gamma_c\|\widetilde{A}\|\|v_j\| + O(u^2), \quad j = 1\colon m. \qquad \square$$

4.2. BLR matrix LU factorization: FCU algorithm. We now turn to the LU factorization of BLR matrices. To compute such factorizations, the classical partitioned LU factorization of full matrices must be adapted by incorporating the compressions of the blocks into LR matrices. Several algorithms have been distinguished depending on when this compression step is performed. Algorithm 4.1 describes the FCU algorithm (standing for factor, compress, update). The algorithm is written in a left-looking fashion: at step k of the FCU algorithm, the kth block-row and block-column are first factored (lines 4–7) before being compressed (lines 9–12), and finally the next block-row and block-column are updated (lines 14–17) using the BLR LU factors "to the left". Note that at step $k = 1$, the update step does not do anything.

Another algorithm, referred to as CFU (compress, factor, update), is analyzed in section 4.3.

We recall that, for all the experiments on the BLR LU factorization, we measure the backward error by solving a linear system as explained in section 2.3.

The next theorem analyzes the FCU algorithm.

Theorem 4.3 (BLR LU factorization: FCU algorithm). *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix partitioned into p^2 blocks of order b. If Algorithm 4.1 runs to completion it produces computed BLR LU factors $\widetilde{L}$ and $\widetilde{U}$ of A satisfying*

$$A = \widetilde{L}\widetilde{U} + \Delta A + F + G,$$

Algorithm 4.1 BLR LU factorization: FCU algorithm

```
1: {**Input**: a $p \times p$ block matrix $A$. **Output**: its BLR LU factors $\widetilde{L}$ and $\widetilde{U}$.}
2: for $k = 1$ to $p$ do
3:   FACTOR:
4:       Compute the LU factorization $L_{kk}U_{kk} = A_{kk}$.
5:       for $i = k+1$ to $p$ do
6:         Solve $L_{ik}U_{kk} = A_{ik}$ for $L_{ik}$ and $L_{kk}U_{ki} = A_{ki}$ for $U_{ki}$.
7:       end for
8:   COMPRESS:
9:       Set $\widetilde{L}_{kk} = L_{kk}$ and $\widetilde{U}_{kk} = U_{kk}$.
10:      for $i = k+1$ to $p$ do
11:        Compute LR approximations $\widetilde{L}_{ik} \approx L_{ik}$ and $\widetilde{U}_{ki} \approx U_{ki}$.
12:      end for
13:  UPDATE:
14:      $A_{kk} \leftarrow A_{kk} - \sum_{j=1}^{k-1} \widetilde{L}_{kj}\widetilde{U}_{jk}$.
15:      for $i = k+1$ to $p$ do
16:        $A_{ik} \leftarrow A_{ik} - \sum_{j=1}^{k-1} \widetilde{L}_{ij}\widetilde{U}_{jk}$ and $A_{ki} \leftarrow A_{ki} - \sum_{j=1}^{k-1} \widetilde{L}_{kj}\widetilde{U}_{ji}$.
17:      end for
18: end for
```

with $\|\Delta A\| \le \gamma_p \|A\|$ *and* $\|F\| \le \gamma_c \|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon)$, *where* $c = b + 2r^{3/2} + p$ *and*

$$G_{ik} = \begin{cases} E_{ik}\widetilde{U}_{kk} & i > k, \\ 0 & i = k, \\ \widetilde{L}_{ii}E_{ik} & i < k, \end{cases} \qquad \|E_{ik}\| \le \varepsilon\beta_{ik}.$$

Proof. The (i,k) block of the L factor is computed by solving

$$L_{ik}\widetilde{U}_{kk} = R_{ik}, \quad R_{ik} = A_{ik} - \sum_{j=1}^{k-1} \widetilde{L}_{ij}\widetilde{U}_{jk}, \quad i > k, \tag{4.7}$$

where $\widetilde{L}$ and $\widetilde{U}$ are the partial BLR LU factors computed in the previous $k-1$ steps (line 11 of Algorithm 4.1). Let $R_{ik}^{(j)} = \widetilde{L}_{ij}\widetilde{U}_{jk}$; by (3.10), the computed $\widehat{R}_{ik}^{(j)}$ satisfies

$$\widehat{R}_{ik}^{(j)} = \widetilde{L}_{ij}\widetilde{U}_{jk} + \Delta R_{ik}^{(j)}, \quad \|\Delta R_{ik}^{(j)}\| \le \gamma_d \|\widetilde{L}_{ij}\|\|\widetilde{U}_{jk}\| + O(u^2) \tag{4.8}$$

with $d = b + 2r^{3/2}$. The computed $\widehat{R}_{ik}$ then satisfies

$$\widehat{R}_{ik} = A_{ik}(1+\theta_k) - \sum_{j=1}^{k-1} \widehat{R}_{ik}^{(j)}(1+\theta_j), \quad |\theta_j| \le \gamma_p. \tag{4.9}$$

By (2.8) we have

$$\widehat{L}_{ik}\widetilde{U}_{kk} = \widehat{R}_{ik} + \Delta R_{ik}^{(k)}, \quad \|\Delta R_{ik}^{(k)}\| \le \gamma_b \|\widehat{L}_{ik}\|\|\widetilde{U}_{kk}\|. \tag{4.10}$$

After compression, we finally obtain the BLR factor $\widetilde{L}_{ik} = \widehat{L}_{ik} + E_{ik}$, with $\|E_{ik}\| \le \varepsilon\beta_{ik}$. Combining (4.8), (4.9), and (4.10) gives

$$A_{ik}(1+\theta_k) - \sum_{j=1}^{k} \widetilde{L}_{ij}\widetilde{U}_{jk} = \sum_{j=1}^{k-1} \Delta R_{ik}^{(j)}(1+\theta_j) - \Delta R_{ik}^{(k)} - E_{ik}\widetilde{U}_{kk}.$$

We therefore obtain

$$A_{ik} - \sum_{j=1}^{k} \widetilde{L}_{ij}\widetilde{U}_{jk} = \Delta A_{ik} + F_{ik} + G_{ik}, \quad \|\Delta A_{ik}\| \le \gamma_p \|A_{ik}\|, \tag{4.11}$$

$$\begin{aligned} \|F_{ik}\| &\le \gamma_{d+p+O(u)} \left(\sum_{j=1}^{k-1} \|\widetilde{L}_{ij}\|\|\widetilde{U}_{jk}\| \right) + \gamma_b \|\widehat{L}_{ik}\|\|\widetilde{U}_{kk}\| \\ &\le \gamma_{d+p+O(\varepsilon)} \left(\sum_{j=1}^{k} \|\widetilde{L}_{ij}\|\|\widetilde{U}_{jk}\| \right), \end{aligned} \tag{4.12}$$

$$\|G_{ik}\| \le \|E_{ik}\widetilde{U}_{kk}\| \le \varepsilon\beta_{ik}\|\widetilde{U}_{kk}\|, \quad i > k. \tag{4.13}$$

This concludes the blocks for $i > k$. For $i = k$, L_{kk} is determined together with U_{kk} on line 4 of Algorithm 4.1, and by Lemma 2.3 we have $\|\widetilde{L}_{kk}\widetilde{U}_{kk} - \widehat{R}_{kk}\| \le \gamma_b \|\widetilde{L}_{kk}\|\|\widetilde{U}_{kk}\|$. Therefore (4.10) holds for $i = k$, too, and hence so does (4.11), with the same bound (4.12) on $\|F_{kk}\|$ and with $G_{kk} = 0$, because diagonal blocks are not compressed. Finally, the case $i < k$ is analogous to the case $i > k$ and yields (4.11) where the bound (4.12) becomes

$$\|F_{ik}\| \le \gamma_{d+p+O(\varepsilon)} \left(\sum_{j=1}^{i} \|\widetilde{L}_{ij}\|\|\widetilde{U}_{jk}\| \right) \tag{4.14}$$

and with

$$\|G_{ik}\| \le \varepsilon\beta_{ik}\|\widetilde{L}_{ii}\|, \quad i < k. \tag{4.15}$$

