Chapter 7

PETROV-GALERKIN METHODS

7.1 Energy Norm Minimization
7.2 Residual Norm Minimization
7.3 General Projection Methods

7.1 Energy Norm Minimization

Saad, Sections 5.3.1, 5.2.1a.

7.1.1 Methods based on approximate minimization

7.1.2 Steepest descent

We assume that $A$ is symmetric positive definite:

$$A = Q\Lambda Q^T, \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_i > 0.$$ 

Unlike SOR, there are no parameters to choose.

7.1.1 Methods based on approximate minimization

Minimization methods minimize some objective function $\phi(x) = \phi(x_1, x_2, \ldots, x_n)$:
Possible objective functions:
(i) norm of error \( \| x - A^{-1} b \|_2 \),
(ii) norm of residual \( \| A(x - A^{-1} b) \|_2 = \| Ax - b \|_2 \),
(iii) energy norm of error \( \| A^{1/2}(x - A^{-1} b) \|_2 = \sqrt{(x - A^{-1} b)^T A(x - A^{-1} b)} \)

where \( A^{1/2} = Q \Lambda^{1/2} Q^T \) is the unique symmetric positive definite square root of \( A \). Each of these is a weighted 2-norm of the rotated error:

\[
\begin{align*}
\| A(x - A^{-1} b) \|_2 &= \| \Lambda Q^T (x - A^{-1} b) \|_2 \\
\| A^{1/2}(x - A^{-1} b) \|_2 &= \| \Lambda^{1/2} Q^T (x - A^{-1} b) \|_2 \\
\| x - A^{-1} b \|_2 &= \| Q^T (x - A^{-1} b) \|_2 \\
\end{align*}
\]

least desirable weights \( \lambda_1, \lambda_2, \ldots, \lambda_n \)

weights \( \lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_n^{1/2} \)

weights 1, 1, \ldots, 1

most desirable

We denote the energy norm by \( ||| w ||| := \sqrt{w^T A w} \). In some applications it measures the square root of energy or power. We would like to minimize \( \| x - A^{-1} b \|_2 \) but any minimization method would need to know \( A^{-1} b \). However,

\[
||| x - A^{-1} b |||^2 = (x - A^{-1} b)^T A (x - A^{-1} b) = x^T Ax - 2 x^T b + b^T A^{-1} b
\]

which attains its minimum at the same value of \( x \) as does

\[
\phi(x) = \frac{1}{2} x^T Ax - x^T b
\]

and this is cheap to compute for given \( x \).

Note

unit circle in energy norm \( = \{ w \mid ||| w ||| = 1 \} \)

\( = \{ w \mid \| \Lambda^{1/2} Q^T w \|_2 = 1 \} \) change variable \( \Lambda^{1/2} Q^T w =: u \)

\( = \{ Q \Lambda^{-1/2} u \mid \| u \|_2 = 1 \} \)

\( = Q \Lambda^{-1/2} \cdot ( \text{unit circle} ) \)

unit circle

\( \Lambda^{-1/2} \cdot ( \text{unit circle} ) \)

\( Q \Lambda^{-1/2} \cdot ( \text{unit circle} ) \)

156
\[
\text{elongation } = \lambda_{\min}^{-1/2} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\|A\|_2\|A^{-1}\|_2} = \sqrt{\kappa_2(A)}
\]

The contours of \(\|x - A^{-1}b\|\) are curves of equal error in energy norm.

For the 5-point difference operator on unit square
\[
\frac{1}{h^2} \begin{pmatrix}
-u_{i-1} & -u_{i+N-1} & -4u_i & -u_{i+1} \\
-\ u_{i-N+1}
\end{pmatrix}, \quad h = \frac{1}{N}.
\]

The eigenvalues are \((2N \sin \frac{2\pi}{2N})^2 + (2N \sin \frac{2\pi}{2N})^2, p, q = 1, 2, \ldots, N - 1.\)
\[
\lambda_{\min} \approx 2\pi^2 \quad \lambda_{\max} \approx 8N^2
\]
\[
\sqrt{\kappa_2(A)} \approx \frac{2}{\pi} N
\]

generic minimization method

\[
x_0 = \text{initial guess}; \quad \{\text{subscripts index iterates not components}\}
\]
\[
\text{for } i = 0, 1, 2, \ldots \text{ do } \{ \quad \text{search direction} \}
\]
\[
\quad \text{choose a direction } p_i; \quad \text{choose a distance } \alpha_i; \quad \text{choose a direction } p_i; \quad x_{i+1} = x_i + \alpha_i p_i;
\]
\[
\text{end for}
\]

