行列計算における確率的誤差解析
～ 行列指数関数の計算を例として ～

2022年12月23日

日本応用数理学会
三部会連携「応用数理セミナー」

電気通信大学　情報理工学研究科
山本有作

はじめに

- 行列指数関数 $\exp(A)$

$$\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}, \quad A \in \mathbb{R}^{N \times N}.$$

  - 任意の行列 A に対して収束
  - 応用: 微分方程式の数値解法

- 数値計算法（$\exp(A)$自体を計算するもの）
  - 打切り型テイラー展開
  - 有理関数近似
  - 複素周回積分
  - 固有値分解／シューア分解
  - Scaling & Squaring

⇒
- Aのノルムが小さいときは効率的
  （ODEの数値解法では $\exp(A\Delta t)$ を計算）
- 行列乗算のみで計算可能
  （HPC の観点から最適）

本発表の目的

- 打切り型テイラー展開による $\exp(A)$ の計算法を例題として，近年 Higham らにより提案された新しい確率的誤差解析[1)]の実際について解説する．
- 併せて，以下の点についても触れる．
  - 混合精度計算の利用
  - ブロッククリロフ部分空間法における残差ギャップの評価

- 背景
  - 誤差制御型数値計算への要求（速度と精度のトレードオフ）
  - 超高速・低精度（半精度など）の計算ハードウェアの登場

1) N. J. Higham & T. Mary: A New Approach to Probabilistic Rounding Error Analysis, *SIAM J. Sci. Comput.*, Vol. 41, No. 5, pp. A2815-A2835 (2019).

目次

- はじめに
- 行列指数関数の計算手法
- 誤差解析
  - 打切り誤差
  - 確定的丸め誤差解析
  - 混合精度演算の場合
  - 確率的丸め誤差解析
- 数値実験
- 事後誤差解析への応用
  - ブロッククリロフ部分空間法における残差ギャップの評価
- おわりに

テイラー展開に基づく行列指数関数の計算

- 計算手順
  - 次の手順により，$\exp(A)$ の近似値 $f_n(A)$ を計算すると仮定

$$
\begin{aligned}
A^k &= A^{k-1}A \quad (k = 2, 3, \ldots, n), \\
B_k &= \frac{1}{k!}A^k \quad (k = 2, 3, \ldots, n), \\
f_n(A) &= I + A + \sum_{k=2}^{n} B_k .
\end{aligned}
$$

- 注意
  - 後に述べる混合精度演算では，ある k_1（$\geqq 2$）に対し，$k > k_1$ のときの B_k の計算を低精度演算で行う．
  - 計算にはホーナー法 $f_n(A) = I + A\left(I + \frac{1}{2}A\left(I + \frac{1}{3}A\left(\cdots\left(I + \frac{1}{n}A\right)\cdots\right)\right)\right)$ を使うほうが一般的であるが，低精度で計算する部分を後回しにしてその影響を局限するため，上記の計算法を用いた．

誤差解析

- 打切り誤差
  - 無限精度で計算したときの打切り誤差は，次のように評価される．

$$
\begin{aligned}
\|f_n(A) - \exp(A)\| &= \left\| \sum_{k=n+1}^{\infty} \frac{A^k}{k!} \right\| \\
&\leq \sum_{k=n+1}^{\infty} \frac{\|A\|^k}{k!} \\
&= \frac{\|A\|^{n+1}}{(n+1)!} \sum_{k=0}^{\infty} \frac{\|A\|^k}{(k+n+1)(k+n)\cdots(n+2)} \\
&\leq \frac{\|A\|^{n+1}}{(n+1)!} \sum_{k=0}^{\infty} \frac{\|A\|^k}{k!} \\
&= \frac{\|A\|^{n+1}}{(n+1)!} e^{\|A\|}.
\end{aligned}
$$

  - ここで，ノルムは p ノルム（$p \geqq 1$）またはフロベニウスノルムである．

(確定的)丸め誤差解析

- 準備
  - 浮動小数点演算で計算した量を, $fl(\cdot)$ で表す.
  - 丸め誤差の単位を **u** とする.
    - $fl(a \odot b) = (1+\delta)(a \odot b)$, ただし, $\odot = +, -, *, /$, $|\delta| \leqq \mathbf{u}$.
  - $\gamma_n = n\mathbf{u} / (1 - n\mathbf{u})$ とおく.
    - $(1+\gamma_m)(1+\gamma_n) \leqq 1+\gamma_{m+n}$ などの関係式が成り立つ.
  - $A = (a_{ij})$ に対し, $|a_{ij}|$を要素とする行列を $|A|$ で表す.
  - 行列 A, B に対し, $A \leqq B$ は要素ごとに不等式が成り立つことを表す.

- 定理: 行列乗算の誤差上界
  - $A \in \mathbf{R}^{m\times n}$, $B \in \mathbf{R}^{n\times p}$ のとき, 行列乗算 AB の丸め誤差の上界は次式で与えられる.

$$| fl(AB) - AB | \leqq \gamma_n |A||B|.$$

A^k の誤差

- $\hat{A}_k = fl(A^k)$ とし，ある正数 $\varepsilon_{k-1} > 0$ について

$$|\hat{A}_{k-1} - A^{k-1}| \leq \epsilon_{k-1}|A|^{k-1}$$

が成り立つと仮定する．このとき，まず，次の式が成り立つ．

$$\begin{aligned} |\hat{A}_{k-1}| &\leq |\hat{A}_{k-1} - A^{k-1}| + |A^{k-1}| \\ &\leq \epsilon_{k-1}|A|^{k-1} + |A|^{k-1} \\ &\leq (1+\epsilon_{k-1})|A|^{k-1} \end{aligned}$$

これと行列乗算の丸め誤差の公式を使うと，

$$|\hat{A}_k - A^k| \leq \underline{|\hat{A}_k - \hat{A}_{k-1}A|} + \underline{|\hat{A}_{k-1}A - A^k|}$$

