§1.5. Approximation polynomiale/rationnelle

En vue du §1.5., avec Ω K-spectrale,
on a envie d'approcher $f(A)$ par $p(A)$
avec $\|f(A)-p(A)\| \leq K \cdot \|f-p\|_\infty$

Q1 comment evaluer $p(A)$?

Q2 comment choisir p ?

La reponse de Q1 depend de la base utilisée pour exprimer p. Base canonique

$$p(A) = \sum_{j=0}^{m} p_j A^j$$

travail principal : Produit de deux matrices $\in \mathbb{C}^{n\times n}$, disons, de complexité $M(n)$.
Evaluation successive $A^{j+1} = A^j \times A \Rightarrow$ complexité $m \cdot M(n)$
même chose par Horner

1.5.1. Algo de Paterson-Stockmeyer

On ecrit $m \geq r\cdot s$ $\left(m \in [r\cdot s, (r+1)\cdot s - 1]\right)$, $p_j := 0 \ \forall j > m$

et $p(A) = \sum_{j=0}^{r} B_j \cdot (A^s)^j$, $B_j = \underbrace{\sum_{k=0}^{s-1} p_{sj+k} \cdot A^k}_{\text{evaluation Horner } (s-1)M(n)}$

total complexité $(r+s)\cdot M(n)$, minimal pour $r = s \approx \sqrt{m}$

Necessite stockage de $s+2$ matrices (au lieu d'une). Reste attractive pour $p(A)\cdot b$!

Nota On peut montrer qu'il existe une constante C de sorte que, pour la sortie S de 1.5.1. en virgule flottante : $\|p(A) - S\| \leq \frac{C\varepsilon \cdot n^2}{1 - C\varepsilon\cdot n^2} \cdot \|\sum |p_j| \cdot |A|^j\|$ (*)

le second membre pouvant être bien plus grand que $\|p(A)\|$! Cancellation, cause evaluation des sommes partielles de $\exp(z)$ en $z = -5$. ! → TP

L'exo suivant nous donne une idée de l'erreur théorique si on utilise des sommes partielles d'un développement de Taylor.

Exo
1.5.2 Soit f analytique dans $\{|z| \leq s(A)\}$, alors

$$E_n := f(A) - \sum_{j=0}^{n} \frac{f^{(j)}(0)}{j!} A^j = \int_0^1 f^{(n+1)}(t \cdot A) \cdot \frac{(1-t)^n}{n!} A^{n+1} \, dt$$

En déduire que $\|E_n\| \leq \frac{\|A\|^{n+1}}{(n+1)!} \cdot \max_{t \in [0,1]} \|f^{(n+1)}(tA)\|$.

Dans l'exo 1.5.2., avec $f(z) = \cos(z)$ et n impair, il reste à estimer $\max_{t \in [0,1]} \|\cos(tA)\| \leq \cosh(\|A\|)$

de même en vue de 1.5.1 : $\|\sum_{j=0}^{n} \frac{f^{(j)}(0)}{j!} A^j\| \leq \cosh(\|A\|)$.

Pour obtenir alors une erreur relative petite, on souhaite alors que $\cos(\|A\|) / \|\cos(A)\|$ ne soit pas trop grand, ce qui est vrai si $\|A\| \leq 1$ mais faux si $\|A\|$ grand. Par conséquent, il vaut mieux appliquer plusieurs fois des formules de réduction d'argument $\cos(A) = 2\cos^2(\frac{A}{2}) - 1$ avant d'utiliser Taylor. (même chose valable pour l'exponentielle).

Au lieu de travailler avec les sommes de Taylor, on peut également calculer des interpolants dans une base de Newton (→ choisir l'ordre de Leja) et l'évaluer avec Horner (ou 1.5.1. si répétition cyclique des points d'interpolation).
Un choix approprié des points d'interpolation dans $\mathbb{Z}$ permet de répondre à Q2, voir plus tard.

