Méthodes itératives pour résoudre des systèmes linéaires.

Tout ce que je raconte ci-dessous est tiré de trois sources :
[A-K] désigne le cours d'*algèbre linéaire numérique* de Grégoire Allaire et Sidi Mahmoud Kaber (Ellipses) ;
[C] est le traité de P.G. Ciarlet *Introduction à l'analyse numérique matricielle et à l'optimisation* (Masson) ;
[S] désigne l'ouvrage *Les matrices théorie et pratique* de Denis Serre (Masson).

Les trois livres me semblent tout à fait recommandables ; très subjectivement, je dirais que les explications informelles sur les questions simples sont souvent plus fournies et éclairantes dans [A-K], les explications sur les questions plus subtiles m'ont parfois paru plus pertinentes dans [C] et les démonstrations sont souvent plus percutantes et plus faciles à lire dans [S].

Dans tout ce qui suit, toutes les matrices sont **carrées** et **réelles** sauf mention contraire. Certains points de ci de là sont généralisables aux matrices rectangulaires (assez peu en fait) mais je n'en parlerai pas ; la restriction aux matrices réelles est une précaution personnelle pour ne pas risquer d'oublier occasionnellement une conjugaison. En fait les résultats sont parfois un peu plus difficiles à prouver dans le cas réel, le passage par $\mathbf{C}$ s'imposant avec une fois au moins une technicité pour redescendre dans le monde réel.

Plus d'une fois, j'abuserai du langage et dirai "$\mathbf{R}^n$" alors que je veux parler de l'espace des matrices-colonnes à n cases ; je noterai si besoin est $(e_1, \ldots, e_n)$ la base canonique de cet espace et $\|X\|_2$ la norme euclidienne canonique d'un vecteur-colonne X, c'est-à-dire $\|X\|_2 = \sqrt{{}^tXX}$.

I - Normes matricielles

1 - Mise au point : spectre, rayon spectral, valeurs singulières

Vous savez évidemment ce que sont les **valeurs propres** ou le **spectre** d'une matrice carrée. Juste un mot pour préciser que je noterai les valeurs propres avec la lettre λ et $\mathrm{Sp}\, A$ pour le spectre de A, et qu'il doit être entendu que le spectre contient aussi les valeurs propres complexes non réelles.

On appellera **rayon spectral** de A le module de la plus grande valeur propre de A ; on notera :

$$\rho(A) = \max_{\lambda \in \mathrm{Sp}\, A} |\lambda|.$$

Moins familière pourra paraître la notion suivante (qui n'est utilisée d'ailleurs que de façon assez anecdotique dans la suite) :

Définition : On appelle **valeurs singulières** de la matrice A les racines carrées des valeurs propres de la matrice (symétrique positive) $A{}^tA$. J'utiliserai systématiquement la lettre μ pour noter les valeurs singulières.

Remarques algébriques assez élémentaires :

* si vous ne voyez pas pourquoi ces racines carrées existeraient, voilà l'explication : $A{}^tA$ est symétrique, donc est diagonalisable avec un spectre réel ; pour tout vecteur X, ${}^tX(A{}^tA)X = {}^t({}^tAX)({}^tAX) = \|{}^tAX\|_2^2 \geq 0$, la matrice $A{}^tA$ est donc symétrique positive, et ses valeurs propres sont mieux que des réels, ce sont des réels positifs.

* si A est dès le départ une matrice symétrique, $A{}^tA = A^2$; on en déduit aussitôt que les valeurs singulières sont les valeurs absolues des valeurs propres de A. Lorsque A pousse la bonté jusqu'à être symétrique positive, les valeurs singulières, c'est la même chose que les valeurs propres.

Explication géométrique des valeurs singulières (pour A inversible)

Notons $\mathcal{S}$ la sphère-unité de $\mathbf{R}^n$ (pour la structure euclidienne canonique), et $\mathcal{E} = A\mathcal{S}$ son image par l'endomorphisme représenté par la matrice A dans la base canonique.

J'affirme que $\mathcal{E}$ est un ellipsoïde, et que les valeurs singulières de A en sont les demi-axes. Il ne reste qu'à le vérifier, ce qui demande de savoir manipuler des matrices mais n'a rien de profond.

Soit P une matrice orthogonale diagonalisant $A{}^tA$; en clair je suppose que :

$$P^{-1}(A^tA)P = \begin{pmatrix} \mu_1^2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \mu_2^2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \mu_{n-1}^2 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \mu_n^2 \end{pmatrix}.$$

Il est facile d'inverser cette identité, au passage je remplace P^{-1} par ${}^tP$, ce qui est raisonnable puisque P est orthogonale :

$${}^tP{}^tA^{-1}A^{-1}P = \begin{pmatrix} \mu_1^{-2} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \mu_2^{-2} & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \mu_{n-1}^{-2} & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \mu_n^{-2} \end{pmatrix}.$$

Soit $(f_1, \cdots, f_n)$ la base de $\mathbf{R}^n$ telle que la matrice de passage de $(e_1, \cdots, e_n)$ à $(f_1, \cdots, f_n)$ soit la matrice P. Comme P est orthogonale, $(f_1, \cdots, f_n)$ est elle aussi orthonormée (pour le produit scalaire canonique). La matrice-colonne Y dans cette nouvelle base du vecteur X est, on le sait, liée à X par la relation $X = PY$.

Il n'y a plus qu'à aligner des équivalences :

$X \in \mathcal{E} \iff PY \in \mathcal{E}$ (relation entre X et Y)

$\iff A^{-1}PY \in \mathcal{S}$ (définition de $\mathcal{E}$ comme image de $\mathcal{S}$)

$\iff {}^t(A^{-1}PY)(A^{-1}PY) = 1$ (formule classique pour décrire la norme canonique)

$\iff {}^tY\left({}^tP{}^tA^{-1}A^{-1}P\right)Y = 1$ (simple développement de la transposée d'un produit)

$\iff \sum \frac{y_i^2}{\mu_i^2} = 1.$ (la matrice qui apparaissait était justement celle qu'on avait préparée plus haut).

On a ainsi calculé l'équation de $\mathcal{E}$ en fonction des coordonnées dans la nouvelle base orthonormale $(f_1, \ldots, f_n)$; au vu de ces équations, il est clair que $\mathcal{E}$ est un ellipsoïde avec les demi-axes annoncés plus haut.

Digression : la décomposition SVD.

