Electronic Transactions on Numerical Analysis.
Volume 65, pp. 579–604, 2026.
© 2026, the Author(s), licensed via CC BY 4.0.
ISSN 1068–9613.
DOI: 10.1553/etna_vol65s579

ETNA
Kent State University and
Johann Radon Institute (RICAM)

# MULTIPRECISION COMPUTATIONS WITH SCHWARZ METHODS*

MICHAL OUTRATA† AND DANIEL B. SZYLD‡

**Abstract.** We explore and analyze the use of multiprecision arithmetic for several classes of Schwarz methods and preconditioners where the approximate solution of the local problems is performed at a lower precision, i.e., with fewer digits of accuracy than in the underlying (double-precision) computation. Conditions for the appropriate round-off criteria for the lower precision are presented. It is found experimentally that for the model problems, about five digits of accuracy are sufficient to achieve the theoretical restrictions, and thus, single precision suffices for the local solves. Several numerical experiments illustrate the obtained results.

**Key words.** classical Schwarz methods, multiprecision computation, M-matrices

**AMS subject classifications.** 65F10

## 1. Introduction.

We consider the solution of systems of linear algebraic equations of the form

$$A\mathbf{u} = \mathbf{f}, \qquad A \in \mathbb{R}^{N\times N}, \mathbf{f} \in \mathbb{R}^{N}. \tag{1.1}$$

In particular, we analyze the use of multiprecision arithmetic for three types of Schwarz methods (see [22]): (i) the *(damped) additive Schwarz method* ((d)AS; see [38, 53]), (ii) the *restricted additive Schwarz method* (RAS; see [9]), and (iii) the *multiplicative Schwarz method* (MS; see [38]). See [32] for an overview of multiprecision/mixed-precision algorithms. Specifically, we study the solution of the local problems (see the precise definitions below) using lower precision. To the best of our knowledge, this is the first time that this approach has been analyzed for Schwarz methods.

The underlying idea of Schwarz methods, as part of the wider family of domain decomposition methods, can be summarized as "divide and conquer", where the solution of a large problem is approximated by subdividing it into many smaller ones that are computationally less demanding than (1.1); these are called the *local* problems (or subproblems or subdomain problems); see Section 2 for a detailed discussion. For matrices obtained by the discretization of a partial differential equation (PDE), the convergence analysis usually focuses on studying the spectral information of the iteration operator. When the method is used as a preconditioner, the convergence analysis usually uses the continuous PDE and its discretization, showing a convergence bound independent of the discretization parameter; see, e.g., [16, 48, 52]. For the algebraic error analysis of the broader class of stationary iterative methods, we refer the reader to [31, Chapter 17] and the references therein.

For *algebraic Schwarz methods*, where the analysis does not take advantage of the provenance of the system matrix, we are usually satisfied with information about the asymptotic convergence factor of the method (see, e.g., [4, 18, 19]), whereas more complete spectral information is often available once we couple the system matrix with further information about its origin, leading to a more complete understanding of the method's behavior; see, e.g., [13, 14, 15, 21, 26, 27]. Importantly, these methods are usually used as *preconditioners*,

*Received September 26, 2025. Accepted September 29, 2026. Published online on November 9, 2026. Recommended by Andreas Frommer. This work was supported by the grants UNCE/24/SCI/005 and PRIMUS/25/SCI/022 of Charles University.

†Corresponding author. Department of Numerical Mathematics, Charles University, Prague, Czech Republic (outrata@karlin.mff.cuni.cz).

‡Department of Mathematics, Temple University, Philadelphia, PA 19122, USA (szyld@temple.edu).

This work is published by ETNA and licensed under the Creative Commons license CC BY 4.0.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



i.e., their convergence is further accelerated using Krylov subspace methods. But in our experience, in order to obtain more insight, it is often very useful to *first* study the Schwarz methods (or other domain decomposition methods) *as stand-alone solvers*. Then, based on their analysis, we can obtain insight into or an estimate of the type of performance we can expect when we accelerate these methods with a Krylov subspace method. Moreover, in this way we often get additional insights into the weak points of the method, which can then be used to propose an improvement such as a coarse space; see, e.g., [13, 23] and also [45]. Schwarz methods as solvers are also fundamental for Schwarz asynchronous iterations; see, e.g., [29, 39].

The goal of multiprecision algorithms is, in general, to reduce the computation, communication, and memory costs by working with some portion of the problem/algorithm in lower precision, e.g., replacing standard double-precision data representation with a single-precision or even half-precision representation in the computationally most challenging part of the algorithm; see, e.g., [32] and the references therein. Intuitively, working (partially) in a lower precision can introduce new issues, e.g., numerical error propagation, and thus introduces a trade-off between computational complexity (by virtue of lowering the precision) and the level of approximation difficulty (e.g., the floating-point precision used). However, as shown in [12], by using multiprecision algorithms, where different parts of the algorithm are carried out in different precisions, it is sometimes possible to get the best of both worlds.

To fix ideas, let us consider having a *working* precision (say, double precision), and denote its *unit round-off* by $u_w$ (say, $u_w \approx 10^{-16}$); see [31, p. 3]. The natural first step is to consider the multiprecision Schwarz methods where the *local* problems are solved in a "lower precision", i.e., in a precision with a unit round-off $u_\ell$ such that $u_\ell > u_w$ (we shall identify the *precision* with its unit round-off, e.g., by "precision $u_\ell$ is lower than $u_w$" we mean $u_\ell > u_w$). Importantly, working in the lower precision $u_\ell$ limits not only the precision of the computation but also its *range*, i.e., without careful scaling, small/large numbers that we can represent in precision $u_w$ may underflow/overflow in precision $u_\ell$; see, e.g., [34]. We comment on the specifics of multiprecision computations relevant to our interest in Section 3.

We note that similar settings have already been considered. In [28], where the authors use the (nonoverlapping) AS as a preconditioner for the conjugate gradient method (CG), they use single precision for the preconditioner solve (i.e., running the AS in single precision) and the rest of the CG method in double precision. In [2], the authors take the (nonoverlapping) block-Jacobi method as a preconditioner for CG, and, based on the 1-norm condition number of each of the diagonal blocks, calculate their inverses in half, single, or double precision and then apply these using dense mat-vec products in parallel for each block; a similar methodology has also been used in [51] and tested for practically relevant and challenging 2D and 3D problems. Similarly, in [46], the authors study (overlapping, with coarse space) AS as preconditioner and the effect of using different precisions and data formats for the subdomain matrices (fixed vs. floating-point precision as well as dense storage of the inverse vs. storage of the Cholesky factor) on the performance of the preconditioned CG method. Note that the focus in all of these papers is on the numerical experiments and observations *about the preconditioned CG method*; that is, the interaction of the domain decomposition method and the different precision choices is present *only implicitly*. The analysis focuses on the (often questionable; see [37, Section 5.6, Corollary 5.6.7 and onward]) condition number bound for CG for the preconditioned system and is not concerned with the domain decomposition method of choice as a stand-alone solver.

In this paper we focus on multiprecision Schwarz methods with a lower precision $u_\ell$ for the local solves, treating both $u_\ell$ and the rounding routine as free variables that can and should be chosen so as to preserve or even improve the convergence of the Schwarz method. To that end, we propose specific rounding routines, derive sufficient conditions for the convergence of

ETNA
Kent State University and
Johann Radon Institute (RICAM)



the resulting multiprecision Schwarz methods, and numerically demonstrate their effectiveness. Later in the paper, we also consider the effectiveness of these methods as preconditioners, i.e., with the Krylov subspace method acceleration.

The rest of the manuscript is organized as follows: in Sections 2 and 3, we give a brief introduction to (algebraic) Schwarz methods and to multiprecision computations. In Section 4, we introduce and analyze the multiprecision Schwarz methods and demonstrate their performance for several model problems. In Section 5, we then explore some additional avenues for the analysis of multiprecision Schwarz methods, and we conclude with some remarks in Section 6.

**2. Algebraic Schwarz methods.** Consider a domain $\Omega$ on which we wish to solve a linear differential equation, and assume that the problem (1.1) corresponds to a discretization on some grid of the domain $\Omega$; this induces the vector space of the unknowns $\overline{W} \cong \mathbb{R}^N$. We can partition these unknowns into groups, obtaining vector spaces $\overline{W}_i, i = 1, \ldots, p$, each of dimension $\overline{N}_i := \dim(\overline{W}_i) < N$. We require these groupings to be *pairwise disjoint*, i.e., $\overline{W}_i \cap \overline{W}_j = \{0\}$ if $i \neq j$, when embedding these into $\mathbb{R}^N$, and *covering*, i.e.,

$$\overline{W} = \sum_{i=1}^{p} \overline{W}_i = \left\{ \mathbf{w} \,|\, \mathbf{w} = \sum_{i=1}^{p} \mathbf{w}_i \text{ for some } \mathbf{w}_i \in \overline{W}_i \right\}.$$

This partition is usually made by exploiting a relation that associates the variables of the discrete problem with the underlying geometry of $\Omega$ so that unknowns corresponding to some $\overline{W}_i$ belong to a single, connected subdomain $\overline{\Omega}_i$ of the physical domain $\Omega$; in such a case, we have $\Omega = \bigcup_{i=1}^{p} \overline{\Omega}_i$. We set the restriction operators $\overline{R}_i : \mathbb{R}^N \to \mathbb{R}^{\overline{N}_i}$, corresponding to $\overline{N}_i$-by-$N$ zero-one matrices with full row rank $\overline{N}_i$, and obtain the prolongation operators as the transposes of the restrictions, i.e., $\overline{R}_i^T : \mathbb{R}^{\overline{N}_i} \to \mathbb{R}^N$. We assume that the restriction matrices are chosen so that

$$\overline{R}_i = \begin{bmatrix} I_{\overline{N}_i} & 0 \end{bmatrix} \Pi_i \in \mathbb{R}^{\overline{N}_i \times N}, \tag{2.1}$$

where $I_{\overline{N}_i}$ is the identity matrix of dimension $\overline{N}_i$, $0 \in \mathbb{R}^{\overline{N}_i \times (N - \overline{N}_i)}$, and $\Pi_i \in \mathbb{R}^{N \times N}$ is a permutation matrix. Note that by composing the prolongation and restriction, we have

$$\overline{R}_i^T \overline{R}_i = \Pi_i^T \begin{bmatrix} I_{\overline{N}_i} & 0 \\ 0 & 0 \end{bmatrix} \Pi_i \in \mathbb{R}^{N \times N}.$$

We also consider a *covering* of $\Omega$ with *overlapping* subdomains by enlarging each subdomain and obtaining $\Omega_i \supset \bar{\Omega}_i$, $i = 1, \ldots, p$. The corresponding subspaces are $W_i \supset \overline{W}_i$, and we set $N_i := \dim(W_i) \geq \overline{N}_i$, with $\sum_{i=1}^{p} N_i > N$. The corresponding matrices are

$$R_i = \begin{bmatrix} I_{N_i} 0 \end{bmatrix} \Pi_i \in \mathbb{R}^{N_i \times N} \in \mathbb{R}^{N_i \times N} \qquad \text{and} \qquad R_i^T R_i = \Pi_i^T \begin{bmatrix} I_{N_i} & 0 \\ 0 & 0 \end{bmatrix} \Pi_i \in \mathbb{R}^{N \times N};$$

see, e.g., [9, 17, 19] for more details. Define also the complementary subspaces $\Omega_{\neg i} = \Omega \setminus \Omega_i$, and the corresponding complementary sets of variables $W_{\neg i}$.

Furthermore, we define the subdomain matrices $A_i$ as the restriction of $A$ to $W_i$, i.e., $A_i := R_i A R_i^T$, and similarly, $A_{\neg i} = R_{\neg i} A R_{\neg i}^T$. We then have

$$A = \Pi_i \begin{bmatrix} A_i & K_i \\ L_i & A_{\neg i} \end{bmatrix} \Pi_i^T \in \mathbb{R}^{N \times N}, \qquad \text{for } i = 1 \ldots, p.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



We then set up the multi-splitting matrices $M_i$ as

$$M_i = \Pi_i \begin{bmatrix} A_i & 0 \\ \times & * \end{bmatrix} \Pi_i^T \in \mathbb{R}^{N\times N}, \qquad \text{for } i = 1\ldots, p,$$

where the blocks $\times$ and $*$ can, in general, be chosen arbitrarily, but the common choice is to set $\times = 0$ and $* = A_{\neg i}$ [18] or $* = \operatorname{diag}(A_{\neg i})$ [4, 19]; for more details on multi-splittings, see the cited works and the references therein.

