FN 5012
5th. class
①

Minimization Algorithms.

When the matrix A is symmetric and positive definite, the problem of obtaining the solution of the linear equation

$$A x = b,$$

is equivalent to finding the point at which the quadratic functional

$$J(y) = \frac{1}{2} y^T A y - y^T b$$

achieves its minimum. We can thus use a minimization algorithm for J to obtain the solution of the linear equation x.

We shall study the gradient method and the conjugate gradient method. We shall see that, even for sparse matrices, the performance of the gradient method is not necessarily better than that of the Cholesky decomposition. We shall also see that for the sparse matrices of finite element approximations, the conjugate gradient method is the method of choice.

The gradient method and the conjugate gradient method have both algorithms of the following form

$x^0 = x_0$, $r^0 = Ax^0 - b$.

For $k = 0, \ldots, M$ do

  If $r^k \approx 0$ then set $x = x^k$ and stop.

  else

    Compute a <u>search direction</u> d^k,

    compute the <u>step length</u> α^k,

    and compute x^{k+1} as follows:

$$x^{k+1} = x^k + \alpha^k d^k.$$

  end If

end For

Notice that once the search direction is fixed we can only play with the value of α^k to set x^{k+1}. the step length α^k is called <u>optimal</u> when

$$J(x^k + \alpha^k d^k) \leq J(x^k + \alpha d^k) \qquad \forall\ \alpha \in \mathbb{R}.$$

Since

$$J(x^k + \alpha d^k) = J(x^k) + \alpha\ d^{kT}(Ax^k - b) + \tfrac{1}{2}\alpha^2\ d^{kT} A d^k$$

we have that the optimal step length is

$$\alpha^k = - \frac{d^{k^T} r^k}{d^{k^T} A d^k}$$

② <u>the Gradient method</u>. This method picks the search direction d^k to be the direction of fastest (<u>local</u>) descent of the functional J, that is,

$$\begin{aligned} d^k &= - \nabla J(x^k) \\ &= - (Ax^k - b) \\ &= - r^k \end{aligned}$$

the algorithm for this method is thus the following:

$x^0 = x_0$, $r^0 = Ax^0 - b$.

For $k = 0, \ldots, M$ do

  If $r^k \approx 0$ then set $x = x^k$ and stop.

  else

$$\alpha^k = \frac{r^{k^T} r^k}{r^{k^T} A r^k},$$

$$x^{k+1} = x^k - \alpha^k r^k$$

  end If

end For

the gradient method generates a minimizing sequence $\{x^k\}_{k\geq 0}$ since

$$\begin{aligned} J(x^{k+1}) &= J(x^k) - \alpha^k r^{kT} r^k + \tfrac{1}{2}\alpha^{k2} r^{kT} A r^k \\ &= J(x^k) - \tfrac{1}{2}\frac{(r^{kT} r^k)^2}{r^{kT} A r^k}. \end{aligned}$$

$$< J(x^k)$$

if $r^k \neq 0$. Moreover, the error in the energy norm decreases exponentially with the number of iterations. To see this, first notice that

$$\begin{aligned} e^{k+1} &= x^{k+1} - x \\ &= (x^k - \alpha^k (Ax^k - Ax)) - x \\ &= e^k - \alpha^k A e^k, \end{aligned}$$

and so

$$\|e^{k+1}\|_A^2 = e^{kT} A e^k - 2\alpha^k e^{kT} A^2 e^k + \alpha_k^2 e^{kT} A^3 e^k.$$

Since

$$\alpha^k = \frac{e^{kT} A^2 e^k}{e^{kT} A^3 e^k},$$

we get

$$\|e^{k+1}\|_A^2 = e^{kT} A e^k - \frac{(e^{kT} A^2 e^k)^2}{e^{kT} A^3 e^k}$$

$$= \left[1 - \left\{\frac{e^{kT} A^2 e^k}{e^{kT} A^3 e^k}\right\}\left\{\frac{e^{kT} A^2 e^k}{e^k A e^k}\right\}\right] \|e^{k-1}\|_A^2$$

and by the so-called Kantorovich inequality,

$$\|e^{k+1}\|_A^2 \leq \left(\frac{K-1}{K+1}\right)^2 \|e^k\|_A^2,$$

where $K = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$.

thus, we have

$$\boxed{\|e^k\|_A \leq \left(\frac{K-1}{K+1}\right)^k \|e^0\|_A.}$$

this inequality tells us that if the matrix A has a very big condition number, which is usually the case for matrices associated with elliptic problems, then the gradient method converges extremely slowly.