We have therefore proved that $A - \widetilde{L}\widetilde{U} = \Delta A + F + G$, where $G_{kk} = 0$ and blockwise bounds on $\|\Delta A\|$, $\|F\|$, and $\|G\|$ are given by (4.11)–(4.15). It thus remains to derive global bounds. Bounding $\|G\|$ is delayed to section 4.2.1, as it depends on the choice of the β_{ik} parameters. The bound $\|\Delta A\| \le \gamma_p \|A\|$ trivially holds. Finally, for matrix F, the Cauchy–Schwarz inequality gives

$$\begin{aligned} \|F\| &\le \gamma_{d+p+O(\varepsilon)} \left(\sum_{i=1}^{p}\sum_{k=1}^{p} \left(\sum_{j=1}^{\min(i,k)} \|\widetilde{L}_{ij}\|\|\widetilde{U}_{jk}\| \right)^2 \right)^{1/2} \\ &\le \gamma_{d+p+O(\varepsilon)} \left(\sum_{i=1}^{p}\sum_{j=1}^{i} \|\widetilde{L}_{ij}\|^2 \sum_{k=1}^{p}\sum_{j=1}^{k} \|\widetilde{U}_{jk}\|^2 \right)^{1/2} \\ &\le \gamma_{d+p+O(\varepsilon)} \left(\left(\sum_{i=1}^{p} \|\widetilde{L}_{i}\|^2 \sum_{k=1}^{p} \|\widetilde{U}_{k}\|^2 \right) \right)^{1/2} \\ &\le \gamma_{d+p} \|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon), \end{aligned} \tag{4.16}$$

where $\widetilde{L}_i$ and $\widetilde{U}_k$ denote the ith block-row of $\widetilde{L}$ and kth block-column of $\widetilde{U}$, respectively. ∎

Theorem 4.3 yields a backward error bound that is comparable to the other results obtained so far: we obtain a term $\|\Delta A\| + \|F\|$ proportional to the unit roundoff u, and a term $\|G\|$ depending on the low-rank threshold ε, the latter likely dominating the former. However, before further commenting on the significance of this bound, we must compute a global bound on $\|G\|$ by examining several possible choices for the β_{ik} parameters.

4.2.1. Bounding $\|G\|$ and choice of β_{ik}. The blockwise bounds on $\|G_{ik}\|$ given in Theorem 4.3 yield the global bound

$$\|G\|^2 \le \varepsilon^2 \sum_{k=1}^{p} \Bigl(\|\widetilde{L}_{kk}\|^2 \sum_{i=1}^{k-1} \beta_{ik}^2 + \|\widetilde{U}_{kk}\|^2 \sum_{i=k+1}^{p} \beta_{ik}^2 \Bigr).$$

With a local threshold $\beta_{ik} = \|A_{ik}\|$, we have

$$\begin{aligned} \|G\| &\le \varepsilon \Bigl(\sum_{k=1}^{p} \max\bigl(\|\widetilde{L}_{kk}\|, \|\widetilde{U}_{kk}\|\bigr)^2 \sum_{i \ne k}^{p} \|A_{ik}\|^2 \Bigr)^{1/2} \\ &\le \varepsilon \max_{k=1:p} \Bigl(\max\bigl(\|\widetilde{L}_{kk}\|, \|\widetilde{U}_{kk}\|\bigr) \Bigr) \|A\|. \end{aligned} \tag{4.17}$$

On the other hand, a global threshold $\beta_{ik} = \|A\|$ yields

$$\begin{aligned} \|G\| &\le \varepsilon \Bigl(\sum_{k=1}^{p} \max\bigl(\|\widetilde{L}_{kk}\|, \|\widetilde{U}_{kk}\|\bigr)^2 \sum_{i \ne k}^{p} \|A\|^2 \Bigr)^{1/2} \\ &\le (p-1)^{1/2} \varepsilon \|D\| \|A\|, \end{aligned} \tag{4.18}$$

where D is the diagonal matrix defined by $D_{kk} = \max\bigl(\|\widetilde{L}_{kk}\|, \|\widetilde{U}_{kk}\|\bigr)$. Bound (4.18) is thus up to a factor $\bigl(p(p-1)\bigr)^{1/2} \approx p$ times larger than (4.17).

Two main observations can be made. First, the use of a global threshold leads again to a bound about p times larger than with a local threshold, just as for Theorem 4.1. This is only a modest loss of accuracy, and indeed Figure 4.2b illustrates that on a Poisson BLR matrix (see (2.10)) a global threshold achieves the best flops–accuracy tradeoff. We will compare global and local thresholds on a wider range of matrices in section 5.

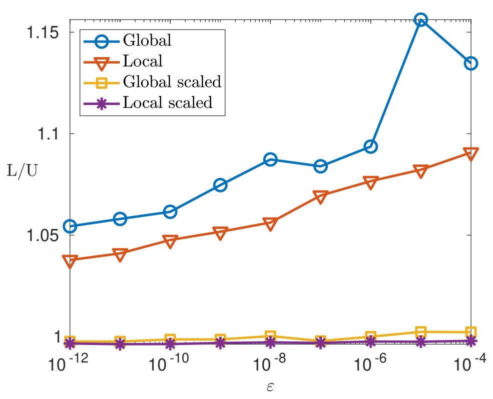
The second important observation is that the error depends on the norm of the diagonal blocks of the LU factors ($\max(\|\widetilde{L}_{kk}\|, \|\widetilde{U}_{kk}\|)$ in (4.17) and $\|D\|$ in (4.18)). This is due to the compress step (lines 9–12 of Algorithm 4.1) being performed after the factor step (lines 4–7). Indeed, the blocks that are compressed are equal to $A_{ik}U_{kk}^{-1}$ and $L_{kk}^{-1}A_{ki}$ and so their norms depend on those of the diagonal blocks. This property of the FCU factorization is undesirable because the L and U factors are often not scaled comparably: the entries of L are bounded by 1 with partial pivoting, whereas those of U are scaled similarly to those of A. As a consequence, the average ranks of the blocks differ depending on whether they belong to the L or U factors, even for symmetric matrices where $U = L^T$. This is illustrated in Figure 4.2a, where both the local and global threshold strategies lead to a U factor with lower ranks than the L factor.

There exist several solutions to avoid the dependence of the ranks of the BLR LU factors and the backward error on the norms of the diagonal blocks. One solution consists in scaling the β_{ik} differently for the L and U factors: specifically, setting $\beta_{ik} = \|A_{ik}\|/\|\widetilde{U}_{kk}\|$ (if $i > k$) and $\beta_{ik} = \|A_{ik}\|/\|\widetilde{L}_{ii}\|$ (if $i < k$) changes bound (4.17) to

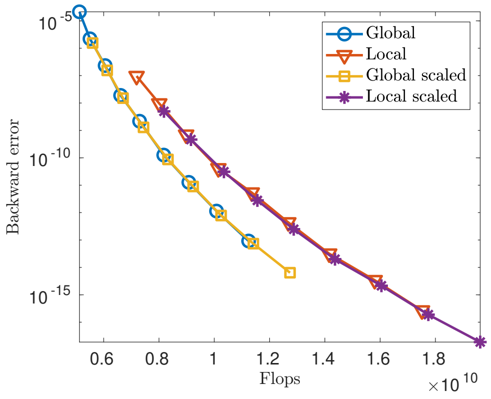
$$\|G\| \le \varepsilon \|A\| \tag{4.19}$$

and setting $\beta_{ik} = \|A\|/\|\widetilde{U}_{kk}\|$ (if $i > k$) and $\beta_{ik} = \|A\|/\|\widetilde{L}_{ii}\|$ (if $i < k$) changes bound (4.18) to

$$\|G\| \le p\varepsilon \|A\|. \tag{4.20}$$



(a) Ratio between the storage for the BLR $\widetilde{L}$ and $\widetilde{U}$ factors.



(b) Flops–accuracy tradeoff. The backward error (2.11) is plotted on the y-axis as a function of the flops for computing the factorization on the x-axis. Each point corresponds to a different ε threshold.

Fig. 4.2: Comparison of several choices for β_{ik} in the BLR FCU factorization (Algorithm 4.1) of a Poisson matrix (see (2.10)), of order $n = 4096$. Local: $\beta_{ik} = \|A_{ik}\|$; local scaled: $\beta_{ik} = \|A_{ik}\|/\|\widetilde{U}_{kk}\|$ $(i > k)$ and $\beta_{ik} = \|A_{ik}\|/\|\widetilde{L}_{ii}\|$ $(i < k)$; global: $\beta_{ik} = \|A\|$; global scaled: $\beta_{ik} = \|A\|/\|\widetilde{U}_{kk}\|$ $(i > k)$ and $\beta_{ik} = \|A\|/\|\widetilde{L}_{ii}\|$ $(i < k)$.