Equipped with this notation, we can formulate the classical algebraic Schwarz methods—AS, RAS, and MS—in matrix form as

$$\mathbf{u}^{(n+1)} = T_\star \mathbf{u}^{(n)} + \mathbf{c}_\star, \qquad \text{or, equivalently, for the errors,} \quad \mathbf{e}^{(n+1)} = T_\star \mathbf{e}^{(n)}, \tag{2.2}$$

where $\star \in \{\mathrm{AS}, \mathrm{RAS}, \mathrm{MS}\}$, $\mathbf{c}_\star \in \mathbb{R}^N$ are some constant vectors, $\mathbf{e}^{(n)} := \mathbf{u} - \mathbf{u}^{(n)} \in \mathbb{R}^N$ is the error vector after $n$ iterations, and the matrices $T_\star$ are the iteration matrices of the respective methods given by[1]

$$\begin{aligned} T_{\mathrm{AS}} &:= I_N - \sum_{i=1}^{p} R_i^T A_i^{-1} R_i A \equiv I_N - M_{\mathrm{AS}}^{-1} A, \\ T_{\mathrm{AS},\theta} &:= I_N - \theta \sum_{i=1}^{p} R_i^T A_i^{-1} R_i A \equiv I_N - M_{\mathrm{AS},\theta}^{-1} A, \\ T_{\mathrm{RAS}} &:= I_N - \sum_{i=1}^{p} \overline{R}_i^T A_i^{-1} R_i A \equiv I_N - M_{\mathrm{RAS}}^{-1} A, \\ T_{\mathrm{MS}} &:= \prod_{i=p}^{1} \left(I_N - R_i^T A_i^{-1} R_i A\right) \equiv I_N - M_{\mathrm{MS}}^{-1} A, \end{aligned} \tag{2.3}$$

where we also included the *(damped) additive Schwarz method* (dAS) with a damping coefficient $\theta$ corresponding to the iteration matrix $T_{\mathrm{AS},\theta}$. Note that we also write each of the iteration matrices $T_\star$ in the form $I - M_\star^{-1} A$, although the notation $M_\star^{-1}$ does not necessarily imply that this matrix is nonsingular. We use this form of $T_\star$ to highlight the fact that the convergence of these stationary methods can be further *accelerated* if we reformulate them as *preconditioners* for a Krylov method. The preconditioners are then the matrices $M_\star^{-1}$, which, for preconditioning, we assume to be nonsingular. Note that these preconditioners are to be "applied" rather than "solved with", i.e., the preconditioner matrix-vector action is given for any vector $\mathbf{v}$ by $\mathbf{v} \mapsto M_\star^{-1}\mathbf{v}$. For the additive-based methods, the definition of $M_\star^{-1}$ is rather straightforward while for the multiplicative Schwarz method the definition becomes seemingly artificial by having

$$M_{\mathrm{MS}}^{-1} = \left(I_N - T_{\mathrm{MS}}\right) A^{-1},$$

which can be further reformulated for practical use; see, e.g., [44, Section 14.3].

Based on (2.2), we see that convergence (or divergence) for a particular choice of the method is determined by the spectral radius $\rho(T_\star)$, which has been studied in detail for certain classes of matrices $A$. To that end, we denote a *symmetric positive definite* or *SPD* matrix $A$ by $A \succ 0$. Recall that $A$ is a *nonsingular $M$-matrix* provided that the off-diagonal elements of $A$ are nonpositive and all elements of the inverse are nonnegative, i.e., $A - \operatorname{diag}(A) \leq 0$ and $A^{-1} \geq 0$, where the inequalities are understood elementwise; see [5, Chapter 6] or [35,

[1] An equivalent multi-splitting formulation of these can be found in [19, Section 2] and [4, Sections 2 and 3].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



Section 2.5] for further details and references. In the rest of the paper we will simply say $M$*-matrix*, meaning a nonsingular $M$-matrix. We finish this section by recalling convergence results for the classical Schwarz methods for these two classes of matrices, where exact arithmetic is assumed.

THEOREM 2.1 ([18, Lemma 2.8], [4, Theorem 3.8]). *Let $A \succ 0$, and let $q \leq p$ be the smallest number of colors such that we can color all the $p$ subspaces $W_1, \ldots, W_p$ so that if $W_i \cap W_j \neq \mathbf{0}$ and $i \neq j$, then $W_i$ and $W_j$ have different colors. Then,*

$$\rho(T_{MS}) \leq \|T_{MS}\|_A < 1, \quad \text{and, for } \theta < 1/q, \text{ it holds that} \quad \rho(T_{AS,\theta}) \leq \|T_{AS,\theta}\|_A < 1.$$

THEOREM 2.2 ([18, Theorem 3.4], [19, Theorem 4.4], [4, Theorem 3.5]). *Let $A$ be an $M$-matrix and $q \leq p$ be the smallest number of colors such that we can color all the $p$ subspaces $W_1, \ldots, W_p$ so that if $W_i \cap W_j \neq \mathbf{0}$ and $i \neq j$, then $W_i$ and $W_j$ have different colors. Then,*

$$\rho(T_{MS}) < 1, \quad \rho(T_{RAS}) < 1, \quad \text{and, for } \theta < 1/q, \text{ it holds that} \quad \rho(T_{AS,\theta}) < 1.$$

We remark that there are other methods closely related to those mentioned. For example, the RAS method has numerous related variants (e.g., WRAS, ASH, RASH, WRASH; see [19, Section 6] for further references). We do not consider them in this paper.

**3. Multiprecision computations.** As we do *not* work with hardware supporting a wider selection of precisions, the different precisions in our multiprecision algorithms need to be simulated in some way. The number of available options is limited, and both, theoretically and practically, two stand out—the `chop` package [33] and the `advanpix` package [1], both implemented in MATLAB. To the best of our knowledge, these are considered the gold standard among the available software for *simulating* various precisions in the numerical analysis and scientific computing community.

**The advanpix package.** Using advanpix, we can specify the number of accurate digits $d_\ell$ for each computation, i.e., the package simulates the precision based on the *decadic* notation of numbers, in contrast to the binary notation that is commonly used in the hardware, software, and also in the IEEE 754-2019 standard for floating-point formats [36]. Note that the standard also addresses decimal formats. Say we want to simulate `half` precision (`fp16`), which corresponds to $u_{\text{half}} \approx 4.88 \times 10^{-4}$. Using `advanpix`, we have to choose to have either four or five accurate digits, neither of which maps precisely onto the standardized format of `fp16`. Moreover, `advanpix` does not include underflow/overflow treatment. However, using `advanpix` allows us to explore also "new" precisions that are not yet standardized or even used, e.g., six or eleven accurate digits, and to frame the computational precision as more of an "integer-continuous" parameter. Moreover, the package is highly optimized software that overwrites the (also highly optimized) standard MATLAB functions to work with the desired number of accurate digits, e.g., the MATLAB LU or QR factorizations for sparse matrices. Without exploiting these, many problems become computationally too demanding (hence the commercial success of this package). In this context we would also like to highlight that a lot of interest has recently been devoted to the *efficient* simulation of arbitrary precisions on GPUs with astonishing results. For example, although the GPU hardware is highly optimized only for low precisions, such as `fp32`, `fp16`, and even lower precisions, a clever way of simulating `fp64` on these GPUs *using these low-precision formats* was competitive with (or even preferable to) the standard hardware implementation of `fp64`; see [41, 42, 43] and the references therein. This opens the door to the possibility of efficiently simulating "new" low precisions in practice.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



TABLE 3.1

| | Signif. | Exp. | $u$ | $x_{min}$ | $x_{max}$ |
|---|---|---|---|---|---|
| q52 | 5 | 2 | $1.25\times10^{-1}$ | $6.10\times10^{-5}$ | $5.73\times10^{4}$ |
| q43 | 4 | 3 | $6.25\times10^{-2}$ | $1.56\times10^{-2}$ | $2.40\times10^{2}$ |
| bfloat16 | 8 | 8 | $3.91\times10^{-3}$ | $1.18\times10^{-38}$ | $3.39\times10^{38}$ |
| fp16 | 11 | 5 | $4.88\times10^{-4}$ | $6.10\times10^{-5}$ | $6.55\times10^{4}$ |
| fp32 | 24 | 8 | $5.96\times10^{-8}$ | $1.18\times10^{-38}$ | $3.40\times10^{38}$ |
| fp64 | 53 | 11 | $1.11\times10^{-16}$ | $2.23\times10^{-308}$ | $1.80\times10^{308}$ |

**The chop toolbox.** The `chop toolbox` is an open-source MATLAB toolbox[2] developed to simulate different precisions using the native `double` precision of MATLAB, essentially by removing a portion of the mantissa of the result after each operation, corresponding to "rounding" back to the simulated precision. For computations in `single` precision or lower, `chop` faithfully simulates the computation in the given precision (see [33, Section 3.1]) and can also simulate underflow/overflow during the computation. Although this toolbox outperforms many other options (see [33, Sections 5 and 6]), it makes some computations prohibitively time-consuming, even after adapting it to sparse matrices. Although it allows for arbitrary user-defined formats (defined by the number of bits allocated to the exponent and the significand), we will restrict ourselves to the currently standard ones, summarized in Table 3.1.

In most cases, the computationally most demanding part of Schwarz methods is the subdomain solves, i.e., the operations involving $A_i^{-1}$. The issue of underflow/overflow is, in our opinion, an important aspect of multiprecision calculations and has been, at least partially, addressed in [34], where the authors propose rescaling procedures so that (close to) the full range of a given precision is utilized (demonstrated for `fp16`). To be concrete, consider a subdomain problem

$$A_i\mathbf{u}_i = \mathbf{f}_i \tag{3.1}$$

(where we omit the iteration index to keep the notation simple) and a precision $u_\ell$ with the positive range $[x_{min}^{(\ell)}, x_{max}^{(\ell)}]$. The authors propose several algorithms for calculating and using diagonal matrices for row and column rescaling of (3.1)—let us denote them by $D_i^r$ and $D_i^c$ (corresponding to $R$ and $S$ in [34]). For any nonsingular $D_i^r$ and $D_i^c$, we then rewrite (3.1) as

$$\mathcal{A}_i\mathbf{v}_i = \mu\mathbf{b}_i, \tag{3.2}$$

with

$$\mathcal{A}_i := \mu D_i^r A_i D_i^c, \qquad \mathbf{b}_i := D_i^r\mathbf{f}_i, \qquad \mathbf{u}_i := D_i^c\mathbf{v}_i, \qquad \text{and} \quad \mu \in \mathbb{R}.$$

The goal is to choose $D_i^r, D_i^c$ so that $|D_i^r A_i D_i^c| \lesssim 1$ entrywise, and then choose $\mu = \nu x_{max}^{(\ell)}$, for some $\nu \in (0,1)$, so that

$$|\mathcal{A}_i| \equiv |\mu D_i^r A_i D_i^c| \lesssim x_{max}^{(\ell)},$$

again in the entrywise sense. A reasonable choice is then to take $D_i^r$ and $D_i^c$ as in [34, Algorithms 2.3 and 2.4], i.e., to take their diagonal entries as the maximum norms of the rows (and then the columns) of $A_i$. According to [34, Table 4.5], the choice of $\nu = 0.1$ (the

[2]Towards the end of the preparation of the manuscript, the `chop toolbox` was also released for `python`; see [10].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



authors use $\theta$ in their notation) is reasonable, and we comment on this choice later. For the system (3.2), we also rescale the right-hand side; namely we write

$$\mu \mathbf{b}_i = \frac{\|\mathbf{b}_i\|_\infty}{\hat{\nu}_i} \hat{\mathbf{b}}_i \qquad \text{with} \qquad \hat{\mathbf{b}}_i := \hat{\nu}_i \frac{\mu}{\|\mathbf{b}_i\|_\infty} \mathbf{b}_i,$$

where, again, $\hat{\nu}_i \in (0,1)$ allows us to tailor how close to $x_{max}^{(\ell)}$ we rescale the entries of the new right-hand-side vector $\hat{\mathbf{b}}_i$. Altogether, we rewrite (3.1) as

$$\mathcal{A}_i \hat{\mathbf{v}}_i = \hat{\mathbf{b}}_i, \tag{3.3}$$

which we solve in the precision $u_\ell$ and then retrieve $\mathbf{u}_i$ in the precision $u_w$ by calculating (also in precision $u_w$)

$$\mathbf{u}_i = \frac{\hat{\nu}_i}{\|\mathbf{b}_i\|_\infty} D_i^c \hat{\mathbf{v}}_i.$$

Importantly, the rescaling preserves the signs of the entries of the matrix, and hence $A_i$ is an $M$-matrix if and only if $\mathcal{A}_i$ is. It can also be adapted to preserve symmetry (see [34, Algorithm 2.5]), and then it also automatically preserves diagonal dominance.

REMARK 3.1. The `advanpix` toolbox is not open source but, as per the documentation, uses several open source libraries in C and C++. To the best of our understanding, the accuracy of the multiprecision calculations in `advanpix` uses a large number of bits for the exponent, notably larger than that in the IEEE 754-2019 standard. Hence, possible low-precision overflow/underflow occurrences, e.g., during the LU factorization, are likely prevented by this implementation without the user's knowledge. Hence, the experiments using `advanpix` should be understood with this caveat.

Nevertheless, we carry out the calculations so as to minimize the chance of underflow, using the appropriate scaling outlined above. For sufficiently small problem sizes, we have compared the `advanpix` results with those obtained by using `chop` and saw no significant difference, suggesting no major overflow/underflow occurrences.

**4. Algebraic analysis of multiprecision Schwarz methods.** In this section, we give results analogous to Theorems 2.1 and 2.2 when the subdomain solves $A_i^{-1}$ are represented using lower precision in some way. The purpose of the numerical experiments here is twofold—to demonstrate the theoretical results and also to build intuition for and understanding of the multiprecision Schwarz methods. Therefore, we will use convergence properties (such as the number of iterations or the convergence factor) to compare the results with their "full-precision" counterparts, as opposed to, e.g., runtimes. All the code used to produce these results is available at https://github.com/MichalOutrata/mpSchwarz, but, naturally, the code assumes that both the `advanpix` and the `chop` toolboxes are available.

We approach the problem from an algebraic point of view, inspired by the results in [4, 18, 19] assuming exact arithmetic, with the primary goal of carrying out the subdomain solves—corresponding to $A_i^{-1}$ (or $M_i^{-1}$)—in a lower precision $u_\ell$, compared to the higher working precision $u_w$. Mixed precision is not explicitly mentioned in either of these works, but the sections on inexact solves on the subdomains in [4, 18, 19] can be reinterpreted in terms of lower-precision solves.