Cet ellipsoïde $A\mathcal{S}$ est aussi utilisable pour justifier une des nombreuses écritures d'une matrice comme produit (ce paragraphe se généralisant particulièrement bien aux matrices rectangulaires).

L'énoncé précis est le suivant : toute matrice A peut se factoriser sous la forme $A = U\Sigma V$ où Σ est une matrice diagonale sur laquelle figurent les valeurs singulières de A tandis que U et V sont des matrices orthogonales.

En voici une explication, que je me sens plus à l'aise pour expliquer en termes d'endomorphismes, et que je rédige pour A inversible quoique le résultat soit vrai pour toute A : soit a l'endomorphisme de $\mathbf{R}^n$ dont la matrice est A dans les bases canoniques (c'est-à-dire la transformation $A \mapsto AX$) ; soit $\mathcal{E}$ l'ellipsoïde $A\mathcal{S} = a(\mathcal{S})$ (éventuellement dégénéré si A n'est pas inversible) ; soit $\mathcal{E}'$ l'ellipsoïde ayant des axes de même longueur $2\mu_1 \geq 2\mu_2 \geq \ldots \geq 2\mu_n$ que $\mathcal{E}$ mais dont les directions des axes sont les axes de coordonnées. Appelons σ l'endomorphisme dont la matrice est diagonale dans la base canonique, avec les μ_i sur la diagonale : cet endomorphisme σ a été construit précisément pour envoyer $\mathcal{S}$ sur $\mathcal{E}'$. On peut ensuite trouver un endomorphisme orthogonal u envoyant $\mathcal{E}'$ sur $\mathcal{E}$ (puisque ces deux ellipsoïdes ont des axes de même longueur) ; l'endomorphisme $v^{-1} = a^{-1} \circ u \circ \sigma$ envoie alors $\mathcal{S}$ sur elle-même (les trois morceaux qu'on compose la transformant successivement en $\mathcal{E}'$ puis en $\mathcal{E}$ avant de la ramener sur elle-même). Cet endomorphisme v^{-1} est donc orthogonal ; son inverse aussi. Il suffit de renvoyer de l'autre côté de l'égalité certaines pièces de la définition de v^{-1} pour décomposer l'endomorphisme a comme annoncé.

Tout ceci est expliqué dans [A-K] (pages 42 à 47) (il déduit la caractérisation des valeurs singulières de la décomposition SVD et non le contraire -ce qui est probablement judicieux pour ne pas être gêné par

les matrices A non inversibles) ; la source mérite d'être consultée pour une intéressante application de la décomposition SVD à la compression des images.

Autre digression : le conditionnement.

Pour A matrice carrée, on appellera **conditionnement** de A le rapport μ_1/μ_n de la plus grande valeur singulière de A par la plus petite. (Notez que le concept de conditionnement dépend du choix d'une norme sur $\mathbf{R}^n$ -je ne parle ici que du conditionnement associé à la norme euclidienne canonique) ; tout plein de détails très clairs sont dans [A-K] pages 89 et suivantes.

Ainsi le conditionnement est toujours plus grand que 1, il vaut 1 exactement pour une matrice orthogonale et vaut $+\infty$ pour une matrice non inversible. Voir une interprétation géométrique du conditionnement rebondissant sur ces remarques page 93 de [A-K].

On remarque aussi immédiatement sur cette définition que les conditionnements de A et de A^{-1} sont égaux (si vous ne voyez pas pourquoi, songez que la plus grande valeur singulière de A^{-1} est $1/\mu_n$ et la plus petite est $1/\mu_1$).

Le conditionnement a un rapport étroit avec la résolution des systèmes linéaires : plus une matrice est mal conditionnée, plus la résolution de $AX = Y$ est sensible à des petites erreurs sur la connaissance de la donnée Y. Voir le dessin que j'ai fait au tableau en exposant tout ça.

2 - Normes sur l'espace des matrices

L'objectif de cette section est de bien faire comprendre qu'on peut avoir trois niveaux d'exigence pour une norme sur l'espace des matrices carrées :

$$\{\text{Normes induites sur } \mathcal{M}_n(\mathbf{R})\} \subset \{\text{Normes matricielles sur } \mathcal{M}_n(\mathbf{R})\} \subset \{\text{Normes sur } \mathcal{M}_n(\mathbf{R})\}.$$

Le plus général des trois concepts, c'est celui de norme quelconque. Ce n'est pas forcément inadapté : les normes fournies par les formules les plus simples sont quelconques et définissent tout de même –équivalence des normes en dimension finie– la même topologie que les autres.

Rentrent dans cette catégorie les normes 1 et ∞ sur l'ensemble des matrices, je veux dire celles données par les formules simples $\sum |a_{ij}|$ et $\text{Max}(|a_{ij}|)$.

On dit ensuite qu'une norme est **matricielle** lorsqu'elle vérifie en outre la relation suivante : pour toutes matrices A et B, $\|AB\| \leq \|A\|\|B\|$.

Le hérault de cette catégorie sera la norme de Frobenius (ou norme de Schur) :

$$\|A\|_F = \sqrt{\sum |a_{ij}|^2}.$$

(Vérifier que cette norme est matricielle est un calcul facile, à coup de Cauchy-Schwarz). Notez tout de suite que pour cette norme $\|I\|_F = \sqrt{n}$; ce détail permet de prouver aussitôt qu'elle n'est pas subordonnée.

Enfin on appelle **norme subordonnée à** (ou **norme induite par**) **une norme** $\|.\|$ **sur** $\mathbf{R}^n$ la norme donnée par la formule :

$$\|A\|_s = \underset{\|X\|=1}{\text{Sup}} \|AX\| \left(= \underset{\|X\|\neq 0}{\text{Sup}} \frac{\|AX\|}{\|X\|} \right).$$

3 - Le théorème de Householder

Quelque chose qui paraît bien dire si une matrice est grosse ou non, c'est son rayon spectral. Pourtant, le rayon spectral n'est évidemment pas une norme : pour le bloc de Jordan le plus simple $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, le rayon spectral est nul alors que la matrice ne l'est pas.

On arrivera tout de même plus loin à prouver des résultats où le rayon spectral s'interprète comme indicateur de la taille de la matrice grâce au

Théorème : (de Householder) Pour toute matrice carrée A

$$\rho(A) = \underset{\|.\|}{\text{Inf}} \|A\|$$

(où l'Inf est calculé sur l'ensemble des normes induites).

