STAT 309: MATHEMATICAL COMPUTATIONS I
FALL 2015
LECTURE 17

1. Richardson method

- unlike the splitting methods in the previous lecture, the iterative methods here do not require splitting A into a sum of two matrices but they are a bit like SOR in that there is a scalar parameter involved at each step
- this scalar parameter can either be fixed throughout (e.g., Richardson) or can vary from one interation to the next (e.g., steepest descent, Chebyshev) or there can even be two scalar parameters at each step (e.g., conjugate gradient)
- the simplest one is known as the *Richardson method*, where the iteration is simply

$$\begin{aligned} \mathbf{x}^{(k+1)} &= (I - \alpha A)\mathbf{x}^{(k)} + \alpha \mathbf{b} \\ &= \mathbf{x}^{(k)} + \alpha(\mathbf{b} - A\mathbf{x}^{(k)}) \\ &= \mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)} \end{aligned} \tag{1.1}$$

  where $\mathbf{r}^{(k)} := \mathbf{b} - A\mathbf{x}^{(k)}$ is the *residual* at the kth step
- note that if $\mathbf{x} = A^{-1}\mathbf{b}$, then we trivially have

$$\mathbf{x} = (I - \alpha A)\mathbf{x} + \alpha \mathbf{b} \tag{1.2}$$

- as usual we define the error $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$, then subtracting (1.1) from (1.2) yields

$$\mathbf{e}^{(k+1)} = B_\alpha \mathbf{e}^{(k)}$$

  where the iteration matrix is $B_\alpha = I - \alpha A$
- we want to choose the parameter $\alpha > 0$ a priori so as to minimize $\|B_\alpha\|_2$
- suppose A is symmetric positive definite, with eigenvalues

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n > 0$$

- since $B_\alpha = I - \alpha A$, we have $\lambda_i = 1 - \alpha \mu_i$ for $i = 1, \ldots, n$
- also since B_α is symmetric we have $\|B_\alpha\|_2 = \rho(B_\alpha)$ (not true if A is not symmetric) and so in this case, miminizing $\|B_\alpha\|_2$ is the same as minimizing $\rho(B_\alpha)$, i.e., finding the optimal α so that the error goes to zero as rapidly as possible
- if we want α so that $\|B_\alpha\|_2$ is minimized, i.e.,

$$\min_\alpha \max_{1 \leq i \leq n} |\lambda_i(\alpha)| = \min_\alpha \max_{1 \leq i \leq n} |1 - \alpha \mu_i|$$

  the optimal parameter $\hat{\alpha}$ is found by solving

$$1 - \hat{\alpha}\mu_n = -(1 - \hat{\alpha}\mu_1)$$

  which yields

$$\hat{\alpha} = \frac{2}{\mu_1 + \mu_n}$$

- this is in fact the ∞-norm case in 2013's Homework **3**, Problem **1**(g) (I omitted part (g) this year; you are welcomed to do it as an exercise)
- note that when $1 - \alpha\mu_1 = -1$, the iteration diverges for some choice of $\mathbf{x}^{(0)}$

Date: December 2, 2015, version 1.0. Comments, bug reports: `lekheng@galton.uchicago.edu`.

- hence the method converges for

$$0 < \alpha < \frac{2}{\mu_1}$$

- however this iteration is sensitive to perturbation and therefore bad numerically
- for example, if $\mu_1 = 10$ and $\mu_n = 10^{-4}$, then the optimal α is $2/(10 + 10^{-4})$, but this is close to a value of α for which the iteration diverges, $\alpha = 2/10$
- also, note that

$$\lambda_1(\hat{\alpha}) = 1 - \frac{2}{\mu_1 + \mu_n}\mu_1 = \frac{\mu_n - \mu_1}{\mu_1 + \mu_n} = \frac{1 - \kappa(A)}{1 + \kappa(A)} \leq 0,$$

and similarly,

$$\lambda_n(\hat{\alpha}) = \frac{\mu_1 - \mu_n}{\mu_1 + \mu_n} = \frac{\mu_1/\mu_n - 1}{\mu_1/\mu_n + 1} = \frac{\kappa(A) - 1}{\kappa(A) + 1} \geq 0$$

