

MULTIGRID PRIMER: THEORY, PRINCIPLES, AND ALGEBRAIC MULTIGRID

STEPHEN F. MCCORMICK* AND RASMUS TAMSTORF†

Abstract. The Multigrid Primer: Basic Principles [8] gives a relatively short exposition of basic multigrid methodology that is simplified by posing the problem in a *variational* setting: a quadratic *energy* minimization formulation of a symmetric positive definite matrix equation. The present paper extends this presentation to include basic theory for this variational form of multigrid together with practical principles and heuristics that can guide their development. The focus is on understanding the general principles that lead to multigrid methods that work rather than the details of specific algorithms. Such an understanding is crucial to the development of multigrid methods in more difficult settings than what is assumed in [8]. To appreciate the advantages that the basic principles offer, the presentation then turns to algebraic multigrid, a solver designed for matrix equations where very little if anything is known about their origin. The focus is on how the principles guide the development of an effective algebraic algorithm.

Key words. multigrid, full multigrid, algebraic multigrid, multigrid theory

MSC codes. 65-01, 65F10, 65F50, 65N22, 65N55

1. Introduction. The Multigrid Primer: Basic Principles [8] focusses on a discretized linear partial differential equation (PDE) as expressed by the following matrix equation:

$$\mathbf{L}^h \mathbf{u}^h = \mathbf{f}^h, \quad (1)$$

where $\mathbf{L}^h \in \mathbb{R}^{n \times n}$ is a symmetric positive definite (SPD) matrix, $\mathbf{f}^h \in \mathbb{R}^n$ is a given source vector, $\mathbf{u}^h \in \mathbb{R}^n$ is the unknown vector to be approximated, and h is the mesh width of the underlying grid. That Primer assumes a uniform rectangular grid in two or three dimensions and coarse grids in the multigrid algorithm formed simply by taking every other grid line in each coordinate direction. It further assumes that the intergrid transfers and coarse-grid correction equation are a direct and compatible result of the finite element Rayleigh-Ritz discretization of the PDE.

This basic setting facilitates a simplified description of multigrid and its principles, but it did not give much insight into how to develop an effective multigrid solver in more challenging situations. The PDE might have characteristics that make effective coarsening and intergrid transfers difficult to design, such as highly variable coefficients, grids or coefficients that leave the matrix with very small connections in a particular coordinate direction (i.e., anisotropy), grids that have been locally refined to create difficulties at the refinement interfaces, or grids that may be very irregular as depicted in Figure 1. It may even be that the matrix does not correspond to any grid or even any geometry at all. Such is the case with the matrices that come from page rank problems, where nodes are pages on the internet that are connected to other nodes via links on the page that a visitor can click, with connection weights influenced by the number of clicks each link receives.

To describe a multigrid approach that aims to deal with these and other challenging settings, section 2 first presents existing convergence theory for the algorithm in [8] to clarify the principles and properties of the solver that make it work. In this theory, we review the analysis of two coarse-level approximation properties that confirm

*University of Colorado, Boulder

†Independent Researcher

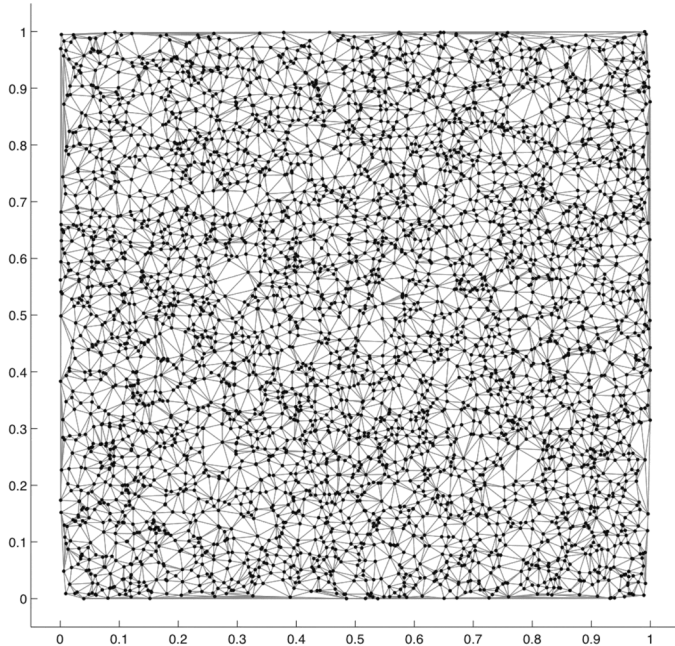


Fig. 1: *A very irregular 2D grid on which a PDE might be discretized.*

theoretical optimality of the multigrid solvers, including full multigrid (FMG). Then, in section 3, we set the scene for section 4 that focuses on applying the insights gained in the theory to algebraic multigrid (AMG) [3, 4]. AMG is a matrix equation solver that requires little if any knowledge of the origin of the discrete problem. This focus serves the purpose of showing how basic multigrid principles can be used to solve a wide variety of problems, and it provides a way to gain a deeper understanding of the nature of multigrid methods and their fundamental principles. We end in section 5 with a summary of our observations.

2. Theory. This section presents theory of convergence for multigrid cycles and of accuracy for FMG. The focus is on certain approximation properties of the methods, which are used later to facilitate development of multigrid for more challenging problems.

2.1. Preliminaries. While this Primer departs from the simpler setting considered in the basic Primer, we assume familiarity with the formulations and notations in [8] on which the development here is based. However, we recall some of its material for completeness, first by noting that the problem we begin to address is given by

$$\mathbf{L}^h \mathbf{v}^h = \mathbf{g}^h. \quad (2)$$

As in [8], while this is just a change in notation from (1), it is useful because it allows the possibility to think of $\mathbf{g}^h$ being the residual $\mathbf{L}^h \mathbf{u}^h - \mathbf{f}^h$, which would mean that we are seeking a *correction* to the current approximate solution $\mathbf{u}^h$.

One of the goals in this paper is to understand how an effective full rank interpolation matrix $\mathbf{P}_{2h}^h$ from grid $2h$ to grid h might be constructed. To this end, the initial

assumption is that the interpolation operator is already given so that we can explore its properties. We assume further that the restriction operator and coarse-grid matrix are defined respectively by the following so-called *variational conditions*:

$$\mathbf{R}_h^{2h} \equiv \alpha^h (\mathbf{P}_{2h}^h)^t \quad \text{and} \quad \mathbf{L}^{2h} \equiv \mathbf{R}_h^{2h} \mathbf{L}^h \mathbf{P}_{2h}^h. \quad (3)$$

The definition of restriction allows for any desired scaling $\alpha^h > 0$; the definition of the coarse-grid operator is called the *Galerkin condition*. As shown in [8], these choices guarantee optimality of the exact coarse-grid correction in the *energy norm*, which applied to a grid h vector $\mathbf{w}^h$ is given by

$$\|\mathbf{w}^h\|_{\mathbf{L}^h} = \langle \mathbf{w}^h, \mathbf{w}^h \rangle_{\mathbf{L}^h}^{\frac{1}{2}} = \langle \mathbf{L}^h \mathbf{w}^h, \mathbf{w}^h \rangle_{\frac{1}{2}},$$

where the angle brackets denote the Euclidean inner product with induced Euclidean norm $\|\cdot\|$. It is also important to remember that the error in an approximation to the solution of (2) is defined by $\mathbf{e}^h = \mathbf{v}^h - (\mathbf{L}^h)^{-1} \mathbf{g}^h$. Finally, we need to remember that the following *weak* and *strong* measures are very useful for measuring the algebraic smoothness of a given nonzero error $\mathbf{e}^h$:

$$\mathcal{M}_w(\mathbf{e}^h) = \frac{\langle \mathbf{L}^h \mathbf{e}^h, \mathbf{e}^h \rangle}{\|\mathbf{L}^h\| \langle \mathbf{e}^h, \mathbf{e}^h \rangle} \quad \text{and} \quad \mathcal{M}_s(\mathbf{e}^h) = \frac{\langle \mathbf{L}^h \mathbf{e}^h, \mathbf{L}^h \mathbf{e}^h \rangle}{\|\mathbf{L}^h\| \langle \mathbf{L}^h \mathbf{e}^h, \mathbf{e}^h \rangle}.$$

We call errors with a large measure in either definition *algebraically oscillatory* (or simply oscillatory), while errors with small measure are called *algebraically smooth* (or simply smooth).

We begin to develop convergence theory by defining the properties of the following operators that lay the theoretical groundwork for multigrid:

$$\mathbf{T}^h \equiv \mathbf{I}^h - \mathbf{S}^h \quad \text{and} \quad \mathbf{S}^h \equiv \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h. \quad (4)$$

$\mathbf{T}^h$ is the *error propagation* operator for an exact coarse-grid correction given by

$$\mathbf{v}^h \leftarrow \mathbf{v}^h - \mathbf{P}_{2h}^h (\mathbf{L}_{2h}^h)^{-1} \mathbf{P}_h^{2h} (\mathbf{L}^h \mathbf{v}^h - \mathbf{g}^h). \quad (5)$$

To prove this claim, first note that the error is $\mathbf{e}^h = \mathbf{v}^h - (\mathbf{L}^h)^{-1} \mathbf{g}^h$. Next, we use an important relationship between this error and the residual given by

$$\mathbf{L}^h \mathbf{e}^h = \mathbf{L}^h (\mathbf{v}^h - (\mathbf{L}^h)^{-1} \mathbf{g}^h) = \mathbf{L}^h \mathbf{v}^h - \mathbf{g}^h. \quad (6)$$

Using this relation to replace the residual in (5) and subtracting $(\mathbf{L}^h)^{-1} \mathbf{g}^h$ from both sides yields

$$\mathbf{e}^h \leftarrow \mathbf{e}^h - \mathbf{P}_{2h}^h (\mathbf{L}_{2h}^h)^{-1} \mathbf{P}_h^{2h} \mathbf{L}^h \mathbf{e}^h = \mathbf{T}^h \mathbf{e}^h,$$

which establishes the claim.

