MULTIGRID SMOOTHERS FOR ULTRA-PARALLEL COMPUTING: ADDITIONAL THEORY AND DISCUSSION

ALLISON H. BAKER[*], ROBERT D. FALGOUT[*], TZANIO V. KOLEV[*], AND ULRIKE MEIER YANG[*]

Abstract. This paper investigates the properties of smoothers in the context of algebraic multigrid (AMG) running on parallel computers with potentially millions of processors. The development of multigrid smoothers in this case is challenging, because some of the best relaxation schemes, such as the Gauss-Seidel (GS) algorithm, are inherently sequential. Based on the sharp two-grid multigrid theory from [17, 18] we characterize the smoothing properties of a number of practical candidates for parallel smoothers, including several C-F, polynomial, and hybrid schemes. We show, in particular, that the popular hybrid GS algorithm has multigrid smoothing properties which are independent of the number of processors in many practical applications, provided that the problem size per processor is large enough. This is encouraging news for the scalability of AMG on ultra-parallel computers. We also introduce the more robust ℓ_1 smoothers, which are always convergent and have already proven essential for the parallel solution of some electromagnetic problems [23].

1. Introduction.

Multigrid (MG) linear solvers are optimal methods because they require $O(N)$ operations to solve a sparse system with N unknowns. Consequently, multigrid methods have good scaling potential on parallel computers, since we can bound the work per processor as the problem size and number of processors are proportionally increased (weak scaling). Near ideal weak scaling performance has been demonstrated in practice. For example, the algebraic multigrid (AMG) solver BoomerAMG [20] in the *hypre* software library [21] has been shown to run effectively on more than 125 thousand processors [16, 5].

One critical component of MG is the smoother, a simple iterative method such as Gauss-Seidel (GS). In the classical setting, the job of the smoother is to make the underlying error smooth so that it can be approximated accurately and efficiently on a coarser grid. More generally, the smoother must eliminate error associated with large eigenvalues of the system, while the coarse-grid correction eliminates the remaining error associated with small eigenvalues.

Some of the best smoothers do not parallelize well, e.g., lexicographical GS. Others used today, while effective on hundreds of thousands of processors, still show some dependence on parallelism and may break down on the millions of processors expected in the next generation machines (we use the term processor here in a generic sense, and distinguish it from cores only when necessary). One such smoother is the *hybrid GS* smoother used in BoomerAMG, which uses GS independently on each processor and updates in a Jacobi-like manner on processor boundaries. In practice hybrid GS is effective on many problems. However, because of its similarity to a block Jacobi method, there is no assurance of obtaining the good convergence of lexicographical GS. In fact, Hybrid GS may perform poorly or even diverge on certain problems, and its scalability has often been cited as a concern as the number of blocks increase with increasing numbers of processors or as block sizes decrease (see, e.g. [1, 15, 31]). For these reasons, previous papers have studied alternatives such as using polynomial smoothers [1] or calculating weighting parameters for hybrid GS [32]. Yet despite its shortcomings, hybrid GS remains the default option in *hypre* because of its overall

[*] Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, P.O. Box 808, L-561, Livermore, CA 94551. This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-TR-489114).

efficiency and robustness. Therefore, one of the main purposes of this paper is to better understand the potential of block smoothers like hybrid GS on millions of processors. We show that these hybrid smoothers can in fact exhibit good smoothing properties independent of parallelism, as long as the blocks satisfy certain properties (e.g., the blocks have some minimal size).

There are many other well-known smoothers that exhibit parallel-independent smoothing properties. In particular, methods like weighted Jacobi (both pointwise and blockwise), red/black GS, Chebyshev and Krylov-based polynomial methods have been extensively studied in classical works such as [12, 29, 19, 6]. In practice, each of these methods have their drawbacks. For example, weighted Jacobi requires the estimation of an ideal weight [32] and Chebyshev involves estimating an eigenvalue interval [1]. For multi-colored GS [1], the number of parallel communications required per iteration is proportional to the number of colors, hence it tends to be slow, especially on coarser grids in AMG where the number of colors is difficult to control. Therefore, the secondary purpose of this paper is to study and identify smoothers that are practical for AMG in the context of millions of processors. To this end, we analyze a variety of candidates for smoothing, revisiting some of the classics as well, under a common framework based on the recent two-grid theory in [17, 18]. Numerical results complementing the theory can be found in [4].

The structure of the paper is as follows. In Section 2, we introduce our approach for doing smoothing analysis in general, and we then analyze several specific classes of smoothers in Section 3 through Section 6, including C-F, polynomial, hybrid, and ℓ_1 smoothers. We make concluding remarks in Section 7.

2. Smoothing Analysis.

Our smoothing analysis is based on the two-grid variational multigrid theory from [17], which was developed for general relaxation and coarsening processes. In this section, we first summarize this theory and then describe our general approach for applying it to smoother analysis. We represent the standard Euclidean inner product by $\langle \cdot, \; \cdot \rangle$ with associated norm, $\|\cdot\| := \langle \cdot, \; \cdot \rangle^{1/2}$. The A-norm (or energy norm) is defined by $\|\cdot\|_A := \langle A \; \cdot, \; \cdot \rangle^{1/2}$ for vectors, and as the corresponding induced operator norm for matrices.

Consider solving the linear system of equations

$$A\mathbf{u} = \mathbf{f}, \tag{2.1}$$

where $\mathbf{u}, \mathbf{f} \in \mathbb{R}^n$ and A is a symmetric positive definite (SPD) matrix. Define the *smoother* (relaxation) error propagator by

$$I - M^{-1}A, \tag{2.2}$$

and assume that the smoother is convergent (in energy norm $\| \cdot \|_A$), i.e. assume that $M^T + M - A$ is SPD. Note that we often refer to the matrix M as the smoother. Denote the symmetrized smoother by

$$\widetilde{M} = M^T(M^T + M - A)^{-1}M, \tag{2.3}$$

so that $I - \widetilde{M}^{-1}A = (I - M^{-1}A)(I - M^{-T}A)$. Let $P : \mathbb{R}^{n_c} \mapsto \mathbb{R}^n$ be the *interpolation* (or *prolongation*) operator, where $\mathbb{R}^{n_c}$ is some lower-dimensional (coarse) vector space of size n_c. The two-grid multigrid error transfer operator with no post-smoothing steps is then given by

$$E_{TG} = (I - P(P^TAP)^{-1}P^TA)(I - M^{-1}A), \tag{2.4}$$

where P^T is the *restriction* operator and $A_c = P^T AP$ is the Galerkin *coarse-grid operator*. Note that coarse-grid correction involves an A-orthogonal projection onto range(P).

Let $R : \mathbb{R}^n \mapsto \mathbb{R}^{n_c}$ be any matrix for which $RP = I_c$, the identity on $\mathbb{R}^{n_c}$, so that PR is a projection onto range(P). We can think of R as defining the *coarse-grid variables*, i.e., $\mathbf{u}_c = R\mathbf{u}$. Also, let $S : \mathbb{R}^{n_s} \mapsto \mathbb{R}^n$ be any full-rank matrix for which $RS = 0$, where $n_s = n - n_c$. Here, the unknowns $\mathbf{u}_s = S^T\mathbf{u}$ are analogous to the fine-grid-only variables (i.e., F-points) in AMG. In addition, R and S form an orthogonal decomposition of $\mathbb{R}^n$: any $\mathbf{e}$ can be expressed as $\mathbf{e} = S\mathbf{e}_s + R^T\mathbf{e}_c$, for some $\mathbf{e}_s$ and $\mathbf{e}_c$. The next theorem summarizes one of the main convergence results in [17].

Theorem 2.1. *(see Theorem 2.2 in [17])*

$$\|E_{TG}\|_A^2 \le 1 - \frac{1}{K}, \quad \textit{where} \;\; K = \sup_{\mathbf{e}} \frac{\|(I - PR)\mathbf{e}\|_{\widetilde{M}}^2}{\|\mathbf{e}\|_A^2} \ge 1. \tag{2.5}$$

Theorem 2.1 gives conditions that P must satisfy in order to achieve a fast uniformly convergent multigrid method. It is clear that to make K small, eigenvectors of A belonging to small eigenvalues must either be interpolated accurately by P or else attenuated efficiently by the smoother (since the denominator is small for these eigenvectors). For brevity, we refer to these as small eigenvectors. The choice of which small eigenvectors to eliminate by smoothing and which to eliminate by coarse-grid correction depends on the "localness" of the modes. Essentially, modes that can be eliminated by a local process (i.e., one that is equivalent to applying an operator with a comparable sparse nonzero structure to A) should be handled by the smoother.

