Numerical Analysis - Part II

Anders C. Hansen

Lecture 19

Iterative methods for linear algebraic systems

Linear systems in elliptic PDEs

As we have seen in the previous sections linear systems $A\mathbf{x} = \mathbf{b}$, where A is a real symmetric positive (negative) definite matrix, frequently occur in numerical methods for solving elliptic partial differential equations.

Poisson's equation on a square

A typical example we already encountered is Poisson's equation on a square where the *five-point formula* approximation yields an $n \times n$ system of linear equations with $n = m^2$ unknowns $u_{p,q}$:

$$u_{p-1,q} + u_{p+1,q} + u_{p,q-1} + u_{p,q+1} - 4u_{p,q} = h^2 f(ph, qh) \qquad (1)$$

In the *natural ordering*, when the grid points are arranged by columns, A is the following block tridiagonal matrix:

$$A = \begin{bmatrix} B & I & & & \\ I & B & I & & \\ & \ddots & \ddots & \ddots & \\ & & I & B & I \\ & & & I & B \end{bmatrix}, \qquad B = \begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix}. \qquad (2)$$

The matrix A is symmetric and negative definite

Lemma 1

For any ordering of the grid points, the matrix A of the system (1) is symmetric and negative definite.

Poisson's equation on a square

Note that when p or q is equal to 1 or m, then the values $u_{0,q}$, $u_{p,0}$ or $u_{p,m+1}$, $u_{m+1,q}$ are known boundary values and they should be moved to the right-hand side, thus leaving fewer unknowns on the left.

For any ordering of the grid points (ph, qh) we have shown in Lemma 1 that the matrix A of this linear system is symmetric and negative definite.

Corollary 2

For the linear system (1), for any ordering of the grid, both Jacobi and Gauss-Seidel methods converge.

Proof. By Lemma 1, A is symmetric and negative definite, hence convergence of Gauss-Seidel. To prove convergence of the Jacobi method, we need negative definiteness of the matrix $2D - A$, and that follows by the same arguments as in Lemma 1: recall that the proof operates with the modulus of the off-diagonal elements and does not depend on their sign. □

Relaxation

It is often possible to improve the efficiency of the splitting method by *relaxation*. Specifically, instead of letting $(A - B)\mathbf{x}^{(k+1)} = -B\mathbf{x}^{(k)} + \mathbf{b}$, we let

$$(A-B)\widehat{\mathbf{x}}^{(k+1)} = -B\mathbf{x}^{(k)}+\mathbf{b}, \quad \text{and then} \quad \mathbf{x}^{(k+1)} = \omega\widehat{\mathbf{x}}^{(k+1)}+(1-\omega)\mathbf{x}^{(k)}$$

where ω is a real constant called the *relaxation parameter*.

Note that $\omega = 1$ corresponds to the standard "unrelaxed" iteration. Good choice of ω leads to a smaller spectral radius of the iteration matrix (compared with the "unrelaxed" method), and the smaller the spectral radius, the faster the iteration converges.

Relaxation

To this end, let us express the relaxation iteration matrix H_ω in terms of $H = -(A-B)^{-1}B$. We have

$$\widehat{\mathbf{x}}^{(k+1)} = H\mathbf{x}^{(k)} + \mathbf{v} \quad \Rightarrow \quad \begin{aligned} \mathbf{x}^{(k+1)} &= \omega\widehat{\mathbf{x}}^{(k+1)} + (1-\omega)\mathbf{x}^{(k)} \\ &= \omega H\mathbf{x}^{(k)} + (1-\omega)\mathbf{x}^{(k)} + \omega\mathbf{v}, \end{aligned}$$

hence

$$H_\omega = \omega H + (1-\omega)I.$$

It follows that the spectra of H_ω and H are related by the rule $\lambda_\omega = \omega\lambda + (1-\omega)$, therefore one may try to choose $\omega \in \mathbb{R}$ to minimize

$$\rho(H_\omega) = \max\{|\omega\lambda + (1-\omega)| : \lambda \in \sigma(H)\}.$$