Pour evaluer $R(A)$ pour une fraction rationnelle $R = P/Q$, $\deg P, Q \leq m$ on dispose de au moins 3 possibilités

(a) evaluation $P(A), Q(A) \Rightarrow$ une inversion de complexité $I(n)$

(b) décomposition en termes simples

$$R(A) = \sum a_j (A - z_j I)^{-1}$$ par exemple $\Rightarrow m \cdot I(n)$

attention, les résidus a_j devraient être de taille modérée

(c) Manipulation d'une FC

1.5.3. Fractions continues (FC):

avec α_i, β_i polynômes

L'écriture $C(z) = \beta_0(z) + \cfrac{\alpha_1(z)}{\beta_2(z)} + \cdots$ formelle

désigne une suite de convergents

$$C_m = \beta_0 + \cfrac{\alpha_1}{\beta_1} + \cdots + \cfrac{\alpha_m}{\beta_m} = \frac{P_m}{Q_m}$$ calculé par

$P_{-1} = 1$, $P_0 = \beta_0$ $Q_{-1} = 0$, $Q_0 = 1$ et pour $n \geq 0$

$$\begin{Bmatrix} P_{m+1} \\ Q_{m+1} \end{Bmatrix} = \beta_{m+1} \begin{Bmatrix} P_m \\ Q_m \end{Bmatrix} + \alpha_{m+1} \begin{Bmatrix} P_{m-1} \\ Q_{m-1} \end{Bmatrix}.$$

Avec les choix

(i) $\alpha_{m+1} = -\dfrac{C_{n+1} - C_n}{C_n - C_{n-1}}$, $\beta_{n+1} = 1 - \alpha_{n+1}$

(ii) $\alpha_{m+1} = \dfrac{P_n Q_{n+1} - Q_n P_{n+1}}{P_n Q_{n-1} - Q_n P_{n-1}}$, $\beta_{n+1} = \dfrac{P_{n+1} Q_{n-1} - Q_{n-1} P_{n+1}}{P_n Q_{n-1} - Q_n P_{n-1}}$

on construit une FC avec convergents $C_n = P_n/Q_n$.

Une evaluation numériquement stable se fait à l'aide d'une remontée:

$$C_m^{(n)} = \beta_m, \quad k < m \quad C_k^{(n)} = \beta_{k-1} + \frac{\alpha_{k+1}}{C_{k+1}^{(n)}}, \quad C_0^{(n)} = C_m$$

$\rightarrow$ complexité revient à $m \cdot I(n) + O(n) \cdot M(n)$?

1.5.4. Interpolants à poles prescrits

Pour Q polynôme et $z_1, \ldots, z_m \in \mathbb{C}$, $Q(z_j) \neq 0$, on cherche P polynôme de degré $< m$ de sorte que $R_{m,Q} = P/Q$ interpole f aux points $z_1, \ldots, z_m$.

Formule de Lagrange : si $z_1, \dots, z_m$ distincts

$$R_{m,Q}(z) = B(z) \cdot \sum_{j=1}^{m} \frac{f(z_j)}{(z-z_j)\,B'(z_j)}, \quad B(z) = \frac{\prod_{k=1}^{m}(z-z_k)}{Q(z)}$$

$\Rightarrow$ existence/unicité.

Lien avec interpolation polynomiale

$P = Q \cdot R_{m,Q}$ polynôme d'interpol. de f aux points $z_1, \dots, z_m$

donc $f(z) - R_{m,Q}(z) = B(z) \cdot [z_1, \dots, z_m, z](Q \cdot f)$

Formule d'Hermite

Si f analytique dans Ω et $z_1, \dots, z_m \in \mathcal{K}(\Omega)$

$$f(z) - R_{m,Q}(z) = \frac{B(z)}{2\pi i} \int_{\partial\Omega} \frac{f(\zeta)}{B(\zeta)} \cdot \frac{d\zeta}{\zeta - z}$$

Choix des points d'interpolation ? Idée : Q simule les s-j de f.

Si on se fixe Q et on souhaite erreur petite sur compact $\mathbb{F} \subset \Omega$, il faudra que $\|B\|_{\mathbb{F}} \cdot \left\| \frac{1}{B} \right\|_{\partial\Omega}$ soit le plus petit possible dit "problème de Zolotarev" (si on durcie z_j et Q).

1.5.5. Interpolants à pôles libres/rationnelles

ici de type $[m-1|m]$

Ici pour $z_1, \dots, z_{2m} \in \mathbb{C}$ on cherche à trouver $R_m = P_m/Q_m$ avec $P_m \in \mathcal{P}_{m-1}$, $Q_m \in \mathcal{P}_m \setminus \{0\}$ de sorte $fQ_m - P_m$ s'annule aux $z_1, \dots, z_{2m}$ comptant multiplicités $\Rightarrow$ si $Q_m(z_j) \neq 0$ (ce qui est vrai génériquement)

R_m interpole f aux $z_1, \dots, z_{2m}$: pôles de R_m dignes de singularités de f.