- therefore

$$\|B_{\hat{\alpha}}\|_2 = \rho(B_{\hat{\alpha}}) = \frac{\kappa(A) - 1}{\kappa(A) + 1}$$

and we see that the convergence rate depends on $\kappa(A)$

- for an actual example from applications, consider the Helmholtz equation on a rectangle R,

$$\begin{aligned} -\Delta \mathbf{u}^{(k+1)} + \sigma(x, y)\mathbf{u}^{(k)} &= \mathbf{f}, \quad (x, y) \in R \\ \mathbf{u} &= \mathbf{g}, \quad (x, y) \in \partial R \end{aligned}$$

- using a finite difference approximation for Δ gives

$$A = \begin{bmatrix} T & -I & & \\ -I & \ddots & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & T \end{bmatrix}$$

and thus the iteration has the form

$$A\mathbf{u}^{(k+1)} + h^2 \Sigma \mathbf{u}^{(k)} = \mathbf{f}$$

where

$$\Sigma = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{nn} \end{bmatrix}, \quad \sigma_{ij} = \sigma(x_i, y_j)$$

- to determine the rate of convergence, we define the *error operator* by

$$\mathbf{e}^{(k+1)} = (h^2 A^{-1} \Sigma)\mathbf{e}^{(k)}$$

- therefore

$$\|\mathbf{e}^{(k+1)}\|_2 \leq h^2 \|A^{-1}\|_2 \|\Sigma\|_2 \|\mathbf{e}^{(k)}\|_2$$

but

$$\|\Sigma\|_2 = \max_{i,j} |\sigma_{ij}|$$

and

$$\lambda_{\min} = 4 - 4\cos \pi h = 4(1 - \cos \pi h) = 8 \sin^2 \left(\frac{\pi h}{2} \right)$$

- therefore
$$\|\mathbf{e}^{(k+1)}\|_2 \leq \frac{\max_{i,j}|\sigma_{ij}|}{2\left(\frac{\sin xh/2}{h/2}\right)^2}\|\mathbf{e}\|_2 \approx \frac{\max_{i,j}|\sigma_{ij}|}{2\pi^2}\|\mathbf{e}^{(k)}\|_2$$
and thus the size of the problem mesh has disappeared, and the method converges if $\max_{i,j}|\sigma_{ij}| \leq 20$
- the rate of convergence is essentially independent of h, which is very desirable

2. METHOD OF STEEPEST DESCENT

- to speed up Richardson method, we consider varying the parameter α from one iteration to the next, i.e.,
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)} \tag{2.1}$$
where α_k is to be chosen at the kth iteration
- again we will assume that A is symmetric positive definite
- given that
$$\mathbf{r}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k)} - \alpha_k A\mathbf{r}^{(k)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{r}^{(k)}$$
- we wish to choose α_k so that $\mathbf{r}^{(k+1)\mathsf{T}}A^{-1}\mathbf{r}^{(k+1)}$ is minimized
- note that this is just the Mahalanobis norm $\|\mathbf{r}^{(k+1)}\|_{A^{-1}}^2$ (why not minimize the 2-norm $\|\mathbf{r}^{(k+1)}\|_2^2$ instead?)
- now
$$\begin{aligned}\mathbf{r}^{(k+1)\mathsf{T}}A^{-1}\mathbf{r}^{(k+1)} &= (\mathbf{r}^{(k)\mathsf{T}} - \alpha_k \mathbf{r}^{(k)\mathsf{T}}A)A^{-1}(\mathbf{r}^{(k)} - \alpha_k A\mathbf{r}^{(k)} \\ &= \mathbf{r}^{(k)\mathsf{T}}A^{-1}\mathbf{r}^{(k)} - 2\alpha_k \mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)} + \alpha_k^2 \mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)}\end{aligned} \tag{2.2}$$
- to find the minimium, we differentiate with respect to α_k and obtain
$$\frac{d}{d\alpha_k}\mathbf{r}^{(k+1)\mathsf{T}}A^{-1}\mathbf{r}^{(k+1)} = -2\mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)} + 2\alpha_k \mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)}$$
which yields
$$\hat{\alpha}_k = \frac{\mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)}}{\mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)}}$$
- note that this is well-defined (denominator not zero) since A is symmetric positive definite
- with this choice of α_k, this method is known as the *method of steepest descent*
- note that
$$0 < \lambda_{\min}(A) \leq \frac{\mathbf{x}^{\mathsf{T}}A\mathbf{x}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} \leq \lambda_{\max}(A)$$
and therefore
$$\frac{1}{\lambda_{\max}(A)} \leq \hat{\alpha}_k \leq \frac{1}{\lambda_{\min}(A)}$$
- substituting $\hat{\alpha}_k$ into (2.2) yields
$$\begin{aligned}\mathbf{r}^{(k+1)\mathsf{T}}A^{-1}\mathbf{r}^{(k+1)} &= \mathbf{r}^{(k)\mathsf{T}}A^{-1}\mathbf{r}^{(k)} - 2\mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)}\frac{\mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)}}{\mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)}} + \left(\frac{\mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)}}{\mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)}}\right)^2 \mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)} \\ &= \mathbf{r}^{(k)\mathsf{T}}A^{-1}\mathbf{r}^{(k)} - \frac{(\mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)})^2}{\mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)}}\end{aligned}$$
and therefore
$$\frac{\|\mathbf{r}^{(k+1)}\|_{A^{-1}}^2}{\|\mathbf{r}^{(k)}\|_{A^{-1}}^2} = 1 - \frac{(\mathbf{r}^{(k)\mathsf{T}}\mathbf{r}^{(k)})^2}{(\mathbf{r}^{(k)\mathsf{T}}A^{-1}\mathbf{r}^{(k)})(\mathbf{r}^{(k)\mathsf{T}}A\mathbf{r}^{(k)})}$$

