Czechoslovak Mathematical Journal

Václav Fabian

L'influence de l'arrondissement sur les évaluations numériques linéaires

Czechoslovak Mathematical Journal, Vol. 8 (1958), No. 2, 203–221

Persistent URL: http://dml.cz/dmlcz/100295

Чехословацкий математический журнал, т. 8 (83) 1958, Прага

L'INFLUENCE DE L'ARRONDISSEMENT SUR LES ÉVALUATIONS NUMÉRIQUES LINÉAIRES

VÁCLAV FABIAN, Praha

(Reçu le 10 mai 1957)

Les méthodes analytiques d'estimer les erreurs de chute s'accumulant au cours de calculs numériques compliqués surestiment souvent trop l'erreur actuelle. L'applicabilité des méthodes statistiques supposant que les erreurs élémentaires soient des variables aléatoires non-corrélées et de moyenne nulle, est problématique si l'on arrondit de la manière habituelle. L'influence de l'arrondissement aléatoire (suggeré par G. E. FORSYTHE [6]) est étudiée; cette manière d'arrondir permet d'estimer les erreurs de chute par les méthodes dont l'applicabilité est, dans le cas de l'arrondissement habituel, douteuse ou totalement impossible.

1. L'introduction

Soient $x_0, y_1, y_2, \ldots$ des vecteurs p-dimensionnels, $A_1, A_2, \ldots$ des matrices de type $p \times p$; les éléments des vecteurs et matrices soient des nombres rationels. Si l'on veut calculer le vecteur x_n, défini par les relations

$$x_i = A_i x_{i-1} + y_i \quad (i = 1, 2, \ldots, n)\,, \tag{1.1}$$

on est obligé, par des raisons techniques, à l'arrondir. Alors on calcule les vecteurs ξ_i satisfaisant aux équations

$$\xi_0 = x_0\,, \quad \xi_i = A_i \xi_{i-1} + y_i + \varepsilon_i\,, \quad (i = 1, 2, \ldots, n)\,; \tag{1.2}$$

nous appelons ε_i erreur élémentaire — autrement dit erreur arrondissante — (au i-ième pas) et $\delta_n = \xi_n - x_n$ erreur de chute (au n-ième pas).

La supposition que les ε_i soient des vecteurs aléatoires non corrélés et de moyenne nulle (on fait souvent la supposition plus forte d'indépendence) est problématique si l'on arrondit de la manière habituelle (voir p. ex. [7], [4]), et légitime, comme nous verrons, dans le cas de l'arrondissement aléatoire.

Un nombre a sera arrondi aléatoirement par une observation faite à une variable aléatoire ζ n'acquérant que des valeurs entières[1]) et telle que $\mathbf{E}\zeta = a$, $\mathbf{E}(\zeta - a)^2 < +\infty$. Dans le cas le plus important,

$$P(\zeta = 0) = 1 - a\,, \quad P(\zeta = 1) = a$$

pour $0 \leqq a < 1$ et analogiquement pour les autres a. On arrondit alcatoirement

[1]) Si l'on arrondit aux nombres entiers.

un vecteur en arrondissant aléatoirement (mais pas nécessairement indépendemment) chacun de ses éléments.

Le processus (1.1) sera arrondi de la manière suivante: On pose $\xi_0(\omega) = \xi_0 = x_0$,[2]) on calcule le vecteur $A_1\xi_0(\omega) = y_1$ et, à l'aide d'une expérience $\mathscr{E}_1$, on l'arrondit aléatoirement; le résultat de l'arrondissement est l'observation $\xi_1(\omega)$ de ξ_1.

Supposons que, à l'aide des expériences $\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_{i-1}$, on a déjà observé les valeurs $\xi_0(\omega), \xi_1(\omega), \ldots, \xi_{i-1}(\omega)$. Alors on calcule le vecteur $A_i\xi_{i-1}(\omega) + y_i$ et, à l'aide d'une expérience $\mathscr{E}_i$, indépendante de $\mathscr{E}_1, \ldots, \mathscr{E}_{i-1}$, on l'arrondit; voilà comment on obtient le vecteur $\xi_i(\omega)$, l'observation du vecteur aléatoire ξ_i.

Le processus stochastique ξ_n jouit des propriétés suivantes:

1. L'espérance mathématique $\mathbf{E}\xi_n$ de ξ_n est x_n, c'est-à-dire ξ_n est ce qu'on appelle estimateur non-biaisé de x_n. La supposition de la normalité rend possible la construction d'intervalles de confiance pour x_n, fondée sur la répétition indépendante des observations (deux au moins) de ξ_n. En réalité, l'hypothèse de normalité de ξ_n n'est pas vérifiée de sorte que le critère-t ne peut être appliqué qu'à titre approximatif, cependant, à notre avis, cette approximation est souvent convenable (voir aussi § 5, remarque 4). On s'aperçoit que cette méthode d'estimer n'a aucune analogie dans le cas où l'on arrondit de la manière habituelle.

2. Nous avons donné une formule explicite pour un estimateur D_n de la matrice de covariance du vecteur δ_n ce qui permet de construire des intervalles de confiance exacts pour x_n. Malheureusement le calcul numérique de D_n peut être bien compliqué et peut exiger de nouveaux arrondissements; mais la méthode du no 1 peut être employée pour estimer les éléments de D_n. (Voir pour exemple [4] où D_n a été estimé dans un cas du problème de Dirichlet.)

3. On peut construire une suite Δ_n de vecteurs aléatoires, majorante (dans un sens précisé dans (5.10)) de la suite δ_n. L'observation de Δ_n est plus facile à faire que celle de ξ_n ou de δ_n ce qui donne une autre méthode d'estimer δ_n, la méthode qui peut être utile s'il s'agit, par exemple, du choix du nombre de décimales avec lequel le calcul doit être fait pour être le plus économique possible. Cette méthode non plus n'a aucune analogie dans le cas de l'arrondissement habituel, ce qui est aussi le cas pour les propriétés suivantes.

4. Comme l'auteur a démontré dans [3], si l'on arrondit de la manière P^0 (voir § 3), si (1.1) est la méthode itérative de Seidel (voir [5]) de résolution d'une équation linéaire $x = Ax + y$, si $A_{km+i} = A_i$ pour $i = 1, \ldots, m$; $k = 1, 2, \ldots$, $(A_mA_{m-1} \ldots A_1)^S \to \mathbf{0}$, alors on a

$$P\left(\frac{1}{N}\sum_{n=1}^{N}\xi_n \to x\right) = 1\,,$$

où x est la solution.

[2]) $\xi_n(\omega)$ désigne ici une valeur particulière du vecteur aléatoire ξ_n.

5. De plus, comme nous verrons au § 6, si la solution x est telle que les éléments de $10^m x$ sont des nombres entiers, si l'on arrondit à m décimales, on a

$$P(\xi_n \neq x \text{ seulement pour un nombre fini d'indices } n) = 1\,,$$

aussitôt que certaines conditions assez générales sont satisfaites.

Les résultats de ce travail-ci ont été appliqués pour estimer l'erreur de chute dans la résolution du problème de Dirichlet pour un carré 10×10 par la méthode itérative de Seidel dans [4], où l'auteur discute aussi l'applicabilité des méthodes statistiques dans les cas de diverses manières d'arrondir.

Quelques résultats que nous démontrons ici ((4.3), (4.7)) ont été établis et employés déjà dans [3], mais seulement pour le cas d'un processus-P^0 qui n'y a été défini qu'intuitivement.

Qu'il me soit permis d'exprimer ici mes remercîments vifs à MM. Vladimír Knichal et Ivo Babuška à l'instigation desquels j'ai commencé à m'occuper des questions traitées ici.

2. Remarques préliminaires

Une matrice avec une seule colonne est dite vecteur; les éléments des matrices sont, sauf avis contraire, des nombres réels. Soit p un nombre naturel; s'il n'est dit rien d'autre du type d'une matrice ou d'un vecteur, la matrice sera du type $p \times p$ et le vecteur sera à p dimensions.

Si M est une matrice de type quelconque, $M^{(\alpha\beta)}$ désigne l'élément de la ligne α et de la colonne β, si, en particulier, x est un vecteur, nous écrivons $x^{(\alpha 1)} = x^{(\alpha)}$; par M' nous désignons la matrice transposée à M. Alors si x est un vecteur, xx' est une matrice et $x'x$ est une matrice du type 1×1 — un nombre réel. Les vecteurs particuliers $\mathbf{0}$ et $\mathbf{I}$ sont définis par $\mathbf{0}^{(\alpha)} = 0$ et $\mathbf{I}^{(\alpha)} = 1$ pour $\alpha = 1, \ldots, p$; la matrice $\mathbf{0}$ vérifie $\mathbf{0}^{(\alpha\beta)} = 0$; pour la matrice $\mathbf{1}$ on a $\mathbf{1}^{(\alpha\beta)} = 0$ pour $\alpha \neq \beta$ et $\mathbf{1}^{(\alpha\alpha)} = 1$. Pour deux matrices (ou vecteurs) nous écrivons $M \leqq N$, si $M^{(\alpha\beta)} \leqq N^{(\alpha\beta)}$; nous écrivons $M \prec N$, si $N - M$ est semidéfinie positive, une matrice D étant semidéfinie positive, si $a'Da \geqq 0$ pour chaque vecteur a. Si M est une matrice, nous désignons par diag M le vecteur vérifiant (diag $M)^{(\alpha)} = M^{(\alpha\alpha)}$ pour $\alpha = 1, \ldots, p$; nous posons tr $M = \mathbf{I}'$ diag M. Nous rappelons que $D_1 \prec D_2$ entraîne en particulier diag $D_1 \leqq$ diag D_2 et tr $D_1 \leqq$ tr D_2.