For example, if we want to make sure that

$$\|e^M\|_A \leq \epsilon,$$

Then we must take

$$M \geq 2 \frac{\log(\epsilon / \|e^0\|_A)}{\log(1 - 1/K)}$$

$$\simeq \left[-2 \log\left(\frac{\epsilon}{\|e^0\|_A}\right)\right] K.$$

$$= c_0 K.$$

This estimate is sharp. To see this, let us consider the following simple case:

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

If $K = \frac{\lambda_1}{\lambda_2} = 1$, then it is easy to verify that the gradient method converges in one iteration, as the error estimate predicts.

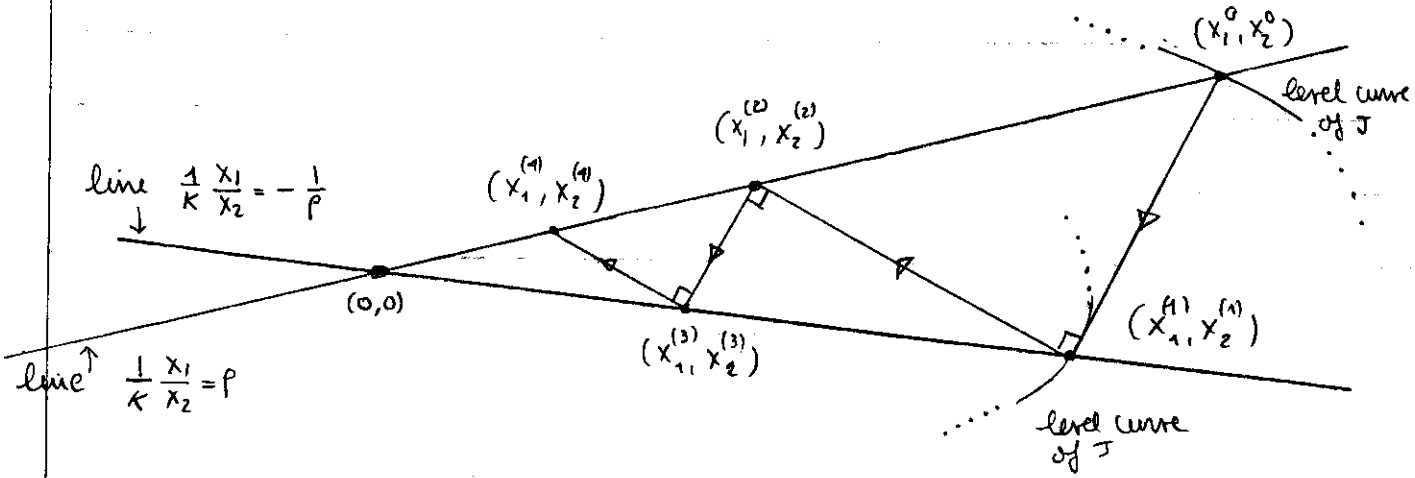
If $K > 1$, the gradient method still converges in a single iteration if $x_1^0 = 0$ or if $x_2^0 = 0$. However, in the other cases, we have that

$$\frac{1}{K} \frac{x_1^{(2n)}}{x_2^{(2n)}} = \rho \,, \qquad \frac{1}{K} \frac{x_1^{(2n+1)}}{x_2^{(2n+1)}} = -\frac{1}{\rho} \,,$$

where

$$\rho = \frac{1}{K} \frac{x_1^{(0)}}{x_2^{(0)}} \,.$$

this implies that we have the following situation:



It can be shown that

$$x_1^{(k+2)} = \left(\frac{1 - 2/k + 1/k^2}{1 + (\rho^2 + \rho^{-1})/k + 1/k^2}\right) x_1^{(k)},$$

and so,

$$|x_1^{(k+2)}| \leq \left(\frac{1 - 1/k}{1 + 1/k}\right)^2 |x_1^{(k)}|.$$

this shows that the number of iterations for convergence is proportional to k when $k \gg 1$, as expected.