第k回の乗算の誤差　A_{k-1} が持っていた誤差

$$\begin{aligned} &\leq \gamma_N|\hat{A}_{k-1}||A| + |\hat{A}_{k-1} - A^{k-1}||A| \\ &\leq \gamma_N(1+\epsilon_{k-1})|A|^{k-1}|A| + \epsilon_{k-1}|A|^{k-1}|A| \\ &= \underline{\{(1+\gamma_N)\epsilon_{k-1} + \gamma_N\}}\,|A|^k. \end{aligned}$$

これを ε_k と置ける

A^k の誤差（続き）

- ε_k に関する漸化式
  - 前ページの結果より，

$$\epsilon_k = (1+\gamma_N)\epsilon_{k-1} + \gamma_N.$$

  - $\varepsilon_1 = 0$ に注意してこれを解くと，ε_k に対する次の上界が得られる．

$$\begin{aligned} \epsilon_k &= (1+\gamma_N)^{k-1}(1+\epsilon_1) - 1 \\ &= (1+\gamma_N)^{k-1} - 1 \\ &\leq \gamma_{(k-1)N}. \end{aligned}$$

$(1+\gamma_m)(1+\gamma_n) \leqq 1+\gamma_{m+n}$

- A^k の誤差
  - 以上より，A^k の丸め誤差の上界が次のように得られる．

$$|\hat{A}_k - A^k| \leq \gamma_{(k-1)N}|A|^k \quad (k \geq 2).$$

$1/k!$ 倍の誤差

- $B_k = (1/k!)A^k$, $\hat{B}_k = fl(B^k)$ とすると,

$$\left|\hat{B}_k - B_k\right| = \left|fl\left(\frac{1}{k!}\hat{A}_k\right) - \frac{1}{k!}\hat{A}_k\right| + \frac{1}{k!}\left|\hat{A}_k - A^k\right|$$

$|fl(a/b) - a/b| \leqq (1+\mathbf{u})(a/b)$ を利用　前ページの結果を利用

$$\leq \mathbf{u}\cdot\frac{1}{k!}\left|\hat{A}_k\right| + \gamma_{(k-1)N}\cdot\frac{1}{k!}|A|^k$$

$|\hat{A}_k| \leqq |A_k| + |\hat{A}_k - A_k|$ として前ページの結果を利用

$$\leq \mathbf{u}\left(1+\gamma_{(k-1)N}\right)\cdot\frac{1}{k!}|A|^k + \gamma_{(k-1)N}\cdot\frac{1}{k!}|A|^k$$

$$= \left\{(1+\mathbf{u})(1+\gamma_{(k-1)N}) - 1\right\}\cdot\frac{1}{k!}|A|^k$$

$$\leq \gamma_{(k-1)N+1}\cdot\frac{1}{k!}|A|^k \quad (k \geq 2).$$

$(1+\gamma_m)(1+\gamma_n) \leqq 1+\gamma_{m+n}$

級数の和の誤差

- $n+1$ 個の行列の加算

$$\begin{aligned} S_0 &= C_0, \\ S_k &= S_{k-1} + C_k \quad (k = 1, 2, \ldots, n) \end{aligned}$$

に対する丸め誤差の上界は，次の公式により与えられる．

$$\left| \hat{S}_n - S_n \right| \leq \gamma_n \sum_{k=0}^{n} |C_k|$$

- ここで，$C_k = \hat{B}_k = fl((1/k!)A^k)$ なので，

$$|C_k| \leqq |B_k| + |\hat{B}_k - B_k| = (1/k!)|A^k| + |\hat{B}_k - B_k|$$

として前ページの結果を用いる．

全体の誤差

- 全体の丸め誤差

$$
\begin{aligned}
\|fl(f_n(A)) - f_n(A)\|_F &\leq \left| fl\left(\sum_{k=0}^{n} \hat{B}_k\right) - \sum_{k=0}^{n} \hat{B}_k \right| && \text{級数の和の誤差} \\
&\quad + \sum_{k=0}^{n} \left| \hat{B}_k - \frac{1}{k!}\hat{A}_k \right| && \text{1/k! 倍の誤差} \\
&\quad + \sum_{k=0}^{n} \frac{1}{k!} \left| \hat{A}_k - A^k \right| && A_k \text{ の計算の誤差} \\
&\leq \sum_{k=0}^{n} \gamma_{(k-1)N+n+1} \cdot \frac{1}{k!} |A|^k .
\end{aligned}
$$

- 全体の誤差

$$
\|f_n(A) - \exp(A)\|_F \leq \left\{ \frac{\|A\|_F^{n+1}}{(n+1)!} + \gamma_{n+(n-1)N+1} \right\} \exp(\|A\|_F).
$$

混合精度計算の場合

- 計算法
  - $\exp(A)$ のテイラー展開では，高次の項ほど寄与が小さくなるため，この部分を低精度で計算しても，全体の精度が落ちないと期待される．
  - $\hat{A}_k = fl(A^k)$ の計算において，$k \leqq k_1$ のときは高精度で計算し，$k > k_1$ のときは（高精度で計算した A_{k1} に基づき）低精度で計算を行う．

- A_k の誤差
  - 低精度演算での丸め誤差の単位を $\mathbf{u}^s$ とし，$\gamma_n^s = n\mathbf{u}^s/(1-n\mathbf{u}^s)$ とおく．
  - このとき，$\hat{A}_k$ $(k > k_1)$ の丸め誤差の上界は次のように与えられる．

$$|\hat{A}_k - A^k| \leq \left\{(1+\gamma_{(k_1-1)N})(1+\gamma^s_{(k-k_1)N}) - 1\right\}|A|^k.$$

混合精度計算の場合（続き）

- 全体の丸め誤差[2)]

$$|fl(f_n(A)) - f_n(A)| \leq \gamma_n \left(I + \sum_{k=1}^{n} \frac{1}{k!}|A|^k \right) + (1+\gamma_n) \sum_{k=2}^{k_1} \gamma_{(k-1)N+1} \cdot \frac{1}{k!}|A|^k$$

高精度計算部分の誤差

$$+(1+\gamma_n) \sum_{k=k_1+1}^{n} \left\{ (1+\gamma_{(k_1-1)N})(1+\gamma^s_{(k-k_1)N+1}) - 1 \right\} \cdot \frac{1}{k!}|A|^k.$$