Following the notation used there, we denote with a tilde *quantities that have been obtained by precision reduction in some sense*. For example, if we assume that the matrix $A_i$ is stored in the working precision $u_w$, and we then store it only in a lower precision $u_\ell$, then the new matrix will be denoted by $\tilde{A}_i$, and replacing all $A_i$ with $\tilde{A}_i$ in the definition of $\mathcal{A}_i$ or $T_\star$ gives us $\tilde{\mathcal{A}}_i$ or $\tilde{T}_\star$ for $\star \in \{\mathrm{AS}, \mathrm{RAS}, \mathrm{MS}\}$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



For clarity, we call the *standard rounding procedure* the rounding to the nearest floating-point number as defined in the IEEE-754 standard [36]. We emphasize that the symbol $\sim$ does not mean that the quantity was obtained by the standard rounding procedure; quite on the contrary—we always consider a particular way of obtaining $\tilde{\mathcal{A}}_i$ from $\mathcal{A}_i$ that suits the situation and is clear from the context. However, we keep a single notation for all these cases (using $\sim$) to highlight their lower-precision nature. We denote the error in the subdomain matrices by $\mathtt{E}_i$, i.e., we have

$$\tilde{A}_i = A_i + \mathtt{E}_i.$$

As is standard in the algebraic convergence theory of Schwarz methods, we are interested in properties of the splittings

$$A_i = \tilde{A}_i - \left(\tilde{A}_i - A_i\right), \qquad i = 1, \dots, p\,. \tag{4.1}$$

**4.1. The general case.** Assuming that $A$ is an $M$-matrix, we recall that a sufficient condition for the convergence of (damped) AS, RAS, and MS is

$$\tilde{A}_i^{-1} \geq 0 \qquad \text{and} \qquad \tilde{A}_i^{-1}\left(\tilde{A}_i - A_i\right) = \tilde{A}_i^{-1}\mathtt{E}_i \geq 0. \tag{4.2}$$

These conditions characterize when the splitting (4.1) is weak regular (of the first type; see [19, Section 4]). Thus, if (4.2) holds for all $i = 1, \dots, p$, then (damped) AS, RAS, and MS with $A_i^{-1}$ ($M_i^{-1}$) replaced with $\tilde{A}_i^{-1}$ ($\tilde{M}_i^{-1}$) converge, i.e., $\rho(\tilde{T}_\star) < 1$.

In light of the rescaling (3.1) to (3.3), we see that the rounding error is committed at the level of the rescaled system; i.e., instead of solving (3.3), we solve

$$\tilde{\mathcal{A}}_i \hat{\mathbf{v}}_i = \tilde{\mathbf{b}}_i,$$

where $\tilde{\mathcal{A}}_i$ (and $\tilde{\mathbf{b}}_i$) is obtained by applying a rounding technique of our choice to $\mathcal{A}_i$ (and $\hat{\mathbf{b}}_i$). In other words, we have

$$\tilde{A}_i = \mu^{-1}(D_i^r)^{-1}\tilde{\mathcal{A}}_i(D_i^r)^{-1}, \tag{4.3}$$

and thus the error matrix $\mathtt{E}_i$ is given by

$$\mathtt{E}_i = \tilde{A}_i - A_i = \mu^{-1}(D_i^r)^{-1}\left(\tilde{\mathcal{A}}_i - \mathcal{A}_i\right)(D_i^r)^{-1} = \mu^{-1}(D_i^r)^{-1}\mathtt{F}_i(D_i^r)^{-1},$$

where we define $\mathtt{F}_i := \tilde{\mathcal{A}}_i - \mathcal{A}_i$ as the rounding error matrix.

Here we would like to recall a useful observation concerning the diagonal rescaling of general stationary iterative methods[3]: if the stationary iterative method is based on a splitting $A = M - N$ such that the entries of $M$ are multiples of the corresponding entries of $A$, then the iteration matrices for $A$ and $D_1 A D_2$ are similar (i.e., have the same convergence factor) for any nonsingular diagonal matrices $D_1, D_2$. On the one hand, this shows that for $u_\ell = u_w$, the rescaling above does not affect the asymptotic convergence rate in our case. On the other hand, we are clearly interested in cases where the rounding *does* make a difference, and through this observation we see that we can expect the convergence factor to be affected by the rescaling.

Revisiting (4.2), a direct calculation shows that the weak regular splitting conditions are invariant with respect to a diagonal scaling with positive entries; i.e., the conditions (4.2) are equivalent to

$$\tilde{\mathcal{A}}_i^{-1} \geq 0 \qquad \text{and} \qquad \tilde{\mathcal{A}}_i^{-1}\left(\tilde{\mathcal{A}}_i - \mathcal{A}_i\right) = \tilde{\mathcal{A}}_i^{-1}\mathtt{F}_i \geq 0. \tag{4.4}$$

[3] We came across this observation in [31, Section 17.2, below equation (17.3)], but this is likely not the original reference.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



The first ingredient in the analysis of (4.4) is rewriting $\tilde{\mathcal{A}}_i^{-1}$ as

$$\tilde{\mathcal{A}}_i^{-1} = \mathcal{A}_i^{-1}\left(I + \mathtt{F}_i\mathcal{A}_i^{-1}\right)^{-1} \tag{4.5}$$

and expanding the inverse matrix there into its Neumann series under the assumption that the spectral radius satisfies

$$\rho(\mathcal{A}_i^{-1}\mathtt{F}_i) < 1. \tag{4.6}$$

Assuming (4.6), the Neumann series expansion reads

$$\tilde{\mathcal{A}}_i^{-1} = \mathcal{A}_i^{-1}\sum_{k=0}^{\infty}\left(-\mathtt{F}_i\mathcal{A}_i^{-1}\right)^k,$$

and we further rearrange it as

$$\begin{aligned}\tilde{\mathcal{A}}_i^{-1} &= \mathcal{A}_i^{-1}\left(I - \mathtt{F}_i\mathcal{A}_i^{-1}\right) + \mathcal{A}_i^{-1}\left(I - \mathtt{F}_i\mathcal{A}_i^{-1}\right)\mathtt{F}_i\mathcal{A}_i^{-1}\mathtt{F}_i\mathcal{A}_i^{-1} + \dots \\ &= \mathcal{A}_i^{-1}\left(I - \mathtt{F}_i\mathcal{A}_i^{-1}\right)\sum_{k=0}^{\infty}\left(\mathtt{F}_i\mathcal{A}_i^{-1}\right)^{2k} = \left(\mathcal{A}_i^{-1} - \mathcal{A}_i^{-1}\mathtt{F}_i\mathcal{A}_i^{-1}\right)\sum_{k=0}^{\infty}\left(\mathtt{F}_i\mathcal{A}_i^{-1}\right)^{2k}.\end{aligned} \tag{4.7}$$

In order to ensure (4.4), we focus on ensuring $\mathtt{F}_i \geq 0$ as well as $\tilde{\mathcal{A}}_i^{-1} \geq 0$. Notice that the latter should be natural, as we have $A_i^{-1} \geq 0$ and thereby also $\mathcal{A}_i^{-1} \geq 0$, while the condition $\mathtt{F}_i \geq 0$ can be accomplished, at least in theory, by virtue of choosing an appropriate $u_\ell$ and the rounding procedure (see details below). In fact, assuming $\mathtt{F}_i \geq 0$, the natural condition for ensuring also $\tilde{\mathcal{A}}_i^{-1} \geq 0$ (and hence (4.4)) becomes

$$\mathcal{A}_i^{-1} \geq \mathcal{A}_i^{-1}\mathtt{F}_i\mathcal{A}_i^{-1}, \tag{4.8}$$

i.e., a second condition for the choice of $u_\ell$ in addition to (4.6). Notice that both (4.6) and (4.8) are in some sense generalizations of the standard relative rounding error assumption

$$|\mathtt{F}_i| \leq u_\ell|\mathcal{A}_i|, \tag{4.9}$$

which is generally guaranteed (the absolute value is to be understood componentwise). Also, similarly to [2, Sections 4 and 5], both (4.6) and (4.8) invite us to choose $(u_\ell)_i$ for each subdomain independently, based on the relevant quantities or their estimates[4]. We retain (4.6) and (4.8) as assumptions, coupling the subdomain problems and the choice of the lower precision $u_\ell$, and we turn our attention to the condition $\mathtt{F}_i \geq 0$.

If we use the standard round-to-the-nearest rounding as specified by IEEE 754-2019, then we are unlikely to satisfy $\mathtt{F}_i \geq 0$, except in some special cases. However, the rounding process is very often fully under our control. Since, by definition, the off-diagonal entries of $\mathcal{A}_i$ are nonpositive while its diagonal entries are nonnegative, a simple way to ensure $\mathtt{F}_i \geq 0$ is to take $\tilde{\mathcal{A}}_i$ as the "sign-informed round up" of $\mathcal{A}_i$.

To this end we assume that for any precision $u_\ell$, we have at our disposal the functions $\mathtt{rd}_{u_\ell}$ and $\mathtt{ru}_{u_\ell}$ that round towards zero (down) and towards plus/minus infinity (up)[5]. We

[4] In [2], the authors work with a similar idea but calculate explicitly the analogue of the inverses $\tilde{\mathcal{A}}_i^{-1}$ in different precisions based on their conditioning. This is somewhat complementary to our approach as our interest lies in the analysis of the resulting method rather than in the practical aspects, which have been covered in [2]; we do not comment further on how to choose $u_\ell$ (or $(u_\ell)_i$) for the subdomain problems.

[5] In the `chop` toolbox, these are already implemented, and for the `advanpix` package, they are straightforward to implement, as we deal with the precision $u_\ell$ corresponding to $d_\ell$ accurate decimal digits (as opposed to dealing with bits in the case of `chop`).

ETNA
Kent State University and
Johann Radon Institute (RICAM)



then introduce the rounding procedure $\mathrm{round}_{Mmtrx}()$ that, for any matrix $X$, yields its low-precision approximation $\mathrm{round}_{Mmtrx}(X)$ given by *upward rounding*, i.e.,

$$(\mathrm{round}_{Mmtrx}(X))_{mn} \equiv (\mathrm{round}_{Mmtrx}(X,u_\ell))_{mn} := \begin{cases} \mathtt{ru}_{u_\ell}\left((X)_{mn}\right), & \text{if } (X)_{mn} > 0, \\ \mathtt{rd}_{u_\ell}\left((X)_{mn}\right), & \text{if } (X)_{mn} < 0. \end{cases}$$

Taking

$$\tilde{\mathcal{A}}_i = \mathrm{round}_{Mmtrx}(\mathcal{A}_i), \tag{4.10}$$

we get $\mathtt{F}_i \geq 0$ and obtain convergent multiprecision Schwarz methods under the assumptions (4.6) and (4.8); we summarize these results in Theorem 4.1 below.

In the following theorem, we use the term *full precision* to mean exact arithmetic, which is the setting in which Theorems 2.1 and 2.2 hold.

THEOREM 4.1. *Let $A$ be an $M$-matrix and $q \leq p$ be the smallest number of colors such that we can color all the $p$ subspaces $W_1, \ldots, W_p$ so that if $W_i \cap W_j \neq \mathbf{0}$ and $i \neq j$, then $W_i$ and $W_j$ have different colors. Moreover, assume that for each $i = 1, \ldots, p$, we replace the subdomain solver $A_i^{-1}$ in full precision with the subdomain solver $\tilde{A}_i^{-1}$ in a (lower) precision $u_\ell$, obtaining the multiprecision (damped) AS, RAS, and MS methods with the iteration matrices $\tilde{T}_{AS,\theta}$, $\tilde{T}_{RAS}$, and $\tilde{T}_{MS}$, respectively. Taking $\tilde{A}_i$ as in (4.3) with $\tilde{\mathcal{A}}_i$ given as in (4.10), if (4.6) and (4.8) are satisfied, then*

$$\rho(\tilde{T}_{MS}) < 1, \quad \rho(\tilde{T}_{RAS}) < 1, \qquad \textit{and, for } \theta < 1/q, \textit{ it holds that} \quad \rho(\tilde{T}_{AS,\theta}) < 1,$$

*and the multiprecision versions of the classical Schwarz methods are convergent.*

REMARK 4.2. In practice, in addition to the error $\mathtt{E}_i$, an additional error due to round-off would occur, say $\mathtt{G}_i$. This very small error may change the signs of the entries. To avoid this problem, one could explicitly add an additional small fixed perturbation, say $\mathtt{H}_i$, throughout so that $A_i + \mathtt{E}_i + \mathtt{G}_i + \mathtt{H}_i$ has the desired signs in its entries (and in the other matrices studied). In this manner, our analysis would carry over to this practical situation. We note, however, that in our experiments, no such issues were detected.

Following [18, Section 4], [19, Section 7], and [4, Section 4], we also obtain comparisons for different choices of $u_\ell$. More specifically, given an $M$-matrix $X$ and two different low precisions $u_\ell^{(1)} \leq u_\ell^{(2)}$, we obtain $\mathrm{round}_{Mmtrx}(X, u_\ell^{(1)}) \leq \mathrm{round}_{Mmtrx}(X, u_\ell^{(2)})$, and hence

$$\left(\mathrm{round}_{Mmtrx}(X, u_\ell^{(1)})\right)^{-1} \geq \left(\mathrm{round}_{Mmtrx}(X, u_\ell^{(2)})\right)^{-1}.$$

Using this for the Schwarz methods, we obtain

$$\rho\left(\tilde{T}_{\star,u_\ell^{(1)}}\right) \leq \rho\left(\tilde{T}_{\star,u_\ell^{(2)}}\right), \qquad \text{where } \star \in \{\mathrm{AS}, \mathrm{RAS}, \mathrm{MS}\}. \tag{4.11}$$

In other words, the better the precision, the faster the convergence. The important questions then become

- *Are the conditions (4.6) and (4.8) in some sense sharp or descriptive in the context of convergence of the multiprecision Schwarz methods?*
- *When is (4.11) strict?*
- *Among those $u_\ell$ for which (4.11) is strict, which should we use, i.e., to what extent are there diminishing returns as we approach equality in (4.11)?*
- *What computational speedups can we obtain for practically relevant problems using modern HPC architectures?*

ETNA
Kent State University and
Johann Radon Institute (RICAM)



Next, we investigate the first three questions numerically for model problems arising from discretizations of reaction-advection-diffusion equations. While the last question is itself important, it goes beyond the scope of this manuscript. Regarding speedups, we note that naive and frequent switching between rounding modes can be extremely costly, as it generally breaks the pipelining of the hardware implementation. However, with modern GPU architectures and with the rise of simulated arithmetic (see the aforementioned Ozaki schemes [43]), it is plausible that even this issue can be addressed. That being said, we do not further comment on this matter, leaving it as a possible direction for future work. In our experiments, the low-precision computations are simulated, and hence the slowdown of changing the rounding technique is generally negligible compared to the slowdown due to precision simulations.