Soient $A_1, A_2, \ldots, y_1, y_2, \ldots$, comme nous avons déjà dit, des matrices et des vecteurs dont les éléments sont des nombres rationnels, soit x_0 un vecteur aux éléments entiers; la suite x_n soit définie par (1.1). Désignons par $\mathbf{A}$ l'ensemble de tous les vecteurs dont les éléments sont des nombres entiers, écrivons

$$\mathbf{R} = \{A_i a + y_i\,;\ \ a \in \mathbf{A},\ i = 1, 2, \ldots\}\,.$$

Désignons aussi par $\mathbf{N}$ l'ensemble des nombres naturels.

Soit $(\Omega, \mathscr{F}, P)$ un espace de probabilité, c'est-à-dire, soit Ω un ensemble non vide, $\mathscr{F}$ und σ-corps de sous-ensembles de Ω (les éléments de $\mathscr{F}$ sont appelés aussi événements), P une mesure définie sur $\mathscr{F}$, $P(\Omega) = 1$. Un événement A est dit positif, si $P(A) > 0$; nous désignons par $\mathscr{F}_+$ le système de rous les événements positifs. Une variable aléatoire est une fonction réelle définie et finie sur Ω et mesurable $(\mathscr{F})$; un vecteur aléatoire ξ est une fonction définie sur Ω dont les valeurs sont des vecteurs réels et pour laquelle $\xi^{(\alpha)}$ est une variable aléatoire pour chaque $\alpha = 1, \ldots, p$ (analogiquement nous introduisons aussi la notion de matrice aléatoire).

L'espérance mathématique d'une variable aléatoire ξ sera désignée par $\mathbf{E}\xi$. L'espérance mathématique $\mathbf{E}\xi$ d'un vecteur aléatoire ξ est définie par

$$(\mathbf{E}\xi)^{(\alpha)} = \mathbf{E}\xi^{(\alpha)} ;$$

la matrice de covariance $\mathbf{D}\xi$ est définie par

$$(\mathbf{D}\xi)^{(\alpha\beta)} = \mathbf{E}\xi^{(\alpha)}\xi^{(\beta)} .$$

D'une manière analogue nous désignons, pour une matrice aléatoire A, par $\mathbf{E}A$ la matrice satisfaisant à

$$(\mathbf{E}A)^{(\alpha\beta)} = \mathbf{E}A^{(\alpha\beta)} ;$$

alors nous pouvons écrire

$$\mathbf{D}\xi = \mathbf{E}\xi\xi' .$$

Désignons par $\mathscr{M}$ l'ensemble de tous les vecteurs aléatoires ξ dont les valeurs appartiennent à l'ensemble $\boldsymbol{A}$ et pour lesquels $\mathbf{E}[\xi^{(\alpha)}]^2 < +\infty$ pour $\alpha = 1, \ldots, p$; écrivons $\mathscr{M}(a) = \{\zeta; \zeta \in \mathscr{M}, \mathbf{E}\zeta = a\}$ pour chaque vecteur a.

ζ étant une matrice aléatoire (de type quelconque) dont l'espérance mathématique existe, W appartenant à $\mathscr{F}_+$, nous désignons par $\mathbf{E}_W\zeta^*$ l'espérance mathématique de ζ, conditionelle par rapport à W. Un système $\mathscr{A}$ forme une décomposition mesurable de $W \in \mathscr{F}$, si $\mathscr{A} \subset \mathscr{F}$; $\bigcup \mathscr{A} = W$; $A \in \mathscr{A}$, $B \in \mathscr{A}$, $A \neq B \Rightarrow A \cap B = \emptyset$. Nous rappelons la propriété suivante:

(2.1) Lemme. *Si* $\mathbf{E}_W\zeta \geqq a$ (ou $\mathbf{E}_W\zeta = a$) *pour chaque élément* $W \in \mathscr{F}_+$ *d'une décomposition dénombrable mesurable de* $A \in \mathscr{F}_+$, *on a* $\mathbf{E}_A\zeta \geqq a$ $(\mathbf{E}_A\zeta = a)$. *En particulier s'il s'agit d'une décomposition de* Ω, *on a* $\mathbf{E}\zeta \geqq a$ $(\mathbf{E}\zeta = a)$.

Nous aurons aussi besoin du

(2.2) Lemme. *Soit* $\mathscr{Z}$ *un système dénombrable de vecteurs aléatoires de dimension quelconque, l'ensemble des valeurs de chaque* $\zeta \in \mathscr{Z}$ *soit dénombrable. Soit* ξ *une fonction vectorielle à dimension quelconque définie sur* Ω. *Supposons que pour chaque* $\omega \in \Omega$ *il existe un système fini* $\mathscr{Z}(\omega) \subset \mathscr{Z}$ *tel que si* $\omega \in \Omega$, $\tilde{\omega} \in \Omega$ *et* $\zeta(\omega) = \zeta(\tilde{\omega})$ *pour chaque* $\zeta \in \mathscr{Z}$ (ω) *on a necessairement aussi* $\xi(\omega) = \xi(\tilde{\omega})$.

Dans ce cas ξ *est un vecteur aléatoire.*

Démonstration: Pour chaque $\omega \in \Omega$ soit

$$A(\omega) = \{\tilde{\omega};\ (\tilde{\omega} \in \Omega,\ \zeta \in \mathscr{Z}(\omega)) \Rightarrow \zeta(\omega) = \zeta(\tilde{\omega})\}\ . \tag{2.2.1}$$

De la dénombrabilité évidente de $\{\mathscr{Z}(\omega);\ \omega \in \Omega\}$ découle celle du système $\{A(\omega);\ \omega \in \Omega\} \subset \mathscr{F}$. ξ étant constant sur chaque $A(\omega)$, l'ensemble des valeurs de ξ est dénombrable lui-aussi et il suffit de montrer que si a est un vecteur, l'ensemble $A = \{\omega;\ \xi(\omega) = a\}$ appartient à $\mathscr{F}$. Mais cela découle immédiatement de ce que A est l'union du système dénombrable $\{A(\omega);\ \xi(\omega) = a\} \subset \mathscr{F}$.

Enfin nous *désignons* par L_i la transformation définie par la relation $L_i x = A_i x + y_i$ pour chaque vecteur x. Alors on peut écrire $x_n = L_n L_{n-1} \ldots \ldots L_1 x_0$. Nous notons que les résultats du § 4 sont indépendants de la supposition $x_i = A_i x_{i-1} + y_i$ et restent valables pour chaque transformation L_i de $\boldsymbol{A}$ dans $\boldsymbol{R}$ (où $\boldsymbol{R} = \bigcup_{i=1}^{\infty} L_i(\boldsymbol{A})$).

3. Definition du processus ξ_n

(3.1) Définition. ξ_n est un *processus-Z* (un processus que l'on obtient en arrondissant aléatoirement le processus (1.1)) si ξ_n sont des fonctions vectorielles sur Ω et que pour chaque $i \in \boldsymbol{N}$, $r \in \boldsymbol{R}$ il existe un vecteur aléatoire $\zeta(i, r)$[3]) et une variable aléatoire m_i satisfaisant aux conditions suivantes:

(3.1.1) Pour chaque $i \in \boldsymbol{N}$, $\omega \in \Omega$ on a $m_i(\omega) \in \boldsymbol{N}$ et

$$\xi_0(\omega) = x_0\ , \quad \xi_i(\omega) = \zeta(m_i(\omega),\ L_i \xi_{i-1}(\omega),\ \omega)\ .$$

(3.1.2) Pour chaque $i \in \boldsymbol{N}$, $\omega \in \Omega$ le vecteur aléatoire

$$\zeta(m_i(\omega),\ L_i \xi_{i-1}(\omega)) \quad \text{appartient à } \mathscr{M}(L_i \xi_{i-1}(\omega))\ .^4)$$

(3.1.3) Pour chaque $k, n \in \boldsymbol{N}$, $a_i \in \boldsymbol{A}$ $(i = 1, \ldots, n-1)$ le vecteur aléatoire $\zeta(k, L_n a_{n-1})$ et l'événement

$$W = \{\omega;\ \xi_i(\omega) = a_i \quad (i = 1, \ldots, n-1)\ ,\quad m_n(\omega) = k\} \tag{3.1.3.1}$$

sont indépendants.