2.1. Smoothing Analysis with Ideal Interpolation. One approach for using the above theory to do smoothing analysis is to consider the best K in Theorem 2.1 by substituting the P that minimizes the following for a given R

$$K_\star = \inf_{P\,:\,RP=I_c} \sup_{\mathbf{e}} \frac{\|(I - PR)\mathbf{e}\|_{\widetilde{M}}^2}{\|\mathbf{e}\|_A^2}. \tag{2.6}$$

The following theorem evaluates this inf-sup problem.

Theorem 2.2. *(see Theorem 3.1 in [17]) Assume that R, S, and P satisfy $RS = 0$ and $RP = I_c$ as above. Then $K_\star$ in* (2.6) *is given by*

$$K_\star = \sup_{\mathbf{e}_s} \frac{\langle S^T \widetilde{M} S\mathbf{e}_s,\ \mathbf{e}_s\rangle}{\langle S^T AS\mathbf{e}_s,\ \mathbf{e}_s\rangle} = \frac{1}{\lambda_{\min}((S^T\widetilde{M}S)^{-1}(S^TAS))}, \tag{2.7}$$

and the corresponding minimizer is

$$P_\star = (I - S(S^TAS)^{-1}S^TA)R^T. \tag{2.8}$$

Equation (2.8) defines the so-called *ideal interpolation* operator. Notice that, if $K_\star$ is uniformly bounded with respect to parameters such as the mesh spacing, then using $P_\star$ as the interpolation operator results in a uniformly convergent two-grid method. Since the inverse of S^TAS may not be sparse, this is generally not a good practical choice for interpolation. However, it is reasonable to use $P_\star$ (and hence $K_\star$) to analyze smoothing.

We will consider two settings in the analysis that follows, depending on the particular smoother. The first is the classical AMG setting where the coarse-grid variables $R\mathbf{u}$ are a subset of the fine-grid variables:

$$R^T = \begin{bmatrix} 0 \\ I_c \end{bmatrix}; \quad S = \begin{bmatrix} I_f \\ 0 \end{bmatrix}; \quad P_\star = \begin{bmatrix} -A_{ff}^{-1}A_{fc} \\ I_c \end{bmatrix}. \tag{2.9}$$

The second setting corresponds more closely to the classical smoothing factor analysis [12], where the coarse-grid variables span the space of the n_c "smallest" (they do not have to strictly be the smallest as we discuss later) eigenvectors of A:

$$R^T = [\mathbf{v}_1, \ldots, \mathbf{v}_{n_c}]; \quad S = [\mathbf{v}_{n_c+1}, \ldots, \mathbf{v}_n]; \quad P_\star = R^T. \tag{2.10}$$

2.2. Comparative Smoothing Analysis. Direct evaluation of $K_\star$ in (2.6) is not always straightforward. However, one useful technique that we use below is to compare the $K_\star$ for one smoother to that of another with well-known smoothing properties (e.g., Gauss-Seidel). Writing $K = K(M)$ in (2.5) as a function of the smoother (similarly for $K_\star$), we articulate this approach in the next lemma.

LEMMA 2.3. *Suppose that M_1 and M_2 are two convergent smoothers for A that satisfy*

$$\langle \widetilde{M}_1\mathbf{x},\ \mathbf{x}\rangle \le c\langle \widetilde{M}_2\mathbf{x},\ \mathbf{x}\rangle \tag{2.11}$$

for all $\mathbf{x}$, with a fixed constant c. Then, for any choice of the interpolation operator in the two-grid multigrid method, we have that

$$K(M_1) \le cK(M_2),$$

and in particular, $K_\star(M_1) \le cK_\star(M_2)$. In other words, multigrid methods using M_1 and M_2 will have comparable parallel scalability properties, provided c is independent of the problem size and the number of processors. Therefore, when (2.11) *holds, we say that M_1 has multigrid smoothing properties comparable to M_2.*

Proof. The proof follows immediately from (2.5) and (2.6). □

REMARK 2.1. *Note that the above result can also be analogously stated in terms of the sharp two-grid theory of [18] since we can write the constant $K_\sharp$ in that theory as*

$$K_\sharp = \sup_{\mathbf{e}} \frac{\left\|(I - \pi_{\widetilde{M}})\mathbf{e}\right\|_{\widetilde{M}}^2}{\left\|\mathbf{e}\right\|_A^2} = \sup_{\mathbf{v}\in\text{range}(I-\pi_A)} \ \inf_{\mathbf{w}\,:\,\mathbf{v}=(I-\pi_A)\mathbf{w}} \frac{\langle \widetilde{M}\mathbf{w},\ \mathbf{w}\rangle}{\langle A\mathbf{v},\ \mathbf{v}\rangle},$$

where $\pi_X = P(P^TXP)^{-1}P^TX$ denotes the X-orthogonal projection onto range(P) *for any SPD matrix X.*

In some cases, we can directly determine the constant c in Lemma 2.3. However, we can also bound c in terms of a few general, yet insightful, constants as shown in the theorem below. First, we state a useful Lemma, which is of general interest.

LEMMA 2.4. *Suppose that A is SPD and B is arbitrary. Then*

$$\langle A\mathbf{x},\ \mathbf{x}\rangle \le c\langle B\mathbf{x},\ \mathbf{x}\rangle \qquad \textit{implies} \qquad \langle B^{-1}\mathbf{x},\ \mathbf{x}\rangle \le c\langle A^{-1}\mathbf{x},\ \mathbf{x}\rangle.$$

Proof. Note that by the given inequality B is invertible, and $B^{-1}+B^{-T}$ is SPD. Using Cauchy-Schwarz and the assumption above, we have

$$\begin{aligned}\langle B^{-1}\mathbf{x},\ \mathbf{x}\rangle^2 &= \langle A^{1/2}B^{-1}\mathbf{x},\ A^{-1/2}\mathbf{x}\rangle^2 \\ &\le \langle AB^{-1}\mathbf{x},\ B^{-1}\mathbf{x}\rangle\langle A^{-1}\mathbf{x},\ \mathbf{x}\rangle \\ &\le c\ \langle B^{-1}\mathbf{x},\ \mathbf{x}\rangle\langle A^{-1}\mathbf{x},\ \mathbf{x}\rangle.\end{aligned}$$

Dividing both sides by $\langle B^{-1}\mathbf{x},\ \mathbf{x}\rangle$ gives the desired result. □

The above Lemma implies, in particular, that if B is positive definite, i.e. its symmetric part $\sigma(B)=(B^T+B)/2$ is SPD, then

$$\langle B^{-1}\mathbf{x},\ \mathbf{x}\rangle \le \langle \sigma(B)^{-1}\mathbf{x},\ \mathbf{x}\rangle.$$

This inequality has appeared previously and can be found for example in [2], Lemma 3.5.

THEOREM 2.5. *Suppose that M_1 and M_2 are two convergent smoothers. Then,*

$$K(M_1) \le \frac{2\,\delta\Delta^2}{2-\omega}\ K(M_2), \tag{2.12}$$

where Δ, ω and δ are given by

$$\Delta = \|\sigma(M_1)^{-\frac{1}{2}}M_1\sigma(M_1)^{-\frac{1}{2}}\|, \quad \omega=\lambda_{\max}(\sigma(M_1)^{-1}A), \quad \delta=\sup_{\mathbf{v}}\frac{\langle M_1\mathbf{v},\ \mathbf{v}\rangle}{\langle M_2\mathbf{v},\ \mathbf{v}\rangle}\,. \tag{2.13}$$

This also holds with K replaced by $K_\star$ and with δ replaced by $\delta_s=\sup_{\mathbf{v}}\frac{\langle S^TM_1S\mathbf{v},\ \mathbf{v}\rangle}{\langle S^TM_2S\mathbf{v},\ \mathbf{v}\rangle}$.