Another claim is that $\mathbf{S}^h$ and $\mathbf{T}^h$ are energy-orthogonal projections onto respectively the *range* of $\mathbf{P}_{2h}^h$, denoted $R(\mathbf{P}_{2h}^h)$, and the *null space* or *kernel* of $\mathbf{R}_h^{2h} \mathbf{L}^h$, denoted $N(\mathbf{R}_h^{2h} \mathbf{L}^h)$. To see this, first note that $\mathbf{S}^h$ maps onto $R(\mathbf{P}_{2h}^h)$: it is into $R(\mathbf{P}_{2h}^h)$ because its lead operator (i.e., what is applied last) is $\mathbf{P}_{2h}^h$ and it is onto because $(\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h$ is full rank. $\mathbf{S}^h$ is also a projection because

$$\begin{aligned} (\mathbf{S}^h)^2 &= \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h \\ &= \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{L}^{2h} (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h \\ &= \mathbf{S}^h. \end{aligned}$$

Similarly, $\mathbf{T}^h$ maps onto $N(\mathbf{R}_h^{2h} \mathbf{L}^h)$: it is into $N(\mathbf{R}_h^{2h} \mathbf{L}^h)$ because

$$\begin{aligned} \mathbf{R}_h^{2h} \mathbf{L}^h \mathbf{T}^h &= \mathbf{R}_h^{2h} \mathbf{L}^h (\mathbf{I}^h - \mathbf{S}^h) \\ &= \mathbf{R}_h^{2h} \mathbf{L}^h (\mathbf{I}^h - \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h) \\ &= \mathbf{R}_h^{2h} \mathbf{L}^h - \mathbf{L}^{2h} (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h \\ &= \mathbf{0}^{2h}, \end{aligned}$$

and it is onto because $\mathbf{T}^h + \mathbf{S}^h = \mathbf{I}^h$. $\mathbf{T}^h$ is also a projection because

$$(\mathbf{T}^h)^2 = (\mathbf{I}^h - \mathbf{S}^h)^2 = (\mathbf{I}^h)^2 - 2\mathbf{S}^h + (\mathbf{S}^h)^2 = \mathbf{I}^h - 2\mathbf{S}^h + \mathbf{S}^h = \mathbf{T}^h.$$

Finally, these projections are energy orthogonal to each other: for any $\mathbf{v}^h$ and $\mathbf{z}^h$, we have that

$$\begin{aligned} \langle \mathbf{T}^h \mathbf{v}^h, \mathbf{S}^h \mathbf{z}^h \rangle_{\mathbf{L}^h} &= \langle \mathbf{L}^h (\mathbf{I}^h - \mathbf{S}^h) \mathbf{v}^h, \mathbf{S}^h \mathbf{z}^h \rangle \\ &= \langle \mathbf{v}^h, \mathbf{L}^h \mathbf{S}^h \mathbf{z}^h \rangle - \langle \mathbf{v}^h, (\mathbf{S}^h)^t \mathbf{L}^h \mathbf{S}^h \mathbf{z}^h \rangle \\ &= \langle \mathbf{v}^h, \mathbf{L}^h \mathbf{S}^h \mathbf{z}^h \rangle - \langle \mathbf{v}^h, \mathbf{L}^h \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h \mathbf{z}^h \rangle \\ &= \langle \mathbf{v}^h, \mathbf{L}^h \mathbf{S}^h \mathbf{z}^h \rangle - \langle \mathbf{v}^h, \mathbf{L}^h \mathbf{P}_{2h}^h (\mathbf{L}^{2h})^{-1} \mathbf{R}_h^{2h} \mathbf{L}^h \mathbf{z}^h \rangle \\ &= \langle \mathbf{v}^h, \mathbf{L}^h \mathbf{S}^h \mathbf{z}^h \rangle - \langle \mathbf{v}^h, \mathbf{L}^h \mathbf{S}^h \mathbf{z}^h \rangle \\ &= 0, \end{aligned}$$

which proves the claim.

Keep in mind that the range of $\mathbf{P}_{2h}^h$ consists of vectors that can be exactly represented by interpolation from the coarse grid, while the kernel of $\mathbf{R}_h^{2h} \mathbf{L}^h$ consists of error vectors that cannot be represented at all by passing their residuals to the coarse grid. That is, $\mathbf{S}^h$ and $\mathbf{T}^h$ provide a decomposition into errors that *can* be eliminated by the coarse grid and those that *cannot* (and must therefore be dealt with on the fine grid).

Given these definitions, we have the following:

- Any grid h vector $\mathbf{e}^h$ can be expressed as $\mathbf{e}^h = \mathbf{s}^h + \mathbf{t}^h$, where $\mathbf{s}^h \in R(\mathbf{P}_{2h}^h)$, $\mathbf{t}^h \in N(\mathbf{R}_h^{2h} \mathbf{L}^h)$, and $\langle \mathbf{s}^h, \mathbf{t}^h \rangle_{\mathbf{L}^h} = 0$ (that is, $\mathbf{s}^h$ and $\mathbf{t}^h$ are orthogonal in the energy inner product).
- $\mathbf{T}^h \mathbf{s}^h = \mathbf{0}$, $\mathbf{S}^h \mathbf{t}^h = \mathbf{0}$, $\mathbf{T}^h \mathbf{t}^h = \mathbf{t}^h$, $\mathbf{S}^h \mathbf{s}^h = \mathbf{s}^h$, $\mathbf{T}^h \mathbf{e}^h = \mathbf{t}^h$, $\mathbf{S}^h \mathbf{e}^h = \mathbf{s}^h$, $\langle \mathbf{T}^h \mathbf{v}^h, \mathbf{S}^h \mathbf{z}^h \rangle_{\mathbf{L}^h} = 0 \forall \mathbf{v}^h, \mathbf{z}^h$, and $\|\mathbf{S}^h\|_{\mathbf{L}^h} = \|\mathbf{T}^h\|_{\mathbf{L}^h} = 1$.
- $\|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h} = \min_{\mathbf{v}^{2h}} \|\mathbf{e}^h - \mathbf{P}_{2h}^h \mathbf{v}^{2h}\|_{\mathbf{L}^h}$ (that is, coarse-grid correction minimizes the energy norm of the error over all possible corrections in the range of interpolation).

These assertions are used repeatedly in the following.

2.2. Theory for basic multigrid cycles. The convergence theory developed here begins with the two-grid scheme applied to (2) that applies coarse-grid correction first and then one Richardson sweep given by

$$\mathbf{v}^h \leftarrow \mathbf{v}^h - \frac{1}{\|\mathbf{L}^h\|} (\mathbf{L}^h \mathbf{v}^h - \mathbf{g}^h).$$

The error propagation matrix for relaxation is $\mathbf{G}^h = \mathbf{I}^h - \frac{1}{\|\mathbf{L}^h\|} \mathbf{L}^h$, so it follows that the error propagation matrix for the two-grid scheme is $\mathbf{G}^h \mathbf{T}^h$. Each cycle converges

with uniform energy bound for any initial guess if and only if

$$\|\mathbf{G}^h \mathbf{T}^h\|_{\mathbf{L}^h}^2 = \left\| \left(I - \frac{1}{\|\mathbf{L}^h\|} \mathbf{L}^h \right) \mathbf{T}^h \right\|_{\mathbf{L}^h}^2 \leq \gamma \quad (7)$$

for some fixed $\gamma \in [0, 1)$. This condition is the same as asking that the strong measure be bounded away from zero:

$$\mathcal{M}_s(\mathbf{t}^h) \geq \delta \quad (8)$$

for some fixed $\delta > 0$ and for all $\mathbf{t}^h$ in the range of $\mathbf{T}^h$. To see this, note first that

$$\begin{aligned} & \|\mathbf{G}^h \mathbf{T}^h\|_{\mathbf{L}^h}^2 \\ &= \max_{\mathbf{v}^h \neq \mathbf{0}^h} \frac{\langle \mathbf{L}^h (I - \frac{1}{\|\mathbf{L}^h\|} \mathbf{L}^h) \mathbf{T}^h \mathbf{v}^h, (I - \frac{1}{\|\mathbf{L}^h\|} \mathbf{L}^h) \mathbf{T}^h \mathbf{v}^h \rangle}{\langle \mathbf{L}^h \mathbf{v}^h, \mathbf{v}^h \rangle} \\ &= \max_{\mathbf{t}^h = \mathbf{T}^h \mathbf{t}^h \neq \mathbf{0}^h, \mathbf{s}^h = \mathbf{S}^h \mathbf{s}^h} \frac{\langle \mathbf{L}^h \mathbf{t}^h, \mathbf{t}^h \rangle - \frac{2}{\|\mathbf{L}^h\|} \langle \mathbf{L}^h \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle + \frac{1}{\|(\mathbf{L}^h)^2\|} \langle (\mathbf{L}^h)^2 \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle}{\langle \mathbf{L}^h \mathbf{t}^h, \mathbf{t}^h \rangle + \langle \mathbf{L}^h \mathbf{s}^h, \mathbf{s}^h \rangle} \\ &= \max_{\mathbf{t}^h = \mathbf{T}^h \mathbf{t}^h \neq \mathbf{0}^h} \frac{\langle \mathbf{L}^h \mathbf{t}^h, \mathbf{t}^h \rangle - \frac{1+\xi}{\|\mathbf{L}^h\|} \langle \mathbf{L}^h \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle}{\langle \mathbf{L}^h \mathbf{t}^h, \mathbf{t}^h \rangle} \\ &= 1 - (1 + \xi) \mathcal{M}_s(\mathbf{t}^h), \end{aligned}$$

where ξ is the value of $\frac{\frac{1}{\|\mathbf{L}^h\|} \langle (\mathbf{L}^h)^2 \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle}{\langle \mathbf{L}^h \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle}$ at the maximizing $\mathbf{t}^h$. Note further that

$$0 < \frac{\frac{1}{\|\mathbf{L}^h\|} \langle (\mathbf{L}^h)^2 \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle}{\langle \mathbf{L}^h \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle} \leq 1,$$

so $\xi \in [0, 1)$. As asserted, we thus have that (7) holds if and only if $1 - (1 + \xi) \mathcal{M}_s(\mathbf{t}^h) \leq \gamma$, that is, $\mathcal{M}_s(\mathbf{t}^h) \geq \delta \equiv \frac{1-\gamma}{1+\xi} > 0$. Note that $\delta \in (\frac{1-\gamma}{2}, 1 - \gamma]$.

