Lecture 2: Sci. Comp. for DPhil Students

Nick Trefethen, Friday 19.10.18

Today

- I.4 Conjugate gradients
- I.5 Convergence of CG

Good reading for today's material: Trefethen & Bau chap. 38 (pp. 293-302).

Handouts

- Quiz 2
- extra copies of handouts from last lecture
- p. 1 of Hestenes & Stiefel paper
- `cg.m` and `m3_CGconvergence.m`
- Lists of books and journals

Announcements

- Please do the quiz now (not to hand in).
- There is trouble with the web site. We hope to get this fixed soon.
- Circulate sheets for course signup. Online signup is not needed.
- Assignment 1 due 11:00 Tuesday. You are encouraged to write your solutions with `publish`. (Not required, just recommended.)
- Course web page: `https://courses.maths.ox.ac.uk/node/39511/`
- Discuss journals and books list.

I want to emphasize: you need to spend regular time looking at course materials, both handouts and online. For example, the list of books contains many gems — be aware of these books. And get in the habit of downloading my brief lecture notes after each lecture. Buy a cup of coffee and spend half an hour looking over this material at leisure.

I.4 Conjugate gradients

Last lecture we saw that $N = 10^3$ is easy, 10^4 is hard, 10^5 is out of range via $O(N^3)$ ("direct") algorithms such as GE (Gaussian elimination) on ordinary machines. So we seek alternatives, of which the prototype is CG.

History of CG

See Golub & O'Leary, *SIAM Review*, 1989.

Hestenes and Stiefel 1952 (originally independently) `[handout]`

```
  [ Swiss chap, Stiefel,         Just one trouble:
    Feeling gleeful,             He's got a double.
    Announces with radiance:     Soon his destiny's
    Conjugate gradients!         Coupled with Hestenes'. ]
```

For 20 years CG was not so important. The explanation is complexity:

```
            |                      GE
   Work     |                     o                     *  CG
            |                   o                 *
            |                 o             *
            |               o         *
            |             o     *
            |           o *
            |       * o
            | *     o
            |     o
            |   o
            | o
            |_________________________________  dimension N
                      1971
```

As the years have passed, we have moved up this diagram.

“Rediscovery of CG” — John Reid 1971

Hugely successful and important. Google search on “conjugate gradient” gives 923,000 hits.

The first hit is the Wikipedia article. The second is Jonathan Shewchuk’s “An introduction to the conjugate gradient method without the agonizing pain”, which you may enjoy. Shewchuk won the 2003 Wilkinson Software Prize for his software package `TRIANGLE`. Let A ($N \times N$ real matrix) be

```
          SPD = symmetric positive definite.
                    |            |
                      T          |
                 A = A     all eigenvalues > 0

                             or equivalently
                          T
                         x Ax > 0 for all nonzero x
```

Many linear systems that arise in applications can be reduced to SPD form — notably those from self-adjoint PDEs.

Let b by an $N \times 1$ vector. We want to solve $Ax = b$ for x. A is nonsingular, so there is a unique solution. Call it x_*.

Define the **A-norm** of a vector x by $\|x\|_A = (x^T Ax)^{1/2}$.

CG is a recurrence that finds x_* iteratively.

Define x_0 = initial guess = 0 (vector) for simplicity.

CG constructs iterates $x_1, x_2, x_3, ... \to x$.

x_n belongs to the **Krylov space**

$$K = \text{span}(b, Ab, A^2b, \ldots, A^{n-1}b).$$

Define the nth **error** and **residual** vectors by

$$e_n = x_* - x_n, \qquad r_n = Ae_n = b - Ax_n.$$

The CG iterates satisfy a powerful optimality property:

Theorem 1. *Among all $x \in K_n$, x_n is the unique one that minimizes $\|e_n\|_A$.*

Corollary.

$$\|e_0\|_A \geq \|e_1\|_A \geq \|e_2\|_A \geq \cdots .$$

CG iteration

CG comes from ideas in optimization, i.e., minimization, a subject on which we'll say more later in this term. If we define the quadratic form

$$f(x) = \frac{1}{2}x^T Ax - x^T b,$$

then the gradient (vector) of f is

$$\nabla f(x) = Ax - b.$$

$f(x)$ takes a minimum uniquely at x_*, with $\nabla f(x_*) = 0$.   CG searches for x_* iteratively.