This scaling of the threshold yields similar ranks in both L and U, as shown in Figure 4.2a. Even though this strategy does not achieve a visibly better flops–accuracy tradeoff, as illustrated in Figure 4.2b, in the following we consider scaled thresholds since they simplify the bounds.

Note that there are alternative strategies to scaling the threshold. For example computing an LDU rather than LU factorization, where both the L and U factors have entries bounded by one. Interestingly, as we show in section 4.3, another solution is to perform a CFU factorization, which avoids this issue by compressing the blocks before factorizing them.

4.2.2. General comments on Theorem 4.3. Outside the technical discussion of the previous section on how to choose the β_{ik} parameters, some higher level conclusions can be drawn from Theorem 4.3, which we summarize in the following result.

Corollary 4.4. *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix partitioned into p^2 blocks of order b. If Algorithm 4.1 runs to completion, it produces BLR LU factors $\widetilde{L}$ and $\widetilde{U}$ of A satisfying*

$$A = \widetilde{L}\widetilde{U} + \Delta A, \quad \|\Delta A\| \le \big(\xi_p\varepsilon + \gamma_p\big)\|A\| + \gamma_c\|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon), \tag{4.21}$$

where $c = b + 2r^{3/2} + p$, and $\xi_p = 1$ or $\xi_p = p$ for a scaled local or global threshold, respectively.

Proof. Directly follows from Theorem 4.3, (4.16), (4.19), and (4.20). ∎

Corollary 4.4 states that the backward error $\|A - \widetilde{L}\widetilde{U}\|$ is of order $O(\varepsilon\|A\| + u\|\widetilde{L}\|\|\widetilde{U}\|)$. The first term corresponds to the low-rank truncations errors and the second term to the floating-point errors. If we set $\varepsilon = 0$, we recover the backward error bound (2.9) for classical LU factorization (with a slightly higher constant). If we set $u = 0$, we obtain a bound on the error introduced by BLR approximations in exact arithmetic. In the general case, since $u \ll \varepsilon$, the low-rank error term dominates if $\|\widetilde{L}\|\|\widetilde{U}\|$ is not too large compared with $\|A\|$. For a stable LU factorization, using for example partial pivoting, $\|\widetilde{L}\|\|\widetilde{U}\|$ is bounded and therefore the BLR factorization is also stable. This is the main conclusion drawn from Corollary 4.4: it proves the long conjectured rule of thumb that the backward error for the BLR LU factorization is proportional to the low-rank threshold ε. This is a very desirable theoretical guarantee that can now be given to the users of BLR solvers. This is illustrated by the numerical experiments on a Poisson matrix (see (2.10)) reported in Figure 4.3a, which additionally show that the unit roundoff u has no impact on the error as long as $u \ll \varepsilon$. We will verify experimentally that this crucial result holds for a wide range of matrices in section 5.

The low-rank error term $\xi_p \varepsilon \|A\|$ grows at most linearly with the number of blocks p. Figure 4.3b illustrates that a roughly linear growth (indicated by the dashed lines) is indeed observed in practice when using a global threshold (except for the largest of the tested thresholds, $\varepsilon = 10^{-4}$, for which the error growth is sublinear). This seems quite acceptable, especially considering that the constant in the error bound for traditional LU factorization is usually of order n^3 [19, Thm. 9.5].

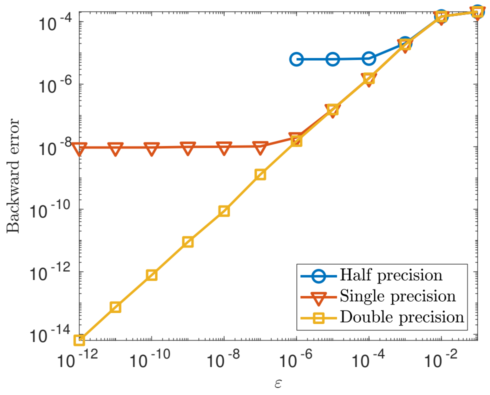
Finally, we briefly comment on the use of fast matrix operations, such as Strassen's algorithm [32]. Similarly to the LR matrix algorithms analyzed in section 3.1, fast matrix algorithms are especially attractive with BLR matrices since only normwise stability is expected, and because they only affect the floating-point error term, which is negligible compared with the low-rank error term for large enough ε. Our analysis therefore theoretically explains why the stability of the BLR factorization is much less sensitive to the use of fast matrix arithmetic, as experimentally observed in [24]. We mention that the algorithm proposed in [24] recasts the operations so as to work on matrices of larger dimensions, which allows for exploiting fast matrix arithmetic more efficiently. The error analysis of this new algorithm is outside our scope, but we expect it to retain the same backward stability as the algorithms analyzed here.

4.2.3. Impact of intermediate recompressions. We now discuss the impact on the accuracy of performing intermediate recompressions during the update step, as described and analyzed in Lemma 3.4.

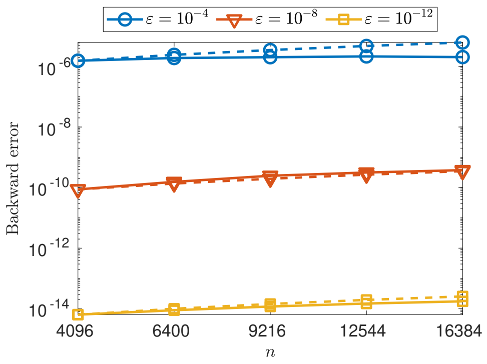
Adapting the proof of Theorem 4.3 to the use of intermediate recompressions is straightforward. It suffices to invoke Lemma 3.4 instead of Lemma 3.3, which changes the expression for R_{ik} in (4.7) to

$$R_{ik} = A_{ik} - \sum_{j=1}^{k-1} \Big(\widetilde{L}_{ij}\widetilde{U}_{jk} + H_{ik}^{(j)} \Big),$$

which has an extra term $H_{ik}^{(j)}$ satisfying $\|H_{ik}^{(j)}\| \le \varepsilon \beta_{ik}^{(j)}$. For simplicity, let us consider the same choice $\beta_{ik}^{(j)} = \beta_{ik}^{H}$ for all j. The rest of the proof carries over and we obtain



(a) Error for varying low-rank thresholds ε and floating-point precisions.



(b) Error for increasing n. Dashed lines indicate a linear growth.

Fig. 4.3: Backward error (2.11) for computing the BLR LU factorization of a Poisson matrix (see (2.10)) of order $n = 4096$ with the FCU algorithm (Algorithm 4.1) and with a global threshold.

$A = \widetilde{L}\widetilde{U} + \Delta A + F + G$, where we now have

$$G_{ik} = \begin{cases} E_{ik}\widetilde{U}_{kk} + H_{ik}, & i > k, \\ 0, & i = k, \\ \widetilde{L}_{ii}E_{ik} + H_{ik}, & i < k, \end{cases} \qquad \|E_{ik}\| \le \varepsilon\beta_{ik}, \quad \|H_{ik}\| \le (\min(i,k)-1)\varepsilon\beta_{ik}^H.$$

To proceed further we must consider specific choices for the β_{ik} and β_{ik}^H parameters, as in section 4.2.1. For a scaled local threshold (4.19) and for $\beta_{ik}^H = \varepsilon\|A_{ik}\|$, we have $\|G_{ik}\| \le \min(i,k)\varepsilon\|A_{ik}\|$, whereas for a scaled global threshold (4.20) and $\beta_{ik}^H = \varepsilon\|A\|$, we obtain instead $\|G_{ik}\| \le \min(i,k)\varepsilon\|A\|$.

Tedious but straightforward computations lead to the bound

$$\|G\| \le \xi_p \varepsilon \|A\|, \tag{4.22}$$

where $\xi_p = p$ and $\xi_p = \frac{1}{\sqrt{6}}p^2$ for a local and global threshold, respectively. The low-rank error is therefore a factor roughly p times larger when intermediate recompressions are performed than when they are not, for both local and global thresholds.