Remarque 1. Pour que la définition ait un sens, il faut que W soit vraiment un ensemble mesurable $\mathscr{F}$. Mais cela découle immédiatement du fait que les ξ_i sont des vecteurs aléatoires ce que nous allons démontrer. En effet, en désignant $\xi = \xi_n$,

$$\mathscr{Z}(\omega) = \{m_i;\ i = 1,2,\ \ldots,\ n\} \cup \{\zeta(m_i(\omega), L_i \xi_{i-1}(\omega));\ i = 1.\ \ldots,\ n\}\ ,$$

$\mathscr{Z} = \bigcup\{\mathscr{Z}(\omega);\ \omega \in \Omega\}$, on peut appliquer (2.2) et en obtenir que ξ_n est un vecteur aléatoire.

[3]) La valeur de $\zeta(i, r)$ pour $\omega \in \Omega$ est désignée par $\zeta(i, r, \omega)$.

[4]) $\zeta(m_i(\omega),\ L_i \xi_{i-1}(\omega))$ est naturellement la fonction qui fait correspondre au point $\tilde{\omega}$ de Ω la valeur $\zeta(m_i(\omega),\ L\xi_{i-1}(\omega),\ \tilde{\omega})$.

Remarque 2. Nous avons formulé la définition dans une généralité considérable afin qu'elle embrasse de divers choix possibles des manières d'arrondir. La supposition que $\zeta(i, r) \in \mathscr{M}(r)$ correspond à l'arrondissement aux nombres entiers. La variable aléatoire m_i donne une règle d'après laquelle on choisit le vecteur arrondissant $\zeta(m_i(\omega), r)$ au i-ième pas. La supposition (3.1.3) est satisfaite si l'on arrondit au n-ième pas à l'aide d'une éxpérience qui a été choisie selon les valeurs des vecteurs $\xi_1, \ldots, \xi_{n-1}$ jusqu'ici observés, mais qui en est indépendante. Dans le théorème suivant nous considérons un cas particulier important du processus-Z.

(3.2) Théorème. *Soit $\zeta(i, r) \in \mathscr{M}(r)$ pour chaque $i \in \mathbf{N}$, $r \in \mathbf{R}$; soit m_i (pour chaque $i \in \mathbf{N}$) une fonction définie sur Ω dont les valeurs appartiennent à $\mathbf{N}$. ξ_n soient définis par* (3.1.1). *Soit $m_n(\omega) = m_n(\tilde{\omega})$ aussitôt que $\xi_i(\omega) = \xi_i(\tilde{\omega})$ pour $i = 1, \ldots, n-1$.*

Supposons que pour chaque n, $k \in \mathbf{N}$, $r \in \mathbf{R}$ le vecteur $\zeta(k, r)$ et le système $\mathscr{Z} = \bigcup\{\mathscr{Z}(\omega); \omega \in W\}$ où W est défini par (3.1.3.1) *et*

$$\mathscr{Z}(\omega) = \{\zeta(m_i(\omega), L_i\xi_{i-1}(\omega)); \ i = 1, 2, \ldots, n-1\},$$

soient indépendants.

Alors ξ_n est un processus-Z.

Remarque. Dans le théorème on suppose que le choix du vecteur arrondissant $\zeta(k, r)$ au n-ième pas est déterminé totalement par les valeurs des $\xi_1, \ldots, \xi_{n-1}$ et que le vecteur choisi $\zeta(k, r)$ est indépendant des vecteurs arrondissants déjà employés.

Démonstration. Evidemment ξ_0 est un vecteur aléatoire et m_0 est une variable aléatoire. Supposons que $\xi_0, \xi_1, \ldots, \xi_{n-1}$ soient des variables aléatoires. D'après (2.2), $m_1, m_2, \ldots, m_n$ et aussi ξ_n sont aléatoires. Alors ξ_n sont des vecteurs aléatoires, m_n des variables aléatoires. La condition (3.1.2) étant évidemment satisfaite, il suffit de démontrer (3.1.3).

Choisissons $k, n \in \mathbf{N}$, $a_i \in \mathbf{A}$ $(i = 1, \ldots, n-1)$, écrivons $r = L_n a_{n-1}$. Si les ensembles $A(\omega)$ sont définis par (2.2.1), nous avons $A(\omega) \subset W$ ou $A(\omega) \cap W = \emptyset$ pour chaque $\omega \in \Omega$.

Alors nous pouvons écrire $W = \bigcup_{i=1}^{\infty} A_i$ où $A_i \in \{A(\omega); \omega \in W\}$. Donc chaque ensemble $B_n = \bigcup_{i=1}^{n} A_i - \bigcup_{i=1}^{n-1} A_i$ est du type

$$\{\omega_i ; \zeta_i(\omega) = r_i \ (i = 1, \ldots, n_1), \quad \eta_i(\omega) \neq s_i \ (i = 1, \ldots, n_2)\},$$

où ζ_i et η_i appartiennent à $\mathscr{Z}$ qui est indépendant de $\zeta(k, r)$; alors B_n sont indépendants de $\zeta(k, r)$ aussi. De là il résulte que $W = \bigcup A_n$ est indépendant de $\zeta(k, r)$, (3.1.3) est vérifié et la démonstration du théorème est terminée.

(3.3) Définition. Si ξ_n est un processus-Z et que les $\zeta(i, r)$ satisfassent aux conditions de la définition (3.1), nous appelons $\zeta(i, r)$ *vecteurs arrondissants*, les variables aléatoires m_i seront appelées *variables déterminant le choix des vecteurs arrondissants*, les vecteurs aléatoires

$$\varepsilon_i = \xi_i - L_i\xi_{i-1} \tag{3.3.1}$$

erreurs arrondissantes et, finalement,

$$\delta_n = \xi_n - x_n \tag{3.3.2}$$

seront les *erreurs de chute*.

Dans la suite nous ne nous servons des symboles ξ_n, ε_n, δ_n, $\zeta(i, r)$, m_n, que dans le sens qu'ils ont dans cette définition—là.

(3.4) Définition. ξ_n est dit *processus- P* si les conditions suivantes sont satisfaites pour chaque $i \in \mathbf{N}$, $r \in \mathbf{R}$, $\mathbf{0} \leqq r \leqq \mathbf{I}$, $a \in \mathbf{A}$, $\alpha = 1, \ldots, p$; $\beta = 1, \ldots, p$; $\alpha \neq \beta$:

$$\zeta(i, r + a) = \zeta(i, r) + a\,,$$
$$P(\zeta^{(\alpha)}(i, r) = 0) = 1 - r^{(\alpha)}\,, \quad P(\zeta^{(\alpha)}(i, r) = 1) = r^{(\alpha)}\,,$$
$$\mathbf{D}^{(\alpha\beta)}[\zeta(i, r) - r] \leqq 0\,.$$

(3.5) Définition. ξ_n est dit *processus-P^0*, s'il est un processus-P et que la matrice $\mathbf{D}(\zeta(i, r) - r)$ soit diagonale.

(3.6) Définition[5]**.** ξ_n est dit *processus-P^-* s'il est un processus-P mais n'est pas un processus-P^0.

(3.7) Lemme. *ξ_n étant un processus-P, on a*

$$\mathbf{D}(\zeta(i, r) - r) \leqq \tfrac{1}{4}\mathbf{1}\,;$$

ξ_n étant un processus-P^0, on a

$$\mathbf{D}(\zeta(i, r) - r) \prec \tfrac{1}{4}\mathbf{1}$$

pour chaque $i \in \mathbf{N}$, $r \in \mathbf{R}$.

Démonstration: Le lemme découle immédiatement des définitions (3.4) et (3.5).

4. Erreurs arrondissantes ε_n

(4.1) *Désignons,* pour chaque $n \in \mathbf{N}$,

$$\mathscr{Z}_n = \{\zeta(k, r) - r;\ \omega \in \Omega,\ m_n(\omega) = k,\ r = L_n\xi_{n-1}(\omega)\}\,.$$

(4.2) Lemme. *Soit Λ une fonction définie sur $\{\zeta(\omega);\ \omega \in \Omega,\ \zeta \in \mathscr{Z}_n\}$ dont les valeurs sont des matrices (de type quelconque), soit $n \in \mathbf{N}$, soit M une matrice. Alors on a*

$$(\zeta \in \mathscr{Z}_n \Rightarrow \mathbf{E}\Lambda\zeta \geqq M) \Rightarrow \mathbf{E}\Lambda\varepsilon_n \geqq M\,,^6) \tag{4.2.1}$$

$$(\zeta \in \mathscr{Z}_n \Rightarrow \mathbf{E}\Lambda\zeta = M) \Rightarrow \mathbf{E}\Lambda\varepsilon_n = M\,. \tag{4.2.2}$$

[5]) C'est M. *Jaroslav Hájek* qui a appelé mon attention à la possibilité d'employer des vecteurs arrondissants pour lesquels $\mathbf{D}^{(\alpha\beta)}(\zeta(i, r) - r) < 0$ pour $\alpha \neq \beta$.