Idée : Pôles de R_m dignes de singularités de f.

Lemme (i) Un tel interpolant existe.

(ii) La fraction P_m/Q_m est unique

(iii) $\forall Q$ de degré $\leq m$: $R_m = R_{2m,Q_m} \cdot Q$ (au cas $Q \neq 0$ générique)

Preuve ad (i) système homogène à $2m$ équations et $2m+1$ inconnues

ad (ii) Soient P_m/Q_m et $\tilde{P}_m/\tilde{Q}_m$ deux solutions à

$$\omega = P_m\tilde{Q}_m - \tilde{P}_m Q_m = (P_m - fQ_m)\tilde{Q}_m - (\tilde{P}_m - f\tilde{Q}_m)Q_m$$

Si $z_1 = z_2 = \dots = z_{2m}$ on parle d'un approximant de Padé en z_1 : $f(z)Q(z) - P(z) = O((z-z_1)^{2m})$ de type $[m-1|m]$

Suivant la preuve du lemme (ii), on arrive à montrer dans le cas "générique" (on a égalité des degrés) que

$$P_{n+j+1}Q_n - P_nQ_{n+j+1} = \gamma_{n,j}\cdot(z-z_1)\cdot\ldots\cdot(z-z_{2n})$$

avec $\gamma_{n,j}$ des polynômes de degré $\le j$, et donc d'après 1.5.3 (ii) $\alpha_{n+1} = -\underbrace{\frac{\gamma_{n,0}}{\gamma_{n-1,0}}}_{\in\mathbb{R}}\cdot(z-z_{2n+1})(z-z_{2n})$

et $\beta_{n+1}\in\mathbb{R}$ dans la CF ayant les convergents R_{nn}.

L'idée des interpolants rationnelles à pôles libres est que les pôles trouvent eux-même la singularité de la fonction, par exemple R_{nn} de $f(z)=\log(z+1)$ pour les approximants de Padé en 0 convergent uniformément dans tout compact de $\mathbb{C}\setminus[-\infty,-1]$, l'intervalle $(-\infty,-1]$ contenant tous les pôles. Ceci est une convergence bien plus intéressante que celle des sommes partielles de Taylor.

Notons que l'on connait explicitement les FC pour les approximants de Padé en 0 pour une grande quantité de fonctions [Baker, Graves-Morris]

Pour une fonction de Markov on peut être plus précis.

1.5.6. Lemme: Soient $f(z)=\int_a^b \frac{d\mu(x)}{z-x}$ (μ pas une fonction rationnelle) et $z_1,\ldots,z_{2n}\in\mathbb{C}\setminus[a,b]$ de sorte que $w_n(z)=\prod_{j=1}^{2n}(z-z_j)$ soit réel sur $\mathbb{R}$. Alors le dénominateur Q_n de l'interpolant rationnel $R_n=P_n/Q_n$ aux points $z_1,\ldots,z_{2n}$ vérifie la relation d'orthogonalité

$$(*)\qquad \int_a^b \frac{Q_n(x)}{w_n(x)}x^j\,d\mu(x)=0,\quad j=0,\ldots,n-1$$

en particulier, tous les racines $r_1,\ldots,r_n$ de Q_n sont simples et appartiennent à $]a,b[$, et

$$P_n(z)/Q_n(z)=\sum_{j=1}^{n}\frac{a_j}{z-r_j}$$ avec $a_1,\ldots,a_n>0$.

Exemple : $f(z) = \frac{1}{z}\log(1+z) = \int_{-\infty}^{-1} \frac{(-1/x)}{z-x}\,dx$ $\quad z_1 = \dots = z_{2m} = 0$

$(*)$ devient $\int_{-\infty}^{-1} Q_m(x)\cdot x^{-2m-1+j}\,dx = 0 \quad j=0,\dots,m-1$

ou encore $\int_0^{-\infty} \underbrace{Q_m\left(-\frac{1}{y}\right)\cdot y^m}_{= \text{const}\cdot L_m(-1+2y)} \cdot y^k\,dy = 0 \quad k=0,1,\dots,m-1$