Proof. From Lemma 2.3 in [17], and since $\langle\sigma(M)\mathbf{x},\ \mathbf{x}\rangle=\langle M\mathbf{x},\ \mathbf{x}\rangle$, we have

$$K(M_1)\le\frac{\Delta^2}{2-\omega}\ K_\sigma(M_1)\le\frac{\delta\Delta^2}{2-\omega}\ K_\sigma(M_2)\,,$$

where

$$K_\sigma(M)=\sup_{\mathbf{e}}\frac{\|(I-PR)\mathbf{e}\|^2_{\sigma(M)}}{\|\mathbf{e}\|^2_A}.$$

Since A is SPD, from (2.3) we have

$$\langle\widetilde{M}^{-1}\mathbf{x},\ \mathbf{x}\rangle=\langle(M^{-1}+M^{-T}-M^{-1}AM^{-T})\mathbf{x},\ \mathbf{x}\rangle\le 2\langle M^{-1}\mathbf{x},\ \mathbf{x}\rangle.$$

Hence, Lemma 2.4 implies that

$$\langle M\mathbf{x},\ \mathbf{x}\rangle\le 2\langle\widetilde{M}\mathbf{x},\ \mathbf{x}\rangle\,, \tag{2.14}$$

which completes the proof of (2.12). The result for $K_\star$ follows similarly from (2.7) and the definition of δ_s. □

The quantity Δ in (2.13) measures the deviation of M from its symmetric part, while $\omega\in(0,2)$ should be bounded away from two. In particular, $\omega\le 1$ is equivalent to

$$\langle A\mathbf{x},\ \mathbf{x}\rangle\le\langle M\mathbf{x},\ \mathbf{x}\rangle, \tag{2.15}$$

for all $\mathbf{x}$. When M is symmetric, this is a seemingly natural multigrid smoother condition, since it implies that $I - M^{-1}A$ will damp the (high-frequency) components of the error corresponding to the large eigenvalues of $M^{-1}A$. This is in contrast with the condition $\langle A\mathbf{x},\ \mathbf{x}\rangle \leq 2\langle M\mathbf{x},\ \mathbf{x}\rangle$ which is equivalent with M being convergent but allows ω to be close to 2, leading to minimal damping of the corresponding eigenvector. The difference is clearly illustrated in the case of Richardson's smoother $M = rI$, where $r = (\lambda_{\min} + \lambda_{\max})/2$ is optimal in terms of convergence, but $r = \lambda_{\max}$ has significantly better smoothing properties. An inequality like (2.15) also holds when M is not symmetric, in the sense that for any symmetrized smoother $\widetilde{M}$ we have

$$\langle A\mathbf{x},\ \mathbf{x}\rangle \leq \langle \widetilde{M}\mathbf{x},\ \mathbf{x}\rangle. \tag{2.16}$$

This can be seen in a couple of ways, for example, by introducing the SPD matrix $D_M = M + M^T - A$ and noting that

$$\begin{aligned}\widetilde{M} &= M^T D_M^{-1} M = (D_M + A - M) D_M^{-1} (D_M + A - M^T)\\ &= A + (A - M) D_M^{-1} (A - M^T).\end{aligned}$$

In particular, (2.15) holds for M defined as two sweeps of any convergent symmetric smoother, such as Jacobi for diagonally dominant and irreducible A. Since two sweeps of Jacobi is really no better as a smoother than one sweep, this example also illustrates the fact that ω alone is not in general a good measure of smoothing properties.

2.3. Historical Notes on Smoothing Analysis.

Our approach for analyzing smoothers has many similarities with previous approaches. As mentioned in Section 2.1, the idea of measuring (or bounding) the two-grid convergence factor by assuming an ideal interpolation operator is essentially what is done in classical smoothing factor analysis introduced in [12]. The smoothing factor measures the effectiveness of relaxation on the oscillatory Fourier modes, which is motivated by the assumption that interpolation (our ideal interpolation) eliminates the smooth Fourier modes. An important aspect of this approach is that it is explicitly tied to the (ideal) coarse-grid correction.

The approach described in Section 2.2 is similar to most other smoother analyses, where either weighted Richardson or Jacobi relaxation is used for M_2 in Lemma 2.3 [19, 8, 9, 26, 27, 25, 28, 10, 11]. A general comparison lemma was stated in [25]. One limitation of this approach is that coarse-grid correction is not explicitly taken into account, so in cases such as Maxwell's equation, care must be taken to compare with a suitable smoother.

For example, a number of multilevel smoothing conditions for multigrid were considered in Appendix B of [11]. The first smoothing condition there, (SM.1), combines (2.16) with a comparative condition of the form (2.11), where M_2 is a Richardson smoother. Since (SM.1) is the only requirement for the smoother (on each level) in results such as the classical Braess-Hackbusch Theorem 3.1 in [11], it is reasonable to expect that any analysis based on Lemma 2.3 will be applicable to the full multilevel multigrid algorithm (even though (2.11) was motivated by a two-grid theory). Another condition from [11] is (SM.2), which is a weighted version of (2.15). We note that (2.15) is not a new condition, and has been imposed on symmetric smoothers in previously published theories, e.g. [22].

3. The *C*-*F* Smoother.

In this section, we apply the smoothing analysis theory from the previous section to the so-called *C*-*F* smoother. *C*-*F* smoothing corresponds to applying an AMG smoother first to the coarse points (C-points) and then to the fine points (F-points). That C-F smoothers can be effective in practice is evident if one considers, for example, that C-F smoothing with Gauss-Seidel on a structured grid is equivalent to red-black Jacobi.

More formally, the C-F smoother is defined by

$$I - M_{CF}^{-1}A; \quad M_{CF} = \begin{bmatrix} M_{ff} & A_{fc} \\ 0 & M_{cc} \end{bmatrix}. \tag{3.1}$$

This smoother converges if and only if the following are convergent:

$$I_f - M_{ff}^{-1}A_{ff}; \quad I_c - M_{cc}^{-1}A_{cc}.$$

Therefore, one can consider using any of the convergent smoothers discussed in the following sections as the M_{ff} and M_{cc} matrices of a C-F smoother. This is typically advantageous since the principle submatrices A_{ff} and A_{cc} have better properties than A in terms of conditioning and diagonal dominance. The following theorem shows that C-F smoothing is good if F-relaxation is fast to converge.

THEOREM 3.1. *Define S as in* (2.9)*. Then $K_\star$ in* (2.6) *for the C-F smoother satisfies*

$$K_\star = \frac{1}{1 - {\varrho_f}^2}; \quad \varrho_f = \|I_f - M_{ff}^{-1}A_{ff}\|_{A_{ff}}.$$

Proof. Similarly to (2.3), define

$$\widetilde{M_{ff}} = M_{ff}^T(M_{ff}^T + M_{ff} - A_{ff})^{-1}M_{ff}\,. \tag{3.2}$$

From (2.3) and the definition of M_{CF} above, we have

$$\widetilde{M} = \begin{bmatrix} M_{ff}^T & 0 \\ A_{cf} & M_{cc}^T \end{bmatrix} \begin{bmatrix} (M_{ff}^T + M_{ff} - A_{ff})^{-1} & 0 \\ 0 & (M_{cc}^T + M_{cc} - A_{cc})^{-1} \end{bmatrix} \begin{bmatrix} M_{ff} & A_{fc} \\ 0 & M_{cc} \end{bmatrix},$$

and therefore $S^T\widetilde{M}S = \widetilde{M_{ff}}$. This implies by Theorem 2.2

$$K_\star = \frac{1}{\lambda_{\min}(\widetilde{M_{ff}}^{\,-1}A_{ff})} = \frac{1}{1 - \lambda_{\max}[(I - M_{ff}^{-1}A_{ff})(I - M_{ff}^{-T}A_{ff})]}.$$

Let $E_{ff} = I_{ff} - M_{ff}^{-1}A_{ff}$ and let $\rho(\cdot)$ denote the spectral radius of a matrix. Then, using the definition of ϱ_f and the fact that $\|B\| = \|B^T\|$ for any matrix B, we have

$$\begin{aligned} {\varrho_f}^2 = \|E_{ff}\|_{A_{ff}}^2 &= \|A_{ff}^{1/2}E_{ff}A_{ff}^{-1/2}\|^2 = \|A_{ff}^{-1/2}E_{ff}^TA_{ff}^{1/2}\|^2 \\ &= \rho(A_{ff}^{1/2}E_{ff}A_{ff}^{-1}E_{ff}^TA_{ff}^{1/2}) = \rho(E_{ff}A_{ff}^{-1}E_{ff}^TA_{ff}) \\ &= \rho[(I - M_{ff}^{-1}A_{ff})(I - M_{ff}^{-T}A_{ff})]\,, \end{aligned}$$

which completes the proof. □

From the above, we see that C-F smoothing is a natural smoother to use when coarse grids are selected based on compatible relaxation (CR) [13, 17], because ϱ_f is estimated as part of the CR coarsening algorithm.

4. Polynomial Smoothers.

Polynomial smoothers are of practical interest for parallel computing for a couple of reasons. First, their application requires only the matrix-vector multiply routine, which is often highly-optimized on modern parallel machines. Second, they are unaffected by the parallel partitioning of the matrix, the number of parallel processes, and the ordering of the unknowns. However, as mentioned previously, one drawback is the need to calculate eigenvalue estimates. Unlike the smoothed aggregation variant of AMG, eigenvalue estimates are not needed by classical AMG, so this computational cost is extra.

We now apply the smoothing analysis from Section 2 to polynomial smoothers. Let $p_\nu(x)$ be a polynomial of degree $\nu \geq 0$ such that $p_\nu(0) = 1$, and consider the smoother

$$I - M^{-1}A = p_\nu(A). \tag{4.1}$$

The following theorem gives conditions for a good polynomial smoother.