[6]) Par $\Lambda\zeta$ nous avons désigné la fonction composée ($\Lambda\zeta(\omega) = \Lambda(\zeta(\omega))$). Evidemment $\Lambda\zeta$ est une matrice aléatoire.

Démonstration. On s'apperçoit de ce que (4.2.2) résulte de la relation (4.2.1), que nous allons démontrer. Le système des ensembles W (définis au (3.1.3)), pour divers k, a_i, $L_n a_{n-1} = r$, forme une décomposition mesurable de Ω; alors pour montrer (4.2.1) il suffit, en raison du lemme (2.1), de montrer que $\mathbf{E}\Lambda\zeta \geqq M$ étant satisfaite pour chaque $\zeta \in \mathscr{Z}_n$ on a $\mathbf{E}_W \Lambda\varepsilon_n \geqq M$ pour chaque W positif. Pour $\omega \in W$, $m_n(\omega) = k$, $L_n \xi_{n-1}(\omega) = r$ on a $(\zeta(k, r) - r) \in \mathscr{Z}_n$ et le vecteur $\zeta(k, r)$ est indépendant de W (voir (3.1.3)). On a aussi $\varepsilon_n(\omega) = \xi_n(\omega) - L_n \xi_{n-1}(\omega) = \zeta(k, r, \omega) - r$ (voir (3.1.1)). Alors $\mathbf{E}_W \Lambda\varepsilon_n = \mathbf{E}_W \Lambda(\zeta(k, r) - r) = \mathbf{E}\Lambda(\zeta(k, r) - r) \geqq M$ et la démonstration de (4.2.1) et du lemme est terminée.

(4.3) Théorème. *Pour chaque $n \in \mathbf{N}$, on a*

$$\mathbf{E}\varepsilon_n = \mathbf{0} .$$

Démonstration. Posons $\Lambda(r) = r$. De (3.1.2) il s'ensuit que $\mathbf{E}\zeta = \mathbf{0}$ pour chaque $n \in \mathbf{N}$, $\zeta \in \mathscr{Z}_n$ et d'après (4.2.2) il en résulte que $\mathbf{E}\varepsilon_n = \mathbf{0}$.

(4.4) Lemme. *n appartenant à $\mathbf{N}$, D_1 et D_2 étant deux matrices vérifiant, pour chaque $\zeta \in \mathscr{Z}_n$, l'inégalité $D_1 \leqq \mathbf{D}\zeta \leqq D_2$, on a $D_1 \leqq \mathbf{D}\varepsilon_n \leqq D_2$.*

Démonstration. On démontre le théorème en appliquant (4.2.1) pour $\Lambda(r) = rr'$ et $M = D_1$ et pour $\Lambda(r) = -rr'$ et $M = -D_2$.

(4.5) Lemme. *n étant un nombre naturel, D_1, D_2 deux matrices vérifiant $D_1 \prec \mathbf{D}\zeta \prec D_2$ pour chaque $\zeta \in \mathscr{Z}_n$, on a $D_1 \prec \mathbf{D}\varepsilon_n \prec D_2$.*

Démonstration. Soit b un vecteur. Posons $\Lambda(r) = b'(rr' - D_1) b$; on a en vertu de la supposition pour chaque $\zeta \in Z_n$

$$\mathbf{E}\Lambda\zeta = \mathbf{E}b'(\zeta\zeta' - \mathrm{D}_1) b = b'(\mathbf{D}\zeta - D_1) b \geqq 0 .$$

Le lemme (4.2) impliquant $\mathbf{E}\Lambda\varepsilon_n \geqq 0$, on a $b'(\mathbf{D}\varepsilon_n - D_1) b \geqq 0$ pour chaque vecteur b et, par conséquence, $\mathbf{D}\varepsilon_n - D_1$ est semidéfinie positive, c'est-à-dire $D_1 \prec \mathbf{D}\varepsilon_n$. La démonstration de $\mathbf{D}\varepsilon_n \prec D_2$ est tout à fait analogue.

(4.6) Théorème. *ξ_n étant un processus-P, on a $\mathbf{D}\varepsilon_n \leqq \frac{1}{4}\mathbf{1}$ pour chaque $n \in \mathbf{N}$. ξ_n étant un processus-P^0, la matrice $\mathbf{D}\varepsilon_n$ est diagonale et $\mathbf{D}\varepsilon_n \prec \frac{1}{4}\mathbf{1}$ pour chaque $n \in \mathbf{N}$.*

Démonstration. Le théorème est une conséquence immédiate de (3.5), (3.7), (4.4) et (4.5).

(4.7) Théorème. *Pour chaque j, $n \in \mathbf{N}$, $j \neq n$ on a*

$$\mathbf{E}\varepsilon_n \varepsilon_j' = \mathbf{0} . \tag{3.7.1}$$

Démonstration. Supposons que $j < n$. Comme dans la démonstration du lemme (4.2), il suffit de montrer que $\mathbf{E}_W \varepsilon_n \varepsilon_j' = \mathbf{0}$ pour chaque ensemble positif W du type considéré dans (3.1.3). Pour $\omega \in W$ nous avons $\varepsilon_j(\omega) = \xi_j(\omega) -$

$- L_j\xi_{j-1}(\omega) = a_j - L_j a_{j-1}$ et $\varepsilon_n(\omega) = \zeta(k, L_n a_{n-1}, \omega) - L_n a_{n-1}$, le vecteur $\zeta(k, L_n a_{n-1})$ étant indépendant de W. Désignons $L_n a_{n-1} = r$. On a alors

$$\mathbf{E}_W \varepsilon_n \varepsilon_j' = [\mathbf{E}_W(\zeta(k, r) - r)](a_j - L_j a_{j-1})' = \mathbf{0}$$

en vertu de l'égalité

$$\mathbf{E}_W(\zeta(k, r) - r) = \mathbf{E}(\zeta(k, r) - r) = \mathbf{0}$$

(voir (3.1.2)).

Pour $n < j$ on procéde d'une manière analogue.

5. Erreur de chute

(5.1) Théorème. *Pour chaque $n \in \mathbf{N}$ on a $\mathbf{E}\xi_n = x_n$, $\mathbf{E}\delta_n = \mathbf{0}$.*

(5.2) Pour simplifier des expressions que nous allons obtenir pour δ_n nous introduisons la convention suivante. Le symbole $A_n A_{n-1} \dots A_{n+1}$ (ou $A_n \dots A_{n+1}$) désigne la matrice $\mathbf{1}$.

(5.3) Théorème. *Pour chaque $n \in \mathbf{N}$ on a*

$$\delta_n = \sum_{i=1}^{n} A_n A_{n-1} \dots A_{i+1} \varepsilon_i .$$

Démonstration de (5.1) et (5.3). Nous allons démontrer (5.3) d'où (5.1) découle en vertu du théorème (4.3). L'égalité (5.3) est satisfaite pour $n = 1$. Supposons qu'elle le soit pour $n = k - 1$. On a donc

$$\delta_k = \xi_k - x_k = A_k \xi_{k-1} + y_k + \varepsilon_k - A_k x_{k-1} - y_k =$$

$$= A_k \delta_{k-1} + \varepsilon_k = A_k \sum_{i=1}^{k-1} A_{k-1} A_{k-2} \dots A_{i+1} \varepsilon_i + \varepsilon_k =$$

$$= \sum_{i=1}^{k} A_k A_{k-1} \dots A_{i+1} \varepsilon_i ,$$

et (5.3) est satisfaite pour $n = k$ aussi et alors pour chaque $n \in \mathbf{N}$.

(5.4) Lemme. *Pour chaque $n \in \mathbf{N}$ on a*

$$\mathbf{D}\delta_n = \sum_{i=1}^{n} A_n \dots A_{i+1} \mathbf{D}\varepsilon_i A_{i+1}' \dots A_n' .$$

Démonstration.

$$\mathbf{D}\delta_n = \mathbf{E}\delta_n \delta_n' = \mathbf{E}\Big(\sum_{i=1}^{n} A_n \dots A_{i+1} \varepsilon_i\Big)\Big(\sum_{j=1}^{n} A_n \dots A_{j+1} \varepsilon_j\Big)' =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} A_n \dots A_{i+1} (\mathbf{E}\varepsilon_i \varepsilon_j') A_{j+1}' \dots A_n' ,$$

d'où le lemme découle en vertu du théorème (4.7).

Remarque. Le lemme (5.4) ne donne pas la possibilité d'évaluer la matrice de covariance $\mathbf{D}\delta_n$ parce que nous ne connaissons pas les matrices $\mathbf{D}\varepsilon_i$. Il n'est pas permis de poser, $\xi_i(\omega)$ et $m_i(\omega)$ étant observés, $\mathbf{D}\varepsilon_i = \mathbf{D}\zeta(m_i(\omega), L_i\xi_{i-1}(\omega))$. Cependant les résultats du § 4 estimant $\mathbf{D}\varepsilon_i$ rendent possible l'estimation de $\mathbf{D}\delta_n$.