We will state the formulas but not derive them.

```
 -----------------------------------------------------------------
|                                                                 |
|   x  = 0 ,   r  = b ,   p  = r                                  |
|    0          0          0    0                                 |
|                                                                 |
|   For n = 1,2,3,...                                             |
|                                                                 |
|              T             T                                    |
|        a = (r    r   ) / (p   A p   )   step length             |
|         n    n-1  n-1       n-1   n-1                           |
|                                                                 |
|        x  = x    + a  p                  approximate solution   |
|         n    n-1    n  n-1                                      |
|                                                                 |
|        r  = r    - a  A p                residual               |
|         n    n-1    n    n-1                                    |
|                                                                 |
|               T         T                                       |
```

```
|        c  = (r  r ) / (r    r   )        improvement this step   |
|         n     n  n      n-1  n-1                                  |
|                                                                   |
|        p  = r  + c  p                     search direction        |
|         n    n    n  n-1                                           |
|                                                                   |
|   end                                                             |
|___________________________________________________________________|
```

```
Note that $A$ enters only via a single matrix-vector multiplication
at each step.  As a result CG takes advantage of sparsity easily.


    [run the first part of m3_CGconvergence.m]
```

I.5 Convergence of CG

The convergence of CG can be analysed beautifully by ideas of approximation theory.

Note that

$$e_0 = x_* - x_0 = x_*, \qquad e_n = x_* - x_n = e_0 - x_n$$

and therefore

$$x_n \in \operatorname{span}(b, Ab, \ldots, A^{n-1}b) = \operatorname{span}(Ae_0, A^2e_0, \ldots, A^ne_0)$$

This implies

$$e_n = p_n(A)e_0$$

where p_n is a polynomial of degree $\leq n$ with $p(0) = 1$.

Thus we can restate Theorem 1 like this:

Theorem 1'. $e_n = p_n(A)e_0$, *where* $\|p_n(A)e_0\|_A =$ *minimum among all polynomials* p_n *of degree* $\leq n$ *with* $p(0) = 1$.

Connection with eigenvalues of A

Let A have eigenvalues (> 0) and orthonormal eigenvectors

$$\lambda_1, \ldots, \lambda_N, \qquad v_1, \ldots, v_N.$$

Expand e_0 in terms of these eigenvectors:

$$e_0 = \sum_{k=1}^{N} a_k v_k.$$

Then

$$e_n = p_n(A)e_0 = \sum_{k=1}^{N} a_k p_n(\lambda_k) v_k.$$

Thus if there exists a polynomial p_n that is small at the eigenvalues, then e_n will be small.

From the definition of $\| \cdot \|_A$, with a little work we derive:

Theorem 2.

$$\frac{\|e_n\|_A}{\|e_0\|_A} \leq \min_{p_n} \max_{\lambda_k} |p_n(\lambda)k)|.$$

where the minimum is over polynomials p_n *of degree* $\leq n$ *with* $p(0) = 1$.

A famous corollary

The **condition number** of an SPD matrix is

$$\kappa(A) = \lambda_{max}(A)/\lambda_{min}(A).$$

By considering Chebyshev polynomials shifted to $[\lambda_{min}, \lambda_{max}]$ we can derive:

Corollary.

$$\frac{\|e_n\|_A}{\|e_0\|_A} \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^n.$$

The right-hand side is approximately

$$2 \exp \left(\frac{-2n}{\sqrt{\kappa(A)}} \right).$$

This implies that to any fixed precision, CG converges in $O(\sqrt{\kappa(A)})$ steps.

`[the rest of m3_CGconvergence.m]`