Theorem 4.5. *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix partitioned into p^2 blocks of order b. If Algorithm 4.1 runs to completion, it produces BLR LU factors $\widetilde{L}$ and $\widetilde{U}$ of A satisfying*

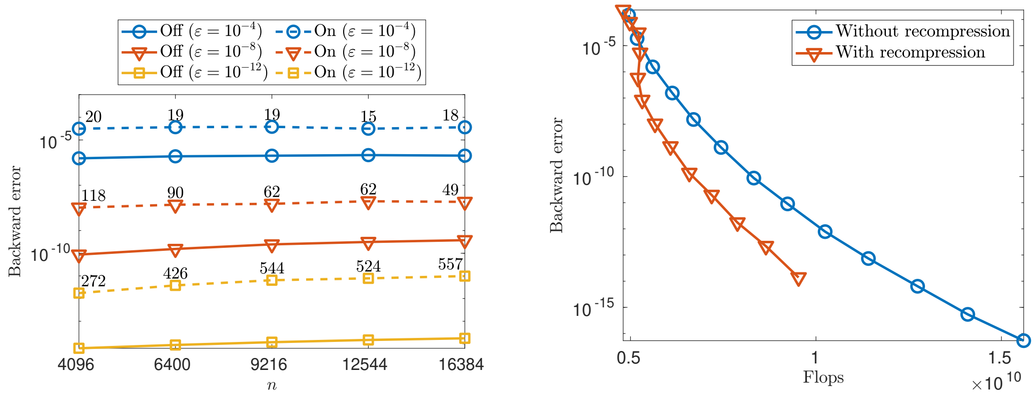
$$A = \widetilde{L}\widetilde{U} + \Delta A, \quad \|\Delta A\| \le \big(\xi_p\varepsilon + \gamma_p\big)\|A\| + \gamma_c\|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon), \tag{4.23}$$

where $c = b + 2r^{3/2} + p$, and where ξ_p is given in Table 4.1.

We illustrate the above analysis with some numerical experiments in Figure 4.4, using a scaled global threshold (results with a local threshold are similar and omitted). In Figure 4.4a, we compare the error growth with and without recompressions. Recompression increases the error by a noticeable factor; however, this factor increases

Table 4.1: Expression of ξ_p in Theorems 4.5, 4.6, and 4.8, depending on whether a local or global threshold is used, and on whether intermediate recompressions are performed during the LU factorization.

	Local threshold	Global threshold
Without recompressions	1	p
With recompressions	p	$p^2/\sqrt{6}$



(a) Backward error (2.11) for increasing n, depending on whether recompressions are "off" or "on". The numbers indicate the ratio between the corresponding errors.

(b) Flops–accuracy tradeoff. The backward error (2.11) is plotted on the y-axis as a function of the flops for computing the factorization on the x-axis. Each point corresponds to a different ε threshold.

Fig. 4.4: Impact of intermediate recompressions on the backward error for a Poisson matrix (see (2.10)) of order $n = 4096$ with the FCU algorithm (Algorithm 4.1) and with a global threshold.

relatively slowly with n. The error growth of the factorization with recompression therefore remains contained. To determine whether this increased error pays off in terms of flops, we plot in Figure 4.4b the error as a function of the flop count for the factorization for several values of ε. Clearly, for this Poisson matrix (see (2.10)), the strategy using recompressions achieves a better tradeoff. We will show that this remains true for a wide range of matrices in section 5.

4.3. BLR matrix LU factorization: CFU algorithm. In the FCU algorithm (Algorithm 4.1), the compress step is performed after the factor step and thus the latter does not exploit the LR property of the blocks. The CFU algorithm, described in Algorithm 4.2, is based on the idea of performing the compress step earlier, before the factor step, so that the off-diagonal blocks may be factored in LR form, as shown on line 10 and as analyzed in Lemma 3.5. This reduces the number of flops needed for the factor step, which is especially important because this step is asymptotically dominant in the FCU algorithm. The CFU algorithm is thus necessary to achieve an

Algorithm 4.2 BLR LU factorization: CFU algorithm

```
1: {Input: a $p \times p$ block matrix $A$. Output: its BLR LU factors $\widetilde{L}$ and $\widetilde{U}$.}
2: for $k = 1$ to $p$ do
3:   COMPRESS:
4:     for $i = k + 1$ to $p$ do
5:       Compute LR approximations $\widetilde{A}_{ik} \approx A_{ik}$ and $\widetilde{A}_{ki} \approx A_{ki}$.
6:     end for
7:   FACTOR:
8:     Compute the LU factorization $\widetilde{L}_{kk}\widetilde{U}_{kk} = A_{kk}$.
9:     for $i = k + 1$ to $p$ do
10:      Solve $\widetilde{L}_{ik}U_{kk} = \widetilde{A}_{ik}$ for $\widetilde{L}_{ik}$ and $L_{kk}\widetilde{U}_{ki} = \widetilde{A}_{ki}$ for $\widetilde{U}_{ki}$.
11:    end for
12:  UPDATE:
13:    $A_{kk} \leftarrow A_{kk} - \sum_{j=1}^{k-1} \widetilde{L}_{kj}\widetilde{U}_{jk}$.
14:    for $i = k + 1$ to $p$ do
15:      $A_{ik} \leftarrow A_{ik} - \sum_{j=1}^{k-1} \widetilde{L}_{ij}\widetilde{U}_{jk}$ and $A_{ki} \leftarrow A_{ki} - \sum_{j=1}^{k-1} \widetilde{L}_{kj}\widetilde{U}_{ji}$.
16:    end for
17: end for
```

optimal complexity [4].

However, the CFU algorithm has not yet been widely accepted as the method of choice, and some BLR solvers still use the FCU algorithm by default, such as MUMPS [5]. There are two reasons for this. The first is that the impact on the accuracy of switching from FCU to CFU was not fully understood and quantified. The next theorem provides an answer to this open question. The second reason is related to numerical pivoting. Performing pivoting requires access to the entries of the entire row or column being factored; however, in the CFU algorithm, the blocks have already been compressed and so the entries of the original block are no longer available. Strategies estimating these entries based on the entries of the LR blocks have been proposed [4, 5, 25] and appear to deliver satisfying results. This is however the object of ongoing research and is outside our scope.

THEOREM 4.6 (BLR LU factorization: CFU algorithm). *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix partitioned into p^2 blocks of order b. If Algorithm 4.2 runs to completion, it produces BLR LU factors of A satisfying*

$$A = \widetilde{L}\widetilde{U} + \Delta A, \quad \|\Delta A\| \le \big(\xi_p\varepsilon + \gamma_p\big)\|A\| + \gamma_c\|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon), \tag{4.24}$$

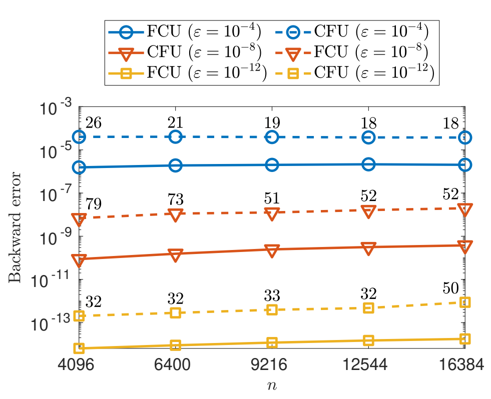
where $c = b + 2r^{3/2} + p$, and where ξ_p is given in Table 4.1.

Proof. In contrast with (4.7) in the proof of Theorem 4.3, the (i, k) block of the L factor is now computed by solving instead

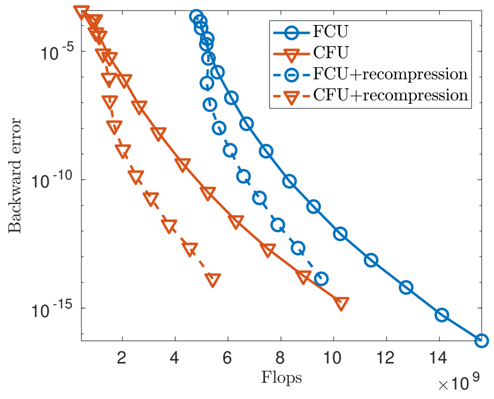
$$L_{ik}\widetilde{U}_{kk} = \widetilde{R}_{ik}, \quad i > k, \tag{4.25}$$

where $\widetilde{R}_{ik}$ is a LR approximation to $\widehat{R}_{ik}$ satisfying $\widetilde{R}_{ik} = \widehat{R}_{ik} + E_{ik}$, with $\|E_{ik}\| \le \varepsilon\beta_{ik}$, and where $\widehat{R}_{ik}$ still satisfies (4.9). (4.25) takes the form of a triangular solve with an LR right-hand side, and thus by (3.17) we have

$$\widetilde{L}_{ik}\widetilde{U}_{kk} = \widetilde{R}_{ik} + \Delta R_{ik}^{(k)}, \quad \|\Delta R_{ik}^{(k)}\| \le \gamma_b\|\widetilde{L}_{ik}\|\|\widetilde{U}_{kk}\|. \tag{4.26}$$



(a) Backward error (2.11) for increasing n (without intermediate recompression). The numbers indicate the ratio between the errors for the CFU and FCU algorithms.