(5.5) Définition. La suite des matrices M_i est dite (ξ_n) réduisante, si chaque M_i est diagonale, ne contient d'autres éléments que 0 et 1 et que, finalement, $M_i\varepsilon_i = \varepsilon_i$ pour chaque i.

Remarque. ξ_n étant un processus-P, $M_i^{(\alpha\beta)} = 0$ pour $\alpha \neq \beta$, $M_i^{(\alpha\alpha)} = 0$ si $A_i^{(\alpha 1)}, \ldots, A_i^{(\alpha p)}$, $y_i^{(\alpha)}$ sont entiers et $M_i^{(\alpha\alpha)} = 1$ autrement, alors $\{M_i\}$ est (ξ_n) réduisante.

(5.6) Théorème. *$\{M_i\}$ étant une suite (ξ_n) réduisante, on a pour chaque $n \in \mathbf{N}$*

$$\mathbf{D}\delta_n = \sum_{i=1}^{n} A_n \ldots A_{i+1} M_i \mathbf{D}\varepsilon_i M_i A'_{i+1} \ldots A'_n .$$

Le théorème n'est qu'une reformulation du lemme (5.4).

(5.7) Théorème. *$\{M_i\}$ étant une suite (ξ_n) réduisante, D_1, D_2 deux matrices vérifiant la relation $D_1 \prec \mathbf{D}(\zeta(k, r) - r) \prec D_2$ pour chaque $k \in \mathbf{N}$, $r \in \mathbf{R}$, on a*

$$\sum_{i=1}^{n} A_n \ldots A_{i+1} M_i D_1 M_i A'_{i+1} \ldots A'_n \prec \mathbf{D}\delta_n \prec \sum_{i=1}^{n} A_n \ldots A_{i+1} M_i D_2 M_i A'_{i+1} \ldots A'_n .$$

Démonstration. Les relations suivantes sont satisfaites pour des matrices B, C, D, E:

$$(B \prec C \Rightarrow DBD' \prec DCD'), \quad (B \prec C, D \prec E \Rightarrow B + D \prec C + E) .$$

De là, du lemme (4.5) et du théorème (5.6), on déduit la proposition du théorème.

(5.8) Théorème. *Soit $\{M_i\}$ une suite (ξ_n) réduisante, soit $A_i \geqq 0$ pour chaque $i \in \mathbf{N}$. Supposons que les matrices D_1, D_2 satisfassent à l'inégalité $D_1 \leqq \mathbf{D}(\zeta(i, r) - r) \leqq D_2$ pour chaque $i \in \mathbf{N}$, $r \in \mathbf{R}$. Alors on a*

$$\sum_{i=1}^{n} A_n \ldots A_{i+1} M_i D_1 M_i A'_{i+1} \ldots A'_n \leqq \mathbf{D}\delta_n \leqq \sum_{i=1}^{n} A_n \ldots A_{i+1} M_i D_2 M_i A'_{i+1} \ldots A'_n .$$

Démonstration. Le théorème découle du lemme (4.2) et du théorème (5.6).

(5.9) Théorème. *Soit $\{M_i\}$ une suite (ξ_n) réduisante. ξ_n étant un processus P^0 on a*

$$\mathbf{D}\delta_n \prec \tfrac{1}{4} \sum_{i=1}^{n} A_n \ldots A_{i+1} M_i A'_{i+1} \ldots A'_n ;$$

ξ_n étant un processus P, l'inégalité $A_i \geqq 0$ étant valable pour chaque $i = 1, 2, \ldots, n$ on a

$$\mathbf{D}\delta_n \leqq \tfrac{1}{4} \sum_{i=1}^{n} A_n \ldots A_{i+1} M_i A'_{i+1} \ldots A'_n .$$

Démonstration. On déduit la première proposition des théorèmes (4.6) et (5.7), la seconde de (4.6) et (5.8) en vertu de $M_i M_i' = M_i$.

Remarque 1. Soit ξ_n^1 un processus-P^0 et ξ_n^2 un processus-P^-, soit $A_i \geqq 0$ pour chaque $i \in \mathbf{N}$. Il semble que le processus ξ_n^2 soit d'une qualité supérieure au processus ξ_n^2 dans le sens qu'on a $\mathbf{D}\delta_n^2 \leqq \mathbf{D}\delta_n^1$ mais non $\mathbf{D}\delta_n^1 \leqq \mathbf{D}\delta_n^2$ où $\delta_n^i = \xi_n^i - x_n$. Malheureusement nous n'avons pas réussi à démontrer une telle relation pour $n > 1$.

Remarque 2. Nous avons déjà dit que les expressions estimant $\mathbf{D}\delta_n$ dans (5.7), (5.8) et (5.9) sont assez compliquées. D'autre part, ces expressions ne dépendent de $x_0, y_1, y_2, \ldots$ que par l'intermédiaire de la suite réduisante $\{M_i\}$, qui peut être commune à une large classe des valeurs $x_0, y_1, y_1, \ldots$ Par exemple le processus (1.1) étant le processus itératif de Ritz ou de Seidel pour résoudre une équation vectorielle $x = Ax + y$, nos expressions estimant $\mathbf{D}\delta_n$ ne dépendent point de y; dans le cas de la méthode de Seidel et du processus-P^0 on peut choisir $M_{i+np}^{(\alpha,\beta)} = 1$ pour $\alpha = \beta = i$ et $M_{i+np}^{(\alpha,\beta)} = 0$ autrement ($i = 1, 2, \ldots, p$).

Remarque 3. Si x_n est un processus itératif de Ritz pour résoudre une équation $x = Ax + y$, on a $A_i = A$, $y_i = y$, $M_i = \mathbf{1}$ pour chaque $i \in \mathbf{N}$. Dans ce cas l'expression $D_n = \sum_{i=1}^{n} A_n \ldots A_{i+1} M_i A_{i+1}' \ldots A_n'$ de (5.9) peut être simplifiée à $\sum_{i=1}^{n} A^{n-i} A'^{n-i}$; si A est symétrique, on a $D_n = \sum_{i=0}^{n-1} A^{2i}$; si $(\mathbf{1} - A^2)^{-1}$ existe, on a $D_n = (\mathbf{1} - A^{2n})(\mathbf{1} - A^2)^{-1}$; si $A \geqq 0$ on a $D_n \leqq (\mathbf{1} - A^2)^{-1}$, ce que sont les expressions considérées déjà par Abramov [1].

Remarque 4. A l'introduction, nous avons suggéré d'employer la supposition de normalité de δ_n pour construire des intervalles de confiance sans connaissance de $\mathbf{D}\delta_n$.

Nous signalons qu'on peut arrondir d'une telle manière que les vecteurs $\zeta^{(1)}(i, r), \ldots, \zeta^{(p)}(i, r)$ soient indépendants et que $\mathbf{E}[\zeta^{(\alpha)}(i, r) - r^{(\alpha)}]^3 = 0$ pour chaque α, i, r. (Par exemple on arrondit le nombre $\frac{1}{4}$ à -1 avec la probabilité $\frac{1}{16}$, à 0 avec la probabilité $\frac{10}{16}$ et à 1 avec la probabilité $\frac{5}{16}$.) Dans ce cas on déduit facilement du lemme (3.1) qu'on a aussi $\mathbf{E}[\varepsilon_i^{(\alpha)}]^3 = 0$ et $\mathbf{E}[\delta_n^{(\alpha)}]^3 = 0$, donc que les erreurs ε_i et δ_n sont approximativement symmétriques. Il nous semble qu'ils le sont aussi dans le cas du processus-P^0 et que l'employement du critère-t robuste donnera des résultats satisfaisants.

Remarque 5. L'indépendance de x_0, y_i de l'expression estimant $\mathbf{D}\delta_n$ dans (5.7) nous permet de construire un processus-Z Δ_n pour les valeurs particulières $x_0 = y_1 = \ldots = \mathbf{0}$, qui majore δ_n dans le sens de la relation $\succ$. L'observation de Δ_n peut être essentiellement plus facile que celle de ξ_n ou que le calcul numérique de $\mathbf{D}\delta_n$ et donne ainsi une méthode simple d'estimer δ_n. La construction et les propriétés de Δ_n seront décrites dans le théorème suivant.

(5.10) Théorème. *Soit ξ_n un processus-P^0, $\{M_i\}$ une suite (ξ_n) réduisante, δ_n l'erreur de chute du processus ξ_n.*

Soit Δ_n un processus-Z correspondant aux valeurs particulières $x_0 = y_1 = \ldots = \mathbf{0}$; supposons ensuite que les vecteurs arrondissants $\varrho(i, r)$ et les variables n_i déterminant le choix des vecteurs arrondissants (du processus Δ_n) satisfassent à la condition

$$\mathbf{D}(\varrho(k, r) - r) \succcurlyeq K M_i \tag{5.10.1}$$

pour chaque $\omega \in \Omega$, $i = 1, 2, \ldots, n$, $k = n_i(\omega)$, $r = A_i \Delta_{i-1}(\omega)$, K étant un nombre positif indépendant de ω, i.