Quelques remarques pour aider à la lecture de la démonstration :

la démonstration découpe l'égalité en deux inégalités :
d'abord on prouve que pour toute norme induite, $\rho(A) \leq \|A\|$;
ensuite on prouve que pour tout $\epsilon > 0$ il existe une norme induite telle que $\|A\| \leq \rho(A) + \epsilon$.

La première partie peut sembler évidente si on est distrait. Elle l'est effectivement lorsqu'on travaille sur l'espace des matrices complexes, la définition de la norme induite étant adaptée par l'usage de X complexes ; ou lorsqu'on travaille avec une matrice A dont le rayon spectral est atteint sur une valeur propre réelle. Mais il faut être un peu astucieux si on veut la prouver pour toutes les matrices réelles. C'est pourquoi je renvoie particulièrement à [S] (page 44) pour cette démonstration. (Les deux autres sources citées éludent cette difficulté).

Pour la deuxième partie, il peut être instructif d'essayer de se convaincre qu'on saurait trouver une norme induite telle que $\|A\| = \rho(A)$ quand la matrice A est diagonalisable. Mais on se convaincra en lisant [S] (remarque de la page 45) qu'en tirer le théorème par un argument de densité est illusoire quoique tentant ; la fin de la preuve n'est pas très palpitante et nécessite de trigonaliser A.

Remarque : il est assez raisonnable de se demander si le théorème serait vrai ou non si on prenait l'Inf sur toutes les normes matricielles et non sur le plus petit ensemble des normes subordonnées. C'est une bonne gymnastique mentale de se demander dans quel sens on peut déduire aussitôt une inégalité de l'énoncé donné plus haut ; pour l'autre sens, il faut jongler avec les sources ! l'énoncé est écrit noir sur blanc dans [A-K] page 57 mais la preuve (très simple) ne me convainc pas si on s'intéresse aux matrices réelles, alors que la démonstration de [S] page 44 prouve bien l'inégalité y compris pour des normes matricielles, mais ne l'annonce pas.

4 - Itération d'une matrice

Le rapport du théorème de Householder avec le titre de ce polycopié est qu'il permet de prouver la :

Proposition : Les trois propriétés suivantes sont équivalentes :

(1) $A^k \to 0$ quand $k \to \infty$;
(2) Pour toute colonne $X \in \mathbf{C}^n$, $A^k X \to 0$ quand $k \to \infty$;
(3) $\rho(A) < 1$.

Démonstration : les implications (1) $\Rightarrow$ (2) et (2) $\Rightarrow$ (3) sont évidentes. La seule subtile est celle qui permet de remonter de (3) à (1). Pour celle-ci, supposons que $\rho(A) < 1$; on peut alors, en utilisant Householder, mettre sur $\mathbf{R}^n$ une norme appropriée de sorte que, pour la norme subordonnée, $\|A\| < 1$. En utilisant cette norme et l'inégalité, $\|A^k\| \leq \|A\|^k$, la convergence de A^k vers 0 est claire.

Remarque : en fait, on peut aussi montrer (3) $\Rightarrow$ (1) "à la main" en utilisant la décomposition $S + N$ de A.

5 - Un peu plus sur la suite $(\|A^k\|)$

On aimerait bien écrire que la suite $\|A^k\|$ converge vers zéro à la même vitesse que $(\rho(A))^k$, lorsque $\rho(A) < 1$.

À ce sujet, l'énoncé disponible dans les sources citées ci-dessus n'est pas très précis (mais a le charme d'être encore vrai dans un Banach de dimension infinie)

Proposition : Pour toute norme sur $\mathcal{M}_n(\mathbf{R})$,

$$\lim_{k \to +\infty} \|A^k\|^{1/k} = \rho(A).$$

On en trouvera une démonstration très lisible et très courte dans [C] page 22 (elle dépend de la proposition précédente, donc de Householder), et une démonstration plus longue, fort plaisante (avec une intervention de la formule de Cauchy) et appropriée à la dimension infinie dans [S] page 48. (Les deux sources énoncent le théorème en supposant la norme "matricielle" mais c'est un exercice facile que de se convaincre que cette restriction n'est pas nécessaire, en dimension finie bien évidemment).

II - Méthodes itératives

1 - Retour sur le théorème d'inversion locale

Supposant (ou feignant de supposer) que la preuve du théorème d'inversion locale vous est "bien connue", j'en rappelle le principe avec quelques notations un peu bizarres, pour ensuite montrer les analogies avec la résolution de systèmes linéaires.

Si elles ne sont pas nettes dans votre tête, c'est le moment d'assimiler deux idées :
* lorsqu'on veut chercher un point fixe d'une application φ, il est souvent malin de considérer une suite récurrente $x_{k+1} = \varphi(x_k)$. Si on a bien choisi x_0 et surtout φ, on aura souvent des arguments d'analyse permettant d'assurer la convergence de la suite, vers un point fixe de φ ;
* lorsqu'on veut résoudre une équation, il peut être opportun de faire une ou deux transformations algébriques sans aucune subtilité, de sorte que le problème devienne une recherche de point fixe.

Soit a une application de classe $\mathcal{C}^1$ définie sur un ouvert U de $\mathbf{R}^n$, à valeurs dans $\mathbf{R}^n$; pour alléger les notations on suppose que U est un voisinage de 0 et que $a(0) = 0$. On notera M la différentielle de a au point 0, qui est donc une application linéaire de $\mathbf{R}^n$ vers $\mathbf{R}^n$. On fait l'hypothèse supplémentaire de l'inversibilité de M. Le théorème d'inversion locale assure alors qu'il existe un ouvert $V \subset U$ voisinage de 0 et un voisinage ouvert W de 0 tel que pour y dans W, l'équation d'inconnue x

$$(E) \qquad a(x) = y$$

admette une et une seule solution dans V (le théorème annonce aussi la différentiabilité de cette solution comme fonction de y, mais on l'oubliera ici).

Pour alléger les notations encore plus, supposons dans un premier temps que $\mathrm{d}a(0) = Id$; on notera $n = Id - a$ –la preuve repose sur le fait que $Id = \mathrm{d}a(0)$ est une bonne approximation de a au voisinage de 0, et que n est en conséquence petite, en fait précisément que ses dérivées partielles sont petites.