L_m polynôme de Legendre $\perp$ s/r $[-1,1]$

Preuve : $\forall j=0,\dots,m-1$: $[z_1,\dots,z_{2m}]_z\, z^j\left[Q_m(z)f(z) - P_m(z)\right] = 0$

$$= [z_1,\dots,z_{2m}]_z \left(\int \frac{x^j Q_m(x)\,d\mu(x)}{z-x} - \underbrace{\int \frac{x^j Q_m(x) - z^j Q_m(z)}{z-x}\,d\mu(x) - z^j P_m(z)}_{\text{polynôme en } z \text{ de degré} < m+j \le 2m-1} \right)$$

$$= [z_1,\dots,z_{2m}]_z \int \frac{x^j Q_m(x)\,d\mu(x)}{z-x} \overset{\text{rec}}{=} -\int \frac{x^j Q_m(x)}{w_m(x)}\,d\mu(x) \Rightarrow (*).$$

Pour montrer la dernière partie observons que w_m est réel et ne change pas de signe sur $[a,b]$ par hypothèse. $(*)$ donne un système homogène à m inconnues et $m-1$ équations à coefficients réels, avec

$$\det\left(\int \frac{x^{j+k}}{w_m(x)}\,d\mu(x)\right)_{j,k=0,\dots,m-1} = !$$

$$\int \frac{d\mu(x_1)}{w_m(x_1)} \cdots \int \frac{d\mu(x_m)}{w_m(x_m)} \cdot x_1^0 x_2^1 \cdots x_m^{m-1} \cdot \det\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & & x_m \\ \vdots & & \vdots \\ x_1^{m-1} & & x_m^{m-1} \end{bmatrix}$$

$$= \frac{1}{m!} \underbrace{\int \cdots \int}_{\ge 0} \det\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_m \\ \vdots & & \vdots \\ x_1^m & & x_m^m \end{bmatrix}^2 > 0.$$

$\Rightarrow Q_m$ de degré $= m$, et Q_m à coeff. réels. Si la propriété sur les racines serait fausse alors il existerait P de degré $< m$ de sorte que $Q_m \cdot P$ est réel et ne change pas de signe sur $]a,b[\Rightarrow \int \frac{Q_m(x)P(x)}{w_m(x)}\,d\mu(x) > 0$ en contradiction avec $(*)$. Pour trouver le signe des a_j dans la décomposition en termes simples, on remarque que, d'après 1.5.4. et le lemme (iii),

$$(**) \quad f(z) - \frac{P_m(z)}{Q_m(z)} = \frac{w_m(z)}{Q_m(z)^2}[z_1,\dots,z_{2m},z]\,(Q_m\cdot(Q_m f - P_m))$$

$$= \text{comme avant } \frac{w_m(z)}{Q_m(z)^2}\int \frac{Q_m(x)^2}{w_m(x)}\cdot\frac{d\mu(x)}{z-x}$$

En écrivant $\tilde{Q}_m(z) = Q_m(z)/(z-z_j)$:

$$\frac{P_m(z)}{Q_m(z)} \overset{(**)}{=} \int \left[1 - \left(\frac{x-\overline{r_j}}{z-r_j}\right)^2\right] \frac{W_m(z)}{\widetilde{Q}_m(z)^2} \cdot \frac{\widehat{Q}_m(x)^2}{W_m(x)} \frac{d\mu(x)}{z-x}$$

$$+ \int \left[1 - \frac{W_m(z)}{\widetilde{Q}_m(z)^2} \cdot \frac{\widehat{Q}_m(x)^2}{W_m(x)}\right] \cdot \frac{d\mu(x)}{z-x}$$

fonction rationnelle en z n'admettant pas de pole en z_j

avec $\frac{1}{z-x}\left[1 - \frac{(x-\overline{r_j})^2}{(z-r_j)^2}\right] = \frac{z+x-2\overline{r_j}}{(z-r_j)^2} = \frac{1}{z-r_j} + \frac{x-\overline{r_j}}{(z-r_j)^2}$

donc $\frac{P_m(z)}{Q_m(z)} = \frac{W_m(z)}{\widetilde{Q}_m(z)^2}\left[\frac{1}{z-r_j}\int \frac{\widetilde{Q}_m^2(x)}{W_m(x)} d\mu(x) + \underbrace{\frac{1}{(z-r_j)^2}\int \frac{\widehat{Q}_m(x)\widehat{Q}_m(x)}{W_m(x)} d\mu(x)}_{=0}\right]$

