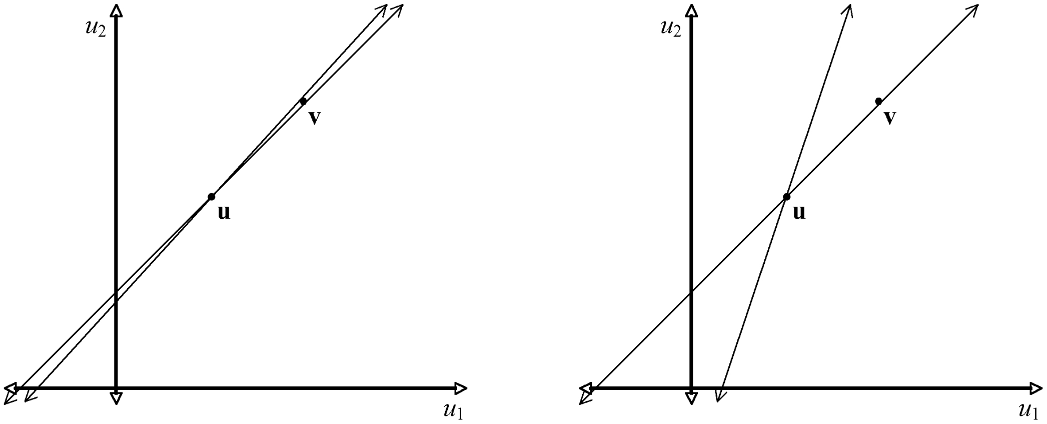

2.1 Residual vs. error. Consider the two systems of linear equations given in the box on residuals and errors in this chapter (pg 8). Make a sketch showing the pair of lines represented by each system. Mark the exact solution **u** and the approximation **v**. Explain why, even thought the error is the same in both cases, the residual is small in one case and large in the other.



$$\begin{pmatrix} 1 & -1 \\ 21 & -20 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} -1 \\ -19 \end{pmatrix} \qquad \begin{pmatrix} 1 & -1 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{1}$$

Figure 1: The appropriate sketches and their respective systems. On the left, we have the lines from the equations in system 1, $A_1\mathbf{u} = \mathbf{f}_1$. On the right, we have the lines from the equations in system 2, $A_2\mathbf{u} = \mathbf{f}_2$.

As stated in the box on pg 8, both these systems have exact solution $\mathbf{u} = (1, 2)^T$. For the same approximation $\mathbf{v} = (1.95, 3)^T$, both systems have the same error $\mathbf{e} = \mathbf{u} - \mathbf{v} = (-.95, -1)^T$ and the Euclidean norm of the error for both systems is $\|\mathbf{e}\|_2 = 1.379$. However, the size of residuals is different for system 1 than it is for system 2. The residual for system 1 is $\mathbf{r}_1 = \mathbf{f}_1 - A_1\mathbf{v} = (.05, .05)^T$ with norm $\|\mathbf{r}_1\|_2 = 0.071$, while the residual for system 2 is $\mathbf{r}_2 = \mathbf{f}_2 - A_2\mathbf{v} = (.05, -1.85)^T$ with norm $\|\mathbf{r}_2\|_2 = 1.851$.

The reason for this discrepancy can be seen in the sketches in figure 1. The approximation **v** is very close to both lines in system 1 on the left side of the figure. Another way of thinking about this is that both equations are individually *almost satisfied*. This gives small residual components and, hence, a small residual vector. For system 2, approximation **v** is only close to one line and fairly far from the other one. This means one small residual component and one large one, and a fairly large residual norm. The lesson here is that the residual does provide an error estimation, but the quality of the estimation is highly dependent on the system.

Note that this is analagous to root finding in scalar equations $y(x) = 0$ and asking how close x is to a root (*actual error*) or how close y is to zero (*residual error*). The actual error is a measure of how close we actually are to what we are looking for and is not computable in practice. The residual error is easy to compute but only gives a rough estimation of how close we are to the solution.

