SIAM J. SCI. COMPUT.
Vol. 43, No. 5, pp. SM1-SM4

SUPPLEMENTARY MATERIALS: ALGEBRAIC ERROR ANALYSIS FOR MIXED-PRECISION MULTIGRID SOLVERS*

STEPHEN F. MCCORMICK†, JOSEPH BENZAKEN‡, AND RASMUS TAMSTORF‡

SM1. Double relaxation sweeps. The theory in the main paper is restricted to V(1,0)-cycles, meaning that each $\mathcal{V}$ uses one *pre-smoothing* sweep on the way down through the coarse grids and no *post-smoothing* sweeps on the way back up to the finest. This specific cycle simplifies the analysis because post-smoothing sweeps substantially complicate the estimates by the accumulation of errors in the transfer of residuals to the coarse levels and, to a lesser extent, because a single pre-smoothing sweep is easier to analyze than multiple sweeps. On the other hand, while the analysis of a general V-cycle would be too complicated to present here, we can more easily analyze multiple pre-smoothing sweeps as we illustrate now for a V(2,0)-cycle.

A relatively simple way to handle multiple sweeps is to combine them into one. To this end, for each $j \in \{1,2\}$, consider a monotonically energy-convergent stationary linear iteration $x \leftarrow x - M^{(j)}(Ax-b)$, where $M^{(j)} \in \mathbb{R}^{n\times n}$, and let α_j be a constant such that computing $M^{(j)}z$ for a vector $z \in \mathbb{R}^n$ in $\dot{\boldsymbol{\varepsilon}}$-precision yields the result $M^{(j)}z + \delta$, $\|\delta\| \le \alpha_j \dot{\boldsymbol{\varepsilon}}\|z\|$. Then the key point here is that the error propagation matrix for relaxation with preconditioner $M^{(1)}$ followed by relaxation with preconditioner $M^{(2)}$ can be written as $(I - M^{(2)}A)(I - M^{(1)}A) = I - M^{(2)}A$, where $M^{(2)} = M^{(1)} + M^{(2)} - M^{(2)}AM^{(1)}$. We can therefore think of, and implement, two relaxation sweeps as just the *one* sweep $y \leftarrow y - M^{(2)}(Ay - r)$, which means that we can analyze a V(2,0)-cycle as just a V(1,0)-cycle with this $M^{(2)}$. The implication is that we just need to provide estimates for $\|M^{(2)}\|$ and a constant $\alpha_{M^{(2)}}$ such that computing $M^{(2)}z$ for any vector z in $\dot{\boldsymbol{\varepsilon}}$-precision yields the result $M^{(2)}z + \delta_{M^{(2)}}$, $\|\delta_{M^{(2)}}\| \le \alpha_{M^{(2)}}\dot{\boldsymbol{\varepsilon}}\|z\|$. This is done in our next theorem.

THEOREM SM1.1. Double Sweeps. *For a double sweep in the ordering specified by* $M^{(2)}z = ((M^{(2)}z) + (M^{(1)}z) - (M^{(2)}(A(M^{(1)}z))))$, *the constant* $\alpha_{M^{(2)}}$ *can be chosen as follows:*

$$\begin{aligned}\alpha_{M^{(2)}} &= \|M^{(2)}\| + (1+\dot{\boldsymbol{\varepsilon}})[(\|M^{(2)}\| + \alpha_2\dot{\boldsymbol{\varepsilon}})(\|A\|\alpha_1 + \psi\dot{m}_A\alpha_1\dot{\boldsymbol{\varepsilon}} + \psi\dot{m}_A\|M^{(1)}\|)\\ &\quad + \|A\|\cdot\|M^{(1)}\|\alpha_2 + \|M^{(1)}\| + \|M^{(2)}\| + 2\alpha_1 + 2\alpha_2].\end{aligned} \tag{SM1.1}$$

Moreover, in general, $\|M^{(2)}\| \le \|M^{(1)}\| + \|M^{(2)}\| + \|A\|\cdot\|M^{(1)}\|\cdot\|M^{(2)}\|$. *The special case* $M^{(1)} = M^{(2)} = \frac{\omega}{\|A\|}I$, $0 < \omega < 2$ *yields the sharp estimate* $\|M^{(2)}\| \le \frac{2\omega}{\|A\|}$.