低精度計算部分の誤差

- 切り替え点 k_1 の決定
  - （高精度部分の最大誤差）$\simeq$（低精度部分の最大誤差）となるように k_1 を決めると，$\|A\|_F \leqq 1$ の場合，

$$\gamma_{N+1} \cdot \frac{1}{2}\|A\|_F^2 \simeq \left\{ (1+\gamma_{(k_1-1)N})(1+\gamma^s_{N+1}) - 1 \right\} \cdot \frac{1}{(k_1+1)!}\|A\|_F^{k_1+1},$$

  - すなわち，

$$\frac{\mathbf{u}}{\mathbf{u}^s} \simeq \frac{2}{(k_1+1)!}\|A\|_F^{k_1-1}.$$

2) Y. Yamamoto, S. Kudo and T. Hoshi: Error analysis of the truncated Taylor series expansion method for computing matrix exponential, *JSIAM Lett*, to appear.

確率的丸め誤差解析

- 確定的丸め誤差解析の問題点
  - 最悪ケースの解析
    - 計算過程での丸め誤差が，すべて最大値を取り，かつ，全体の誤差を増大させる方向の符号を持つと仮定.
  - 長さ N のベクトルの内積の丸め誤差上界は $O(N\mathbf{u})$.
    - 実際の数値結果で多く見られる $O(N^{1/2}\mathbf{u})$ の振る舞いと乖離.
  - N が非常に大きい場合，あるいは低精度計算の場合は，意味のある誤差上界が得られないことがある.

- 確率的丸め誤差解析
  - 丸め誤差を乱数として扱い，確率的な打ち消し合いを考慮.
  - 長さ N のベクトルの内積の丸め誤差上界は $O(N^{1/2}\mathbf{u})$.
  - 上界は，ある（1に近い）確率で成り立つ確率的上界として得られる.

新しい確率的丸め誤差解析手法

- 従来の確率的丸め誤差解析
  - 丸め誤差に一様分布や正規分布などの簡単な分布を仮定.
    - 実際の丸め誤差の分布を必ずしも反映していない.
  - $O(\mathbf{u}^2)$ の項を無視して成り立つ解析
  - 中心極限定理に基づくため，N が大きい場合にのみ適用可能.

- 新しい確率的丸め誤差解析[1)]
  - 丸め誤差に特定の分布を仮定せず，期待値 0 と独立性のみを仮定.
  - 高次の項についても厳密な解析
  - 集中不等式[3)]に基づくため，任意の N について適用可能.

1) N. J. Higham & T. Mary: A New Approach to Probabilistic Rounding Error Analysis, *SIAM J. Sci. Comput.*, Vol. 41, No. 5, pp. A2815-A2835 (2019).

3) 確率変数の和がその期待値からずれる確率の上界を，ずれの関数として与える不等式.

新しい確率的丸め誤差解析手法(1)

- 定義(確率的誤差解析での γ_n)

$$\tilde{\gamma}_n(\lambda) = \exp\left(\lambda\sqrt{n}\mathbf{u} + \frac{n\mathbf{u}^2}{1-\mathbf{u}}\right) - 1 \quad (\simeq \lambda\sqrt{n}\mathbf{u})$$

- 定理1([1], Theorem 2.4)
  - $\delta_1, \delta_2, \ldots, \delta_n$ を平均 0 で絶対値が $\mathbf{u}$ 以下の独立な乱数とし, $\rho_i = \pm 1$ $(1 \leqq i \leqq n)$とする. このとき, 任意の $\lambda > 0$ に対して, 次の式

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \widetilde{\theta}_n, \quad |\widetilde{\theta}_n| \leq \widetilde{\gamma}_n(\lambda)$$

が次の $P(\lambda)$ 以上の確率で成り立つ.

$$P(\lambda) = 1 - 2\exp\left(-\frac{\lambda^2(1-\mathbf{u})^2}{2}\right).$$

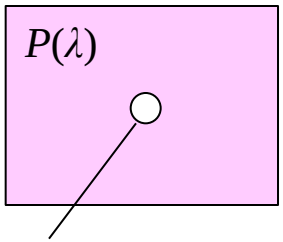
誤差解析で頻繁に現れる式 $\prod_{i=1}^{n}(1+\delta_i)^{\rho_i}$ の確率的上界を与える.

新しい確率的丸め誤差解析手法(2)

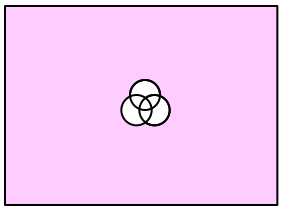
- 定義

$$Q(\lambda, n) = 1 - n(1 - P(\lambda)) = 1 - 2n \exp\left(-\frac{\lambda^2(1-u)^2}{2}\right).$$

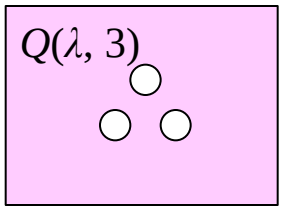
- 意味と利用法
  - 各々が確率$P(\lambda)$で起こる独立な事象n個が同時に起こる確率の下界
  - $\prod_{i=1}^{n}(1+\delta_i)^{\rho_i}$の形の式が複数個所で現れる場合に，その全てが誤差上界以下となる確率を計算するのに使われる.
  - 性質: $1 - \{(1 - Q(\lambda, m)) + (1 - Q(\lambda, n))\} = Q(\lambda, m+n)$



確率 $1 - P(\lambda)$ 以下で起こる事象
（誤差が上界を超える）



$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i}$の形の式が
3カ所で現れる場合



新しい確率的丸め誤差解析手法(3)

- 定理2([1], Theorem 3.1 とその後の注)
  - $\mathbf{a}, \mathbf{b} \in \mathbf{R}^n$ とし, $y = \mathbf{a}^{\mathrm{T}}\mathbf{b}$ とする. このとき, $|fl(y) - y| \leqq \tilde{\gamma}_n(\lambda)|\mathbf{a}|^{\mathrm{T}}|\mathbf{b}|$ が少なくとも確率 $Q(\lambda, n)$ で成り立つ.