2.2 Residual Equation. Use the definition of the algebraic error and the residual to derive the residual equation $A\mathbf{e} = \mathbf{r}$.

The algebraic error for a current approximation $\mathbf{v}$ is defined as $\mathbf{e} := \mathbf{u} - \mathbf{v}$, where $\mathbf{u}$ is the actual solution to $A\mathbf{u} = \mathbf{f}$. Also, the residual for a current approximation is defined as $\mathbf{r} := \mathbf{f} - A\mathbf{v}$. Using these two definitions and the fact that $\mathbf{u}$ is the actual solution, we can derive the desired equation:

$$A\mathbf{e} = A(\mathbf{u} - \mathbf{v}) = A\mathbf{u} - A\mathbf{v} = \mathbf{f} - A\mathbf{v} =: \mathbf{r} \tag{2}$$

2.3 Weighted Jacobi Iteration.

Throughout this problem, recall the matrix splitting $A = D - L - U$, where D is the diagonal part of A, and $-L$ and $-U$ are the strictly lower and strictly upper parts of A, respectively.

(a) Starting with the component form of the weighted Jacobi method, show that it can be written in matrix form as $\mathbf{v}^{(1)} = [(1-\omega)I + \omega R_J]\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f}$.

The component form of weighted Jacobi update is the weighted average between a Jacobi update and the old iterate (from pg. 9):

$$v_j^{(1)} = (1-\omega)v_j^{(0)} + \omega v_j^*, \tag{3}$$

where $\mathbf{v}^*$ is the Jacobi update given in the component form,

$$v_j^* = \frac{b_j - \sum_{i\neq j}^{n} a_{ji} v_i^{(0)}}{a_{jj}}, \tag{4}$$

or equivalently in vector-matrix form,

$$\mathbf{v}^* = D^{-1}(L+U)\mathbf{v}^{(0)} + D^{-1}\mathbf{f} =: R_J\mathbf{v}^{(0)} + D^{-1}\mathbf{f}. \tag{5}$$

Now, we can rewrite the weighted Jacobi update in vector-matrix form,

$$\begin{aligned} \mathbf{v}^{(1)} &= (1-\omega)\mathbf{v}^{(0)} + \omega\mathbf{v}^* \\ &= (1-\omega)I\mathbf{v}^{(0)} + \omega(R_J\mathbf{v}^{(0)} + D^{-1}\mathbf{f}) \\ &= [(1-\omega)I + \omega R_J]\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f}. \end{aligned} \tag{6}$$

(b) Show that the weighted Jacobi method may also be written in the form:

$$\mathbf{v}^{(1)} = R_\omega\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f}.$$

Starting from where we were on part (a), and using the definiton on pg. 9 $R_\omega := (1-\omega)I + \omega R_J$, we simply have to do a substitution:

$$\mathbf{v}^{(1)} = [(1-\omega)I + \omega R_J]\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f} = R_\omega\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f}. \tag{7}$$

(c) Show that the weighted Jacobi method may also be written in the form:

$$\mathbf{v}^{(1)} = \mathbf{v}^{(0)} + \omega D^{-1}\mathbf{r}^{(0)}.$$

Starting from where we were on part (a) with the definitions $R_J := D^{-1}(L+U)$ and $\mathbf{r}^{(0)} := \mathbf{f} - (D-L-U)\mathbf{v}^{(0)}$, and the fact that $I = D^{-1}D$, we derive:

$$\begin{aligned} \mathbf{v}^{(1)} &= [(1-\omega)I + \omega D^{-1}(L+U)]\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f} \\ &= (1-\omega)\mathbf{v}^{(0)} + \omega D^{-1}(L+U)\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f} \\ &= \mathbf{v}^{(0)} - \omega D^{-1}D\mathbf{v}^{(0)} + \omega D^{-1}(L+U)\mathbf{v}^{(0)} + \omega D^{-1}\mathbf{f} \\ &= \mathbf{v}^{(0)} + \omega D^{-1}(\mathbf{f} - (D-L-U)\mathbf{v}^{(0)}) \\ &= \mathbf{v}^{(0)} + \omega D^{-1}\mathbf{r}^{(0)}. \end{aligned} \tag{8}$$

(d) Assume that A is the matrix from the (*1D*) model problem (*with the* h^2 *on the right hand side with* $\mathbf{f}$). Show that the weighted Jacobi iteration matrix can be expressed as $R_\omega = I - \frac{\omega}{2}A$.