Faisons quelques manipulations tout à fait élémentaires sur l'équation (E) :

$$a(x) = y \iff (Id - n)(x) = y \iff x = n(x) + y.$$

Ainsi résoudre (E) équivaut à trouver un point fixe de l'application $\varphi_y : x \mapsto n(x) + y$. C'est le moment de faire de l'analyse : en utilisant la petitesse des dérivées partielles de n, on montre que φ_y est contractante, du moins si on la restreint judicieusement. En utilisant le théorème du point fixe pour les applications contractantes dans un espace métrique complet, on a la garantie que φ_y possède un point fixe unique, mais aussi que ce point fixe peut être reconstitué comme limite d'une suite récurrente $x_{k+1} = \varphi_y(x_k)$.

Voyons maintenant ce qu'il faut écrire si la différentielle de a n'est pas Id mais une application linéaire bijective M quelconque (l'analyse est la même, les équations sont seulement un peu plus lourdes) : on pose encore $n = M - a$; les manipulations sur (E) s'écrivent alors :

$$a(x) = y \iff (M - n)(x) = y \iff M(x) = n(x) + y \iff x = (M^{-1} \circ n)(x) + M^{-1}(y).$$

La résolution de (E) est donc cette fois ramenée à l'itération de $\varphi_y : x \mapsto (M^{-1} \circ n)(x) + M^{-1}(y)$. La formule prend plus de place, mais le principe est exactement le même.

Cela fournirait-il un procédé pour résoudre $A(x) = y$ (d'inconnue x) lorque $a = A$ est linéaire ? Manifestement pas. En effet l'endomorphisme $M = \mathrm{d}a(0)$ coïncide alors avec l'application $a = A$; la petite application contractante n est tellement petite qu'elle est nulle ; la fonction à itérer est alors simplement la fonction constante $A^{-1}(y)$. Certes, la suite récurrente destinée à approcher $A^{-1}(y)$ convergera très vite, puisqu'elle sera constante dès son second terme, mais le calcul de l'application à itérer n'est pas plus simple que le calcul direct de la solution $x = A^{-1}(y)$ (non seulement ce n'est pas plus simple mais c'est exactement la même chose !!!)

Alors que faire ? L'idée est de décomposer $A = M - n$ (qu'on écrira alors $A = M - N$ pour mettre en relief la linéarité de n) pour une M qui n'est pas égale à A mais qui n'en est tout de même pas trop éloignée, tout en étant beaucoup plus facile à inverser. Dès lors que M n'est pas trop éloignée de A, N sera petit, pas de façon aussi flagrante que le n de tout à l'heure (qui avait une différentielle nulle en 0) mais suffisamment tout de même pour garantir la convergence d'une suite récurrente bien construite vers la solution de $A(x) = y$.

Après cette longue introduction, il est temps de rendre précis ce plan (en glissant des endomorphismes, que j'aime mieux pour expliquer, aux matrices, que j'aime mieux pour écrire les calculs avec sobriété).

2 - Principe des méthodes de résolution itérative

Soit A une matrice carrée réelle inversible (n,n). Soit Y un vecteur-colonne fixé. On veut calculer l'unique solution X_∞ de l'équation $AX = Y$.

On décompose A comme somme de deux matrices : $A = M - N$, où on choisit M et N pour que :
* M soit inversible, et beaucoup plus facile à inverser que A ;
* N soit petite, au sens précis suivant : $\rho(M^{-1}N) < 1$.

On choisit un X_0 arbitrairement ; on note φ l'application affine de $\mathbf{R}^n$ vers $\mathbf{R}^n$ définie par :

$$\varphi(X) = (M^{-1}N)X + M^{-1}Y.$$

On considère alors la suite récurrente définie à partir de X_0 par la relation $X_{k+1} = \varphi(X_k)$. Alors la suite (X_k) converge vers X_∞.

Justification : remarquons tout d'abord que X_∞ est un point fixe de φ : en effet

$$\begin{aligned}\varphi(X_\infty) - X_\infty &= [(M^{-1}N)X_\infty + M^{-1}Y] - X_\infty \\ &= [(M^{-1}N)X_\infty + M^{-1}AX_\infty] - M^{-1}MX_\infty \\ &= M^{-1}(N + A - M)X_\infty \\ &= 0.\end{aligned}$$

(ce calcul matriciel bien opaque reprend de façon condensée les manipulations détaillées au 1)).

Soustrayons alors membre à membre les deux identités $X_{k+1} = \varphi(X_k)$ et $X_\infty = \varphi(X_\infty)$: seule la partie linéaire $(M^{-1}N)$ de l'application affine φ reste après soustraction, et on obtient :

$$X_{k+1} - X_\infty = (M^{-1}N)(X_k - X_\infty).$$

Donc pour tout k, $X_k - X_\infty = (M^{-1}N)^k(X_0 - X_\infty)$. Or l'hypothèse de petitesse de N a précisément été posée sous la forme même qui garantit la convergence vers 0 de cette suite vectorielle (allez vous référer au I-4 si vous l'avez déjà oublié).

3 - Description sommaire des méthodes les plus simples

Selon la forme du M choisie pour être inversée à la place de A, la méthode itérative porte un nom différent :
* pour une matrice M scalaire, on parle de méthode du gradient à pas constant, ou de méthode de Richardson ;
* on peut ensuite penser à utiliser une M diagonale ; l'idée la plus simple étant de prendre $M = D$, où la matrice diagonale D recopie la diagonale de A. On parle alors de méthode de Jacobi. Une méthode plus subtile est d'utiliser $M = \dfrac{D}{\omega}$ pour un paramètre ω réel à choisir judicieusement ; on parle alors de méthode de Jacobi relaxée ;
* le niveau de complexité supérieur consiste à utiliser une M triangulaire, disons triangulaire inférieure. Dans ce paragraphe, on notera $A = D - E - F$ avec D diagonale, E triangulaire strictement inférieure et F triangulaire strictement supérieure. L'idée la plus simple est alors de prendre $M = D - E$; on parle alors de méthode de Gauss-Seidel. On peut la compliquer en utilisant plutôt $M = \dfrac{D}{\omega} - E$ pour un paramètre ω réel : on parlera alors de méthode de Gauss-Seidel relaxée.
* et on peut faire pire, mais ce n'est plus alors une stricte application du 2) : on peut ne plus faire une itération à proprement parler, mais modifier l'application affine utilisée de pas en pas ; ce seront alors les méthodes du gradient à pas variable, voire du gradient conjugué. Je n'en parlerai pas pour la bonne raison que je n'ai pas lu les chapitres qui les traitent et n'y connais donc rien ; si ça vous intéresse, c'est fait dans les trois sources que j'ai utilisées.