To understand the two-grid condition (8) a little better, we make a (generally unrealistic) assumption that $\mathbf{t}^h$ is an eigenvector of $\mathbf{L}^h$ with corresponding eigenvalue λ . Then

$$\mathcal{M}_s(\mathbf{t}^h) = \frac{\langle \mathbf{L}^h \mathbf{t}^h, \mathbf{L}^h \mathbf{t}^h \rangle}{\|\mathbf{L}^h\| \langle \mathbf{L}^h \mathbf{t}^h, \mathbf{t}^h \rangle} = \frac{\lambda^2}{\lambda \|\mathbf{L}^h\|} = \frac{\lambda}{\|\mathbf{L}^h\|},$$

so the two-grid condition requires that λ be theoretically comparable (up to δ) to $\|\mathbf{L}^h\|$, the largest eigenvalue of $\mathbf{L}^h$. Of course, the practical quality of convergence depends on δ , and a value of δ near zero means that γ is near 1, so it signals poor expected numerical performance. Generally, we want error components $\mathbf{t}^h \in N(\mathbf{R}_h^{2h} \mathbf{L}^h)$, which $\mathbf{T}^h$ leaves untouched, to lie primarily in the upper spectrum of $\mathbf{L}^h$.

Unfortunately, counterexamples (such as standard cell-centered coarsening of a discrete Poisson equation; c.f., [2]) show that this two-grid condition does not suffice to establish uniform convergence of the multilevel V-cycle. However, a condition that is sufficient, as shown below in Theorem 1, is the following *strong approximation property*. See [11] for some conditions on the PDE and its discretization that produce matrices satisfying this property.

DEFINITION 2.1. *Strong Approximation Property (SAP)* A coarse-grid correction scheme satisfies the Strong Approximation Property if and only if

$$\min_{\mathbf{v}^{2h}} \|\mathbf{e}^h - \mathbf{P}_{2h}^h \mathbf{v}^{2h}\|_{\mathbf{L}^h}^2 \leq \frac{C}{\|\mathbf{L}^h\|} \langle \mathbf{L}^h \mathbf{e}^h, \mathbf{L}^h \mathbf{e}^h \rangle$$

for some fixed constant $C < \infty$ and for all grid h vectors $\mathbf{e}^h$.

The SAP can be written a little more compactly as

$$\|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2 \leq \frac{C}{\|\mathbf{L}^h\|} \langle \mathbf{L}^h \mathbf{e}^h, \mathbf{L}^h \mathbf{e}^h \rangle. \quad (9)$$

Yet another way to write this condition is in terms of the coarse-grid correction error propagation matrix $\mathbf{T}^h$ and the strong measure:

$$\frac{\|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2}{\|\mathbf{e}^h\|_{\mathbf{L}^h}^2} \leq C \mathcal{M}_s(\mathbf{e}^h).$$

Note that this property reduces to the two-grid condition (with $C = \frac{1}{\delta}$) when $\mathbf{e}^h$ is in the range of $\mathbf{T}^h$, that is, $\mathbf{e}^h = \mathbf{T}^h \mathbf{e}^h$. (Remember that vectors in the range of $\mathbf{T}^h$ are precisely those that the coarse-grid correction leaves untouched and therefore undiminished.)

THEOREM 1. (V-cycle convergence) *The strong approximation property is a sufficient condition for uniformly bounded convergence in the energy norm of the $V(0, 1)$ -cycle, with Richardson post-relaxation, applied to (2) in that its error propagation matrix $\mathbf{V}^h$ satisfies*

$$\|\mathbf{V}^h\|_{\mathbf{L}^h}^2 \leq 1 - \frac{1}{C}.$$

Proof. We spare the casual reader by relegating the proof to Appendix A. $\square$

The crux of the proof is in showing that the SAP implies that

$$\|\mathbf{G}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2 \leq \|\mathbf{P}_{2^h}^h \mathbf{v}^{2h}\|_{\mathbf{L}^h}^2 + \left(1 - \frac{1}{C}\right) \|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2. \quad (10)$$

This bound states that one relaxation sweep reduces the square of the energy of the error by $\frac{1}{C}$ times the square of the oscillatory part of the error. In this very general form of (10), the theorem applies to any relaxation scheme for which it holds (e.g., Richardson, Chebyshev, Jacobi, steepest descent, or conjugate gradients), thereby guaranteeing optimal per-cycle performance when used in a V-cycle.

The strong approximation property requires that vectors in the kernel of $\mathbf{L}^h$ be exactly eliminated by coarse-grid correction. This follows from noting that if $\mathbf{e}^h$ is in the kernel of $\mathbf{L}^h$, then the right-hand side of the SAP vanishes, and thus so too must the left-hand side. However, we are assuming that $\mathbf{L}^h$ is SPD, so its kernel consists only of $\mathbf{0}^h$ and this requirement is, therefore, trivially satisfied by setting $\mathbf{v}^{2h} = \mathbf{0}^{2h}$. On the other hand, the strong approximation property also clearly shows that any vector that is in the near-kernel (in the sense of yielding a relatively small residual, $\mathbf{L}^h \mathbf{e}^h$) must be nearly eliminated by coarse-grid correction. This again follows from noting that if $\mathbf{e}^h$ is in the near-kernel of $\mathbf{L}^h$, then the right-hand side of the SAP is small, and thus so too must the left-hand side be. *This observation provides the critical insight that the interpolation operator should adequately approximate near-kernel components.* The focus on the near-kernel is generally sufficient because algebraically oscillatory vectors automatically satisfy the SAP: by definition, an algebraically oscillatory error $\mathbf{e}^h$ satisfies $\mathcal{M}_s(\mathbf{e}^h) \geq \delta$ for some $\delta \gg 0$, so it follows that

$$\frac{\|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2}{\|\mathbf{e}^h\|_{\mathbf{L}^h}^2} \leq \frac{\|\mathbf{e}^h\|_{\mathbf{L}^h}^2}{\|\mathbf{e}^h\|_{\mathbf{L}^h}^2} \leq C \mathcal{M}_s(\mathbf{e}^h),$$

where $C = \frac{1}{\delta} \ll \infty$. Finally, since the SAP involves the choice of coarse-grid points and interpolation to the fine grid, it gives us a way to assess the quality of the coarse-grid correction process.

As we have just seen, the strong measure is related to the strong approximation property in the energy inner product. We have a similar relationship for the weak case in the Euclidean norm.

DEFINITION 2.2. *Weak Approximation Property (WAP)* *A coarse-grid correction scheme satisfies the Weak Approximation Property if and only if*

$$\min_{\mathbf{v}^{2h}} \|\mathbf{e}^h - P_{2h}^h \mathbf{v}^{2h}\|^2 \leq \frac{C}{\|\mathbf{L}^h\|} \langle \mathbf{L}^h \mathbf{e}^h, \mathbf{e}^h \rangle.$$

At this point, it should be no surprise that the WAP can be written in terms of the weak measure:

$$\frac{\|\mathbf{T}^h \mathbf{e}^h\|^2}{\|\mathbf{e}^h\|^2} \leq C \mathcal{M}_w(\mathbf{e}^h).$$

By arguments similar to those for the strong approximation property, the weak approximation property confirms uniform convergence of the two-grid form of multigrid in the Euclidean norm. It can also confirm convergence of a multigrid cycling scheme like the W-cycle that is more aggressive than the V-cycle, provided the productive effort spent on the coarse grids is commensurate with the size of C .

Some relationships between the SAP and the WAP, together with an additional approximation property, are developed in Appendix B. They can be useful for designing geometrically based multilevel algorithms, but are especially useful for AMG. Most important is that a tentative AMG method can attempt to achieve the WAP by focusing on the near-kernel, and then a special indirect interpolation used in reformulating P_{2h}^h can be used to aim for the SAP. Unfortunately, classical AMG typically fails to satisfy the SAP. See [7]. Somewhat surprising, though, is that it still achieves optimal V-cycle convergence over a fairly wide class of matrix equations. On the other hand, failure for the coarse grids to achieve adequate approximation of the fine-grid smooth error means that an FMG version of classical AMG usually fails to work efficiently. However, the study in [7] suggests versions of AMG that may overcome this limitation. In light of that and for completeness of the theory presented here, the next section provides an analysis of FMG based on properties that confirm its optimality.

2.3. FMG theory. With the theory of convergence of multigrid cycles as an iterative solver at hand, we now turn to the more direct FMG approach that uses these cycles as basic building blocks to achieve accuracy with bounds that compare to that of the discretization, and to do this at optimal cost. To explain this optimality in some detail, we need to delve into the continuum. To avoid too much abstraction, we settle with illustrating the general case by way of the following model problem given by the two-dimensional Poisson problem:

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{11}$$

where $\Delta = \nabla \cdot \nabla = \partial_{xx} + \partial_{yy}$, Ω is the unit square with boundary $\partial\Omega$, and f is a square-integrable function on Ω . It is tempting to define the energy functional for (11) by $\langle -\Delta v, v \rangle - 2\langle v, f \rangle$, with angle brackets here now denoting the $L^2(\Omega)$ inner product expressed by $\langle w, z \rangle = \int_{\Omega} wv \, dx \, dy$. However, simple functions in the discretization that are pieced together across the elements are generally not twice differentiable in

any sense and are therefore not admissible arguments for such a definition. Instead, this definition is *weakened* using the Gauss Divergence Theorem, given here by

$$\langle -\nabla \cdot \nabla u, u \rangle = \langle \nabla u, \nabla u \rangle. \quad (12)$$

This identity is valid because so-called *full regularity* of (11) and square integrability of f guarantee that u is sufficiently smooth, and because the boundary terms vanish by virtue of their conditions being homogeneous and Dirichlet. So (12) leads to the following definitions of the respective energy functional and energy norm for the weak form of (11):

$$E(u) \equiv \langle \nabla u, \nabla u \rangle - 2\langle u, f \rangle \text{ and } \|u\|_L \equiv \langle \nabla u, \nabla u \rangle^{\frac{1}{2}}.$$

We use subscript L here to suggest that it is derived from the strong form (11). We can thus take the formal view that L and $\|u\|_L$ correspond to weak forms of $-\Delta$ and $\langle Lu, u \rangle^{\frac{1}{2}}$, respectively.

