SOLUTION BY MINIMIZATION

This is a part of a much larger subject, one taken up in much more extended fashion in *optimization theory*. To solve $Ax = b$, we reformulate it as a minimization problem.

Assume A is a real symmetric positive definite matrix of order $\dot{n}$. Define

$$f(x) = \frac{1}{2}x^T Ax - b^T x, \quad x \in \mathbb{R}^n$$

The solution of

$$\min_{x \in \mathbb{R}^n} f(x)$$

is $x = x^* \equiv A^{-1}b$. To see this, we introduce the useful quantity

$$E(x) = \frac{1}{2}\left(x^* - x\right)^T A \left(x^* - x\right), \quad x \in \mathbb{R}^n$$

Sometimes this is referred to as the "energy" associated with $x^* - x$, due to certain physical quantities associated with A.

Claim:

$$f(x) = E(x) - \frac{1}{2} b^T x^*, \quad x \in \mathbb{R}^n$$

Expanding,

$$\begin{aligned} E(x) - \tfrac{1}{2} b^T x^* &= \tfrac{1}{2} (x^* - x)^T A (x^* - x) - \tfrac{1}{2} b^T x^* \\ &= \tfrac{1}{2} (x^*)^T A x^* - \tfrac{1}{2} x^T A x^* - \tfrac{1}{2} (x^*)^T A x \\ &\quad + \tfrac{1}{2} x^T A x - \tfrac{1}{2} b^T x^* \end{aligned}$$

Simplifying,

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

$$\frac{1}{2} (x^*)^T A x^* - \frac{1}{2} b^T x^* = \frac{1}{2} \left[(x^*)^T A x^* - (x^*)^T b \right] = 0$$

Then

$$E(x) - \frac{1}{2} b^T x^* = -b^T x + \frac{1}{2} x^T A x = f(x)$$

Since A is symmetric and positive definite, let its eigenvalues be denoted by

$$0 < \lambda_1 \leq \cdots \leq \lambda_n$$

From Exercise 15 of Chapter 7, we obtain directly that

$$\lambda_1 \|z\|_2^2 \leq z^T A z \leq \lambda_n \|z\|_2^2, \quad z \in \mathbb{R}^n$$

Thus for the function $E(x)$,

$$\lambda_1 \|x^* - x\|_2^2 \leq E(x) \leq \lambda_n \|x^* - x\|_2^2, \quad x \in \mathbb{R}^n$$

Thus

$$E(x) = 0 \quad \Leftrightarrow \quad x = x^*$$

Since

$$f(x) = E(x) - \frac{1}{2} b^T x^*, \quad x \in \mathbb{R}^n$$

we have that $f(x)$ is a minimum if and only if $x = x^*$; and in that case,

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

HOW TO MINIMIZE $f(x)$?

We can choose a basis $\{p_1, ..., p_n\}$ and then look in succession at minimizing $f(x)$ along each direction $p = p_j$:

$$\min_{-\infty<\alpha<\infty} f\left(x^{(0)} + \alpha p\right) = f\left(x^{(0)} + \alpha^* p\right)$$

$$x^{(0)} \leftarrow x^{(0)} + \alpha^* p$$

For example, one could choose the basis $\{p_1, ..., p_n\}$ to be the standard basis $\left\{e^{(1)}, ..., e^{(n)}\right\}$. In fact, there are much better choices.

In optimization theory, we often choose a basis $\{p_1, ..., p_n\}$ of *conjugate directions*. These are a basis for which

$$p_j^T A p_i = 0, \quad i, j = 1, ..., n, \quad i \neq j$$

We say these are 'A-conjugate' or 'A-orthogonal'. Introduce a new inner product and norm

$$(x, y)_A = y^T A x, \quad \|x\|_A = \mathsf{sqrt}\left((x, x)_A\right)$$

Then from Exercise 15 of Chapter 7, as before,

$$\text{sqrt}\,(\lambda_1)\,\|x\|_2 \leq \|x\|_A \leq \text{sqrt}\,(\lambda_n)\,\|x\|_2\,, \quad x \in \mathbb{R}^n$$

With this norm $\|x\|_A$, called the *energy norm*, we have that a basis of conjugate directions is in fact an orthogonal basis with respect to the inner product $(x, y)_A$. Also,

$$E(x) = \frac{1}{2}\,\|x^* - x\|_A^2$$

Using the orthogonality, it is straightforward to obtain

$$x^* = \alpha_1 p_1 + \cdots + \alpha_n p_n$$

$$\alpha_k = \frac{p_k^T b}{p_k^T A p_k}, \quad k = 1, ..., n$$

The main question is how to choose the conjugate directions $\{p_k\}$.