Proof. The following is meant to clarify the stages for computing $M^{(2)}z$, with the

*Received by the editors July 1, 2020; accepted for publication (in revised form) February 18, 2021; published electronically June 24, 2021.
†University of Colorado at Boulder, Boulder, CO (stephen.mccormick@colorado.edu).
‡Walt Disney Animation Studios, Burbank, CA (Joseph.Benzaken@disneyanimation.com, Rasmus.Tamstorf@disneyanimation.com).

subscripted w's standing for the indicated quantities computed in $\dot{\boldsymbol{\epsilon}}$-precision:

$$M^{(2)}z = (\underbrace{(M^{(2)}z)}_{w_4} + \underbrace{(M^{(1)}z)}_{w_1} - (M^{(2)}\,(A\,\underbrace{(M^{(1)}z)}_{w_1}))).$$

w_2 over $(A\,(M^{(1)}z))$; w_3 over $(M^{(2)}\,(A\,(M^{(1)}z)))$; w_5 over the whole expression.

Thus, by definition,

$$w_1 = M^{(1)}z + \delta_1, \quad \|\delta_1\| \le \alpha_1\dot{\boldsymbol{\epsilon}}\|z\|. \tag{SM1.2}$$

We then use (2.3) in the main paper to obtain the slight overestimate

$$w_2 = AM^{(1)}z + A\delta_1 + \delta_2, \quad \|\delta_2\| \le \psi\dot{m}_A\dot{\boldsymbol{\epsilon}}\|w_1\| = \psi\dot{m}_A\dot{\boldsymbol{\epsilon}}\|M^{(1)}z + \delta_1\|.$$

Letting $\delta_3 = A\delta_1 + \delta_2$, where $\|\delta_3\| \le \|A\| \cdot \|\delta_1\| + \psi\dot{m}_A\dot{\boldsymbol{\epsilon}}(\|\delta_1\| + \|M^{(1)}z\|)$, yields

$$w_2 = AM^{(1)}z + \delta_3, \quad \|\delta_3\| \le \Big((\|A\| + \psi\dot{m}_A\dot{\boldsymbol{\epsilon}})\alpha_1 + \psi\dot{m}_A\|M^{(1)}\|\Big)\,\dot{\boldsymbol{\epsilon}}\|z\|.$$

Similarly,

$$w_3 = M^{(2)}AM^{(1)}z + M^{(2)}\delta_3 + \delta_4, \quad \|\delta_4\| \le \alpha_2\dot{\boldsymbol{\epsilon}}\|w_2\| = \alpha_2\dot{\boldsymbol{\epsilon}}\|AM^{(1)}z + \delta_3\|.$$

Letting $\delta_5 = M^{(2)}\delta_3 + \delta_4$, where $\|\delta_5\| \le \|M^{(2)}\| \cdot \|\delta_3\| + \alpha_2\dot{\boldsymbol{\epsilon}}(\|AM^{(1)}z\| + \|\delta_3\|)$, yields $w_3 = M^{(2)}AM^{(1)}z + \delta_5$, $\|\delta_5\| \le \Upsilon\dot{\boldsymbol{\epsilon}}\|z\|$, where

$$\Upsilon = (\|M^{(2)}\|+\alpha_2\dot{\boldsymbol{\epsilon}})(\|A\|\alpha_1+\psi\dot{m}_A\alpha_1\dot{\boldsymbol{\epsilon}}+\psi\dot{m}_A\|M^{(1)}\|)+\|A\|\cdot\|M^{(1)}\|\alpha_2\dot{\boldsymbol{\epsilon}}\|z\|. \tag{SM1.3}$$