4 - Les méthodes itératives résistent aux petites erreurs

C'est très clairement fait page 158 de [A-K], je le paraphrase à l'usage de ceux qui auraient ce polycopié sous la main sans le livre aux alentours.

Dans une situation réelle, on fait une petite erreur de calcul à chaque pas de l'itération (due notamment aux arrondis dans les calculs liés à l'inversion de M).

La formule idéale $X_{k+1} = \varphi(X_k)$ ne modélise donc pas correctement la façon de fonctionner de la machine ; il faut la remplacer par la formule réaliste $X_{k+1} = \varphi(X_k) + \epsilon_k$, où ϵ_k est une petite erreur.

Le bonheur, c'est que ces erreurs ne se propagent pas ! La démonstration en est assez facile (on estime $\|X_k - X_\infty\|$ en tenant compte des ϵ_i, $i < k$, la formule restant à taille humaine. En majorant alors les saletés accumulées dans cette formule par une somme de série géométrique, ça roule).

Évidemment, il ne faut pas espérer converger exactement vers X_∞ en faisant des erreurs, l'énoncé précis qu'on parvient à montrer est le suivant :

Proposition : soit φ une application affine dont la partie linéaire $M^{-1}N$ vérifie $\rho(M^{-1}N) < 1$, et soit X_∞ l'unique point fixe de φ. Il existe une constante C ne dépendant que de $M^{-1}N$ telle que, pour toute valeur de X_0 et toute suite d'erreurs d'arrondi (ϵ_k), si on considère la suite pas tout à fait récurrente $X_{k+1} = \varphi(X_k) + \epsilon_k$, celle-ci vérifie :

$$\limsup \|X_k - X_\infty\|_2 \leq C \operatorname*{Sup}_k \|\epsilon_k\|_2.$$

5 - M est-elle vraiment facile à inverser ?

Dans les calculs de cette section, lorsqu'un vecteur-colonne s'appelle d'une lettre majuscule Z, on note $z^{(i)}$ sa i-ème composante.

D'expérience pour avoir essayé avant vous de les lire, il me semble que les calculs qui suivent sont sans intérêt à lire, mais qu'il est instructif de les écrire. Suivez les donc avec un crayon. Dans les trois sources proposées, c'est dans [C] que je les ai le mieux aimés (pages 98 à 101).

Il s'agit de vérifier que pour chacune des méthodes itératives, la formule donnant X_{k+1} en fonction de X_k est effectivement très facile à calculer effectivement.

Soit $A = (a_{ij})$ la matrice intervenant dans le système linéaire à résoudre ; on notera $A = D - E - F$ avec D diagonale, E triangulaire strictement inférieure et F triangulaire strictement supérieure.

Pour chaque méthode itérative, X_{k+1} est donné en fonction de X_k par la relation $X_{k+1} = (M^{-1}N)X_k + M^{-1}Y$; tenant compte de la définition de $N = M - A$, on regroupera plutôt cette relation sous la forme :

$$(*) \qquad MX_{k+1} = MX_k - AX_k + Y$$

*** Explicitation des calculs pour le gradient à pas constant**

Notons $M = \dfrac{I}{\alpha}$ la matrice scalaire utilisée ; la relation $(*)$ s'écrit alors :

$$X_{k+1} = X_k + \alpha(Y - AX_k)$$

dont la simplicité est indéniable sans qu'il soit besoin de développer complètement le système.

*** Explicitation des calculs pour la méthode de Jacobi**

Pour cette méthode, la relation $(*)$ devient :

$$DX_{k+1} = (E + F)X_k + Y$$

qu'on peut apprécier de développer sous forme d'un système pour y voir quelque chose :

$$\begin{cases} a_{11}x_{k+1}^{(1)} = \quad y^{(1)} \quad -a_{12}x_k^{(2)} \cdots -a_{1,n-1}x_k^{(n-1)} - a_{1n}x_k^{(n)} \\ a_{22}x_{k+1}^{(2)} = -a_{21}x_k^{(1)} \quad +y^{(2)} \quad \cdots -a_{2,n-1}x_k^{(n-1)} - a_{2n}x_k^{(n)} \\ \qquad\qquad\qquad\qquad\qquad \cdots \\ a_{nn}x_{k+1}^{(n)} = -a_{n1}x_k^{(1)} - a_{n2}x_k^{(2)} \cdots -a_{n,n-1}x_k^{(n-1)} \quad +y^{(n)} \end{cases}$$

*** Explicitation des calculs pour la méthode de Gauss-Seidel**

Pour cette méthode, la relation $(*)$ devient :

$$(D - E)X_{k+1} = FX_k + Y$$

qu'on regroupera plutôt en :

$$DX_{k+1} = EX_{k+1} + FX_k + Y$$

ainsi regroupé, ça ressemble beaucoup au précédent :

$$\begin{cases} a_{11}x_{k+1}^{(1)} = y^{(1)} - a_{12}x_k^{(2)} \cdots - a_{1,n-1}x_k^{(n-1)} - a_{1n}x_k^{(n)} \\ a_{22}x_{k+1}^{(2)} = -a_{21}x_{k+1}^{(1)} + y^{(2)} \cdots - a_{2,n-1}x_k^{(n-1)} - a_{2n}x_k^{(n)} \\ \cdots \\ a_{nn}x_{k+1}^{(n)} = -a_{n1}x_{k+1}^{(1)} - a_{n2}x_{k+1}^{(2)} \cdots - a_{n,n-1}x_{k+1}^{(n-1)} + y^{(n)} \end{cases}$$

*** Explicitation des calculs pour la méthode de relaxation**

Faites comme moi ! Testez le sur un bout de brouillon et constatez que ça reste très simple.

6 - Un exemple à suivre pour comparer les diverses méthodes

Collant de près à [A-K], je noterai

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & & \vdots \\ 0 & -1 & 2 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 2 & -1 & 0 \\ \vdots & & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix} \in \mathcal{M}_n(\mathbf{R}).$$

(Notez que j'ai fait deux petites modifications par rapport à ma source ; pour moi la matrice A_n est (n, n) alors que dans [A-K] elle est $(n-1, n-1)$ ce qui se révélait extrêmement malcommode pour en parler longuement ; par ailleurs je n'ai pas reproduit le facteur $(n+1)^2$ que [A-K] a choisi de faire figurer dans cette matrice).