Let us conclude our study of the gradient method by obtaining the number of operations it needs. this number is clearly M times the number of operations needed to compute x^{k+1} from x^k.

We can rewrite the gradient method's algorithm as follows, to minimize the number of operations it requires:

$$x^0 = x_0; \quad r^0 = Ax^0 - b.$$

For $k = 0, \ldots, M$ do

$$p^k = A r^k$$

$$\alpha^k = \frac{r^{kT} r^k}{r^{kT} p^k}$$

$$x^{k+1} = x^k - \alpha^k r^k$$

$$r^{k+1} = r^k - \alpha^k p^k$$

EndFor

Thus, the number of operations is M times the number of operations a multiplication by A requires!

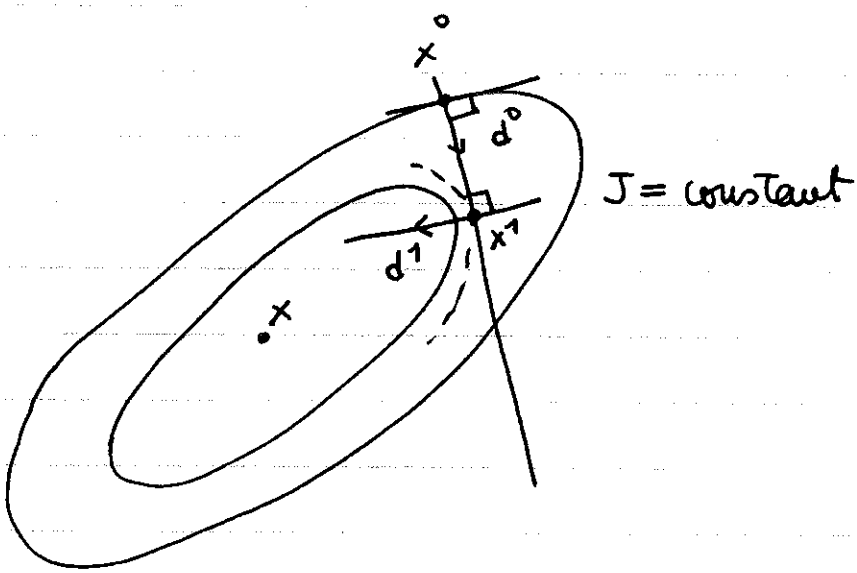
How can we improve the steepest descent method? To see how, let us begin by noting that

$$\begin{aligned} J(y) &= \tfrac{1}{2} y^T A y - b^T y \\ &= \tfrac{1}{2} y^T A y - x^T A y \\ &= \tfrac{1}{2} (y-x)^T A (y-x) - \tfrac{1}{2} x^T A y \end{aligned}$$

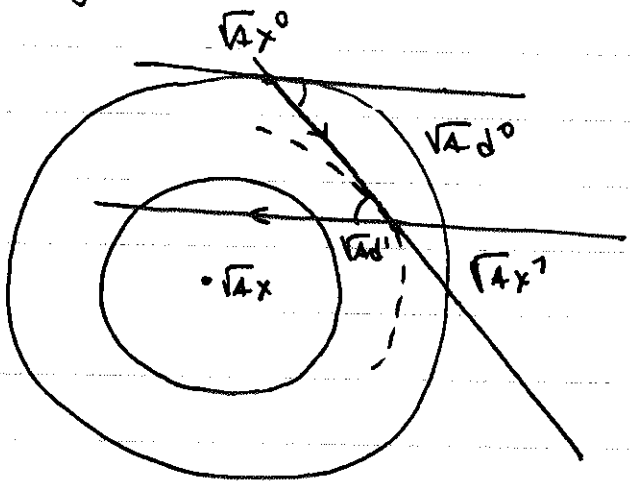
since A is symmetric (verify this computation!). Hence

$$J(y) = J(x) + \tfrac{1}{2} (y-x)^T A (y-x) \qquad \forall\, y \in \mathbb{R}^n.$$

