Electronic Transactions on Numerical Analysis.
Volume 46, pp. 1–xx, 2017.

ISSN 1068–9613.
DOI:

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 5 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-matrix, convergence analysis

**AMS subject classifications.** 65F10, 65F08, 65Y99, 65Y99

## 1. Introduction.

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

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

In particular, we analyze the use of multiprecision arithmetic for three types of Schwarz methods (see [21]): (i) the *(damped) additive Schwarz method* ((d)AS, see [35, 50]), (ii) the *restricted additive Schwarz method* (RAS, see [7]), and (iii) the *multiplicative Schwarz method* (MS, see [35]); see [30] for an overview of multiprecision/mixed-precision algorithms. Specifically, we study the solution of the local problems (see precise definitions below) using lower precision. To the best of our knowledge, this is the first time this approach is analyzed for Schwarz methods.

The underlying idea of Schwarz methods, as a 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 sub-dividing 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., [14, 46, 48]. For the algebraic error analysis of the broader class of stationary iterative methods, we refer the reader to [29, Chapter 17] and the references therein.

For *algebraic Schwarz methods*, where 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., [3, 17, 18]), 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 behavior; see, e.g., [11, 12, 13, 20, 24, 25]. Importantly, these methods are usually used as *preconditioners*, i.e., their convergence is further accelerated using Krylov subspace methods. But in our experience, in order to obtain more insight it is very often very useful to *first* study the Schwarz methods (or other domain decomposition methods) *as stand-alone solvers*. Then, based on

*Received... Accepted... Published online on... Recommended by....
Work supported by the PRIMUS grant PRIMUS/25/SCI/022 of Charles University.
†Department of Numerical Mathematics, Charles University, Prague, Czech republic.
‡Department of Mathematics, Temple University, Philadelphia, PA 19122, USA.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



their analysis, we can obtain an insight into or 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 be then used to propose an improvement such as a coarse space, see, e.g., [11, 22] and also [43]. Schwarz methods as solvers are also fundamental for Schwarz asynchronous iterations; see, e.g., [28, 36].

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 in the computationally most challenging part of the algorithm, see, e.g., [30] 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 [10], using multiprecision algorithms, where different parts of the algorithm are carried out in different precision, 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 [29, 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 $u_w$ may underflow/overflow in $u_\ell$, see, e.g., [32]. We comment on the specifics of multiprecision computations relevant to our interest in Section 3.

We note that similar settings has been already considered, e.g., in [27], the authors use the (non-overlapping) AS as a preconditioner for the conjugate gradient method (CG), using single precision for the preconditioner solve (i.e., running the AS in single precision) and the rest of the CG algorithm in double precision. In [2], the authors take the (non-overlapping) block-Jacobi as a preconditioner for CG and based on the 1-norm condition number of each of the diagonal blocks they calculate their inverses in half, single, or double precision and then apply these using dense mat-vec products in parallel for each block; similar methodology has been used also in [47] and tested for practically relevant and challenging 2D and 3D problems. Similarly, in [44], the authors study (overlapping, with coarse space) AS as the preconditioner and the effect of using different precision 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. Note that the focus in all of these papers is on numerical experiments and observations *about the preconditioned CG*, i.e., the interaction of the domain decomposition method and the different precision choices is present *only implicitly*. The analysis focuses on the (often questionable, see [34, Section 5.6, Corollary 5.6.7 and onward]) condition number bound for CG for the preconditioned system and is not interested in 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 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.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



Thus, our contribution consists of analyzing Schwarz methods where the local problems are solved with lower precision. Our analysis provides sufficient conditions, and when these conditions are met, one can calculate the minimum number of digits needed in the approximation to the solution of the local problem to obtain overall convergence. Our experiments with multiple type of discretized partial differential equations indicate that our conditions are satisfied with about 5 digits of accuracy. The computations shown illustrate that this is indeed the case.

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

**2. Algebraic Schwarz methods.** Consider $p$ subspaces $\overline{W}_i \subset \mathbb{R}^N, i = 1 \dots p$ that form a non-overlapping decomposition of $\mathbb{R}^N$, i.e.,

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

and $\overline{W}_i \cap \overline{W}_j = \{0\}$ if $i \neq j$, and we denote their dimensions by $\bar{N}_i := \dim(\overline{W}_i)$. If the problem (1.1) corresponds to a discretization on some grid on a domain $\Omega$, then $\overline{W}_i$ are often subspaces corresponding to the unknowns in physical subdomains $\Omega_i \subset \Omega$. We set the restriction operators $\bar{R}_i : \mathbb{R}^N \to \mathbb{R}^{\bar{N}_i}$, corresponding to $\bar{N}_i$-by-$N$ zero-one matrices with full row rank $\bar{N}_i$, and obtain the prolongation operators as the transpose of the restrictions, i.e., $\bar{R}_i^T : \mathbb{R}^{\bar{N}_i} \to \mathbb{R}^N$. We assume that the restriction matrices are chosen so that

$$\bar{R}_i = \left[ I_{\bar{N}_i} 0 \right] \overline{\Pi}_i \in \mathbb{R}^{\bar{N}_i \times N}, \tag{2.1}$$

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

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

We also consider the analogous objects for the *overlapping* case $\overline{W}_i \subset W_i$ (by enlarging each of the subspaces $\overline{W}_i$ and omitting the bar in the notation) and set $N_i := \dim(W_i)$ and the matrices

$$R_i = \left[ I_{N_i} 0 \right] \Pi_i \in \mathbb{R}^{N_i \times N} \in \mathbb{R}^{N_i \times N} \quad \text{and} \quad R_i^T R_i = \Pi_i^T \begin{bmatrix} I_{N_i} & 0 \\ 0 & 0 \end{bmatrix} \Pi_i \in \mathbb{R}^{N \times N},$$

see, e.g., [7, 18] and also [15] for more details. 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 we denote the complement of the indices in the range of $R_i$ by $\neg i$, e.g.,

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

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

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

ETNA
Kent State University and
Johann Radon Institute (RICAM)



where the blocks $\times$ and $*$ can be, in general, chosen arbitrarily but the common choice is to set $\times = 0$ and $* = A_{\neg i}$ ([17]) or $* = \operatorname{diag}(A_{\neg i})$ [3, 18]; 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

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

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*

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

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 $I - M_\star^{-1} A$ so as 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 $M_\star^{-1}$, where the inverse highlights that these preconditioners are to be "applied" rather then "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 straight-forward while for the multiplicative Schwarz the definition becomes seemingly artificial by having

$$M_{\mathrm{MS}}^{-1} = \left( I_N - T_{\mathrm{MS}} \right) A^{-1}, \tag{2.4}$$

which can be further reformulated for practical use, see, e.g., [42, 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 say that a matrix $A$ is *symmetric, positive-definite* or *SPD* (denoted by $A \succ 0$), provided that

$$A^T = A \quad \text{and} \quad \mathbf{v}^T A \mathbf{v} > 0 \text{ for all } \mathbf{v} \neq 0. \tag{2.5}$$

Denoting the spectrum of $A$ by $\sigma(A)$, (2.5) is equivalent to

$$A^T = A \quad \text{and} \quad \lambda > 0 \text{ for all } \lambda \in \sigma(A).$$

We say that $A$ is a *nonsingular $M$-matrix*, provided that the off-diagonal elements of $A$ are non-positive and all elements of the inverse are non-negative, i.e., $A - \operatorname{diag}(A) \leq 0$ and $A^{-1} \geq 0$, where the inequalities are understood element-wise, see [4, Chapter 6] or [33, Section 2.5] for further details and references. In the rest of the paper we will simply say *an $M$-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.

THEOREM 2.1 ([17, Lemma 2.8], [3, 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 $W_i \cap W_j \neq \{0\}$, then $W_i$ and $W_j$ have different colors. Then*

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

*An equivalent multi-splitting formulation of these can be found in [18, Section 2] and [3, Sections 2 and 3].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



THEOREM 2.2 ([17, Theorem 3.4], [18, Theorem 4.4], [3, 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 $W_i \cap W_j \neq \{0\}$, then $W_i$ and $W_j$ have different colors. Then*

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

REMARK 2.3. We note that Theorem 2.2 seemingly guarantees RAS convergence even for "no overlap" case where $W_i = \overline{W}_i$ for some (or all) $i$. At first this might seem contradictory to the analytic results for the standard Laplace test problem, where $W_i$ corresponds to discretization of the problem on $\Omega_i$. But there is no contradiction as even the "no overlap" case in the algebraic sense corresponds to the "$h$ overlap" in the analytic sense of the subdomains $\Omega_i$, thanks to the Dirichlet boundary condition enforcement.

We also note that there are other methods closely related to the metnioned ones, e.g., the RAS method has number connected variants (e.g., WRAS,ASH,RASH,WRASH, see [18, Section 6] for further references). We do not consider them in this paper.

**3. Multiprecision computations.** As we do *not* work with hardware with a wider selection of precision, the different precisions in our multiprecision algorithms needs to be simulated in some way. The number of options available is limited and both theoretically and practically, two stand out – the `chop` package [31] and the `advanpix` package [37], both implemented in MATLAB. To the best of our knowledge, these are considered the golden standard among the available software for *simulating* various precisions in the numerical analysis and scientific computing community.

*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 and the definition of the standard precisions `double`, `single` and `half`, see [1]. Say we want to simulate a `half` precision (`fp16`), which corresponds to $u_{\text{half}} \approx 4.88 \times 10^{-4}$. Using `advanpix`, we have to chose 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, this allows to explore also "new" precisions which are not yet standardized or even used, e.g., six or eleven accurate digits, and frames the computation precision as more of a "integer-continuous" parameter. Moreover, the package is a highly optimized software that overwrites the standard (also highly optimized) 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 too computationally demanding (hence the commercial success of this package). In this context we would also like to highlight that a lot of interest has been recently devoted to *efficient* simulation of arbitrary precisions on GPUs with astonishing results. For example, although the hardware of GPUs is highly optimized only for low-precisions, such as `fp32`, `fp16` and even lower, a clever way of simulation of `fp64` on these GPUs *using these low-precision formats* was competitive with (or even preferable to) the standard hardware implementation of `fp64`, see [39, 40, 41] and the references therein. This opens doors to real possibility of efficiently simulating "new" low-precisions in practice.