The discretization of (11) is now a matter of choosing a finite-dimensional subspace of the admissible function space for E . To be specific, suppose that Ω is partitioned by a hierarchy of triangles on a uniform grid, starting with a coarsest single-point mesh size of $\frac{1}{2}$ and continuing with uniform refinement by a factor of 2 to the finest grid with a sufficiently small mesh size. For each h in the hierarchy, let $\mathcal{H}^h$ denote the space of the real-valued finite element *functions* u^h that are linear in each grid h element, continuous across the domain, and zero on the boundary. Then the discretization on each grid h is derived by minimizing $E(u^h)$ over $u^h \in \mathcal{H}^h$, leading to the matrix equation

$$\mathbf{L}^h \mathbf{u}^h = \mathbf{f}^h. \quad (13)$$

defined in terms of the vector $\mathbf{u}^h$ whose entries are the values of u^h at the nodes¹. This *Rayleigh-Ritz* derivation is straightforward though lengthy enough that we leave it to the interested reader to find it in any of the available finite element texts such as [10]. A consequence of this formulation is that the equations and induced intergrid transfers satisfy the variational conditions (3) on every level.

Since FMG approximates $\mathbf{u}^h$ by, say, $\mathbf{v}^h$ with the PDE in mind, we need to have a measure of the *total* error in $\mathbf{v}^h$ as it compares to the function u that minimizes E . Another advantage of this variational formulation is that, instead of the vector $\mathbf{v}^h$, we can compare its equivalent finite element function v^h to u as measured in the energy norm, namely, $\|v^h - u\|_L$. (By equivalent we mean that $\mathbf{v}^h$ and v^h have the same nodal values.) This finite element view simplifies the discussion because all of the approximations and the exact weak solution of the PDE are in the same space, meaning for one that we are able to avoid using intergrid transfer operators. We are also able to avoid worrying about relative errors because all of the errors we consider are measured in terms of the fixed solution u that is absorbed in some form in the constant C under the assumption that

$$\|u^h - u\|_L \leq Ch^q. \quad (14)$$

Here, u^h is the exact minimizer of the functional on level h , C is a fixed constant, and $q > 0$ is the order of the energy discretization error. For standard continuous

¹This raises a subtle point. Note, in general, that u^h does not satisfy (12) because the left side of the identity is usually undefined. But (12) is valid for u and the right side is well-defined for both u^h and u , so both are admissible functions for E , meaning that it is valid to seek u by minimizing $E(u^h)$ over $\mathcal{H}^h$.

piecewise linear finite element discretization of (11) on a uniform triangulation and sufficient smoothness of u , this estimate holds with $q = 1$.

The goal of FMG is to produce a grid h function that approximates the discrete solution to within the same accuracy bounds as that for the discrete solution approximation to the continuum solution, that is,

$$\|v^h - u^h\|_L \leq Ch^q. \quad (15)$$

To see what this requires of the multigrid cycles, suppose for induction purposes that (15) is analogously true for v^{2h} as an approximation to $u^{2h} \in \mathcal{H}^{2h} \subsetneq \mathcal{H}^h$:

$$\|v^{2h} - u^{2h}\|_L \leq C(2h)^q. \quad (16)$$

This holds trivially for the coarsest grid H : the problem there is the scalar equation $a_{1,1}^H u_1^H = f_1^H$, where, at the single point of grid H , $a_{1,1}^H$ is the single entry of the matrix corresponding to L and u_1^H and f_1^H are the respective single entries of the vectors corresponding to u and f ; and that equation is solved exactly by one sweep of Richardson's iteration with the particular scale $\frac{1}{a_{1,1}^H}$ that we have chosen.

Next, we need to appeal to an orthogonality property afforded by our variational setting. To this end, first note that a function w minimizes $E(w)$ over an admissible subspace $\mathcal{S}$ if and only if it minimizes the energy error $\|w - u\|_L$ over $\mathcal{S}$. (This follows directly from the same logic that we used for the discrete case to show that the energy functional and the square of the energy error differ by a constant.) So we can think of any subspace minimizer w of the energy functional also as a minimizer of the energy error $\|w - u\|_L$. But, for any such w , it must also be that $w - u$ is energy orthogonal to $\mathcal{S}$, that is, $\langle \nabla(w - u), \nabla z \rangle = 0$ for any $z \in \mathcal{S}$. (The closest energy approximation to $u \in \mathcal{S}$ is its energy-orthogonal projection onto that subspace.) To see this, let z be any nonzero function in $\mathcal{S}$ and s be any nonzero scalar in $\mathbb{R}$. Then adding sz to w yields

$$\|w + sz - u\|_L^2 = \|w - u\|_L^2 + 2s\langle \nabla(w - u), \nabla z \rangle + s^2\|z\|_L^2.$$

If we assume that $w - u$ is not energy orthogonal to $\mathcal{S}$, then we must be able to choose z so that $\langle \nabla(w - u), \nabla z \rangle \neq 0$. However, if we choose $s = -\|z\|_L^{-2} \langle \nabla(w - u), \nabla z \rangle$, then

$$\begin{aligned} \|w + sz - u\|_L^2 &= \|w - u\|_L^2 - 2\|z\|_L^{-2} \langle \nabla(w - u), \nabla z \rangle^2 + \|z\|_L^{-4} \langle \nabla(w - u), \nabla z \rangle^2 \|z\|_L^2 \\ &= \|w - u\|_L^2 - \|z\|_L^{-2} \langle \nabla(w - u), \nabla z \rangle^2 \\ &< \|w - u\|_L^2. \end{aligned}$$

This is a contradiction because w is assumed to be the minimizer over $\mathcal{S}$. So it must be that $\langle \nabla(w - u), \nabla z \rangle = 0$ for all $z \in \mathcal{S}$, as asserted.

Now, since u^h is the best energy approximation to u from $\mathcal{S} \equiv \mathcal{H}^h$, we can thus conclude that $u^h - u$ is energy orthogonal to $u^{2h} - u^h \in \mathcal{S}$, which in turn implies that

$$\|u^{2h} - u\|_L^2 = \|u^{2h} - u^h\|_L^2 + \|u^h - u\|_L^2 \quad (17)$$

and, hence, that

$$\|u^{2h} - u^h\|_L^2 \leq \|u^{2h} - u\|_L^2. \quad (18)$$

But (17) also implies that u^{2h} minimizes $\|u^{2h} - u^h\|_L^2$ over $\mathcal{H}^{2h}$. Thus, by an energy-orthogonal argument analogous to the above, together with (18), we thus have that

$$\|v^{2h} - u^h\|_L^2 = \|v^{2h} - u^{2h}\|_L^2 + \|u^{2h} - u^h\|_L^2 \leq \|v^{2h} - u^{2h}\|_L^2 + \|u^{2h} - u\|_L^2. \quad (19)$$

Finally, (19), (16), and (14) (with $2h$ replacing h) combine to yield

$$\|v^{2h} - u^h\|_L \leq \left((C(2h)^q)^2 + (C(2h)^q)^2 \right)^{\frac{1}{2}} = 2^{m+\frac{1}{2}} Ch^q. \quad (20)$$

The conclusion is that, for FMG to achieve accuracy in energy comparable to that of the discretization itself, we need only apply enough multigrid cycles on each level to reduce the error by a factor of $2^{-(m+\frac{1}{2})}$. For the model problem with $m = 1$, this requirement amounts to enough V-cycles to reduce the error by about a third. A $V(2,1)$ -cycle is in common use, most likely due to the recommendation by Achi Brandt in the early days of multigrid, and it more than delivers that reduction factor for the model problem.

Other solvers applied to discrete elliptic PDEs have been able to achieve optimal iteration cost in some cases, but it seems improbable that the true optimality that FMG has been able to achieve could be obtained in general without appealing to the relatively inexpensive coarser levels for reasonably accurate initial guesses. In any case, this optimality of FMG came as a welcome surprise in the early days of multigrid research, and it ensured that multigrid would become the method of choice for many applications.

3. Solver considerations. Before we switch gears to talk about AMG, we discuss some practical and theoretical principles that need to be kept in mind in the development of effective solvers.

3.1. Limitations of theory. We need to be clear that the theory does not really ensure that a particular multigrid method is very efficient in practice. This limitation is a common weakness of convergence estimates for numerical solvers because they typically involve indeterminant constants such as the C in the WAP and SAP. This lack of specificity means that we cannot really say exactly what the convergence factors or discretization errors should be. Moreover, because the theory deals only with upper bounds, we cannot even conclude that we should see a particular pattern in the errors as the iterations or refinements proceed. Also, while a lower bound with an unspecified constant might guarantee a particular order of the error in theory, it would not help in any way in practice.

3.2. Usefulness of theory. On the other hand, numerical behavior often follows the patterns that such convergence estimates predict. Moreover, the theory is useful because it suggests certain properties of the algorithm components that make for a good multigrid approach. Our specific aim is therefore to show how one might use the theory developed above to suggest how to choose effective coarse grids and interpolation operators.

To this end, section 4 focuses on the development of classical Algebraic Multigrid because it provides a natural setting for this discussion in that it deals directly with *algebraically smooth* components whose treatments are critical to achieving good approximation properties. AMG is *algebraic* because it can be constructed with little or no knowledge or use of any connection between the matrix equation and any geometric entity such as a grid. Our focus here is a departure from the *geometric* approach that we have so far had in mind. However, this focus not only introduces an algebraic method, but it also has the advantage that understanding the principles that arise in this setting gives a deeper sense of what makes multigrid work in general.

3.3. Algebraic terminology. To avoid potential confusion, we need to be clear about terminology. While algebraic methods do not necessarily involve geometric

information directly, the literature typically uses some geometric nomenclature in the development. In the same way, we use *smooth* for the error in an algebraic context, although it may not look so in whatever geometric setting that the matrix equation might be posed. We also use the standard convention of *nodes* or *points* to refer to the unknowns in the matrix equation, that is, the vector indices or degrees of freedom. We also use *neighborhood* to refer to nodes connected directly in the matrix. None of these concepts necessary hold in the underlying geometry if such even exists, but it makes sense nonetheless as we show below. Finally, we consider errors in the near-kernel of the given matrix to be the smooth errors, but this is because we are focussing on Krylov smoothers that have the property that they stall precisely on errors that are nearly in the matrix null space. A more general treatment would focus on the near-kernel of the correction component of relaxation, which may not be the same as the matrix near-kernel.