- the *Kantorovich inequality*, which comes up very often in applications such as optimization and statistics, states that for a symmetric positive definite A,

$$\frac{\mathbf{x}^\mathsf{T} A\mathbf{x} \cdot \mathbf{x}^\mathsf{T} A^{-1}\mathbf{x}}{(\mathbf{x}^\mathsf{T}\mathbf{x})^2} \leq \left(\frac{\sqrt{\kappa} + \sqrt{\kappa}^{-1}}{2}\right)^2, \quad \kappa = \kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

- it follows from the Kantorovich inequality that

$$\frac{\|\mathbf{r}^{(k+1)}\|_{A^{-1}}^2}{\|\mathbf{r}^{(k)}\|_{A^{-1}}^2} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2$$

- thus,

$$\frac{\|\mathbf{r}^{(1)}\|_{A^{-1}}}{\|\mathbf{r}^{(0)}\|_{A^{-1}}} \cdot \frac{\|\mathbf{r}^{(2)}\|_{A^{-1}}}{\|\mathbf{r}^{(1)}\|_{A^{-1}}} \cdot \cdots \cdot \frac{\|\mathbf{r}^{(k)}\|_{A^{-1}}}{\|\mathbf{r}^{(k-1)}\|_{A^{-1}}} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k$$

  which yields

$$\frac{\|\mathbf{r}^{(k)}\|_{A^{-1}}}{\|\mathbf{r}^{(0)}\|_{A^{-1}}} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k$$

- in other words, the rate of convergence is the same as when the parameter α_k is chosen a priori to be

$$\hat{\alpha} = \frac{2}{\mu_1 + \mu_n}$$

- so it would appear that we might as well have used Richardson's method in the first place
- but that's not quite the case — the problem is that we must know μ_1 and μ_n in order to determine the optimal $\hat{\alpha}$
- steepest descent does not require us to know μ_1 and μ_n
- however the price we pay for steepest descent is that we need to compute α_k at each step
- if you know some optimization, the steepest descent method described above is the same as applying the steepest descent method in continuous optimization to the problem

$$\min \frac{1}{2}\mathbf{x}^\mathsf{T} A\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x}$$

3. Chebyshev iteration

- again we are interested in solving $A\mathbf{x} = \mathbf{b}$
- let us rewrite the steepest descent iteration (2.1) in the form

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)}\\ &= (I - \alpha_k A)\mathbf{x}^{(k)} + \alpha_k \mathbf{b}\end{aligned}$$