確定的誤差解析の γ_n を $\tilde{\gamma}_n(\lambda)$ で置き換えた式

- 略証
  - 通常の誤差解析と同様, $|\varepsilon_i| \leqq \mathbf{u}$, $|\delta_j| \leqq \mathbf{u}$ に対し, 次の式が成り立つ.

$$fl(y) = \sum_{i=1}^{n} a_i b_i (1+\varepsilon_i) \prod_{j=\max(i,2)}^{n} (1+\delta_j)$$

  - ここで, ε_i, δ_j を独立な乱数と見ると, 定理1より, 各項について確率 $P(\lambda)$ 以上で次の不等式が成り立つ.

$$(1+\varepsilon_i) \prod_{j=\max(i,2)}^{n} (1+\delta_j) = 1+\psi_i, \quad |\psi_i| \leq \tilde{\gamma}_{n-\max(i,2)+2}(\lambda)$$

  - 項は n 個あるから, そのすべてが上界を満たす確率は $Q(\lambda, n)$ 以上.

新しい確率的丸め誤差解析手法(4)

- 定理3([1], Theorem 3.4)
  - $A \in \mathbf{R}^{m\times n}$, $A \in \mathbf{R}^{n\times p}$ とし, $C = AB$ とする. このとき, $|fl(C) - C| \leqq \tilde{\gamma}_n(\lambda)\,|A||B|$ が少なくとも確率 $Q(\lambda, mnp)$ で成り立つ.

- 証明
  - 行列乗算 $C = AB$ は長さ n の内積 mp 個からなり, その全てで定理2の上界が成り立たなくてはならない. その確率は $Q(\lambda, mnp)$ 以上.

exp(A) の計算の確率的誤差解析

- 補題4（A_k の誤差）
  - $Q(\lambda, (k-1)N^3)$ 以上の確率で次の不等式が成り立つ．

$$|\hat{A}_k - A^k| \le \underbrace{\tilde{\gamma}_{(k-1)^2 N}(\lambda)}_{\| \cdot} |A|^k \quad (k \ge 2)$$

$$\lambda(k-1)\sqrt{N}\mathbf{u}$$

- 補題4の証明
  - ある正数 $\varepsilon_{k-1} > 0$ について，$Q(\lambda, (k-2)N^3)$ 以上の確率で次の式が成り立つと仮定する．

$$|\hat{A}_{k-1} - A^{k-1}| \le \tilde{\epsilon}_{k-1}(\lambda)|A|^{k-1}$$

  - この式が成り立つとき，次の式も成り立つ．

$$\begin{aligned} |\hat{A}_{k-1}| &\le |\hat{A}_{k-1} - A^{k-1}| + |A^{k-1}| \\ &\le \tilde{\epsilon}_{k-1}(\lambda)|A|^{k-1} + |A|^{k-1} \\ &\le (1 + \tilde{\epsilon}_{k-1}(\lambda))|A|^{k-1} \end{aligned}$$

補題4の証明（続き）

- さらに，定理3より，$\hat{A}^{k-1}$ と A の乗算において，次の不等式が $Q(\lambda, N^3)$ 以上の確率で成り立つ．

$$|\hat{A}_k - \hat{A}_{k-1}A| \leq \tilde{\gamma}_N(\lambda)|\hat{A}_{k-1}||A|.$$

- これらより，

$$\begin{aligned} & 1 - \{(1 - Q(\lambda, (k-2)N^3)) + (1 - Q(\lambda, N^3))\} \\ = \;& 1 - \{(k-2)N^3(1 - P(\lambda)) + N^3(1 - P(\lambda))\} \\ = \;& 1 - (k-1)N^3(1 - P(\lambda)) \;=\; Q(\lambda, (k-1)N^3)) \end{aligned}$$

以上の確率で，次の不等式が成り立つ，

$$\begin{aligned} |\hat{A}_k - A^k| &\leq |\hat{A}_k - \hat{A}_{k-1}A| + |\hat{A}_{k-1}A - A^k| \\ &\leq \tilde{\gamma}_N(\lambda)|\hat{A}_{k-1}||A| + |\hat{A}_{k-1} - A^{k-1}||A| \\ &\leq \tilde{\gamma_N}(\lambda)(1 + \tilde{\epsilon}_{k-1}(\lambda))|A|^{k-1}|A| + \tilde{\epsilon}_{k-1}(\lambda)|A|^{k-1}|A| \\ &= \{\underline{(1 + \tilde{\gamma}_N(\lambda))\tilde{\epsilon}_{k-1}(\lambda) + \tilde{\gamma}_N(\lambda)}\}\,|A|^k. \end{aligned}$$

これを $\tilde{\varepsilon}_k$ と置ける

補題4の証明（続き）

- $\tilde{\varepsilon}_k$ に関する漸化式
  - 前ページの結果より，

$$\tilde{\epsilon}_k(\lambda) = (1+\tilde{\gamma}_N(\lambda))\tilde{\epsilon}_{k-1}(\lambda) + \tilde{\gamma}_N(\lambda).$$

  - $\varepsilon_1 = 0$ に注意してこれを解くと，ε_k に対する次の上界が得られる．

$$\begin{aligned}
\tilde{\epsilon}_k(\lambda) &= (1+\tilde{\gamma}_N(\lambda))^{k-1}(1+\tilde{\epsilon}_1(\lambda)) - 1 \\
&= (1+\tilde{\gamma}_N(\lambda))^{k-1} - 1 \\
&= \exp\left((k-1)\lambda\sqrt{N}\mathbf{u} + \frac{(k-1)N\mathbf{u}^2}{1-\mathbf{u}^2}\right) - 1 \\
&\le \exp\left(\lambda\sqrt{(k-1)^2N}\mathbf{u} + \frac{(k-1)^2N\mathbf{u}^2}{1-\mathbf{u}^2}\right) - 1 \\
&= \tilde{\gamma}_{(k-1)^2N}(\lambda),
\end{aligned}$$