3.4. Roles of the WAP and SAP. We first need to point out that, while V-cycle convergence is usually the desired target in applications because of its superior efficiency, the nature of the energy norm that the SAP is based on makes it more difficult to use as a design tool. The problem stems from the need to determine interpolation *locally*, that is, in an algebraic neighborhood of interconnected nodes. Using the energy norm to measure how well interpolation approximates local errors is problematic because the matrix present in that norm generally reaches to nodes outside of the neighborhood. This lack of locality inhibits the ability to isolate the design of interpolation. The problem comes from the need to determine what is smooth, that is, what is locally in the near-kernel of the matrix. For this reason, it is more common in practice to develop interpolation schemes based on the WAP because estimates involving the Euclidean norm can be wholly restricted to individual neighborhoods. This locality provides the foundation for AMG as shown in the next section, but its nodal-based approach means that the approximation properties are more subtly addressed. This should become apparent as we establish what the near-kernel is like.

3.5. The need for local smoothness estimation. The practical importance of determining coarsening locally cannot be overemphasized. Just as finite elements realized its true potential when it transitioned from a global spectral approach to a local piecewise polynomial methodology, so too is multigrid most effective when coarse-grid corrections are constructed from local approximations of errors based on their local character. The dimension of the space of smooth errors that relaxation cannot effectively reduce is typically a significant fraction of the dimension of the fine-level space. This is generally much too large to permit an exact computation of a global basis for that space. This reasoning is analogous to saying that global spectral discretizations based on eigenvectors or Fourier modes may not be practical in general. So the approach taken by algebraic multigrid methods is to attempt to approximate algebraically smooth errors over small sets of nodes that are interconnected in the matrix by using a significantly smaller number of nodes. The intent is that these approximations can then be pieced together to provide adequate global approximation to all algebraically smooth errors.

4. AMG. We use the above principles now to motivate development of *classical* algebraic multigrid [3, 4, 9] applied to a matrix equation of the form

$$\mathbf{A}^h \mathbf{x}^h = \mathbf{b}^h, \tag{21}$$

where we include the superscript h to accommodate the coarsening descriptions that follow. (Note that we are turning now to the more common matrix/vector notation.) The primary aim of this overview is to emphasize how the principles suggested by the theory are used in an attempt to make AMG efficient for the problems at hand. The following indicates where these principles that underpin AMG are used.

On a historical note, it has been common in the literature to refer to the original algebraic multigrid algorithm as Ruge-Stüben AMG while referencing [9]. This accessible publication is indeed a very important resource for everything about the original method, including extensive numerical tests and a development of supporting theory. On the other hand, while these citations give the impression that the algorithm is new here, it is really just the classical one introduced in [3] and [4].

For a broader picture of AMG applicability, there are many options and variants as noted in the original publications, and many more that have since been developed as a quick internet search would easily uncover. Moreover, classical AMG and its variants have been described in some detail in many publications, from basic introductions to full algorithm descriptions to extensive numerical results. In that light, we are content to provide a summary of classical AMG's basic steps and refer the reader for details to Falgout's *An Introduction to AMG* [6] and to Chapter 8 of the *A Multigrid Tutorial* [5]. It may also be helpful to check out the available code PyAMG [1]. Finally, while the description here is for the original classical version AMG, the many new variants that have emerged and are continuing to be developed have improved AMG's efficiency and dramatically widened its applicability. Finally, the reader interested in a much more extensive and in-depth coverage of algebraic multigrid methods and relevant theory may wish to access [12].

Classical AMG has several components as outlined in the following subsections.

4.1. Set-up phase: relaxation. Geometric multigrid methods are very often blessed with a natural way to coarsen the fine grid by appealing to the discretization of the PDE that created the matrix equation. After all, it is frequently the case that the coarse grids can be constructed in analogy to the fine grid and intergrid transfers can take their cue from that construction. In such typical situations, coarsening is given and the task is to determine a good smoother. In contrast, it is more likely when forming an AMG solver that the smoother is chosen at the outset, leaving the essential design task of constructing coarsening in order to address what errors the smoother cannot effectively reduce. In that vein, we assume that a Krylov method such as Richardson is fixed as the smoother and we thus turn to coarsening that accurately approximates vectors in the near-kernel of $\mathbf{A}^h$.

4.2. Set-up phase: interpolation. It may seem premature to determine interpolation ahead of the coarse grid, but it is essential to understand what is needed in this transfer from coarse to fine grids for it to be effective. For the time being, we therefore assume a given partition of the grid into nodes in the set $\mathcal{C}$, which we use for the coarse grid, and the remaining $\mathcal{F}$ nodes. How this can be done effectively is described in the rest of this and the next subsections.

4.2.1. First smoothness principle. To determine interpolation relevant to this partition, we must understand the first algebraic smoothness principle that, by definition, states that the *correction* due to relaxation is small. For Richardson in particular, this implies that

$$(\mathbf{I}^h - \frac{1}{\|\mathbf{A}^h\|} \mathbf{A}^h) \mathbf{e}^h \approx \mathbf{e}^h \quad \implies \quad \frac{1}{\|\mathbf{A}^h\|} \mathbf{A}^h \mathbf{e}^h \approx \mathbf{0}^h. \quad (22)$$

4.2.2. Second smoothness principle. We must also understand the principle of *operator interpolation* since, with some critical modifications, it can be used as the prolongation matrix. The first smoothness principle suggests the use of the matrix entries to determine the weights of interpolation. Specifically, we want a formula that computes the value of a vector at an $\mathcal{F}$ node i given that we know the values of that vector at all $\mathcal{C}$ nodes j . That is just what interpolation is meant to do. We can now draw on (22) to say that

$$\sum_j a_{ij} e_j \approx 0 \quad \implies \quad a_{ii} e_i \approx \sum_{j \neq i} (-a_{ij}) e_j. \quad (23)$$

(We drop the superscript h for the entries of $\mathbf{A}^h$, and similarly for that of $\mathbf{e}^h$, because all entries we that we specify belong to grid h .) If all of the nodes with index $j \neq i$ with nonzero connections to node i are $\mathcal{C}$ nodes, then we could simply define interpolation to node i to have the weights $\frac{-a_{ij}}{a_{ii}}$. The problem is that it is not generally feasible to arrange the partitioning so that all of the neighbors of every $\mathcal{F}$ node are $\mathcal{C}$ nodes: if this were done consistently to create the hierarchy, it would typically mean very slow coarsening and high complexity with the attendant high cost. So we need to figure out how to eliminate all $\mathcal{F}$ nodes that are connected to i . More than that, we need to be sure that interpolation to i does not involve too many $\mathcal{C}$ nodes either since that too could lead to high complexity. We could simply zero out any connection to the offending $\mathcal{F}$ or $\mathcal{C}$ nodes, but that would risk potential violation of (23), on which smooth error approximation depends. We could instead zero out any connection to the offending nodes and adjust a_{ii} to correct the change, but smoothness gives us a better choice.

4.2.3. Third smoothness principle. This principle is, for many applications, that errors not well reduced by residual-based relaxation are approximately constant on average along *strong connections*, that is, if $-a_{ij}$ is not small compared to the other connections to nodes in the neighborhood of j , then the error at j must be approximately equal to the error at i on average. In that case, we say that j strongly depends on i . We are going to be ignoring positive off-diagonal entries of $\mathbf{A}^h$ with our scope restricted to matrices where their presence does not interfere with our development. In particular, we focus on the model Poisson case discretized by continuous piecewise linear finite elements because it has no positive off-diagonal entries. Many other matrices do, however, but for a fairly large class that are of elliptic type, their entries are all relatively too small to interfere with the coarsening developed below. That interference does tend to become an issue, however, as the approximation order of the finite elements increases, which can limit the usefulness of classical AMG without some modifications.

4.2.4. Third smoothness principle for the model problem. In any event, the task now is to show that this principle holds in the Poisson case. We first note that equation (7) in [8] together with the definition of the weak measure shows that relaxation stalls whenever

$$\langle \mathbf{A}^h \mathbf{e}^h, \mathbf{e}^h \rangle \ll \|\mathbf{A}^h\| \langle \mathbf{e}^h, \mathbf{e}^h \rangle.$$

This bound then confirms the following:

$$\begin{aligned}
& \frac{1}{2} \sum_{i,j} (-a_{ij})(e_i - e_j)^2 + \sum_{i,j} a_{ij} e_i^2 \\
&= \frac{1}{2} \left(- \sum_{i,j} a_{ij} e_i^2 - \sum_{i,j} a_{ij} e_j^2 + 2 \sum_{i,j} a_{ij} e_i e_j \right) + \sum_{i,j} a_{ij} e_i^2 \\
&= \sum_{i,j} a_{ij} e_i e_j = \langle \mathbf{A}^h \mathbf{e}^h, \mathbf{e}^h \rangle \ll \|\mathbf{A}^h\| \langle \mathbf{e}^h, \mathbf{e}^h \rangle = \|\mathbf{A}^h\| \sum_i e_i^2.
\end{aligned}$$

Reorganizing this bound yields

$$\frac{1}{2} \sum_{i,j} (-a_{ij})(e_i - e_j)^2 \ll \|\mathbf{A}^h\| \sum_i (1 - \sigma_i) e_i^2,$$

where $\sigma_i = \frac{1}{\|\mathbf{A}^h\|} \sum_j a_{ij}$. For the model Poisson problem that we have in mind, the σ_i are well below 1 because the matrix has row sums that are zero in the interior of the domain and still significantly less than $\frac{1}{\|\mathbf{A}^h\|}$ at the boundary. Note here our use of smoothness in that constants are in the near-kernel of $\mathbf{A}^h$. We have also used the WAP to establish the claim that the entries of smooth error are almost constant on average across strong connections. More generally, we can say that if $\mathbf{e}^h$ is smooth, then $|a_{ij}|e_j \approx |a_{ij}|e_i$ on average: if j strongly depends to i , then $e_j \approx e_i$; otherwise, $|a_{ij}|e_j \approx |a_{ij}|e_i$ must still be true on average just because $|a_{ij}|$ is small. This supports the third smoothness principle that we need in order to eliminate nonzeros in the matrix that impede operator interpolation.