Since x is fixed, the level curves of J are those of the mapping $y \longmapsto \tfrac{1}{2}(y-x)^T A (y-x)$ which are ellipses centered at the point x:



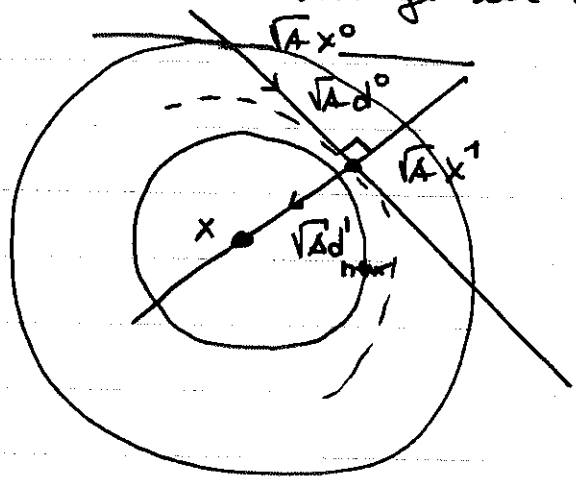
If we carry out the linear transformation $z = \sqrt{A}\, y$, the previous figure becomes



We now see that we could improve the steepest descent method by taking a different search direction d^1. Note that if the new search direction, d^1_{new} is such that

$$0 = \left(\sqrt{A}d^0\right)^T \left(\sqrt{A}\, d^1_{new}\right) = d^{0T} A\, d^1_{new},$$

the method would converge in only two iterations:



The conjugate gradient is based in this idea and is indeed superior to the steepest descent method as we see next.

③ the Conjugate gradient method. This method picks the search direction d^k as follows:

$$d^k = -r^k + \beta^k d^{k-1}, \qquad (d^0 = -r^0),$$

where the factor β^k is obtained by enforcing the A-orthogonality of d^k and d^{k-1}, that is:

$$\beta^k = \frac{r^{kT} A d^{k-1}}{d^{k-1T} A d^{k-1}}.$$

With this choice of search direction, the following remarkable property holds:

If $r^{m+1} \neq 0$ then, for $i < j \leq m+1$,

$$\begin{aligned} d^{iT} A d^j &= 0, \\ r^{iT} r^j &= 0, \\ d^{iT} r^j &= 0. \end{aligned}$$

the most important consequence of this property is the following result:

theorem. The Conjugate gradient method converges in at most n iterations.

Proof. Suppose the result not true. Then $r^n \neq 0$. But $\{r^l\}_{l=0}^{n}$ is a set of orthogonal vectors, by our orthogonality property. Since this is not possible, we must have $r^n = 0$. This completes the proof.

Another consequence is that we can rewrite α^k and β^k as follows:

$$\alpha^k = -\frac{r^{k^T} d^k}{d^{k^T} A d^k}$$

$$= -\frac{r^{k^T}(-r^k + \beta^k d^{k-1})}{d^{k^T} A d^k} \quad \text{(definition of } d^k\text{)}$$

$$= \frac{r^{k^T} r^k}{d^{k^T} A d^k},$$

$$\beta^k = \frac{r^{k^T} A d^{k-1}}{d^{k-1^T} A d^{k-1}}$$

$$= \frac{r^{k^T}\{r^k - r^{k-1}\}}{d^{k-1^T} A d^{k-1}} \cdot \frac{1}{\alpha^{k-1}} \quad \text{(def. of } x^k, r^k\text{)}$$

$$= \frac{r^{k^T} r^k}{d^{k-1^T} A d^{k-1}} \cdot \frac{d^{k-1^T} A d^{k-1}}{r^{k-1^T} r^{k-1}}$$

$$= \frac{r^{k^T} r^k}{r^{k-1^T} r^{k-1}}.$$

We can thus write the algorithm as follows:

$x^0 = x_0, \quad r^0 = Ax^0 - b.$

For $k = 0, \ldots, n-1$ do

If $r^k \approx 0$ then set $x = x^k$ and stop.

else

$$\beta^k = \frac{r^{k^T} r^k}{r^{k-1^T} r^{k-1}} \qquad (\beta^0 = 0)$$

$$d^k = -r^k + \beta^k d^{k-1} \qquad (d^0 = -r^0)$$

$$p^k = A d^k$$

$$\alpha^k = \frac{r^{k^T} r^k}{d^{k^T} p^k}$$

$$x^{k+1} = x^k + \alpha^k d^k$$

$$r^{k+1} = r^k + \alpha^k p^k$$

end If

end For

Thus, we can see that at each iteration, only one multiplication by the matrix A is required.