We can now use the estimate $w_4 = M^{(2)}z + \delta_6$, $\|\delta_6\| \le \alpha_2\dot{\boldsymbol{\epsilon}}\|z\|$ together with (SM1.2) and (SM1.3) to obtain

$$\begin{aligned}
w_5 &= (w_1 + w_4 - w_3) + \delta_7 + \delta_8 = M^{(2)}z + \delta_1 - \delta_5 + \delta_6 + \delta_7 + \delta_8,\\
\|\delta_7\| &\le \dot{\boldsymbol{\epsilon}}\|w_1 + w_4\| \le (\|M^{(1)}\| + \|M^{(2)}\| + \alpha_1 + \alpha_2)\dot{\boldsymbol{\epsilon}}\|z\|,\\
\|\delta_8\| &\le \dot{\boldsymbol{\epsilon}}\|w_1 + w_4 - w_3 + \delta_7\| \le \dot{\boldsymbol{\epsilon}}(\|M^{(2)}\|\|z\| + \|\delta_1\| + \|\delta_5\| + \|\delta_6\| + \|\delta_7\|).
\end{aligned}$$

Letting $\delta = \delta_1 - \delta_5 + \delta_6 + \delta_7 + \delta_8$, then $w_5 = M^{(2)}z + \delta$, where

$$\begin{aligned}
\|\delta\| &\le \|\delta_1\| + \|\delta_5\| + \|\delta_6\| + \|\delta_7\| + \|\delta_8\|\\
&\le \Big(\dot{\boldsymbol{\epsilon}}\|M^{(2)}\| + (1+\dot{\boldsymbol{\epsilon}})(\|\delta_1\| + \|\delta_5\| + \|\delta_6\| + \|\delta_7\|\Big)\,\|z\|\\
&\le \Big(\|M^{(2)}\| + (1+\dot{\boldsymbol{\epsilon}})\,[\alpha_1 + \alpha_2 + \Upsilon + \alpha_1 + \alpha_2]\Big)\,\dot{\boldsymbol{\epsilon}}\|z\|,
\end{aligned}$$

thus establishing (SM1.1). The estimates for $\|M^{(2)}\|$ are straightforward. ∎

SM2. Second-order Chebyshev iteration. The equation in (SM1.1) can be used in a recursive way to analyze any Krylov method, where the error propagation matrix is a polynomial in A. For example, it is fairly straightforward to show that $\alpha_{M^{(2)}} = \mathcal{O}(\frac{\dot{m}_A}{\|A\|})$ for the $\mathcal{K}^{\text{th}}$-order Chebyshev relaxation (cf., [SM1]), although the constant in this order bound depends exponentially on $\mathcal{K}$. On the other hand, a more

direct approach can achieve a somewhat tighter bound, as illustrated for the case $\mathcal{K} = 2$ in our next theorem.

Second-order Chebyshev relaxation can be formed from two sweeps of Richardson iteration with error propagation factors of the form $I - s_j A$, $j = 1, 2$. Assume that the coefficients in the Chebyshev factors are chosen to so that $0 < s_j = \mathcal{O}(\frac{1}{\|A\|})$, $j = 1, 2$. This assumption would generally hold in the multigrid context when the smoothing interval is chosen to be a fixed percentage of the upper spectrum of A. We can thus write Chebyshev iteration in the form $y \leftarrow y - M_C(Ay - r)$, where $M_C = \omega_1 I - \omega_2 A$, $\omega_1 = \mathcal{O}(\frac{1}{\|A\|}) > 0$, and $\omega_2 = \mathcal{O}(\frac{1}{\|A\|^2}) > 0$. Define $\dot{m}_A = \frac{m_A}{1 - m_A \dot{\boldsymbol{\epsilon}}}$, where m_A is the maximum number of nonzeros in the rows of A. Finally, suppose that computing $M_C z$ in $\dot{\boldsymbol{\epsilon}}$-precision for any vector z yields $M_C z + \delta_{M_C}$, $\|\delta_{M_C}\| \leq \alpha_{M_C} \dot{\boldsymbol{\epsilon}} \|z\|$, for some constant α_{M_C}.

