ITERATIVE METHODS FOR THE SOLUTION OF LINEAR SYSTEMS

Many iterative methods for solving the linear system $Ax=b$ originate from the simple splitting

$$A = M - N$$

where M must be invertible. The choice of M and N determines the method.

It is easy to see that the solution x satisfies

$$Mx = Nx + b .$$

This motivates the iterative method

$$\begin{cases} M x^{(k+1)} = N x^{(k)} + b & k=0,1,\dots \\ x^{(0)} \text{ given.} \end{cases}$$

we may rewrite the above as

$$\begin{cases} x^{(k+1)} = T x^{(k)} + c & k=0,1,\dots \\ x^{(0)} \text{ given} \end{cases}$$

where $T = M^{-1} N$ is called the iteration matrix

$c = M^{-1} b$

Examples of iterative methods

(i) Jacobi (or point Jacobi)

Here we split A as $A = \overbrace{D}^{M} - \overbrace{(L+U)}^{N}$

where

D = diagonal part of A

$-L$ = strictly lower triangular part of A

$-U$ = " upper " " "

Assuming D^{-1} exists, Jacobi's method can be expressed as

$$\begin{cases} D x^{(k+1)} = (L+U)\, x^{(k)} + b, & k = 0, 1, \ldots \\ x^{(0)} \text{ given.} \end{cases}$$

In terms of components,

$$\begin{cases} a_{ii}\, x_i^{(k+1)} = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j^{(k)} + b_i, & i = 1, \ldots, n, \quad k = 0, 1, \ldots \\ x^{(0)} \text{ given.} \end{cases}$$

(ii) Gauss-Seidel (or point G.-S.)

Again with $A = D - L - U$,

here we choose $M = D - L$ and $N = U$, $P = (D-L)^{-1} U$

$$\begin{cases} (D-L)\, x^{(k+1)} = U x^{(k)} + b & k = 0, 1, \ldots \\ x^{(0)} \text{ given.} \end{cases}$$

In componentwise form, G.-S. reads

$$\sum_{j=1}^{i} a_{ij}\, x_j^{(k+1)} = \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} + b_i, \qquad \begin{matrix} i=1,\dots,n \\ k=0,1,\dots \end{matrix}$$

Remark

In contrast to Jacobi, when updating $x_i^{(k+1)}$, we use the updated values $x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}$

(iii) Successive Over Relaxation (SOR) method

In this method, $x_i^{(k+1)}$ is computed as the weighted mean of $x_i^{(k)}$ and the Gauss-Seidel iterate for that component. Specifically, for $\omega \neq 0$, a real parameter, the SOR method is given by

$$\begin{cases} x_i^{(k+1)} = (1-\omega)\, x_i^{(k)} + \dfrac{\omega}{a_{ii}} \left(b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) \\ \qquad i = 1,\dots,n; \quad k = 0,1,\dots \\ x_i^{(0)} \text{ given} \qquad i = 1,\dots,n. \end{cases}$$

This corresponds to the splitting $M = -L + \frac{1}{\omega} D$

$$N = \left[\left(\frac{1}{\omega} - 1 \right) D + U \right]$$

$$P = M^{-1} N = \left(D - \omega L \right)^{-1} \left[(1-\omega) D + \omega U \right].$$

The parameter ω is called the relaxation factor or parameter.

$\omega = 1 \Rightarrow$ SOR reduces to Gauss-Seidel

$\omega < 1 \Rightarrow$ underrelaxation

$\omega > 1 \Rightarrow$ overrelaxation

In addition to the above point iterative methods, we can define analogous block iterative methods. Suppose the matrix A can be decomposed into $p \times p$ blocks, i.e.

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & \cdots & \cdots & \vdots \\ \vdots & & & \vdots \\ A_{p1} & \cdots & & A_{pp} \end{bmatrix},$$

where the diagonal blocks A_{ii} are square and are invertible. In analogy to the point case, we partition L, U, D into $p \times p$ blocks as follows

$$-L = \begin{bmatrix} 0 & & \\ A_{21} & \ddots & \\ \vdots & \ddots & 0 \\ A_{p1} & \cdots & A_{p,p-1} \end{bmatrix}, \quad D = \begin{bmatrix} A_{11} & & & \\ & A_{22} & & 0 \\ & & \ddots & \\ 0 & & & A_{pp} \end{bmatrix}, \quad -U = \begin{bmatrix} 0 & A_{12} & \cdots & A_{1p} \\ & \ddots & \ddots & \vdots \\ & & \ddots & A_{p-1,p} \\ & 0 & & \end{bmatrix}$$

we define the block iterative methods by:

Block Jacobi $\quad D x^{(k+1)} = (L + U) x^{(k)} + b$

Block Gauss-Seidel $\quad (D - L) x^{(k+1)} = U x^{(k)} + b$

Block SOR $\quad \left(\frac{1}{\omega} D - L\right) x^{(k+1)} = \left[\left(\frac{1}{\omega} - 1\right) D + U\right] x^{(k)} + b.$

In particular, we should be interested in the case where A is block triangular.