*chop toolbox.* The `chop toolbox` is an open-source MATLAB toolbox* developed for simulating different precisions using the native `double` 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 precision (see [31, Section 3.1]) and can also simulate the

*Towards the end of preparing the manuscript, the `chop toolbox` has been also released for `python`, see [9].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



underflow/overflow during the computation. Although this toolbox outperforms many other options (see [31, 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 below.

| | Signif. | Exp. | $u$ | $x_{min}$ | $x_{max}$ |
|---|---|---|---|---|---|
| `q52` | 5 | 2 | $1.25 \times 10^{-1}$ | $6.10 \times 10^{-5}$ | $5.73 \times 10^{4}$ |
| `q43` | 4 | 3 | $6.25 \times 10^{-2}$ | $1.56 \times 10^{-2}$ | $2.40 \times 10^{2}$ |
| `bfloat16` | 8 | 8 | $3.91 \times 10^{-3}$ | $1.18 \times 10^{-38}$ | $3.39 \times 10^{38}$ |
| `fp16` | 11 | 5 | $4.88 \times 10^{-4}$ | $6.10 \times 10^{-5}$ | $6.55 \times 10^{4}$ |
| `fp32` | 24 | 8 | $5.96 \times 10^{-8}$ | $1.18 \times 10^{-38}$ | $3.40 \times 10^{38}$ |
| `fp64` | 53 | 11 | $1.11 \times 10^{-16}$ | $2.23 \times 10^{-308}$ | $1.80 \times 10^{308}$ |

TABLE 3.1

In most cases, the computationally most demanding part of Schwarz methods are the subdomain solves, i.e., the operations including $A_i^{-1}$. The issue of underflow/overflow is, in our opinion, an important piece in multiprecision calculations and has been, at least partially, addressed in [32], where the authors propose re-scaling procedures so that (close to) the full range of a given precision is utilized (demonstrated for `fp16`). To be concrete, having 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 $D_i^r$ and $D_i^c$ (corresponding to $R$ and $S$ in [32]). For any non-singular $D_i^r$ and $D_i^c$ we then rewrite (3.1) as

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

with

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

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

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

A reasonable choice then is to take $D_i^r$ and $D_i^c$ as in [32, Algorithms 2.3 and 2.4], i.e., as the maximum norms of the rows (and then the columns) of $A_i$. According to [32, Table 4.5], the choice of $\nu = 0.1$ (the 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

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

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 rewrote (3.1) into

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

ETNA
Kent State University and
Johann Radon Institute (RICAM)



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

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

Importantly, the rescaling preserves 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 be also adapted to preserve symmetry (see [32, Algorithm 2.5]) and then it also automatically preserves diagonal dominance.

REMARK 3.1. Since the `advanpix` toolbox is not open source, possible low-precision overflow/underflow appearances, e.g., during the LU factorization, are treated automatically and without the user's knowledge. In other words, it is fair to say that in spite our best efforts, many experiments are carried out without overflow/underflow errors, although we carry out the calculations so as to minimize their appearances by appropriate scaling.

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

We approach the problem from an algebraic point of view, inspired by the results in [3, 17, 18], 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 $u_w$. This direction is not explicitly mentioned in either of the works but follows from the sections focusing on inexact solves; see [17, Sections 2 and 3], [18, Section 7] and [3, Section 4]. Following the notation there, we will denote with tildes *quantities that have been obtained by precision-reduction in some sense*, e.g., 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 $u_\ell$, the new matrix will be denoted by $\tilde{A}_i$ and replacing all $\mathcal{A}_i$ with $\tilde{\mathcal{A}}_i$ in the definition of $A_i$ or $T_\star$ gives us $\tilde{A}_i$ or $\tilde{T}_\star$ (for $\star \in \{\mathrm{AS}, \mathrm{RAS}, \mathrm{MS}\}$). We emphasize that the symbol $\sim$ does not mean that the quantity was obtained by the classic 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 of these cases (using $\sim$) to highlight the lower-precision nature. We denote the error in the subdomain matrices by $\mathtt{E}_i$, i.e., we have

$$\tilde{A}_i = A_i + \mathtt{E}_i. \tag{4.1}$$

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

$$A_i = \tilde{A}_i - \left(\tilde{A}_i - A_i\right), \quad i = 1, \ldots, p\,. \tag{4.2}$$

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

$$\tilde{A}_i^{-1} \geq 0 \quad \text{and} \quad \tilde{A}_i^{-1}\left(\tilde{A}_i - A_i\right) = \tilde{A}_i^{-1}\mathtt{E}_i \geq 0. \tag{4.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



These conditions characterize when the splitting (4.2) is weak regular (of the first type, see [18, Section 4]) and thus if (4.3) holds for all $i = 1, \ldots, p$, 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 a rounding technique of our choice applied to $\mathcal{A}_i$ (and $\hat{\mathbf{b}}_i$). In other words, we have

$$\tilde{A}_i = \mu^{-1}(D_i^r)^{-1}\tilde{\mathcal{A}}_i(D_i^r)^{-1} \tag{4.4}$$

and so 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 for diagonal re-scaling of general stationary iterative methods* – 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 $D_1AD_2$ are similar (i.e., have the same convergence factor) for any non-singular diagonal matrices $D_1, D_2$. On one hand, this shows that for $u_\ell = u_w$, the re-scaling above doesn't affect the asymptotic convergence rate in our case. On the the other, 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 re-scaling.

Revisiting (4.3), a direct calculation shows that the weak regular splitting conditions are invariant with respect to diagonal scaling with positive entries, i.e., the conditions (4.3) are equivalent to

$$\tilde{\mathcal{A}}_i^{-1} \geq 0 \quad \text{and} \quad \tilde{\mathcal{A}}_i^{-1}\left(\tilde{\mathcal{A}}_i - \mathcal{A}_i\right) = \tilde{\mathcal{A}}_i^{-1}\mathtt{F}_i \geq 0. \tag{4.5}$$

The first ingredient for the analysis of (4.5) is rewritting $\tilde{\mathcal{A}}_i^{-1}$ as

$$\tilde{\mathcal{A}}_i^{-1} = \mathcal{A}_i^{-1}\left(I + \mathtt{F}_i\mathcal{A}_i^{-1}\right)^{-1}, \tag{4.6}$$

and expanding the inverse matrix there into its Neumann series under the assumption

$$\|\mathcal{A}_i^{-1}\mathtt{F}_i\| < 1, \tag{4.7}$$

in some induced norm. Assuming (4.7), the Neumann serie expansion reads

$$\tilde{\mathcal{A}}_i^{-1} = \mathcal{A}_i^{-1}\sum_{k=0}^{+\infty}\left(-\mathtt{F}_i\mathcal{A}_i^{-1}\right)^k, \tag{4.8}$$

and we further rearrange it as

$$\begin{aligned}\tilde{\mathcal{A}}_i^{-1} &= \mathcal{A}_i^{-1}\sum_{k=0}^{+\infty}\left(-\mathtt{F}_i\mathcal{A}_i^{-1}\right)^k = \mathcal{A}_i^{-1}\left(I - \mathtt{F}_i\mathcal{A}_i^{-1}\right) + \mathcal{A}_i^{-1}\left(I - \mathtt{F}_i\mathcal{A}_i^{-1}\right)\mathtt{F}_i\mathcal{A}_i^{-1}\mathtt{F}_i\mathcal{A}_i^{-1} + \ldots \\ &= \mathcal{A}_i^{-1}\left(I - \mathtt{F}_i\mathcal{A}_i^{-1}\right)\sum_{k=0}^{+\infty}\left(\mathtt{F}_i\mathcal{A}_i^{-1}\right)^{2k} = \left(\mathcal{A}_i^{-1} - \mathcal{A}_i^{-1}\mathtt{F}_i\mathcal{A}_i^{-1}\right)\sum_{k=0}^{+\infty}\left(\mathtt{F}_i\mathcal{A}_i^{-1}\right)^{2k}.\end{aligned} \tag{4.9}$$

*We came across this observation in [29, Section 17.2, below eqn. (17.3)] but this is likely not the original reference.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



In order to ensure (4.5) we will 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. In fact, assuming $\mathtt{F}_i \geq 0$ the natural condition for ensuring also $\tilde{\mathcal{A}}_i^{-1} \geq 0$ (and hence (4.5)) becomes

$$\mathcal{A}_i^{-1} \geq \mathcal{A}_i^{-1}\mathtt{F}_i\mathcal{A}_i^{-1}, \tag{4.10}$$

a second condition on the choice of $u_\ell$ in addition to (4.7). Notice that both (4.7) as well as (4.10) are in some sense generalizations of the standard relative rounding error assumption

$$|\mathtt{F}_i| \leq u_\ell |\mathcal{A}_i|, \tag{4.11}$$

which is generally guaranteed (the absolute value is to be understood component-wise). Also, similarly to [2, Sections 4 and 5], both (4.7) and (4.10) invite us to choose $(u_\ell)_i$ for each subdomain independently, based on the relevant quantities or their estimates*. We keep (4.7) and (4.10) as assumptions, coupling the subdomain problems and the choice of the lower precision $u_\ell$ and move our attention to the condition $\mathtt{F}_i \geq 0$.

If we use the standard rounding, then we are unlikely to satisfy $\mathtt{F}_i \geq 0$ except for some special cases. However, the process of rounding is very often fully under our control. Since by definition the off-diagonal entries of $\mathcal{A}_i$ are non-positive while its diagonal entries are non-negative, 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)†. We then introduce the rounding procedure $\mathrm{round}_{Mmtrx}()$ that for any matrix $X$ gives its low-precision approximation $\mathrm{round}_{Mmtrx}(X)$ given by

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

Taking

$$\tilde{\mathcal{A}}_i = \mathrm{round}_{Mmtrx}(\mathcal{A}_i), \tag{4.12}$$

we get $\mathtt{F}_i \geq 0$ and obtain a convergent multiprecision Schwarz methods under the assumptions (4.7) and (4.10); we summarize these results in Theorem 4.1 below.

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 $W_i \cap W_j \neq \{0\}$, 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 a precision $u_w$ with the subdomain solver $\tilde{A}_i^{-1}$ in a precision $u_\ell$, with $u_w < u_\ell$, 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 (4.4) with $\tilde{\mathcal{A}}_i$ given as in (4.12), if (4.7) and (4.10) are satisfied, then*

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

*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 aspect, which has been covered in [2] and we do not comment further on how to choose $u_\ell$ (or $(u_\ell)_i$) for the subdomain problems.

†In the `chop` toolbox, these are already implemented and for the `advanpix` package, these are straight-forward 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)



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

Following [17, Sections 4], [18, Section 7] and [3, Section 4], we also obtain the comparisons for different choices of $u_\ell$. To be more specific, having an $M$-matrix $X$ and two different low-precisions $u_\ell^{(1)} \leq u_\ell^{(2)}$ with $u_w \leq u_\ell^{(1)} \leq u_\ell^{(2)}$, 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

$$\rho\left(\tilde{T}_{\star, u_\ell^{(1)}}\right) \leq \rho\left(\tilde{T}_{\star, u_\ell^{(2)}}\right), \quad \text{where } \star \in \{\mathrm{AS}, \mathrm{RAS}, \mathrm{MS}\}. \tag{4.13}$$

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

- *Are the conditions* (4.7) *and* (4.10) *in some sense sharp or descriptive in the context of convergence of the multiprecision Schwarz methods?*
- *When is* (4.13) *strict?*
- *Out of those $u_\ell$ for which* (4.13) *is strict, which should we use, i.e., to what extend are there diminishing returns as we approach equality in* (4.13)*?*

Next, we investigate these questions numerically on three model problems coming from a discretization of a reaction-advection-diffusion equation. Taking the unit square, i.e., $\Omega = \{\mathbf{x} = [x_1, x_2]^T \in [0,1]^2\}$, we consider the partial differential equation

$$\mathcal{L}u = f \quad \text{in } \Omega \quad \text{and} \quad u = g \quad \text{on } \partial\Omega, \tag{4.14}$$

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

$$\mathcal{L}u := \eta(\mathbf{x})u - \mathrm{div}\,(\alpha(\mathbf{x})\nabla u) + \mathbf{b}(\mathbf{x}) \cdot \nabla u. \tag{4.15}$$

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 [26, Figure 2.1])

$$\eta(\mathbf{x}) := x_1^2 \cos(x_1 + x_2)^2, \;\; \alpha(\mathbf{x}) := 20(x_1 + x_2)^2 e^{x_1 - x_2} \quad \text{and} \quad \begin{aligned} b_1(\mathbf{x}) &:= x_2 - 0.5, \\ b_2(\mathbf{x}) &:= x_1 - 0.5. \end{aligned}$$

