# Mixed precision HODLR matrices

Xiaobo Liu[1], Erin Carson[2], Xinye Chen[2]

Abstract

**Introduction and overview**. Hierarchical matrices, often abbreviated as $\mathcal{H}$-matrices [1], comprise a class of dense rank-structured matrices with a hierarchical low-rank structure, which is used to approximate a dense or sparse matrix by dividing it into multiple submatrices in a hierarchical way, where a number of submatrices are selected to be approximated by low-rank factors according to an admissibility condition.

Computations of hierarchical matrices have attracted significant attention in the science and engineering community as exploiting data-sparse structures can significantly reduce the computational complexity of many important kernels such as matrix–vector products, matrix factorizations, etc. One particularly popular option within this class is the *Hierarchical Off-Diagonal Low-Rank* (HODLR) format, whose definition is associated with the binary cluster tree $\mathcal{T}_\ell$ of depth $\ell \in \mathbb{N}^+$ [4].

**Definition 1** ($(\mathcal{T}_\ell, p)$-HODLR matrix)**.** *$H \in \mathbb{R}^{n\times n}$ is $(\mathcal{T}_\ell, p)$-HODLR matrix if every off-diagonal block $H(I_i^k, I_j^k)$ associated with siblings $I_i^k$ and $I_j^k$ in $\mathcal{T}_\ell$, $k = 1, \ldots, \ell$, has rank at most $p$.*

In the proposed talk, we consider constructing HODLR matrices in a mixed precision manner and offer insights into the resulting behavior of finite precision computations. Our analysis confirms what is largely intuitive: the lower the quality of the low-rank approximation, the lower the precision which can be used without detriment. We provide theoretical bounds which determine which precisions can safely be used in order to balance the overall error.

**Practical definition of HODLR matrix.** In order to quantify the error incurred in the low-rank factorization of the off-diagonal blocks, we introduce the practical definition of $(\mathcal{T}_\ell, p, \varepsilon)$-HODLR matrix as in Definition 2. The approximation error in the diagonal blocks of all levels of the $(\mathcal{T}_\ell, p, \varepsilon)$-HODLR matrix $\widetilde{H}$ is immediately obtainable following Definition 2 in the Frobenius norm, and, as a special case, one can show $\|\widetilde{H} - H\|_F \le \varepsilon \|H\|_F$.

**Definition 2** ($(\mathcal{T}_\ell, p, \varepsilon)$-HODLR matrix)**.** *Let $H \in \mathbb{R}^{n\times n}$ be a $(\mathcal{T}_\ell, p)$-HODLR matrix. Then $\widetilde{H} \in \mathbb{R}^{n\times n}$ is defined to be a $(\mathcal{T}_\ell, p, \varepsilon)$-HODLR matrix to $H$, if every off-diagonal block $\widetilde{H}(I_i^k, I_j^k)$ associated with siblings $I_i^k$ and $I_j^k$ in $\mathcal{T}_\ell$, $k = 1, \ldots, \ell$, satisfies $\|\widetilde{H}(I_i^k, I_j^k) - H(I_i^k, I_j^k)\| \le \varepsilon \|H(I_i^k, I_j^k)\|$, where $0 \le \varepsilon < 1$.*

**Mixed-precision representation.** First, we develop a mixed precision algorithm for constructing HODLR matrices. Let us assume that the off-diagonal blocks from the $k$th level of $\widetilde{H}$, $1 \le k \le \ell$, are compressed in the form

$$\widetilde{H}_{ij}^{(k)} = \widetilde{U}_i^{(k)} (\widetilde{V}_j^{(k)})^T, \quad |i - j| = 1, \tag{1}$$

where $\widetilde{U}_i^{(k)} \in \mathbb{R}^{n/2^k \times p}$ has orthonormal columns to precision $u$ and $\widetilde{V}_j^{(k)} \in \mathbb{R}^{n/2^k \times p}$. Our idea is to compress the low-rank blocks $\widetilde{H}_{ij}^{(k)}$ and represent the low-rank factors $\widetilde{U}_i^{(k)}$ and $\widetilde{V}_j^{(k)}$ in precisions potentially lower than the working precision; given a set of available precisions, the same precision, say, $u_k$, is used for the storage of all low-rank factors at level $k$. To keep the global error in the

[1]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.
[2]Department of Numerical Mathematics, Charles University, Prague, Czech Republic.