Les propriétés de A qui serviront pour la suite :

La matrice $A_n = (a_{ij})$ vérifie la propriété suivante : $a_{ij} = 0$ dès que $|i - j| > 1$. On dit qu'elle est **tridiagonale**. Elle est par ailleurs symétrique.

Les valeurs propres de A_n sont les n réels

$$\lambda_r = 4\sin^2(r\frac{\pi}{2(n+1)}) \quad (1 \leq r \leq n).$$

(Cela se montre en explicitant les vecteurs propres correspondants).

Enfin précisons à toutes fins utiles que j'abrègerai A_n en A quand je n'ai pas besoin de mettre en relief la dépendance en n.

On va comparer les diverses méthodes itératives sur A. Notez que, comme la diagonale de A est constante, la méthode de Jacobi (ou la méthode de Jacobi relaxée) ne se distinguent pas de la méthode du gradient à pas constant sur cet exemple.

Pourquoi cette A ? :

Pour comprendre d'où vient cette matrice, on se réfèrera à [A-K] page 83. J'en dis deux mots, probablement pas assez pour être compréhensible.

Lorsqu'on veut résoudre numériquement sur $[0, 1]$ l'équation $-u'' = f$, où f est fixée et u est l'inconnue, assujettie à des conditions aux limites relatives à ses valeurs en 0 et en 1, on est amené à choisir un entier n grand et remplacer les fonctions f et u par les vecteurs-colonnes

$$F = \begin{pmatrix} f(\frac{1}{n+1}) \\ f(\frac{2}{n+1}) \\ \vdots \\ f(\frac{n}{n+1}) \end{pmatrix} \text{ et } U = \begin{pmatrix} u(\frac{1}{n+1}) \\ u(\frac{2}{n+1}) \\ \vdots \\ u(\frac{n}{n+1}) \end{pmatrix}.$$

Soit k fixé. En utilisant un développement limité de u au voisinage de $\frac{k}{n+1}$, on voit facilement que : $(n+1)^2[-u(\frac{k-1}{n+1})+2u(\frac{k}{n+1})-u(\frac{k+1}{n+1})]$ est une bonne approximation de $-u''(\frac{k}{n+1})$. En restant volontairement vaseux sur ce qui se passe aux bornes, on voit donc que le vecteur-colonne $(n+1)^2 A_n U$ représente bien la fonction $-u''$; l'équation fonctionnelle $-u'' = f$ est donc transformée en l'équation vectorielle $A_n U = \frac{F}{(n+1)^2}$. (J'ai abusivement simplifié, la prise en compte correcte des conditions aux bords nécessite de compliquer un peu la définition de F au niveau de ses première et dernière composantes).

Résoudre l'équation $A_n X = Y$ n'est donc pas un exemple arbitraire, mais bien un problème très concret, qu'on peut compliquer ensuite en préférant s'intéresser au laplacien en dimension 2 ou 3 plutôt que 1.

7 - La méthode du gradient à pas constant vue sur A

Appliquons à A la méthode du gradient à pas constant en utilisant la matrice scalaire $\frac{I}{\alpha}$. La première question à se poser est de savoir quels α sont utilisables et lesquels ne le sont pas. La discussion qui suit utilise seulement la diagonalisabilité (dans $\mathbf{R}$) de A.

On sait qu'une méthode itérative converge sous la condition $\rho(M^{-1}N) < 1$. On notera $\mathcal{J}$ le $M^{-1}N$ de cette section ; on peut le calculer explicitement : $\mathcal{J} = M^{-1}N = M^{-1}(M-A) = I - M^{-1}A = I - \alpha A$.

Les valeurs propres de $\mathcal{J}$ sont donc les réels $1-\alpha\lambda_i$, où les λ_i sont les valeurs propres de A. Si les valeurs propres de A n'étaient pas de signe constant, la méthode du gradient à pas constant serait vouée à l'échec : l'un au moins de ces réels serait forcément plus grand que 1. Comme toutes les valeurs propres de A sont positives, il se passe des choses intéressantes pour $\alpha > 0$, même si la méthode échoue pour $\alpha < 0$.

Ne regardons donc que les valeurs $\alpha > 0$ et regardons comment évoluent les n valeurs propres de $\mathcal{J}$, quand α varie. Ces valeurs propres sont : $1-\alpha\lambda_n \leq \cdots \leq 1-\alpha\lambda_2 \leq 1-\alpha\lambda_1$.

Lorsque α est à peine plus grand que 0, ces n valeurs propres sont toutes très proches de 1 –et toutes dans l'intervalle $]-1,1[$: la méthode est convergente. La plus grande des valeurs propres en valeur absolue est $1-\alpha\lambda_1$, et donc $\rho(\mathcal{J}) = 1-\alpha\lambda_1$. Faisons croître par la pensée α ; on voit les valeurs propres de $\mathcal{J}$ filer vers la gauche. Une valeur de α est remarquable : celle où les valeurs propres extrêmes sont opposées, soit le α solution de $1-\alpha\lambda_n = -(1-\alpha\lambda_1)$, c'est-à-dire :

$$\alpha = \frac{2}{(\lambda_1+\lambda_n)} = \frac{1}{2\left(\sin^2(\frac{\pi}{n+1}) + \sin^2(\frac{n\pi}{n+1})\right)} = \frac{1}{2\left(\sin^2(\frac{\pi}{n+1}) + \cos^2(\frac{\pi}{n+1})\right)} = \frac{1}{2}.$$

La méthode du gradient à pas fixe a donc convergé tant que α valait moins de $\frac{1}{2}$, et le $\rho(\mathcal{J})$ correspondant diminuait constamment. Au-delà de $\frac{1}{2}$, la valeur propre la plus négative $1-\alpha\lambda_n$ est plus grande en valeur absolue que la valeur propre la plus positive $1-\alpha\lambda_1$ et c'est donc elle qui mène la danse ; désormais $\rho(\mathcal{J}) = |1-\alpha\lambda_n| = \alpha\lambda_n - 1$. Et ceci jusqu'à ce que $1-\alpha\lambda_n$ atteigne la valeur -1, c'est-à-dire pour $\alpha = \frac{2}{\lambda_n}$. Au-delà de cette valeur, la méthode cesse d'être convergente.