$\tilde{\gamma}$ の形にするために，ここを大きくした．

  - すなわち，$|\hat{A}_k - A^k| \le \tilde{\gamma}_{(k-1)^2N}(\lambda)|A|^k \quad (k \ge 2)$.

exp(*A*) の計算の確率的誤差解析

- 1/*k*! 倍の誤差と級数の和の誤差
  - 1/*k*! 倍の誤差と級数の和の誤差についても考慮し，確率をアップデートすると，次のように全体の丸め誤差の確率的評価が得られる．

- 定理5
  - $Q(\lambda, (n-1)N^3 + (2n-1)N^2)$ 以上の確率で次の不等式が成り立つ．

$$|fl(f_n(A)) - f_n(A)| \leq \tilde{\gamma}_n(\lambda)\left(I + |A| + \sum_{k=0}^{n} \frac{1}{k!}|A|^k\right)$$
$$+(1+\tilde{\gamma}_n(\lambda))\sum_{k=2}^{n} \tilde{\gamma}_{((k-1)\sqrt{N}+1)^2}(\lambda) \cdot \frac{1}{k!}|A|^k.$$

注意

解析過程からわかるように，ある A_k が誤差上界を満たすという事象は，A_l $(k \leqq l)$ が誤差上界を満たすという事象に含まれるので，A_k の誤差の確率については，$k = n$ のときの確率のみを考えればよい．

混合精度計算の場合

- 計算法
  - $\hat{A}_k = fl(A^k)$ の計算において，$k \leqq k_1$ のときは高精度で計算し，$k > k_1$ のときは（高精度で計算した A_{k1} に基づき）低精度で計算を行う．

- 定理6（結果のみ）
  - $Q(\lambda, (n-1)N^3 + (2n-1)N^2)$ 以上の確率で次の不等式が成り立つ．

$$
\begin{aligned}
&|fl(f_n(A)) - f_n(A)| \\
&\quad \leq \tilde{\gamma}_n(\lambda) \left(I + |A| + \sum_{k=0}^{n} \frac{1}{k!} |A|^k \right) \\
&\qquad + (1 + \tilde{\gamma}_n(\lambda)) \sum_{k=2}^{k_1} \tilde{\gamma}_{((k-1)\sqrt{N}+1)^2}(\lambda) \cdot \frac{1}{k!} |A|^k \qquad \text{高精度計算部分の誤差} \\
&\qquad + (1 + \tilde{\gamma}_n(\lambda)) \sum_{k=k_1+1}^{n} \left\{ (1 + \tilde{\gamma}_{(k_1-1)^2 N}(\lambda))(1 + \tilde{\gamma}^s_{((k-k_1)\sqrt{N}+1)^2}(\lambda)) - 1 \right\}. \qquad \text{低精度計算部分の誤差}
\end{aligned}
$$

確率的誤差解析のまとめ

- ポイント

(1) まず，確率を付与すべき事象（アルゴリズム中のある行列積の誤差が一定値以下となるなど）を定義し，その包含関係を整理する.

(2) 通常の誤差解析と同様にして，計算の各ステップで生じる誤差の上界を求める．ただし，γ_n を $\tilde{\gamma}_n(\lambda)$ で置き換える.

(3) 誤差上界を新しく作るごとに，その成立確率についてもアップデートする.

(4) 複数の $\gamma_n(\lambda)$ が現れる式を簡単化するには，次の公式を使う.

$$(1+\tilde{\gamma}_m(\lambda))(1+\tilde{\gamma}_n(\lambda)) \leq 1+\tilde{\gamma}_{(\sqrt{m}+\sqrt{n})^2}(\lambda)$$

・通常の誤差解析における $(1+\gamma_m)(1+\gamma_n) \leq 1+\gamma_{m+n}$ に相当.

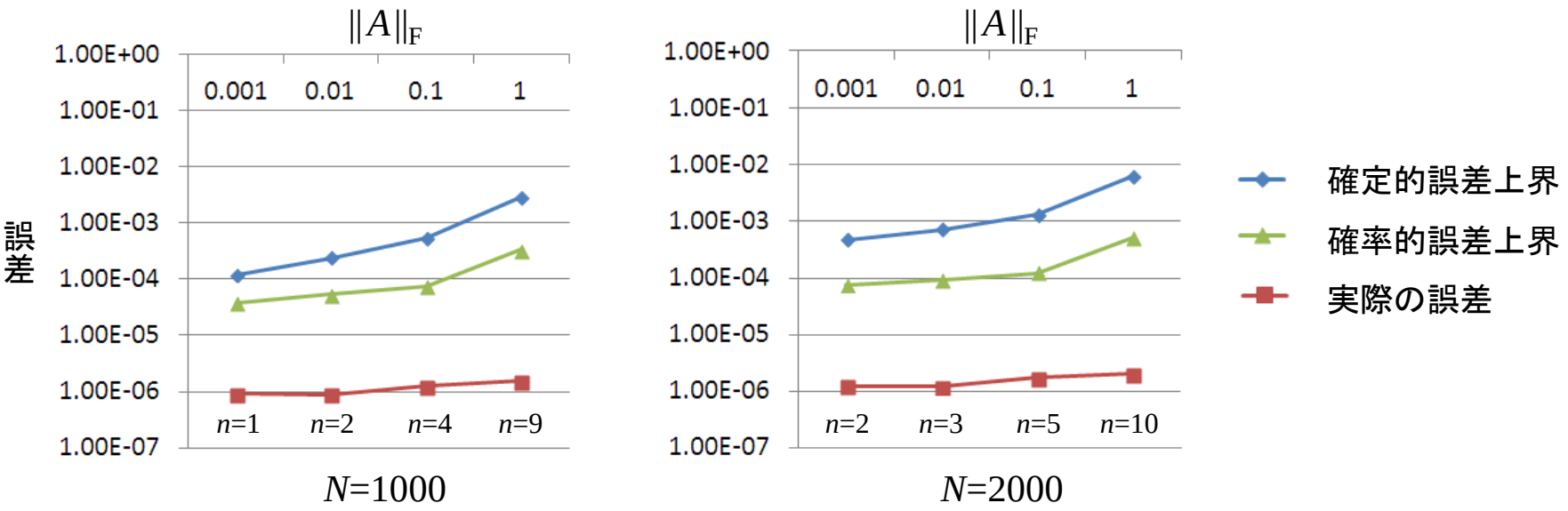
(5) 誤差が求めた上界を超える確率が十分小さな値（10^{-10}以下など）になるよう λ の値を調節する.