THEOREM 4.1. *Let $A = V\Lambda V^T$ be the eigen-decomposition of A with eigenvectors $\mathbf{v}_k$ and associated eigenvalues λ_k, and define S as in* (2.10)*. Then $K_\star$ in* (2.6) *for the polynomial smoother satisfies*

$$K_\star = \left(1 - \max_{k>n_c} p_\nu(\lambda_k)^2\right)^{-1}.$$

Minimizing $K_\star$ over all p_ν, we have

$$\min_{p_\nu} K_\star \leq \left(1 - \left(\min_{p_\nu} \max_{x\in[\alpha,\beta]} |p_\nu(x)|\right)^2\right)^{-1}; \quad \alpha \leq \lambda_{n_c+1} \leq \lambda_n \leq \beta.$$

Proof. Order the eigenvectors in V so that we can write $S = VS_i$, $S_i = [I_s, 0]^T$. Then, since $I - \widetilde{M}^{-1}A = (I - M^{-1}A)(I - M^{-T}A)$, we have

$$\begin{aligned} S^T\widetilde{M}S &= S^T(A^{-1} - (I - M^{-1}A)A^{-1}(I - M^{-1}A)^T)^{-1}S \\ &= S_i^T V^T(A^{-1} - p_\nu(A)^2A^{-1})^{-1}VS_i \\ &= S_i^T(\Lambda^{-1} - p_\nu(\Lambda)^2\Lambda^{-1})^{-1}S_i \\ &= (\Lambda_s^{-1} - p_\nu(\Lambda_s)^2\Lambda_s^{-1})^{-1}. \end{aligned}$$

Since $S^TAS = \Lambda_s$, then

$$(S^T\widetilde{M}S)^{-1}(S^TAS) = I_s - p_\nu(\Lambda_s)^2,$$

and the first result follows from Theorem 2.2. The second result follows trivially from the first since we are maximizing over a larger set $[\alpha, \beta]$ containing λ_k, $k > n_c$. □

In the following two subsections, we first discuss the optimal polynomial smoother according to Theorem 4.1 and then briefly overview several other choices of polynomials that may also be good smoothers for AMG in practice.

4.1. Chebyshev Smoothers.

The min-max problem in Theorem 4.1 has a classical solution $q_\nu(x)$ in terms of Chebyshev polynomials (see, e.g., [2]). Let $T_k(t)$ be the Chebyshev polynomial of degree k defined by the recursion

$$T_0(t) = 1; \quad T_1(t) = t; \quad T_k(t) = 2tT_{k-1}(t) - T_{k-2}(t), \; k = 2, 3, \ldots \tag{4.2}$$

By letting $t = \cos(\xi) \in [-1, 1]$, it is easy to show that the explicit form of these polynomials is $T_k(t) = \cos(k\xi)$. The polynomial $q_\nu(x)$ is given by

$$q_\nu(x) = \frac{T_\nu\left(\frac{\beta+\alpha-2x}{\beta-\alpha}\right)}{T_\nu\left(\frac{\beta+\alpha}{\beta-\alpha}\right)}, \tag{4.3}$$

and has the required property that $q_\nu(0) = 1$. It also satisfies

$$-1 < q_\nu(x) < 1 \text{ for } x \in (0, \beta],$$

which implies that the smoother (4.1) with $p_\nu = q_\nu$ is convergent as long as the spectrum of A is contained in the interval $(0, \beta]$. To show the above inequality with $\alpha, \beta > 0$ observe that the Chebyshev polynomial $T_\nu(x)$ equals 1 for $x = 1$ and is strictly monotonically increasing for $x > 1$ (see, e.g., (5.28) in [2]). Therefore, $x \in [\alpha, \beta]$ implies $T_\nu\left(\frac{\beta+\alpha}{\beta-\alpha}\right) > 1 \geq \left|T_\nu\left(\frac{\beta+\alpha-2x}{\beta-\alpha}\right)\right|$, while $|q_\nu(x)| < 1$ due to $\frac{\beta+\alpha}{\beta-\alpha} > \frac{\beta+\alpha-2x}{\beta-\alpha} \geq 1$ for $x \in (0, \alpha]$.

Since $K_\star$ is a measure of the smoothing properties of the smoother (4.1), then Theorem 4.1 shows that a good choice for polynomial smoothing is $q_\nu(x)$ where the interval $[\alpha, \beta]$ contains the "large" eigenvalues of A. The upper bound β can easily be estimated using a few iterations of conjugate gradient (CG), but choosing a suitable α is not obvious in general. It is clear that α depends on the coarse-grid size, but it should also depend on the distribution of eigenvalues for the problem and possibly even the nature of the associated eigenvectors. To see this, consider a simple Laplace example on a unit domain discretized by standard finite differences. Assume full coarsening so that $n_c/n = 1/2^d$ where d is the dimension. We discuss three possible choices for α below.

First, note that the analysis above does not require that R be made up of the strictly smallest eigenvectors of A. Consider instead that R contains the smooth Fourier modes used in standard local Fourier analysis. In this case, it is easy to see from standard Fourier diagrams that α should be chosen such that

$$\alpha/\beta = 1/2 \text{ (1D)},\ 1/4 \text{ (2D)},\ 1/6 \text{ (3D)}. \tag{4.4}$$

The resulting Chebyshev polynomial smoothers were first derived almost 30 years ago in [29].

Now consider letting R contain the actual n_c smallest eigenvectors for the Laplace equation. Using Matlab, we get the estimates

$$\alpha/\beta \approx 0.5 \text{ (1D)},\ 0.32 \text{ (2D)},\ 0.28 \text{ (3D)}. \tag{4.5}$$

Consider again letting R contain the actual n_c smallest eigenvectors, but assume that the eigenvalues are distributed uniformly. Then, we have

$$\alpha/\beta = 1/2 \text{ (1D)},\ 1/4 \text{ (2D)},\ 1/8 \text{ (3D)}. \tag{4.6}$$

In practice, we set β by estimating $\lambda_{\max}$ with several iterations of CG and set $\alpha = a\beta$ for some fraction $0 \leq a \leq 1$. We use $a = 0.3$ in the numerical experiments in [4]. A similar approach is used in [1], but with a small $a = 1/30$. It is not vital to estimate $\lambda_{\min}$ unless it is large. In that case, no coarse grid is needed, and the smoother should damp all eigenvectors equally well, i.e., α should approximate $\lambda_{\min}$.

4.2. Other Polynomial Smoothers. Although the above theory leads naturally to the Chebyshev polynomial in (4.3), there are several other polynomials in the literature that are also good smoothers. We briefly summarize some of the most notable here.

A related smoother to the Chebyshev polynomial in (4.3) is the following shifted and scaled Chebyshev polynomial used in the AMLI method [3]

$$q_\nu^+(x) = \frac{1 + T_\nu\left(\frac{\beta+\alpha-2x}{\beta-\alpha}\right)}{1 + T_\nu\left(\frac{\beta+\alpha}{\beta-\alpha}\right)}. \tag{4.7}$$

This has the required property that $q_\nu^+(0) = 1$, but satisfies

$$0 < q_\nu^+(x) < 1 \text{ for } x \in (0, \beta].$$

This implies that (2.15) holds, a sometimes desirable property for smoothers.

Another polynomial smoother of interest is used in both the smoothed aggregation (SA) and cascadic multigrid methods [7, 14, 30], and is given by

$$\phi_\nu(x) = (-1)^\nu \left(\frac{1}{2\nu+1}\right)\left(\frac{\sqrt{\beta}}{\sqrt{x}}\right) T_{2\nu+1}\left(\frac{\sqrt{x}}{\sqrt{\beta}}\right). \tag{4.8}$$

Note that (4.8) does not require the estimation of α. It can be shown that ϕ_ν is the minimizer of

$$\min_{p_\nu} \max_{x\in[0,\beta]} |\sqrt{x}\, p_\nu(x)|. \tag{4.9}$$

The weak approximation property in (2.5) shows that coarse-grid correction must eliminate eigenvectors with accuracy proportional to the square root of their associated eigenvalue. The $\sqrt{x}$ term in (4.9) serves the role of coarse-grid correction, so the polynomial ϕ_ν has a certain optimality with respect to the weak approximation property. However, (4.9) does not account for the fact that coarse-grid correction only operates on a subspace of size n_c, so the resulting smoother ϕ_ν does not damp the largest eigenvectors (i.e., those not damped at all by coarse-grid correction) as much as it otherwise would. It could be modified to satisfy (4.9) over the interval $[\alpha, \beta]$ to improve its properties as a smoother, but it is not clear that this is better than using the Chebyshev polynomial in (4.3). Note that the polynomial ϕ_ν makes perfect sense for smoothing a tentative interpolation operator, which is its primary purpose.

The MLS smoother in [1] is the product of ϕ_ν and a complementary (post) smoother of the form

$$I - \frac{\omega}{\lambda_{\max}(\phi_\nu^2 A)}\phi_\nu^2 A.$$

It has better overall smoothing properties than ϕ_ν alone, and it is particularly advantageous when using aggressive coarsening.