Alors on a

$$\mathbf{D}\delta_n \prec \frac{4}{K} \mathbf{D}\, \Delta_n\,; \tag{5.10.2}$$

si $A_i \geqq \mathbf{0}$ pour chaque $i = 1, 2, \ldots, n$, on peut reformuler la proposition en remplaçant le signe $\succcurlyeq$ par $\geqq$ dans les deux relations (5.10.1) *et* (5.10.2).

Démonstration. On s'aperçoit de ce que de $x_0 = y_i = \mathbf{0}$, il s'ensuit que $x_n = \mathbf{0}$ pour chaque $n \in N$, donc que Δ_n est l'erreur de chute du processus Δ_n. γ_i étant les erreurs arrondissantes correspondant au processus Δ_n, on déduit de (5.10.1) et du lemme (4.5) que $\mathbf{D}\gamma_i \succcurlyeq K M_i$ pour chaque $i \in \mathbf{N}$. Alors en vertu du lemme (5.4) et du théorème (5.9) on obtient

$$\mathbf{D}\, \Delta_n = \sum_{i=1}^{n} A_n \ldots A_{i+1} \mathbf{D}\gamma_i A'_{i+1} \ldots A'_n \succcurlyeq K \sum_{i=1}^{n} A_n \ldots A_{i+1} M_i A'_{i+1} \ldots A'_n \succcurlyeq \tfrac{1}{4} K \mathbf{D}\delta_n\,;$$

si $A_i \geqq \mathbf{0}$ on peut remplacer le signe $\succcurlyeq$ par $\geqq$; donc la démonstration est terminée.

Remarque. Évidemment Δ_n ne peut pas être un processus-P^0, parce que, dans ce cas, on aurait $\varrho(n, \mathbf{0}) = \mathbf{0}$ et par conséquence $\Delta_n = \mathbf{0}$ pour chaque $n \in \mathbf{N}$.

Quant à la question de savoir si l'on peut vraiment observer le processus Δ_n, on voit premièrement, que la condition (5.10.1), n_i étant convenablement choisies, ne se rapporte qu'à un nombre fini de vecteurs r.

Si $\Delta_{i-1}(\omega)$ est déjà observé, on déduit de la propriété de la suite réduisante, que $(\mathbf{1} - M_i) A_i \Delta_{i-1}(\omega)$ appartient à l'ensemble $\mathbf{A}$ des vecteurs dont les éléments sont entiers. On pose $\Delta_i^{(\alpha)}(\omega) = A_i \Delta_{i-1}^{(\alpha)}(\omega)$ pour chaque α vérifiant $M_i^{(\alpha)} = 0$. Pour les autres α on arrondit $A_i \Delta_{i-1}^{(\alpha)}(\omega)$ aléatoirement d'une telle manière que, ζ_α étant la variable aléatoire arrondissant $a^{(\alpha)} = A_i \Delta_{i-1}^{(\alpha)}(\omega)$, on ait $\mathbf{E}\zeta_\alpha = a^{(\alpha)}$, $\mathbf{E}(\zeta_\alpha - a^{(\alpha)})^2 \geqq K$, et que de plus ζ_α soient indépendants et que la condition habituelle de l'indépendance des vecteurs arrondissants aux divers pas du processus soit satisfaite.

Par exemple R étant l'ensemble des nombres a_α pour divers α et i, on peut procéder de la manière suivante: On arrondit 0 à -1 avec une probabilité p $(0 < p \leqq \frac{1}{2})$, à 1 avec la même probabilité p et à 0 avec la probabilité $1 - 2p$.

Alors si ζ_a est la variable arrondissant $a \in R$, on a $\mathbf{E}(\zeta_a - a)^2 > 0$ et on peut poser $K = \inf_{a \in R} \mathbf{E}(\zeta_a - a)^2$. R étant fini, K est positif.

6. Une propriété de convergence

(6.1) Définition. ξ_n est dit *processus-JP*, si les conditions suivantes sont satisfaites:

(6.1.1) ξ_n est un processus-P.

(6.1.2) Pour chaque $i \in \mathbf{N}$, $r \in \mathbf{R}$, $a \in \mathbf{A}$, on a

$$P(\zeta(i, r) = a) = P(\zeta(1, r) = a) .$$

(6.1.3) En désignant

$$\mathbf{A}(r) = \{y \in \mathbf{A};\ -1 < y^{(\alpha)} - r^{(\alpha)} < 1 \quad \text{pour} \quad \alpha = 1, \ldots, p\}$$

on a $P(\zeta(1, r) \in \mathbf{A}(r)) = 1$ pour chaque $r \in \mathbf{R}$: (Il s'ensuit que $P(\xi(1, r) = a) > 0$ pour chaque $r \in \mathbf{R}$, $a \in \mathbf{A}(r)$.)

(6.1.4) Il existe un $m \in \mathbf{N}$ tel que $L_{ms+i} = L_i$ (c'est-à-dire $A_{ms+i} = A_i$ et $y_{sm+i} = y_i$) pour chaque $s \in \mathbf{N}$, $i \in \mathbf{N}$ et que $(A_m \ldots A_1)^s \to \mathbf{0}$, $x_n \to x$.[7])

Remarque. La condition concernant A_i, x_n est satisfaite s'il s'agit d'un processus itératif de Ritz ou de Seidel pour résoudre une équation $x = Ax + y$. La condition concernant $\zeta(i, r)$ est satisfaite par exemple dans le cas d'un processus-P^0.

(6.2) Remarque. Soit ξ_n un processus-JP. Désignons

$$F_{ir} = \{\omega;\ \zeta(i, r, \omega) \text{ non } \in \mathbf{A}(r)\} ;$$

il s'ensuit de (6.1.2) et de (6.1.3) que $P(F_{ir}) = 0$ pour chaque $i \in \mathbf{N}$, $r \in \mathbf{R}$. De la dénombrabilité de $\mathbf{N}$ et $\mathbf{R}$ il résulte donc que $P(\Omega_0) = 0$, où

$$\Omega_0 = \mathbf{U}\{F_{ir};\ i \in \mathbf{N},\ r \in \mathbf{R}\} .$$

Pour $\omega \in \Omega - \Omega_0$ on a maintenant

$$\zeta(i, r, \omega) \in \mathbf{A}(r) \tag{6.2.1}$$

pour chaque $i \in \mathbf{N}$, $r \in \mathbf{R}$.

De là on déduit facilement que pour $\omega \in \Omega - \Omega_0$, $n \in \mathbf{N}$, $\alpha = 1, 2, \ldots, p$ on a

$$-1 < \varepsilon_n^{(\alpha)}(\omega) < 1 . \tag{6.2.2}$$

[7]) $\to$ désigne la convergence de chaque coordonnée.

(6.3) Lemme. *ξ_n étant un processus-JP, l'ensemble*

$$F = \{a_n;\ n \in \mathbf{N},\ a_0 = x_0,\ a_i \in \mathbf{A}(L_i a_{i-1}) \quad (i = 1, \ldots, n)\}$$

est fini et égal à

$$V = \{a;\ P\{\xi_n = a\} > 0\,, \quad n \in \mathbf{N}\}\,.$$

Démonstration: Supposons que ξ_n soit un processus-JP. Il résulte de (6.1.3) que $V \supset F$ et de la remarque (6.2) que $F \subset V$, donc $V = F$. De plus, en donnant à Ω_0 le même sens que dans la remarque (6.2), on a

$$V = \{\xi_n(\omega);\ n \in \mathbf{N},\ \omega \in \Omega - \Omega_0\}\,.$$

Soit $\|x\| = \sqrt{x'x}$ pour chaque vecteur x et $\mathbf{I}M\mathbf{I} = \inf_{\|x\| \leqq \mathbf{1}} \|Mx\|$ pour chaque matrice M. Il y a un tel nombre m que $A_{ms+i} = A_i$ pour chaque $s \in \mathbf{N}$, $i \in \mathbf{N}$. Nous pouvons supposer que la matrice $B = A_m A_{m-1} \ldots A_1$ vérifie l'inégalité $\mathbf{I}B\mathbf{I} < 1$. (Autrement, on déduit de la convergence $B^s \to \mathbf{0}$ que $\mathbf{I}B^s\mathbf{I} \to 0$ et que $\mathbf{I}B^s\mathbf{I} < 1$ pour un s proprement choisi: alors on pose $m' = ms$.)