数値実験

- 目的
  - テイラー展開法で $\exp(A)$ を計算し，実際の誤差と誤差上界を比較
- 実験条件
  - テスト行列：[-0.5, 0.5] の一様乱数で生成し，スケーリング
  - 計算精度：単精度（一部は半精度）
  - 実際の誤差：倍精度の固有値分解法での結果を真値として算出
  - 確率的誤差上界については，誤差が上界を超える確率が 10^{-10} 以下になるように，$\lambda = 10$ と設定
- 計算機環境
  - Intel Xeon E2660 V2 processor（2.2MHz, 10コア）
  - Intel Fortran compiler
  - Intel Math Kernel Library（LAPACK/BLAS）

数値結果(1)

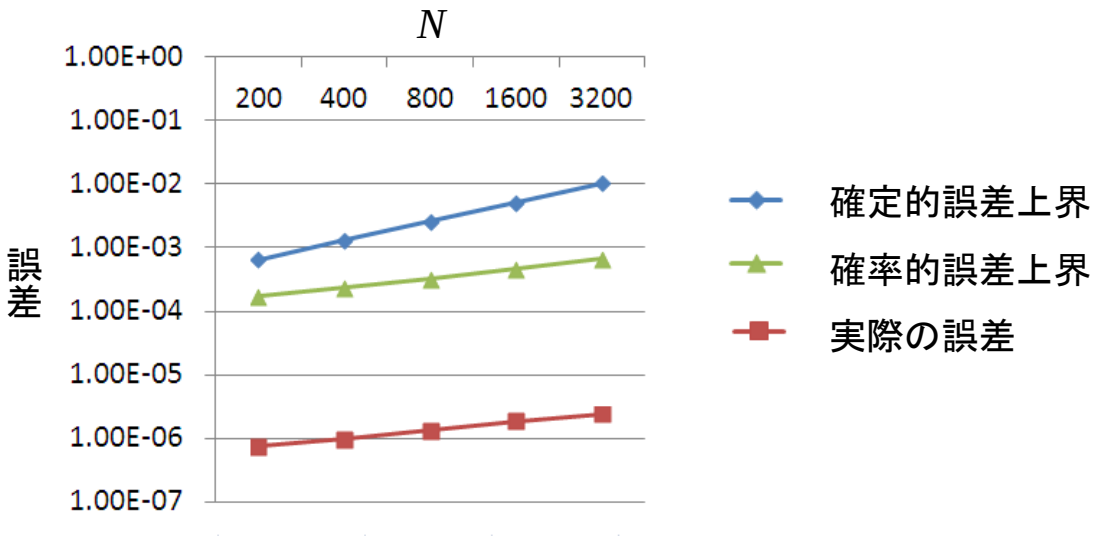
- A のノルムを変えた場合
  - $N = 1000, 2000$ とし，$\|A\|_{\mathrm{F}}$ を 10^{-3} から 1 まで変えて，実際の誤差，確定的誤差上界，確率的誤差上界を比較.
  - テイラー展開の次数 n は，打切り誤差が 10^{-7} 以下になるよう設定.



  - 確率的誤差上界の方が実際の誤差とのギャップが小さい.
  - 実際の誤差とのギャップは，ノルムが大きくなるにつれて増大

数値結果(2)

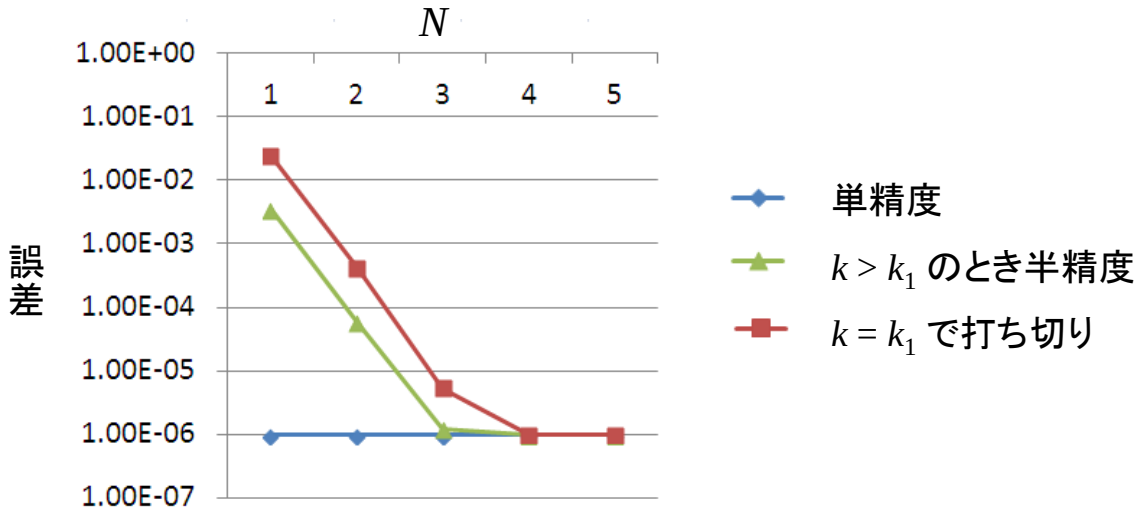
- A のサイズを変えた場合
  - $\|A\|_F = 1$ と固定し，N を 200 から 3200 まで変えて，実際の誤差，確定的誤差上界，確率的誤差上界を比較．



  - 確定的誤差上界は N に比例して増加
  - 確率的誤差上界は $N^{1/2}$ に比例し，実際の誤差により近い結果

数値結果(3)