Next, we show that we do not have to carry out $n-1$ iterations and stop the algorithm a considerable amount of iterations before. To do this, we use the fact that

$$\text{span}\{d^0, \ldots, d^{m-1}\} = \text{span}\{r^0, r^1, \ldots, r^{m-1}\}$$
$$= \text{span}\{r^0, Ar^0, \ldots, A^{m-1} r^0\} =: W_m,$$

to estimate the error $\|x^k - x\|_A$.

Since

$$W_k = \text{span}\{r^0, r^1, \ldots, r^{k-1}\},$$

we have the following error equation:

$$e^{k^T} A w = r^{k^T} w = 0 \qquad \forall\, w \in W_k.$$

thus:

$$(x^k - x^0)^T A w = (x - x^0)^T A w \qquad \forall\, w \in W_k,$$

and since

$$x^k - x^0 = \sum_{j=0}^{k-1} \alpha^j d^j \qquad \in W_k,$$

we see that

> $x^k - x^0$ is the A-orthogonal projection of $x - x^0$ into W_k.

As a consequence, we have

$$\|e^k\|_A \leq \inf_{w \in W_k} \|(x - x_0) - w\|_A.$$

Since

$$W_k = \text{span}\{r^0, Ar^0, \dots, A^{k-1}r^0\}$$

$$= \text{span}\{A(x^0-x), A^2(x^0-x), \dots, A^k(x^0-x)\},$$

we can write

$$\|x^k - x\|_A \leq \inf_{p \in \mathcal{P}_k} \|p(A)(x-x^0)\|_A,$$

where $\mathcal{P}_k$ is the set of polynomials of degree at most k such that $p(0) = 1$. It can be proven that

$$\|x^k - x\|_A \leq 2\left[\frac{\sqrt{K}-1}{\sqrt{K}+1}\right]^k \|x^0 - x\|_A.$$

④ Proof of the orthogonality properties of the CG method.

First, we prove that the following properties are equivalent

(1) $r^{m+1} \neq 0$

(2) $\beta^{m+1} \neq 0$

(3) $d^{m+1} \neq 0$

(4) $\alpha^{m+1} \neq 0$.

Let us prove that (1) $\Rightarrow$ (2). Suppose that $\beta^{m+1} = 0$. By the orthogonality relations,

$$\beta^{m+1} = \frac{r^{m+1^T} r^{m+1}}{r^{m^T} r^m},$$

and so $r^{m+1} = 0$, which is not possible. Hence, (1) $\Rightarrow$ (2).

Let us prove that (2) $\Rightarrow$ (3). Suppose that $d^{m+1} = 0$. Then $r^{m+1} = \beta^{m+1} d^m$ and hence,

$$\begin{aligned} r^{m+1^T} r^{m+1} &= \beta^{m+1} r^{m+1^T} d^m \\ &= 0, \end{aligned}$$

by the orthogonality relations. this implies $r^{m+1} = 0$ and hence $\beta^{m+1} = 0$, which is not possible. Hence (2) $\Rightarrow$ (3).

Let us prove that (3) $\Rightarrow$ (4). Assume that $\alpha^{m+1} = 0$. By the orthogonality properties

$$\begin{aligned} \alpha^{m+1} &= \frac{r^{m+1^T} r^{m+1}}{d^{m+1^T} A d^{m+1}} \\ &= 0, \end{aligned}$$

which implies $r^{m+1} = 0$, which in turn implies $d^{m+1} = 0$. Hence (3) $\Rightarrow$ (4).

Finally, let us prove that (4) $\Rightarrow$ (1). Assume that $r^{m+1} = 0$. Then we clearly have $\alpha^{m+1} = 0$, which is not possible this completes the proof.

this shows that if $r^{m+1} \neq 0$, the corresponding new step of the method is well-defined.