Defn. The iterative method $\begin{cases} x^{(k+1)} = Px^{(k)} + C, & k=0,1,\dots \\ x^{(0)} \text{ given} \end{cases}$

where $Ax=b \iff x=Px+C$,

is called convergent if for all initial values $x^{(0)}$ $x^{(k)}$ converges to x.

Remark

This is a definition of global convergence. i.e. convergence is independent of the choice of initial starting vector $x^{(0)}$. The insistence on the global aspect of convergence is not unrealistic for iterative methods for linear systems. This is however not the case for nonlinear problems where iterative methods cannot be, except in very special cases, expected to converge starting from arbitrary initial vectors.

Theorem Let x be the solution of $Ax=b \iff x=Px+C$. Then, the following statements are equivalent.

(i) The iterative method $x^{(k+1)} = Px^{(k)} + C$ converges.

(ii) The spectral radius of P satisfies $\rho(P) < 1$.

(iii) There exists an induced matrix norm $\|\cdot\|$ such that $\|P\| < 1$.

proof.

(i) $\Rightarrow$ (ii) Let $e^{(k)} = x - x^{(k)}$. $\Rightarrow e^{(k+1)} = Pe^{(k)}$

$\Rightarrow e^{(k)} = P^k e^{(0)}$, $k=0,1,\dots$.

Let $|\lambda| = \rho(P)$ and V a corresponding eigenvector.

Set $e^{(0)} = V \Rightarrow e^{(k)} = \lambda^k e^{(0)} \Rightarrow \|e^{(k)}\| = |\lambda|^k \|e^{(0)}\|$.

since $\|e^{(0)}\| \neq 0$ and $\|e^{(k)}\| \to 0$, we must have $|\lambda| < 1$.

(ii) $\Rightarrow$ (iii) This is an easy consequence of the following result:

Theorem Let A be an $n \times n$ matrix and $\rho(A)$ its spectral radius. Then given any $\varepsilon > 0$, there exists an induced matrix norm $\|\cdot\|_\varepsilon$ such that $\|A\|_\varepsilon < \rho(A) + \varepsilon$.

proof.

Let U be a unitary matrix such that

$$U^* A U = U^{-1} A U = \text{upper triangular matrix} = \begin{bmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ & \lambda_2 & & \\ & & \ddots & t_{n-1,n} \\ 0 & & & \lambda_n \end{bmatrix},$$

where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of A.

For $\delta > 0$, to be chosen appropriately, consider $D_\delta = \operatorname{diag}\{1, \delta, \delta^2, \ldots, \delta^{n-1}\}$. We have

$$(U D_\delta)^{-1} A (U D_\delta) = \begin{bmatrix} \lambda_1 & \delta t_{12} & \cdots & \delta^{n-1} t_{1n} \\ & \lambda_2 & \delta t_{23} & \\ & & \ddots & \delta t_{n-1,n} \\ 0 & & & \lambda_n \end{bmatrix}.$$

Given $\varepsilon > 0$, choose δ sufficiently small, so that

$$(*) \qquad \sum_{j=i+1}^{n} |\delta^{j-i} t_{ij}| \leq \varepsilon \qquad i = 1, \ldots, n-1.$$

Define the map $\|\cdot\|_\varepsilon : \mathbb{C}^{n \times n} \to \mathbb{R}$ by

$$\|A\|_\varepsilon = \|(U D_\delta)^{-1} A (U D_\delta)\|_\infty .$$

On one hand, it can be shown that $\|\cdot\|_\varepsilon$ is indeed a matrix norm induced by the vector norm $\|x\|_\varepsilon \equiv \|(U D_\delta^{-1}) x\|_\infty$, on the other, it follows from the above that

$$\|A\|_\varepsilon = \|(U D_\delta)^{-1} A (U D_\delta)\|_\infty = \max_i \left| \lambda_i + \sum_{j=i+1}^{n} \delta^{j-i} t_{ij} \right|$$
$$\leq \rho(A) + \varepsilon .$$

(iii) $\Rightarrow$ (i). If there exists $\|\cdot\|$ (an induced matrix norm) such that $\|P\| < 1$, then for any $x^{(0)}$,

$$\|e^{(k)}\| = \|P^k e^{(0)}\| \leq \|P\|^k \|e^{(0)}\| \to 0$$

since $\|P\| < 1$. ■

The above results imply that a necessary and sufficient condition for convergence is that $\rho(P) < 1$. We now consider some aspects pertinent to the speed of convergence. First, we note that the speed of convergence depends on the vector norm chosen. Also, an interesting question is whether the errors decrease monotonically.