*Problem 2.* (inspired by [19, Section 4.1])

$$\eta \equiv 0, \;\; \alpha \equiv 1 \quad \text{and} \quad \begin{aligned} b_1(\mathbf{x}) &:= \beta(x_1(x_1 - 1)(1 - 2x_2)), \\ b_2(\mathbf{x}) &:= -\beta(x_2(x_2 - 1)(1 - 2x_1)), \end{aligned} \quad \text{with} \quad \beta = 100.$$

*Problem 3.* (based on Problem 2)

$$\eta \equiv 0, \;\; \alpha(\mathbf{x}) = \begin{cases} 10^6 & \text{if } \|\mathbf{x} - [0.5\;0.1]^T\| < 0.25, \\ 1 & \text{otherwise,} \end{cases} \quad \text{and} \quad \begin{aligned} b_1(\mathbf{x}) &:= \beta(x_1(x_1 - 1)(1 - 2x_2)), \\ b_2(\mathbf{x}) &:= -\beta(x_2(x_2 - 1)(1 - 2x_1)), \end{aligned}$$

again with $\beta = 100$. To discretize we use the standard 5-point stencil finite difference scheme, adapting some of the code from [26] and obtain systems of linear equations (1.1) with $A$ being a non-symmetric $M$-matrix. We then partition $A$ into two overlapping subdomain problems,

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



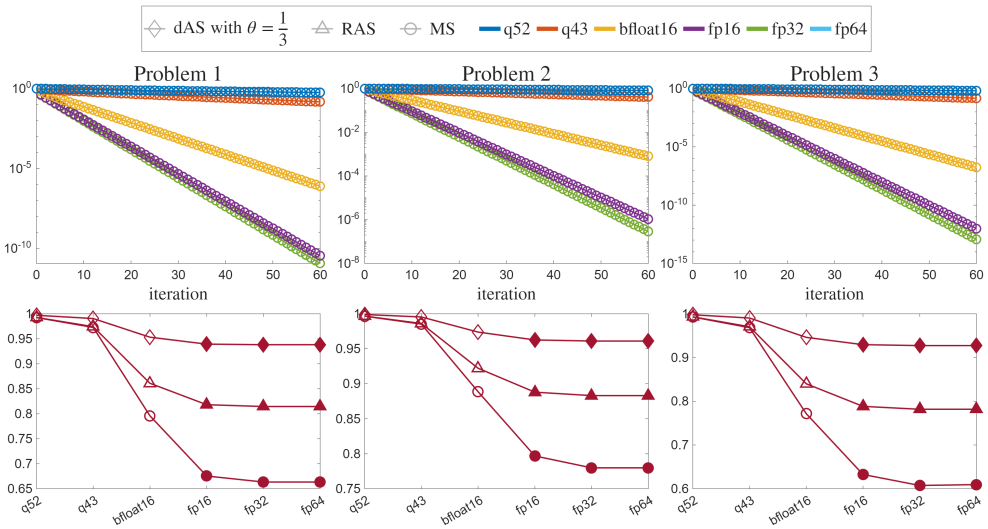


FIG. 4.1. *Top: the 2-norm of the error of the multiplicative Schwarz method for different choices of $u_w$ using the* `chop` *toolbox. For* `fp32` *and* `fp64` *the graphs are indiscernible from each other. Bottom: the observed convergence factor $\rho_{\text{conv}}$ for different methods and choices of $u_\ell$; if the conditions (4.7) and (4.10) are satisfied for both subdomain $i = 1, 2$ for a certain $u_\ell$, then the marker is filled. For example, for Problem 1 the conditions (4.7) and (4.10) are satisfied for both $i = 1, 2$ starting from* `fp16`.

taking the size of the overlap block to correspond to the bandwidth of $A$, i.e., we consider two overlapping subdomains $\Omega_1, \Omega_2 \subset \Omega$ with overlap width* $\mathcal{O}(h)$. We take our right-hand side vector $\mathbf{f}$ and our initial approximation vector $\mathbf{u}^{(0)}$ as random vectors with entries in $(0, 1)$.

First we fix $N = 2500$ and show the convergence curves and the observed convergence factor $\rho_{\text{conv}}$ in Figure 4.1 for the multiplicative Schwarz method and the standard low-precision formats in the `chop` package (see Table 3.1 above), adjusting the scaling from Section 3 so as to use as much of the available range of each precision while not overflowing during the computations. We see that the methods in fact converge in all of the considered precisions, although the conditions (4.7) and (4.10) are satisfied only for `fp16`, `fp32` and `fp64`. Moreover, once the conditions (4.7) and (4.10) are satisfied, they are also satisfied for higher precisions and, more importantly, the observed convergence factor $\rho_{\text{conv}}$ (calculated based on the convergence curves) essentially becomes invariant to increasing the precision further. In other words, we get *very little* additional computational benefits (within the first 60 iterations) by considering higher precisions once the conditions (4.7) and (4.10) are satisfied. This suggests that the conditions (4.7) and (4.10) offer a good guidance on the a-priori choice of the working proecision $u_w$.

We illustrate this further with the analogous experiment but run using the `advanpix` toolbox, which allows us finer tuning of the considered precision for the price of foregoing the control over underflow/overflow situation (however since we encountered no overflow with `chop`, this seems not too worrying) in Figure 4.2. We see that not only the observations from Figure 4.1 still hold true, but the case for the use of the conditions (4.7) and (4.10) as predictors for the suitable precision $u_\ell$ is further strengthened.

REMARK 4.2. Numerically, the experiments suggest that (4.7) is generally weaker than (4.10), although we have not been able to establish this as a theoretical result. However,

*As a result, we expect the convergence factor of Schwarz method to deteriorate as $N$ increases, see [18, Section 5] and [3, Section 5].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



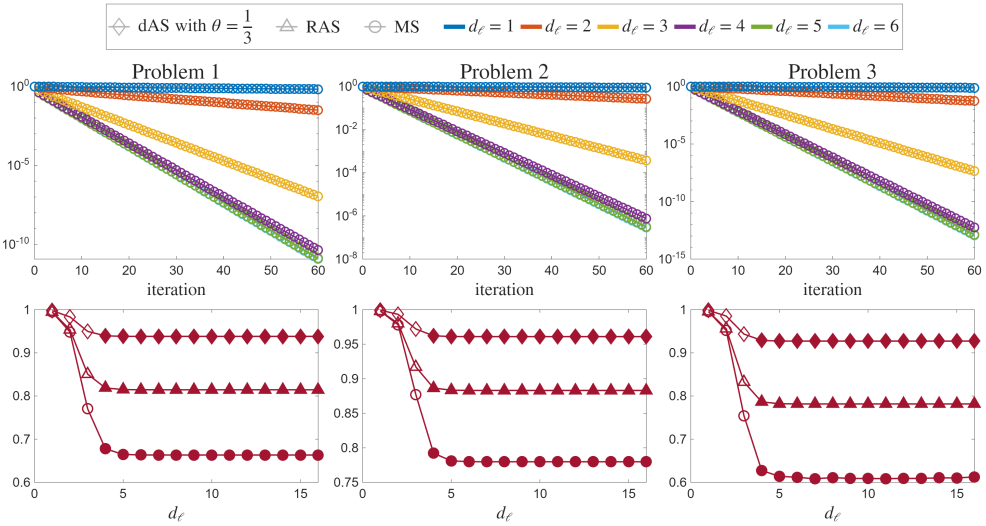


FIG. 4.2. *Top: the 2-norm of the error of the multiplicative Schwarz method for different choices of $d_\ell$, using the* `advanpix` *toolbox. For $d_\ell \geq 6$ the graphs are essentially indiscernible from the green one for $d_\ell = 5$. For Bottom: the observed convergence factor $\rho_{\text{conv}}$ for different methods and choices of $d_\ell$; if the conditions (4.7) and (4.10) are satisfied for both subdomain $i = 1, 2$ for a certain $d_\ell$, then the marker is filled. For example, for Problem 1 the conditions (4.7) and (4.10) are satisfied for both $i = 1, 2$ from $d_\ell = 4$ onward.*

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 (4.7) was - for some problems and mesh-sizes - satisfied for `bfloat16`, while (4.10) wasn't) or `advanpix` (again, (4.7) was rarely satisfied for $d_\ell = 3$, while (4.10) wasn't).

We further illustrate the tipping point of the conditions (4.7) and (4.10) being 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 (but not encountered due to the rescaling). In full precision, we expect the classical two-domain profile of the largest eigenmode of the matrix $T_{\text{MS}}$, smooth on each of the subdomains. Indeed, for `fp16`, `fp32` and `fp64` that 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 establish to the same extend. In other words, for too low precision $u_\ell$ the method effectively looses its continuous level interpretation as a domain decomposition method, although it is still a reasonably effective (even convergent) smoother. Importantly, satisfying the conditions (4.7) and (4.10) 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 kept the problem size relatively small so far, we next experiment also with varying $N$. However, letting $N$ grow, two numerical limitations come forward – (i) the `chop` toolbox becomes too slow and (ii) the verification of the condition (4.10) becomes untenable. Hence,

ETNA
Kent State University and
Johann Radon Institute (RICAM)



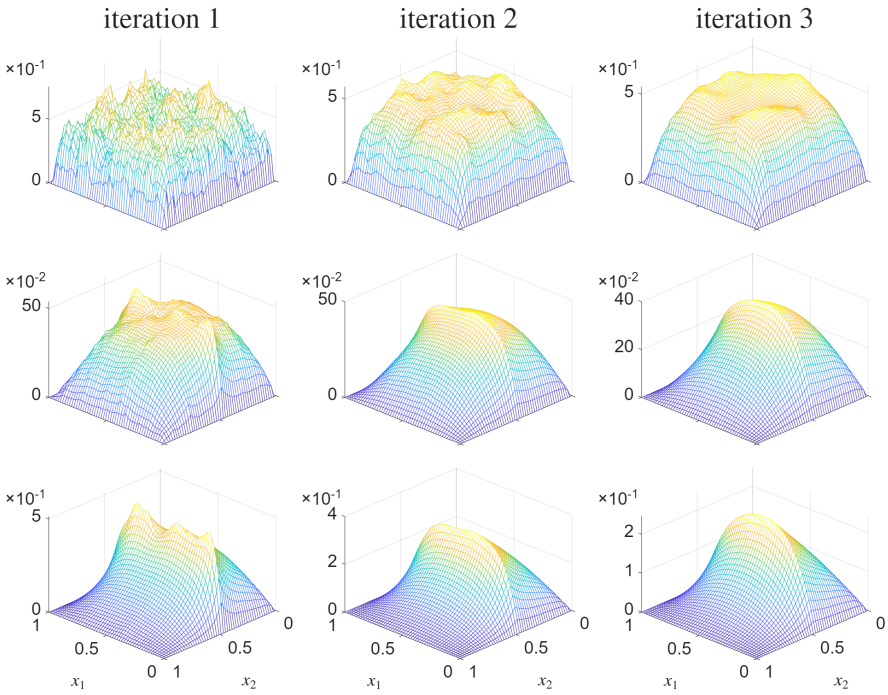


FIG. 4.3. *The errors of the multiplicative Schwarz method used for Problem 1 with* $N = 2500$ *and* `chop` *toolbox after 1, 2 and 3 iterations using* `q43` *(top),* `bfloat16` *and (middle)* `fp16` *(bottom). Up to scaling, the error for* `q52` *is analogous to the top row and the errors for* `fp32` *and* `fp64` *are are analogous to the bottom row.*

for the following experiments we will use only the `advanpix` toolbox and only verify the condition (4.7) (essentially testing whether the observation in Remark 4.2 holds true also for larger $N$).