(b) Flops–accuracy tradeoff ($n = 4096$). The backward error (2.11) is plotted on the y-axis as a function of the flops for computing the factorization on the x-axis. Each point corresponds to a different ε threshold.

Fig. 4.5: Comparison between FCU (Algorithm 4.1) and CFU (Algorithm 4.2) BLR LU factorizations for Poisson matrices (see (2.10)).

We therefore obtain

$$A_{ik}(1+\theta_k) - \sum_{j=1}^{k} \widetilde{L}_{ij}\widetilde{U}_{jk} = \sum_{j=1}^{k-1} \Delta R_{ik}^{(j)} - \Delta R_{ik}^{(k)} - E_{ik} = F_{ik} + G_{ik}, \tag{4.27}$$

with the same bound (4.12) on $\|F_{ik}\|$ as for the FCU algorithm, but a different bound $\|G_{ik}\| \le \varepsilon\beta_{ik}$ instead of (4.13), without the term $\|\widetilde{U}_{kk}\|$. This concludes the proof for the case $i > k$. The cases $i = k$ and $i < k$ are similar and overall we have $A - \widetilde{L}\widetilde{U} = \Delta A + F + G$, where the blockwise bounds on $\|\Delta A\|$ and $\|F\|$ are the same as those for the FCU algorithm, whereas $\|G_{ik}\| \le \varepsilon\beta_{ik}$ for all i (with $\beta_{kk} = 0$). The bound $\|G\| \le \xi_p\|A\|$ trivially follows with $\xi_p = 1$ or $\xi_p = p$ depending on whether $\beta_{ik} = \|A_{ik}\|$ or $\beta_{ik} = \|A\|$, respectively. $\square$

A notable difference of the CFU algorithm is that, unlike the FCU algorithm, it does not depend on the norm of the diagonal blocks of the LU factors. The CFU algorithm thus avoids the issue of having different compression in the L and U factors, as discussed in section 4.2.1, and hence there is no reason to scale the threshold as suggested for the FCU algorithm (see (4.20)). The main conclusion to draw from Theorem 4.6 is therefore that the CFU algorithm satisfies the same error bound as the FCU one when the latter algorithm uses a scaled threshold (as in Theorem 4.5).

This conclusion is supported by numerical experiments in Figure 4.5. Figure 4.5a shows that the FCU and CFU algorithms yield similar errors on Poisson matrices, regardless of the matrix size n. Indeed, even though the CFU algorithm yields a backward error larger by a noticeable factor, this factor does not increase with n. Therefore, since the CFU algorithm achieves a lower flop count, it achieves a much better tradeoff than the FCU one, and this is illustrated in Figure 4.5b. These observations will be extended to a wider range of matrices in section 5.

Finally, we briefly comment on some other BLR LU factorization variants, for which we have performed similar analyses that we omit for the sake of conciseness.

The FUC algorithm [25] performs the compression step only after the completion of the LU factorization. This variant therefore only reduces the storage, not the flops, for performing the LU factorization. Its main advantage is to avoid the low-rank and the floating-point errors accumulating together, replacing the $O(u\varepsilon)$ term in the bounds of Theorems 4.3 and 4.6 by $O(u^2)$. This is however a negligible reduction of the error, and therefore the FUC variant is not competitive with the other variants.

Another widely used algorithm is the CUF variant (see, for example, [4, 28]), which compresses the entire matrix A and then computes its BLR LU factorization. The CUF algorithm is very similar to the CFU one, only differing in that, at line 15 of Algorithm 4.2, the blocks A_{ik} (and A_{ki}) are already in LR form. Therefore the result of the product of the LR factors $\widetilde{L}_{ij}\widetilde{U}_{jk}$ may be obtained directly as an LR matrix, avoiding the last product in Lemma 3.3: this just affects the constant in the error bound and we therefore conclude that the CFU and CUF algorithms achieve similar error bounds. Note that one particularity of the CUF algorithm is that the use of intermediate recompressions (section 4.2.3) is mandatory to contain the growth of the ranks of the BLR LU factors throughout the factorization.

4.4. BLR linear systems. We first analyze the solution of a triangular system $\widetilde{T}x = v$, where $\widetilde{T}$ is a BLR matrix.

THEOREM 4.7 (BLR triangular system). *Let $\widetilde{T} \in \mathbb{R}^{n\times n}$ be a triangular BLR matrix partitioned into p^2 LR blocks $\widetilde{T}_{ij} \in \mathbb{R}^{b\times b}$ and let $v \in \mathbb{R}^n$. If the solution to the system $\widetilde{T}x = v$ is computed by solving the triangular system $T_{ii}x_i = v_i - \sum_{j=1}^{i-1} \widetilde{T}_{ij}x_j$ for each block $x_i = x((i-1)b+1 : ib)$, the computed solution $\widehat{x}$ satisfies*

$$(\widetilde{T} + \Delta\widetilde{T})\widehat{x} = v + \Delta v, \quad \|\Delta\widetilde{T}\| \le \gamma_c\|\widetilde{T}\|, \quad \|\Delta v\| \le \gamma_p\|v\|, \tag{4.28}$$

where $c = b + r^{3/2} + p$.

Proof. Let $w_i^{(j)} = \widetilde{T}_{ij}\widehat{x}_j$, where $\widehat{x}_j$ is the jth block-row of the computed $\widehat{x}$ in the previous $i-1$ steps. By Lemma 3.1, the computed $\widehat{w}_i^{(j)}$ satisfies

$$\widehat{w}_i^{(j)} = (\widetilde{T}_{ij} + F_{ij})\widehat{x}_j, \quad \|F_{ij}\| \le \gamma_d\|\widetilde{T}_{ij}\|,$$

with $d = b + r^{3/2}$. Let $w_i = v_i - \sum_{j=1}^{i-1} w_i^{(j)}$; the computed $\widehat{w}_i$ satisfies

$$\widehat{w}_i = v_i(1+\theta_i) - \sum_{j=1}^{i-1} \widehat{w}_i^{(j)}(1+\theta_j), \quad |\theta_j| \le \gamma_p.$$

By Lemma 2.2, the computed solution $\widehat{x}_i$ to $T_{ii}x_i = \widehat{w}_i$ satisfies

$$(T_{ii} + F_{ii})\widehat{x}_i = \widehat{w}_i, \quad \|F_{ii}\| \le \gamma_b\|T_{ii}\|.$$

Therefore, recalling that $\widetilde{T}_{ii} = T_{ii}$, we have

$$\sum_{j=1}^{i} (\widetilde{T}_{ij} + \Delta\widetilde{T}_{ij})\widehat{x}_j = v_i(1+\theta_i), \tag{4.29}$$

with $\Delta\widetilde{T}_{ij} = \theta_j \widetilde{T}_{ij} + F_{ij}(1+\theta_j)$ and thus $\|\Delta\widetilde{T}_{ij}\| \le \gamma_{d+p+O(u)}\|\widetilde{T}_{ij}\|$. Gathering (4.29) over all block-rows i, we obtain $(\widetilde{T} + \Delta\widetilde{T})\widehat{x} = v + \Delta v$, with $\|\Delta v\| \le \gamma_p\|v\|$ and

$$\|\Delta\widetilde{T}\| \le \gamma_{d+p+O(u)} \Big(\sum_{i=1}^{p}\sum_{j=1}^{i} \|\widetilde{T}_{ij}\|^2 \Big)^{1/2} \le \gamma_{d+p}\|\widetilde{T}\| + O(u^2),$$

as required. ∎

We are ready for our final theorem, which builds upon all our previous analyses to prove the backward stability of the solution to linear systems by BLR LU factorization.