mixed-precision representation at the same level as an unified working-precision representation, we choose

$$u_k \le \varepsilon/(2^{k/2}\xi_k),$$

where $\varepsilon > u$ (since the factorizations are calculated in the working precision $u$) can be thought of as the accuracy threshold in the low-rank factorizations (1) and

$$\xi_k := \max_{|i-j|=1} \|\widetilde{H}_{ij}^{(k)}\|_F / \|\widetilde{H}\|_F, \quad 1 \le k \le \ell,$$

which essentially characterizes the relative importance of the off-diagonal blocks in level-$k$ to the whole matrix in terms of magnitude. This means that, as the tree depth increases, the unit roundoff $u_k$ must be smaller to offset the error between the HODLR matrix and the original matrix and that, since $0 < \xi_k < 1$ holds for $k = 1\colon \ell$, generally no higher-than-working precisions are needed among $u_k$ for a HODLR matrix with mild depth $\ell$, say, $\ell \le 10$ (so $2^{k/2} \le 32$). We then propose an adaptive scheme for precision selection, which dynamically determines what degree of precision is required for the computations in each level of the cluster tree. We show that the error in the resulting mixed-precision representation $\widehat{H}$ satisfies

$$\|H - \widehat{H}\|_F \lesssim (2\sqrt{2\ell} + 1)\varepsilon \|H\|_F.$$

**Matrix–vector products.** Next, we give error bounds on the working precision $u$ so that the backward error in computing the matrix–vector product in finite precision does not exceed the error resulting from inexact representation of the matrix. The key idea is that, if the HODLR matrix $H$ is approximated by the mixed-precision representation $\widehat{H}$, to calculate the matrix–vector product $b \leftarrow \widehat{H}x$ we should try to balance the errors occurring in the approximation of $\widehat{H}$ and in the finite-precision computation, as shown from the following result.

**Lemma 1.** *Let $\widehat{A}_p$ an approximation of $A$ such that $\|A - \widehat{A}_p\|_F \lesssim \eta$ for some $\eta > 0$. Then the error due to finite precision computation of $\widehat{y} = \mathrm{fl}(\widehat{A}_p x)$ will be no larger than the error due to the computed inexact representation when the working precision has unit roundoff $u \le \eta/(n\|\widehat{A}_p\|_F)$.*

Applying Lemma 1 to the computation of the matrix–vector product associated with $\widehat{H}_{ij}^{(k)}$ and ignoring the errors in the summation of the vector elements (which are usually negligible compared with the error in the block matrix–vector products), we can obtain the following result.

**Theorem 1.** *Let $\widetilde{H}$ be a $(\mathcal{T}_\ell, p, \varepsilon)$-HODLR matrix associated with the HODLR matrix $H$, and let $\widehat{H}$ denote our mixed-precision representation. If $b = \widehat{H}x$ is computed in a working $u \le \varepsilon/n$, then the computed $\widehat{b}$ satisfies*

$$\widehat{b} = \mathrm{fl}(\widehat{H}x) = (H + \Delta H)x, \quad \|\Delta H\|_F \le 10 \cdot 2^{\ell/2}\varepsilon \|H\|_F.$$

**LU factorization.** Finally, we derive error bounds on the LU factorization of the mixed-precision HODLR matrix $\widehat{H}$. The factorization is done by a recursive algorithm which computes for all but the bottom level the block LU factorization

$$\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ & U_{22} \end{bmatrix},$$

where $L_{11}$ and $L_{22}$ are lower triangular and $U_{11}$ and $U_{22}$ are upper triangular, and it invokes dense routines on the bottom level. Based on the results from [3, sect. 3.5] and [3, Thm. 8.5], we first look at the backward error in the LU factorization of the HODLR matrices at level $k = \ell - 1$ and then use induction to quantify the backward error in the LU decomposition of diagonal blocks in the other levels, up to the level $k = 0$ ($H_{11}^{(0)} := H$). We arrive at the following result.

**Theorem 2.** *Let $\widehat{H}$ be the mixed-precision $\ell$-level HODLR representation. If the LU decomposition of $\widehat{H}$ is computed in a working precision $u \lesssim \varepsilon/n$, then the LU factorization of the HODLR matrix $\widehat{H}$ satisfies*

$$\widehat{L}\widehat{U} = H + \Delta H, \quad \|\Delta H\|_F \lesssim 2^{\ell+1}\varepsilon\|H\|_F + 11 \cdot 2^{\ell}\varepsilon\|\widehat{L}\|_F\|\widehat{U}\|_F.$$

Noted that our finite precision analysis remains valid in the case where the HODLR matrices are stored in one precision and therefore also provides new results for this case. We will also present the numerical simulations we performed across various datasets to verify our theoretical results.

The talk is based on [2]. We have also developed a MATLAB toolbox called `mhodlr` for matrix computations with HODLR representation and mixed-precision simulations, which supports other important operations within the class of HODLR matrix such as (mixed-precision) matrix multiplication and Cholesky factorization. The documentation webpage of `mhodlr` MATLAB toolbox is at `https://mhodlr.readthedocs.io/en/latest/index.html`.

## References


[1] Steffen Börm, Lars Grasedyck, and Wolfgang Hackbusch. Introduction to hierarchical matrices with applications. *Eng. Anal. Bound. Elem.*, 27(5):405–422, 2003.

[2] Erin Carson, Xinye Chen, and Xiaobo Liu. Mixed precision HODLR matrices. ArXiv:2407.21637 [math.NA], July 2024.

[3] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, Philadelphia, PA, USA, second edition, 2002.

[4] Stefano Massei, Leonardo Robol, and Daniel Kressner. hm-toolbox: MATLAB software for HODLR and HSS matrices. *SIAM J. Sci. Comput.*, 42(2):C43–C68, 2020.