All of the above characteristics remained true with the only change being the first $d_\ell$ so that the condition (4.7) is satisfied. We show these for $N \in \{2500, \dots, 108900\}$ in Figure 4.4. Notably, we see that after the first $d_\ell$ such that (4.7) is satisfied there is little to no change in using additional precision, precisely as observed above for $N = 2500$. In other words, the dominant eigenmodes of $T_\star$ with $\star \in \{\mathrm{AS}, \mathrm{RAS}, \mathrm{MS}\}$ seem to be well-captured already with limited precision *and* the other eigenmodes are not too sensitive with respect to small perturbations of the subdomain solves and stay "non-dominant". In addition, the same type of behavior as showed in Figure 4.3 is present for larger $N$, i.e., for too low precision the methods lose their two-domain nature, converge extremely slow but remain effective smoothers.

We see that the convergence factors $\rho_{\mathrm{conv}}$ are remarkably uniform for the different problems as well as with respect to changing the solve precision $u_\ell$. The condition (4.7) is satisfied either at $d_\ell = 4$ or $d_\ell = 5$, also depending on the size of the problem and once the condition is satisfied, then $\rho_{\mathrm{conv}}$ stabilizes around this final value. In other words, based on these experiments the condition (4.7) still governs the required precision. We note that this is perhaps not too surprising as the condition (4.10) is clearly *only* a sufficient one – if it does not hold, then the entries of the matrix $\tilde{\mathcal{A}}_i^{-1}$ in (4.9) are given as an oscillating sum (rather than a sum of only non-negative numbers), which still can easily sum-up to a non-negative number. On the other hand, if (4.7) doesn't hold, then there's no easy way around it.

We remark that in both [2, 44] 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

ETNA
Kent State University and
Johann Radon Institute (RICAM)



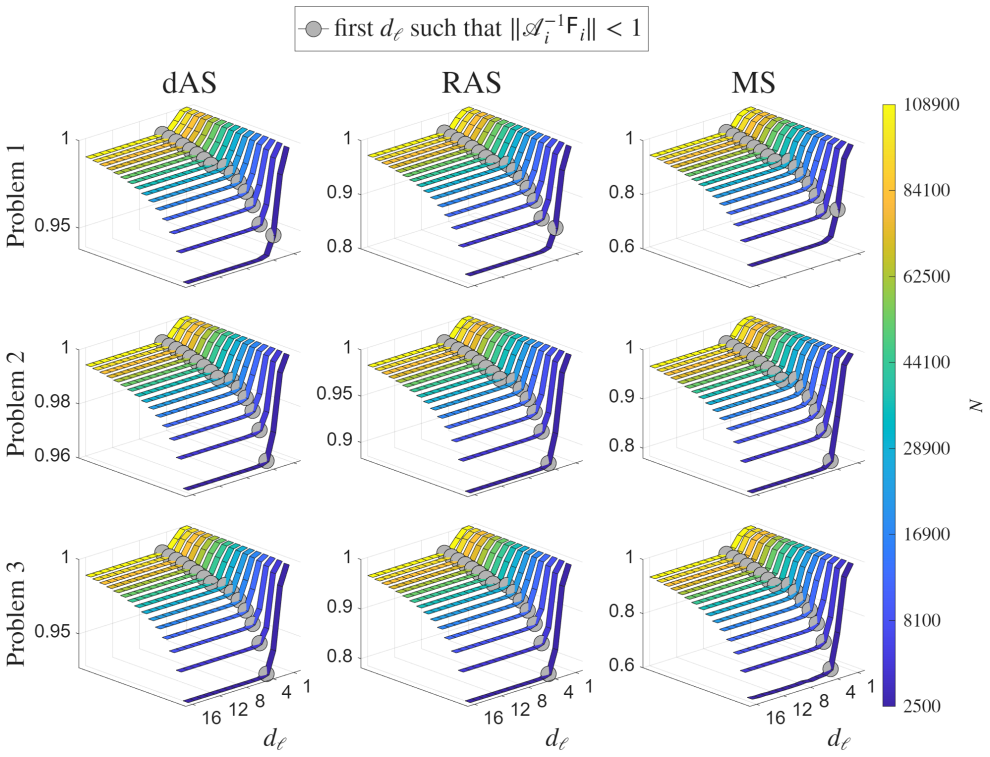


FIG. 4.4. *We show $\rho_{\text{conv}}$ of the Schwarz methods for $d_\ell = 1, 2, \ldots, 16$ and different problem sizes $N$. We highlight the first $d_\ell$ for which the condition (4.7) is satisfied (it is also satisfied for all the following ones).*

Problem 1 and 2 are fairly small within the range $(10^3, 10^5)$, while 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 does explain 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 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 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, which is clearly visible in Figure 4.4.

In addition, note that in condition (4.7) the 2-norm can be replaced by any consistent and equivalent matrix norm and the Neumann series result is still valid; see, e.g., [38, Section 1.3, Lemma 1.3.10]. In other words, the computationally unfeasible condition (4.7) can be replaced by

$$\|\mathcal{A}_i^{-1}\mathsf{F}_i\|_F^2 < 1, \quad \text{or} \quad \|\mathcal{A}_i^{-1}\mathsf{F}_i\|_1^2 < 1.$$

Although these are clearly preferable for the purpose of determining the number of digits $d_\ell$ (or $d_{\ell_i}, i = 1, \ldots, p$), see [2], they might give 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 on the dominant eigenmode and eigenvalue of $T_\star$, as one expects for a fixed-point iteration.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



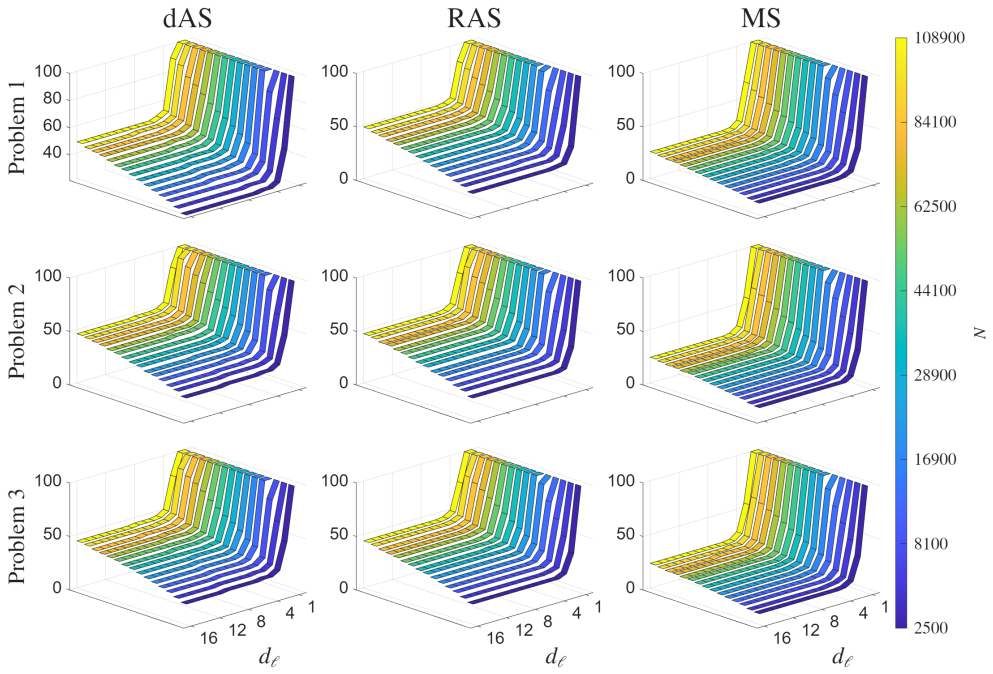


FIG. 4.5. *The number of the preconditioned GMRES iterations to reduce the relative residual below* $10^{-12}$ *capped at 100.*

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 successful preconditioners, calculating in $u_\ell$ instead of $u_w$ on the subdomains *should not* make eigenbasis much more ill-conditioned or the spectrum much more "spread out", otherwise a notable slowdown of GMRES convergence (compared to the appropriate full-precision Schwarz method) can occur. In other words, the above experiments do not necessarily suggest that the multiprecision Schwarz methods will be also efficient when used as precnoditioners. We investigate that next numerically and use the preconditioned GMRES with multiprecision dAS, RAS and MS as the left preconditioners and with preconditioned relative residual tolerance $10^{-12}$, zero initial approximation and maximum number of iterations set to 100. We show the number of GMRES iterations in Figure 4.5.

We observe that the effect of the low-precision does not meaningfully disrupt the number of iterations that the preconditioned GMRES needs to reduce the preconditioned relative residual to the tolerance $10^{-12}$. Moreover, we see the same diminishing returns as we did 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 the increase of the iterations as $N$ increases for similar reasons as in Figure 4.4. While we do not consider the analysis of the multiprecision preconditioned GMRES method, we refer the reader to [6] for analysis and further references.

Summarizing, we can say that we observe that in the model examples *four or five digits suffices* to achieve virtually indistinguishable results to full double precision, i.e., running `fp32` (or even `fp16` for smaller $N$) should be up to twice as fast (four times as fast) to the standard `fp64` Schwarz method without any meaningful drawback. Moreover, the results showcase that running `fp16` or even `bfloat16` should result in a negligible slowdown while offering up to a further two-fold speed-up.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



**4.2. The symmetric case.** We consider now *symmetric $M$-matrices*, sometimes also called Stieltjes matrices (Stieltjes matrices are themselves symmetric positive definite, see [4, 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 is done symmetrically (or if we store and work only on, say, the upper-triangular part of the matrix), then the symmetry is preserved as well. Notably, to achieve symmetrical scaling, the proposed rescaling algorithm needs to be symmetrized, leading to an iterative procedure, see [32, Algorithms 2.5]. Moreover, our rounding routine can be further tailored to preserve other useful properties of the subdomain matrices.

For example, assuming $A$ has dominant entries on the diagonal, in the sense that $a_{ii} \geq |a_{ij}|$ for all $i, j$, [32, Algorithms 2.5] converges in a single step, yielding $D_i^{(r)} = D_i^{(c)} = \operatorname{diag}(a_{11}^{-1/2}, \ldots, a_{N_i N_i}^{-1/2})$ with $D_i^{(r)} A_i D_i^{(c)}$ having all ones on the diagonal and the rest of the entries being bounded in absolute value from above by one. Then, taking $\nu$ as some power of two (or other number we represent exactly in $u_\ell$), we can use the rounding routine

$$(\text{round}_{Diag}(X))_{mn} := \begin{cases} (X)_{mn}, & \text{if } m = n, \\ \mathtt{rd}_{u_\ell}\left((X)_{mn}\right), & \text{if } m \neq n, \end{cases} \tag{4.16}$$

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

Next, we turn our attention to the classical convergence theory for the algebraic Schwarz methods for the symmetric, positive-definite case. As the RAS method is inherently non-symmetric, it is standard to consider the convergence theory only for the (damped) AS and MS methods and as a result we focus only on these two classes*. For these, the driving force behind Theorem 2.1 is the so-called $P$-regular Splitting Theorem, see, e.g., [38, Theorem 7.1.9]. The core assumption there becomes that the splittings in (4.2) are $P$-regular splitting, i.e., that

$$\tilde{A}_i^T + \tilde{A}_i - A_i \succeq 0. \tag{4.17}$$

In order to satify (4.17), the standard assumption in the literature is $\tilde{A}_i \succ A_i$ (see, e.g., [3, 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 of $D_i^{(r)} = D_i^{(c)} =: D_i$, we get

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

and (4.17) becomes equivalent to

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

which is ensured by the two following conditions

$$\tilde{\mathcal{A}}_i \succ 0 \quad \text{and} \quad \lambda_{min}(\tilde{\mathcal{A}}_i) \geq |\lambda_{-\infty}(\mathtt{F}_i)|, \tag{4.18}$$

*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 [43] using the variational methods for a simple model problem. In general, convergence of RAS is usually addressed in combination with the particular problem, see [15], or based on other properties of the system matrix, see [18].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



where $\lambda_{min}(\tilde{\mathcal{A}}_i) \geq 0$ is the smallest eigenvalue of $\tilde{\mathcal{A}}_i$ (as we assume there $\tilde{\mathcal{A}}_i \succ 0$) 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 for the rounding error matrix to be indefinite.