First for general A, we have

$$\begin{aligned} R_\omega &= (1-\omega)I + \omega D^{-1}(L+U) \\ &= I - \omega D^{-1}D + \omega D^{-1}(L+U) \\ &= I - \omega D^{-1}(D-L-U) \\ &= I - \omega D^{-1}A. \end{aligned} \tag{9}$$

And, for the 1D model problem's A, with the h^2 on the right-hand-side, every diagonal element of A is 2, so $D = 2I$. Therefore, $D^{-1} = \frac{1}{2}I$. Thus, we have:

$$R_\omega = I - \omega D^{-1}A = I - \frac{\omega}{2}IA = I - \frac{\omega}{2}A. \tag{10}$$

2.8 Eigenvalues of the (1D) model problem. Compute the eigenvalues of the matrix equation A of the one-dimensional model problem. (Hint: Write out a typical equation of the system $A\mathbf{w} = \lambda\mathbf{w}$ with $w_0 = w_n = 0$. Notice that the vectors form $w_j^k = \sin(\frac{jk\pi}{n}), 1 \leq k \leq n-1, 0 \leq j \leq n$, satisfy the boundary conditions.) How many distinct eigenvalues are there? Compute $\lambda_1, \lambda_2, \lambda_{n-2}, \lambda_{n-1}$, when $n = 32$.

Assuming that $\mathbf{w}^k$ is an eigenvector, it must satisfy the eigenproblem $A\mathbf{w}^k = \lambda_k\mathbf{w}$ which is equivalent to the $n-1$ equations

$$\frac{-w_{j-1}^k + 2w_j^k - w_{j+1}^k}{h^2} = \lambda_k w_j^k. \tag{11}$$

The hint implies that we should try sine vectors, $w_j^k = \sin(\frac{jk\pi}{n}), 1 \leq k \leq n-1$, as eigenvectors. Plugging the sine vectors into the left-hand-side of equation (11) and simplifying the equation to match the form of the right-hand-side will give us the eigenvalues λ_k:

$$-\sin\left(\frac{(j-1)k\pi}{n}\right) + 2\sin\left(\frac{jk\pi}{n}\right) - \sin\left(\frac{(j+1)k\pi}{n}\right). \tag{12}$$

Using the trig identities $\sin(a \pm b) = \sin(a)\cos(b) \pm \cos(a)\sin(b)$, we have

$$\sin\left(\frac{(j-1)k\pi}{n}\right) = \sin\left(\frac{jk\pi}{n}\right)\cos\left(\frac{k\pi}{n}\right) - \cos\left(\frac{jk\pi}{n}\right)\sin\left(\frac{k\pi}{n}\right), \tag{13}$$

and

$$\sin\left(\frac{(j+1)k\pi}{n}\right) = \sin\left(\frac{jk\pi}{n}\right)\cos\left(\frac{k\pi}{n}\right) + \cos\left(\frac{jk\pi}{n}\right)\sin\left(\frac{k\pi}{n}\right). \tag{14}$$

So the left-hand-side of (11) can be simplified to

$$\left[2 - 2\cos\left(\frac{h\pi}{n}\right)\right]\sin\left(\frac{jk\pi}{n}\right). \tag{15}$$

This means that $\lambda_k = 2 - 2\cos\left(\frac{k\pi}{n}\right)$; using the trig identity $\sin^2 a = \frac{1-\cos(2a)}{2}$, we get

$$\lambda_k = 4\sin^2\left(\frac{k\pi}{2n}\right). \tag{16}$$

This is an increasing sequence of eigenvalues in the range $1 \leq k \leq n-1$, so we have $n-1$ distinct eigenvalues. For $n = 32$, $\lambda_1 = .00963, \lambda_2 = .03843, \lambda_{30} = 3.96157, \lambda_{31} = 3.99037$.