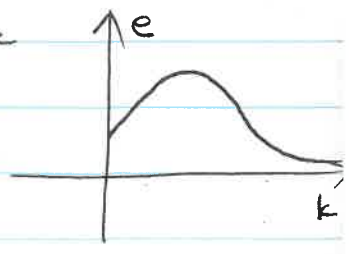
First, we consider the case of a normal iteration matrix and $\|\cdot\|_2$. We have $P = U^* \Lambda U$, where U is unitary and Λ is the diagonal matrix of eigenvalues of P. We have

$$\|P^k\|_2 = \|U^* \Lambda^k U\|_2 = \|\Lambda^k\|_2 = \rho(P)^k.$$

So in this particular case, the speed of convergence is given in terms of $\rho(P)$ and the errors can be seen to decrease monotonically.

If P is not normal and/or $\|\cdot\|$ is not the Euclidean norm, then the errors may not decrease monotonically. Indeed they may behave as in the figure. This phenomenon can be traced to a Jordan block of the form

$$\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}.$$



However, It can be shown that, asymptotically, the error vector $e^{(k)} = P^k e^{(0)}$ behaves "at worst" like $\rho(P)^k$. This is stated more precisely in the following result

Theorem (i) Let $\|\cdot\|$ be any vector norm, and let x be such that $x = Px + c$. Consider the iterative method

$$x^{(k+1)} = Px^{(k)} + c, \quad k = 0, 1, \ldots .$$

Then

$$\lim_{k\to\infty} \left\{ \sup_{\|x^{(0)} - x\| = 1} \|x^{(k)} - x\|^{1/k} \right\} = \rho(P) .$$

(ii) Let x be such that $x = Px + c = \tilde{P}x + \tilde{c}$.

consider the iterative methods

$$x^{(k+1)} = Px^{(k)} + c \quad \text{and} \quad \tilde{x}^{(k+1)} = \tilde{P}\tilde{x}^{(k)} + \tilde{c}, \quad k = 0, 1, \ldots ,$$

with $\rho(P) < \rho(\tilde{P})$, $x^{(0)} = \tilde{x}^{(0)}$.

Then for any number $\varepsilon > 0$, $\exists$ integer $\ell = \ell(\varepsilon)$ such that

$$\sup_{\|x^{(0)} - x\| = 1} \left\{ \frac{\|\tilde{x}^{(k)} - x\|}{\|x^{(k)} - x\|} \right\}^{1/k} \geq \frac{\rho(\tilde{P})}{\rho(P) + \varepsilon} . \qquad \square$$

Convergence of the Jacobi, Gauss-Seidel and relaxation methods

Theorem (Stein-Rosenberg) Let the Jacobi matrix $J = D^{-1}(L+U)$ be nonnegative, i.e. $J_{ij} \geq 0$. Let

$\mathcal{L}_1 = (D-L)^{-1}U$ denote the associated Gauss-Seidel matrix. Then, one and only one of the following mutually exclusive relations is valid

(i) $\rho(J) = \rho(\mathcal{L}_1) = 0$

(ii) $0 < \rho(\mathcal{L}_1) < \rho(J) < 1$

(iii) $1 = \rho(J) = \rho(\mathcal{L}_1)$

(iv) $1 < \rho(J) < \rho(\mathcal{L}_1)$.

Theorem Let A be a strictly diagonally dominant matrix, then both the Jacobi and Gauss-Seidel methods converge.

Remark The last result is not true for Hermitian, pos. definite matrices. Indeed, if

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}, \quad \text{then } \rho(J) > 1.$$

However, we have the following result

Theorem Let A be Hermitian, pos. def. and let $A = M - N$, M invertible. If in addition $M^* + N$ is Hermitian, pos. definite, then $\rho(M^{-1}N) < 1$.

Theorem (Ostrowski-Reich) If A is Hermitian and positive definite, the point or block relaxation method converges if and only if $0 < \omega < 2$.

Remark In proving the necessity of having $0 < \omega < 2$, we use the fact that $s(\mathcal{L}_\omega) \geq |\omega - 1|$, $\omega \neq 0$.

Let $A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$. Then $\rho(J) < 1 < \rho(\mathcal{L}_1)$

Let $A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$. Then $\rho(\mathcal{L}_1) < 1 < \rho(J)$.