- 半精度演算による結果
  - $N = 400, \|A\|_{\mathrm{F}} = 1$ とし，$k > k_1$ のときの A^k の計算を半精度で実行
  - 半精度演算は，単精度演算の結果の仮数部を 11 ビットに切り捨てて エミュレート
  - 次数を $k = k_1$ で打ち切った場合の結果も同時にプロット



  - 4次以上の項を半精度で計算しても，単精度とほぼ同じ精度を達成
  - しかし，完全な打ち切りと比べた精度向上はあまり大きくない

事後誤差解析への応用(1)

- 事前誤差解析
  - 入力データから計算される量(ノルム，条件数など)のみを使って丸め誤差上界を導くタイプの誤差解析
  - 計算を行う前に丸め誤差を推定できる.
  - これまで説明した誤差解析事例はすべて事前誤差解析

- 事後誤差解析
  - 浮動小数点による計算結果を誤差上界に含むタイプの誤差解析
  - 計算後(or 計算途中)でないと誤差上界の値が求められない.
  - 事前誤差解析よりもタイトな誤差上界が得られる場合が多い.
  - 事前誤差解析が困難でも事後誤差解析は可能な場合がある.
    - LU 分解の後退誤差 $A + E = \hat{L}\hat{U}$ など

事後誤差解析への応用(2)

- ブロッククリロフ部分空間法
  - 複数右辺を持つ方程式 $AX = B$ ($A \in \mathbf{R}^{n\times n}$, $B, X \in \mathbf{R}^{n\times s}$)を解くためのクリロフ部分空間法
  - ブロッククリロフ部分空間 $\mathcal{K}_k^B(A, R_0) := \left\{ \sum_{i=0}^{k-1} A^i R_0 \gamma_i \mid \gamma_i \in \mathbb{R}^{s\times s} \right\}$ を用いて解を求める.

- 残差ギャップの問題
  - ブロッククリロフ部分空間法では, 残差ベクトル $\hat{R}_k$ を近似解 $\hat{X}_k$ から $\hat{R}_k = B - A\hat{X}_k$ と計算するのではなく, 漸化式により計算する.
  - そのため, アルゴリズム中の残差 $\hat{R}_k$ と真の残差 $B - A\hat{X}_k$ との間に食い違いが生じることがある. これを残差ギャップと呼ぶ.
  - 残差ギャップは偽収束等の問題をもたらすため, その評価は重要[4)]

    ⇨ 確率的な事後誤差解析を用いた残差ギャップの上界の導出

4) K. Aihara, A. Imakura and K. Morikuni: Cross-interactive residual smoothing for global and llock Lanczos-type solvers for linear systems with multiple right-hand sides, *SIMAX*, Vol. 43, No. 3 (2022).

残差ギャップの上界の導出(1)

- 解の更新式と残差の更新式
  - 列直交なブロック進行方向ベクトル $Q_{k-1} \in \mathbf{R}^{n \times s}$ と係数 $\alpha_{k-1} \in \mathbf{R}^{s \times s}$ を用いて X_k と R_k を次のように更新する.

$$X_k = X_{k-1} + Q_{k-1}\alpha_{k-1}$$
$$R_k = R_{k-1} - (AQ_{k-1})\alpha_{k-1}$$

- α_{k-1} のノルムの評価
  - 漸化式より, $Q_{k-1}\alpha_{k-1} = X_k - X_{k-1}$.
  - Q_{k-1}が列直交行列であることより,

$$\begin{aligned} ||\alpha_{k-1}||_F &= ||Q_{k-1}\alpha_{k-1}||_F \\ &= ||X_k - X_{k-1}||_F \\ &\leq ||X_k||_F + ||X_{k-1}||_F. \end{aligned}$$

事後評価

残差ギャップの上界の導出(2)

- $\hat{X}_k$ の確率的誤差評価
  - 次式で定義される誤差 E_1, E_2 を評価する.

$$\begin{aligned}\hat{X}_k &= fl(\hat{X}_{k-1} + fl(\hat{Q}_{k-1}\hat{\alpha}_{k-1})) \\ &= \hat{X}_{k-1} + (\hat{Q}_{k-1}\hat{\alpha}_{k-1} + E_k^{(1)}) + E_k^{(2)}.\end{aligned}$$

行列乗算の誤差　　加算の誤差

  - 定理3(行列乗算の誤差)より, 確率 $Q(\lambda, ns^2)$ 以上で次が成り立つ.

$$|E_k^{(1)}| \leq \tilde{\gamma}_s(\lambda)|\hat{Q}_{k-1}|\,|\hat{\alpha}_{k-1}|.$$

$$\begin{aligned}\therefore \|E_k^{(1)}\|_F &\leq \tilde{\gamma}_s(\lambda)\||\hat{Q}_{k-1}|\|_F\,\|\hat{\alpha}_{k-1}\|_F \\ &\leq \tilde{\gamma}_s(\lambda)\sqrt{s}(\|\hat{X}_k\|_F + \|\hat{X}_{k-1}\|_F).\end{aligned}$$

確率的事後誤差上界

  - 一方, $|E_k^{(2)}| \leq \mathbf{u}(\hat{X}_{k-1} + fl(\hat{Q}_{k-1}\hat{\alpha}_{k-1})) \simeq \mathbf{u}|\hat{X}_k|$. 確定的事後誤差上界
  - これらより, 確率 $Q(\lambda, ns^2)$ 以上で,

$$\|E_k^{(1)} + E_k^{(2)}\|_F \lesssim (\lambda s + 1)\mathbf{u}\||\hat{X}_k|\|_F + \lambda s\mathbf{u}\||\hat{X}_{k-1}|\|_F.$$

残差ギャップの上界の導出(3)

- $\hat{R}_k$ の確率的誤差評価
  - 次式で定義される誤差 $E_k^{(3)}, E_k^{(4)}, E_k^{(5)}$ を評価する.

$$\begin{aligned}\hat{R}_k &= fl(\hat{R}_{k-1} + fl(fl(A\hat{Q}_{k-1})\hat{\alpha}_{k-1})) \\ &= \hat{R}_{k-1} + ((A\hat{Q}_{k-1} + E_k^{(3)})\hat{\alpha}_{k-1} + E_k^{(4)}) + E_k^{(5)}.\end{aligned}$$