2.10 Jacobi eigenvalues and eigenvectors. Find the eigenvalues of the weighted Jacobi iteration matrix when it is applied to the one-dimensional model problem matrix A, (*with* $\sigma = 0$). Verify that the eigenvectors of R_ω are the same as the eigenvectors of A.

Recall that the eigenvectors of A satisfy $A\mathbf{w}^k = \lambda_k \mathbf{w}^k$ (problem 2.8) and that for our model problem $R_\omega = I - \frac{\omega}{2}A$ (problem 2.3d). These facts can be used to show that

$$R_\omega \mathbf{w}^k = (I - \frac{\omega}{2}A)\mathbf{w}^k = \mathbf{w}^k - \frac{\omega}{2}A\mathbf{w}^k = \mathbf{w}^k - \frac{\omega}{2}\lambda_k\mathbf{w}^k = (1 - \frac{\omega}{2}\lambda_k)\mathbf{w}^k. \tag{17}$$

So $\mathbf{w}^k$ is indeed an eigenvector of R_ω, with corresponding eigenvalue

$$\mu_k = 1 - \frac{\omega}{2}\lambda_k = 1 - 2\omega\sin^2\left(\frac{k\pi}{2n}\right). \tag{18}$$

2.13 Optimal Jacobi. Show that when the weighted Jacobi method is used with $\omega = \frac{2}{3}$, the smothing factor is $\frac{1}{3}$. Show that if ω is chosen to damp the smooth modes effectively, then the oscillatory modes are actually amplified.

The sequence of eigenvalues of R_ω found in problem 2.10 could be modeled with a continuous function of wave number $\mu(k) = 1 - 2\omega\sin^2(\frac{k\pi}{2n})$. If we want to maximize the absolute value of this function over the oscillatory wave numbers $\frac{n}{2} \leq k < n$, we get a good estimate of the smoothing factor. To maximize, we set the derivative to zero:

$$\mu'(k) = -4\omega\sin\left(\frac{k\pi}{2n}\right)\cos\left(\frac{k\pi}{2n}\right) = 0, \tag{19}$$

which implies that $\frac{k\pi}{2n} = 0$ or $\frac{k\pi}{2n} = \frac{\pi}{2}$. Since $k \approx 0 \leq \frac{n}{2}$ is a smooth wavenumber and outside of our range of minimization, we look at $k \approx n$. We also look at the other endpoint: $k \approx \frac{n}{2}$. When $\omega = \frac{2}{3}$,

$$|\mu(n)| = \left|1 - \frac{4}{3}\sin^2\frac{\pi}{2}\right| = \left|-\frac{1}{3}\right| = \frac{1}{3}, \tag{20}$$

and

$$\left|\mu\left(\frac{n}{2}\right)\right| = \left|1 - \frac{4}{3}\sin^2\frac{\pi}{4}\right| = \left|\frac{1}{3}\right| = \frac{1}{3}. \tag{21}$$

So the smoothing factor is $\frac{1}{3}$.

Alternatively, if we require ω to damp the smoothest mode, $k = 1$, by a constant convergence factor $0 < \tau < 1$, we ask that

$$|\mu_1| = \left|1 - 2\omega \sin^2\left(\frac{\pi}{2n}\right)\right| \leq \tau. \tag{22}$$