Theorem 4.8 (BLR linear system). *Let $\widetilde{A} \in \mathbb{R}^{n\times n}$ be a $pb \times pb$ BLR matrix and let $v \in \mathbb{R}^n$. If the linear system $\widetilde{A}x = v$ is solved by solving the triangular systems $\widetilde{L}y = v$, $\widetilde{U}x = y$, where $\widetilde{L}$ and $\widetilde{U}$ are the BLR LU factors computed by either Algorithm 4.1 or 4.2, then the computed solution $\widehat{x}$ satisfies*

$$\big(A + \Delta A\big)\widehat{x} = v + \Delta v, \tag{4.30}$$
$$\|\Delta A\| \le \big(\xi_p\varepsilon + \gamma_p\big)\|A\| + \gamma_{3c}\|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon), \tag{4.31}$$
$$\|\Delta v\| \le \gamma_p\big(\|v\| + \|\widetilde{L}\|\|\widetilde{U}\|\|\widehat{x}\|\big) + O(u^2), \tag{4.32}$$

where $c = b + 2r^{3/2} + p$, and where ξ_p is given in Table 4.1.

Proof. By Theorems 4.5 and 4.6, the BLR LU factors computed by the FCU algorithm or the CFU algorithm satisfy $A + \Delta A = \widetilde{L}\widetilde{U}$, with

$$\|\Delta A\| \le \big(\xi_p\varepsilon + \gamma_p\big)\|A\| + \gamma_c\|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon).$$

By Theorem 4.7, the computed $\widehat{y}$ satisfies

$$\big(\widetilde{L} + \Delta\widetilde{L}\big)\widehat{y} = v + \Delta v, \quad \|\Delta\widetilde{L}\| \le \gamma_c\|\widetilde{L}\|, \quad \|\Delta v\| \le \gamma_p\|v\|.$$

Similarly the computed $\widehat{x}$ satisfies

$$\big(\widetilde{U} + \Delta\widetilde{U}\big)\widehat{x} = \widehat{y} + \Delta\widehat{y}, \quad \|\Delta\widetilde{U}\| \le \gamma_c\|\widetilde{U}\|, \quad \|\Delta\widehat{y}\| \le \gamma_p\|\widehat{y}\|.$$

We therefore obtain on the one hand

$$\big(\widetilde{L} + \Delta\widetilde{L}\big)\big(\widetilde{U} + \Delta\widetilde{U}\big)\widehat{x} = \big(A + \Delta A + \Delta\widetilde{L}\widetilde{U} + \widetilde{L}\Delta\widetilde{U} + \Delta\widetilde{L}\Delta\widetilde{U}\big)\widehat{x} = \big(A + \Delta A'\big)\widehat{x}$$

and on the other hand

$$\big(\widetilde{L} + \Delta\widetilde{L}\big)\big(\widetilde{U} + \Delta\widetilde{U}\big)\widehat{x} = v + \Delta v + \widetilde{L}\Delta\widehat{y} + \Delta\widetilde{L}\Delta\widehat{y} = v + \Delta v',$$

yielding $\big(A + \Delta A'\big)\widehat{x} = v + \Delta v'$, with

$$\begin{aligned}
\|\Delta A'\| &= \|\Delta A + \Delta\widetilde{L}\widetilde{U} + \widetilde{L}\Delta\widetilde{U} + \Delta\widetilde{L}\Delta\widetilde{U}\| \\
&\le \big(\xi_p\varepsilon + \gamma_p\big)\|A\| + \gamma_{3c}\|\widetilde{L}\|\|\widetilde{U}\| + O(u\varepsilon), \\
\|\Delta v'\| &= \|\Delta v + \widetilde{L}\Delta\widehat{y} + \Delta\widetilde{L}\Delta\widehat{y}\| \le \gamma_p\big(\|v\| + \|\widetilde{L}\|\|\widetilde{U}\|\|\widehat{x}\|\big) + O(u^2).
\end{aligned}$$ ∎

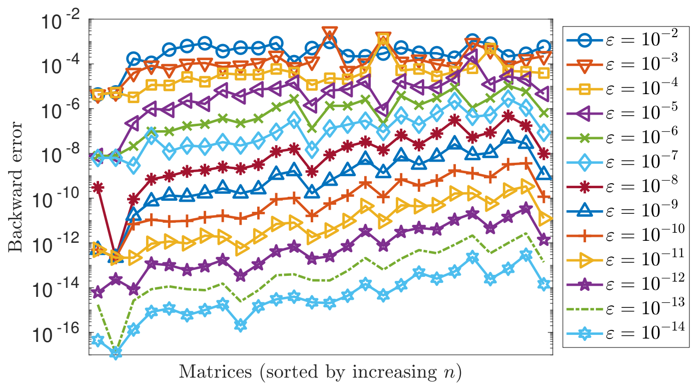


Fig. 5.1: Backward error (2.11) for solving a linear system by BLR LU factorization for the matrices from Table 5.1. We use the CFU algorithm (Algorithm 4.2) with a global threshold and with intermediate recompressions.

5. Additional experiments and discussion.

In the previous sections we have illustrated our analysis with numerical experiments performed on Poisson matrices (see (2.10)). In this final section, we provide some additional experiments to demonstrate that the conclusions drawn in the previous sections extend to many other kind of problems coming from various real-life applications. We use 26 root separators (Schur complements of sparse matrices) obtained from matrices from the SuiteSparse collection [13] listed in Table 5.1.

The main conclusions drawn from our analysis and experiments in the previous sections were the following.

- As predicted by our analysis, we have observed a tight correlation between the low-rank threshold ε and the backward error. We show that this crucial result remains true for a wide range of matrices in section 5.1.
- We experimentally determined using a Poisson matrix of relatively small order $n = 4096$ that the use of a global threshold, the CFU algorithm, and intermediate recompressions achieves a better flops–accuracy tradeoff than the use of a local threshold, the FCU algorithm, and no recompressions, respectively. In section 5.2, we first analyze using Poisson matrices how this comparison evolves as n increases. Then, in section 5.3, we extend these conclusions to a wide range of matrices.

5.1. Impact of ε for a wide range of matrices.

For each matrix A listed in Table 5.1, we solve a linear system $Ax = v$ via BLR LU factorization, using the CFU algorithm with a global threshold and intermediate recompressions. We report the backward error (2.11) in Figure 5.1, which shows that for all these matrices there is a good and often even excellent correlation between the threshold ε and the measured backward error.

The main conclusion of our analysis, which is that the backward error is directly determined by ε, is therefore confirmed experimentally for a wide range of matrices.

5.2. Flops–accuracy tradeoff for increasing n.

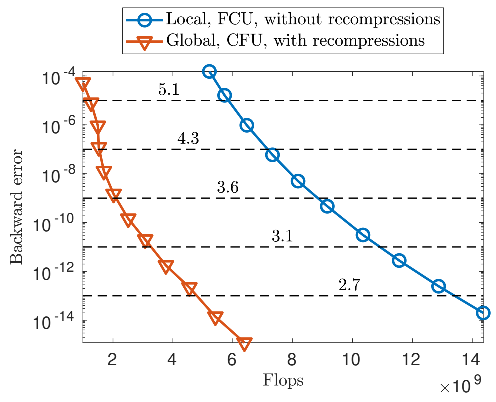
On the one hand, the use of a global threshold and intermediate recompressions both lead to a constant ξ_p larger by about a factor $p = n/b$, as shown in Table 4.1. On the other hand, intermediate

Table 5.1: List of matrices from the SuiteSparse collection [13]; n and p denote the order and the number of block-rows and block-columns of the root separator on which we perform the experiments.