  - まず，行列 A の1行当たりの非ゼロ要素数の最大値を m とすると，

$$|E_k^{(3)}| \leq \tilde{\gamma}_m(\lambda)|A|\,|\hat{Q}_{k-1}|.$$

$$\therefore \|E_k^{(3)}\|_F \leq \tilde{\gamma}_m(\lambda)\||A|\|_F\|\hat{Q}_{k-1}\|_F = \tilde{\gamma}_m(\lambda)\sqrt{s}\|A\|_F.$$ （確率≧ $Q(\lambda, nms)$）

  - 次に，$\mathbf{u}^2$の項を無視すると，次の評価が成り立つ.

$$|E_k^{(4)}| \leq \tilde{\gamma}_s(\lambda)|A\hat{Q}_{k-1} + E_k^{(3)}|\,|\hat{\alpha}_{k-1}| \simeq \tilde{\gamma}_s(\lambda)|A\hat{Q}_{k-1}|\,|\hat{\alpha}_{k-1}|.$$

$$\begin{aligned}\therefore \|E_k^{(4)}\|_F &\leq \tilde{\gamma}_s(\lambda)\|A\hat{Q}_{k-1}\|_F\|\hat{\alpha}_{k-1}\|_F \\ &\leq \tilde{\gamma}_s(\lambda)\|A\|_F\|\hat{Q}_{k-1}\|_2\|\hat{\alpha}_{k-1}\|_F \\ &\leq \tilde{\gamma}_s(\lambda)\|A\|_F(\|\hat{X}_k\|_F + \|\hat{X}_{k-1}\|_F).\end{aligned}$$ （確率≧ $Q(\lambda, ns^2)$）

残差ギャップの上界の導出(4)

- $\hat{R}_k$ の確率的誤差評価(続き)
  - また, $|E_k^{(5)}| \leq \mathbf{u}(\hat{R}_{k-1} + fl(fl(A\hat{Q}_{k-1})\hat{\alpha}_{k-1})) \simeq \mathbf{u}|\hat{R}_k|$.
  - 以上より,

$$\begin{aligned}
\|E_k^{(3)}\alpha_{k-1} + E_k^{(4)} + E_k^{(5)}\|_F &\leq \|E_k^{(3)}\|_F\|\hat{\alpha}_{k-1}\|_F + \|E_k^{(4)}\|_F + \|E_k^{(5)}\|_F \\
&\leq \tilde{\gamma}_m(\lambda)\sqrt{s}\|A\|_F(\|\hat{X}_k\|_F + \|\hat{X}_{k-1}\|_F) \\
&\quad +\tilde{\gamma}_s(\lambda)\|A\|_F(\|\hat{X}_k\|_F + \|\hat{X}_{k-1}\|_F) + \mathbf{u}\|\hat{R}_k\|_F \\
&\simeq \lambda\sqrt{s}(\sqrt{m}+1)\mathbf{u}\|A\|_F(\|\hat{X}_k\|_F + \|\hat{X}_{k-1}\|_F) + \mathbf{u}\|\hat{R}_k\|_F.
\end{aligned}$$

(確率$\geqq Q(\lambda, n(m+s)s)$)

- $\hat{X}_k$ と $\hat{R}_k$ の表式

$$\hat{X}_k = \hat{X}_0 + \sum_{\ell=1}^{k}(\hat{Q}_{\ell-1}\hat{\alpha}_{\ell-1} + E_\ell^{(1)} + E_\ell^{(2)}),$$

$$\hat{R}_k = \hat{R}_0 + \sum_{\ell=1}^{k}(A\hat{Q}_{\ell-1}\hat{\alpha}_{\ell-1} + E_\ell^{(3)}\hat{\alpha}_{\ell-1} + E_\ell^{(4)} + E_\ell^{(5)}).$$

残差ギャップの上界の導出(5)

- 残差ギャップの確率的上界

$$
\begin{aligned}
\|(B - A\hat{X}_k) - \hat{R}_k\|_F &= \left\| -A\sum_{\ell=1}^{k}(E_\ell^{(1)} + E_\ell^{(2)}) - \sum_{\ell=1}^{k}(E_\ell^{(3)}\hat{\alpha}_{\ell-1} + E_\ell^{(4)} + E_\ell^{(5)}) \right\|_F \\
&\leq \sum_{\ell=1}^{k}\left(\|A\|_F \|E_\ell^{(1)} + E_\ell^{(2)}\|_F + \|E_\ell^{(3)}\hat{\alpha}_{\ell-1} + E_\ell^{(4)} + E_\ell^{(5)}\|_F \right) \\
&\leq \left\{ 2\lambda(\sqrt{ms} + s + \sqrt{s}) + 1 \right\} \mathbf{u}\|A\|_F \sum_{\ell=0}^{k}\|X_\ell\|_F + \mathbf{u}\sum_{\ell=1}^{k}\|\hat{R}\|_F .
\end{aligned}
$$

（確率≧ $Q(\lambda, n(m+2s)sk)$）

- 定理7
  - ブロッククリロフ部分空間法において，解 $\hat{X}_k$ と残差 $\hat{R}_k$ を列直交なブロック方向ベクトル $\hat{Q}_k$ を用いて更新するとき，残差ギャップ $(B - A\hat{X}_k)$
    - $\hat{R}_k$ は $Q(\lambda, n(m+2s)sk)$ 以上の確率で上記の上界を満たす．

⇨ 真の残差の推定や偽収束の検出への応用

おわりに

- 打切り型テイラー展開に基づく行列指数関数の計算法を例題として，Higham らによって提案された確率的誤差解析の実際を解説した.
- 確率的誤差解析では，$N^{1/2}$（N: 行列の次元数）に比例する上界が得られた．これは，実際の誤差の振る舞いと一致している.
- テイラー展開の高次の項を半精度で計算する混合精度計算，及びブロッククリロフ部分空間法における残差ギャップについても，誤差上界を導出した.