Solving for th ω that cause this to happen, we get

$$\frac{1-\tau}{2\sin^2\left(\frac{\pi}{2n}\right)} \leq \omega \leq \frac{1+\tau}{2\sin^2\left(\frac{\pi}{2n}\right)}, \tag{23}$$

a range of values that grow arbitrarily large as n gets big. If we use any ω in this range, the absolute value of the eigenvalues of R_ω for oscillatory wave numbers, $\mu_{\frac{n}{2}}$ through μ_{n-1}, also get arbitrarily large. For the best-case example, consider the lowest oscillatory eigenvalue, $\mu_{\frac{n}{2}}$, with lowest ω, $\omega = \frac{1-\tau}{2\sin\left(\frac{\pi}{2n}\right)}$:

$$|\mu_{\frac{n}{2}}| = \left|1 - \frac{2(1-\tau)\sin^2\left(\frac{\pi}{4}\right)}{2\sin^2\left(\frac{\pi}{2n}\right)}\right| = \left|1 - \frac{(1-\tau)}{2\sin^2\left(\frac{\pi}{2n}\right)}\right|. \tag{24}$$

For $\tau = .9$ and $n = 32$, and choosing ω as the best case, $|\mu_{\frac{n}{2}}| = 19.77$. This means that this relaxation would multiply the amplitude of oscillatory modes by about 20 per iteration. Also, note that $\tau = .9$ is a modest convergence factor and $n = 32$ is a *very* small problem. This effect is much more dramatic for larger problems.

2.16 Properties of Gauss-Seidel. Assume A is symmetric, positive definite.

(a) Show that the ith step of a *single sweep* of the Gauss-Seidel method applied to $A\mathbf{u} = \mathbf{f}$ may be expressed as

$$v_i \leftarrow v_i + \frac{r_i}{a_{ii}}. \tag{25}$$

First note that i was swapped with j throughout this problem for notational convenience. The ith step in a single sweep of Gauss-Seidel may be thought of as solving the ith equation for v_i, the ith unknown, using the most current information for the rest of the components of $\mathbf{v}$. Doing so gives the *component form* of the ith step in a single Gauss-Seidel iteration, a common way of stating the Gauss-Seidel iteration:

$$\begin{array}{rcl} \sum_{j=1}^{n} a_{ij}v_j & = & f_i \\ a_{ii}v_i & = & f_i - \sum_{j\neq i} a_{ij}v_j \\ v_i & \leftarrow & \frac{1}{a_{ii}}[f_i - \sum_{j\neq i} a_{ij}v_j] \quad \text{for} \quad i = 1, ..., n. \end{array} \tag{26}$$

Now adding and subtracting v_i to this iteration allows us to rewrite this in the desired form:

$$\frac{1}{a_{ii}}[f_i - \sum_{j\neq i} a_{ij}v_j] + v_i - v_i = \frac{1}{a_{ii}}[f_i - \sum_{i=1}^{n} a_{ij}v_j] + v_i = v_i + \frac{r_i}{a_{ii}}. \tag{27}$$

(b) Show that the ith step of a *single sweep* of the Gauss-Seidel method can be expressed in vector form as

$$\mathbf{v} \leftarrow \mathbf{v} + \frac{(\mathbf{r}, \hat{\mathbf{e}}_i)}{(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i)}\hat{\mathbf{e}}_i, \tag{28}$$

where $\hat{\mathbf{e}}_i$ is the ith (elementary) unit vector.

First, note that taking an inner product of the elementary unit vector $\hat{\mathbf{e}}_i$ with any vector just yeilds the ith component of that vector. In particular, $(\mathbf{r}, \hat{\mathbf{e}}_i) = r_i$. Also, note that the A inner product of a elementary unit vector gives a diagonal component of A: $(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i) = a_{ii}$. Then, using the fact that the ith step is only changing the ith component of $\mathbf{v}$, we can rewrite the iteration in *vector form*:

$$\mathbf{v} \leftarrow \mathbf{v} + \frac{r_i}{a_{ii}}\hat{\mathbf{e}}_i = \mathbf{v} + \frac{(\mathbf{r}, \hat{\mathbf{e}}_i)}{(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i)}\hat{\mathbf{e}}_i. \tag{29}$$

(c) Show that each sweep of Gauss-Seidel decreases the quantity $(A\mathbf{e}, \mathbf{e})$ (the A-norm of the error), where $\mathbf{e} = \mathbf{u} - \mathbf{v}$.