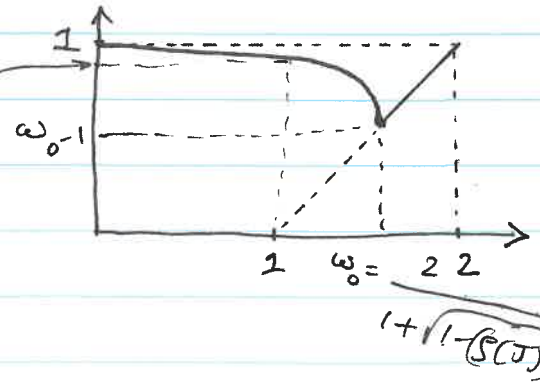
Theorem (Comparison of the Jacobi and Gauss-Seidel methods) Let A be a block tridiagonal matrix. Then the spectral radii of the corresponding block Jacobi and Gauss-Seidel methods are related by the equation

$$\rho(\mathcal{L}_1) = (\rho(J))^2;$$

so that the two methods converge or diverge simultaneously. If they converge, the Gauss-Seidel method converges more rapidly than the Jacobi method. □

Theorem Let A be a block tridiagonal matrix, such that all the eigenvalues of the corresponding block Jacobi method are real. Then, the block Jacobi and the block relaxation methods for $0 < \omega < 2$ converge or diverge simultaneously. If they converge, the function $\rho(\mathcal{L}_\omega)$ has the form

$\rho(\mathcal{L}_1) = (\rho(J))^2$



Remarks (1) Under the stated conditions, the optimal value of the relaxation parameter ω is given by

$$\omega_0 = \frac{2}{1+\sqrt{1-(\rho(J))^2}}.$$

This optimal choice of ω leads to dramatic increase in the speed of convergence. However, only in very special cases can ω_0 be computed exactly.

(2) The shape of $\rho(\mathcal{L}_\omega)$, $0<\omega<2$ shows that it is better to overestimate ω_0 than to underestimate it.

In particular, assume that A is Hermitian, positive definite, block tridiagonal matrix. By the Ostrowski-Reich theorem, the block (as well as point) relaxation method converges for $0<\omega<2$. In order to apply the last theorem, we need to show that the corresponding block Jacobi matrix has real eigenvalues.

Now, if $v \neq 0$

$$D^{-1}(L+U)v = \alpha v, \quad \text{i.e.} \quad (L+U)v = \alpha D v.$$

A Hermitian $\Rightarrow$ $D > 0$ (real) and $U^* = L$.

$$Av = (D-L-U)v = (1-\alpha)Dv \Rightarrow v^*Av = (1-\alpha)v^*Dv$$

Hence,

$1-\alpha$ is real and positive $\Rightarrow$ α real.

Iterative methods based on minimization techniques

We consider once again the problem of solving the linear system $Ax = b$. Suppose we can find a function $f: R^n \to R$ such that the solution x of $Ax = b$ is a (global) minimum of f. Then, any minimization algorithm can be used to approximate x. Of course, the minimization procedure should not lead to a problem more difficult than the solution of the original system and should converge with "reasonably" fast.

It turns out that if A is symmetric, positive definite, then an associated function f can be readily found.

Theorem Let A be an $n \times n$, symmetric, positive definite matrix. Then x is a solution of the linear system $Ax = b$ if and only if it is a global minimizer of the function $f(x): R^n \to R$ given by

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

Proof

Fix $x \in R^n$. Then, every vector $y \in R^n$ can be expressed as $y = x + \varepsilon v$ for some scalar ε and $v \in R^n$. We have

$$f(y) = f(x + \varepsilon v) = \frac{1}{2} (x + \varepsilon v)^T A (x + \varepsilon v) - b^T (x + \varepsilon v)$$

$$= \frac{1}{2} x^T A x - b^T x + \varepsilon \left[Ax - b \right] + \frac{1}{2} \varepsilon^2 v^T A v$$

$$= f(x) + \varepsilon v^T \left[Ax - b \right] + \frac{1}{2} \varepsilon^2 v^T A v.$$

Now if x solves $Ax = b$, then

$$f(y) = f(x) + \frac{1}{2} \varepsilon^2 v^T A v$$
$$\geq f(x), \qquad \forall y \in R^n$$

since A is s.p.d. Hence x is a global minimizer of f.

Conversely, assume that x is a global minimum of f but that $Ax - b \neq 0$. We can find $v \neq 0$ such that $v^T [Ax - b] = \alpha < 0$.