The first condition in (4.18) can be ensured by rounding as in Section 4.1 so that $\tilde{\mathcal{A}}_i$ is still a Stieltjes matrix and hence symmetric, positive-definite. The second condition can be further expanded on, 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) for $\tilde{\mathcal{A}}_i = \mathcal{A}_i + \mathtt{F}_i$, we obtain

$$\lambda_{min}(\tilde{\mathcal{A}}_i) \geq \lambda_{min}(\mathcal{A}_i) + \lambda_{-\infty}(\mathtt{F}_i), \tag{4.19}$$

so that to ensure the second condition in (4.18), it is enough to require

$$\lambda_{min}(\mathcal{A}_i) \geq 2\,|\lambda_{-\infty}(\mathtt{F}_i)|\,. \tag{4.20}$$

We summarize the results in the following theorem.

THEOREM 4.3. *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 $W_i \cap W_j \neq \{0\}$, 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 a precision $u_w$ with the subdomain solver $\tilde{A}_i^{-1}$ with $u_w < u_\ell$, 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 (4.4) with $\tilde{\mathcal{A}}_i$ given as in (4.12) with a symmetric scaling, if (4.7), (4.10) and (4.20) are satisfied, then*

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

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

We note that analysis for SPD matrices and multiprecision additive Schwarz methods has been considered elsewhere; see [47, 2, 44, 27]. However, in all of these papers the authors consider the (non-damped) 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 on the condition number of the preconditioned system. We also note that the assumptions look somewhat similar*. Indeed, to get a coarser version of (4.20) 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 (4.20) we would require

$$\lambda_{min}(\mathcal{A}_i) \geq 2\rho(\mathtt{F}_i). \tag{4.21}$$

Since we still have the entry-wise comparison of $u_\ell|\mathtt{F}_i|$ and $|\mathcal{A}_i|$, see (4.11), and we have knowledge of the sign distribution of the entries of these matrices, we could arrive at some comparison theorem for $\rho(\mathtt{F}_i)$ and $\rho(\mathcal{A}_i)$, so that (4.21) would relate the condition number $\rho(\mathcal{A}_i)/\lambda_{min}(\mathcal{A}_i)$ with the used precision $u_\ell$, obtaining the type of condition we encounter in [44, Section 3.2.2, equation (14)] or [2, Section 5]. The above derivation illustrates that our results are more nuanced compared to the existing ones, also in treating Schwarz methods (and their convergence) as standalone 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 the existing literature and we carefully exploit this extra information by the specialized rounding techniques.

Next, we show results analogous to the experiments considered in Section 4.1. We consider the same problem as in (4.14)–(4.15) but omit the advection terms so that we obtain Stieltjes matrices after discretization, using the following parameters.

*Compare [44, equations (12) and (14)] and [2, Section 5] with (4.7) and (4.20).

ETNA
Kent State University and
Johann Radon Institute (RICAM)



*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 non-constant reaction and diffusion coefficients (otherwise omitting the advection term leads to the standard Poisson problem).

The same questions as before are of interest. We fix $N = 2500$ and show the convergence curves and the observed convergence factor $\rho_{\text{conv}}$ in Figure 4.6 (using the `chop` toolbox) and Figure 4.7 (using `advanpix` toolbox). We draw very similar conclusions to the ones in Section 4.1. We note that the additional condition (4.20) was almost always weaker than (4.10) and comparable to (4.7). However, just as in the non-symmetric case in Section 4.1, the differences were small (e.g., for `fp16` and `bfloat16` for `chop` or for neighboring precisions for `advanpix`).

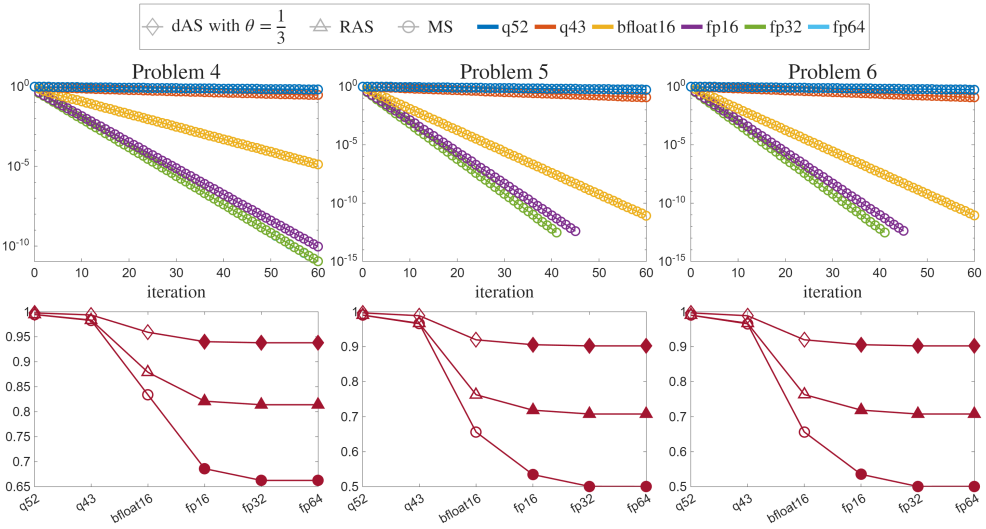


FIG. 4.6. *Top: the 2-norm of the error of the multiplicative Schwarz method for different choices of $u_w$ using the* `chop` *toolbox. For* `fp32` *and* `fp64` *the graphs are indiscernible from each other. Bottom: the observed convergence factor $\rho_{\text{conv}}$ for different methods and choices of $u_\ell$; if the conditions (4.7), (4.10) and (4.20) are satisfied for both subdomain $i = 1, 2$ for a certain $u_\ell$, then the marker is filled. For example, for Problem 4 the conditions (4.7), (4.10) and (4.20) are satisfied for both $i = 1, 2$ starting from* `fp16`.

In Figure 4.8 we plot 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 see, generally speaking, similar results to 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, similarly to the non-symmetric case, the dominant eigenmodes of $T_\star$ with $\star \in \{\text{AS}, \text{RAS}, \text{MS}\}$ appear to be

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



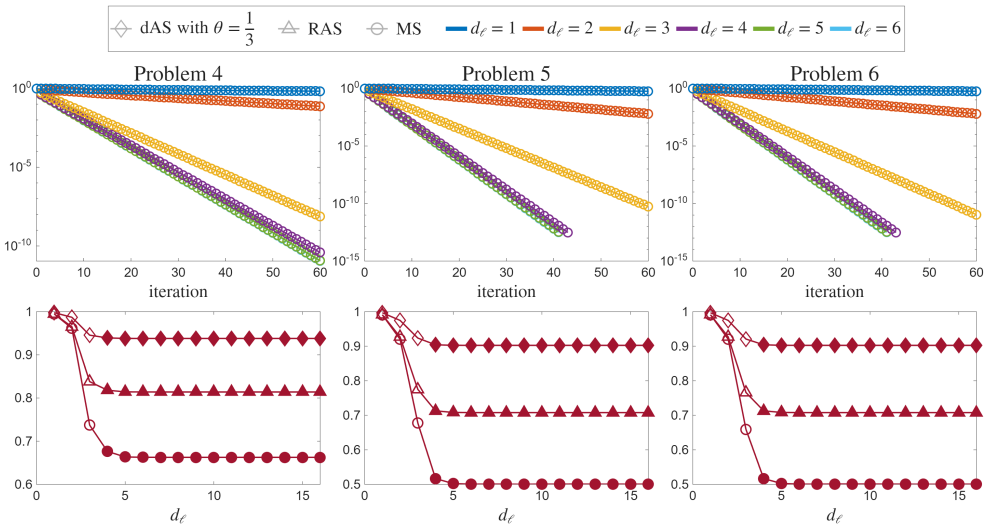


FIG. 4.7. *Left: the 2-norm of the error of the multiplicative Schwarz method for different choices of* $d_\ell$. *Right: the observed convergence factor* $\rho_{\text{conv}}$ *for different methods and choices of* $d_\ell$*; if the conditions* (4.7), (4.10) *and* (4.20) *are satisfied for both subdomains* $i = 1, 2$ *for a certain* $d_\ell$*, then the marker is filled.*

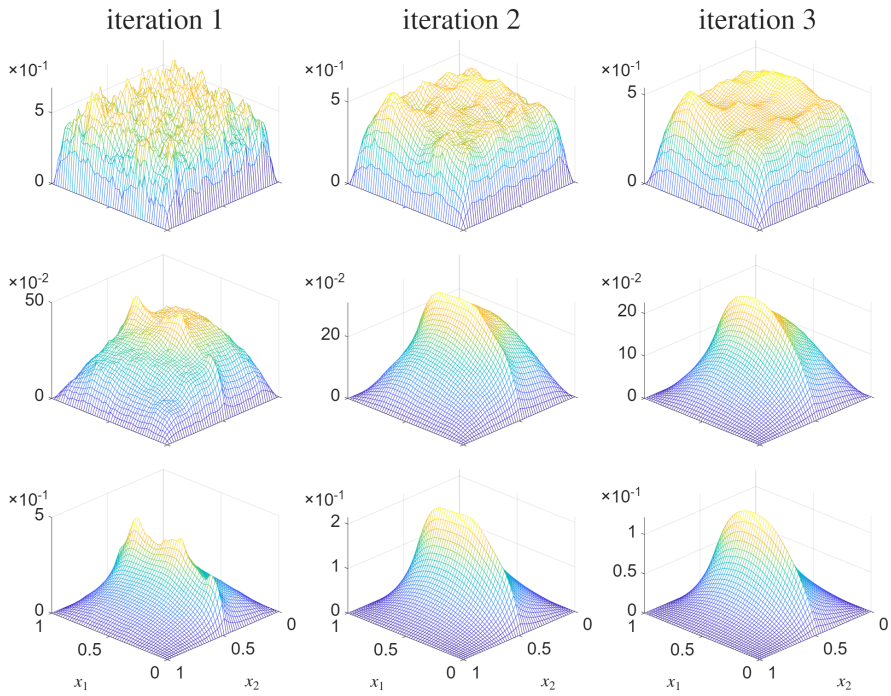


FIG. 4.8. *The errors of the multiplicative Schwarz method used for Problem 5 with* $N = 2500$ *during the initial iterations using* `q43` *(top),* `bfloat16` *and (middle)* `fp16` *(bottom). Up to scaling, the error for* `q52` *is analogous to the top row and the errors for* `fp32` *and* `fp64` *are are analogous to the bottom row.*

well-captured already with limited precision, e.g., $d_\ell = 4 \sim 6$, *and* the other eigenmodes are not too sensitive with respect to small perturbations of the subdomain solves and stay

ETNA
Kent State University and
Johann Radon Institute (RICAM)