One could choose \(\alpha_i\) to minimize \(\phi(x_i + \alpha p_i)\). We have
\[
\frac{d}{d\alpha} \phi(x_i + \alpha p_i) = \frac{d}{d\alpha} \left( \frac{\alpha^2}{2} p_i^T A p_i + \alpha (p_i^T A x_i - p_i^T b) + \frac{1}{2} x_i^T A x_i - x_i^T b \right)
\]
\[
= \alpha p_i^T A p_i - p_i^T r_i
\]
where \( r_i = b - Ax_i \), the residual. Hence

\[
\alpha_i = \frac{p_i^T r_i}{p_i^T A p_i}
\]

Note that

\[
r_{i+1} = r_i - \alpha_i A p_i,
\]

which is very cheap because we need to compute \( A p_i \) anyway to get \( \alpha_i \). Also note

\[
x_{i+1} = x_i + \frac{p_i p_i^T}{p_i^T A p_i} \cdot r_i.
\]

The matrix with the underbrace is a rank one approximation to \( A^{-1} \).

**cyclic coordinate descent** Use search directions \( e_1, e_2, \ldots, e_n, e_1, e_2, \ldots, e_n, \ldots \):

\[
\begin{align*}
x_1 &= x_0 + e_1 \frac{e_1^T r_0}{a_{11}}, \\
x_2 &= x_1 + e_2 \frac{e_2^T r_1}{a_{22}}, \\
&\vdots \\
x_{\text{new}} &= x_{\text{old}} + e_i \frac{e_i^T r_{\text{old}}}{a_{ii}}.
\end{align*}
\]

For descent in the \( i \)th direction only the \( i \)th component of residual is computed and only the \( i \)th component is updated. This is part of a Gauss-Seidel sweep—the relaxation of the \( i \)th unknown.

\( n \) steps of cyclic coordinate descent \( \equiv 1 \) sweep of Gauss-Seidel
You can see why GS always converges if \( A \) is symmetric positive definite. Also, you can see the reason for overrelaxation, i.e., for multiplying the correction by \( \omega > 1 \). The graph below shows that

![Graph showing phi(x_{old} + alpha p_{old})](image)

the objective function is reduced if we use \( \omega \alpha_k, 0 < \omega < 2 \).

### 7.1.2 Steepest descent

What is the direction of steepest descent? Consider

\[
\phi(x_{\text{old}} + \varepsilon p)
\]

where \( \varepsilon > 0 \) is infinitesimal and \( ||p||_2 \) is fixed. We have

\[
\phi(x_{\text{old}} + \varepsilon p) = \frac{1}{2} (x_{\text{old}} + \varepsilon p)^T A(x_{\text{old}} + \varepsilon p) - (x_{\text{old}} + \varepsilon p)^T b
\]

\[
= \frac{1}{2} x_{\text{old}}^T A x_{\text{old}} - x_{\text{old}}^T b - \varepsilon p^T (b - A x_{\text{old}}) + O(\varepsilon^2).
\]

We can get greatest decrease in \( p^T (b - A x_{\text{old}}) = p^T r \) for fixed \( ||p||_2 \)

by choosing \( p = b - A x_{\text{old}} = r_{\text{old}} \). I.e., the residual points in the direction of steepest descent:

\( x_0 = \) initial guess; \( r_0 = b - A x_0 \);

for \( i = 0, 1, 2, \ldots \) do {
  \( \alpha_i = \frac{r_i^T r_i}{r_i^T A r_i} \);
  \( x_{i+1} = x_i + \alpha_i r_i \);
  \( r_{i+1} = r_i - \alpha_i A r_i \);
}

159
Convergence can be slow if $\kappa_2(A)$ is large.

The trouble with these methods is that when we minimize in direction $p_i$, we lose the minimization in directions $p_1, p_2, \ldots, p_{i-1}$. This can be avoided if we use conjugate directions

$$p_i^T A p_j = 0 \text{ if } i \neq j.$$ 

The conjugate gradient method constructs conjugate directions\(^1\) from the gradients $r_0, r_1, r_2, \ldots$. It is the subject of Section 8.4.