Thus

$$f(y) = f(x) + \varepsilon \alpha + \frac{1}{2} \varepsilon^2 v^T A v.$$
$$= f(x) + \varepsilon \left[\alpha + \frac{1}{2} \varepsilon v^T A v \right]$$

Since v is fixed and $\alpha < 0$, we can choose $\varepsilon > 0$ sufficiently small so that $\alpha + \frac{1}{2} \varepsilon v^T A v < 0$. This implies that $f(y) = f(x) +$ negative number

$\Rightarrow y = x + \varepsilon v \neq x$, $f(y) < f(x)$ which contradicts the fact that x is a global minimum of f. ■

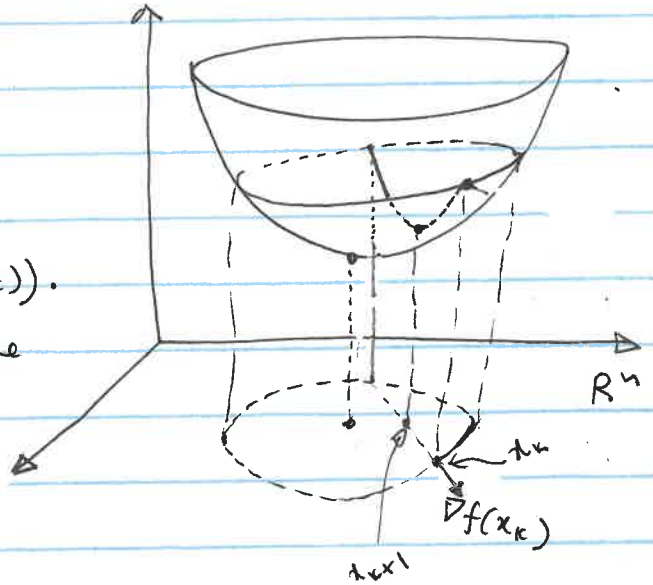
Thus, in the special case where A is s.p.d., any minimization algorithm applied to the iterative function $f(x) = \frac{1}{2} x^T A x - b^T x$, is in fact, an algorithm for approximating the solution x of $Ax = b$.

One of the most simple minded (and slowest)

algorithms is the Steepest descent method.

Suppose we are given x_k. We know that $(x_k, f(x_k))$, f increase (decreases) fastest in the direction of $\nabla f(x_k)$ $(-\nabla f(x_k))$. Hence, it makes sense to define

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$



where $\alpha_k \geq 0$ is a parameter to be determined. We now choose α_k so as to minimize $f(x_{k+1})$, i.e.

α_k is such that the map $\alpha \longmapsto g(\alpha) = f(x_k - \alpha \nabla f(x_k))$ is minimized. For general f, g may not have a minimum. However, in the special case of

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

we see that

$$g(\alpha) = \frac{1}{2} (x_k - \alpha \nabla f(x_k))^T A (x_k - \alpha \nabla f(x_k)) - b^T (x_k - \alpha \nabla f(x_k))$$

with $\nabla f(x_k) = A x_k - b \equiv r_k$.

$$\Rightarrow \quad g(\alpha) = \frac{1}{2} x_k^T A x_k - b^T x_k - \alpha \left[x_k^T A r_k - b^T r_k \right] + \frac{1}{2} \alpha^2 r_k^T A r_k$$

$$= f(x_k) - \alpha r_k^T r_k + \frac{1}{2} \alpha^2 r_k^T A r_k.$$

Note that $g(\alpha)$ is quadratic in α such that

$$g''(\alpha) = r_k^T A r_k > 0$$

Thus g has a global minimum $\Leftrightarrow g'(\alpha)=0$, i.e.

$$\alpha\, r_k^T A r_k - r_k^T r_k = 0 \quad \Rightarrow \quad \boxed{\alpha = \frac{r_k^T r_k}{r_k^T A r_k}}.$$

So, the steepest descent algorithm for $f(x)=\frac{1}{2}x^TAx-b^Tx$ is

$$x_{k+1} = x_k - \alpha_k r_k, \qquad k=0,1,\cdots$$

where $r_k = Ax_k - b$ and $\alpha_k = \dfrac{r_k^T r_k}{r_k^T(Ar_k)}$.

Note that if $r_k \neq 0$, then x_{k+1} is well defined. If $r_k=0$, this means $x_k = x$ and there is no need to compute x_{k+1} !!.

We next study the convergence properties of the steepest descent method.

<u>Lemma</u> The iterative process $x_{k+1} = x_k - \dfrac{r_k^T r_k}{r_k^T A r_k} r_k$, $k=0,\cdots$ $r_k = Ax_k - b$, satisfies

$$E(x_{k+1}) = \left\{ 1 - \frac{(r_k^T r_k)^2}{(r_k^T A r_k)(r_k^T A^{-1} r_k)} \right\} E(x_k), \tag{1}$$

where $E(y) = \frac{1}{2}(x-y)^T A (x-y)$.