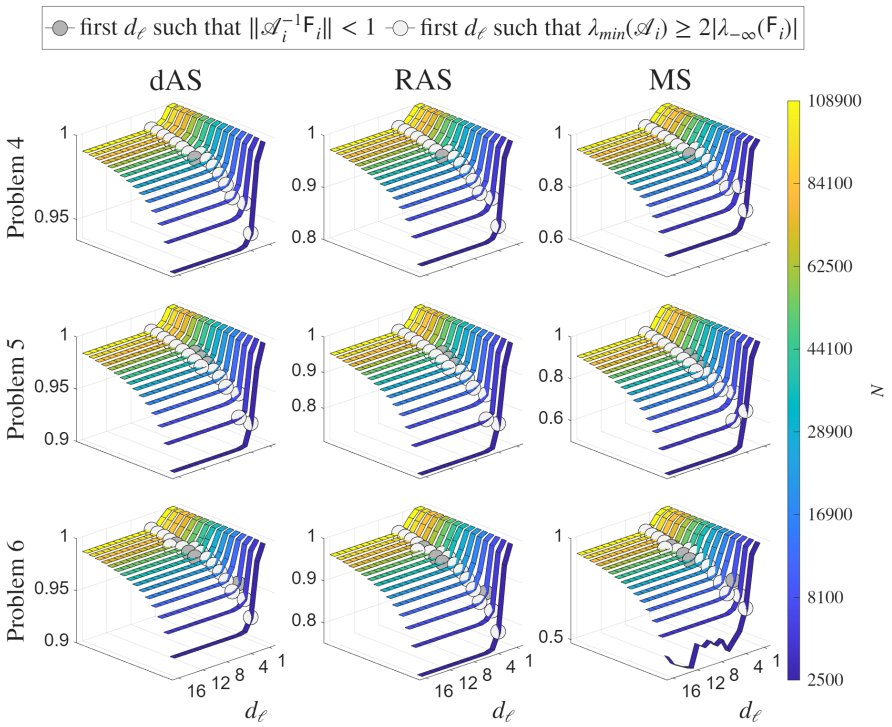


FIG. 4.9. *We show $\rho_{\text{conv}}$ of the Schwarz method for $d_\ell = 1, 2, \ldots, 16$ and different problem sizes $N$. We highlight the first $d_\ell$ for which the conditions (4.7) and (4.20) are satisfied by ● (or by ○).*

"non-dominant".

The theoretical results only hold if the all of the conditions (4.7), (4.10) and (4.20) hold true but the model problems suggest that either of the conditions (4.7) or (4.20) give a good indicator. However, we note that the condition (4.20) becomes *much* more pessimistic, if we omit the re-scaling, e.g., if we take $D_i^{(r,c)} = I_{N_i}$ for Problem 6, then (4.20) is satisfied only for $d_\ell \gtrapprox 10$, i.e., long after the convergence factor has in fact stabilized at the final value. The same is true if we replace the condition (4.20) with a cruder version relating to the condition number of $\mathcal{A}_i$, see (4.21) and below. The condition (4.7), however, has been fairly robust, localizing fairly accurately the optimal $d_\ell$ regardless of the employed scaling. Also, similarly to the non-symmetric case, the convergence factor graph becomes notably "less smooth" but otherwise qualitatively similar. The "non-smoothness" of the convergence factor for Problem 6 and the smallest mesh resolution, i.e., $N = 2500$ also stands out. The reason 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 check by inspecting the error plots directly.

Last, we use the preconditioned GMRES with multiprecision dAS*, RAS and MS as the left preconditioners and with relative residual tolerance $10^{-12}$, zero initial approximation and maximum number of iterations set to 100. We show the GMRES convergence curves and the number of iterations in Figure 4.10. We see that the number of iterations again stays mostly

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

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



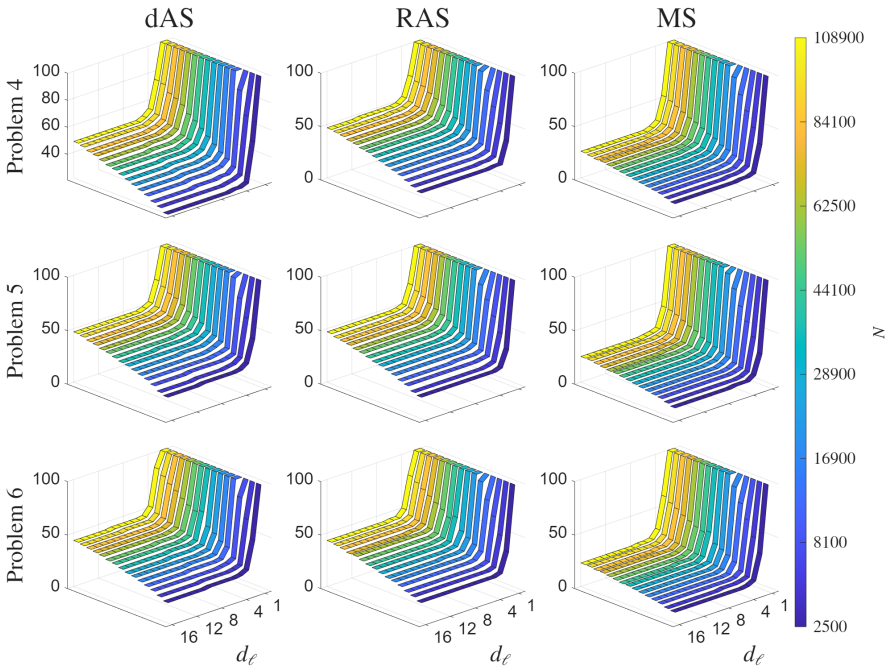


FIG. 4.10. *The number of the preconditioned GMRES iterations to reduce the relative residual below* $10^{-12}$ *capped at 100.*

stable with respect to changing $d_\ell$ for a fixed $N$ and, moreover, the conditions (4.7) or (4.20) still work as a reasonably accurate indicator for the choice of the number of digits $d_\ell$.

**5. Comparison of standard and multiprecision 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. However, another approach would be to use perturbation theory for the analysis, and in fact, for the case of Schwarz methods as a preconditioners this seems to be a more viable path.

In this section we compare the convergence of the "exact" method, i.e., with $u_\ell = u_w$, say, double precision, and that with the multiprecision approach. We consider the general, non-symmetric case and first focus on the easier-to-analyze additive methods, i.e., (damped) AS and RAS, and comment on the extension for the multiplicative case later.

*Additive Schwarz methods.* We start by assuming the set-up of Section 4.1, namely, similarly to (4.6) we can write

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

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

$$\tilde{A}_i^{-1} - A_i^{-1} = \mathcal{E}_i A_i^{-1}. \tag{5.1}$$

As a result, if we assume*

$$\|\mathcal{A}_i^{-1} \mathtt{F}_i\| \leq \epsilon < \frac{1}{2}, \tag{5.2}$$

*This assumption is analogous to (4.7). In fact, the derivations requiring (4.7) can be carried out analogously even if we assume (5.2) instead of (4.7) but the derivation becomes somewhat more lengthy.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



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

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

and so

$$\|\tilde{A}_i^{-1} - A_i^{-1}\| = 2\epsilon\kappa \left( D_i^{(c)} \right) \|A_i^{-1}\|, \tag{5.3}$$

where $\kappa(\cdot)$ denotes the condition number with respect to the norm $\|\cdot\|$. In other words, the (small) perturbation of the scaled subdomain matrices can perturb the subdomain solves proportionally to the given subdomain solve norm, i.e., to $\|A_i^{-1}\|$, and to the condition number $\kappa \left( D_i^{(c)} \right)$ of the column-scaling matrix. This shows that in the ideal scenario, we either get a well-scaled subdomain matrices *or* we can mostly fix the scaling by row-scaling (recall that, conveniently, the row-scaling takes precedence in [32, Algorithms 2.3 and 2.4]). Also, notice that (5.2) is similar to the assumption [44, equation (12)], i.e., to the norm-wise equivalent of (4.11) but for the inverses and after the scaling.

Next, we insert (5.2) into the definition of the additive Schwarz methods in (2.3) and get

$$\tilde{M}_\star^{-1} A = M_\star^{-1} A + E_\star \quad \text{where} \quad E_\star := \begin{cases} \theta \sum_{i=1}^{p} R_i^T \mathcal{E}_i A_i^{-1} R_i A, & \text{for (damped) AS,} \\ \sum_{i=1}^{p} \bar{R}_i^T \mathcal{E}_i A_i^{-1} R_i A, & \text{for RAS.} \end{cases} \tag{5.4}$$

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

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

In fact, we can further rewrite this as

$$E_\star = \begin{cases} AS, \theta: & \theta \sum_{i=1}^{p} \Pi_i^T \begin{bmatrix} \mathcal{E}_i & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_{N_i} & A_i^{-1} K_i \\ 0 & 0 \end{bmatrix} \Pi_i, \\ RAS: & \sum_{i=1}^{p} \overline{\Pi}_i^T \begin{bmatrix} (\mathcal{E}_i)_{1:\bar{N}_i,:} & 0 \\ 0_{\bar{N}_i+1:N_i,:} & \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_{N_i} & A_i^{-1} K_i \\ 0 & 0 \end{bmatrix} \Pi_i, \end{cases} \tag{5.5}$$

where we clearly see the main difference between the two methods in the "double-counting of the overlap", see [15, 18]. The structure also highlights that the (damped) additive Schwarz can be written in a purely block diagonal preconditioner while the restricted version cannot.

Next, we proceed with the analysis for the simpler case of only two subdomains, i.e., $p = 2$, which directly generalizes to the case of $p$ subdomains with $q = 1$, i.e., to the "no cross-points" case. In such case we can choose a *global* permutation matrix $\Pi$ so that

$$E_{AS,\theta} = \theta \Pi^T \begin{bmatrix} \mathcal{E}_1 & 0 \\ 0 & \mathcal{E}_2 \end{bmatrix} \begin{bmatrix} I_{N_1} & A_1^{-1} K_1 \\ A_2^{-1} K_2 & I_{N_2} \end{bmatrix} Z \Pi, \tag{5.6}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



where $Z$ acts as the "zipper" for the overlap, i.e.,

$$Z = \begin{bmatrix} I_{\bar{N}_1} & & \\ & I_{N_1-\bar{N}_1} & \\ & I_{N_2-\bar{N}_2} & \\ & & I_{\bar{N}_2} \end{bmatrix}.$$

We notice that this way we managed to factor the error matrix $E^{(k)}_{AS,\theta}$ so that the first term carries the multiprecision error while the second term carries the Schwarz-method structure. As a result, bounding the norm of $E^{(k)}_{AS,\theta}$ and $E_{AS,\theta}$ becomes easier*. Assuming (5.2) for both $i = 1, 2$ and noticing that $\|Z\| \leq 2$ we observe that

$$\|E_{AS,\theta}\| \leq 2\epsilon\kappa\left(D_i^{(c)}\right)\|M^{-1}_{AS,\theta}A\|. \tag{5.7}$$

Unfortunately, similar approach does not work for the restricted additive Schwarz method as the "structure matrix" in (5.5) is that of the additive Schwarz method, rather than of the restricted version.

