STAT 309: MATHEMATICAL COMPUTATIONS I
FALL 2019
LECTURE 16

1. Chebyshev iteration (continue)

- recall that we wanted to solve
$$\min_{P_k(1)=1} \max_{\alpha \le \mu \le \beta} |P_k(\mu)| \tag{1.1}$$
- in general we may transform the interval $[\alpha, \beta]$ to $[-1, 1]$ by a change of variable
$$[\alpha, \beta] \ni t \mapsto \frac{2t - (\beta + \alpha)}{\beta - \alpha} \in [-1, 1]$$
so it is enough to solve (1.1) for $\alpha = -1$ and $\beta = +1$
- we claim that the solution is given by Chebyshev polynomials
$$C_k(x) = \cos(k \cos^{-1}(x)) \tag{1.2}$$
and the details are in Homework **5**
- plugging in $x = 1$ shows that
$$C_k(1) = \cos(k \cos^{-1}(1)) = \cos(0) = 1$$
so C_k meets the condition in (1.1)
- also C_k are by definition bounded by 1 in absolute value on the interval $|x| \le 1$
- since as a function, $\cos^{-1}(x)$ is not defined when $|x| > 1$, a more careful version of (1.2) would be
$$C_k(x) = \begin{cases} \cos(k \cos^{-1}(x)) & \text{if } |x| \le 1 \\ \cosh(k \cosh^{-1}(x)) & \text{if } x > 1 \\ (-1)^k \cosh(k \cosh^{-1}(-x)) & \text{if } x < -1 \end{cases} \tag{1.3}$$
but almost nobody would use (1.3) — in practice, C_k are obtained from a recurrence relation that we derive next
- 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 $k \ge 1$, it is convenient to define a variant that we will call *monic Chebyshev polynomials*:
$$T_k(x) := \frac{C_k(x)}{2^{k-1}}, \quad k = 1, 2, 3, \dots$$

Date: November 20, 2019, version 1.0. Comments, bug reports: `lekheng@galton.uchicago.edu`.

- as an example, we will solve a variant of (1.1)

$$\min_{P_k \text{ monic}} \max_{-1 \le \mu \le +1} |P_k(\mu)| \tag{1.4}$$

- we claim that for $k = 2$, the solution to (1.4) is given by

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

- 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
- doing this for arbitrary k on $[\alpha, \beta]$ gives the following

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

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

- suppose the eigenvalues of A are contained in the interval $[\alpha, \beta]$, then since

$$\frac{\|\mathbf{e}^{(k)}\|_2}{\|\mathbf{e}^{(0)}\|_2} \le \|P_k(A)\|_2 \le \max_{1 \le i \le n} |P_k(\mu_i)| \le \max_{\alpha \le \mu \le \beta} |P_k(\mu)|,$$

  we want to choose P_k so that the last term on the right is minimized
- if we fix k, then we have

$$\alpha_j^{(k)} = \left[\frac{\beta + \alpha}{2} - \left(\frac{\beta - \alpha}{2}\right)\cos\frac{(2j+1)\pi}{2k}\right]^{-1}, \quad j = 0, \ldots, k-1,$$

  details are as in Homework **5**
- 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 $\|\mathbf{r}^{(k)}\| / \|\mathbf{r}^{(0)}\| \le \varepsilon$, 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, with steepest descent, we get

$$\frac{\|\mathbf{e}^{(k)}\|_2}{\|\mathbf{e}^{(0)}\|_2} \approx \left(\frac{\kappa - 1}{\kappa + 1}\right)^k$$

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

whereas using Chebyshev polynomials yields

$$\frac{\|\mathbf{e}^{(k)}\|_2}{\|\mathbf{e}^{(0)}\|_2} \leq \frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k} \approx \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k = \left(\frac{\kappa-1}{\kappa+2\sqrt{\kappa}+1}\right)^k$$

2. Classical 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

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k-1)} + \omega_{k+1}(\alpha_k \mathbf{z}^{(k)} - \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) \tag{2.1}$$

  where

$$M\mathbf{z}^{(k)} = \mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)} \tag{2.2}$$

  for some M
- in particular, if we choose $\omega_k = 1$ and $\alpha_k = 1$ for all $k = 0, 1, \dots,$ 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 (2.1) 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 (2.2), 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

$$\langle \mathbf{z}^{(p)}, M\mathbf{z}^{(q)} \rangle = 0, \quad p \neq q, \quad p = 1, 2, \dots, k$$