Le paramètre optimal à utiliser est celui qui minimise $\rho(\mathcal{J})$; il sera donc opportun d'utiliser la méthode du gradient à pas constant pour la matrice $M = 2I$. Tiens c'est justement la partie diagonale de A : ici la méthode à pas constant optimale coïncide avec la méthode de Jacobi.

Pour une telle méthode, combien de fois faudra-t-il itérer pour obtenir une bonne approximation de X_∞ ? [A-K] fait le calcul pour $n = 100$ en supposant qu'on considère comme "bonne" une approximation à 10^{-2} près et qu'on sait que $\|X_0 - X_\infty\|_2 \leq 1$. Le calcul est faisable sans tricher, car $\mathcal{J}$ est symétrique, donc (pour la norme induite par la norme euclidienne de $\mathbf{R}^n$) $\|\mathcal{J}\|$, qui est égal à la plus grande valeur singulière de $\mathcal{J}$ est aussi égal à $\rho(\mathcal{J})$. On est donc sûr que pour tout k, $\|\mathcal{J}\|^k \leq [\rho(\mathcal{J})]^k$. On sait par ailleurs calculer $\rho(\mathcal{J}) = 1-\frac{1}{2}\lambda_1 = \cos\left(\frac{\pi}{n+1}\right) = 1-\frac{1}{2}\frac{\pi^2}{n^2} + o\left(\frac{1}{n^2}\right)$, et comme 100 est grand, on considère le o final comme négligeable. Tout ceci garantit que l'erreur après la k-ème itération vérifie à peu près l'inégalité :

$$\|X_k - X_\infty\| \leq \left(1-\frac{1}{2}\frac{\pi^2}{n^2}\right)^k \qquad (\text{où} \quad n = 100),$$

cette inégalité permettant à [A-K] de conclure que 9342 itérations seront suffisantes.

8 - La méthode de Jacobi vue sur A

Dans cet exemple particulier, elle coïncide avec la méthode du gradient à pas fixe optimal. On en a donc déjà fait le tour. Je signalerai simplement qu'on pourrait faire semblant de ne pas remarquer qu'on sait calculer les valeurs propres de $\mathcal{J}$ explicitement en fonction de celles de A et ne pas être bloqué pour autant. Voir la remarque 8.2.1 page 160 de [A-K].

9 - La méthode de Gauss-Seidel vue sur A

Dans cette section, comme on l'a déjà fait plusieurs fois, on notera

$$A = D - E - F \quad \text{où} \quad D = 2I,\ E = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \text{ et } F = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

la notation $\mathcal{J}$ continuera à désigner la partie linéaire de l'application affine itérée pour la méthode de Jacobi, c'est-à-dire $\mathcal{J} = I - \frac{1}{2}A = \frac{1}{2}(E+F) = D^{-1}(E+F)$.

L'application affine à itérer pour la méthode de Gauss-Seidel a pour partie linéaire $M^{-1}N$ pour $M = D-E$ et $N = F$; on notera $\mathcal{G}$ cette partie linéaire. Contrairement à ce qui se passait pour Jacobi, le bénéfice de la symétrie de A est perdu : $\mathcal{G}$ n'est, elle, pas symétrique.

On voit mal comment calculer le rayon spectral de $\mathcal{G}$... et pourtant c'est faisable !

Proposition : parce que A est tridiagonale, on a la relation

$$\rho(\mathcal{G}) = [\rho(\mathcal{J})]^2.$$

(on peut même être plus précis : les valeurs propres de $\mathcal{G}$ sont exactement les carrés des valeurs propres de $\mathcal{J}$ et la valeur propre nulle.)

Démonstration : comment peut-on donc utiliser une hypothèse aussi peu familière que la tridiagonalité ? C'est un bidouillage purement matriciel, que j'ai recopié en espérant que ça me permettrait de le comprendre, mais qui m'intrigue toujours autant maintenant que je l'ai sous la main... Voyons si ça va mieux quand je le tape :

on bidouille sur les polynômes caractéristiques :

$$\chi_{\mathcal{J}} = \det(D^{-1}(E+F) - XI) = (-1)^n \det(D^{-1}) \det(XD - E - F);$$

$$\chi_{\mathcal{G}}(X^2) = \det((D-E)^{-1}F - X^2 I) = (-1)^n \det((D-E)^{-1}) \det(X^2(D-E) - F).$$

On introduit alors les matrices :

$$B = X(XD - E - F) = \begin{pmatrix} 2X^2 & -X & 0 & \cdots & \cdots & \cdots & 0 \\ -X & 2X^2 & -X & \ddots & & & \vdots \\ 0 & -X & 2X^2 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 2X^2 & -X & 0 \\ \vdots & & & \ddots & -X & 2X^2 & -X \\ 0 & \cdots & \cdots & \cdots & 0 & -X & 2X^2 \end{pmatrix}$$

et

$$C = X^2(D-E) - F = \begin{pmatrix} 2X^2 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ -X^2 & 2X^2 & -1 & \ddots & & & \vdots \\ 0 & -X^2 & 2X^2 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 2X^2 & -1 & 0 \\ \vdots & & & \ddots & -X^2 & 2X^2 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -X^2 & 2X^2 \end{pmatrix}.$$

Avec ces notations, $\det(B) = X^n \det(XD - E - F)$ et donc $\det(B)$ est égal à une constante réelle près à $X^n \chi_{\mathcal{J}}$. De même et plus directement encore, $\det(C)$ est égal à une constante près à $\chi_{\mathcal{G}}(X^2)$.

Mais -et c'est là que la tridiagonalité intervient- les matrices B et C sont semblables dans $\mathbf{R}(X)$: c'est un simple calcul, la matrice de passage à utiliser étant

$$P = \begin{pmatrix} X & 0 & \cdots & \cdots & 0 \\ 0 & X^2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & X^{n-1} & 0 \\ 0 & \cdots & \cdots & 0 & X^n \end{pmatrix}.$$

donc $\det(B) = \det(C)$, donc $\chi_{\mathcal{G}}(X^2) = X^n \chi_{\mathcal{J}}(X)$.

De cette dernière identité polynomiale résulte ce qui est affirmé sur la relation entre les spectres de $\mathcal{J}$ et de $\mathcal{G}$.

(J'ai fini de taper et ça m'est toujours aussi opaque !).