- this time, instead of picking only α_k to minimize some quantity at the kth step, we will pick $\alpha_0, \alpha_1, \ldots, \alpha_k$ simultaneously to minimize some quantity at the kth step
- since the exact solution $\mathbf{x}$ satisfies

$$\mathbf{x} = (I - \alpha_k A)\mathbf{x} + \alpha_k \mathbf{b},$$

  it follows that

$$\mathbf{e}^{(k+1)} = (I - \alpha_k A)\mathbf{e}^{(k)}$$

  where $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$
- so we have

$$\begin{aligned}\mathbf{e}^{(1)} &= (I - \alpha_0 A)\mathbf{e}^{(0)}\\ &\vdots\\ \mathbf{e}^{(k)} &= (I - \alpha_{k-1} A)(I - \alpha_{k-2} A)\cdots(I - \alpha_0 A)\mathbf{e}^{(0)}\end{aligned}$$

- in other words,
$$\mathbf{e}^{(k)} = P_k(A)\mathbf{e}^{(0)}$$
where
$$P_k(A) = (I - \alpha_{k-1}A)(I - \alpha_{k-2}A)\cdots(I - \alpha_0 A).$$
is a polynomial of degree k
- by the Cayley–Hamilton theorem, the minimal polynomial $\psi(x)$ of A has the following property:
$$\psi(A) = \prod_{k=0}^{d-1}(A - \mu_k I) = 0$$
where d is the number of distinct eigenvalues μ_k of A, when $A = A^{\mathsf{T}}$
- in other words
$$\prod_{k=0}^{d-1}\left(I - \frac{1}{\mu_k}A\right) = 0$$
so we could choose $\alpha_k = 1/\mu_k$, but this is a bad choice because we almost never know the eigenvalues of A and even if we do, this choice is unstable because μ_k can vary immensely in magnitude
- however, we have
$$\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|} \leq \|P_k(A)\|_2,$$
which allows us to use approximation theory to find a suitable P_k
- if $A = Q\Lambda Q^{\mathsf{T}}$ where $\Lambda = \operatorname{diag}(\mu_1, \ldots, \mu_n)$, then $P_k(A) = QP_k(\Lambda)Q^{\mathsf{T}}$, and therefore
$$\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|} \leq \|P_k(\Lambda)\|_2$$
- since
$$P_k(\Lambda) = \begin{bmatrix} P_k(\mu_1) & & \\ & \ddots & \\ & & P_k(\mu_n) \end{bmatrix},$$
it follows that
$$\|P_k(\Lambda)\|_2 = \max_{1\leq i\leq n}|P_k(\mu_i)|$$
- note that $P_k(O) = I$ and so we want a polynomial that solves the minimization problem
$$\min_{p_k(0)=1}\max_{1\leq i\leq n}|p_k(\mu_i)|$$
- we start by finding a monic[1] polynomial $\hat{p}_k(\mu)$ such that
$$\max_{1\leq i\leq n}|\hat{p}_k(\mu_i)| = \min_{p_k \text{ monic}}\max_{1\leq i\leq n}|p_k(\mu_i)|$$
- why monic? note that if we do not restrict the coefficients of p_k in some way, the solution to the above minimization problem will be trivial (the minimum can be made arbitrarily small)
- in the homework we will see another way to restrict the coefficients of p_k
- clearly,
$$\min_{p_k \text{ monic}}\max_{1\leq i\leq n}|p_k(\mu_i)| \leq \min_{p_k \text{ monic}}\max_{\mu_n\leq\mu\leq\mu_1}|p_k(\mu)| \leq \min_{p_k \text{ monic}}\max_{\alpha\leq\mu\leq\beta}|p_k(\mu)|$$
where $\alpha \leq \mu_n \leq \mu_1 \leq \beta$

[1] a monic polynomial is one with leading coefficient 1, i.e., $x^k + a_{k-1}x^{k-1} + \cdots + a_1 x + a_0$

- ideally we want $\alpha = \mu_n$ and $\beta = \mu_1$ but often we don't have the eigenvalues but only lower and upper bounds
- therefore, we will try to find a monic polynomial $\hat{p}_k$ that is of minimum absolute value on the interval $[\alpha, \beta]$
- the solution to this problem is given by the *Chebyshev polynomials*, suitably normalized so that it is monic
- the Chebyshev polynomial of degree k is most easily defined to be the polynomial expression that gives the expansion of $\cos(k\theta)$ in terms of $\cos(\theta)$, formally,

$$C_k(\cos x) = \cos kx$$

or

$$C_k(x) = \cos(k \cos^{-1}(x))$$

- but since as a function, $\cos^{-1}(x)$ is not defined when $|x| > 1$, a more careful definition would be

$$C_k(x) = \begin{cases} \cos(k \cos^{-1}(x)) & \text{if } |x| \le 1 \\ \cosh(k \cosh^{-1}(x)) & \text{if } x > 1 \\ (-1)^k \cosh(k \cosh^{-1}(-x)) & \text{if } x < -1 \end{cases}$$