Taking the unit square, i.e., $\Omega = \{\mathbf{x} = [x_1, x_2]^T \in [0,1]^2\}$, we consider the partial differential equation

$$\mathcal{L}u = f \quad \text{in } \Omega \qquad \text{and} \qquad u = g \quad \text{on } \partial\Omega, \tag{4.12}$$

with the differential operator $\mathcal{L}$ given by

$$\mathcal{L}u := \eta(\mathbf{x})u - \operatorname{div}(\alpha(\mathbf{x})\nabla u) + \mathbf{b}(\mathbf{x}) \cdot \nabla u. \tag{4.13}$$

We take the coefficient functions $\eta(\mathbf{x})$, $\alpha(\mathbf{x})$, and $\mathbf{b}(\mathbf{x}) = [b_1(\mathbf{x}), b_2(\mathbf{x})]^T$ as follows:

*Problem 1.* (inspired by [24, Figure 2.1])

$$\eta(\mathbf{x}) := x_1^2 \cos(x_1 + x_2)^2, \quad \alpha(\mathbf{x}) := 20(x_1 + x_2)^2 e^{x_1 - x_2}, \quad \text{and} \quad \mathbf{b}(\mathbf{x}) = \begin{bmatrix} x_2 - 0.5, \\ x_1 - 0.5 \end{bmatrix}.$$

*Problem 2.* (inspired by [20, Section 4.1])

$$\eta \equiv 0, \quad \alpha \equiv 1, \quad \text{and} \quad \mathbf{b}(\mathbf{x}) = \begin{bmatrix} \beta x_1(x_1 - 1)(1 - 2x_2) \\ -\beta x_2(x_2 - 1)(1 - 2x_1) \end{bmatrix}, \quad \text{with} \quad \beta = 100.$$

*Problem 3.* (based on Problem 2)

$$\eta \equiv 0, \quad \alpha(\mathbf{x}) = \begin{cases} 10^6 & \text{if } \|\mathbf{x} - [0.5\ 0.1]^T\| < 0.25, \\ 1 & \text{otherwise,} \end{cases}$$
$$\text{and} \quad \mathbf{b}(\mathbf{x}) = \begin{bmatrix} \beta x_1(x_1 - 1)(1 - 2x_2) \\ -\beta x_2(x_2 - 1)(1 - 2x_1) \end{bmatrix},$$

again with $\beta = 100$. To discretize the problem, we use the standard five-point stencil finite-difference scheme, adapting some of the code from [24] and obtain systems of linear equations (1.1) with $A$ being a nonsymmetric $M$-matrix. We then partition $A$ into two overlapping subdomain problems, taking the size of the overlap block to correspond to the bandwidth[6] of $A$, that is, we consider two overlapping subdomains $\Omega_1, \Omega_2 \subset \Omega$ with overlap width[7] $\mathcal{O}(h)$. We take the right-hand-side vector $\mathbf{f}$ and the initial approximation vector $\mathbf{u}^{(0)}$ to be random vectors with entries in $(0,1)$.

First we fix $N = 50^2$ and display the convergence curves and the observed convergence factor[8] $\rho_{\text{conv}}$ in Figure 4.1 for the multiplicative Schwarz method and the standard

[6] We order the unknowns in the MATLAB's `numgrid` style, which is column-oriented and results in a banded matrix.

[7] As a result, we expect the convergence factor of the Schwarz method to deteriorate as $N$ increases; see [19, Section 5] and [4, Section 5].

[8] Given the relative error norms of a Schwarz method, we computed $\rho_{\text{conv}}$ based on a least-squares fit of the linear convergence factor.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



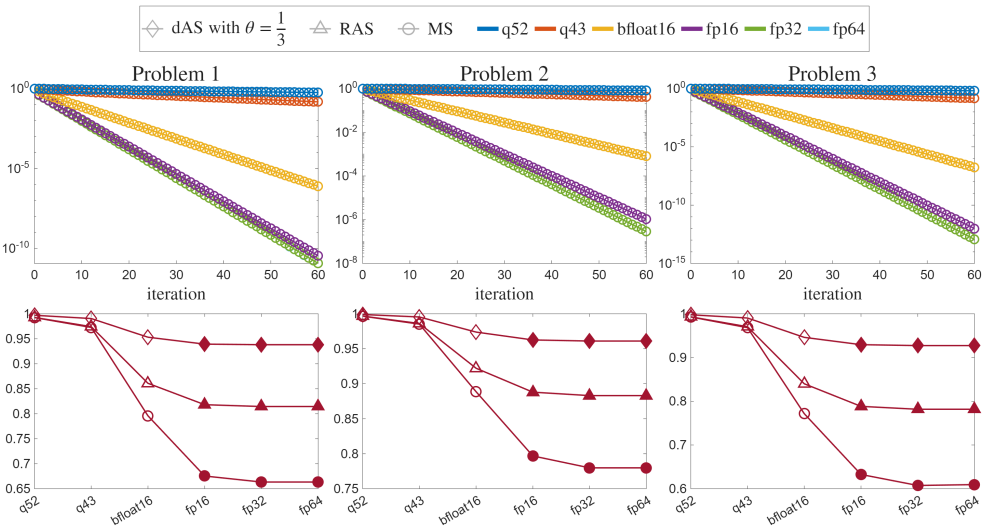



FIG. 4.1. *Top: the 2-norm of the error of the multiplicative Schwarz method with $u_w$ corresponding to* `fp64` *and different choices of $u_\ell$, using the* `chop` *toolbox. For* `fp32` *and* `fp64`*, the graphs are indistinguishable from each other. Bottom: the observed convergence factor $\rho_{\text{conv}}$ for all Schwarz methods and choices of $u_\ell$; if the conditions* (4.6) *and* (4.8) *are satisfied for both subdomains $i = 1, 2$ for a certain $u_\ell$, then the marker is filled. For example, for Problem 1, the conditions* (4.6) *and* (4.8) *are satisfied for both $i = 1, 2$ starting from* `fp16`*.*

low-precision formats in the `chop` package (see Table 3.1 above), adjusting the scaling from Section 3 to use as much of the available range of each precision without overflow during the computations. We see that the methods in fact converge in all of the considered precisions, although the conditions (4.6) and (4.8) are satisfied only for `fp16`, `fp32`, and `fp64`. Moreover, once the conditions (4.6) and (4.8) are satisfied, they are also satisfied for higher precisions, and, more importantly, the observed convergence factor $\rho_{\text{conv}}$ essentially becomes invariant as the precision is increased further. In other words, we obtain *very little* additional computational benefit (within the first 60 iterations) from considering higher precisions once the conditions (4.6) and (4.8) are satisfied. This suggests that the conditions (4.6) and (4.8) provides a good guidance for the a-priori choice of the lower precision $u_\ell$.

We illustrate this further by showing the analogous experiment using the `advanpix` toolbox, which allows finer tuning of the considered precision at the cost of giving up the control over underflow/overflow situations (however, since we encountered no overflow with `chop`, this seems not too worrying) in Figure 4.2. We see that not only do the observations from Figure 4.1 still hold, but the case for using the conditions (4.6) and (4.8) as predictors of a suitable precision $u_\ell$ is further strengthened.

REMARK 4.3. Numerically, the experiments suggest that (4.6) is generally weaker than (4.8), although we have not been able to establish this as a theoretical result. However, we have never observed this discrepancy to be large, using either `chop` (e.g., for `chop`, the difference is only present for `bfloat16` and `fp16`, where (4.6) was—for some problems and mesh sizes—satisfied for `bfloat16`, whereas (4.8) was not) or `advanpix` (again, (4.6) was rarely satisfied for $d_\ell = 3$, whereas (4.8) was not).

We further illustrate the tipping point at which the conditions (4.6) and (4.8) are met or violated by plotting the error of the multiplicative Schwarz method throughout the initial iterations for different choices of $u_\ell$ for Problem 1 in Figure 4.3, again using `chop` with overflow enabled (although no overflow was encountered due to the rescaling). In full precision, we expect to see the classical two-domain profile of the largest eigenmode of the matrix $T_{\text{MS}}$,

ETNA
Kent State University and
Johann Radon Institute (RICAM)



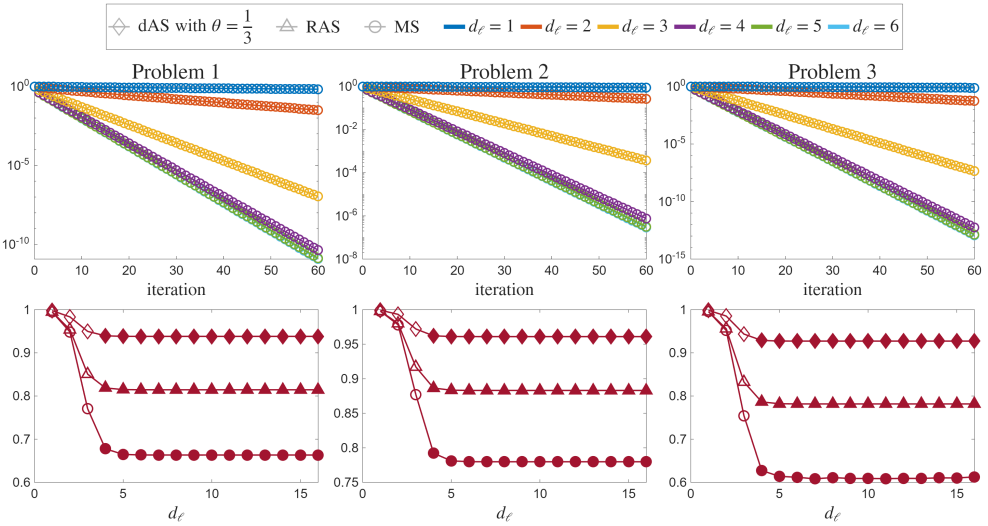



FIG. 4.2. *Top: the 2-norm of the error of the multiplicative Schwarz method with $u_w$ corresponding to* `fp64` *and different choices of $d_\ell$, using the* `advanpix` *toolbox. For $d_\ell \geq 6$, the graphs are essentially indistinguishable from the green one for $d_\ell = 5$. Bottom: the observed convergence factor $\rho_{\text{conv}}$ for all Schwarz methods and choices of $d_\ell$; if the conditions (4.6) and (4.8) are satisfied for both subdomains $i = 1, 2$ for a certain $d_\ell$, then the marker is filled. For example, for Problem 1, the conditions (4.6) and (4.8) are satisfied for both $i = 1, 2$ from $d_\ell = 4$ onward.*

which is smooth on each of the subdomains. Indeed, for `fp16`, `fp32`, and `fp64` this is what we observe. However, for `q52` and `q43`, the ridge in the middle that separates the two subdomains never forms, and for `bfloat16`, it takes several iterations to be established to the same extent. In other words, for sufficiently low precision $u_\ell$, the method effectively loses its continuous-level interpretation as a domain decomposition method, although it is still a reasonably effective (and even convergent) smoother. Importantly, satisfying the conditions (4.6) and (4.8) is visible not only in the rate of convergence but also in the *nature of it*.

The above observations remained true when changing

- the `chop` and `advanpix` toolboxes,
- the problem (we experimented with various settings of reaction-advection-diffusion problems such that $A$ is an $M$-matrix),
- the method (although, e.g., for dAS we observe an initial period before the error converges to the dominant eigenmode),
- the initial approximation (the only change is in the initial period before the error converges to the dominant eigenmode).

As we have kept the problem size relatively small so far, we next consider an experiment with varying $N$. However, letting $N$ grow, two numerical limitations arise: (i) the `chop` toolbox becomes too slow and (ii) the verification of the condition (4.8) becomes untenable. Hence, for the following experiments, we use only the `advanpix` toolbox and verify only condition (4.6) (essentially testing whether the observation in Remark 4.3 also holds for larger $N$).