The polynomial smoother in [24] minimizes an equation like (4.9) over the interval $[\alpha, \beta]$, but with $\sqrt{x}$ in the equation replaced by $1/x$. This means that the amplitude of the polynomial increases over the interval $[\alpha, \beta]$. The polynomial is computed through a three-term recurrence.

The conjugate gradient method is also a good smoother [6]. Note that it converges to the Chebyshev polynomial in (4.3), but over the entire eigenvalue interval

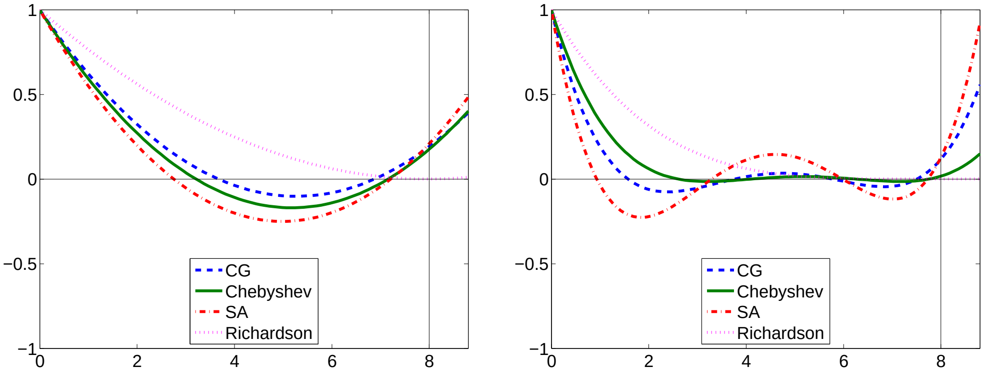

FIG. 4.1. *Various polynomials of order two (left) and four (right). The CG polynomial was generated by solving a 2D Laplace problem on a* 25×25 *grid with a random initial error. The Chebyshev polynomial* (4.3) *uses* $a = 0.3$*. The SA polynomial is given by* (4.8).

$[\alpha, \beta] = [\lambda_{\min}, \lambda_{\max}]$. Even though this is not a good value for α, it is only relevant asymptotically; for small ν, CG has good smoothing properties. Other Krylov methods such as conjugate residual (also called minimum residual or MINRES) typically have good smoothing properties as well [6].

In Figure 4.1 we plot several polynomials over the eigenvalue interval. Focusing on the fourth-order figure, note that the polynomial tails for $x > \beta = 8$ turn up steeply. For this reason, it is important not to underestimate $\lambda_{\max}$ in practice. Note also that the CG polynomial closely approximates Chebyshev. As previously mentioned, the SA polynomial does not damp the large eigenvectors as well as the others. The Richardson polynomial is given by $(I - \lambda_{\max}^{-1})^{\nu}$. We include it in the figure because it is the simplest smoother to understand and it is used in most classical smoothing analysis. In the interest of keeping the figure readable, we do not plot all of the polynomials in this section. Note, however, that they all have good smoothing properties, with mostly minor differences between them as noted in the text.

5. The Hybrid Smoother. The class of so-called hybrid smoothers can be viewed as the result of the straightforward parallelization of a smoother. For example, the easiest parallelization of GS is to have each process independently use GS on its domain and then exchange information from neighbor processors after each iteration, resulting in a Jacobi-like update at the processor boundaries. As noted in Section 1, hybrid smoothers, like hybrid GS in particular, are of interest because they are easy to implement and often quite effective in practice, even though convergence may not be guaranteed. In this section, we first formally define hybrid smoothers and apply the smoothing analysis theory from Section 2. We then discuss two particular hybrid smoothers, hybrid GS and Block Jacobi, in more detail, and, finally, we discuss the use of weights with hybrid smoothers.

We define the *hybrid smoother* to be essentially an inexact block Jacobi method. Specifically, let $\Omega = \{1, \ldots, n\}$ and consider the non-overlapping partition of Ω,

$$\Omega = \bigcup_{k=1}^{p} \Omega_k.$$

Of particular practical interest in this paper is the case where Ω_k represents the unknowns on processor k so that p is the total number of processors, but the analysis below is for the general setting. Let A be partitioned into blocks A_{kl} of size $n_k \times n_l$ where the rows of A_{kl} are in Ω_k and the columns are in Ω_l. Let $I - B_k^{-1} A_{kk}$ be a smoother for A_{kk}. Then, the hybrid smoother is defined by

$$(5.1) \qquad I - M_H^{-1} A; \quad M_H = \mathrm{diag}\{B_k\},$$

where $\mathrm{diag}\{B_k\}$ denotes the block-diagonal matrix with blocks B_k.

If $B_k = A_{kk}$, then (5.1) is block Jacobi. As p increases, the convergence of block Jacobi approaches that of pointwise Jacobi. However, although (unweighted) pointwise Jacobi is often not a good smoother, we show below that block Jacobi and other hybrid smoothers can have good smoothing properties independent of p, as long as the blocks are sufficiently large. We also show that this threshold block size can be quite small. We first discuss the convergence properties of the hybrid smoother.

Assume that the block smoothers are convergent in the sense of (2.15), that is

$$\langle B_k \mathbf{v}_k,\ \mathbf{v}_k \rangle \ge \langle A_{kk} \mathbf{v}_k,\ \mathbf{v}_k \rangle .$$

Then, $\langle (B_k^T + B_k - A_{kk}) \mathbf{v}_k,\ \mathbf{v}_k \rangle \ge \langle A_{kk} \mathbf{v}_k,\ \mathbf{v}_k \rangle$. To show that the hybrid smoother is convergent, we need to show that $M_H^T + M_H - A$ is SPD. With $\mathbf{v}$ composed of blocks $\mathbf{v}_k \in \mathbb{R}^{n_k}$, we have that

$$\begin{aligned} \langle (M_H^T + M_H - A)\mathbf{v},\ \mathbf{v} \rangle &= \sum_k \langle (B_k^T + B_k - A_{kk})\mathbf{v}_k,\ \mathbf{v}_k \rangle - \sum_k \sum_{l \ne k} \langle A_{kl} \mathbf{v}_l,\ \mathbf{v}_k \rangle \\ &\ge \sum_k \langle A_{kk} \mathbf{v}_k,\ \mathbf{v}_k \rangle - \sum_k \sum_{l \ne k} \langle A_{kl} \mathbf{v}_l,\ \mathbf{v}_k \rangle . \end{aligned}$$

One class of matrices for which the latter is positive is the class of block red-black matrices, i.e., when A admits the following two-by-two form

$$A = \begin{bmatrix} A_{rr} & A_{rb} \\ A_{br} & A_{bb} \end{bmatrix},$$

with block-diagonal matrices A_{rr} and A_{bb}. To see this, note that SPD A implies

$$\sum_k \sum_l \langle A_{kl} \mathbf{v}_l,\ \mathbf{v}_k \rangle > 0.$$

Replacing $\mathbf{v}_k$ with $\epsilon_k \mathbf{v}_k$ for $\epsilon_k = 1$ or $\epsilon_k = -1$, we obtain

$$\begin{aligned} \sum_k \langle A_{kk} \mathbf{v}_k,\ \mathbf{v}_k \rangle &> - \sum_k \sum_{l \ne k} \epsilon_k \epsilon_l \ \langle A_{kl} \mathbf{v}_l,\ \mathbf{v}_k \rangle \\ &= \sum_k \sum_{l \ne k} \langle A_{kl} \mathbf{v}_l,\ \mathbf{v}_k \rangle, \end{aligned}$$

where the last equality holds by choosing $\epsilon_k = 1$ for the "red" blocks and $\epsilon_k = -1$ for the "black" blocks. In that case, $\epsilon_k \epsilon_l = -1$ for any $k \ne l$ where $A_{kl} \ne 0$. As a practical example of a block red-black matrix, consider a structured (i.e. topologically Cartesian) partitioning of a 5-point discretization in 2D.

To analyze the smoothing properties of the hybrid smoother, we introduce a constant, $\theta \ge 0$, which is a measure of the relative size of the block off-diagonal portion of A. First, define the sets

$$(5.2) \qquad \Omega^{(i)} = \{ j \in \Omega_k \ :\ i \in \Omega_k \}; \quad \Omega_o^{(i)} = \{ j \notin \Omega_k \ :\ i \in \Omega_k \}.$$

Hence, $\Omega^{(i)}$ is the set of columns in the diagonal block for row i while $\Omega_o^{(i)}$ contains the remaining “off-diagonal” columns in row i. Now, with a_{ij} denoting the coefficients of A, define θ such that

$$a_{ii} \geq \theta \sum_{j \in \Omega_o^{(i)}} |a_{ij}| \quad \text{for all rows } i. \tag{5.3}$$

Under weak scaling, θ will quickly stabilize to a value independent of the number of processors. In many applications this value will satisfy $\theta > 1$. This, for example, is the case when A is diagonally dominant and each A_{kk} has at least two non-zero entries per row (in particular the block sizes are large enough). Another example is the 5-point discretization of the Laplacian in 2D, where $\theta = 2$. In general θ is large whenever most of the strong connections for each i (relatively large $|a_{ij}|$) are contained inside its block. For finite element discretizations, better values for θ are obtained when the blocks correspond to element partitioning (as opposed to random partitioning of the degrees of freedom, see Section 7.3 in [4].