- for example,

$$C_0(x) = 1, \quad C_1(x) = x, \quad C_2(x) = 2x^2 - 1, \quad C_3(x) = 4x^3 - 3x, \quad C_4(x) = 8x^4 - 8x^2 + 1$$

- these polynomials are by definition bounded by 1 in absolute value on the interval $|x| \le 1$
- if $\theta = \cos^{-1} x$ then, using the trigonometric identities

$$\cos(k+1)\theta = \cos k\theta \cos\theta - \sin k\theta \sin\theta$$
$$\cos(k-1)\theta = \cos k\theta \cos\theta + \sin k\theta \sin\theta$$

we obtain

$$\cos(k+1)\theta = 2\cos k\theta \cos\theta - \cos(k-1)\theta$$

which yields the three-term recurrence relation of the Chebyshev polynomials

$$C_{k+1}(x) = 2xC_k(x) - C_{k-1}(x)$$

- since this relation leads to a leading coefficient of 2^{k-1} for $C_k(x)$ when $k \ge 1$, we need to normalize so that it is monic:

$$T_k(x) := \frac{C_k(x)}{2^{k-1}}, \quad k = 1, 2, 3, \ldots$$

- we now claim that for $k = 2$, $\hat{p}_2(x)$ is

$$T_2(x) = x - \frac{1}{2},$$

scaled and translated appropriately so as to be small on the interval $[\alpha, \beta]$ and monic
- we may transform the interval $[\alpha, \beta]$ to $[-1, 1]$ by a change of variable

$$[\alpha, \beta] \ni t \mapsto \frac{2t - (\beta + \alpha)}{\beta - \alpha} \in [-1, 1]$$

- note that on $[-1, 1]$, $T_2(x)$ has a maximum at $x = -1$ and $x = 1$, and a local minimum at $x = 0$
- now, suppose that there is another polynomial $p_2(x) = x^2 + bx + c$ such that $p_2(-1) < T_2(-1)$, $p_2(1) < T_2(1)$, and $p_2(0) > T_2(0)$
- then the polynomial $q_1(x) = T_2(x) - p_2(x)$ has three sign changes in the interval $[-1, 1]$, but since $T_2(x)$ and $p_2(x)$ have the same leading coefficient, $q_1(x)$ can have degree at most 1, so it must be identically zero

- we obtain the following

Theorem 1. *The monic polynomial of degree exactly k having smallest uniform norm*[2] *in $C[\alpha, \beta]$ is*

$$\left(\frac{\beta-\alpha}{2}\right)^k T_k\left(\frac{2x-\beta-\alpha}{\beta-\alpha}\right).$$