All the above characteristics remained true, with the only change being the value of $\ell$ for which the precision $d_\ell$ first satisfies the condition (4.6). We display these for $N \in \{50^2, \ldots, 330^2\}$ in Figure 4.4. Notably, we see that after the first $d_\ell$ for which (4.6) is satisfied, there is little to no change from using additional precision, precisely as observed previously for $N = 50^2$. In other words, the dominant eigenmodes of $T_\star$ with $\star \in \{\text{AS}, \text{RAS}, \text{MS}\}$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



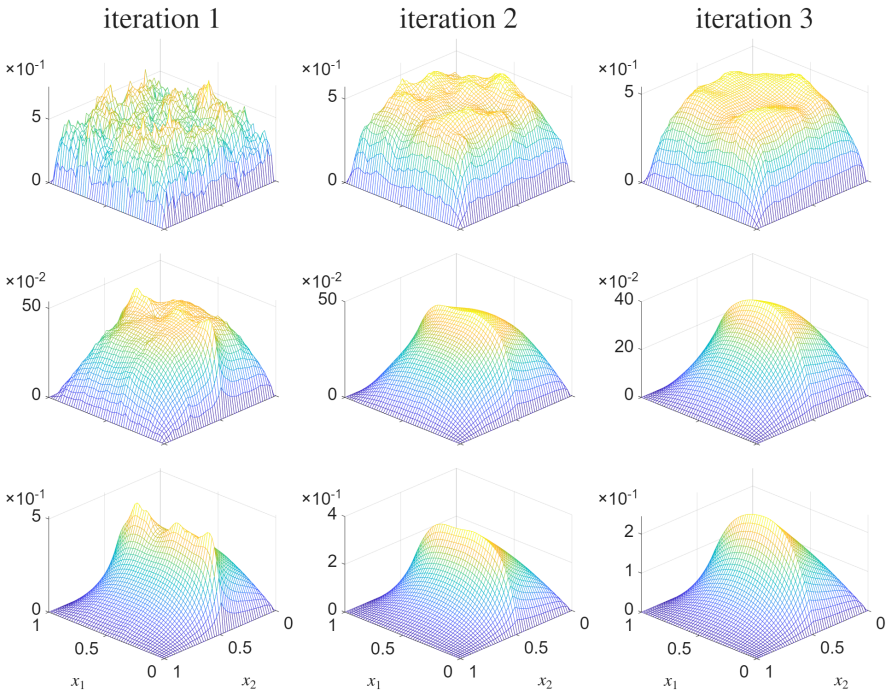



FIG. 4.3. *The errors of the multiplicative Schwarz method used for Problem 1 with* $N = 50^2$ *and the* `chop` *toolbox after 1, 2, and 3 iterations using* `q43` *(top),* `bfloat16` *(middle), and* `fp16` *(bottom). Up to scaling, the error for* `q52` *is analogous to that in the top row, and the errors for* `fp32` *and* `fp64` *are analogous to those in the bottom row.*

seem to be well captured already with limited precision, *and* the other eigenmodes are not too sensitive to small perturbations of the subdomain solves and remain "non-dominant". In addition, the same type of behavior as shown in Figure 4.3 is present for larger $N$; i.e., for too low precision, the methods lose their two-domain nature, converge extremely slowly, but remain effective smoothers.

We see that the convergence factors $\rho_{\text{conv}}$ are remarkably uniform across the different problems and with respect to changes in the solve precision $u_\ell$. Condition (4.6) is satisfied either at $d_\ell = 4$ or $d_\ell = 5$, depending also on the size of the problem, and once the condition is satisfied, $\rho_{\text{conv}}$ stabilizes around this final value. In other words, based on these experiments, condition (4.6) still governs the required precision. We note that this is perhaps not too surprising, as condition (4.8) is clearly *only* a sufficient condition—if it does not hold, then the entries of the matrix $\tilde{\mathcal{A}}_i^{-1}$ in (4.7) are given by an oscillating sum (rather than a sum of only nonnegative numbers), which can easily sum up to a nonnegative number. On the other hand, if (4.6) does not hold, then there is no easy way around it.

We remark that in both [2, 46], the authors use the condition numbers of the subdomain matrices for choosing $u_\ell$. The condition numbers of the subdomain matrices $A_i$ for both Problems 1 and 2 are fairly small, within the range $(10^3, 10^5)$, whereas for Problem 3, the subdomain matrices $A_i$ have condition numbers within the range $(10^9, 10^{11})$. We see that the conditioning of the subdomain problem and the precision $u_\ell$ seems to interact very little. This remained true for other, similarly focused experiments. We note that the rescaling process explains part of this observation, but we note that the analogous plots for experiments *without the rescaling*, i.e., taking $D_i^{(r,c)} = I_{N_i}$, look fairly similar, although the plots are "less

ETNA
Kent State University and
Johann Radon Institute (RICAM)



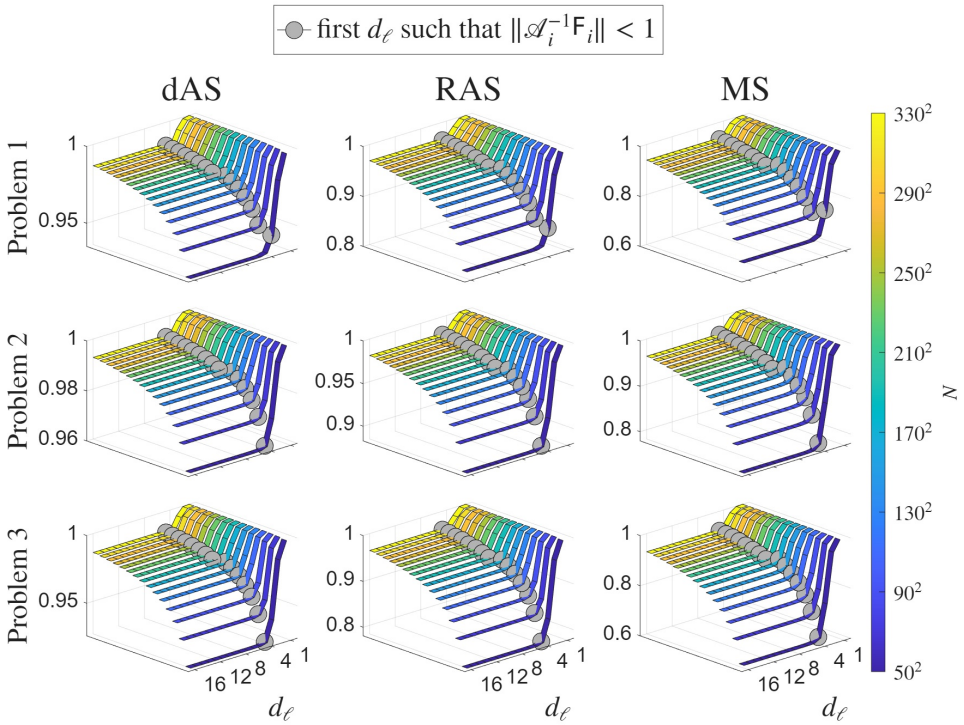



FIG. 4.4. *We display $\rho_{\text{conv}}$ for the Schwarz methods for $d_\ell = 1, 2, \ldots, 16$ and different problem sizes $N$. The gray circles highlight the first $d_\ell$ for which condition (4.6) is satisfied (it is also satisfied for all the subsequent values).*

smooth".

Notice that, as $N$ increases, $\rho_{\text{conv}}$ tends towards 1. This is a feature of Schwarz methods—the convergence factor depends on the physical width of the overlap of the subdomains $\Omega_1$ and $\Omega_2$. As noted above, in our setting, the overlap width is proportional to $h \sim 1/N$, and therefore this effect is expected even in full precision, as is clearly visible in Figure 4.4.

In addition, note that in condition (4.6), the 2-norm can be replaced by any consistent and equivalent matrix norm, and the Neumann series result is still valid; see, e.g., [40, Section 1.3, Lemma 1.3.10]. In other words, the computationally infeasible condition (4.6) can be replaced by

$$\|\mathcal{A}_i^{-1}\mathrm{F}_i\|_F^2 < 1 \qquad \text{or} \qquad \|\mathcal{A}_i^{-1}\mathrm{F}_i\|_1^2 < 1.$$

Although these are clearly preferable for determining the number of digits $d_\ell$ (or $d_{\ell_i}$, $i = 1, \ldots, p$) (see [2]), they might provide a worse indication of whether or not a given precision is suitable for a given $N$.

When running Schwarz methods, we can see the effect of the lower precision $u_\ell$ only in the dominant eigenmode and eigenvalue of $T_\star$, as one expects for a fixed-point iteration. However, in practice, we usually accelerate Schwarz methods using Krylov subspace methods, i.e., we use Schwarz methods as preconditioners for Krylov methods. In order to be effective preconditioners, calculations in the precision $u_\ell$ instead of the precision $u_w$ on the subdomains *should not* make the eigenbasis of $M_*^{-1}A$ much more ill-conditioned or the spectrum much more "spread out"; otherwise a notable slowdown of GMRES convergence (compared to

ETNA
Kent State University and
Johann Radon Institute (RICAM)



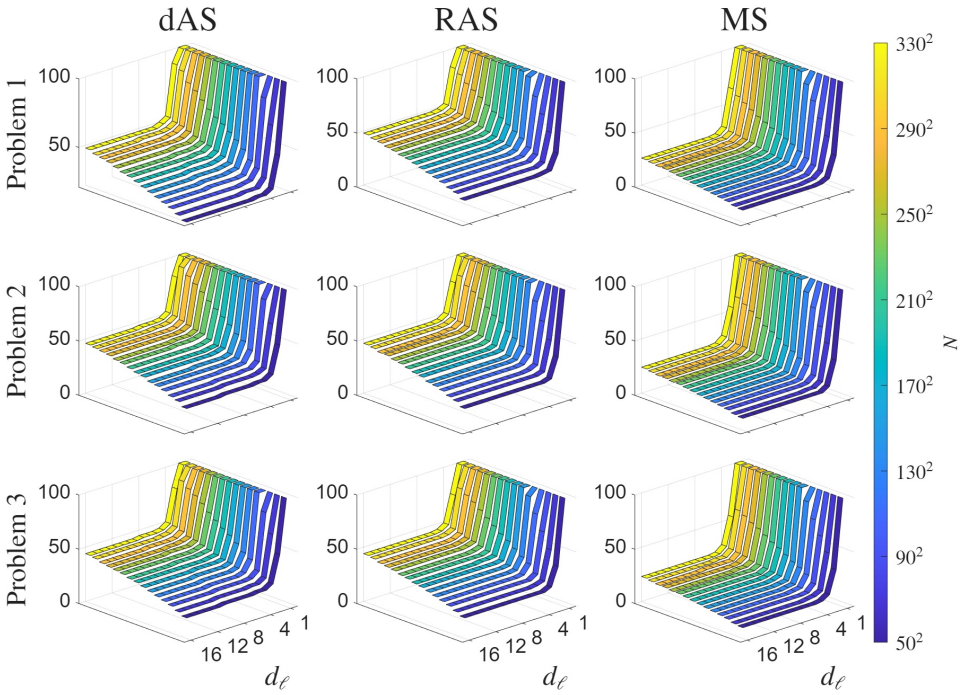



FIG. 4.5. *The number of preconditioned GMRES iterations to reduce the relative residual norm below* $10^{-12}$ *capped at 100.*

that of the appropriate full-precision Schwarz method) can occur[9]. In other words, the above experiments do not necessarily suggest that the multiprecision Schwarz methods will also be efficient when used as preconditioners. We investigate this next numerically and use preconditioned GMRES with multiprecision dAS, RAS, and MS as the left preconditioners, with preconditioned relative residual tolerance $10^{-12}$, a zero initial approximation, and a maximum number of iterations set to 100. We display the number of GMRES iterations in Figure 4.5.

We observe that the effect of the low precision does not meaningfully affect the number of iterations that the preconditioned GMRES method requires to reduce the preconditioned relative residual to the tolerance $10^{-12}$. Moreover, we observe the same diminishing returns as for the convergence factors of the methods in Figure 4.4, *and* these occur mostly at the same thresholds, i.e., for the same precisions $u_\ell$. We again observe an increase in the number of iterations as $N$ increases, for similar reasons as for the stand-alone solves. Although we do not provide an analysis of the multiprecision preconditioned GMRES method, we refer the reader to [7] for an analysis and further references.

REMARK 4.4. In the context of (very) low-precision preconditioners, the classical GMRES method should be replaced by the *flexible GMRES* method (FGMRES; see [44, Section 9.4.1]), as the low-precision nature of the preconditioner (in our case, the low-precision solves) can no longer be associated with a single matrix representation across the iterations. Moreover, as noted in [11], FGMRES in the multiprecision context tends to be more robust. In

[9] We note that this widely used rule is only a heuristic; see [37, Section 5.7] and the references therein for more details on GMRES convergence analysis.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



this context, our result show that, even for the standard GMRES method, which is not adapted to the specifics of the preconditioners, we obtain the stated results. For more details on the FGMRES analysis within the multiprecision context, see [3, 11].

In summary , we observe that in the model examples, *four or five digits suffice* to achieve results virtually indistinguishable from those obtained with full double precision. That is, running `fp32` (or even `fp16` for smaller $N$) should be up to twice as fast (four times as fast) as the standard `fp64` Schwarz method, without any meaningful drawback. Moreover, the results show that running `fp16` or even `bfloat16` should result in a negligible slowdown while offering up to a further twofold speedup.

**4.2. The symmetric case.** We now consider *symmetric $M$-matrices*, sometimes also called Stieltjes matrices (Stieltjes matrices are symmetric positive definite; see [5, Chapter 6, Theorem 2.3 ($D_{16}$)]).

First, we would like to highlight that Theorem 4.1 applies to this case "as is". Moreover, if the rescaling and rounding are performed symmetrically (or if we store and work only with, say, the upper-triangular part of the matrix), then symmetry is preserved for the tilde quantities as well. Notably, to achieve symmetric scaling, the proposed rescaling algorithm needs to be symmetrized, leading to an iterative procedure; see [34, Algorithm 2.5]. Moreover, our rounding routine can be further tailored to preserve other useful properties of the subdomain matrices.

For example, assuming that $A$ has dominant diagonal entries in the sense that $a_{ii} \geq |a_{ij}|$ for all $i, j$, Algorithm 2.5 from [34] converges in a single step, yielding that $D_i^{(r)} = D_i^{(c)} = \operatorname{diag}(a_{11}^{-1/2}, \ldots, a_{N_i N_i}^{-1/2})$, with $D_i^{(r)} A_i D_i^{(c)}$ having all ones on the diagonal and all other entries being bounded in absolute value by one. Then, taking $\nu$ to be a power of two (or another number represented exactly in precision $u_\ell$), we can use the rounding routine

$$\left(\operatorname{round}_{Diag}(X)\right)_{mn} := \begin{cases} (X)_{mn} \equiv 1, & \text{if } m = n, \\ \mathtt{rd}_{u_\ell}\left((X)_{mn}\right), & \text{if } m \neq n, \end{cases}$$

in order to preserve this property (or, e.g., diagonal dominance) also for the rescaled, rounded matrix $\tilde{\mathcal{A}}_i$. In either case, any reasonable rounding should satisfy $\mathtt{F}_i^T = \mathtt{F}_i$ (as $\mathcal{A}_i$ is symmetric), which suffices for now.