5.1. Hybrid Gauss-Seidel. In this section we consider the hybrid Gauss-Seidel smoother M_{HGS}, which is obtained when the blocks B_k in (5.1) are chosen to be Gauss-Seidel sweeps for A_{kk}. This smoother is of practical importance, for example, because it is the default option in the BoomerAMG code.

Let $A = D + L + L^T$, where D is the diagonal of A and L and L^T are its strictly lower and upper triangular parts. We first remark that M_{HGS} is convergent if $\theta > 1$ or if A is red-black both with and without the block partitioning. Indeed, the θ condition implies

$$\langle D\mathbf{v},\ \mathbf{v}\rangle \leq \frac{\theta}{\theta - 1} \langle (M_{HGS}^T + M_{HGS} - A)\mathbf{v},\ \mathbf{v}\rangle\,, \tag{5.4}$$

while in the red-black case we have that both regular and block Jacobi are convergent, and therefore

$$2\langle M_{HGS}\mathbf{v},\ \mathbf{v}\rangle = \sum_k \langle A_{kk}\mathbf{v}_i,\ \mathbf{v}_k\rangle + \langle D\mathbf{v},\ \mathbf{v}\rangle > \langle A\mathbf{v},\ \mathbf{v}\rangle\,.$$

Note that if A has large positive off-diagonal entries, such as in discretizations of definite Maxwell problems, M_{HGS} may be divergent, even for large block sizes. This was the motivation in [23] to develop the ℓ_1 smoothers considered in the next section.

In the next two theorems, we compare the smoothing properties of M_{HGS} to that of the standard Gauss-Seidel smoother $M_{GS} = D + L$. We first estimate the constants in Theorem 2.5 to do the comparison.

THEOREM 5.1. *Assume that A is diagonally dominant and M_{HGS} corresponds to hybrid red-black Gauss-Seidel. Then,*

$$K(M_{HGS}) \leq \frac{4(3\theta - 1)}{3(\theta - 1)}\, K(M_{GS}).$$

Proof. We use Theorem 2.5. First, note that if γ satisfies $\langle A\mathbf{v},\ \mathbf{v}\rangle \leq \gamma\langle D\mathbf{v},\ \mathbf{v}\rangle$ for all $\mathbf{v}$, then

$$\omega \leq \frac{2}{1 + \frac{\theta - 1}{\gamma\theta}}.$$

		$\|E_{TG}\|_A^2$		$K_\star$	
m	p	BJac	HGS	BJac	HGS
512	1	0.00	0.20	1.00	1.25
256	2	0.50	0.32	65.12	1.81
128	4	0.50	0.32	110.62	1.81
32	16	0.51	0.32	418.96	1.81
16	32	0.53	0.32	834.93	1.81
4	128	0.56	0.41	3334.24	1.81
2	256	0.56	0.39	6667.23	2.33
1	512	1.00	1.00	26664.93	26664.93

Table 5.1

Convergence factors and constants from Theorem 2.1 for (unweighted) block Jacobi (BJac) and hybrid GS (HGS) for a 1D Laplace problem with m unknowns per block and p blocks.

This follows from the definition of ω in (2.13), the fact that $2\sigma(M_{HGS}) = A+(M_{HGS}^T+M_{HGS}-A)$, and (5.4). From the assumptions, $\delta \leq 2$ and $\gamma \leq 2$, so $\omega \leq (4\theta)/(3\theta-1)$. It is not difficult to show that $\Delta^2 \leq 4/3$ for hybrid red-black GS. □

Next, we compute the constant in Lemma 2.3 directly. Note that this approach requires less assumptions, but also gives a worse estimate when θ is close to one.

Theorem 5.2. *Assume that $\theta > 1$. Then*

$$K(M_{HGS}) \leq \frac{\theta}{\theta-1}\left(1+\frac{2}{\theta}\right)^2 K(M_{GS}).$$

Proof. Analogous to Theorem 6.2 from the next section, using (5.4) and the fact that

$$\|(M_{HGS}-M_{GS})\mathbf{x}\|_{D^{-1}}^2 \leq \frac{1}{\theta^2}\langle D\mathbf{x},\ \mathbf{x}\rangle \leq \frac{4}{\theta^2}\langle \widetilde{M_{GS}}\mathbf{x},\ \mathbf{x}\rangle.$$

□

By Theorem 5.1 and Theorem 5.2 we can conclude that hybrid Gauss-Seidel will be a convergent smoother with smoothing properties comparable to full Gauss-Seidel provided that $\theta > 1$, e.g. if A is diagonally dominant and each block is large enough to have at least two non-zero entries per row.

5.2. Block Jacobi. As mentioned in the beginning of Section 5, the hybrid smoother can be thought of as an inexact block Jacobi method. Since hybrid GS can be shown to have smoothing properties comparable to GS under certain conditions, it seems plausible that (unweighted) block Jacobi might have even better smoothing properties. In fact, block Jacobi is not a particularly good smoother, though it can exhibit smoothing properties independent of the number of blocks (processors).

As an example, consider again a standard Laplace problem on a unit domain with homogeneous Dirichlet boundary conditions . In Table 5.1, we report $\|E_{TG}\|_A^2$ from Theorem 2.1 for the ideal interpolation operator $P_\star$ in (2.9) for a coarsening factor of two in 1D. We also report the corresponding $K_\star$. From the table, we see that hybrid GS is a better smoother than block Jacobi, while both methods appear to have p-independent convergence factors for $m > 1$ (this is easily confirmed by fixing m and increasing p; not shown). At $m = 1$, both methods degenerate into unweighted pointwise Jacobi, which is known to have poor smoothing properties. We also see from the table that $K_\star$ is stable for hybrid GS but unbounded for block Jacobi (additional numerics shows that $K_\star$ depends on both m and p). This implies that the theoretical tools in Section 2 are not adequate for analyzing block Jacobi. Although not the best smoother choice in practice, we would like to get a deeper understanding of block Jacobi's smoothing properties. One approach might be to base the analysis on the sharp theory in [18], but we have not yet pursued this.

The observations from Table 5.1 also carry over to 2D (we have not done 3D experiments), but they are more pronounced. In particular, the convergence factor for $m \geq (2 \times 2)$ approaches 0.76 for block Jacobi instead of 0.56 as in 1D, while hybrid GS stays at 0.39. Another item worth noting is that the convergence of both methods degrades for larger coarsening factors, as one would expect. In addition, for block Jacobi, the minimum block size needed to yield good smoothing properties increases with increasing coarsening factor. It remains the same for hybrid GS as indicated by the theory.

5.3. Using Weights in Hybrid Smoothers.

While we have shown that for many problems hybrid smoothers converge well, there are various situations where this is not the case, see e.g. Section 7.3 in [4]. Convergence can be achieved by multiplying M_H with a weight ω as follows:

$$M_\omega = \omega M_H.$$

If M_H is SPD and $\omega = \lambda_{max}(M_H^{-1/2} A M_H^{-1/2})$, we immediately get (2.15). In practice, ω can be obtained by the use of Lanczos or CG iterations. For further details on the use of relaxation weights in hybrid smoothers, see [32].

6. The ℓ_1 Smoother.

While weighted hybrid smoothers are an attempt to fix hybrid smoothers by multiplying them with a suitable parameter, ℓ_1 *smoothers* do so by adding an appropriate diagonal matrix, which also leads to guaranteed convergence. They have the additional benefit of not requiring eigenvalue estimates. The ℓ_1 *smoother* is defined by

$$I - M_{\ell_1}^{-1} A; \quad M_{\ell_1} = M_H + D^{\ell_1} = \operatorname{diag}\{B_k + D_k^{\ell_1}\}, \tag{6.1}$$

where D^{ℓ_1} is a diagonal matrix with entries

$$d_{ii}^{\ell_1} = \sum_{j \in \Omega_o^{(i)}} |a_{ij}| \,.$$

Note that with this notation (5.3) is simply $D \geq \theta D^{\ell_1}$. Furthermore, D^{ℓ_1} has the important property that

$$\langle A\mathbf{v},\ \mathbf{v}\rangle \leq \sum_k \langle A_{kk}\mathbf{v}_k,\ \mathbf{v}_k\rangle + \langle D^{\ell_1}\mathbf{v},\ \mathbf{v}\rangle \,, \tag{6.2}$$

which follows from the Schwarz inequality $2|a_{ij}v_iv_j| \leq |a_{ij}|v_i^2 + |a_{ij}|v_j^2$.