THEOREM SM2.1. Chebyshev. *For one second-order Chebyshev iteration in the ordering specified by $M_C z = (\omega_1 z) - (\omega_2 (Az))$, we can choose*

$$\alpha_{M_C} = \|M_C\| + (\omega_1 + (1 + \dot{\boldsymbol{\epsilon}})\omega_2 \psi \dot{m}_A + \omega_2 \|A\|)(1 + \dot{\boldsymbol{\epsilon}}).$$

Note that $\|M_C\| = \mathcal{O}(\frac{1}{\|A\|})$ and, if $\psi \approx \mathcal{O}(\|A\|)$, then $\alpha_{M_C} = \mathcal{O}\left(\frac{\dot{m}_A}{\|A\|}\right)$.

Proof. Computing $M_C z$ according to $M_C z = (\underbrace{(\underbrace{\omega_1 z}_{w_1}) - (\underbrace{\omega_2 (\underbrace{Az}_{w_2})}_{w_3})}_{w_4})$, we have

$$\begin{aligned}
w_1 &= \omega_1 z + \delta_1, \quad \|\delta_1\| \leq \omega_1 \dot{\boldsymbol{\epsilon}} \|z\|, \\
w_2 &= Az + \delta_2, \quad \|\delta_2\| \leq \psi \dot{m}_A \dot{\boldsymbol{\epsilon}} \|z\|, \\
w_3 &= -\omega_2 w_2 + \delta_3, \quad \|\delta_3\| \leq \omega_2 \dot{\boldsymbol{\epsilon}} \|Az + \delta_2\| \leq \omega_2 \dot{\boldsymbol{\epsilon}} (\|A\| + \psi \dot{m}_A \dot{\boldsymbol{\epsilon}}) \|z\|, \\
w_4 &= w_1 + w_3 + \delta_4, \quad \|\delta_4\| \leq \dot{\boldsymbol{\epsilon}} \|w_1 + w_3\| = \dot{\boldsymbol{\epsilon}} \|M_C z + \delta_1 + \omega_2 \delta_2 + \delta_3\|.
\end{aligned}$$

This implies that $w_4 = M_C z + \delta_C$, $\delta_C = \delta_1 - \omega_2 \delta_2 + \delta_3 + \delta_4$. The theorem now follows from noting that

$$\begin{aligned}
\|\delta_C\| &\leq \|\delta_1 - \omega_2 \delta_2 + \delta_3\| + \dot{\boldsymbol{\epsilon}} \|M_C z + \delta_1 + \omega_2 \delta_2 + \delta_3\| \\
&\leq \|M_C\| \|z\| \dot{\boldsymbol{\epsilon}} + (\|\delta_1\| + \|\omega_2 \delta_2\| + \|\delta_3\|)(1 + \dot{\boldsymbol{\epsilon}}).
\end{aligned}$$

Remark SM2.2. Chebyshev iteration based on A preconditioned by its diagonal D can be formed from two sweeps of damped Jacobi with error propagation operators $I - s_j D^{-1} A = I - M^{(j)} A$, where $M^{(j)} = M^{(j)}_{ii} = \mathcal{O}\left(\frac{\kappa(D)}{\|A\|}\right)$, $j = 1, 2$. To estimate α_{M_C} for this case, we can mimic the proof of Theorem SM2.1, but with the understanding now that the ω_j are matrices: $\omega_1 = M^{(2)} + M^{(1)}$ and $\omega_2 = M^{(2)} A M^{(1)}$. If the diagonal matrices $M^{(j)}$ have been formed accurately beforehand, perhaps in the setup phase at higher precision, then the line of reasoning is much the same as the above proof with two extra steps to account for the increased complexity of M_C. The resulting estimate is of the same order as that in Theorem SM2.1 with the exception that a power of $\kappa(D)$ appears in the implied constant. When the diagonal entries of A are widely varying, it may be more effective to construct $D^{-1} A$ in higher precision before it is used in V-cycles in order to avoid an explicit dependence on the condition number of D appearing in the estimate for α_{M_C}.

REFERENCES

[SM1] M. Adams, M. Brezina, J. Hu, and R. Tuminaro, *Parallel Multigrid Smoothing: Polynomial Versus Gauss–Seidel*, Journal of Computational Physics, 188 (2003), pp. 593–610, https://doi.org/10.1016/S0021-9991(03)00194-3.