Now, we prove that we have

$$\begin{aligned} &\text{span}\{d^0, d^1, \ldots, d^{m+1}\} \\ &= \text{span}\{r^0, r^1, \ldots, r^{m+1}\} \\ &= \text{span}\{r^0, Ar^0, \ldots, A^m r^0\}, \end{aligned}$$

provided that $r^{m+1} \neq 0$. We proceed by induction on m. For $m = -1$, the result is trivially true since $d^0 = -r^0$. Now assume the result true for $m = n$ and let us prove it is true for $m = n+1$. But

$$\begin{aligned} \text{span}\{d^0, d^1, \ldots, d^{n+1}\} &= \text{span}\{d^0, d^1, \ldots, -r^{n+1} + \beta_{n+1} d^n\} \\ &= \text{span}\{d^0, d^1, \ldots, d^n, -r^{n+1}\} \\ &= \text{span}\{r^0, r^1, \ldots, r^n, r^{n+1}\} \\ &= \text{span}\{r^0, r^1, \ldots, r^n, r^n + \alpha_n A d^n\} \\ \text{(since } \alpha^n \neq 0!) \quad &= \text{span}\{r^0, r^1, \ldots, r^n, A d^n\} \\ &= \text{span}\{r^0, Ar^0, \ldots, A^n r^0, A d^n\} \\ \text{(since } \beta^n \neq 0!) \quad &= \text{span}\{r^0, Ar^0, \ldots, A^n r^0, A^{n+1} r^0\}. \end{aligned}$$

this proves the result.

Now, we prove the orthogonality property by induction on m. For $m=0$, we have

$$d^{0T} A d^{1} = 0,$$

since this is enforced to compute β^1. We also have, that

$$\begin{aligned} r^{0T} r^{1} &= r^{0T} (r^{0} + \alpha^{0} A d^{0}) \\ &= r^{0T} r^{0} + \alpha^{0} \; r^{0T} A d^{0} \\ &= r^{0T} r^{0} + \frac{r^{0T} r^{0}}{r^{0T} A r^{0}} \cdot (r^{0T} A (-r^{0})) \\ &= 0. \end{aligned}$$

Finally

$$d^{0T} r^{1} = - r^{0T} r^{1} = 0.$$

Suppose that the orthogonality properties hold for $m=n$, and let us prove they also hold for $m=n+1$. We have to prove that the property holds for $j=n+1$ and $i \leq n$. We have that

$$d^{nT} A d^{n+1} = 0,$$

by definition of β^{n+1}. Also, for $i<n$,

$$
\begin{aligned}
d^{i^T} A d^{n+1} &= d^{i^T} A \left(-r^{n+1} + \beta^n d^n\right) \\
&= -d^{i^T} A r^{n+1} \\
&= \frac{1}{\alpha^i} \left(r^i - r^{i+1}\right)^T r^{n+1} \\
&= 0
\end{aligned}
$$

provided that $r^{j^T} r^{n+1} = 0 \quad \forall\, j \leq n$. But

$$
\begin{aligned}
r^{j^T} r^{n+1} &= r^{j^T} \left(r^n + \alpha^n A d^n\right) \\
&= r^{j^T} r^n + \alpha^n r^{j^T} A d^n \\
&= r^{j^T} r^n + \alpha^n \left(-d^j + \beta^j d^{j-1}\right) A d^n \\
&= 0
\end{aligned}
$$

for $j < n$, by the inductive hypothesis. For $j = n$, we have

$$
\begin{aligned}
r^{n^T} r^{n+1} &= r^{n^T} r^n - \alpha^n d^n A d^n \\
&= r^{n^T} r^n + r^{n^T} d^n \\
&= r^{n^T} \left(r^n + d^n\right) \\
&= r^{n^T} d^{n-1} \cdot \beta_n \\
&= 0,
\end{aligned}
$$

since $d^{n-1} \in \text{span}\{r^0, \ldots, r^{n-1}\}$ and by the inductive hypothesis.

Finally, the last property follows from the second and the "span" property.