Next, we turn our attention to the classical convergence theory for the algebraic Schwarz methods in the symmetric positive-definite case. As the RAS method is inherently nonsymmetric, it is standard to consider the convergence theory only for the (damped) AS and MS methods, and thus we focus only on these two classes[10]. For these methods, the driving force behind Theorem 2.1 is the so-called $P$-regular Splitting Theorem; see, e.g., [40, Theorem 7.1.9]. The core assumption there becomes that the splitting in (4.1) is a $P$-regular splitting, i.e., that

$$\tilde{A}_i^T + \tilde{A}_i - A_i \succeq 0\,. \tag{4.14}$$

In order to satisfy (4.14), the standard assumption in the literature is $\tilde{A}_i \succeq A_i$ (see, e.g., [4, equation (39), p. 621]), but, unfortunately, this is not an "easy-to-ensure condition" for a specific rounding routine. Instead, we first observe that, for a symmetric scaling, i.e., the case

[10] Some theory for SPD matrices has been developed for variants of the RAS method (see [8]), and recently, the convergence of RAS for SPD matrices was studied in [45] using variational methods for a simple model problem. In general, convergence of RAS is usually addressed in combination with the particular problem (see [17]) or based on other properties of the system matrix; see [19].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



of $D_i^{(r)} = D_i^{(c)} =: D_i$, we obtain

$$\tilde{A}_i^T + \tilde{A}_i - A_i = \mu^{-1} D_i^{-1} \left( \tilde{\mathcal{A}}_i^T + \tilde{\mathcal{A}}_i - \mathcal{A}_i \right) D_i^{-1} = \mu^{-1} D_i^{-1} \left( \tilde{\mathcal{A}}_i + \mathtt{F}_i \right) D_i^{-1},$$

and (4.14) becomes equivalent to

$$\tilde{\mathcal{A}}_i + \mathtt{F}_i \succeq 0,$$

which is ensured by the following two conditions:

$$\tilde{\mathcal{A}}_i \succ 0 \qquad \text{and} \qquad \lambda_{min}(\tilde{\mathcal{A}}_i) \geq |\lambda_{-\infty}(\mathtt{F}_i)|\,, \tag{4.15}$$

where $\lambda_{min}(\tilde{\mathcal{A}}_i) > 0$ is the smallest eigenvalue of $\tilde{\mathcal{A}}_i$ (as we assume $\tilde{\mathcal{A}}_i \succ 0$) and $\lambda_{-\infty}(\mathtt{F}_i)$ is the smallest eigenvalue of $\mathtt{F}_i$ (on the real line, *not* in absolute value, since $\mathtt{F}_i^T = \mathtt{F}_i$). Notice that these conditions differ substantially, as we allow the rounding error matrix to be indefinite.

The first condition in (4.15) can be ensured by rounding as in Section 4.1 so that $\tilde{\mathcal{A}}_i$ remains a Stieltjes matrix and hence is symmetric positive definite. The second condition can be further expanded, using the standard perturbation theory of eigenvalues for symmetric matrices (as both $\mathcal{A}_i$ and $\mathtt{F}_i$ are symmetric). Indeed, using Weyl's theorem (see [49, Theorem 4.8 and Corollary 4.9]) for $\tilde{\mathcal{A}}_i = \mathcal{A}_i + \mathtt{F}_i$, we obtain

$$\lambda_{min}(\tilde{\mathcal{A}}_i) \geq \lambda_{min}(\mathcal{A}_i) + \lambda_{-\infty}(\mathtt{F}_i),$$

so that, to ensure the second condition in (4.15), it is enough to require

$$\lambda_{min}(\mathcal{A}_i) \geq 2\,|\lambda_{-\infty}(\mathtt{F}_i)|\,. \tag{4.16}$$

We summarize the results in the following theorem:

THEOREM 4.5. *Let $A$ be a Stieltjes matrix and $q \leq p$ be the smallest number of colors such that we can color all the $p$ subspaces $W_1, \ldots, W_p$ so that if $W_i \cap W_j \neq \{0\}$ and $i \neq j$, then $W_i$ and $W_j$ have different colors. Moreover, assume that for each $i = 1, \ldots, p$, we replace the subdomain solver $A_i^{-1}$ in full precision with the subdomain solver $\tilde{A}_i^{-1}$ with (lower) precision $u_\ell$, obtaining the multiprecision (damped) AS and MS methods with the iteration matrices $\tilde{T}_{AS,\theta}$ and $\tilde{T}_{MS}$. Taking $\tilde{A}_i$ as in (4.3) with $\tilde{\mathcal{A}}_i$ given as in (4.10) with a symmetric scaling, if (4.6), (4.8), and (4.16) are satisfied, then*

$$\rho(\tilde{T}_{MS}) < 1, \quad \textit{and, for } \theta < 1/q, \textit{ it holds that} \quad \rho(\tilde{T}_{AS,\theta}) < 1,$$

*and the multiprecision versions of the classical Schwarz methods are convergent.*

We note that an analysis for SPD matrices and multiprecision additive Schwarz methods has been considered elsewhere; see [2, 28, 46, 51]. However, in all these papers the authors consider the (nondamped) AS *as a preconditioner for CG*, and hence the analysis and/or numerical investigation focus on the preconditioned CG method, e.g., using variational techniques that allow establishing a bound for the condition number of the preconditioned system. We also note that the assumptions look somewhat similar[11]. Indeed, to obtain a less restrictive version of (4.16), we can replace $\lambda_{-\infty}(\mathtt{F}_i)$ with $-\max_{\lambda \in \sigma(\mathtt{F}_i)} |\lambda| \equiv -\rho(\mathtt{F}_i)$, so that instead of (4.16), we would require

$$\lambda_{min}(\mathcal{A}_i) \geq 2\rho(\mathtt{F}_i). \tag{4.17}$$

[11] Compare [46, equations (12) and (14)] and [2, Section 5] with (4.6) and (4.16).

ETNA
Kent State University and
Johann Radon Institute (RICAM)



Since we still have the entrywise comparison of $u_\ell|\mathtt{F}_i|$ and $|\mathcal{A}_i|$ (see (4.9)) and we have knowledge of the sign distribution of the entries of these matrices, we could arrive at a comparison theorem for $\rho(\mathtt{F}_i)$ and $\rho(\mathcal{A}_i)$, so that (4.17) would relate the condition number $\rho(\mathcal{A}_i)/\lambda_{min}(\mathcal{A}_i)$ with the used precision $u_\ell$, yielding the type of condition we encounter in [46, Section 3.2.2, equation (14)] or [2, Section 5]. The above derivation illustrates that our results are more nuanced than the existing ones, also in treating Schwarz methods (and their convergence) as stand-alone methods. We focus on more particular systems (in the sense of the $M$-matrix property, which, together with symmetry, constitutes a subclass of SPD matrices) compared to those in the existing literature, and we carefully exploit this extra information through the specialized rounding techniques.

Next, we present results analogous to the experiments considered in Section 4.1. We consider the same problem as in (4.12)–(4.13) but omit the advection terms so that we obtain Stieltjes matrices after discretization, using the following parameters:

*Problem 4.* (analogue of Problem 1)

$$\eta(\mathbf{x}) := x_1^2\cos(x_1+x_2)^2,\quad \alpha(\mathbf{x}) := (x_1+x_2)^2 e^{x_1-x_2},\quad \text{and}\quad b_1(\mathbf{x}) = b_2(\mathbf{x}) = 0.$$

*Problem 5.* (analogue of Problem 2)

$$\eta(\mathbf{x}) := 500x_1 + x_2,\quad \alpha(\mathbf{x}) := 1 + 9(x_1+x_2),\quad \text{and}\quad b_1(\mathbf{x}) = b_2(\mathbf{x}) = 0.$$

*Problem 6.* (analogue of Problem 3)

$$\eta \equiv 0,\quad \alpha(\mathbf{x}) = \begin{cases} 10^6 & \text{if } \|\mathbf{x} - [0.5\ 0.1]^T\| < 0.25, \\ 1 & \text{otherwise,} \end{cases} \quad \text{and}\quad b_1(\mathbf{x}) = b_2(\mathbf{x}) = 0.$$

We note that, for Problem 4, we added nonconstant reaction and diffusion coefficients (otherwise, omitting the advection term leads to the standard Poisson problem).

The same questions as before are of interest[12]. We fix $N = 50^2$ and display the convergence curves and the observed convergence factor $\rho_{\text{conv}}$ in Figure 4.6 (using the `chop` toolbox) and Figure 4.7 (using the `advanpix` toolbox). We draw conclusions very similar to those in Section 4.1. We note that the additional condition (4.16) was almost always weaker than (4.8) and comparable to (4.6). However, just as in the nonsymmetric case in Section 4.1, the differences were small (e.g., for `fp16` and `bfloat16` with `chop` or for neighboring precisions with `advanpix`).

In Figure 4.8, we display the error of the multiplicative Schwarz method at iterations 1, 2, and 3 for different choices of the precision $u_\ell$ for Problem 5 and observe, generally speaking, similar results to those in Figure 4.3. Our experience with dAS and RAS is fairly similar.

Looking at the observed convergence factors $\rho_{\text{conv}}$ in Figure 4.9, we find, similarly to the nonsymmetric case, that the dominant eigenmodes of $T_\star$ with $\star \in \{\text{AS}, \text{RAS}, \text{MS}\}$ appear to be well captured already with limited precision, e.g., $d_\ell = 4 \sim 6$, *and* that the other eigenmodes are not too sensitive with respect to small perturbations of the subdomain solves and remain "non-dominant".

The theoretical results only hold if all of the conditions (4.6), (4.8), and (4.16) hold, but the model problems suggest that either of the conditions (4.6) or (4.16) provides a good indicator. However, we note that the condition (4.16) becomes *much* more pessimistic if we omit the rescaling. For example, if we take $D_i^{(r,c)} = I_{N_i}$ for Problem 6, then (4.16) is satisfied only for $d_\ell \gtrapprox 10$, i.e., long after the convergence factor has stabilized at its final

[12] The same caveat regarding the loss of efficiency due to routine and naive swapping between rounding modes also applies; see page 588, below Theorem 4.1.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



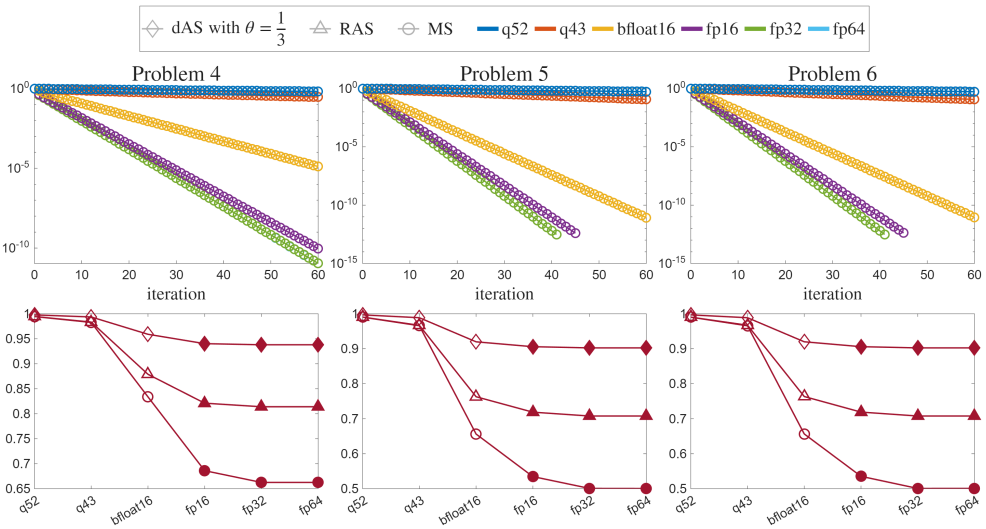



FIG. 4.6. *Top: the 2-norm of the error of the multiplicative Schwarz method for different choices of $u_\ell$ and $u_w$ corresponding to* `fp64`*, using the* `chop` *toolbox. For* `fp32` *and* `fp64`*, the graphs are indistinguishable from each other. Bottom: the observed convergence factor $\rho_{\text{conv}}$ for all Schwarz methods and choices of $u_\ell$; if the conditions* (4.6)*,* (4.8)*, and* (4.16) *are satisfied for both subdomains $i = 1, 2$ for a certain $u_\ell$, then the marker is filled. For example, for Problem 4, the conditions* (4.6)*,* (4.8)*, and* (4.16) *are satisfied for both $i = 1, 2$ starting from* `fp16`*.*