<u>proof.</u>

The proof is by direct computation. We have, setting $y_k = x_k - x$,

$$\frac{E(x_k) - E(x_{k+1})}{E(x_k)} = \frac{2\alpha_k r_k^T r_k - \alpha_k^2 r_k^T A r_k}{y_k^T A y_k}.$$

Using $r_k = A y_k$, we have

$$\frac{E(x_k) - E(x_{k+1})}{E(x_k)} = \frac{\frac{2(r_k^T r_k)^2}{r_k^T A r_k} - \frac{(r_k^T r_k)^2}{r_k^T A r_k}}{r_k^T A^{-1} r_k} = \frac{(r_k^T r_k)^2}{r_k^T A^{-1} r_k},$$

From which the result follows. ∎

In order to obtain a bound on the rate of convergence, we need a bound on the r.h.s. of (1). The best bound is due to Kantorovich and his lemma stated below.

lemma (Kantorovich inequality) Let A be a symmetric, positive definite matrix. For any vector x there holds

$$\frac{(x^T x)^2}{(x^T A x)(x^T A^{-1} x)} \geq \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}$$

where λ_1 and λ_n are respectively the smallest and largest eigenvalues of A.

proof.

Let the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A satisfy

$$0 < \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n.$$

Since A is symmetric, A has a complete set of corresponding or orthonormal eigenvectors $v_1, \ldots, v_n$. Indeed, if $Q = (v_1 | \cdots | v_n)$ then $Q^T A Q = \Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$.

Now let $x = \sum_{i=1}^{n} \alpha_i v_i$. Then

$$\frac{(x^T x)^2}{(x^T A x)(x^T A^{-1} x)} = \frac{\left(\sum_{i=1}^{n} \alpha_i^2\right)^2}{\left(\sum_i \lambda_i \alpha_i^2\right)\left(\sum_i \lambda_i^{-1} \alpha_i^2\right)},$$

which can be written as

$$\frac{(x^T x)^2}{(x^T A x)(x^T A^{-1} x)} = \frac{1}{\sum_i \xi_i \lambda_i} \Big/ \sum_i \left(\frac{\xi_i}{\lambda_i}\right) \equiv \frac{\phi(\xi)}{\psi(\xi)}, \quad \text{where } \xi_i = \frac{\alpha_i^2}{\sum_j \alpha_j^2}.$$

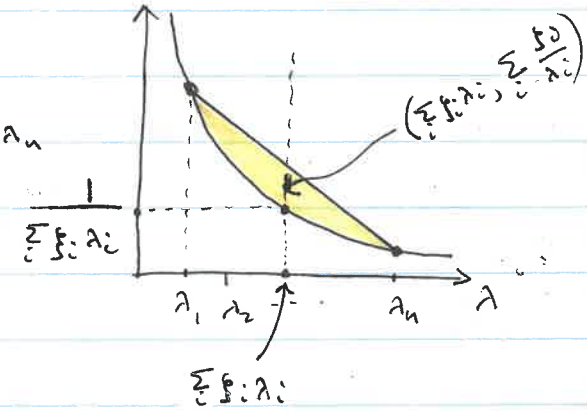
(without loss, one could take $\sum_i \alpha_i^2 = 1$!!)

Note that $\sum_i \xi_i^2 = 1$.

$\Rightarrow$ $\sum_i \xi_i \lambda_i$ is a point between λ_1 and λ_n

Also, the point $\left(\sum_i \xi_i \lambda_i, \sum_i \frac{\xi_i}{\lambda_i}\right)$ is a convex combination of the points $(\lambda_1, \frac{1}{\lambda_1}), \ldots, (\lambda_n, \frac{1}{\lambda_n})$, hence it belongs to the convex hull of these pts., i.e. the shaded region. It also belongs to the same vertical line as $\left(\sum_i \xi_i \lambda_i, \frac{1}{\sum_i \xi_i \lambda_i}\right)$. Thus, an appropriate lower bound for $\frac{\phi(\xi)}{\psi(\xi)}$ is

$$\min_{\lambda_1 \le \lambda \le \lambda_n} \frac{\frac{1}{\lambda} \leftarrow \text{pt. on curve}}{(\lambda_1 + \lambda_n - \lambda)/\lambda_1 \lambda_n \leftarrow \text{pt. on line.}}$$



which is minimized at $\lambda = \frac{\lambda_1 + \lambda_n}{2}$ and has value there equal to $\frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}$. ∎

Using Kantorovich's bound in (1) we get

Theorem (Steepest descent – quadratic case) Assume that A is s.p.d. Then for any $x_0 \in \mathbb{R}^n$, the method of steepest descent converges to the unique minimum of

