Prof. A. C. Hansen

Mathematical Tripos Part II: Michaelmas Term 2024

Numerical Analysis – Lecture 18

Approach 4.20 (Minimization of quadratic function) The methods we considered so far for solving $A\boldsymbol{x} = \boldsymbol{b}$, namely Jacobi, Gauss-Seidel, and those with relaxation, fit into the scheme

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + c_k \boldsymbol{d}^{(k)} ,$$

where we were aimed at getting $\rho(H) < 1$ for the iteration matix H. Say, for Jacobi with relaxation, we set $c_k = \omega$ and $\boldsymbol{d}^{(k)} = D^{-1}(\boldsymbol{b} - A\boldsymbol{x}^{(k)})$.

For solving $A\boldsymbol{x} = \boldsymbol{b}$ with a (positive definite) matrix $A > 0$, there is a different approach to constructing good iterative methods. It is based on succesive minimization of the quadratic function

$$F(\boldsymbol{x}^{(k)}) := \|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|_A^2 = \|\boldsymbol{e}^{(k)}\|_A^2 ,$$

since the minimizer is clearly the exact solution. Here, $\|\boldsymbol{y}\|_A := (A\boldsymbol{y}, \boldsymbol{y})^{1/2} := \sqrt{\boldsymbol{y}^T A \boldsymbol{y}}$ is a Euclidean-type distance which is well-defined for $A > 0$. So, at each step k, we are decreasing the A-distance between $\boldsymbol{x}^{(k)}$ and the exact solution $\boldsymbol{x}^*$. Thus, for a symmetric positive definite $A > 0$, we choose an iterative method that provides the descent condition

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + c_k \boldsymbol{d}^{(k)} \quad \Rightarrow \quad F(\boldsymbol{x}^{(k+1)}) < F(\boldsymbol{x}^{(k)}) . \tag{4.5}$$

An equivalent approach is to minimize the quadratic function

$$F_1(\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{b} ,$$

which attains its minimum when $\nabla F_1(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b} = 0$, and which does not involve the unknown $\boldsymbol{x}^*$. It is easy to check that $F_1(\boldsymbol{x}) = \frac{1}{2}F(\boldsymbol{x}) - \frac{1}{2}c$, where $c = \boldsymbol{x}^{*T} A \boldsymbol{x}^*$ is a constant independent of k, hence equivalence.

Example 4.21 Both the Jacobi and the Gauss–Seidel methods satisfy (4.5), precisely

$$(A\boldsymbol{e}^{(k+1)}, e^{(k+1)}) = (A\boldsymbol{e}^{(k)}, \boldsymbol{e}^{(k)}) - (C\boldsymbol{y}^{(k)}, \boldsymbol{y}^{(k)}) < (A\boldsymbol{e}^{(k)}, \boldsymbol{e}^{(k)}) ,$$

where for Gauss-Seidel: $C = D > 0, \quad \boldsymbol{y}^{(k)} := (L_0 + D)^{-1} A\boldsymbol{e}^{(k)}$;
and for Jacobi: $C = 2D - A > 0, \quad \boldsymbol{y}^{(k)} := D^{-1} A\boldsymbol{e}^{(k)}$.

Method 4.22 (A-orthogonal projection) Next, we strengthen the descent condition (4.5), namely given $\boldsymbol{x}^{(k)}$ and some $\boldsymbol{d}^{(k)}$ (called a *search direction*), we will seek $\boldsymbol{x}^{(k+1)}$ from the set of vectors on the line $\ell = \{\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\}_{\alpha \in \mathbb{R}}$ such that it makes the value of $F(\boldsymbol{x}^{(k+1)})$ not just smaller than $F(\boldsymbol{x}^{(k)})$, but as small as possible (with respect to this set), namely

$$\boldsymbol{x}^{(k+1)} := \arg\min_{\alpha} F(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) . \tag{4.6}$$

Lemma 4.23 *The minimizer in (4.6) is given by the formula*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)} , \qquad \alpha_k = \frac{(\boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)})}{(A\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)})} . \tag{4.7}$$

(This choice of α_k is referred to as exact line search.)