**Review questions**

1. Define the energy norm associated with a symmetric positive definite matrix $A$.

2. Solving $Ax = b$ where $A$ is symmetric positive definite is equivalent to minimizing what readily computable scalar function of $x$? Relate this to the energy norm of the error.

3. What is the spectral condition number $\kappa_2(A)$ of a symmetric positive definite matrix $A$?

4. Given a search direction $p_i$ and an approximation $x_i$ to the value $x^*$ that minimizes $\phi(x) = \frac{1}{2}x^T A x - b^T x$ with $A$ symmetric positive definite, what is the best choice for $x_{i+1}$?

5. How are search directions chosen in cyclic coordinate descent?

6. What known method do we get if we apply cyclic coordinate descent to objective function $\phi(x) = \frac{1}{2}x^T A x - b^T x$ where $A$ is symmetric positive definite?

7. For objective function $\phi(x) = \frac{1}{2}x^T A x - b^T x$ where $A$ is symmetric positive definite, what is the direction of steepest descent?

\(^1\)Often $(x, y)_A = x^T A y$ is called the energy inner product.
8. On what property of the matrix does the rate of convergence of steepest descent depend?

Exercises

1. Let

\[
A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.
\]

Plotted below are the level curves (contours) of the function \(\phi(x) = \frac{1}{2}x^TAx - b^Tx\) where \(x = [x_1 \ x_2]^T\) is variable. Draw the straight line

\[a_{11}x_1 + a_{12}x_2 = b_1\]

and the straight line

\[a_{21}x_1 + a_{22}x_2 = b_2.\]

Explain in words how you constructed these two lines. Recall that relaxing the \(i\)th variable so as to satisfy the \(i\)th equation is equivalent to minimizing \(\phi(x)\) in the direction of the \(i\)th coordinate axis.

2. Draw level curves for \(|||x - A^{-1}b|||\) where

\[
A = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 3 \\ 0 \end{bmatrix}.
\]

Model your solution after page 156 of notes. Label various key quantities and do a reasonably careful drawing.

3. Show that in the generic minimization method if \(\alpha_i\) is chosen to minimize \(\phi(x_i + \alpha p_i)\), then \(r_{i+1}\) is orthogonal to \(p_i\). Interpret this geometrically.
7.2 Residual Norm Minimization

Saad, Sections 5.3.3, 5.3.2, 5.2.1b.

For a general matrix only the residual norm $\|b - Ax\|$ can serve as an objective function for minimization. This attains its minimum at the same value of $x$ as does

$$\phi(x) = \frac{1}{2} x^T (A^T A)x - x^T (A^T b).$$

This is the objective function for minimizing the energy norm for the normal equations $(A^T A)x = A^T b$.

A generic minimization method with an analytic line search takes the form

$$x_0 = \text{initial guess};$$
$$r_0 = b - Ax_0;$$

for $i = 0, 1, 2, \ldots$ do {
    choose a direction $p_i$;
    $\alpha_i = ((Ap_i)^T r_i)/((Ap_i)^T Ap_i)$;
    $x_{i+1} = x_i + \alpha_i p_i$;
    $r_{i+1} = r_i - \alpha_i Ap_i$;
}

7.1.1 Residual Norm Steepest Descent

7.1.2 Minimal Residual

7.2.1 Residual Norm Steepest Descent

At a given point $x_i$ the direction of steepest descent for the residual norm is

$$p_i = A^T r_i,$$

leading to an algorithm requiring a multiplication by $A$ and by $A^T$ at each step.

7.2.2 Minimal Residual

It is often the case that the matrix $A$ is available only implicitly via a function call and it is inconvenient to request a matrix-vector product with $A^T$. This is avoided in the minimal residual method by using

$$p_i = r_i$$

for the search direction. The disadvantage is that convergence for an arbitrary nonsingular $A$ is no longer guaranteed. Convergence, however, can be proved for the case where

$$A + A^T$$

is s.p.d.
Review questions

1. The 2-norm of the residual for $Ax = b$ is equal to the energy norm for what system of linear equations?

2. Given a search direction $p_i$ and an approximation $x_i$ to the value $x^*$ that minimizes $\|b - Ax\|_2$ with $A$ nonsingular, what is the best choice for $x_{i+1}$?