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

i.e. the unique solution of the system $Ax = b$.
Furthermore, with $E(x_k) = \frac{1}{2}(x_k - x)^T A (x_k - x)$, there holds

$$E(x_{k+1}) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right)^2 E(x_k), \quad k = 0, 1, \ldots$$

$$= \left(\frac{1 - \lambda_1/\lambda_n}{1 + \lambda_1/\lambda_n} \right)^2 E(x_k). \quad \blacksquare$$

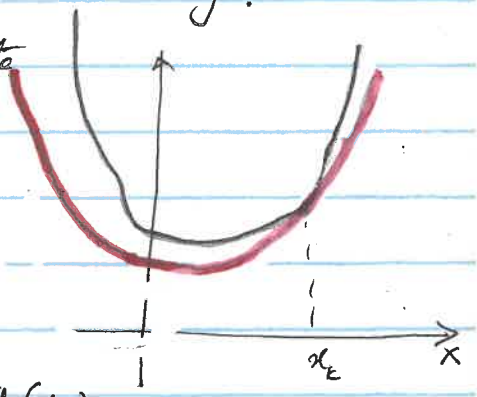
Remark If the matrix A is such that $0 < \frac{\lambda_1}{\lambda_n} \ll 1$, then the ratio

$$\frac{1 - \lambda_1/\lambda_n}{1 + \lambda_1/\lambda_n} \text{ will be very close to } 1$$

and consequently, convergence will be slow.

Newton's method The idea behind Newton's method is that the function f being minimized is approximated locally by a quadratic function, and this quadratic function is minimized exactly.

Thus, near x_k, we can approximate f by the quadratic Taylor polynomial according to



$$f(x) \approx f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k) \equiv q(x_k)$$

where $\nabla^2 f(z)$ is the "Hessian" matrix of f at z

$$\left(\nabla^2 f(z)\right)_{ij} = \left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_z .$$

Under certain condition on f, the quadratic $q: R^n \to R$ will have a unique minimizer. We let

$$x_{k+1} = \min_{x \in R^n} q(x)$$

In general, we have

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) .$$

Note that this is precisely the classical Newton's method for finding roots applied to the fn. $\nabla f: R^n \to R^n$.

Conjugate direction methods : The conjugate Gradient method

Defn. Given a symmetric matrix A, two vectors d_1 and d_2 are said to be *A-orthogonal* or *conjugate with respect to A* if

$$d_1^T A d_2 = 0 .$$

If $A = I$, conjugacy is equiv. to the usual notion of orthogonality.
A set of vectors $d_0, \dots, d_k$ is said to be A-orthogonal if

$$d_i^T A d_j = 0 \quad \text{if } i \neq j$$

Proposition If A is s.p.d. and the set of non zero vectors $d_0, \dots, d_k$ are A-orthogonal, then these vectors are linearly independent.

Proof.

Suppose $\exists\ \alpha_0, \dots, \alpha_k$ such that $\alpha_0 d_0 + \dots + \alpha_k d_k = 0$
then, for any j, $0 \leq j \leq k$,

$$0 = d_j^T A \cdot 0 = d_j^T A \left(\alpha_0 d_0 + \dots + \alpha_j d_j + \dots \alpha_k d_k \right)$$

$$= \alpha_0 d_j^T A d_0 + \dots + \alpha_j d_j^T A d_j + \dots + \alpha_k d_j^T A d_k$$

$$= \alpha_j d_j^T A d_j .$$

Since A is s.p.d., $d_j^T A d_j > 0$ since $d_j \neq 0$.
This shows that $\alpha_j = 0$. Since j is arbitrary, $\alpha_j = 0$, $j = 0, \dots, k$
$\Rightarrow$ $d_0, \dots, d_k$ are lin. indep. ∎

Conjugate direction Theorem Let $d_0, \dots, d_{n-1}$ be a set of nonzero A-orthogonal vectors. For any $x_0 \in R^n$, the sequence $\{x_k\}$ generated by

$$\text{(CD)} \quad \begin{cases} x_{k+1} = x_k - \alpha_k d_k, & k = 0, 1, \dots \\ \alpha_k = \dfrac{r_k^T d_k}{d_k^T A d_k}, & r_k = A x_k - b \end{cases}$$

converges to the unique solution x of $Ax = b$ in at most n steps. i.e. $x_m = x$ for some $m \le n$.

proof.