The error before performing the ith step is $\mathbf{u} - \mathbf{v}$ and the error afterwards is

$$\mathbf{u} - \left(\mathbf{v} + \frac{r_i}{a_{ii}}\hat{\mathbf{e}}_i\right) = \mathbf{e} - \frac{r_i}{a_{ii}}\hat{\mathbf{e}}_i. \tag{30}$$

Looking at the A-norm of the error afterward and simplifying gives

$$\begin{aligned}\left(A(\mathbf{e} - \tfrac{r_i}{a_{ii}}\hat{\mathbf{e}}_i), \mathbf{e} - \tfrac{r_i}{a_{ii}}\hat{\mathbf{e}}_i\right) &= \left(A\mathbf{e} - \tfrac{r_i}{a_{ii}}A\hat{\mathbf{e}}_i, \mathbf{e} - \tfrac{r_i}{a_{ii}}\hat{\mathbf{e}}_i\right) \\ &= (A\mathbf{e}, \mathbf{e}) - 2\tfrac{r_i}{a_{ii}}(A\mathbf{e}, \hat{\mathbf{e}}_i) + \left(\tfrac{r_i}{a_{ii}}\right)^2 (A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i) \\ &= (A\mathbf{e}, \mathbf{e}) - 2\tfrac{r_i^2}{a_{ii}} + \tfrac{r_i^2}{a_{ii}} \\ &= (A\mathbf{e}, \mathbf{e}) - \tfrac{r_i^2}{a_{ii}}.\end{aligned} \tag{31}$$

The quantity $\frac{r_i^2}{a_{ii}}$ is always positive because both the numerator and denominator are positive. The numerator is the square of a real number and the denominator is a diagonal element of a symmetric positive definite matrix. All this shows that the norm is nonincreasing for each single step of Gauss-Seidel and, therefore, for the entire sweep. All we need is just one of the r_i to be nonzero for it to actually decrease, which must be true because we assumed that the error is nonzero, so the residual must be nonzero too.

(d) Show that Gauss-Seidel is optimal in the sense that the quantity $\|\mathbf{e} - s\hat{\mathbf{e}}_i\|_A$ is minimized for each $1 \leq j \leq n$ when $s = (\mathbf{r}, \hat{\mathbf{e}}_i)/(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i)$, which is precisely a Gauss-Seidel step. To show that the s used in the ith step of Gauss-Seidel is

the minimizer for the quantity $\|\mathbf{e} - s\hat{\mathbf{e}}_i\|_A$, we instead look at the value of s that minimizes the square of that quantity:

$$\begin{aligned} f(s) &= \|\mathbf{e} + s\hat{\mathbf{e}}_i\|_A^2 \\ &= (A(\mathbf{e} + s\hat{\mathbf{e}}_i), \mathbf{e} + s\hat{\mathbf{e}}_i) \\ &= (A\mathbf{e}, \mathbf{e}) - 2s(A\mathbf{e}, \hat{\mathbf{e}}_i) + s^2(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i). \end{aligned} \tag{32}$$

We minimize $f(s)$ by setting $f'(s)$ equal to zero:

$$f'(s) = -2(A\mathbf{e}, \hat{\mathbf{e}}_i) + 2s(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i) = 0. \tag{33}$$

The solution to this equation is $s = (\mathbf{r}, \hat{\mathbf{e}}_i)/(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i)$, which is where our minimum is attained. We know this is a minimum because the function $f(s)$ is concave up due to the fact that A is symmetric positive definite:

$$f''(s) = 2(A\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i) = 2a_{ii} > 0. \tag{34}$$

The minimum of $f(s)$ occurs at the same place that the minimum of $\sqrt{f(s)} = \|\mathbf{e} - s\hat{\mathbf{e}}_i\|_A$ does. Notice that we have shown that each step of Gauss-Seidel moves the iterate in the $\hat{\mathbf{e}}_i$ direction by the amount s that minimizes the new error in the A-norm.