Relaxation

In general, $\sigma(H)$ is unknown, but often we have some information about it which can be utilized to find a "good" (rather than "best") value of ω. For example, suppose that it is known that $\sigma(H)$ is real and resides in the interval $[\alpha, \beta]$ where $-1 < \alpha < \beta < 1$. In that case we seek ω to minimize

$$\max\{|\omega\lambda + (1-\omega)| : \lambda \in [\alpha, \beta]\}.$$

Relaxation

It is readily seen that, for a fixed $\lambda < 1$, the function $f(\omega) = \omega\lambda + (1-\omega)$ is decreasing, therefore, as ω increases (decreases) from 1 the spectrum of H_ω moves to the left (to the right) of the spectrum of H. It is clear that the optimal location of the spectrum $\sigma(H_\omega)$ (or of the interval $[\alpha_\omega, \beta_\omega]$ that contains $\sigma(H_\omega)$) is the one which is centralized around the origin:

$$-[\omega\alpha+(1-\omega)] = \omega\beta+(1-\omega) \quad \Rightarrow \quad \begin{aligned} \omega_{\mathrm{opt}} &= \frac{2}{2-(\alpha+\beta)} \\ -\alpha_{\omega_{\mathrm{opt}}} &= \beta_{\omega_{\mathrm{opt}}} = \frac{\beta-\alpha}{2-(\alpha+\beta)}\,. \end{aligned}$$

Minimization of quadratic function

The methods we considered so far for solving $A\mathbf{x} = \mathbf{b}$, namely Jacobi, Gauss-Seidel, and those with relaxation, fit into the scheme

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

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 $\mathbf{d}^{(k)} = D^{-1}(\mathbf{b} - A\mathbf{x}^{(k)})$.
For solving $A\mathbf{x} = \mathbf{b}$ 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

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

since the minimizer is clearly the exact solution. Here, $\|\mathbf{y}\|_A := (A\mathbf{y}, \mathbf{y})^{1/2} := \sqrt{\mathbf{y}^T A\mathbf{y}}$ is a Euclidean-type distance which is well-defined for $A > 0$.

Minimization of quadratic function

So, at each step k, we are decreasing the A-distance between $\mathbf{x}^{(k)}$ and the exact solution $\mathbf{x}^*$. Thus, for a symmetric positive definite $A > 0$, we choose an iterative method that provides the descent condition

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + c_k \mathbf{d}^{(k)} \quad \Rightarrow \quad F(\mathbf{x}^{(k+1)}) < F(\mathbf{x}^{(k)}) \,. \tag{3}$$

Minimization of quadratic function

An equivalent approach is to minimize the quadratic function

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

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

Quadratic function – Jacobi and Gauss–Seidel

Both the Jacobi and the Gauss–Seidel methods satisfy (3), precisely

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

where for Gauss-Seidel: $C = D > 0$, $\mathbf{y}^{(k)} := (L_0 + D)^{-1} A\mathbf{e}^{(k)}$;

and for Jacobi: $C = 2D - A > 0$, $\mathbf{y}^{(k)} := D^{-1} A\mathbf{e}^{(k)}$.

A-orthogonal projection

A-orthogonal projection method: Next, we strengthen the descent condition (3), namely given $\mathbf{x}^{(k)}$ and some $\mathbf{d}^{(k)}$ (called a *search direction*), we will seek $\mathbf{x}^{(k+1)}$ from the set of vectors on the line $\ell = \{\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}\}_{\alpha \in \mathbb{R}}$ such that it makes the value of $F(\mathbf{x}^{(k+1)})$ not just smaller than $F(\mathbf{x}^{(k)})$, but as small as possible (with respect to this set), namely

$$\mathbf{x}^{(k+1)} := \arg\min_{\alpha} F(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}) \,. \tag{4}$$

Lemma 3

The minimizer in (4) is given by the formula

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\,, \qquad \alpha_k = \frac{(\mathbf{r}^{(k)}, \mathbf{d}^{(k)})}{(A\mathbf{d}^{(k)}, \mathbf{d}^{(k)})}\,. \tag{5}$$

A-orthogonal projection

Proof. From the definition of F, it follows that in (4) we should choose the point $\mathbf{x}^{(k+1)} \in \ell$ that minimizes the A-distance between $\mathbf{x}^*$ and the points $\mathbf{y} \in \ell$. Geometrically, it is clear that the minimum occurs when $\mathbf{x}^{(k+1)}$ is the A-orthogonal projection of $\mathbf{x}^*$ onto the line $\ell = \{\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}\}$, i.e., when

$$\begin{aligned} \mathbf{x}^* - \mathbf{x}^{(k+1)} \perp_A \mathbf{d}^{(k)} \quad &\Rightarrow \quad A(\mathbf{x}^* - \mathbf{x}^{(k+1)}) \perp \mathbf{d}^{(k)} \\ &\Rightarrow \quad \mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)} \perp \mathbf{d}^{(k)} . \end{aligned}$$