We first show that M_{ℓ_1} is A-convergent, i.e., that $M_{\ell_1}^T + M_{\ell_1} - A$ is SPD. In the case where $B_k = A_{kk}$, we can actually show more, since (6.2) implies (2.15). In general, if the block smoothers B_k are non-divergent in the A_{kk}-norm with at least one of them being convergent, then

$$\langle A_{kk}\mathbf{v}_k,\ \mathbf{v}_k\rangle \leq \langle (B_k^T + B_k)\mathbf{v}_k,\ \mathbf{v}_k\rangle$$

with strict inequality holding for at least one k. Hence, from (6.2),

$$\langle A\mathbf{v},\ \mathbf{v}\rangle < \sum_k \langle (B_k^T + B_k + D_k^{\ell_1})\mathbf{v}_k,\ \mathbf{v}_k\rangle \le \langle (M_{\ell_1}^T + M_{\ell_1})\mathbf{v},\ \mathbf{v}\rangle.$$

REMARK 6.1. *The following scaled ℓ_1 smoother is also A-convergent:*

$$M_{\ell_1} = diag\{B_k + \frac{1}{2}D_k^{\ell_1}\}.$$

6.1. ℓ_1 Jacobi. We first consider the ℓ_1 point Jacobi smoother $M_{\ell_1 J} = D + D^{\ell_1}$ with blocks of size one. From above, this smoother is always convergent and satisfies (2.15). In the next theorem we compare $M_{\ell_1 J}$ to standard GS using Theorem 2.5. Note that since the blocks are of size one, θ satisfies $a_{ii} \ge \theta \sum_{j\neq i} |a_{ij}|$.

THEOREM 6.1. *Without any restrictions, we have*

$$K(M_{\ell_1 J}) \le 4\left(1 + \frac{1}{\theta}\right) K(M_{GS}).$$

In particular, ℓ_1 Jacobi has multigrid smoothing properties comparable to full Gauss-Seidel for any A, for which θ is bounded away from zero.

Proof. Since $M_{\ell_1 J}$ is symmetric and satisfies (2.15) we can take $\Delta = 1$ and $\omega = 1$. To estimate δ we observe that

$$\langle M_{\ell_1 J}\mathbf{x},\ \mathbf{x}\rangle \le \left(1 + \frac{1}{\theta}\right)\langle D\mathbf{x},\ \mathbf{x}\rangle \le \left(1 + \frac{1}{\theta}\right) 2\langle M_{GS}\mathbf{x},\ \mathbf{x}\rangle\,.$$

□

6.2. ℓ_1 Gauss-Seidel. Finally, let $M_{\ell_1 GS} = M_{HGS} + D^{\ell_1}$ be the ℓ_1 Gauss-Seidel smoother. This is the default smoother used in the AMS code [23]. As shown earlier in this section, this smoother is always convergent, and we analyze it by directly computing the constant in Lemma 2.3.

THEOREM 6.2. *Without any restrictions, we have*

$$K(M_{\ell_1 GS}) \le \left(1 + \frac{4}{\theta}\right)^2 K(M_{GS}).$$

In particular, ℓ_1 Gauss-Seidel has multigrid smoothing properties comparable to full Gauss-Seidel for any A, for which θ is bounded away from zero, independently of the number of blocks (processors) or the block sizes.

Proof. First, observe that $\langle M_{\ell_1 GS}\mathbf{x},\ \mathbf{x}\rangle \ge \langle M_{GS}\mathbf{x},\ \mathbf{x}\rangle$ implies

$$\langle D\mathbf{x},\ \mathbf{x}\rangle \le \langle (M_{\ell_1 GS}^T + M_{\ell_1 GS} - A)\mathbf{x},\ \mathbf{x}\rangle.$$

Therefore,

$$\langle \widetilde{M_{\ell_1 GS}}\mathbf{x},\ \mathbf{x}\rangle = \langle (M_{\ell_1 GS}^T + M_{\ell_1 GS} - A)^{-1} M_{\ell_1 GS}\mathbf{x},\ M_{\ell_1 GS}\mathbf{x}\rangle \le \|M_{\ell_1 GS}\mathbf{x}\|_{D^{-1}}^2.$$

By the triangle inequality in the D^{-1}-inner product,

$$\|M_{\ell_1 GS}\mathbf{x}\|_{D^{-1}} \le \|M_{GS}\mathbf{x}\|_{D^{-1}} + \|(M_{\ell_1 GS} - M_{GS})\mathbf{x}\|_{D^{-1}}.$$

The first term above is simply $\langle \widetilde{M_{GS}}\mathbf{x},\ \mathbf{x}\rangle^{1/2}$, while the second can be estimated as follows using the Schwarz inequality (in lines 2 and 4), the symmetry of A (in line 5), and the fact that A is SPD together with (2.14) (in line 6):

$$
\begin{aligned}
\|(M_{\ell_1 GS} - M_{GS})\mathbf{x}\|^2_{D^{-1}} &= \sum_i \frac{1}{a_{ii}} \Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| x_i - \sum_{\substack{j \in \Omega_o^{(i)} \\ j<i}} a_{ij} x_j \Bigg)^2 \\
&\leq \sum_i \frac{1}{a_{ii}} \Bigg[\Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| \Bigg)^{1/2} \Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| x_i^2 \Bigg)^{1/2} + \Bigg(\sum_{\substack{j \in \Omega_o^{(i)} \\ j<i}} |a_{ij}| \Bigg)^{1/2} \Bigg(\sum_{\substack{j \in \Omega_o^{(i)} \\ j<i}} |a_{ij}| x_j^2 \Bigg)^{1/2} \Bigg]^2 \\
&\leq \sum_i \frac{1}{a_{ii}} \Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| \Bigg) \Bigg[\Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| x_i^2 \Bigg)^{1/2} + \Bigg(\sum_{\substack{j \in \Omega_o^{(i)} \\ j<i}} |a_{ij}| x_j^2 \Bigg)^{1/2} \Bigg]^2 \\
&\leq 2 \sum_i \frac{1}{a_{ii}} \Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| \Bigg) \Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| x_i^2 + \sum_{\substack{j \in \Omega_o^{(i)} \\ j<i}} |a_{ij}| x_j^2 \Bigg) \\
&\leq \frac{2}{\theta} \sum_i \Bigg(\sum_{j \in \Omega_o^{(i)}} |a_{ij}| + \sum_{\substack{j \in \Omega_o^{(i)} \\ j>i}} |a_{ij}| \Bigg) x_i^2 \\
&\leq \frac{4}{\theta^2} \langle D\mathbf{x},\ \mathbf{x}\rangle \leq \frac{8}{\theta^2} \langle M_{GS}\mathbf{x},\ \mathbf{x}\rangle \leq \frac{16}{\theta^2} \langle \widetilde{M_{GS}}\mathbf{x},\ \mathbf{x}\rangle.
\end{aligned}
$$

The desired bound now follows by assembling the above estimates together. Note that in the last line we derived the inequality $\langle D\mathbf{x},\ \mathbf{x}\rangle \leq 4\langle \widetilde{M_{GS}}\mathbf{x},\ \mathbf{x}\rangle$, which has appeared previously in [33], Lemma 3.3 and in [30], Proposition 6.12. □

The scaled variant of ℓ_1 GS based on Remark 6.1 is given by $M_{\ell_1 GS} = M_{HGS} + \frac{1}{2} D^{\ell_1}$. A result similar to Theorem 6.2 also holds for this smoother, as shown below.

Theorem 6.3. *Without any restrictions, we have for the scaled $\ell_1 GS$ smoother*

$$
K(M_{\ell_1 GS}) \leq \left(1 + \frac{\sqrt{10}}{\theta} \right)^2 K(M_{GS}).
$$

Proof. The proof follows as above, except that we have that

$$
\|(M_{\ell_1 GS} - M_{GS})\mathbf{x}\|^2_{D^{-1}} = \sum_i \frac{1}{a_{ii}} \Bigg(\frac{1}{2} \sum_{j \in \Omega_o^{(i)}} |a_{ij}| x_i - \sum_{\substack{j \in \Omega_o^{(i)} \\ j<i}} a_{ij} x_j \Bigg)^2 .
$$