- we want
$$\frac{\|\mathbf{e}^{(k)}\|_2}{\|\mathbf{e}^{(0)}\|_2} \leq \|P_k(A)\|_2 \leq \max_{1\leq i\leq n}|P_k(\mu_i)| \leq \max_{\alpha\leq\mu\leq\beta}|P_k(\mu)|$$
where $P_k(0) = I$ and the eigenvalues of A are contained in the interval $[\alpha, \beta]$
- if we fix k, then we have
$$\alpha_j^{(k)} = \left[\frac{\beta+\alpha}{2} - \left(\frac{\beta-\alpha}{2}\right)\cos\frac{(2j+1)\pi}{2k}\right]^{-1}, \quad j = 0, \ldots, k-1$$
- note that
$$\alpha_0^{(1)} = \frac{2}{\beta+\alpha},$$
which is the same optimal parameter obtained using a different analysis
- therefore, we can select k and then use the parameters $\alpha_0^{(k)}, \ldots, \alpha_{k-1}^{(k)}$
- if $\|\mathbf{r}^{(k)}\|/\|\mathbf{r}^{(0)}\| \leq \varepsilon$, we can stop; otherwise, we simply recycle these parameters
- the process should not be stopped before the full cycle, because a partial polynomial may not be small on the interval $[\mu_n, \mu_1]$
- also, using the parameters in an arbitrary order may lead to numerical instabilities even though mathematically the order does not matter
- for a long time, the determination of a suitable ordering was an open problem, but it has now been solved
- it has been shown that when solving Laplace's equation using 128 parameters, a simple left-to-right ordering results in $\|\mathbf{e}^{(128)}\| \approx 10^{35}$, while the optimal ordering yields $\|\mathbf{e}^{(128)}\| \approx 10^{-7}$
- in the absence of roundoff error, using Chebyshev polynomials yields
$$\frac{\|\mathbf{e}^{(k)}\|_2}{\|\mathbf{e}^{(0)}\|_2} \leq \frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k} \approx \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$$
whereas, with steepest descent,
$$\frac{\|\mathbf{e}^{(k)}\|_2}{\|\mathbf{e}^{(0)}\|_2} \approx \left(\frac{\kappa-1}{\kappa+1}\right)^k$$

4. conjugate gradient method

- up till this point we have only considered semi-iterative methods for solving $A\mathbf{x} = \mathbf{b}$ with just one parameter α_k at each step
- now we will consider a method that depends on two parameters α_k and ω_k at each step
- we consider iterations defined by
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k-1)} + \omega_{k+1}(\alpha_k\mathbf{z}^{(k)} - \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) \tag{4.1}$$
where
$$M\mathbf{z}^{(k)} = \mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)} \tag{4.2}$$
for some M

[2]Recall that the uniform norm of a continuous function f on $[\alpha, \beta]$ is just $\|f\| = \max_{x\in[\alpha,\beta]}|f(x)|$.

- in particular, if we choose $\omega_k = 1$ and $\alpha_k = 1$ for all $k = 0, 1, \ldots$, then this reduces to
$$\mathbf{x}^{(k+1)} = M^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}) - \mathbf{x}^{(k)}$$
or
$$M\mathbf{x}^{(k+1)} = \mathbf{b} - (A - M)\mathbf{x}^{(k)} = N\mathbf{x}^{(k)} + \mathbf{b}$$
where $A = M - N$
- in other words, this includes features from both splitting methods and semi-iterative methods
- our goal is to choose the parameters α_k and ω_k so that $\|P_k(M^{-1}A)\mathbf{e}^{(0)}\|_2$ is minimized, where
$$\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)} = P_k(M^{-1}A)\mathbf{e}^{(0)}$$
- in the following we will write
$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{T}\mathbf{y}$$
- suppose we can impose the condition that
$$\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)} \rangle = \delta_{jk}$$
where both M and A are $n \times n$ and required to be symmetric positive definite
- if this is possible, then it follows that $\mathbf{z}^{(n+1)} = \mathbf{0}$, and therefore $\mathbf{r}^{(n+1)} = \mathbf{0}$, implying convergence in n iterations
- it follows from (4.1) that
$$\mathbf{b} - A\mathbf{x}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k-1)} - \omega_{k+1}(\alpha_k A\mathbf{z}^{(k)} + A\mathbf{x}^{(k)} - \mathbf{b} + \mathbf{b} - A\mathbf{y}^{(k-1)})$$
which simplifies to
$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k-1)} - \omega_{k+1}(\alpha_k A\mathbf{z}^{(k)} - \mathbf{r}^{(k)} + \mathbf{r}^{(k-1)})$$
- from (4.2), we obtain
$$M\mathbf{z}^{(k+1)} = M\mathbf{z}^{(k-1)} - \omega_{k+1}(\alpha_k A\mathbf{z}^{(k)} - M\mathbf{z}^{(k)} + M\mathbf{z}^{(k-1)})$$
- we use the induction hypothesis
$$\langle \mathbf{z}^{(p)}, M\mathbf{z}^{(q)} \rangle = 0, \quad p \neq q, \quad p = 1, 2, \ldots, k$$
- then
$$\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k+1)} \rangle = \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-1)} \rangle - \omega_{k+1}[\langle \alpha_k \mathbf{z}^{(k)}, A\mathbf{z}^{(k)} \rangle - \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)} \rangle + \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-1)} \rangle]$$
which yields
$$\alpha_k = \frac{\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)} \rangle}{\langle \mathbf{z}^{(k)}, A\mathbf{z}^{(k)} \rangle}$$
- similarly,
$$\langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k+1)} \rangle = \langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k-1)} \rangle - \omega_{k+1}[\langle \alpha_k \mathbf{z}^{(k-1)}, A\mathbf{z}^{(k)} \rangle - \langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k)} \rangle + \langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k-1)} \rangle]$$
which yields
$$\omega_{k+1} = \frac{\langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k-1)} \rangle}{\alpha_k \langle \mathbf{z}^{(k-1)}, A\mathbf{z}^{(k)} \rangle + \langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k-1)} \rangle}$$
- we can simplify this expression for ω_{k+1} by noting that by symmetry,
$$\langle \mathbf{z}^{(k-1)}, A\mathbf{z}^{(k)} \rangle = \langle \mathbf{z}^{(k)}, A\mathbf{z}^{(k-1)} \rangle$$