4.2.5. Using the principles to partition the grid. To this end, we return to (23), but we partition the nodes now into the set $\mathcal{C}_i$ of some of the nodes to which j strongly depends (freedom to eliminate any strong connection that seems inconvenient is described in the next subsection), the set $\mathcal{D}_i^s$ of the remainder to which j strongly depends, and the set $\mathcal{D}_i^w$ corresponding to the rest of the nodes along *weak* connections (i.e., relatively small $|a_{ij}|$):

$$a_{ii} e_i \approx \sum_{j \in \mathcal{C}_i} (-a_{ij}) e_j + \sum_{j \in \mathcal{D}_i^s} (-a_{ij}) e_j + \sum_{j \in \mathcal{D}_i^w} (-a_{ij}) e_j.$$

The fact that $|a_{ij}|$ for $j \in \mathcal{D}_i$ is relatively large justifies replacing e_j by e_i for such j , which then leads to

$$\left(a_{ii} + \sum_{j \in \mathcal{C}_i} a_{ij} \right) e_i \approx \sum_{j \in \mathcal{C}_i^s} (-a_{ij}) e_j + \sum_{j \in \mathcal{D}_i^w} (-a_{ij}) e_j. \quad (24)$$

But we are still left with the unwanted nodes in $\mathcal{D}_i^w$. We could replace the corresponding e_j by e_i as we did for the $\mathcal{C}_i^s$ nodes. But experience has shown that a better way is to interpolate from the $\mathcal{C}_i^s$ nodes to each $\mathcal{D}_i^w$ node so that the right-hand side of (24) would be in terms only of $\mathcal{C}_i^s$ nodes. For this to be feasible, connections to $\mathcal{C}_i^s$ must be present in $\mathbf{A}^h$ for every $\mathcal{D}_i^w$ node. This is a requirement for coarsening that we adhere to in the next subsection. So we are free to determine this interpolation by

letting the weights to $j \in \mathcal{D}_i^w$ be their connecting coefficients $-a_{jk}$, $k \in \mathcal{D}_i^s$, divided by their sum, yielding the approximation

$$\mathbf{e}_j^h \approx \frac{\sum_{k \in \mathcal{C}_i} a_{jk} \mathbf{e}_k}{\sum_{k \in \mathcal{D}_i^s} a_{jk}}. \quad (25)$$

This expression is an interpolation of the nodes in $\mathcal{D}_i^s$ that is *indirect* in that interpolation to i does not involve the $j \in \mathcal{D}_i^s$ directly, but instead allows us to eliminate the $\mathcal{D}_i^s$ nodes from (24). Thus, after a bit of algebra and taking the approximation in (24) to mean equality, we obtain the following formula for the interpolation to $i \in \mathcal{F}$:

$$w_{ij} = - \frac{a_{ij} + \sum_{\ell \in \mathcal{D}_i^s} \left(\frac{a_{i\ell} a_{\ell j}}{\sum_{k \in \mathcal{C}_i} a_{\ell k}} \right)}{a_{ii} + \sum_{\ell \in \mathcal{D}_i^w} a_{i\ell}}. \quad (26)$$

This construction of interpolation rested heavily on knowledge of the near-kernel of A in that the near-kernel locally is well represented by constants. For other PDE systems such as linear elasticity that involve more than one kind of unknown, this development would require much more effort.

4.3. Set-up phase: coarse grid. With interpolation in hand based on a given coarse grid, our focus now is on how to construct that grid so that it effectively supports good approximation properties. More specifically, the principles associated with strong connections and smooth errors are to be used as guides in the construction of a coarse grid, with the aim that it supports approximating smooth fine-grid error accurately, reduces fine-grid smooth error effectively, and has substantially fewer nodes and lower matrix complexity than the fine grid has. Given the way that interpolation is to be determined, these aims lead to two heuristic criteria used in this construction:

- **H_1 :** Each node j that strongly depends on node $i \in \mathcal{F}$ should either be in $\mathcal{C}_i$ or strongly depend on at least one node in $\mathcal{C}_i$.
- **H_2 :** $\mathcal{C}$ should be a maximal subset of all the fine-grid nodes with the property that no $\mathcal{C}$ node depends strongly on any other $\mathcal{C}$ node.

Heuristic **H_1** stems from the use of the indirect interpolation formula in (25). The connections for each $j \in \mathcal{D}_i^s$ to the $\mathcal{C}_i$ node used in (25) should be strong enough to support the third smoothness principle, meaning that the error at node j is approximately equal to the errors at the other nodes in (25). Heuristic **H_2** aims to make processing on the coarse grid much less costly than that on the fine grid.

It is not always possible to achieve both heuristics at once, so instead, the construction process summarized next enforces **H_1** rigorously while attempting to address **H_2** as much as possible. Experience has shown that this objective tends to achieve the best trade-off between accuracy, which **H_1** addresses, and cost, which **H_2** addresses.

The actual coarse-grid construction process is fairly detailed. Instead of providing a lengthy description of it, we refer the reader to the details in *The Coloring Scheme* section of Chapter 8 in [5]. We settle here with a summary of the selection of $\mathcal{C}$ as follows:

- **Initialize:** *This step computes a measure that identifies how advantageous it would be for a node to be on the coarse grid because a large value for it would mean could support the most $\mathcal{F}$ nodes.* For each node i , compute the number λ_i of nodes that strongly depend on it.

- **Pass 1:** *This preliminary partition of the grid into $\mathcal{C}$ and $\mathcal{F}$ nodes attempts to ensure that $\mathcal{C}$ supports good approximation while enforcing $\mathbf{H}_2$.* This pass cycles through undecided nodes by making an i with maximal λ_i a $\mathcal{C}$ node, placing undecided nodes j dependent on i in $\mathcal{F}$, incrementing λ_k of the undecided nodes k that j depends strongly on to raise their value as potential $\mathcal{C}$ nodes, and decrementing λ_ℓ for every undecided node ℓ that i depends strongly on to reduce its $\mathcal{C}$ desirability. This cycle is repeated until each node has been assigned to $\mathcal{C}$ or $\mathcal{F}$. At this point, the partition is likely to have allowed good approximation to the near-kernel components while reining in the size of $\mathcal{C}$.
- **Pass 2:** *$\mathcal{F}$ nodes are changed to $\mathcal{C}$ nodes as necessary to enforce $\mathbf{H}_1$, while possibly violating $\mathbf{H}_2$ but minimizing the chances of $\mathcal{C}$ -to- $\mathcal{C}$ connections, which would tend to increase coarse-grid complexity.* The process is to cycle through each $\mathcal{F}$ node i that depends strongly on another $\mathcal{F}$ node j to make sure that they both strongly depend on a common $\mathcal{C}$ node. If they do not, then j is *tentatively* changed to be in $\mathcal{C}$. If there are any other $\mathcal{F}$ nodes that i depends strongly on where both do not strongly depend on a common $\mathcal{C}$ node, then move j back to $\mathcal{F}$ and i to $\mathcal{C}$, thus ensuring that i and j are not both $\mathcal{C}$ nodes. Continue this process until $\mathbf{H}_1$ is fully enforced.

Figure 2 displays the initial grid and four subsequent steps of the coloring algorithm. The connections shown are for the nine-point stencil based on continuous bilinear finite elements on a uniform grid applied to our model Poisson problem. This example is simplified somewhat because the decrement step in Pass 1 is never invoked, and Pass 2 never has to make any reassignments. Figure 8.4 in Chapter 8 of [5] includes all of the coloring steps for this example, while more complicated cases are described later in that chapter.

4.4. Set-up phase: restriction and coarse-grid matrix. With grid $2h$ and the interpolation operator P_{2h}^h from grid $2h$ to grid h now given, the restriction and coarse-grid matrices are defined automatically by the variational conditions in (3).

4.5. Solve phase: multigrid cycles. AMG typically uses a standard V-cycle based on the constructs as noted above. Assume that we are given a suitable relaxation scheme applied to (21) such as Richardson represented with μ sweeps by

$$\mathbf{x}^h \leftarrow \mathbf{G}^h(\mathbf{x}^h, \mathbf{b}^h, \mu, h),$$

where the $\mathbf{x}^h$ on the right is a current grid h approximation and $\mathbf{b}^h$ is a given source term. Then an AMG cycle with μ relaxation sweeps before and ν after coarsening, starting with initial guess $\mathbf{x}^h$, is given by

$$\mathbf{x}^h \leftarrow \text{AMG}^h(\mathbf{x}^h, \mathbf{b}^h, \mu, \nu, h),$$

and defined in Algorithm 1.

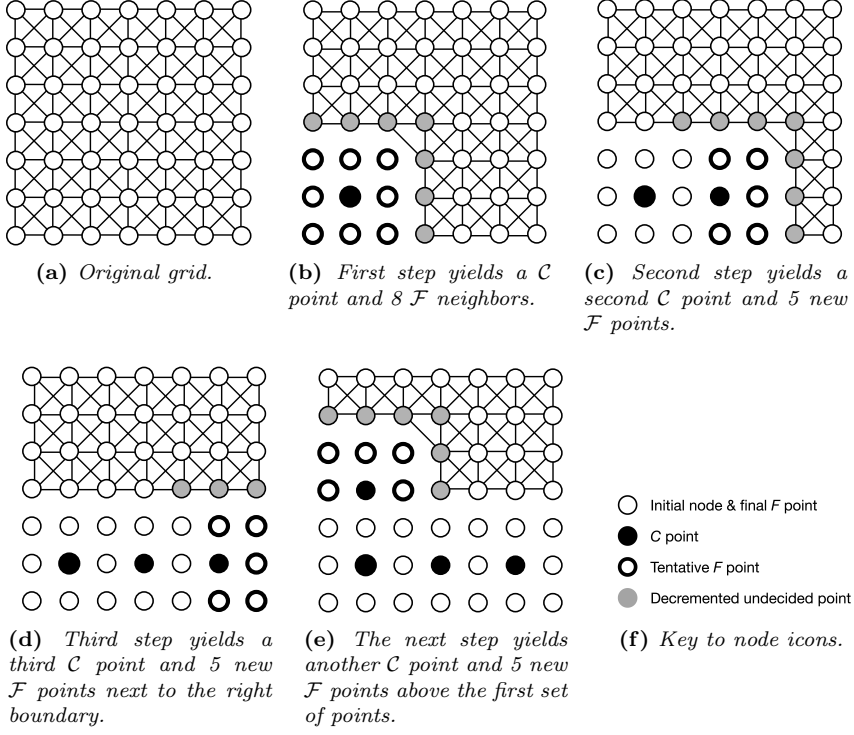


Fig. 2: Coloring example on an initial uniform grid of mesh size $h = \frac{1}{8}$ for the nine-point Laplacian, with only the interior nodes depicted. All the connecting lines have the same value, so the λ 's are initially equal to 8 for all nodes except those next to the boundary. The upper middle diagram results from the black dot becoming a C point and all of its neighbors becoming F points, with the next layer of neighbors receiving incremented λ 's. Because of the lexicographic ordering of the search, the coloring process continues in a similar way to the right up to the boundary (upper right to lower left diagrams), and then to the upper far left as depicted in the bottom middle diagram.