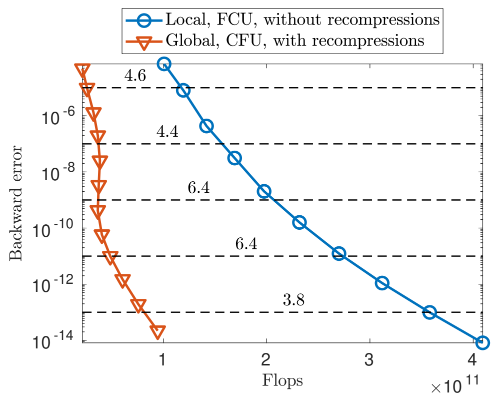
Matrix	n	p	Application
thermomech_dM	388	2	Thermal
shallow_water1	543	2	Computational fluid dynamics
Dubcova3	607	2	2D/3D
oilpan	644	3	Structural
hood	763	3	Structural
ship_003	990	4	Structural
pwtk	1080	4	Structural
af_shell3	1455	6	Structural
cfd2	1772	7	Computational fluid dynamics
x104	1992	8	Structural
Flan_1565	2916	11	Structural
ML_Geer	3006	12	Structural
2cubes_sphere	3325	13	Electromagnetics
audikw_1	3768	15	Structural
Transport	4991	19	Structural
Hook_1498	6414	25	Structural
nd24k	7785	30	2D/3D
Fault_639	7983	31	Structural
atmosmodd	8584	34	Computational fluid dynamics
Emilia_923	9843	38	Structural
Geo_1438	13254	52	Structural
nlpkkt80	14080	55	Optimization
Long_Coup_dt0	15448	60	Structural
Serena	15552	61	Structural
Cube_Coup_dt0	21072	82	Structural
cage13	21478	79	DNA electrophoresis

recompressions and the CFU algorithm both reduce the asymptotic complexity of the LU factorization [4]. Even a global threshold may provide an asymptotic improvement, because the proportion of blocks of small norm with respect to the norm of the global matrix increases with n.

It is therefore important to investigate whether the strategy achieving the best flops–accuracy tradeoff depends on n. Figure 5.2 compares the tradeoff achieved by two strategies: the first uses the FCU algorithm with a local threshold and without recompressions, whereas the second uses the CFU algorithm with a global threshold and with recompressions. We compare these strategies for two Poisson matrices of order $n = 4096$ and $n = 16384$. Not only is the second strategy the best choice for both matrices, but the gap between the two is larger for $n = 16384$ than for $n = 4096$. Indeed, for five different values of ε (from 10^{-13} to 10^{-3}, indicated by the dashed lines on the figure), we measure the flops required by each strategy and plot the ratio between the two. The figure shows that, except for $\varepsilon = 10^{-3}$, this ratio is larger for the larger matrix. Additional experiments (not shown) on matrices of intermediate



(a) $n = 4096$.



(b) $n = 16384$.

Fig. 5.2: Flops–accuracy tradeoff for two opposite strategies for the solution of a BLR linear system, using two Poisson matrices of different orders. The numbers indicate the ratio between the flops required by the two strategies for a given cutoff value of ε (corresponding to the dashed lines).

order between 4096 and 16384 show that this ratio gradually increases with n. We conclude that the second strategy becomes more and more beneficial with respect to the first strategy as n increases.

This experimental observation may in fact be justified theoretically for some classes of matrices. For instance, for Poisson matrices, the ranks of the blocks are known to be logarithmically dependent on the threshold ε, that is, $r = O(\log 1/\varepsilon)$ [8,9]. The impact of a larger ξ_p on the error can be compensated by simply using a smaller threshold $\varepsilon' = \varepsilon/\xi_p$, which in turn yields a larger rank $r' = O(\log \xi_p/\varepsilon)$. Therefore, as $p = n/b$ increases, compensating for the error increase due to a larger ξ_p only increases the cost of the factorization by logarithmic factors of n. Since a global threshold, intermediate recompressions, and the CFU algorithm all reduce this cost by factors $O(n^\alpha)$, with $\alpha > 0$, we conclude that the use of these strategies must eventually become beneficial for large enough n.

5.3. Flops–accuracy tradeoff for a wide range of matrices. We finally compare the flops–accuracy tradeoff achieved by global and local thresholds, the FCU and CFU algorithms, and the use of intermediate recompressions on the set of matrices listed in Table 5.1. We compare the eight possible strategies depending on the combination of parameters.

For each matrix, we run each strategy multiple times with slightly different values of ε between 10^{-7} and 10^{-9}. We then select a cutoff value and choose for each strategy the largest ε producing a backward error smaller than this cutoff. This is done to guarantee that all eight strategies achieve roughly the same backward error. We can then measure the number of flops required by each strategy and plot the result as a performance profile in Figure 5.3.

As the figure shows, the strategy using the CFU algorithm with a global threshold and with recompressions achieves the best tradeoff for *all* 26 matrices, thereby confirming our previous observations that this is the best parameter setting. This

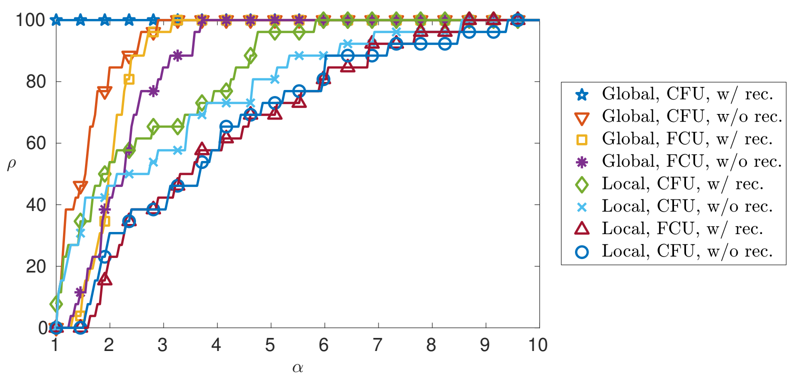


Fig. 5.3: Performance profile of the eight possible strategies to solve a BLR linear system over the matrices listed in Table 5.1. ρ (y-axis) indicates the percentage of matrices for which a given strategy requires less than α (x-axis) times the flops required by the best strategy. For any given matrix, all eight strategies achieve roughly the same backward error.

strategy can reduce the number of flops by a factor up to 10 with no loss of accuracy. The performance of the remaining seven strategies gives some indication of the relative importance of each parameter: using a global threshold has the greatest impact, followed by the CFU algorithm, and finally the intermediate recompressions.

6. Conclusions.

We have analyzed the errors introduced in various matrix algorithms by the use of block low-rank (BLR) approximations. Our analysis provides important new theoretical guarantees, as well as some new insights into the numerical behavior of BLR matrix algorithms in floating-point arithmetic. We now gather and summarize our main conclusions.

6.1. Summary.

We have derived a set of normwise error bounds that share a common point: they are expressed as the sum of two terms: one term is associated with the low-rank truncation errors, whose magnitude can be controlled via the low-rank threshold ε; the other term is associated with the floating-point errors, whose magnitude depends on the unit roundoff u.

Usually, we have $u \ll \varepsilon$, and therefore the error is mainly determined by ε. In particular, we have proved in Theorem 4.8 that BLR linear systems $Ax = v$ can be solved with a backward error proportional to $\xi_p \varepsilon \|A\|$, where ξ_p is a small constant growing at most quadratically with the number of block-rows and block-columns $p = n/b$. Our analysis therefore proves for the first time the backward stability of the solution of BLR linear systems and provides a theoretical justification for the empirical observation that the backward error is closely related to the low-rank threshold. Users can therefore control the numerical behavior of BLR solvers simply by setting ε to the target accuracy.

When $u \ll \varepsilon$, the unit roundoff has only a limited impact. This remark is of particular relevance in the context where BLR solvers are used as preconditioners for iterative methods [20], for which the low-rank threshold may be set to very large values (such as $\varepsilon = 0.01$ or even 0.1). In this setting, our analysis indicates that the use of low precisions, such as half precision, should be very attractive; see [22] and the references therein for details of half precision and how it can be exploited in standard

LU factorization.

We have analyzed several key parameters in the BLR LU factorization and assessed how to choose them to obtain the best possible tradeoff between flops and accuracy. First, we have shown that the use of a global threshold (block A_{ij} is compressed such that $\|A_{ij} - \widetilde{A}_{ij}\| \le \varepsilon \|A\|$) should be preferred to that of a local one ($\|A_{ij} - \widetilde{A}_{ij}\| \le \varepsilon \|A_{ij}\|$). Second, the use of intermediate recompressions in the update step (so-called LUAR strategy in [4]) only impacts the constant in the error bound and is therefore recommended. Finally, we have compared two different factorization variants, the FCU and CFU algorithms, which differ in when the BLR compression is incorporated in the LU algorithm, and we have shown that they yield similar error bounds; the CFU algorithm, which achieves the best complexity, should therefore be preferred.

We have supported all of these conclusions with numerical experiments on a wide range of matrices from various real-life applications.

6.2. Perspectives. There exist numerous structured matrix representations other than BLR, such as multilevel [6] and hierarchical [16] representations, block-separable matrices [15], HSS matrices [35], and so on. Our analysis could be extended to these other type of matrices (we note the existing work regarding HSS matrices [34]), although we expect that the resulting analyses would yield similar results.