Since $d_0, \dots, d_{n-1}$ are lin. indep., we can write

$$x - x_0 = \beta_0 d_0 + \cdots + \beta_{n-1} d_{n-1}$$

for some $\beta_0, \dots, \beta_{n-1}$. Multiplying with A and taking inner product with d_k, we find

$$\text{(1)} \quad \beta_k = \frac{d_k^T A (x - x_0)}{d_k^T A d_k}.$$

Now following the iterative process (CD) from x_0 to x_k gives

$$x_k - x_0 = -(\alpha_0 d_0 + \cdots + \alpha_{k-1} d_{k-1}).$$

Hence, by the A-orthogonality of $d_0, \dots, d_{n-1}$ it follows that

$$\text{(2)} \quad d_k^T A (x_k - x_0) = 0.$$

Substituting (2) in (1), we get

$$\beta_k = \frac{d_k^T A (x - x_0)}{d_k^T A d_k} = \frac{d_k^T A (x - x_k)}{d_k^T A d_k} + \frac{d_k^T A (x_k - x_0)}{d_k^T A d_k}$$

($d_k^T A (x_k - x_0) = 0$)

$$= \frac{- r_k^T d_k}{d_k^T A d_k} = - \alpha_k .$$

Hence,

$$x - x_0 = - \alpha_0 d_0 - \cdots - \alpha_{n-1} d_{n-1} = x_n - x_0 ,$$

$\Rightarrow x_n = x$.

It is possible that $x_m = x$ for $m < n$. in that case $r_m = 0 \Rightarrow \alpha_m = 0 \Rightarrow x_{m+1} = x_m = x$. ∎

Expanding subspace theorem Let $d_0, \ldots, d_{n-1}$ be a sequence of nonzero A-orthogonal vectors in R^n. Then, for any $x_0 \in R^n$, the sequence x_k generated according to

$$x_{k+1} = x_k - \alpha_k d_k$$

$$\alpha_k = + \frac{r_k^T d_k}{d_k^T A d_k}$$

has the property that x_k minimizes $f(x) = \frac{1}{2} x^T A x - b^T x$ on the line $x = x_{k-1} + \alpha d_{k-1}$, $-\infty < \alpha < \infty$, as well as the linear variety $x_0 + B_k = x_0 + \text{span}\{d_0, \ldots, d_{k-1}\}$.

Corollary In the method of conjugate directions, the gradients (residuals) r_k, $k = 0, \ldots n$ satisfy

$$r_k^T d_i = 0 \quad \text{for} \quad i < k$$

The Conjugate Gradient method (CG)

The conjugate gradient method is the conjugate direction method that is obtained by selecting the successive direction vectors as a conjugate version of the successive gradients obtained as the method progresses. Thus, the directions are not specified beforehand but rather are determined sequentially at each step of the iteration. At step k one evaluates the current negative gradient vector and adds to it a linear combination of the previous direction vectors to obtain a new conjugate direction vector along which to move.

CG Algorithm

Starting at any $x_0 \in R^n$, define $d_0 = -r_0 = b - Ax_0$ and ...

$$x_{k+1} = x_k - \alpha_k d_k$$

(CG)
$$\alpha_k = \frac{r_k^T d_k}{d_k^T A d_k}$$

$$d_{k+1} = -r_{k+1} + \beta_k d_k, \qquad \beta_k = \frac{r_{k+1}^T A d_k}{d_k^T A d_k}$$

To verify that the CG algorithm is a conjugate direction algorithm, we need to verify that the vectors $d_0, \ldots, d_{n-1}$ generated are A-orthogonal. In this direction, we have

Conjugate Gradient Theorem: The CG algorithm above is a conjugate direction method. If it does not terminate at x_k, then

a) $\text{span}\{r_0, r_1, \ldots, r_k\} = \text{span}\{r_0, Ar_0, \ldots, A^k r_0\}$

b) $\text{span}\{d_0, d_1, \ldots, d_k\} = \text{span}\{d_0, Ad_0, \ldots, A^k d_0\}$

✓ c) $d_k^T A d_i = 0 \quad \text{for} \quad i \le k-1$

d) $\alpha_k = -\dfrac{r_k^T r_k}{d_k^T A d_k}$

e) $\beta_k = \dfrac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$.

In particular, part c) establishes the A-orthogonality of the vectors $d_0, \ldots, d_{n-1}$. Thus the CG algorithm terminates (in exact arithmetic (*)) after at most n steps. As such, this is not an interesting algorithm since in practical applications n is large.

However, the CG is more interesting in view of the following estimate

$$E(x_k) \le 4\left(\frac{1-\sqrt{\lambda_1/\lambda_n}}{1+\sqrt{\lambda_1/\lambda_n}}\right)^{2k} E(x_0).$$

Note the similarity between this and the estimate for the Steepest descent algorithm.

(*) In inexact arithmetic, CG does not terminate after a finite number of steps. Moreover

On one hand, this estimate does not guarantee an error reduction for each k, it is rather "global" in nature. On the other, for $\alpha \lambda_1/\lambda_n \ll 1$, the convergence rate of CG is much faster than that of S.D. due to the fact that $\sqrt{\lambda_1/\lambda_n}$ is much closer to 1 than $\frac{\lambda_1}{\lambda_n}$!!