$+$ analytique en z_j

$$= \underbrace{\frac{W_m(r_j)}{\widetilde{Q}_m(r_j)} \cdot \int \frac{\widehat{Q}_m(x)^2}{W_m(x)}}_{= \alpha_j \geq 0} \cdot \frac{1}{z-r_j} + \text{analytique en } z_j$$

∎

1.5.7. Théorème de Markov :

Sous les hypothèses 1.5.6., si $\liminf_{j\to\infty} \operatorname{dist}(z_j, [a,b]) > 0$ alors $R_m \to f$ uniformément sur tout compact $C \subset \mathbb{C}\setminus[a,b]$.

Preuve : M.q.

(***) $\exists C > 0 : \forall m \geq 1, \forall z \in \mathbb{C}\setminus[a,b] : |R_m(z)| \leq \frac{C}{\operatorname{dist}(z,[a,b])}$

Montel (livre de Schiffer)
$\Rightarrow$ (R_m) est normale sur $\mathbb{C}\setminus[a,b]$

Preuve de (***) :

$$|f(z_1)| = \left|\frac{P_m(z_1)}{Q_m(z_1)}\right| = \left|\sum_{j=1}^m \frac{\alpha_j}{z_1 - r_j}\right|$$

$$\geq \left|\operatorname{Im}\sum_{j=1}^m \frac{\alpha_j}{z_1-r_j}\right| = \sum_{j=1}^m \frac{|\alpha_j|\,|\operatorname{Im}(z_1)|}{|z_1-r_j|^2}$$

$\Rightarrow$ si $z_1 \notin \mathbb{R}$: $\sum_{j=1}^m |\alpha_j| \leq \frac{|f(z_1)|}{\operatorname{Im}(z_1)} =: C$

sinon, par exemple $z_1 > b$: $|f(z_1)| \geq \sum_{j=1}^m \frac{|\alpha_j|}{|z_1 - r_j|} \geq \sum \frac{|\alpha_j|}{|z_1 - b|}$

et donc $\sum |\alpha_j| \leq \frac{|f(z_1)|}{\operatorname{dist}(z_1,[a,b])} =: C$ (aussi si $z_1 < a$)

et alors $|R_m(z)| \leq \sum \frac{|a_j|}{|z-r_j|} \leq \frac{\sum_{j=1}^{m} |a_j|}{\text{dist}(z,[a,b])} \leq \frac{C}{\text{dist}(z,[a,b])}$

si Thm est faux alors $\exists \Lambda$ infini $\exists \hat{z}_0 \in \mathbb{C} \setminus [a,b]$ avec $R_m \xrightarrow{\Lambda} c \neq f(\hat{z}_0)$. D'après Montel, $\exists \hat{\Lambda} \subset \Lambda$ infini, $\exists g$ analytique dans $\mathbb{C} \setminus [a,b]$:

$R_m \xrightarrow{\hat{\Lambda}} g$ unif. sur tout sous-ensemble fermé de $\mathbb{C} \setminus [a,b]$.

Notons que $g(z_j) = f(z_j)$ pour $j = 1, 2, 3, \ldots$ comptant multiplicité, et les z_j accumulent sur $\mathbb{C} \setminus [a,b]$ par hypothèse $\Rightarrow g \equiv f \Rightarrow g(\hat{z}) = f(\hat{z})$ contradiction. ∎

1.5.8. Corollaire :

Sous les hypothèses du 1.5.6.,
si $z_1 = z_2 = \ldots \in \mathbb{R}$ (par exemple Padé)
et $r > 0$ t.q. $z_1 - r > b$ alors

$$\| f - R_m \|_{\{|z-z_1| \leq r\}} = |(f - R_m)(z_1 - r)|$$

en particulier, si $\| A - z_1 I \| < z_1 - b$

$$\| f(A) - R_m(A) \| \leq |(f - R_m)(z_1 - \| A - z_1 I \|)|.$$

Preuve $(f - R_m)(z) = \frac{(z-z_1)^{2m}}{Q_m(z)^2} \int \frac{Q_m(x)}{(x-z_1)^{2m}} \cdot \frac{dw(x)}{z-x}$

chaque terme atteint son max dans $\{|z-z_1| \leq r\}$ en module en $z = z_1 - r$. Pour la deuxième partie, voir théorème 1.4.3. de Neumaier. ∎