This gives expression for α_k in (5). □

The steepest descent method

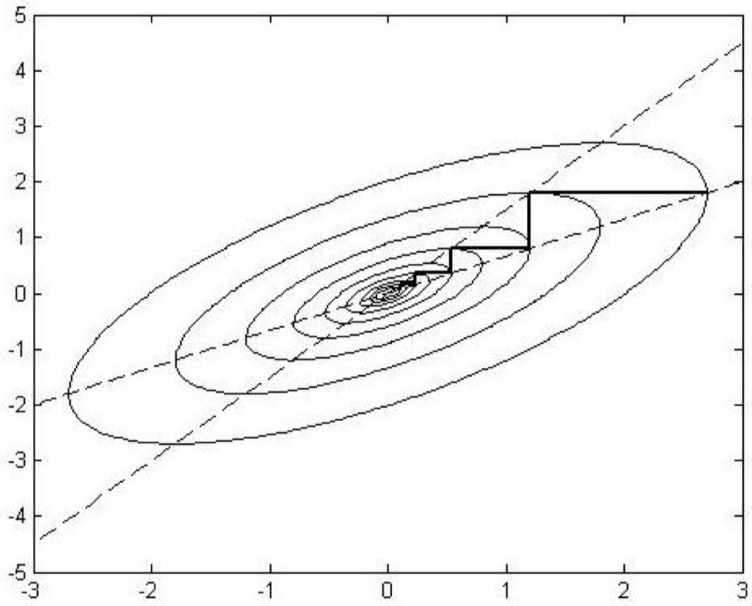
The steepest descent method: This method takes $\mathbf{d}^{(k)} = -\nabla F_1(\mathbf{x}^{(k)}) = \mathbf{b} - A\mathbf{x}^{(k)}$ for every k, the reason being that, locally, the negative gradient of a quadratic function shows the direction of the (locally) steepest descent at a given point. Thus, the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

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k(\mathbf{b} - A\mathbf{x}^{(k)}), \qquad k \geq 0\,. \tag{6}$$

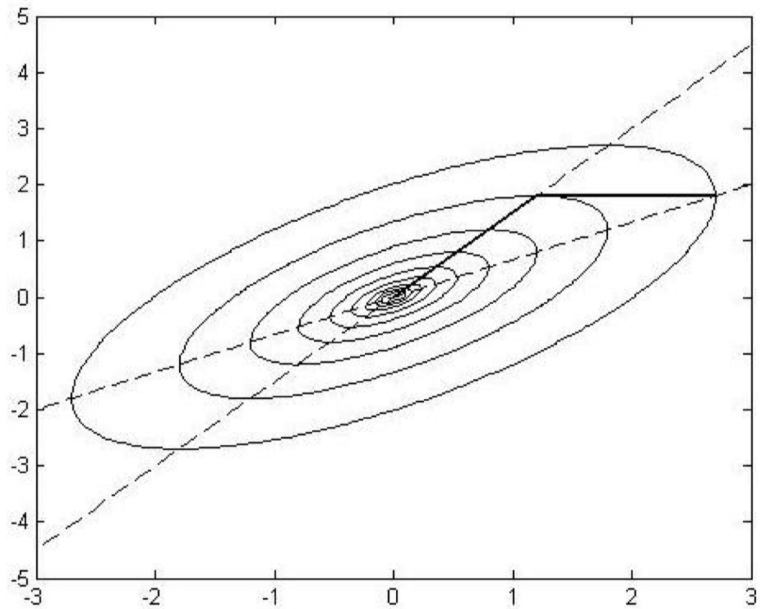
It can be proved that the sequence $(\mathbf{x}^{(k)})$ converges to the solution $\mathbf{x}^*$ of the system $A\mathbf{x} = \mathbf{b}$ as required, but usually the speed of convergence is rather slow.