If the threshold ε is chosen too close to the unit roundoff u, Assumption 2.4 no longer holds and we are unable to accurately detect the numerical rank of the blocks, which dramatically increases the cost of the factorization. In order to avoid this issue, which will become of growing importance with the rise of low precisions, we are forced to use a higher precision. In this context, recent advances that reduce rounding error accumulation [10, 11] can help decrease the threshold of ε at which we are forced to switch to a higher precision.

REFERENCES

[1] Kadir Akbudak, Hatem Ltaief, Aleksandr Mikhalev, and David Keyes. Tile low rank Cholesky factorization for climate/weather modeling applications on manycore architectures. In *High Performance Computing*, Julian M. Kunkel, Rio Yokota, Pavan Balaji, and David Keyes, editors, Cham, 2017, pages 22–40. Springer International Publishing.

[2] P. R. Amestoy, R. Brossier, A. Buttari, J.-Y. L'Excellent, T. Mary, L. Métivier, A. Miniussi, and S. Operto. Fast 3D frequency-domain full waveform inversion with a parallel Block Low-Rank multifrontal direct solver: application to OBC data from the North Sea. *Geophysics*, 81(6):R363–R383, 2016.

[3] Patrick Amestoy, Cleve Ashcraft, Olivier Boiteau, Alfredo Buttari, Jean-Yves L'Excellent, and Clément Weisbecker. Improving multifrontal methods by means of block low-rank representations. *SIAM J. Sci. Comput.*, 37(3):A1451–A1474, 2015.

[4] Patrick Amestoy, Alfredo Buttari, Jean-Yves L'Excellent, and Theo Mary. On the complexity of the block low-rank multifrontal factorization. *SIAM J. Sci. Comput.*, 39(4):A1710–A1740, 2017.

[5] Patrick R. Amestoy, Alfredo Buttari, Jean-Yves L'Excellent, and Theo Mary. Performance and scalability of the block low-rank multifrontal factorization on multicore architectures. *ACM Trans. Math. Software*, 2018. Accepted for publication.

[6] Patrick R. Amestoy, Alfredo Buttari, Jean-Yves L'Excellent, and Theo Mary. Bridging the gap between flat and hierarchical low-rank matrix formats: The multilevel block low-rank format. *SIAM J. Sci. Comput.*, 41(3):A1414–A1442, 2019.

[7] Julie Anton, Cleve Ashcraft, and Clément Weisbecker. A Block Low-Rank multithreaded factorization for dense BEM operators. In *SIAM Conference on Parallel Processing (SIAM PP16)*, Paris, France, April 2016.

[8] Mario Bebendorf. *Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems*, volume 63 of *Lecture Notes in Computational Science and Engineering (LNCSE)*. Springer-Verlag, 2008. ISBN ISBN 978-3-540-77146-3.

[9] Mario Bebendorf and Wolfgang Hackbusch. Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. *Numerische Mathematik*, 95(1):1–28, 2003.

[10] Pierre Blanchard, Nicholas J. Higham, Florent Lopez, Theo Mary, and Srikara Pranesh. Mixed precision block fused multiply-add: Error analysis and application to GPU tensor cores. 2019. In preparation.

[11] Pierre Blanchard, Nicholas J. Higham, and Theo Mary. A class of fast and accurate summation algorithms. MIMS EPrint 2019.6, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, April 2019. 16 pp.

[12] Ali Charara, David Keyes, and Hatem Ltaief. Tile low-rank GEMM using batched operations on GPUs. In *Euro-Par 2018: Parallel Processing*, Marco Aldinucci, Luca Padovani, and Massimo Torquati, editors, Cham, 2018, pages 811–825. Springer International Publishing.

[13] Timothy A. Davis and Yifan Hu. The University of Florida Sparse Matrix Collection. *ACM Trans. Math. Software*, 38(1):1:1–1:25, 2011.

[14] J. A. George. Nested dissection of a regular finite-element mesh. *SIAM J. Numer. Anal.*, 10 (2):345–363, 1973.

[15] Adrianna Gillman, Patrick M. Young, and Per-Gunnar Martinsson. A direct solver with $O(N)$ complexity for integral equations on one-dimensional domains. *Frontiers of Mathematics in China*, 7(2):217–247, 2012.

[16] Wolfgang Hackbusch. *Hierarchical Matrices : Algorithms and Analysis*, volume 49 of *Springer series in computational mathematics*. Springer, Berlin, 2015. xxv, 511 pp.

[17] Nicholas J. Higham. Exploiting fast matrix multiplication within the level 3 BLAS. *ACM Trans. Math. Software*, 16(4):352–368, 1990.

[18] Nicholas J. Higham. Stability of a method for multiplying complex matrices with three real matrix multiplications. *SIAM J. Matrix Anal. Appl.*, 13(3):681–687, 1992.

[19] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. xxx+680 pp. ISBN 0-89871-521-0.

[20] Nicholas J. Higham and Theo Mary. A new preconditioner that exploits low-rank approximations to factorization error. *SIAM J. Sci. Comput.*, 41(1):A59–A82, 2019.

[21] Nicholas J. Higham and Srikara Pranesh. Simulating low precision floating-point arithmetic. MIMS EPrint 2019.4, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, March 2019. 18 pp. Revised July 2019. To appear in SIAM J. Sci. Comput.

[22] Nicholas J. Higham, Srikara Pranesh, and Mawussi Zounon. Squeezing a matrix into half precision, with an application to solving linear systems. *SIAM J. Sci. Comput.*, 41(4): A2536–A2551, 2019.

[23] Akihiro Ida, Hiroshi Nakashima, and Masatoshi Kawai. Parallel hierarchical matrices with block low-rank representation on distributed memory computer systems. In *Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region*, ACM, 2018, pages 232–240.

[24] C.-P. Jeannerod, T. Mary, C. Pernet, and D. Roche. Exploiting fast matrix arithmetic in block low-rank factorizations. *SIAM J. Matrix Anal. Appl.*, 2019. Submitted.

[25] Théo Mary. *Block Low-Rank Multifrontal Solvers: Complexity, Performance, and Scalability*. PhD thesis, Université de Toulouse, Toulouse, France, November 2017.

[26] Satoshi Ohshima, Ichitaro Yamazaki, Akihiro Ida, and Rio Yokota. Optimization of hierarchical matrix computation on GPU. In *Supercomputing Frontiers*, Rio Yokota and Weigang Wu, editors, Springer International Publishing, Cham, 2018, pages 274–292.

[27] G. Pichon. *On the use of low-rank arithmetic to reduce the complexity of parallel sparse linear solvers based on direct factorization techniques*. Ph.D. thesis, Université de Bordeaux, November 2018.

[28] Grégoire Pichon, Eric Darve, Mathieu Faverge, Pierre Ramet, and Jean Roman. Sparse supernodal solver using block low-rank compression: Design, performance and analysis. *Journal of Computational Science*, 27:255–270, 2018.

[29] J. L. Rigal and J. Gaches. On the compatibility of a given solution with the data of a linear system. *J. Assoc. Comput. Mach.*, 14(3):543–548, 1967.

[30] M. Sergent, D. Goudin, S. Thibault, and O. Aumage. Controlling the memory subscription of distributed applications with a task-based runtime system. In *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, volume 00, May 2016, pages 318–327.

[31] D. V. Shantsev, P. Jaysaval, S. de la Kethulle de Ryhove, P. R. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. Mary. Large-scale 3D EM modeling with a Block Low-Rank multifrontal direct solver. *Geophys. J. Int.*, 209(3):1558–1571, 2017.

[32] Volker Strassen. Gaussian elimination is not optimal. *Numerische mathematik*, 13(4):354–356, 1969.

[33] Clément Weisbecker. *Improving multifrontal solvers by means of algebraic block low-rank representations*. PhD thesis, Institut National Polytechnique de Toulouse, October 2013.

[34] Y. Xi and J. Xia. On the stability of some hierarchical rank structured matrix algorithms. *SIAM J. Matrix Anal. Appl.*, 37(3):1279–1303, 2016.

[35] Jianlin Xia, Shivkumar Chandrasekaran, Ming Gu, and Xiaoye S. Li. Fast algorithms for hierarchically semiseparable matrices. *Numer. Linear Algebra Appl.*, 17(6):953–976, 2010.