Ainsi, en gros, une étape d'itération de Gauss-Seidel comprime l'erreur d'un facteur $[\rho(\mathcal{J})]^2$, soit autant que deux étapes d'itération de Jacobi. J'écris "en gros" parce que désormais la matrice qu'on itère à chaque étape n'est pas symétrique, donc, contrairement à ce qui se passait avec Jacobi, $\|\mathcal{G}^k\|$ pourrait bien être significativement plus gros que $[\rho(\mathcal{G})]^k$. Ce point m'a inquiété en lisant [A-K] qui glisse sur la question... et je n'y apporte pas de réponse. Toujours est-il que, si on est prêt à fermer les yeux, on conclura comme dans [A-K] page 167 que le nombre d'itérations nécessaire pour approcher la solution à 10^{-2} près, pour le système venant de A_{100}, et sachant que l'erreur initiale était de 1 au plus est désormais de 4671 (ils ne se sont pas foulés, ils ont exactement divisé par deux la valeur donnée pour Jacobi).

10 - La méthode de relaxation

Là ça devient franchement très ennuyeux, mais je l'écris quand même ; il me semble en effet instructif de se rendre compte que des manipulations de coupeur de cheveux en quatre ne font pas seulement gratter quelques itérations de plus mais fournissent un vrai saut qualitatif : l'ordre de grandeur du nombre d'itérations nécessaire est la racine carrée de celui qui était nécessaire pour Jacobi ou Gauss-Seidel.

Donc ici, il faudra bien choisir un paramètre réel ω et (avec les mêmes notations D, E et F qu'au début du 8) on découpera A en

$$A = M - N \qquad \text{où} \qquad M = \frac{D}{\omega} - E \quad \text{et donc} \quad N = \left(\frac{1}{\omega} - 1\right) D + F.$$

Si on prenait $\omega = 1$, on appliquerait donc la méthode de Gauss-Seidel, mais on va choisir un ω plus tordu.

Avant de choisir ω, il faut comprendre dans quelle gamme de valeurs on doit le choisir ; notons $\mathcal{G}_\omega$ la matrice $M^{-1}N$, on sait que les ω acceptables sont ceux pour lequel $\rho[\mathcal{G}_\omega]$ est strictement plus petit que 1. La réponse est que ceci arrive pour $\omega \in]0,2[$, voyons sommairement d'où ça vient.

Proposition : Si $\omega \notin]0,2[$, la méthode de relaxation diverge pour certains choix de la valeur initiale X_0 (la preuve n'utilisant aucune particularité de la matrice A).

La démonstration en est insolemment simple : le déterminant de la matrice triangulaire inférieure M^{-1} est $\omega^n (\det D)^{-1}$, le déterminant de la matrice triangulaire supérieure N est $\left(\frac{1}{\omega} - 1\right)^n \det D$. Donc on

sait calculer sans mal le $\det[\mathcal{G}_\omega] = (1-\omega)^n$. Si ω n'est pas dans l'intervalle $]0,2[$, la matrice itérée a un déterminant supérieur ou égal à 1 en valeur absolue : il est impossible que toutes ses valeurs propres aient un effet contractant.

Proposition : Parce que A est symétrique définie positive, si $\omega \in]0,2[$ la méthode de relaxation converge.

La démonstration est basée sur le théorème 8.1.5 page 157 dans [A-K] dont la preuve semble beaucoup plus lisible dans [S] page 104 ; le tout étant un calcul de quelques lignes dont je serais bien à mal d'expliquer le sens profond.

C'est maintenant que les calculs deviennent odieux. La suite du programme est de calculer le rayon spectral de $\mathcal{G}\omega$, et la méthode est la même que celle qui a permis de calculer celui de $\mathcal{G} = \mathcal{G}_1$ à la section précédente (elle fonctionne parce que A est tridiagonale). On fait un calcul liant les polynômes caractéristiques de $\mathcal{G}_\omega$ et de $\mathcal{J}$ du même esprit qu'à la section précédente, mais beaucoup plus laid.

Alors que pour $\omega = 1$ on pouvait annoncer assez simplement que les valeurs propres de $\mathcal{G}$ étaient 0 et les β^2 (β parcourant le spectre de $\mathcal{J}$), cette fois on trouve que les valeurs propres de $\mathcal{G}_\omega$ sont les

$$\frac{1}{2}(\beta^2\omega^2 - 2\omega + 2) \pm \frac{\beta\omega}{2}\sqrt{\beta^2\omega^2 - 4(\omega-1)}$$

(suggestion : testez comment ça dégénère si $\omega = 0$, si $\omega = 1$, si $\omega = 2$).

Puis chaque source noircit une bonne page de calculs à partir de cette formule pour déterminer lequel de ces complexes a le plus grand module –calculs que je n'ai vraiment pas eu envie de lire ! Sauf méprise de ma part, ils ne contiennent rien de subtil, c'est un exercice monstrueux de calculs élémentaires sur des complexes. À l'issue du calcul, on a une formule pour $\rho(\mathcal{G}_\omega)$; on étudie alors cette fonction (ça devient un exercice niveau terminale, mais toujours aussi ennuyeux), on trace son graphe (ou plutôt on fait confiance à ceux qui l'ont tracé, vu que c'est le même sur les trois sources, ils doivent ne pas s'être trompés).

L'étude montre alors que la valeur idéale à utiliser, celle ω_{opt} qui minimise $\rho(\mathcal{G}_\omega)$ est une certaine valeur $\omega_{\text{opt}} \in]1,2[$ et on a une formule explicite pour la valeur que prend alors $\rho(\mathcal{G}_{\omega_{\text{opt}}})$.

Dans l'exemple de A, cette valeur est finalement simple, puisque c'est :

$$\rho(\mathcal{G}_{\omega_{\text{opt}}}) = \frac{2}{1+\sin(\pi/(n+1))} - 1$$

dont un développement limité pour n tendant vers l'infini est donc $\rho(\mathcal{G}_{\omega_{\text{opt}}}) = 1 - \dfrac{2\pi}{n} + o\left(\dfrac{1}{n}\right)$.

Comparez ce développement limité à celui concernant la méthode de Jacobi, écrit dans les dernières lignes de la section 7 : on voit que le gain entre les deux méthodes n'est pas un simple facteur constant sur un terme en $1/n^2$, comme ç'avait été le cas en passant à Gauss-Seidel, mais un vrai saut qualitatif.

Glissant encore sur les problèmes dus à la non-symétrie de $\mathcal{G}_\omega$, [A-K] conclut que cette fois, 75 itérations suffiront là où tout à l'heure 9342 étaient nécessaires.