and therefore

$$\begin{aligned}\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle &= \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-2)}\rangle \\ &\quad + \omega_k(\alpha_{k-1}\langle \mathbf{z}^{(k)}, A\mathbf{z}^{(k-1)}\rangle - \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-1)}\rangle + \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-2)}\rangle) \\ &= \omega_k\alpha_{k-1}\langle \mathbf{z}^{(k)}, A\mathbf{z}^{(k-1)}\rangle\end{aligned}$$

which yields

$$\begin{aligned}\omega_{k+1} &= \frac{\langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k-1)}\rangle}{-\dfrac{\alpha_k}{\alpha_{k+1}}\dfrac{1}{\omega_k}\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle + \langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k-1)}\rangle} \\ &= \left[1 - \frac{\alpha_k}{\alpha_{k-1}}\frac{1}{\omega_k}\frac{\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle}{\langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k-1)}\rangle}\right]^{-1}\end{aligned}$$

- we have shown that
$$\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k+1)}\rangle = \langle \mathbf{z}^{(k-1)}, M\mathbf{z}^{(k+1)}\rangle = 0$$
- it can easily be shown that
$$\langle \mathbf{z}^{(\ell)}, M\mathbf{z}^{(k+1)}\rangle = 0, \quad \ell < k-1$$
- we now state the *classical conjugate gradient* algorithm:

  $\mathbf{x}^{(0)}$ given
  solve $M\mathbf{z}^{(0)} = \mathbf{r}^{(0)}$
  $\mathbf{p}^{(0)} = \mathbf{z}^{(0)}$
  for $k = 0, \ldots$
    $\alpha_k = \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle / \langle \mathbf{p}^{(k)}, A\mathbf{p}^{(k)}\rangle$
    $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{p}^{(k)}$
    $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} + \alpha_k A\mathbf{p}^{(k)}$
    test for convergence
    solve $M\mathbf{z}^{(k+1)} = \mathbf{r}^{(k+1)}$
    $\beta_{k+1} = \langle \mathbf{z}^{(k+1)}, M\mathbf{z}^{(k+1)}\rangle / \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle$
    $\mathbf{p}^{(k+1)} = \mathbf{z}^{(k+1)} + \beta_{k+1}\mathbf{p}^{(k)}$
  end
- it can be shown that
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(0)} + P_k(K)\mathbf{z}^{(0)}$$
where $K = M^{-1}A$
- furthermore, amongst all methods which generate a polynomial for a given $\mathbf{x}^{(0)}$, the conjugate gradient method minimizes the quantity
$$\varepsilon^{k+1} = \mathbf{e}^{(k+1)\mathsf{T}} A\mathbf{e}^{(k+1)}$$
- most notable of all is that if A has p distinct eigenvalues, then the conjugate gradient method converges in p steps
- this is particularly useful in *domain decomposition*, where the interface between two subdomains consists of only a small number of points
- the way we developed conjugate gradient here is somewhat unusual, in order to illustrate the connection with the earlier discussions
- modern ways of deriving conjugate gradient (cf. Homework **5**, Problem **6**) usually involve consideration of *Krylov subspaces* — it is in fact the first Krylov subspace iterative method