The reason is that the iteration (6) decreases the value of $F(\mathbf{x}^{(k+1)})$ locally, relatively to $F(\mathbf{x}^{(k)})$, but the global decrease, with respect to $F(\mathbf{x}^{(0)})$, is often not that large. The use of *conjugate directions* provides a method with a global minimization property.

Steepest descent and conjugate gradient



(a) Worst case scenario of steepest descent



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

Conjugate directions

Definition 4 (Conjugate directions)

The vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are *conjugate* with respect to a symmetric positive definite matrix A if they are nonzero and A-orthogonal: $(\mathbf{u}, \mathbf{v})_A := (A\mathbf{u}, \mathbf{v}) = 0$.

Theorem 5 (Non-examinable)

Given $A \in \mathbb{R}^{n\times n}$, $A > 0$, *let* $\{\mathbf{d}^{(k)}\}_{k=0}^{n-1}$ *be a set of the conjugate directions, i.e.,* $(A\mathbf{d}^{(k)}, \mathbf{d}^{(i)}) = 0$ *for* $i < k$. *Then the value of* $F(\mathbf{x}^{(m+1)})$ *obtained through step-by-step minimization for each* $k = 0..m$ *as described in (5) coincides with the minimum of* $F(\mathbf{y})$ *taken over all* $\mathbf{y} = \mathbf{x}^{(0)} + \sum_{k=0}^{m} c_k \mathbf{d}^{(k)}$ *simultaneously, namely*

$$\arg \min_{c_0,\ldots,c_m} F(\mathbf{y}) = \mathbf{x}^{(m+1)} = \mathbf{x}^{(0)} + \sum_{k=0}^{m} \alpha_k \mathbf{d}^{(k)} .$$

Proof. Again, it is clear geometrically that the minimal A-distance between the exact solution $\mathbf{x}^*$ and the points $\mathbf{y}$ on the plane $\mathcal{P} := \{\mathbf{x}^{(0)} + \sum_{k=0}^{m} c_k \mathbf{d}^{(k)} : c_k \in \mathbb{R}\}$ is attained when $\mathbf{x}^{(m+1)} \in \mathcal{P}$ is the A-orthogonal projection of $\mathbf{x}^*$ onto $\mathcal{P}$, i.e.,

$$\arg\min_{\mathbf{y}\in\mathcal{P}} F(\mathbf{y}) = \mathbf{x}^{(m+1)} \quad \Leftrightarrow \quad \mathbf{x}^* - \mathbf{x}^{(m+1)} \perp_A \{\mathbf{d}^{(k)}\}_{k=0}^{m}.$$

It can be shown then, that (for conjugate $\{\mathbf{d}^{(k)}\}$) the latter conditions provide expressions for α_k as given in (5). □

Conjugate gradient - Warm up

So, if a sequence $(\mathbf{d}^{(k)})$ 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 $\{\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, ..., A^{n-1}\mathbf{r}_0\}$ with $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$. 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 6

It is possible to extend the methods for solving $A\mathbf{x} = \mathbf{b}$ with symmetric positive definite A to any other matrices by a simple trick. Suppose we want to solve $B\mathbf{x} = \mathbf{c}$, 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 $A = B^T B$, $\mathbf{b} = B^T \mathbf{c}$ and then solving $A\mathbf{x} = \mathbf{b}$ with the conjugate gradient algorithm (or any other method for positive definite A).

The conjugate gradient method

Here it is.

(A) For any initial vector $\mathbf{x}^{(0)}$, set $\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$;

(B) For $k \geq 0$, calculate $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$ and the residual

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)}, \quad \text{with}$$

$$\alpha_k := \{\, \mathbf{r}^{(k+1)} \perp \mathbf{d}^{(k)} \} = \frac{(\mathbf{r}^{(k)}, \mathbf{d}^{(k)})}{(A\mathbf{d}^{(k)}, \mathbf{d}^{(k)})}, \quad k \geq 0\,. \tag{7}$$

(C) For the same k, the next conjugate direction is the vector

$$\mathbf{d}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{d}^{(k)}, \quad \text{with}$$

$$\beta_k := \{\, \mathbf{d}^{(k+1)} \perp A\mathbf{d}^{(k)} \} = -\frac{(\mathbf{r}^{(k+1)}, A\mathbf{d}^{(k)})}{(\mathbf{d}^{(k)}, A\mathbf{d}^{(k)})}, \quad k \geq 0\,. \tag{8}$$