□

Another convergent option, which also takes advantage of the local estimation of θ in (5.3) is

$$
M_{\ell_1 GS*} = M_{HGS} + D^{\ell_1 *}, \qquad \text{where} \qquad d_{ii}^{\ell_1 *} = \begin{cases} 0, & \text{if } a_{ii} \geq \eta d_{ii}^{\ell_1}; \\ d_{ii}^{\ell_1}/2, & \text{otherwise.} \end{cases} \tag{6.3}
$$

and η is a fixed parameter satisfying $\eta > 1$. This smoother locally switches to ℓ_1 GS if hybrid GS is not appropriate, and we have found that the value $\eta = 1.5$ works well in practice. Some results with this smoother are shown in Section 7 of [4]. Note that the smoother $M_{\ell_1 GS*}$ is identical to M_{HGS} when $\theta \geq \eta$, and reduces to the scaled version of $M_{\ell_1 GS}$ based on Remark 6.1 when θ is uniformly small relative to η. Furthermore, the general conclusion of Theorem 6.2 still holds for $M_{\ell_1 GS*}$, since $\eta = 1.5$, for example, implies

$$\langle D\mathbf{x},\ \mathbf{x}\rangle \leq 3\langle (M_{\ell_1 GS*}^T + M_{\ell_1 GS*} - A)\mathbf{x},\ \mathbf{x}\rangle.$$

7. Concluding Remarks.

In this paper we reviewed and analyzed a number of practical parallel multigrid smoothers, and evaluated their potential for scalability on ultra-parallel computers with millions of processors. Based on the framework from [17, 18] we proposed both a direct (Theorem 2.2) and a comparative (Lemma 2.3) approach for smoother analysis. Using these approaches, we showed that C-F smoothing is good if F-relaxation is fast to converge (Theorem 3.1), that Chebyshev is the optimal polynomial smoother (Theorem 4.1), and that hybrid Gauss-Seidel exhibits multigrid smoothing properties which are independent of the number of processors in many practical applications, e.g. if the matrix is diagonally dominant and the problem size per processor is large enough (Theorem 5.2). For the more robust ℓ_1 smoothers described in Section 6, we were able to prove processor-independent equivalence with full Gauss-Seidel with minimal restrictions on the matrix (Theorem 6.2).

Numerical experiments for the most promising of the above parallel smoothers using both message-passing (with MPI) and threading (with OpenMP) can be found in [4].

Acknowledgments. We thank Panayot Vassilevski for his helpful advice.

REFERENCES

[1] M. Adams, M. Brezina, J. Hu, and R. Tuminaro, *Parallel multigrid smoothing: Polynomial versus Gauss-Seidel*, J. Comput. Phys., 188 (2003), pp. 593–610. 1, 2, 9, 10

[2] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, 1996. 5, 8, 9

[3] O. Axelsson and P. Vassilevski, *Algebraic multilevel preconditioning methods I*, Numer. Math., 56 (1989), pp. 157–177. 10

[4] A. H. Baker, R. D. Falgout, T. V. Kolev, and U. M. Yang, *Multigrid smoothers for ultra-parallel computing*, SIAM J. Sci. Comput., (2011). (to appear), LLNL-JRNL-435315. 2, 9, 13, 15, 18

[5] A. H. Baker, T. Gamblin, M. Schulz, and U. M. Yang, *Challenges of scaling algebraic multigrid across modern multicore architectures*, in Proceedings of the 25th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2011), 2011. To appear. Also available as LLNL Tech. Report LLNL-CONF-458074. 1

[6] R. E. Bank and C. C. Douglas, *Sharp estimates for multigrid rates of convergence with general smoothing and acceleration*, SIAM J. Numer. Anal., 22 (1985), pp. 617–633. 2, 10, 11

[7] F. A. Bornemann and P. Deuflhard, *The cascadic multigrid method for elliptic problems*, Numer. Math., 75 (1996), pp. 135–152. 10

[8] D. Braess, *The convergence rate of a multigrid method with Gauss–Seidel relaxation for the Poisson equation*, in Multigrid Methods, W. Hackbusch and U. Trottenberg, eds., vol. 960 of Lecture Notes in Mathematics, Berlin, 1982, Springer-Verlag, pp. 368–386. 6

[9] D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V cycle*, SIAM J. Numer. Anal., 20 (1983), pp. 967–975. 6

[10] J. H. Bramble, *Multigrid Methods*, vol. 294 of Pitman Research Notes in Mathematical Sciences, Longman Scientific & Technical, Essex, England, 1993. 6

[11] J. H. Bramble and X. Zhang, *The analysis of multigrid methods*, Handb. Numer. Anal., VII (2000), pp. 173–415. 6

[12] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390. 2, 4, 6

[13] J. J. Brannick and R. D. Falgout, *Compatible relaxation and coarsening in algebraic multigrid*, SIAM J. Sci. Comput., 32 (2010), pp. 1393–1416. LLNL-JRNL-417122. 7

[14] M. Brezina, C. Heberton, J. Mandel, and P. Vaněk, *An iterative method with convergence rate chosen a priori*, Tech. Report UCD CCM Report 140, Center for Computational Mathematics, University of Colorado at Denver, February 1999. 10

[15] E. Chow, R. Falgout, J. Hu, R. Tuminaro, and U. Yang, *A survey of parallelization techniques for multigrid solvers*, in Parallel Processing for Scientific Computing, M. Heroux, P. Raghavan, and H. Simon, eds., SIAM Series on Software, Environments, and Tools, SIAM, 2006, ch. 10. 1

[16] R. D. Falgout, *An introduction to algebraic multigrid*, Computing in Science and Engineering, 8 (2006), pp. 24–33. Special issue on Multigrid Computing. UCRL-JRNL-220851. 1

[17] R. D. Falgout and P. S. Vassilevski, *On generalizing the algebraic multigrid framework*, SIAM J. Numer. Anal., 42 (2004), pp. 1669–1693. UCRL-JC-150807. 1, 2, 3, 5, 7, 18

[18] R. D. Falgout, P. S. Vassilevski, and L. T. Zikatanov, *On two-grid convergence estimates*, Numer. Linear Algebra Appl., 12 (2005), pp. 471–494. UCRL-JRNL-203843. 1, 2, 4, 15, 18

[19] W. Hackbusch, *Multi-grid convergence theory*, in Multigrid Methods, W. Hackbusch and U. Trottenberg, eds., vol. 960 of Lecture Notes in Mathematics, Springer-Verlag, 1982, pp. 177–219. 2, 6

[20] V. Henson and U. Yang, *BoomerAMG: A parallel algebraic multigrid solver and preconditioner*, Appl. Numer. Math., 41 (2002), pp. 155–177. 1

[21] *hypre: High performance preconditioners*. http://www.llnl.gov/CASC/hypre/. 1

[22] T. Kolev and P. Vassilevski, *AMG by element agglomeration and constrained energy minimization interpolation*, Numer. Linear Algebra Appl., 13 (2006), pp. 771–788. UCRL-JC-219462. 6

[23] ———, *Parallel auxiliary space AMG for H(curl) problems*, J. Comput. Math., 27 (2009), pp. 604–623. Special issue on Adaptive and Multilevel Methods in Electromagnetics. UCRL-JRNL-237306. 1, 13, 16

[24] J. Kraus, V. Pillwein, and L. Zikatanov, *Algebraic multilevel iteration methods and the best approximation to $1/x$ in the uniform norm*, Tech. Report 2009-17, Johann Radon Institute for Computational and Applied Mathematics (RICAM), 2010. http://arxiv.org/abs/1002.1859v1. 10

[25] J. Mandel, S. F. McCormick, and J. W. Ruge, *An algebraic theory for multigrid methods for variational problems*, SIAM J. Numer. Anal., 25 (1988), pp. 91–110. 6

[26] S. F. McCormick, *Multigrid methods for variational problems: further results*, SIAM J. Numer. Anal., 21 (1984), pp. 255–263. 6

[27] ———, *Multigrid methods for variational problems: general theory for the V–cycle*, SIAM J. Numer. Anal., 22 (1985), pp. 634–643. 6

[28] J. W. Ruge and K. Stüben, *Algebraic multigrid (AMG)*, in Multigrid Methods, S. F. McCormick, ed., vol. 3 of Frontiers in Applied Mathematics, SIAM, Philadelphia, PA, 1987, pp. 73–130. 6

[29] K. Stüben and U. Trottenberg, *Multigrid methods: Fundamental algorithms, model problem analysis and applications*, in Multigrid Methods, W. Hackbusch and U. Trottenberg, eds., vol. 960 of Lecture Notes in Mathematics, Springer-Verlag, 1982, pp. 1–176. 2, 9

[30] P. S. Vassilevski, *Multilevel block factorization preconditioners: Matrix-based analysis and algorithms for solving finite element equations*, Springer, New York, 2008. 10, 17

[31] U. Yang, *Parallel algebraic multigrid methods - high performance preconditioners*, in Numerical Solution of Partial Differential Equations on Parallel Computers, A. Bruaset and A. Tveito, eds., vol. 51, Springer-Verlag, 2006, pp. 209–236. 1

[32] U. M. Yang, *On the use of relaxation parameters in hybrid smoothers*, Numer. Linear Algebra Appl., 11 (2004), pp. 155–172. UCRL-JC-151575. 1, 2, 15

[33] L. T. Zikatanov, *Two-sided bounds on the convergence rate of two-level methods*, Numer. Linear Algebra Appl., 15 (2008), pp. 439–454. 17