- 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
$$\begin{aligned}\langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle &= \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-2)}\rangle \\ &\qquad + \omega_k(\alpha_{k-1}\langle \mathbf{z}^{(k)}, A\mathbf{z}^{(k-1)}\rangle - \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-1)}\rangle + \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k-2)}\rangle) \\ &= \omega_k \alpha_{k-1}\langle \mathbf{z}^{(k)}, A\mathbf{z}^{(k-1)}\rangle\end{aligned}$$
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 $k = 0, \ldots$
  $\quad\alpha_k = \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle / \langle \mathbf{p}^{(k)}, A\mathbf{p}^{(k)}\rangle$
  $\quad\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$
  $\quad\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} + \alpha_k A\mathbf{p}^{(k)}$
  $\quad$*test for convergence*
  $\quad$solve $M\mathbf{z}^{(k+1)} = \mathbf{r}^{(k+1)}$
  $\quad\beta_{k+1} = \langle \mathbf{z}^{(k+1)}, M\mathbf{z}^{(k+1)}\rangle / \langle \mathbf{z}^{(k)}, M\mathbf{z}^{(k)}\rangle$
  $\quad\mathbf{p}^{(k+1)} = \mathbf{z}^{(k+1)} + \beta_{k+1}\mathbf{p}^{(k)}$
  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 usually involve consideration of *Krylov subspaces* — it is in fact the first Krylov subspace iterative method

3. Modern conjugate gradient method

- how to get choice of coefficients: want $\alpha_1, \dots, \alpha_k$ so that

$$\min_{\alpha_1,\dots,\alpha_k} \|\mathbf{x}_k - \mathbf{x}_*\|_A^2, \qquad \mathbf{x}_k = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_k \mathbf{v}_k, \qquad \mathbf{v}_i = A^i \mathbf{b}$$

- expand

$$\|\mathbf{x}_k - \mathbf{x}_*\|_A^2 = (\mathbf{x}_k - \mathbf{x}_*)^\mathsf{T} A (\mathbf{x}_k - \mathbf{x}_*) = \mathbf{x}_k^\mathsf{T} A \mathbf{x}_k - 2\mathbf{x}_k^\mathsf{T} A \mathbf{x}_* + \text{constant}$$

- using A-orthogonality of $\mathbf{v}_1, \dots, \mathbf{v}_k$

$$\min_{\alpha_1,\dots,\alpha_k} \sum_{i,j=1}^{k} \alpha_i \alpha_j \mathbf{v}_i^\mathsf{T} A \mathbf{v}_j^\mathsf{T} - 2\sum_{i=1}^{k} \alpha_i \mathbf{v}_i^\mathsf{T} A \mathbf{x}_* = \min_{\alpha_1,\dots,\alpha_k} \sum_{i=1}^{k} \alpha_i^2 - 2\sum_{i=1}^{k} \alpha_i \mathbf{v}_i^\mathsf{T} \mathbf{b}$$

  and so

$$\alpha_i = \frac{\mathbf{v}_i^\mathsf{T} \mathbf{b}}{\mathbf{v}_i^\mathsf{T} A \mathbf{v}_i}$$

- how to get three-term recurrence:

$$\{\mathbf{b}, A\mathbf{b}, \dots, A^{n-1}\mathbf{b}\} \xrightarrow{\text{Gram–Schmidt in } \langle \cdot,\cdot \rangle_A} \{\mathbf{v}_0, \dots, \mathbf{v}_n\}$$

- since $\mathbf{v}_j \in \operatorname{span}\{\mathbf{b}, A\mathbf{b}, \dots, A^j\mathbf{b}\}$, so

$$\begin{aligned} A\mathbf{v}_j \in \operatorname{span}\{A\mathbf{b}, A^2\mathbf{b}, \dots, A^{j+1}\mathbf{b}\} &\subseteq \operatorname{span}\{\mathbf{b}, A\mathbf{b}, \dots, A^{j+1}\mathbf{b}\} \\ &= \operatorname{span}\{\mathbf{v}_0, \dots, \mathbf{v}_{j+1}\} \subseteq \operatorname{span}\{\mathbf{v}_0, \dots, \mathbf{v}_i\} \end{aligned}$$

  if $j + 1 < i$
- hence if $j \leq i - 2$,

$$\mathbf{v}_i^\mathsf{T} A (A\mathbf{v}_j) = \mathbf{v}_j^\mathsf{T} A (A\mathbf{v}_i) = 0$$