Proof. From the definition of F, it follows that in (4.6) we should choose the point $\boldsymbol{x}^{(k+1)} \in \ell$ that minimizes the A-distance between $\boldsymbol{x}^*$ and the points $\boldsymbol{y} \in \ell$. Geometrically, it is clear that the minimum occurs when $\boldsymbol{x}^{(k+1)}$ is the A-orthogonal projection of $\boldsymbol{x}^*$ onto the line $\ell = \{\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\}$, i.e., when

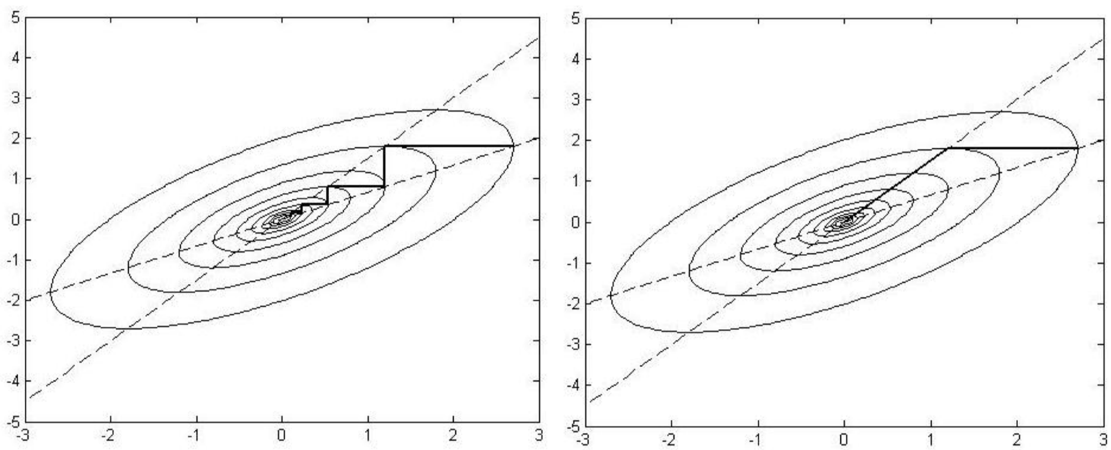
$$\boldsymbol{x}^* - \boldsymbol{x}^{(k+1)} \perp_A \boldsymbol{d}^{(k)} \quad \Rightarrow \quad A(\boldsymbol{x}^* - \boldsymbol{x}^{(k+1)}) \perp \boldsymbol{d}^{(k)} \quad \Rightarrow \quad \boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k A\boldsymbol{d}^{(k)} \perp \boldsymbol{d}^{(k)} .$$

This gives expression for α_k in (4.7). □

Method 4.24 (The steepest descent method) This method takes $\boldsymbol{d}^{(k)} = -\nabla F_1(\boldsymbol{x}^{(k)}) = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ for every k, the reason being that, locally, the negative gradient of a quadratic function shows the direction of the (locally) steepest descent at a given point. Thus, the iterations have the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k(\boldsymbol{b} - A\boldsymbol{x}^{(k)}), \qquad k \geq 0\,. \tag{4.8}$$

It can be proved that the sequence $(\boldsymbol{x}^{(k)})$ converges to the solution $\boldsymbol{x}^*$ of the system $A\boldsymbol{x} = \boldsymbol{b}$ as required, but usually the speed of convergence is rather slow. The reason is that the iteration (4.8) decreases the value of $F(\boldsymbol{x}^{(k+1)})$ locally, relatively to $F(\boldsymbol{x}^{(k)})$, but the global decrease, with respect to $F(\boldsymbol{x}^{(0)})$, is often not that large. The use of *conjugate directions* provides a method with a global minimization property.



(a) Worst case scenario of steepest descent

(b) Conjugate gradient method applied to the same problem as in (a)

Conjugate directions Let's revisit equation (4.7) for a general direction $\boldsymbol{d}$ (i.e., not necessarily equal to the negative gradient). Assume $\boldsymbol{x} = \boldsymbol{x}^{(k)}$, and let $\boldsymbol{e}^{(k)} = \boldsymbol{x}^* - \boldsymbol{x}^{(k)}$ be the error and $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)} = A\boldsymbol{e}^{(k)}$ be the residual. Then we can write $\langle \boldsymbol{r}^{(k)}, \boldsymbol{d}\rangle = \langle \boldsymbol{e}^{(k)}, \boldsymbol{d}\rangle_A$, and so for a general search direction $\boldsymbol{d}$ with an exact line search, the iterate takes the form $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \frac{\langle \boldsymbol{e}^{(k)}, \boldsymbol{d}\rangle_A}{\langle \boldsymbol{d}, \boldsymbol{d}\rangle_A}\boldsymbol{d}$. By substracting $\boldsymbol{x}^*$, the iterates in terms of the error $\boldsymbol{e}^{(k+1)}$ are given by:

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{e}^{(k)} - \frac{\langle \boldsymbol{e}^{(k)}, \boldsymbol{d}\rangle_A}{\langle \boldsymbol{d}, \boldsymbol{d}\rangle_A}\boldsymbol{d}. \tag{4.9}$$

Geometrically, this means that $\boldsymbol{e}^{(k+1)}$ is the projection of $\boldsymbol{e}^{(k)}$ onto the hyperplane that is A-orthogonal to $\boldsymbol{d}$, i.e., we have

$$\langle \boldsymbol{e}^{(k+1)}, \boldsymbol{d}\rangle_A = 0. \tag{4.10}$$

Definition 4.25 (Conjugate directions) The vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are *conjugate* with respect to a symmetric positive definite matrix A if they are nonzero and A-orthogonal: $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_A := \langle \boldsymbol{u}, A\boldsymbol{v}\rangle = 0$.

The observation above allows us to prove the following important result.

Theorem 4.26 *Let $\boldsymbol{d}^{(0)}, \boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(n-1)}$ be n nonzero pairwise conjugate directions, and consider the sequence of iterates*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}, \qquad \alpha_k = \frac{\langle \boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)}\rangle}{\langle \boldsymbol{d}^{(k)}, A\boldsymbol{d}^{(k)}\rangle}.$$

Let $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ be the residual. Then for each $k = 1, \ldots, n$, $\boldsymbol{r}^{(k)}$ is orthogonal to $\operatorname{span}\{\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k-1)}\}$. In particular $\boldsymbol{r}^{(n)} = 0$.

Proof. Since $\boldsymbol{r}^{(k)} = A\boldsymbol{e}^{(k)}$, it suffices to show that $\boldsymbol{e}^{(k)}$ is A-orthogonal to $\mathrm{span}\{\boldsymbol{d}^{(0)}, \dots, \boldsymbol{d}^{(k-1)}\}$. The proof is by induction on k. For $k = 0$ there is nothing to prove. Assume the statement is true for $k \geq 0$, and consider the equation (4.9) (with $\boldsymbol{d} = \boldsymbol{d}^{(k)}$). From the induction hypothesis, and the fact that the $\boldsymbol{d}^{(i)}$ are pairwise conjugate directions, we see that $\boldsymbol{e}^{(k+1)}$ is A-orthogonal to $\boldsymbol{d}^{(0)}, \dots, \boldsymbol{d}^{(k-1)}$. Furthermore, we have already seen in (4.10) that $\langle \boldsymbol{e}^{(k+1)}, \boldsymbol{d}^{(k)} \rangle_A = 0$. Thus this shows that $\boldsymbol{e}^{(k+1)}$ is A-orthogonal to $\boldsymbol{d}^{(0)}, \dots, \boldsymbol{d}^{(k)}$ as desired. □

So, if a sequence $(\boldsymbol{d}^{(k)})$ of conjugate directions is at hands, we have an iterative procedure with good approximation properties.

The (A-orthogonal) basis of conjugate directions is constructed by A-orthogonalization of the sequence $\{\boldsymbol{r}_0, A\boldsymbol{r}_0, A^2\boldsymbol{r}_0, ..., A^{n-1}\boldsymbol{r}_0\}$ with $\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$. This is done in the way similar to orthogonalization of the monomial sequence $\{1, x, x^2, ..., x^{n-1}\}$ using a recurrence relation.

Remark 4.27 It is possible to extend the methods for solving $A\boldsymbol{x} = \boldsymbol{b}$ with symmetric positive definite A to any other matrices by a simple trick. Suppose we want to solve $B\boldsymbol{x} = \boldsymbol{c}$, where $B \in \mathbb{R}^{n \times n}$ is nonsingular. We can convert the above system to the symmetric and positive definite setting by defining $A = B^T B$, $\boldsymbol{b} = B^T \boldsymbol{c}$ and then solving $A\boldsymbol{x} = \boldsymbol{b}$ with the conjugate gradient algorithm (or any other method for positive definite A).