Recall

$$f(x) = \frac{1}{2}x^T Ax - b^T x, \quad x \in \mathbb{R}^n$$

Introduce the partial solutions $x_0 = 0$,

$$x_k = \alpha_1 p_1 + \cdots + \alpha_k p_k, \quad k = 1, ..., n$$

$$\alpha_j = \frac{p_j^T b}{p_j^T A p_j}, \quad j = 1, ..., k$$

$$r_k = b - Ax_k = -\nabla f(x_k)$$

Then $r_0 = b$, and

$$x_k = x_{k-1} + \alpha_k p_k, \quad r_k = r_{k-1} - \alpha_k A p_k$$

For $k = n$, we will have $x_n = x^*$, the true solution. Often, we may have $x_k = x^*$ with $k < n$; or x_k may nearly equal x^*, accurately enough for practical purposes.

There are a number of properties with the use of the conjugate directions in minimizing $f(x)$, and these are given in Lemmas 1 and 2 on page 565. For example,

$$r_k^T p_i = 0, \quad i = 1, ..., k$$

and

$$\min_{-\infty<\alpha<\infty} f(x_{k-1} + \alpha p_k)$$

is solved uniquely with

$$\alpha = \alpha_k \equiv \frac{p_k^T b}{p_k^T A p_k}$$

Let $\mathcal{S}_k$ be the span of $\{p_1, ..., p_k\}$. Then

$$\min_{x \in \mathcal{S}_k} f(x)$$

is solved uniquely by $x = x_k$.

THE CONJUGATE GRADIENT METHOD

Given an initial guess, the direction of steepest descent on the graph of $z = f(x)$ is given by

$$-\nabla f(x_0) = r_0$$

and we choose this as our first conjugate direction p_1. In our case, we choose $x_0 = 0$ for simplicity, and then

$$p_1 = b$$

We construct the iterates x_k and the conjugate directions p_k simultaneously. Assume the iterates $x_1, ..., x_k$ and the conjugate directions $p_1, ..., p_k$ have been generated. A new direction p_{k+1} must be generated, and it must be A-conjugate to $p_1, ..., p_k$.

Assume $x_k \neq x^*$, as otherwise we would be done. Therefore, $r_k \neq 0$. We set

$$p_{k+1} = r_k + \beta_{k+1} p_k$$

Then the condition

$$p_k^T A p_{k+1} = 0$$

implies

$$\beta_{k+1} = -\frac{p_k^T A r_k}{p_k^T A p_k}$$

Together with

$$x_{k+1} = x_k + \alpha_{k+1} p_{k+1}, \quad \alpha_{k+1} = \frac{p_{k+1}^T b}{p_{k+1}^T A p_{k+1}}$$

this defines the *congugate gradient iteration* method.

The method is guaranteed to converge after at most n iterations, although it often gets there much sooner; or an acceptably small error is obtained with some x_k for some k much less than n. There are many optimality properties to this iteration, and we give only one here. Let

$$c = \frac{\lambda_1}{\lambda_n} = \frac{1}{\|A\|_2 \left\|A^{-1}\right\|_2} = \frac{1}{\text{cond}(A)_2}$$

with λ_1 and λ_n the smallest and largest eigenvalues of A. Then

$$\|x^* - x_k\|_A \leq 2 \left[\frac{1 - \text{sqrt}\,(c)}{1 + \text{sqrt}\,(c)}\right]^k \|x^*\|_A$$

The closer to 1 is $\text{cond}(A)_2$, the faster is the convergence.

NUMERICAL EXAMPLE

Consider solving a discretization of the integral equation

$$3x(s) - \int_0^1 \cos(\pi st)\, x(t)\, dt = 1, \quad 0 \le s \le 1$$

Convert this to an approximating linear system by applying the midpoint numerical integration rule with $n = 100$ subdivisions of $[0, 1]$. Let $h = 1/n$, and let t_i be the midpoint of the i^{th} subinterval of width h. Then the linear system is

$$3z_i - h \sum_{j=1}^{n} \cos\left(\pi t_i t_j\right) z_j = 1, \quad i = 1, ..., n$$

k	$\|x^* - x_k\|_A$	$\|x^* - x_k\|_\infty$
1	7.48E−1	7.11E−2
2	4.60E−3	7.60E−4
3	7.75E−7	1.46E−7
4	1.41E−12	2.83E−13
5	4.04E−15	6.11E−16

PRECONDITIONERS

Find a nonsingular matrix Q and rewrite $Ax = b$ as

$$\left(Q^{-1}AQ^{-T}\right)\left(Q^{T}x\right) = Q^{-1}b$$

with $Q^{-T} = \left(Q^{-1}\right)^{T}$. Introduce

$$\widetilde{A} = Q^{-1}AQ^{-T}, \quad \widetilde{x} = Q^{T}x, \quad \widetilde{b} = Q^{-1}b$$

Then solve $\widetilde{A}\widetilde{x} = \widetilde{b}$ by conjugate gradient iteration.