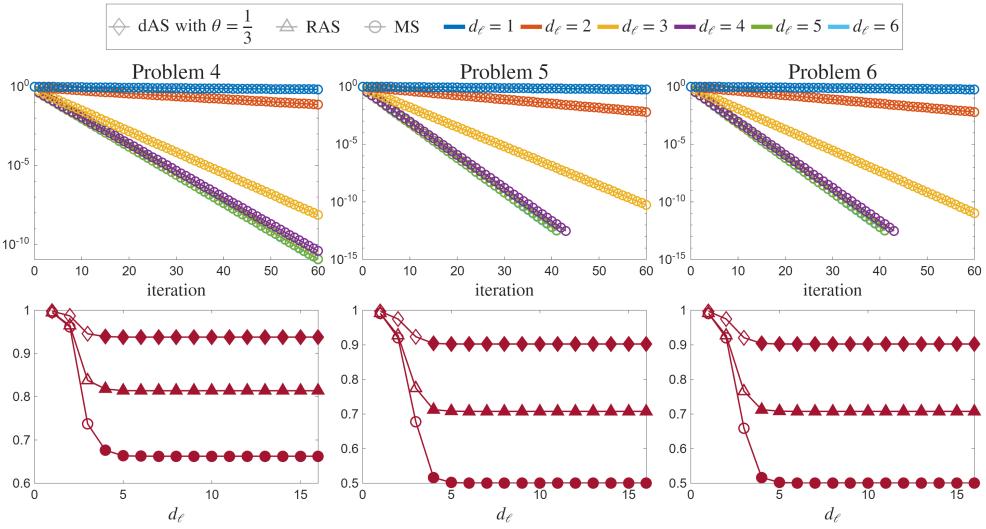



FIG. 4.7. *Top: the 2-norm of the error of the multiplicative Schwarz method for different choices of $d_\ell$ with $u_w$ corresponding to* `fp64`*, using the* `advanpix` *toolbox. Bottom: the observed convergence factor $\rho_{\text{conv}}$ for different methods and choices of $d_\ell$; if the conditions* (4.6)*,* (4.8)*, and* (4.16) *are satisfied for both subdomains $i = 1, 2$ for a certain $d_\ell$, then the marker is filled.*

value. The same is true if we replace condition (4.16) with a cruder version involving the condition number of $\mathcal{A}_i$; see (4.17) and below. Condition (4.6), however, has been fairly robust, localizing fairly accurately the optimal $d_\ell$ regardless of the scaling used. Also, similarly to the nonsymmetric case, the convergence-factor graph becomes notably "less smooth" but is otherwise qualitatively similar. The "nonsmoothness" of the convergence factor for Problem 6

ETNA
Kent State University and
Johann Radon Institute (RICAM)



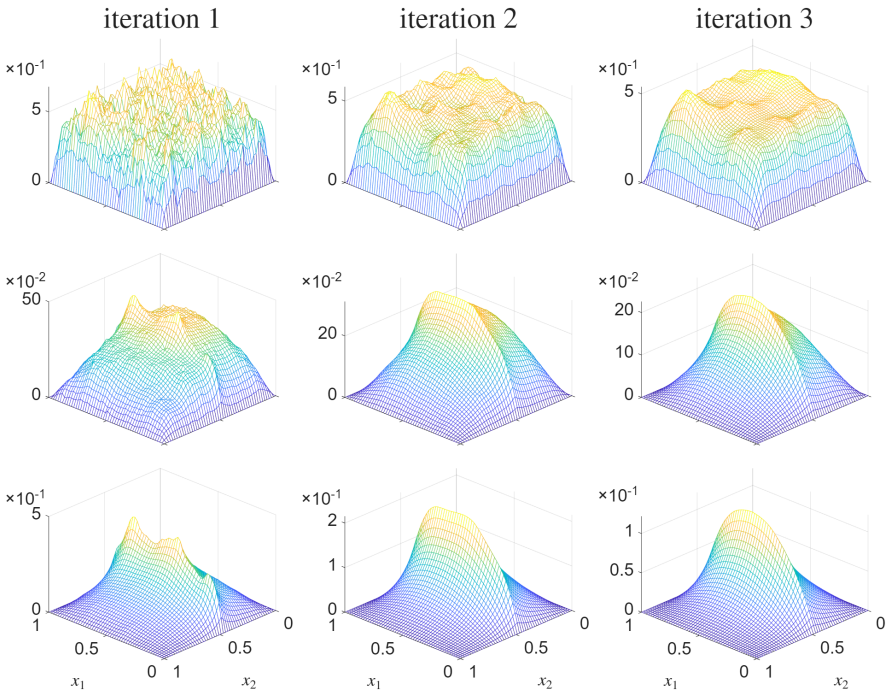



FIG. 4.8. *The errors of the multiplicative Schwarz method used for Problem 5 with $N = 50^2$ during the initial iterations using* `q43` *(top),* `bfloat16` *(middle), and* `fp16` *(bottom). Up to scaling, the error for* `q52` *is analogous to the top row, and the errors for* `fp32` *and* `fp64` *are are analogous to the bottom row.*

and the smallest mesh resolution, i.e., $N = 50^2$ also stands out. This is not due to the low-precision use—the algorithm has simply essentially converged after 60 iterations as we have $0.55^{60} \approx 2.6 \times 10^{-16}$. This is also easy to verify by inspecting the error plots directly.

Finally, we use the preconditioned GMRES method with multiprecision dAS[13], RAS, and MS as the left preconditioners, with relative residual tolerance $10^{-12}$, a zero initial approximation, and a maximum number of iterations set to 100. We display the GMRES convergence curves and the number of iterations in Figure 4.10. We see that the number of iterations again remains mostly stable as $d_\ell$ varies for a fixed $N$, and, moreover, the conditions (4.6) or (4.16) still serve as reasonably accurate indicators for the choice of the number of digits $d_\ell$.

**5. Comparison of the standard and the multiprecision additive Schwarz methods.** So far, we have studied multiprecision Schwarz methods as solvers and provided convergence conditions based on the iteration operator of the multiprecision Schwarz methods. In this section we briefly explore the natural complementary approach, namely, interpreting the multi-precision Schwarz methods as modifications of the full-precision methods using perturbation theory. We present the idea only for the additive Schwarz methods and leave a detailed analysis (also including other methods) as an open problem for future research.

We start by assuming the setup of Section 4.1, namely, similarly to (4.5), we write

$$\tilde{A}_i^{-1} - A_i^{-1} = \mu D_i^{(c)} \left( \tilde{\mathcal{A}}_i^{-1} - \mathcal{A}_i^{-1} \right) D_i^{(r)} = \mu D_i^{(c)} \left( (I + \mathcal{A}_i^{-1} \mathrm{F}_i)^{-1} - I \right) \mathcal{A}_i^{-1} D_i^{(r)},$$

[13] In practice, we would take advantage of the symmetry of dAS as a preconditioner and would run a left-preconditioned CG. Here we use GMRES simply to keep the preconditioner results comparable with those of all the other methods.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



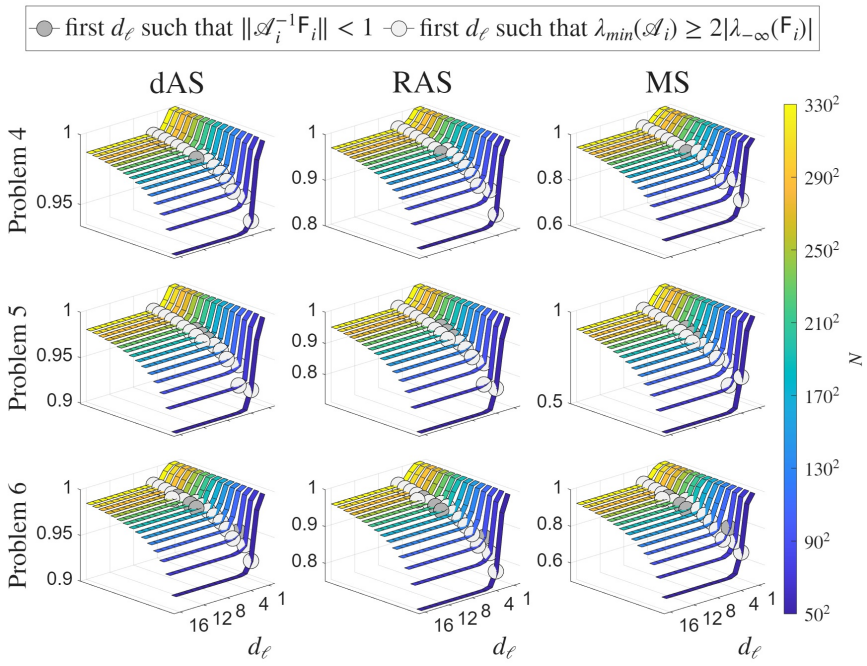



FIG. 4.9. *We display $\rho_{\text{conv}}$ for the Schwarz method for $d_\ell = 1, 2, \ldots, 16$ and different problem sizes $N$. We highlight the first $d_\ell$ for which conditions (4.6) and (4.16) are satisfied by ● and ○, respectively.*

and denoting $\mathcal{E}_i := D_i^{(c)} \left((I + \mathcal{A}_i^{-1}\mathsf{F}_i)^{-1} - I\right) (D_i^{(c)})^{-1}$, we obtain

$$\tilde{A}_i^{-1} - A_i^{-1} = \mathcal{E}_i A_i^{-1}.$$

As a result, if we assume[14]

$$\|\mathcal{A}_i^{-1}\mathsf{F}_i\| \leq \epsilon < \frac{1}{2}, \tag{5.1}$$

for some $\epsilon \in (0, 1/2)$, then

$$\|\mathcal{E}_i\| = \left\| D_i^{(c)} \left( \sum_{k=1}^{\infty} (-1)^k \left(\mathcal{A}_i^{-1}\mathsf{F}_i\right)^k \right) (D_i^{(c)})^{-1} \right\| \leq \kappa\left(D_i^{(c)}\right) \epsilon \frac{1}{1-\epsilon} < 2\epsilon\kappa\left(D_i^{(c)}\right),$$

and thus

$$\|\tilde{A}_i^{-1} - A_i^{-1}\| \leq 2\epsilon\kappa\left(D_i^{(c)}\right) \|A_i^{-1}\|,$$

where $\kappa(\cdot)$ denotes the condition number with respect to the norm $\|\cdot\|$. In words, a (small) perturbation of the scaled subdomain matrices can perturb the subdomain solves proportionally to the norm of the given subdomain solve and the condition number of the column-scaling matrix. Hence, we hope to obtain well-scaled subdomain matrices, for which the scaling can be mostly done by row scaling (recall that, conveniently, the row scaling takes precedence

[14] This assumption is analogous to (4.6). In fact, the derivations requiring (4.6) can be carried out analogously even if we assume (5.1) instead of (4.6), but the derivation becomes somewhat more lengthy.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



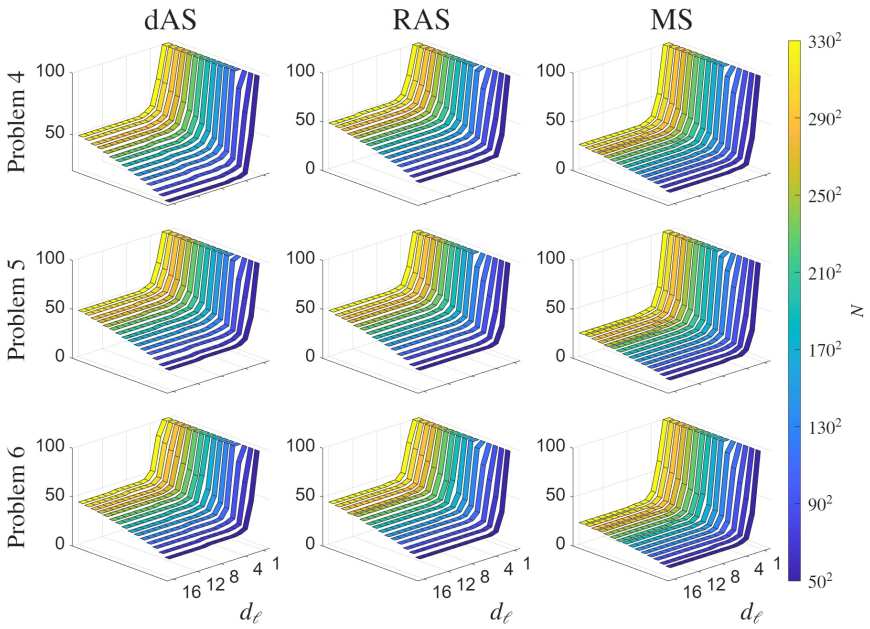



FIG. 4.10. *The number of preconditioned GMRES iterations to reduce the relative residual below* $10^{-12}$ *capped at 100.*

in [34, Algorithms 2.3 and 2.4]). In the case of (damped) AS, one of the main advantages is the symmetry, and hence symmetrical scaling is of particular interest, where we would also have to hope for a well-scaled matrix. Also, notice that (5.1) is similar to the assumption in [46, equation (12)] but for the inverses and after scaling.

Next, we insert (5.1) into the definition of the additive Schwarz methods in (2.3) and obtain

$$\tilde{M}_\star^{-1}A = M_\star^{-1}A + E_\star, \quad \text{where } E_\star := \begin{cases} \theta \sum_{i=1}^{p} R_i^T \mathcal{E}_i A_i^{-1} R_i A, & \text{for (damped) AS,} \\ \sum_{i=1}^{p} \overline{R}_i^T \mathcal{E}_i A_i^{-1} R_i A, & \text{for RAS.} \end{cases} \tag{5.2}$$

Recalling the matrix definitions in (2.1) and below, we can write

$$A_i^{-1} R_i A = A_i^{-1} \begin{bmatrix} I_{N_i} & 0 \end{bmatrix} \Pi_i \Pi_i^T \begin{bmatrix} A_i & K_i \\ L_i & A_{\neg i} \end{bmatrix} \Pi_i = \begin{bmatrix} I_{N_i} & A_i^{-1} K_i \end{bmatrix} \Pi_i,$$

and hence,

$$E_\star = \begin{cases} AS, \theta: & \theta \sum_{i=1}^{p} \Pi_i^T \begin{bmatrix} \mathcal{E}_i & \mathcal{E}_i A_i^{-1} K_i \\ 0 & 0 \end{bmatrix} \Pi_i, \\ RAS: & \sum_{i=1}^{p} \overline{\Pi}_i^T \begin{bmatrix} (\mathcal{E}_i)_{1:\overline{N}_i,:} & \left(\mathcal{E}_i A_i^{-1} K_i\right)_{1:\overline{N}_i,:} \\ 0 & 0 \end{bmatrix} \Pi_i. \end{cases}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



In fact, we can further rewrite this as

$$E_\star = \begin{cases} AS, \theta: & \theta \sum_{i=1}^{p} \Pi_i^T \begin{bmatrix} \mathcal{E}_i & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_{N_i} & A_i^{-1}K_i \\ 0 & 0 \end{bmatrix} \Pi_i, \\ RAS: & \sum_{i=1}^{p} \overline{\Pi}_i^T \begin{bmatrix} \begin{bmatrix} (\mathcal{E}_i)_{1:\overline{N}_i,:} \\ 0_{\overline{N}_i+1:N_i,:} \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_{N_i} & A_i^{-1}K_i \\ 0 & 0 \end{bmatrix} \Pi_i, \end{cases}$$

where we clearly see the main difference between the two methods in the "double counting of the overlap"; see [17, 19]. We outline the importance of such an analysis in the following remarks.