Algorithm 1 $\text{AMG}^h(x^h, b^h, \mu, \nu, h)$; V-cycle

Given: A^h, R_h^{2h}, P_{2h}^h

- | | |
|--|--|
| 1: $x^h \leftarrow G^h(x^h, b^h, \mu, h)$ | ▷ Relax μ times. |
| 2: if $h < H$ then | ▷ Keep coarsening. |
| 3: $b^{2h} = R_h^{2h}(A^h x^h - b^h)$ | ▷ Transfer grid h residual to grid $2h$. |
| 4: $x^{2h} \leftarrow \text{AMG}^{2h}(0^{2h}, b^{2h}, \mu, \nu, 2h)$ | ▷ Apply AMG^{2h} with zero initial guess. |
| 5: $x^h \leftarrow x^h - P_{2h}^h x^{2h}$ | ▷ Correct fine-grid iterate. |
| 6: end if | |
| 7: $x^h \leftarrow G^h(x^h, b^h, \nu, h)$ | ▷ Relax ν times. |
| 8: return x^h | |
-

5. Summary. With elliptic PDEs in mind, to achieve optimal multigrid performance, the following properties should be considered:

- The algebraically smooth errors, which relaxation by definition cannot quickly eliminate, should provide a pattern that can be effectively exploited in the construction of interpolation. For Krylov methods such as Richardson and Chebyshev, this means that interpolation should adequately approximate near-kernel components of the operator.
- There should be some sense of locality, such as that afforded by the WAP, since that is what makes it possible and practical to determine an adequate representation of the algebraically smooth errors. Moreover, coarsening should be constructed based on strongly interconnected points, and it should carefully avoid (near) redundancy. The aim is that strong connections would guarantee some sense of a smoothing property of relaxation and a lack of any serious redundancy would avoid artificial ill-conditioning.

Acknowledgement. The authors would like to thank John Ruge for his help in making sure that the description of classical AMG is correct.

REFERENCES

- [1] Nathan Bell, Luke N. Olson, Jacob Schroder, and Ben Southworth. PyAMG: Algebraic multigrid solvers in python. *Journal of Open Source Software*, 8(87):5495, 2023. URL: <https://pyamg.readthedocs.io/en/latest>, doi:10.21105/joss.05495.
- [2] James H Bramble, Richard E Ewing, Joseph E Pasciak, and Jian Shen. The analysis of multigrid algorithms for cell centered finite difference methods. *Advances in Computational Mathematics*, 5(1):15–29, 1996. doi:10.1007/BF02124733.
- [3] A. Brandt, S. McCormick, and J. Ruge. Algebraic multigrid (AMG) for automatic multigrid solution with application to geodetic computations. Technical report, Institute for Computational Studies, Colorado State University, POB 1852, Fort Collins, Colorado, 1982. Also available at: <http://amath.colorado.edu/~steven/pubs/fosls-mg/Multigrid.Publications.files/amg0.pdf>. URL: http://www.researchgate.net/publication/239066414_Algebraic_Multigrid_AMG_for_Automatic_Multigrid_Solution_with_Application_to_Geodetic_Computations.
- [4] A. Brandt, S. F. McCormick, and J. W. Ruge. Algebraic multigrid (AMG) for sparse matrix equations. In David John Evans, editor, *Sparsity and its Applications*, pages 257–284. Cambridge University Press, 1985. Also available at: <http://amath.colorado.edu/~steven/pubs/fosls-mg/Multigrid.Publications.files/amg1.pdf>. URL: http://www.researchgate.net/publication/243769570_Algebraic_multigrid_AMG_for_sparse_matrix_equations.
- [5] W. Briggs, V. Henson, and S. McCormick. *A Multigrid Tutorial, Second Edition*. Society for Industrial and Applied Mathematics, 2000. doi:10.1137/1.9780898719505.
- [6] Robert D. Falgout. An introduction to algebraic multigrid. *Comput. Sci. Eng.*, 8:24–33, 2006. URL: <https://api.semanticscholar.org/CorpusID:233207226>.
- [7] Robert D. Falgout, Wayne B. Mitchell, Thomas A. Manteuffel, and Stephen F. McCormick. Discretization-accuracy convergence for full algebraic multigrid. Technical Report LLNL-JRNL-696160, Lawrence Livermore National Laboratory, 2016. Student Competition Paper, Copper Mountain Conference on Multigrid Methods. URL: <https://www.osti.gov/servlets/purl/1840118>.
- [8] Stephen F. McCormick and Rasmus Tamstorf. Multigrid Primer: Basic Principles. Technical report, The University of Colorado at Boulder, CU-Boulder, 2026. URL: <https://arxiv.org/pdf/2605.17828>.
- [9] J. W. Ruge and K. Stüben. Algebraic Multigrid. In *Multigrid Methods*, chapter 4, pages 73–130. SIAM, 1987. doi:10.1137/1.9781611971057.ch4.
- [10] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [11] Panayot S. Vassilevski. Approximation properties of coarse spaces by algebraic multigrid. Technical Report LLNL-PROC-422574, Lawrence Livermore National Laboratory, 2010. URL: <http://www.osti.gov/servlets/purl/1114745/>.
- [12] Jinchao Xu and Ludmil Zikatanov. Algebraic multigrid methods. *Acta Numerica*, 26:591–721, 2017. URL: <https://arxiv.org/pdf/1611.01917>, doi:10.1017/S0962492917000083.

A. Proof of Theorem 1. Remembering that $e^h = S^h e^h + T^h e^h$ and that $S^h e^h$ is in the range of P_{2h}^h , then we can write $e^h = P_{2h}^h w^{2h} + T^h e^h$ for some grid $2h$ vector w^{2h} . With (9), we thus have that

$$\begin{aligned} \|G^h e^h\|_{L^h}^2 &= \|e^h\|_{L^h}^2 - \frac{2}{\|L^h\|} \|L^h e^h\|^2 + \frac{1}{\|L^h\|^2} \langle (L^h)^2 e^h, L^h e^h \rangle \\ &\leq \|S^h e^h\|_{L^h}^2 + \|T^h e^h\|_{L^h}^2 - \frac{1}{\|L^h\|} \|L^h e^h\|^2 \\ &\leq \|P_{2h}^h w^{2h}\|_{L^h}^2 + \left(1 - \frac{1}{C}\right) \|T^h e^h\|_{L^h}^2. \end{aligned} \quad (27)$$

To obtain a recursion formula for V^h in terms of its grid $2h$ counterpart V^{2h} , note by (6) that

$$(L^{2h})^{-1} R_h^{2h} (L^h v^h - g^h) = (L^{2h})^{-1} R_h^{2h} L^h e^h.$$

Letting e^{2h} denote the error in the zero initial guess on grid $2h$, we thus have the error expansion

$$0^{2h} = (L^{2h})^{-1} R_h^{2h} L^h e^h + e^{2h}.$$

Thus, since V^{2h} is the error propagator for the grid $2h$ V-cycle, then using the relation

$$(L^{2h})^{-1} R_h^{2h} L^h e^h = (L^{2h})^{-1} R_h^{2h} L^h P_{2h}^h w^{2h} = w^{2h}$$

shows that the new approximate grid $2h$ solution is

$$v^{2h} = (L^{2h})^{-1} R_h^{2h} L^h e^h + V^{2h} e^{2h} = w^{2h} - V^{2h} (L^{2h})^{-1} R_h^{2h} L^h e^h.$$

Finally, $P_{2h}^h v^{2h}$ is subtracted from v^h to obtain the new error, leading to

$$\begin{aligned} e^h &\leftarrow e^h - P_{2h}^h v^{2h} \\ &= P_{2h}^h w^{2h} + T^h e^h - P_{2h}^h \left(w^{2h} - V^{2h} (L^{2h})^{-1} R_h^{2h} L^h e^h \right) \\ &= \left(P_{2h}^h V^{2h} (L^{2h})^{-1} R_h^{2h} L^h + T^h \right) e^h. \end{aligned}$$

This somewhat intricate argument thus shows that the error propagation matrix for coarse-grid correction is $P_{2h}^h V^{2h} (L^{2h})^{-1} R_h^{2h} L^h + T^h$, which in turn means that the error propagation matrices of the fine and coarse V-cycles are related according to the recursion²

$$V^h = G^h \left(\left(P_{2h}^h V^{2h} (L^{2h})^{-1} R_h^{2h} L^h \right) + T^h \right). \quad (28)$$

Note that

$$\begin{aligned} &\left(\left(P_{2h}^h V^{2h} (L^{2h})^{-1} R_h^{2h} L^h \right) + T^h \right) e^h \\ &= \left(\left(P_{2h}^h V^{2h} (L^{2h})^{-1} R_h^{2h} L^h \right) + T^h \right) (P_{2h}^h w^{2h} + T^h e^h) \\ &= P_{2h}^h V^{2h} (L^{2h})^{-1} R_h^{2h} L^h P_{2h}^h w^{2h} + T^h e^h \\ &= P_{2h}^h V^{2h} w^{2h} + T^h e^h. \end{aligned} \quad (29)$$

²It is useful to consider the two extreme cases of V^{2h} here by reflecting on the form of (28). If the V-cycle on grid $2h$ happens to be an exact solver, then V^{2h} would be the zero matrix, so the grid h V-cycle becomes an optimally convergent two-grid cycle. At the other extreme, if the grid $2h$ V-cycle makes no change to the grid $2h$ approximation (i.e., 0^{2h}), then V^{2h} would be the grid $2h$ identity and the grid h V-cycle would devolve effectively into fine-grid relaxation only. It is the intermediate case that can be expected (and is assumed in (30)), where the grid $2h$ V-cycle reduces a significant fraction of grid $2h$ error.