A natural next step would be to carry out this reformulation also for the iteration matrix $\tilde{T}_{AS,\theta}$ as it is its spectral radius that asymptotically governs the convergence. However, since the spectral radius is not sub-additive or even stable with respect to perturbations, this wouldn't be directly useful for quantifying the slow-down of the Schwarz method convergence, unless we consider a more specific situation. An extreme example is the case when the rounding is in nature only *scalar* (highlighting the "nicest" case included in the above setting). That is, there exist a scalar $\alpha$ such that $\mathcal{A}_i = \alpha\mathtt{A}_i$ where $\tilde{\mathtt{A}}_i = \mathtt{A}_i$, i.e., the subdomain matrices are scalar multiples of a matrix that can be stored "exactly"† in the considered precision $u_\ell$. Then

$$\mathtt{F}_i = (\tilde{\alpha} - \alpha)\,\mathcal{A}_i, \tag{5.8}$$

and denoting $\tau := |(\tilde{\alpha} - \alpha)/\alpha| \in (0, 1)$ we get

$$\mathcal{E}_i = -\frac{\tau}{1+\tau}I, \tag{5.9}$$

and hence

$$\tilde{T}_{AS,\theta} = T_{AS,\theta} + \frac{\tau}{1+\tau}M^{-1}_{AS,\theta}A = I - \frac{1}{1+\tau}M^{-1}_{AS,\theta}A.$$

In words, the inexactness of the local solves can be interpreted as an (additional) damping for the additive Schwarz method. This factor can be conveniently included into the damping factor $\theta$, i.e., the damping factor $\theta$ can be chosen with this in mind, e.g., in our case taking $\theta = 1/2$ still guarantees convergence convergence. Notice though that $\tau$ corresponds to the relative rounding error for $\alpha$ in the precision $u_\ell$ and hence $1/(1+\tau)$ tends to 1 as $u_\ell$ decreases.

*In many areas of interest it is often more suitable to bound the norm of the iteration matrix (and hence of the error) over two or more iterations due to the nature of the underlying PDE analysis, see, e.g., [23, 26]. Further research in this direction might be useful here as well.

†Here "exactly" means to the same precision we store the solution. Also, notice that verifying (4.3) becomes trivial. Notice that such problems arise, e.g., when discretizing "nice" Poisson-like problems with finite differences so that $\alpha = 1/h^2$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



*Multiplicative Schwarz methods.* For the multiplicative Schwarz method we start by writing the error matrix for the iteration matrix $\tilde{T}_{\mathrm{MS}}$, i.e., we write

$$\tilde{T}_{\mathrm{MS}} = T_{\mathrm{MS}} - E_{\mathrm{MS}}.$$

We choose to introduce the sign in this way because then we have

$$T_{\mathrm{MS}} = I - M_{\mathrm{MS}}^{-1}A \quad \text{and} \quad \tilde{T}_{\mathrm{MS}} = I - \tilde{M}_{\mathrm{MS}}^{-1}A,$$

and hence we get a consistent notation with (5.5), i.e.,

$$\tilde{M}_{\mathrm{MS}}^{-1}A = M_{\mathrm{MS}}^{-1}A + E_{\mathrm{MS}}.$$

Considering the two-subdomains setting, i.e., $p = 2$, we get

$$E_{\mathrm{MS}} = \Bigg\{ \overbrace{-\left(I - R_2^T A_2^{-1} R_2 A\right) R_1^T \mathcal{E}_1 A_1^{-1} R_1 A}^{=:G_1} \overbrace{- R_2^T \mathcal{E}_2 A_2^{-1} R_2 A \left(I - R_1^T A_1^{-1} R_1 A\right)}^{=:G_2} \underbrace{+ R_2^T \mathcal{E}_2 A_2^{-1} R_2 A R_1^T \mathcal{E}_1 A_1^{-1} R_1 A}_{=:G_3},$$

and notice that the situation becomes more complicated than for the additive methods as the matrices $\mathcal{E}_i$ now interact with the other subdomain solves. This is a consequence of the sequential nature of MS as opposed to (damped) AS and RAS and makes *fully general* and yet *insightful* analysis not possible, precisely because of the unknown interaction. To visualize this, let us assume that $A$ has been (symmetrically) permuted so that

$$A = \begin{bmatrix} A_I & A_{I,o} & A_{I,II} \\ A_{o,I} & A_o & A_{o,II} \\ A_{II,I} & A_{II,o} & A_{II} \end{bmatrix} \quad \text{with} \quad \begin{aligned} A_1 &= \begin{bmatrix} A_I & A_{I,o} \\ A_{o,I} & A_o \end{bmatrix}, K_1 = \begin{bmatrix} A_{I,II} \\ A_{o,II} \end{bmatrix}, \\ K_2 &= \begin{bmatrix} A_{o,I} \\ A_{II,I} \end{bmatrix}, A_2 = \begin{bmatrix} A_o & A_{o,II} \\ A_{II,o} & A_{II} \end{bmatrix}. \end{aligned}$$

A direct calculation then gives the formulas*

$$T_{\mathrm{MS}} = \begin{bmatrix} 0 & 0 & -\left(A_1^{-1}K_1\right)_{I,:} \\ 0 & 0 & \\ 0 & 0 & A_2^{-1}K_2\left(A_1^{-1}K_1\right)_{I,:} \end{bmatrix}, \quad G_1 = \begin{bmatrix} I & 0 & 0 \\ A_2^{-1}K_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{E}_1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I & 0 & \\ 0 & I & A_1^{-1}K_1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$G_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & & \mathcal{E}_2 \\ 0 & & \end{bmatrix} \left( \begin{bmatrix} 0 & 0 & \left(A_1^{-1}K_1\right)_{I,:} \\ 0 & 0 & \\ 0 & 0 & -A_2^{-1}K_2\left(A_1^{-1}K_1\right)_{I,:} \end{bmatrix} + \begin{bmatrix} 0 & 0 & -A_1^{-1}K_1 \\ 0 & 0 & \\ 0 & 0 & I \end{bmatrix} \right),$$

$$G_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & & \mathcal{E}_2 \\ 0 & & \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ A_2^{-1}K_2 & I & 0 \\ & 0 & I \end{bmatrix} \begin{bmatrix} \mathcal{E}_1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I & 0 & \\ 0 & I & A_1^{-1}K_1 \\ 0 & 0 & 0 \end{bmatrix}.$$

All hope of obtaining an analogous bound to (5.7) is lost here as any attempt to get a common factor of $\epsilon \geq \|\mathcal{E}_i\|$ necessarily *separates* the subdomain matrices $A_1^{-1}K_1$ and $A_2^{-1}K_2$ in both $G_1$ and $G_3$. Hence, any such general bound *breaks* the "continuity" of one MS iteration. However, looking at $G_2$ we see the structure of $T_{\mathrm{MS}}$ appearing, with an additional term. It is useful to notice that the same structure can be retained also for $G_1$ and $G_3$, *provided we have additional knowledge* about the interaction of the the subdomain matrices $A_i^{-1}K_i$ and

*Many of the following calculations are similar to the ones presented in [26, Section 3.2] for the *modified restricted* Schwarz method.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



the matrices $\mathcal{E}_i$. For example, considering the simplest rounding setting as in (5.8)–(5.9) a straight-forward calculation gives us

$$G_1 = \frac{\tau}{1+\tau}\left(T_{\mathrm{MS}} + \begin{bmatrix} I & 0 & 0 \\ -A_2^{-1}K_2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right), \quad G_2 = \frac{\tau}{1+\tau}\left(T_{\mathrm{MS}} + \begin{bmatrix} 0 & 0 & -A_1^{-1}K_1 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}\right),$$
$$G_3 = \left(\frac{\tau}{1+\tau}\right)^2\left(T_{\mathrm{MS}} + \begin{bmatrix} 0 & 0 & A_1^{-1}K_1 \\ A_2^{-1}K_2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right),$$

and we see that an adapted version of (5.7) can be established in this particular case, i.e.,

$$\begin{aligned}\|E_{\mathrm{MS}}\| \leq \quad & \frac{\tau}{1+\tau}\left(2(\|M_{\mathrm{MS}}^{-1}A\|+1) + \|A_1^{-1}K_1\| + \|A_2^{-1}K_2\|\right) \\ & + \left(\frac{\tau}{1+\tau}\right)^2\left(\|M_{\mathrm{MS}}^{-1}A\| + \|A_1^{-1}K_1\| + \|A_2^{-1}K_2\| + 2\right).\end{aligned}$$

*Accelerated Schwarz methods.* Next, we analyze the methods as *preconditioners*, using the Schwarz methods as preconditioners for GMRES. First, we note that the GMRES convergence in the above examples was *almost linear* (as opposed to *(strongly) superlinear*) and this was true also when we used dAS or RAS instead of MS and in all of our experiments. A recent result in [5, Theorem 1.1 and Corollary 1.2] shows that for a linear GMRES convergence a perturbation of the system matrix that is *sufficiently* small in norm slows the linear convergence only negligibly. Adapted to our case, let us assume that the GMRES preconditioned with a standard Schwarz method converged linearly with the convergence factor $\rho_\star^{GMRES}$. Then, running the GMRES preconditioned with a multiprecision Schwarz method and obtaining the residual vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots$, we obtain the following bound

$$\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \leq \left(\rho_\star^{GMRES} + \frac{1}{\sqrt{k}}(1+\rho_\star^{GMRES})\|A^{-1}M_\star\|\|E_\star\|_F\right)^k.$$

Observe that for any of the Schwarz methods this bound uses the inverse of the norm of the preconditioned system, which is of course one of the factors of the *condition number of the preconditioned system*. For the symmetric case, the condition number is a historically classical quantity used to bound the convergence behavior of Krylov subspace methods. For the nonsymmetric case, the connection between the condition number and the convergence is in general not present. For MS the bound also includes additional terms.

Alternatively, we can use the pseudospectra-based bound. As the pseudospectrum of a matrix is stable with respect to perturbations (as oppose to the the spectrum; see [49, The second definition of pseudospectra, p. 14]), these are often useful in situation like ours, i.e., when trying to analyze the effect of a (small) perturbation to the system matrix on the convergence behavior of GMRES, see [45]. First, we recall that the $\delta$-pseudospectrum of a matrix $X$, denoted by $\sigma_\delta(\mathrm{X})$, is defined as

$$\sigma_\delta(\mathrm{X}) = \left\{z \in \mathbb{C} \mid \|(z\mathrm{I}-\mathrm{X})^{-1}\| > \frac{1}{\delta}\right\} = \{z \in \sigma(\mathrm{X}+\mathrm{E}) \text{ for some } \mathrm{E} \text{ with } \|\mathrm{E}\| < \delta\},$$

for any $\delta > 0$ and, clearly, for $\delta = 0$ we recover the spectrum, i.e., $\sigma_0(\mathrm{X}) = \sigma(\mathrm{X})$. Moreover, for any $\delta > 0$, $\sigma_\delta(\mathrm{X})$ forms a union of Jordan curves enclosing $\sigma(\mathrm{X})$. Assuming we are solving a problem $X\mathbf{v} = \mathbf{b}$, using the $\delta$-pseudospectrum, we get the standard ideal GMRES bound

$$\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \leq \frac{L_\delta}{2\pi\delta} \min_{\substack{\deg(\varphi)\leq k \\ \varphi(0)=1}} \max_{z\in\sigma_\delta(\mathrm{X})} |\varphi(z)|, \tag{5.10}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



where $L_\delta$ denotes the arc length of the boundary of $\sigma_\delta(\mathrm{X})$ and $\varphi(z)$ is a polynomial of the degree up to $k$ and normalized at the origin; for more details on pseudospectra see [49] and references therein and for their use in the context of Krylov subspace methods (and GMRES in particular) see [34, Sections 4.9 and 5.7.3] but also [16, Section 2.3] and the work cited there. We also note that (5.10) is in fact a *family* of bounds based on $\delta$, rather than a single bound. The common wisdom is that larger values of $\delta$ tend to be more descriptive at the initial convergence phase (up to a certain $\delta_0$ for which the bound stops being useful at all) while smaller values of $\delta$ give a more accurate prediction for later stages of the GMRES convergence, see [16, Section 2.3].