3. For the (square of the) residual for $Ax = b$, what is the direction of steepest descent?

4. How are search directions chosen in the minimal residual method?

5. What is the disadvantage of the residual norm steepest descent method compared to the minimal residual method?

6. For which nonsymmetric matrices $A$ can it be proved that the residual norm steepest descent method converges to the solution of $Ax = b$?

7. For which nonsymmetric matrices $A$ can it be proved that the minimal residual method converges to the solution of $Ax = b$?

7.3 General Petrov-Galerkin Methods

Saad, Section 5.1. (i) The term “projection method” is unhelpful (a better term is “restriction”); it is not used in the 1997 book by Anne Greenbaum nor in the 2003 book by van der Vorst. (ii) The statement bridging pages 130 and 131 does not seem correct except in a very contrived way.

Consider a method based on minimizing the (square of the) energy norm

$$(b - Ax)^T A^{-1} (b - Ax), \quad x \in x_0 + \mathcal{K}$$

where $A$ is symmetric positive definite and $\mathcal{K}$ is the span of one or several search directions. Let $\tilde{x}$ be the minimizer, let $r = b - A\tilde{x}$, and consider another feasible value $x' = \tilde{x} + w$, $w \in \mathcal{K}$:

$$(b - Ax')^T A^{-1} (b - Ax') = (r - Aw)^T A^{-1} (r - Aw) = r^T A^{-1} r - 2 w^T r + w^T A w.$$

Hence, as discussed previously, $w = 0$ minimizes the energy norm of the error if and only if $w^T r = 0$ for all $w \in \mathcal{K}$. Therefore the minimization problem can be expressed:

$$\text{find } x \in x_0 + \mathcal{K} \text{ such that } b - Ax \perp \mathcal{K}.$$
This alternative formulation is sometimes called the Galerkin conditions. If $\mathcal{K}$ has a basis $U$, then $\tilde{x} = x_0 + Uy$ and $y$ is the solution of the system

$$U^T A U y = U^T r_0$$

where $r_0 = b - A x_0$.

Similarly, minimizing the residual norm can be expressed

$$\text{find } x \in x_0 + \mathcal{K} \text{ such that } b - A x \perp \mathcal{K}.$$ 

The general Petrov-Galerkin method takes the form

$$\text{find } x \in x_0 + \mathcal{K} \text{ such that } b - A x \perp \mathcal{L}$$

where $\mathcal{L}$ is a subspace of the same dimension as $\mathcal{K}$. If $\mathcal{K}$ has a basis $U$ and $\mathcal{L}$ has a basis $V$, this yields the system

$$V^T A U y = V^T r_0$$

to solve for $y$ where $r_0 = b - A x_0$.

**Review questions**

1. Let $A$ be symmetric positive definite, and let $\mathcal{K}$ be a linear subspace. Express the problem of minimizing the (square of the) energy norm

$$(b - A x)^T A^{-1} (b - A x), \quad x \in x_0 + \mathcal{K}$$

as a Petrov-Galerkin (or Galerkin) method.

2. What is the general form of a Galerkin method for solving $Ax = b$?

3. Consider the Galerkin conditions

$$\text{find } x \in x_0 + \mathcal{K} \text{ such that } b - A x \perp \mathcal{K}$$

for seeking an approximate solution to $Ax = b$. Express these as a linear system of equations given that $\mathcal{K} = \mathcal{R}(U)$ where $U$ is of full rank.

4. Let $A$ be symmetric positive definite, and let $\mathcal{K}$ be a linear subspace. Express the problem of minimizing the (square of the) residual norm

$$(b - A x)^T (b - A x), \quad x \in x_0 + \mathcal{K}$$

as a Petrov-Galerkin (or Galerkin) method.

5. What is the general form of a Petrov-Galerkin method for solving $Ax = b$?
6. Consider the Petrov-Galerkin conditions

\[ \text{find } x \in x_0 + \mathcal{K} \text{ such that } b - Ax \perp \mathcal{L} \]

for seeking an approximate solution to \( Ax = b \). Express these as a linear system of equations given that \( \mathcal{K} = \mathcal{R}(U) \) and \( \mathcal{L} = \mathcal{R}(v) \) where \( U \) and \( V \) are of full and equal rank.