REMARK 5.1. In our opinion, a more detailed study of the *interaction* between the error introduced by low-precision subdomain solves and the structure of the Schwarz methods could lead to a more insightful analysis. In this way, we would be able to obtain better guidance for choosing the precision $u_\ell$.

REMARK 5.2. The formulation (5.2) allows us to use results from the convergence theory of Krylov subspace methods for perturbed systems. In particular, results such as [6, Theorem 1.1 and Corollary 1.2] or [47, Theorems 2.1 and 2.3] give direct recipes for bounding the convergence delay of GMRES due to the subdomain solves in the precision $u_\ell$, assuming that $E_\star$ is small in some sense. Perturbation results also exist for CG and can be useful for the (damped) AS; see the seminal papers [30, 50] and also [37, Section 5.9] for further references.

**6. Conclusion and future work.** We have proposed and analyzed multiprecision Schwarz methods specifically tailored to problems where we can guarantee the methods' convergence—namely, problems where the system matrix is an $M$-matrix. Using specific rounding techniques, we were able to preserve the convergence property and suggest several natural conditions for choosing a suitable precision depending on the problem. We presented several numerical experiments for PDE model problems that support our theoretical results and further illustrate that our proposed conditions identify the ideal precision for the subdomain solves. As future work we intend to consider generalizations to multiple subdomains and/or "interface conditions" in the sense of [24].

An understanding of the interaction between the subdomain matrices $A_i^{-1}K_i$ and the matrices $\mathcal{E}_i$ for all three classical Schwarz methods for a wider variety of problems would certainly be interesting, and we leave it as a possibility for future research. Also, it has been shown that it is often more suitable to bound the norm of the iteration matrix (and hence that of the error) over two or more iterations due to the nature of the underlying PDE analysis; see, e.g., [25, 24]. Exploiting this to get a better grasp on the multiprecision Schwarz methods as stand-alone solvers would be useful. Naturally, extending this analysis to preconditioning or rather understanding how to do that would also be of clear interest. This would be likely overlap with the so-called double-sweeping preconditioners and their analysis.

Finally, the authors would like to thank Petr Vacek for many stimulating and fruitful discussions on the topic of multiprecision computations, Theo Mary for enlightening discussions that led to our Remark 4.2, and Andreas Frommer for insightful feedback on the submitted manuscript, which led to its notable improvement.

REFERENCES

[1] ADVANPIX, *Multiprecision Computing Toolbox for MATLAB v.5.2.5.15470*, Advanpix LLC, Yokohama, Japan, 2025. https://www.advanpix.com/

**ETNA**
Kent State University and
Johann Radon Institute (RICAM)



[2] H. ANZT, J. DONGARRA, G. FLEGAR, N. J. HIGHAM, AND E. S. QUINTANA-ORTÍ, *Adaptive precision in block-Jacobi preconditioning for iterative sparse linear system solvers*, Concurrency Computat. Pract. Exper., 31 (2019), Paper No. e4460, 12 pages.

[3] M. ARIOLI AND I. S. DUFF, *Using FGMRES to obtain backward stability in mixed precision*, Electron. Trans. Numer. Anal., 33 (2008/09), pp. 31–44. https://etna.ricam.oeaw.ac.at/vol.33.2008-2009/pp31-44.dir/pp31-44.pdf

[4] M. BENZI, A. FROMMER, R. NABBEN, AND D. B. SZYLD, *Algebraic theory of multiplicative Schwarz methods*, Numer. Math., 89 (2001), pp. 605–639.

[5] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979. Reprinted by SIAM, Philadelphia, 1994.

[6] J. BLECHTA, *Stability of linear GMRES convergence with respect to compact perturbations*, SIAM J. Matrix Anal. Appl., 42 (2021), pp. 436–447.

[7] A. BUTTARI, N. J. HIGHAM, T. MARY, AND B. VIEUBLÉ, *A modular framework for the backward error analysis of GMRES*, IMA J. Numer. Anal., 46 (2026), pp. 1939–1993.

[8] X.-C. CAI, M. DRYJA, AND M. SARKIS, *Restricted additive Schwarz preconditioners with harmonic overlap for symmetric positive definite linear systems*, SIAM J. Numer. Anal., 41 (2003), pp. 1209–1231.

[9] X.-C. CAI AND M. SARKIS, *A restricted additive Schwarz preconditioner for general sparse linear systems*, SIAM J. Sci. Comput., 21 (1999), pp. 792–797.

[10] E. CARSON AND X. CHEN, *pychop: Emulating low-precision arithmetic in numerical methods and neural networks*, ACM Trans. Math. Softw., published online, (2026). https://doi.org/10.1145/3840284

[11] E. CARSON AND I. DAUŽICKAITĖ, *The stability of split-preconditioned FGMRES in four precisions*, Electron. Trans. Numer. Anal., 60 (2024), pp. 40–58. https://etna.ricam.oeaw.ac.at/vol.60.2024/pp40-58.dir/pp40-58.pdf

[12] E. CARSON AND N. J. HIGHAM, *Accelerating the solution of linear systems by iterative refinement in three precisions*, SIAM J. Sci. Comput., 40 (2018), pp. A817–A847.

[13] F. CUVELIER, M. J. GANDER, AND L. HALPERN, *Fundamental coarse space components for Schwarz methods with crosspoints*, in Domain Decomposition Methods in Science and Engineering XXVI, S. C. Brenner, E. Chung, A. Klawonn, F. Kwok, J. Xu, and J. Zou, eds., vol. 145 of Lect. Notes Comput. Sci. Eng., Springer, Cham, 2022, pp. 41–52.

[14] V. DOLEAN, M. J. GANDER, AND L. GERARDO-GIORDA, *Optimized Schwarz methods for Maxwell's equations*, SIAM J. Sci. Comput., 31 (2009), pp. 2193–2213.

[15] V. DOLEAN, M. J. GANDER, S. LANTERI, J.-F. LEE, AND Z. PENG, *Effective transmission conditions for domain decomposition methods applied to the time-harmonic curl-curl Maxwell's equations*, J. Comput. Phys., 280 (2015), pp. 232–247.

[16] V. DOLEAN, P. JOLIVET, AND F. NATAF, *An Introduction to Domain Decomposition Methods*, SIAM, Philadelphia, 2015.

[17] E. EFSTATHIOU AND M. J. GANDER, *Why restricted additive Schwarz converges faster than additive Schwarz*, BIT Numer. Math., 43 (2003), pp. 945–959.

[18] A. FROMMER AND D. B. SZYLD, *Weighted max norms, splittings, and overlapping additive Schwarz iterations*, Numer. Math., 83 (1999), pp. 259–278.

[19] ———, *An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms*, SIAM J. Numer. Anal., 39 (2001), pp. 463–479.

[20] ———, *On the convergence of randomized and greedy relaxation schemes for solving nonsingular linear systems of equations*, Numer. Algorithms, 92 (2023), pp. 639–664.

[21] M. J. GANDER, *Optimized Schwarz methods*, SIAM J. Numer. Anal., 44 (2006), pp. 699–731.

[22] ———, *Schwarz methods over the course of time*, Electron. Trans. Numer. Anal., 31 (2008), pp. 228–255. https://etna.ricam.oeaw.ac.at/vol.31.2008/pp228-255.dir/pp228-255.pdf

[23] M. J. GANDER AND L. HALPERN, *Piece-wise constant, linear and oscillatory: a historical introduction to spectral coarse spaces with focus on Schwarz methods*, in Domain Decomposition Methods in Science and Engineering XXVII, Z. Dostál, T. Kozubek, A. Klawonn, U. Langer, L. F. Pavarino, J. Šístek, and O. B. Widlund, eds., vol. 149 of Lect. Notes Comput. Sci. Eng., Springer, Cham, 2024, pp. 189–198.

[24] M. J. GANDER, S. LOISEL, AND D. B. SZYLD, *An optimal block iterative method and preconditioner for banded matrices with applications to PDEs on irregular domains*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 653–680.

[25] M. J. GANDER AND M. OUTRATA, *On algebraic bounds for POSM and MRAS*, in Domain Decomposition Methods in Science and Engineering XXVII, Z. Dostál, T. Kozubek, A. Klawonn, U. Langer, L. F. Pavarino, J. Šístek, and O. B. Widlund, eds., vol. 149 of Lect. Notes Comput. Sci. Eng., Springer, Cham, 2024, pp. 383–390.

[26] M. J. GANDER AND H. ZHANG, *A class of iterative solvers for the Helmholtz equation: factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods*, SIAM Rev., 61 (2019), pp. 3–76.

**ETNA**
**Kent State University and**
**Johann Radon Institute (RICAM)**



[27] ———, *Schwarz methods by domain truncation*, Acta Numer., 31 (2022), pp. 1–134.
[28] L. GIRAUD, A. HAIDAR, AND L. T. WATSON, *Mixed-precision preconditioners in parallel domain decomposition solvers*, in Domain Decomposition Methods in Science and Engineering XVII, U. Langer, M. Discacciati, D. E. Keyes, O. B. Widlund, and W. Zulehner, eds., vol. 60 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 2008, pp. 357–364.
[29] C. GLUSA, E. G. BOMAN, E. CHOW, S. RAJAMANICKAM, AND D. B. SZYLD, *Scalable asynchronous domain decomposition solvers*, SIAM J. Sci. Comput., 42 (2020), pp. C384–C409.
[30] A. GREENBAUM, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.
[31] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 2002.
[32] N. J. HIGHAM AND T. MARY, *Mixed precision algorithms in numerical linear algebra*, Acta Numer., 31 (2022), pp. 347–414.
[33] N. J. HIGHAM AND S. PRANESH, *Simulating low precision floating-point arithmetic*, SIAM J. Sci. Comput., 41 (2019), pp. C585–C602.
[34] N. J. HIGHAM, S. PRANESH, AND M. ZOUNON, *Squeezing a matrix into half precision, with an application to solving linear systems*, SIAM J. Sci. Comput., 41 (2019), pp. A2536–A2551.
[35] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1994.
[36] IEEE, *IEEE Standard for floating-point arithmetic*, IEEE Std 754-2019 (Revision of IEEE 754-2008), (2019), pp. 1–84.
[37] J. LIESEN AND Z. STRAKOŠ, *Krylov Subspace Methods*, Oxford University Press, Oxford, 2013.
[38] P. L. LIONS, *On the Schwarz alternating method. I.*, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations (Paris, 1987), R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., SIAM, Philadelphia, 1988, pp. 1–42.
[39] F. MAGOULÈS, D. B. SZYLD, AND C. VENET, *Asynchronous optimized Schwarz methods with and without overlap*, Numer. Math., 137 (2017), pp. 199–227.
[40] J. M. ORTEGA, *Numerical Analysis*, 2nd ed., SIAM, Philadelphia, 1990.
[41] K. OZAKI, T. OGITA, S. OISHI, AND S. M. RUMP, *Error-free transformations of matrix multiplication by using fast routines of matrix multiplication and its applications*, Numer. Algorithms, 59 (2012), pp. 95–118.
[42] ———, *Generalization of error-free transformation for matrix multiplication and its application*, Nonlinear Theory Appl., 4 (2013), pp. 2–11.
[43] K. OZAKI, Y. UCHINO, AND T. IMAMURA, *Ozaki scheme II: A GEMM-oriented emulation of floating-point matrix multiplication using an integer modular technique*, Int. J. High Performance Comput., in press, 2026. https://doi.org/10.1177/10943420261467
[44] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[45] M. SARKIS AND M. DRYJA, *Convergence bounds for one-dimensional ASH and RAS*, in Domain Decomposition Methods in Science and Engineering XXVII, Z. Dostál, T. Kozubek, A. Klawonn, U. Langer, L. F. Pavarino, J. Šístek, and O. B. Widlund, eds., vol. 149 of Lect. Notes Comput. Sci. Eng., Springer, Cham, 2024, pp. 39–50.
[46] J. SCHNECK, M. WEISER, AND F. WENDE, *Impact of mixed precision and storage layout on additive Schwarz smoothers*, Numer. Linear Algebra Appl., 28 (2021), Paper No. e2366, 22 pages.
[47] J. A. SIFUENTES, M. EMBREE, AND R. B. MORGAN, *GMRES convergence for perturbed coefficient matrices, with application to approximate deflation preconditioning*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1066–1088.
[48] B. SMITH, P. BJØRSTAD, AND W. GROPP, *Domain Decomposition*, Cambridge University Press, Cambridge, 1996.
[49] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[50] Z. STRAKOŠ, *On the real convergence rate of the conjugate gradient method*, Linear Algebra Appl., 154/156 (1991), pp. 535–549.
[51] N. TIAN, S. HUANG, AND X. XU, *Mixed precision block-Jacobi preconditioner: algorithms, performance evaluation and feature analysis*, CCF Trans. High Performance Comput., 7 (2025), pp. 114–128.
[52] A. TOSELLI AND O. WIDLUND, *Domain Decomposition Methods—Algorithms and Theory*, Springer, Berlin, 2005.
[53] O. WIDLUND AND M. DRYJA, *An additive variant of the Schwarz alternating method for the case of many subregions*, Tech. Rep. 339, Ultracomputer Note 131, Department of Computer Science, Courant Institute, New York University, 1987.