Importantly, in [45, Section 2.2] the authors give two results relevant to our situation, which we summarize below.

PROPOSITION 5.1 ([45, Theorems 2.1 and 2.3]). *Adopting the above notation, let $\mathbf{r}_k$ ($\boldsymbol{\rho}_k$) be the preconditioned GMRES residual with the full-precision (multiprecision, with the solve precision $u_\ell$) Schwarz method preconditioner. Assuming that $\epsilon := \|E_*\| < 1$, then for any $\delta \in (0, \epsilon)$ we have*

$$\frac{\|\boldsymbol{\rho}_k\|}{\|\mathbf{b}\|} \leq \left(1 + \frac{\epsilon}{\delta - \epsilon}\right) \frac{L_\delta}{2\pi\delta} \min_{\substack{\deg(\varphi) \leq k \\ \varphi(0)=1}} \max_{z \in \sigma_\delta(\mathrm{M}_\star^{-1}\mathrm{A})} |\varphi(z)|. \tag{5.11}$$

In words, the pseudospectral bound is stable with respect to small perturbations and so an accurate pseudospectral bound on the full-precision system leads to only a slightly more pessimistic bound – a delayed version of the full-precision one – for the multiprecision preconditioner.

We also note that in [45, Corollary 2.2], the authors show that for a *fixed* perturbation matrix multiplied by a magnitude factor, i.e., for the case $E_\star(d_\ell) = \epsilon_{d_\ell} Z_\star$, we can expect that the (precondtioned) residual norms will level-off from a certain precision onward. Moreover, this specific threshold can be estimated using the pseudospectra of the original (full-precision preconditioner) system. However, the question of calculating the pseudospectra of the original (full-precision) preconditioned system as well as a reasonable choice of (several) value(s) of $\delta$ remains highly problem dependent and will be a key factor in determining the accuracy of these bounds.

**6. Conclusion and future work.** We have proposed and analyzed multiprecision Schwarz methods that are specifically tailored for problems where we can guarantee the methods convergence – problems where the system matrix is a so-called $M$-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 on PDE model problems that support our theoretical results and further illustrate aptness of our proposed conditions. As future work we intend to consider generalizations for multiple subdomains and/or "interface conditions" in the sense of [26].

An understanding of the interaction of 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 be certainly interesting and we leave it open 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 of the error) over two or more iterations due to the nature of the underlying PDE analysis, see, e.g., [23, 26]. 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 be also of clear interest. This would be likely overlapping with the so-called double-sweeping preconditioners and their analysis.

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



**References.**

[1] IEEE Standard for floating-point arithmetic. *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pages 1–84, 2019. doi: 10.1109/IEEESTD.2019.8766229.
[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 and Computation: Practice and Experience*, 31:e4460, 2019.
[3] M. Benzi, A. Frommer, R. Nabben, and D. B. Szyld. Algebraic theory of multiplicative Schwarz methods. *Numerische Mathematik*, 89:605–639, 2001.
[4] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979. Reprinted by SIAM, Philadelphia, 1994.
[5] J. Blechta. Stability of linear GMRES convergence with respect to compact perturbations. *SIAM Journal on Matrix Analysis and Applications*, 42:436–447, 2021.
[6] A. Buttari, N. J. Higham, T. Mary, and B. Vieublé. A modular framework for the backward error analysis of GMRES. *IMA Journal of Numerical Analysis; to appear*, 2025.
[7] X.-C. Cai and M. Sarkis. A restricted additive Schwarz preconditioner for general sparse linear systems. *SIAM Journal on Scientific Computing*, 21:792–797, 1999.
[8] X.-C. Cai, M. Dryja, and M. Sarkis. Restricted additive Schwarz preconditioners with harmonic overlap for symmetric positive definite linear systems. *SIAM Journal on Numerical Analysis*, 41:1209–1231, 2003.
[9] E. Carson and X. Chen. Pychop: Emulating low-precision arithmetic in numerical methods and neural networks, 2025. URL https://arxiv.org/abs/2504.07835.
[10] E. Carson and N. J. Higham. Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM Journal on Scientific Computing*, 40:A817–A847, 2018.
[11] F. Cuvelier, M. J. Gander, and L. Halpern. Fundamental coarse space components for Schwarz methods with crosspoints. In S. C. Brenner, E. T. S. Chung, A. Klawonn, F. Kwok, J. Xu, and J. Zou, editors, *Domain Decomposition Methods in Science and Engineering XXVI*, volume 145 of *Lecture notes in Computer Science and Engineering*, Cham, Switzerland, 2023. Springer.
[12] V. Dolean, M. J. Gander, and L. Gerardo-Giorda. Optimized Schwarz methods for Maxwell's equations. *SIAM Journal on Scientific Computing*, 31:2193–2213, 2009.
[13] 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. *Journal of Ccomputational Physics*, 280:232–247, 2015.
[14] Victorita Dolean, Pierre Jolivet, and Frédéric Nataf. *An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation*. SIAM, Philadelphia, 2015.
[15] E. Efstathiou and M. J. Gander. Why restricted additive Schwarz converges faster than additive Schwarz. *BIT Numerical Mathematics*, 43:945–959, 2003.
[16] M. Embree. How descriptive are GMRES convergence bounds?, 2023. arXiv preprint: 2209.01231.
[17] A. Frommer and D. B. Szyld. Weighted max norms, splittings, and overlapping additive Schwarz iterations. *Numerische Mathematik*, 83:259–278, 1999.
[18] A. Frommer and D. B. Szyld. An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms. *SIAM Journal on Numerical Analysis*, 39: 463–479, 2001.
[19] A. Frommer and D. B. Szyld. On the convergence of randomized and greedy relaxation schemes for solving nonsingular linear systems of equations. *Numerical Algorithms*, 92:

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



639–664, 2023.

[20] M. J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44: 699–731, 2006.

[21] M. J. Gander. Schwarz methods over the course of time. *Electronical Transactions on Numerical Analysis*, 31:228–255, 2008.

[22] 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 Z. Dostál, T. Kozubek, A. Klawonn, L. Ulrich, L. F. Pavarino, J. Šístek, and O. B. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXVII*, volume 149 of *Lecture notes in Computer Science and Engineering*, Cham, Switzerland, 2024. Springer.

[23] M. J. Gander and M. Outrata. On algebraic bounds for POSM and MRAS. In Z. Dostál, T. Kozubek, A. Klawonn, L. Ulrich, L. F. Pavarino, J. Šístek, and O. B. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXVII*, volume 149 of *Lecture notes in Computer Science and Engineering*, Cham, Switzerland, 2024. Springer.

[24] 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 Review*, 61:3–76, 2019.

[25] M. J. Gander and H. Zhang. Schwarz methods by domain truncation. *Acta Numerica*, 31:1–134, 2022.

[26] 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 Journal on Matrix Analysis and Applications*, 33:653–680, 2012.

[27] L. Giraud, A. Haidar, and L. T. Watson. Mixed-precision preconditioners in parallel domain decomposition solvers. In U. Langer, M. Discacciati, D. Keyes, O. Widlund, and W. Zulehner, editors, *Domain Decomposition Methods in Science and Engineering XVII*, volume 60 of *Lecture notes in Computer Science and Engineering*, pages 357–364, Berlin, Heidelberg, 2008. Springer.

[28] C. Glusa, E. G. Boman, E. Chow, S. Rajamanickam, and D. B. Szyld. Scalable asynchronous domain decomposition solvers. *SIAM Journal on Scientific Computing*, pages C384–C409, 2020.

[29] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, USA, 2002.

[30] N. J. Higham and T. Mary. Mixed precision algorithms in numerical linear algebra. *Acta Numerica*, 31:347–414, 2022.

[31] N. J. Higham and S. Pranesh. Simulating low precision floating-point arithmetic. *SIAM Journal on Scientific Computing*, 41:C585–C602, 2019.

[32] N. J. Higham, S. Pranesh, and M. Zounon. Squeezing a matrix into half precision, with an application to solving linear systems. *SIAM Journal on Scientific Computing*, 41: A2536–A2551, 2019.

[33] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, Cambridge, 1994.

[34] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis*. Oxford University Press, Oxford, 2013.

[35] P. L. Lions. On the Schwarz alternating method. In R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, editors, *Domain Decomposition Methods for Partial Differential Equations*, pages 1–42. SIAM, 1988.

[36] F. Magoulès, D. B. Szyld, and C.Venet. Asynchronous optimized Schwarz methods with and without overlap. *Numerische Mathematik*, 137:199–227, 2017.

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



[37] Multiprecision Computing Toolbox for MATLAB 5.2.5.15470. Advanpix LLC., Yokohama, Japan.

[38] J. M. Ortega. *Numerical Analysis: A Second Course*. Classics in Applied Mathematics. SIAM, Philadelpia, 1990.

[39] K. Ozaki, T. Ogita, S. Oishi, and S. Rump. Error-free transformations of matrix multiplication by using fast routines of matrix multiplication and its applications. *Numerical Algorithms*, 59:95–118, 2012.

[40] K. Ozaki, T. Ogita, S. Oishi, and S. M. Rump. Generalization of error-free transformation for matrix multiplication and its application. *Nonlinear Theory and Its Applications*, 4: 2–11, 2013.

[41] K. Ozaki, Y. Uchino, and T. Imamura. Ozaki scheme II: A GEMM-oriented emulation of floating-point matrix multiplication using an integer modular technique. arXiv preprint arXiv:2504.08009, 2025.

[42] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Other Titles in Applied Mathematics. SIAM, Philadelphia, Second edition, 2003. ISBN 978-0-89871-534-7.

[43] M. Sarkis and M. Dryja. Convergence bounds for one-dimensional ASH and RAS. In Z. Dostál, T. Kozubek, A. Klawonn, L. Ulrich, L. F. Pavarino, J. Šístek, and O. B. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXVII*, Lecture notes in Computer Science and Engineering. Springer Berlin, Heidelberg, 2023.

[44] J. Schneck, M. Weiser, and F. Wende. Impact of mixed precision and storage layout on additive Schwarz smoothers. *Numerical Linear Algebra with Applications*, 28:e2366, 2021.

[45] J. A. Sifuentes, M. Embree, and R. B. Morgan. GMRES convergence for perturbed coefficient matrices, with application to approximate deflation preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 34:1066–1088, 2013.

[46] Barry F. Smith, Petter E. Bjørstad, and William D. Gropp. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, Cambridge, New York, Melbourne, 1996.

[47] N. Tian, S. Huang, and X. Xu. Mixed precision block-Jacobi preconditioner: algorithms, performance evaluation and feature analysis. *CCF Transactions on High Performance Computing*, 7:114–128, 2025.

[48] Andrea Toselli and Olof Widlund. *Domain Decomposition Methods - Algorithms and Theory*, volume 34 of *Series in Computational Mathematics*. Springer, Berlin, Heidelberg, New York, 2005.

[49] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behaviour of Non-Normal Matrices and Operators*. Princeton University Press, Princeton, New Jersey, 2005.

[50] O. Widlund and M. Dryja. An additive variant of the Schwarz alternating method for the case of many subregions. Technical Report 339, Ultracomputer Note 131, Department of Computer Science, Courant Institute, New York University, 1987.