We try to choose Q such that

$$\text{cond}(\widetilde{A})_2 \ll \text{cond}(A)_2$$

and thus have the conjugate gradient iteration converge more rapidly. In applying this technique, the matrix $\widetilde{A}$ is never produced explicitly.

There is an "industry" that develops such *preconditioners*.

KRYLOV SUBSPACE METHODS

Look at the formulas for p_{k+1} and x_{k+1}:

$$p_{k+1} = r_k + \beta_{k+1} p_k$$

$$x_{k+1} = x_k + \alpha_{k+1} p_{k+1}$$

with $p_1 = b$, $x_0 = 0$, $r_0 = b$. Then

$$x_1 = \alpha_1 p_1, \quad r_1 = b - A x_1$$

$$p_2 = b - A x_1 + \beta_2 b = c_1 b + c_2 A b$$

for some c_1, c_2. In general, we can show

$$p_k = \sum_{j=0}^{k-1} c_j A^j b, \quad x_k = \sum_{j=0}^{k-1} d_j A^j b$$

for some constants $\left\{c_j\right\}$ and $\left\{d_j\right\}$.

Consider the subspace

$$\mathcal{S}_k = \text{span}\left\{b, Ab, A^2b, ..., A^{k-1}b\right\}$$

This is called the *Krylov subspace* of order k; and we are seeking our solution x_k from this subspace. There are methods other than the conjugate gradient method which seek solutions from $\mathcal{S}_k$. When A is no longer symmetric, one such method is called *GMRES*, and it is quite popular for such purposes. For a reference for such methods, see

R. Freund, G. Golub, and N. Nachtigal (1992) Iterative solution of linear systems, in *Acta Numerica 1992*, Cambridge University Press, pp. 57-100.

OPTIMALITY OF CG METHOD

Theorem. The iterates $\{x_k\}$ of the CG method satisfy

$$\|x^* - x_k\|_A = \min_{\deg(q)<k} \|x^* - q(A)b\|_A$$

From this many convergence results can be obtained, including one given earlier using the condition number of A.

Let $\lambda_1, ..., \lambda_n$ be the eigenvalues of A, with orthonormal eigenvectors $\underline{e}_1, ..., \underline{e}_n$. Then

$$q(A)e_j = q(\lambda_j)e_j$$

Write

$$x^* = \sum_{j=1}^{n} \xi_j \underline{e}_j$$

Then

$$q(A) = \sum_{j=1}^{n} \xi_j q(\lambda_j) \underline{e}_j$$

This leads to

$$\|x^* - x_k\|_A = \min_{\deg(q)<k} \sum_{j=1}^{n} \xi_j^2 \lambda_j \left[1 + \lambda_j q(\lambda_j)\right]^2$$

As an example of the use of this, suppose that A has only 4 distinct eigenvalues, say $\lambda_1, ..., \lambda_4$. Then let q be a degree 3 polynomial for which

$$1 + \lambda_j q(\lambda_j) = 0, \quad j = 1, 2, 3, 4$$

The above expression will then be zero and $x_4 = x^*$. What happens when the eigenvalues cluster around a few points?

CONJUGATE GRADIENT THEOREM

Assume $x_k \neq x^*$, and therefore $r_k = b - Ax_k \neq 0$. Then:

(a) $\text{span}\{r_0, r_1, ..., r_k\} = \text{span}\left\{b, Ab, A^2b, ...A^kb\right\}$

(b) $\text{span}\{p_1, ..., p_{k+1}\} = \text{span}\left\{b, Ab, A^2b, ...A^kb\right\}$

(c) $p_{k+1}^T p_i = 0,\ i = 1, ..., k$

(d) $\alpha_{k+1} = \dfrac{r_k^T r_k}{p_{k+1}^T A p_{k+1}}$

(e) $\beta_{k+1} = \dfrac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$

A NONLINEAR GENERALIZATION

Consider solving

$$\min_{x \in \mathbb{R}^n} f(x)$$

for a general scalar nonlinear function $f(x)$ defined on $\mathbb{R}^n$. Following is the Fletcher-Reeves generalization of the conjugate gradient iteration.

A. Given x_0, define $r_0 = p_1 = -\nabla f(x_0)$.
B. For $k = 1, ..n$:

Set $x_k = x_{k-1} + \alpha_k p_k$ with α_k the minimizer of

$$f(x_{k-1} + \alpha p_k)$$

Set $r_k = -\nabla f(x_k)$
For $k < n$, set $p_{k+1} = r_k + \beta_{k+1} p_k$ with

$$\beta_{k+1} = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$$

C. Set $x_0 := x_n$ and return to Step A.