On a évidemment (voir le théorème (5.3))

$$\delta_{nm} = \sum_{i=1}^{m} A_m A_{m-1} \ldots A_{i+1} \varepsilon_{m(n-1)+i} + B\delta_{m(n-1)}\,. \tag{6.3.1}$$

ξ_n étant un processus-JP, nous avons $-\mathbf{I} \leqq \varepsilon_i(\omega) \leqq \mathbf{I}$ pour chaque $i \in \mathbf{N}$, $\omega \in \Omega - \Omega_0$. On en déduit qu'il existe un nombre C_1 tel que

$$\|\sum_{i=1}^{m-j} A_m \ldots A_{i+1} \varepsilon_{nm+i}\| < C_1 \tag{6.3.2}$$

pour chaque $n \in \mathbf{N}$ et $j = 0, 1, \ldots, m-1$. (Nous désignons, pour chaque vecteur aléatoire ζ, par $\|\zeta\|$ la borne supérieure de l'ensemble $\{\|\zeta(\omega)\|;\ \omega \in \Omega - \Omega_0\}$.)

Nous allons montrer, qu'on a, pour chaque $n \in \mathbf{N}$, $\|\delta_{nm}\| < \dfrac{C_1}{1 - \mathbf{I}B\mathbf{I}}$. La proposition est correcte pour $n = 0$. En supposant qu'elle l'est pour $n = k - 1$ on déduit de (6.3.1) et (6.3.2) que

$$\|\delta_{km}\| \leqq C_1 + \mathbf{I}B\mathbf{I}\,\frac{C_1}{1 - \mathbf{I}B\mathbf{I}} = \frac{C_1}{1 - \mathbf{I}B\mathbf{I}}\,;$$

donc la proposition est établie pour chaque $n \in \mathbf{N}$. On en déduit en vertu de (6.3.2) que pour chaque $n \in \mathbf{N}$, $s = 1, \ldots, m$, on a

$$\|\delta_{nm+s}\| = \|\sum_{j=1}^{s} A_s A_{s-1} \ldots A_{j+1} \varepsilon_{nm+j} + A_s A_{s-1} \ldots A_1 \delta_{nm}\| \leqq C_1 +$$

$$+ \mathbf{I}A_s A_{s-1} \ldots A_1\mathbf{I}\,\|\delta_{nm}\| \leqq C_1 + (C_2 + 1)\,\frac{C_1}{1 - \mathbf{I}B\mathbf{I}}\,,$$

où nous avons posé $C_2 = \max_{s=1,\ldots,p-1} \mathbf{I}A_s \ldots A_1\mathbf{I}$.

Donc nous avons montré, que l'ensemble $\{\delta_n(\omega);\ \omega \in \Omega - \Omega_0,\ n \in \mathbf{N}\}$ est borné par rapport à la norme $\|.\|$. De la convergence $x_n \to x$ on déduit que de même l'ensemble $\{x_n;\ n \in \mathbf{N}\}$ et donc aussi l'ensemble V, sont bornés par rapport à $\|.\|$. Évidemment chaque sous-ensemble borné de $\mathbf{A}$ est fini, ce qui est le cas pour $V = F$.

(6.4) Définition. Une suite ξ_n de vecteurs aléatoires est dite un *processus*

(6.4.1) *avec un nombre fini d'états,* si l'ensemble

$$\{a;\ P(\xi_n = a) > 0,\ n \in \mathbf{N}\} \text{ est fini;}$$

(6.4.2) *markovien,* si pour chaque $n \in \mathbf{N}$, $a_1, \ldots, a_n$,

$$W = \bigcap_{i=1}^{n-1} \{\omega;\ \xi_i(\omega) = a_i\}\,, \quad P(W) > 0$$

on a

$$P_W\{\xi_n = a_n\} = P_{\{\xi_{n-1} = a_{n-1}\}}\{\xi_n = a_n\}\,;$$

(6.4.3) *markovien homogène,* s'il est markovien et qu'il existe une fonction $f(a, b)$ telle que

$$P\{\xi_{n-1} = a\} > 0 \Rightarrow P_{\{\xi_{n-1} = a\}}\{\xi_n = b\} = f(a, b)\,. \tag{6.4.3.1}$$

(6.5) Lemme. *Chaque processus-JP ξ_n est un processus markovien avec un nombre fini d'états.*

Démonstration. ξ_n est un processus avec un nombre fini d'états d'après le lemme (6.3).

Pour chaque $n \in \mathbf{N}$, $k \in \mathbf{N}$, $a_i \in \mathbf{A}$ $(i = 1, \ldots, n)$, l'ensemble

$$V = \{\omega;\ \xi_i(\omega) = a_i \quad (i = 1, \ldots, n-1),\ m_n(\omega) = k\}$$

est, en vertu de (3.1.3), indépendant de $\zeta(k, L_n a_{n-1})$.

Si $P(V) > 0$, on a

$$P_V\{\xi_n = a_n\} = P_V\{\zeta(k, L_n a_{n-1}) = a_n\} = P\{\zeta(k, L_n a_{n-1}) = a_n\}$$

ce qui revient à

$$P_V\{\xi_n = a_n\} = P\{\zeta(1, L_n a_{n-1}) = a_n\} \tag{6.5.1}$$

(voir la définition (6.1)).

Soit maintenant $\mathbf{W}$ le système des ensembles V pour divers $k \in \mathbf{N}$, soit $\mathbf{W}_1$ le système des ensembles V pour divers $k, a_1, \ldots, a_{n-2}$. Alors $\mathbf{W}$ est une décomposition mesurable de l'ensemble

$$W = \{\omega;\ \xi_i(\omega) = a_i \quad (i = 1, \ldots, n-1)\}$$

(supposons que $P(W) > 0$) et $\mathbf{W}_1$ est une décomposition measurable de l'ensemble

$$W_1 = \{\omega;\ \xi_{n-1}(\omega) = a_{n-1}\}\,.$$

Par une double application du lemme (2.1) il s'en ensuit que

$$P_{W}\{\xi_n = a_n\} = P\{\zeta(1, L_n a_{n-1}) = a_n\} = P_{W_1}\{\xi_n = a_n\} .$$

Mais, d'après (6.4.2), c'est justement ce qu'il fallait démontrer.

(6.6) Lemme. *Soit ξ_n un processus-JP, $m \in \mathbf{N}$, $A_{sm+i} = A_i$ pour chaque $s \in \mathbf{N}$, $i \in \mathbf{N}$. Alors ξ_0, ξ_m, ξ_{2m}, ... est un processus markovien homogène avec un nombre fini d'états.*

Démonstration. Une suite partielle d'un processus markovien avec un nombre fini d'états est elle-aussi un processus markovien. Donc il suffit de montrer l'homogénéité de ξ_{nm}, les autres propriétés s'ensuivant du lemme précédent.

Or en raison de (6.5.1) on déduit du lemme (6.5) que

$$P_{\{\xi_{(s-1)m}=a_0\}}\{\xi_{sm} = a_m\} = \sum_{a_1, a_2, \ldots, a_{m-1}} \prod_{i=1}^{m} P_{\{\xi_{(s-1)m+i-1}=a_{i-1}\}}\{\xi_{(s-1)m+i} = a_i\} =$$

$$= \sum_{a_1, a_2, \ldots, a_{m-1}} \prod_{i=1}^{m} P\{\zeta(1, L_{(s-1)m+i}(a_{i-1})) = a_i\} =$$

$$= \sum_{a_1, a_2, \ldots, a_{m-1}} \prod_{i=1}^{m} P\{\zeta(1, L_i(a_{i-1})) = a_i\} .$$

La dernière expression ne dépend pas de s; on peut donc poser

$$f(a, b) = \sum_{a_1, \ldots, a_{m-1}} \prod_{i=1}^{m} P\{\zeta(1, L_i(a_{i-1})) = a_i\} ,$$

où $a_0 = a$, $a_m = b$ et la condition (6.4.3.1) est vérifiée.

(6.7) Définition. *Soit $a \in \mathbf{A}$, $b \in \mathbf{A}$.* Alors b est dit *(a, L)-accessible* s'il existe une suite finie $a_0, a_1, \ldots, a_n$, telle que $a_0 = a$, $a_n = b$, $a_i \in \mathbf{A}(L_i a_{i-1})$ pour $i = 1, 2, \ldots, n$. Si la condition $a_i \in \mathbf{A}(L_i a_{i-1})$ est remplacée par $a_i \in \mathbf{A}(A_i a_{i-1})$, le vecteur b est dit *(a, A)-accessible.*

(6.8) Théorème. *Pour qu'on ait*

$P\{\omega;\ \xi_n(\omega) \neq x$ pour un nombre fini d'indices n seulement$\} = 1$ (6.8.1) *pour chaque choix de $x_0 \in \mathbf{A}$ et pour chaque processus-JP ξ_n, il faut et il suffit que pour chaque $x_0 \in \mathbf{A}$ le vecteur x soit (x_0, L)-accessible.*

Démonstration: Si pour un x_0 (6.8.1) est valable, il existe un tel $\omega \in \Omega - \Omega_0$ que $\xi_n(\omega) = x$. Alors $\xi_0(\omega) = x_0$, $\xi_i(\omega) \in \mathbf{A}_i(L_i \xi_{i-1}(\omega))$ et le vecteur x est (x_0, L)-accessible; donc la condition est nécessaire. Supposons que x soit (a, L)-accessible quel que soit $a \in \mathbf{A}$. On sait déjà que $\gamma_n = \xi_{nm}$ est un processus markovien homogène. Evidemment, x est ce qu'on appelle état sans retour, c'est-à-dire, on a, en raison de (6.2.2) et des relations $L_i x = x$ $(i = 1, 2, \ldots)$, $\xi_n(\omega) = x$ lorsque $\omega \in \Omega - \Omega_0$ et $\xi_i(\omega) = x$ pour un $i < n$.