If we substitute into (27) the new error $\mathbf{P}_{2h}^h \mathbf{V}^{2h} \mathbf{w}^{2h} + \mathbf{T}^h \mathbf{e}^h$ and its smooth component $\mathbf{P}_{2h}^h \mathbf{V}^{2h} \mathbf{w}^{2h}$ in place of $\mathbf{e}^h$ and $\mathbf{P}_{2h}^h \mathbf{w}^{2h}$, respectively, then (27)-(29) together with the induction hypothesis

$$\|\mathbf{V}^{2h}\|_{\mathbf{L}^{2h}}^2 \leq 1 - \frac{1}{C} \quad (30)$$

shows that

$$\begin{aligned} \|\mathbf{V}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2 &= \|\mathbf{G}^h (\mathbf{P}_{2h}^h \mathbf{V}^{2h} \mathbf{w}^{2h} + \mathbf{T}^h \mathbf{e}^h)\|_{\mathbf{L}^h}^2 \\ &\leq \|\mathbf{P}_{2h}^h \mathbf{V}^{2h} \mathbf{w}^{2h}\|_{\mathbf{L}^h}^2 + \left(1 - \frac{1}{C}\right) \|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2 \\ &= \langle \mathbf{L}^h \mathbf{P}_{2h}^h \mathbf{V}^{2h} \mathbf{w}^{2h}, \mathbf{P}_{2h}^h \mathbf{V}^{2h} \mathbf{w}^{2h} \rangle + \left(1 - \frac{1}{C}\right) \|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2 \\ &= \langle \mathbf{R}_h^{2h} \mathbf{L}^h \mathbf{P}_{2h}^h \mathbf{V}^{2h} \mathbf{w}^{2h}, \mathbf{V}^{2h} \mathbf{w}^{2h} \rangle + \left(1 - \frac{1}{C}\right) \|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2 \\ &= \|\mathbf{V}^{2h} \mathbf{w}^{2h}\|_{\mathbf{L}^{2h}}^2 + \left(1 - \frac{1}{C}\right) \|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2 \\ &\leq \left(1 - \frac{1}{C}\right) (\|\mathbf{w}^{2h}\|_{\mathbf{L}^{2h}}^2 + \|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2) \\ &= \left(1 - \frac{1}{C}\right) (\|\mathbf{P}_{2h}^h \mathbf{w}^{2h}\|_{\mathbf{L}^h}^2 + \|\mathbf{T}^h \mathbf{e}^h\|_{\mathbf{L}^h}^2) \\ &= \left(1 - \frac{1}{C}\right) \|\mathbf{e}^h\|_{\mathbf{L}^h}^2. \end{aligned}$$

This proves our assertion that $\|\mathbf{V}^h\|_{\mathbf{L}^h}^2 \leq 1 - \frac{1}{C}$.

B. Approximation properties. Following is a list of three approximation properties that an interpolation operator $\mathbf{P} \in \mathbb{R}^{n \times n_c}$ might possess in the context of approximating the solution of an equation in the SPD matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$. (For simplicity here, n and n_c denote the respective fine- and coarse-grid mesh sizes, and the matrices and vectors are used without superscripts.) Each property assumes that for every fine-grid vector $\mathbf{v}$,

WAP $_{\mathbf{L}^2}$: there exists a coarse-grid vector $\mathbf{w}$ such that $\|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 \leq \frac{C}{\|\mathbf{L}\|^2} \|\mathbf{L}\mathbf{v}\|^2$.

WAP $_{\mathbf{L}}$: there exists a coarse-grid vector $\mathbf{w}$ such that $\|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 \leq \frac{C}{\|\mathbf{L}\|} \|\mathbf{v}\|_{\mathbf{L}}^2$.

SAP $_{\mathbf{L}}$: there exists a coarse-grid vector $\mathbf{w}$ such that $\|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}}^2 \leq \frac{C}{\|\mathbf{L}\|} \|\mathbf{L}\mathbf{v}\|^2$.

The following results show that the WAP for $\mathbf{L}^2$ (WAP $_{\mathbf{L}^2}$) implies the WAP for $\mathbf{L}$ (WAP $_{\mathbf{L}}$) and is *equivalent* to the SAP for $\mathbf{L}$ (SAP $_{\mathbf{L}}$). We first show for completeness of proof that SAP $_{\mathbf{L}}$ implies WAP $_{\mathbf{L}}$. These properties are assumed to hold for any fine-grid vector $\mathbf{v}$ with C a fixed constant that may vary in meaning with each occurrence.

- SAP $_{\mathbf{L}} \implies$ WAP $_{\mathbf{L}}$: Let $\mathbf{P}\mathbf{w}$ and $\mathbf{P}\mathbf{z}$ be the best approximation to $\mathbf{v}$ and $\mathbf{L}^{-1}(\mathbf{v} - \mathbf{P}\mathbf{w})$, respectively, from the range of $\mathbf{P}$ in energy. Then, because

$\mathbf{v} - \mathbf{P}\mathbf{w}$ is energy orthogonal to the range of $\mathbf{P}$, we have that

$$\begin{aligned}\|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 &= \langle \mathbf{v} - \mathbf{P}\mathbf{w}, \mathbf{L}^{-1}(\mathbf{v} - \mathbf{P}\mathbf{w}) \rangle_{\mathbf{L}} \\ &= \langle \mathbf{v} - \mathbf{P}\mathbf{w}, \mathbf{L}^{-1}(\mathbf{v} - \mathbf{P}\mathbf{w}) - \mathbf{P}\mathbf{z} \rangle_{\mathbf{L}} \\ &\leq \|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}} \cdot \|\mathbf{L}^{-1}(\mathbf{v} - \mathbf{P}\mathbf{w}) - \mathbf{P}\mathbf{z}\|_{\mathbf{L}}.\end{aligned}$$

The $\text{SAP}_{\mathbf{L}}$ for $\mathbf{L}^{-1}(\mathbf{v} - \mathbf{P}\mathbf{w})$ then implies that

$$\begin{aligned}\|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 &\leq \|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}} \cdot \sqrt{\frac{C}{\|\mathbf{L}\|}} \|\mathbf{L}\mathbf{L}^{-1}(\mathbf{v} - \mathbf{P}\mathbf{w})\| \\ &= \|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}} \cdot \sqrt{\frac{C}{\|\mathbf{L}\|}} \|\mathbf{v} - \mathbf{P}\mathbf{w}\|.\end{aligned}$$

Dividing both sides by $\|\mathbf{v} - \mathbf{P}\mathbf{w}\|$, squaring, and applying the $\text{SAP}_{\mathbf{L}}$ for $\mathbf{v}$ yields the $\text{WAP}_{\mathbf{L}}$ (note the C^2):

$$\|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 \leq \frac{C}{\|\mathbf{L}\|} \|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}}^2 \leq \frac{C^2}{\|\mathbf{L}\|^2} \|\mathbf{L}\mathbf{v}\|^2 \leq \frac{C^2}{\|\mathbf{L}\|} \|\mathbf{v}\|_{\mathbf{L}}^2.$$

- $\text{WAP}_{\mathbf{L}^2} \implies \text{WAP}_{\mathbf{L}}$: For the $\mathbf{w}$ guaranteed by $\text{WAP}_{\mathbf{L}^2}$, we have that

$$\|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 \leq \frac{C}{\|\mathbf{L}\|^2} \|\mathbf{L}\mathbf{v}\|^2 \leq \frac{C}{\|\mathbf{L}\|} \|\mathbf{v}\|_{\mathbf{L}}^2.$$

- $\text{WAP}_{\mathbf{L}^2} \implies \text{SAP}_{\mathbf{L}}$: For the $\mathbf{w}$ guaranteed by $\text{WAP}_{\mathbf{L}^2}$, we have that

$$\|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}}^2 \leq \|\mathbf{L}\| \cdot \|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 \leq \frac{C}{\|\mathbf{L}\|} \|\mathbf{L}\mathbf{v}\|^2.$$

- $\text{SAP}_{\mathbf{L}} \implies \text{WAP}_{\mathbf{L}^2}$: For the $\mathbf{w}$ guaranteed by $\text{SAP}_{\mathbf{L}}$, we have that

$$\frac{C}{\|\mathbf{L}\|^2} \|\mathbf{L}\mathbf{v}\|^2 \geq \frac{\|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}}^2}{\|\mathbf{L}\|}.$$

But the $\text{WAP}_{\mathbf{L}}$, which is implied by the $\text{SAP}_{\mathbf{L}}$, in turn implies that there exists a coarse-grid $\mathbf{z}$ $\mathbf{v}$ such that

$$\frac{\|\mathbf{v} - \mathbf{P}\mathbf{w}\|_{\mathbf{L}}^2}{\|\mathbf{L}\|} \geq \frac{1}{C} \|(\mathbf{v} - \mathbf{P}\mathbf{w}) - \mathbf{P}\mathbf{z}\|^2 = \frac{1}{C} \|\mathbf{v} - \mathbf{P}(\mathbf{w} + \mathbf{z})\|^2.$$

The constants C in these two bounds may be different, but we can replace both by their largest. Thus, putting these bounds together, with $\mathbf{w}$ replacing $\mathbf{w} + \mathbf{z}$, implies the $\text{WAP}_{\mathbf{L}^2}$ with C^2 instead of C : $\|\mathbf{v} - \mathbf{P}\mathbf{w}\|^2 \leq \frac{C^2}{\|\mathbf{L}\|^2} \|\mathbf{L}\mathbf{v}\|^2$.