Mais de plus, pour chaque état $a \in V \subset \mathbf{A}$ (où V est l'ensemble défini au lemme 6.3) il y a une probabilité positive d'une transition à x, ce qui s'ensuit de la supposition (6.1.3) et du fait que x est (a, L)-accessible.

Donc pour chaque a tel que $P\{\gamma_n = a\} > 0$ il existe un nombre $s_a \in \mathbf{N}$ et un nombre positif p_a de façon que

$$P_{\{\gamma_n = a\}}\{\gamma_{n+s_a} \neq x\} \leqq 1 - p_a .$$

Choisissant $p = \min_{a \in V} p_a$, $s = \max_{a \in V} s_a$ (V est fini), et procédant par la voie bien connue, on obtient que

$$P\{\gamma_{ns+s} \neq x\} = \sum_{a \neq x} P\{\gamma_{ns} = a\} \cdot P_{\{\gamma_{ns} = a\}}\{\gamma_{ns+s} \neq x\} \leqq$$
$$\leqq (1 - p) \sum_{a \neq x} P\{\gamma_{ns} = a\} = (1 - p)\, P\{\gamma_{ns} \neq x\} \leqq (1 - p)^{n+1} ,$$

d'où l'on déduit $\sum_{n=0}^{\infty} P\{\gamma_{ns} \neq x\} < +\infty$, ce qui donne d'après le lemme de Borel-Cantelli (voir [2], Chap. III, § 2, th. 1.2)

$$P\{\omega;\ \gamma_{ns}(\omega) \neq x \text{ pour un nombre fini d'indices } n \text{ seulement}\} = 1;$$

désignant l'événement dans { } par A, on a

$\{\omega;\ \xi_n(\omega) \neq x$ pour un nombre fini d'indices n seulement $\} \supset A - \Omega_0$, $P(A - \Omega_0) = 1$; donc la condition est suffisante.

Remarque. Les considérations suivantes servent à éclaircir la condition employée dans le théorème précédent.

(6.9) Lemme. *Soit $x \in \mathbf{A}$. Pour que x soit (a, L)-accessible il faut et il suffit que $\mathbf{0}$ soit $(a - x, A)$-accessible.*

Démonstration. Avant tout nous appelons l'attention sur la relation $L_i x = x$ pour chaque $i \in \mathbf{N}$, qui s'ensuit de la supposition (6.1.4). Donc si l'on a $b_0 = a - x$, $b_i \in \mathbf{A}\ (A_i b_{i-1})$ $(i = 1, \ldots, n-1)$, $b_n = \mathbf{0}$ on peut définir $a_i = b_i + x$. De là on a $a_0 = a$, $a_n = x$, $L_i a_{i-1} = A_i(b_{i-1} + x) + y_i = A_i b_{i-1} + L_i x = A_i b_{i-1} + x$. Donc, b_i appartenant à $\mathbf{A}(A_i b_{i-1})$, $a_i = b_i + x$ appartient à $\mathbf{A}(A_i b_{i-1} + x) = \mathbf{A}(L_i a_{i-1})$. La seconde partie du lemme peut être démontrée d'une manière tout à fait analogue.

(6.10) Lemme. *Soit f une fonction réelle non négative définie sur $\mathbf{A}$ ne s'annulant qu'en $\mathbf{0}$; soit q un nombre, $0 \leqq q < 1$. Supposons que pour chaque $a \in \mathbf{A}$, $i \in \mathbf{N}$ il existe un vecteur $b_i \in \mathbf{A}(A_i a)$ de façon que $f(b_i) \leqq q f(a)$.*

Alors $\mathbf{0}$ est (aA)-accessible pour chaque $a \in \mathbf{A}$.

Démonstration. Soit $a \in \mathbf{A}$; définissons une suite a_n, $a_0 = a$, $a_i \in \mathbf{A}(A_i a_{i-1})$, $f(a_i) \leqq q f(a_{i-1})$. On a $f(a_n) \to 0$, mais du lemme (6.3) il s'ensuit

que l'ensemble $\{f(a_i);\ i \in \mathbf{N}\}$ est fini; donc $f(a_n) = 0$ et $a_n = \mathbf{0}$ pour les indices suffisamment grands.

(6.11) Exemple: Supposons que $A_i = A$, $y_i = y$, ce qui est le cas correspondant à la méthode itérative de Ritz pour résoudre l'équation $x = Ax + y$. Supposons de plus que $\|.\|$ soit une norme des vecteurs telle que pour chaque $r \in \mathbf{R}$ il existe un $a \in \mathbf{A}(r)$ tel que $\|a\| \leqq \|r\|$.[8]) Supposons que $\mathbf{|}A\mathbf{|} < 1$, où $\mathbf{|}\ .\ \mathbf{|}$ est la norme correspondant à $\|\ .\ \|$.

Dans ce cas, $\mathbf{0}$ est (a, A)-accessible pour chaque $a \in \mathbf{A}$, ce qui s'ensuit du lemme précédent si l'on y pose $f = \|\ .\ \|$.

Alors si la solution x de l'équation $x = Ax + y$ appartient à $\mathbf{A}$, si ξ_n est un processus-JP (pour les A_i, y_i particuliers) on a

$P\{\omega;\ \xi_n(\omega) \neq x$ pour un nombre fini d'indices n seulement$\} = 1$.

Malheureusement, nous ne savons pas si cette propriété reste valable aussi dans le cas général où la supposition $\mathbf{|}A\mathbf{|} < 1$ est remplacée par la condition $A^s \to \mathbf{0}$ pour $s \to \infty$. Pour la méthode de Seidel nous n'avons pas non plus déduit des conditions plus simples que celle du lemme (6.9) pour que $\mathbf{0}$ soit (a, A)-accessible pour chaque $a \in \mathbf{A}$.

LITTÉRATURE

[1] *А. А. Абрамов:* О влиянии ошибок округления при решении уравнения Лапласа, Вычислительная математика и вычислительная техника, Сб. *1* АН СССР, Москва 1953.

[2] *J. L. Doob:* Stochastic processes, New York 1953.

[3] *V. Fabian:* Zufälliges Abrunden und die Konvergenz des linearen (Seidelschen) Iterationverfahrens, Mathematische Nachrichten, *16* (1957), 265—270.

[4] *V. Fabian:* Odhad chyby zaokrouhlování při lineárních iteračních procesech, zejména při Seidelově řešení Dirichletova problému pro čtverec 10×10, Aplikace matematiky, *3* (1958), 22—44.

[5] *В. Н. Фадеева:* Вычислительные методы линейной алгебры, Москва 1950.

[6] *G. E. Forsythe:* Note on rounding-off errors, Nat. Bur. Stand., Los Angeles, Calif., 3 pp. (1950).

[7] *H. D. Huskey:* On the precision of a certain procedure of numerical integration, J. Research Nat. Bur. Stand. *42* (1949), 57—62.

[8]) Par exemple les normes $\|x\|_1 = \sqrt{x'x}$, $\|x\|_2 = \operatorname{Max}_{\alpha} |x^{(\alpha)}|$, $\|x\|_3 = \sum_{\alpha=1}^{\infty} |x^{(\alpha)}|$ jouissent de cette propriété.

Резюме

ВЛИЯНИЕ ОКРУГЛЕНИЯ НА ЛИНЕЙНЫЕ ИТЕРАЦИОННЫЕ ВЫЧИСЛЕНИЯ

ВАЦЛАВ ФАБИАН (Václav Fabian), Прага

(Поступило в редакцию 10/V 1957 г.)

В статье исследуется влияние случайного округления (предложенного Г. Е. Форсайтом [6]) на линейный векторный итерационный процесс вида

$$x_i = A_i x_{i-1} + y_i ;$$

округляя на i-м шагу путем прибавления ошибки округления ε_i, получаем последовательность ξ_n, которая удовлетворяет соотношениям

$$\xi_0 = x_0 , \quad \xi_i = A_i \xi_{i-1} + y_i + \varepsilon_i .$$

При случайном округлении ε_i, а также ξ_i, являются векторными случайными переменными, обладающими следующими свойствами:

1. Ожидаемое значение $\mathbf{E}\xi_n$ равно x_n. Итак, ξ_n является несмещенной оценкой вектора x_n.

2. Матрицу ковариантности $\mathbf{D}(\xi_n - x_n)$ можно оценить непосредственно аналитическим путем (теоремы (5.7), (5.8)) (что может представить серьезные трудности при численном расчете), или (теорема (5.10)) простым методом типа Монте Карло.

3. Свойство сходимости, доказанное автором в [3], дополнено в параграфе 6 дальнейшим результатом более специального характера.