日本応用数理学会論文誌
Vol. 26, No. 2, 2016, pp. 182〜212

浮動小数点演算による内積の丸め誤差解析

樋口 裕幸 * 尾崎 克久 †
* 芝浦工業大学大学院理工学研究科システム理工学専攻
† 芝浦工業大学システム理工学部数理科学科

概要. 数値計算に幅広く採用されている浮動小数点演算は高速である一方で誤差の問題を抱えている．そのために，計算結果と真値の距離の上限，すなわち絶対誤差の上限を求める丸め誤差解析の研究が進んでいる．本稿では，内積に対する丸め誤差解析の最近の結果をその特徴を含めて紹介しつつ，著者らにより改善できた誤差評価式とその特徴を紹介する．

Notes on Error Estimates for Dot Product by Floating-Point Arithmetic

Hiroyuki Higuchi* Katsuhisa Ozaki†
*Division of Systems Engineering and Science,
Graduate School of Engineering and Science,
Shibaura Institute of Technology
†Department of Mathematical Science,
College of Systems Engineering and Science,
Shibaura Institute of Technology

Abstract. This paper is concerned with a rounding error analysis of dot product by floating-point arithmetic. Floating-point arithmetic is fast performed on recent architectures. However, since the number of total bits in floating-point format is finite, rounding errors may occur in arithmetic operations. Recently, new analysis of rounding errors for dot product was developed. We review the results of the error analysis of dot product and revise them.

1. はじめに

本稿では，浮動小数点演算による内積の丸め誤差解析について最近の成果を紹介するとともに，先行研究の改善を行う．浮動小数点数とその演算は IEEE 754 規格 [3] によって定められた 2 進数フォーマットを使用する．オーバーフロー・アンダーフローという用語の定義は IEEE 754 規格 [3] に従うものとする．$\mathbb{F}$ を固定された精度における浮動小数点数の正規化数・非正規化数・零の和集合とする．fl(·) や float(·) という表記は文献 [4] と同じ使い方をする．すなわち，fl(·) は括弧内におけるすべての 2 項演算を最近点への丸め

モード（偶数丸め：roundTiesToEven を採用）による浮動小数点演算で評価した結果を意味する．2 数の積に対して「·」記号を基本的に用いない．ただし，浮動小数点演算中に誤差が発生する可能性がある場合や，「·」記号を入れたほうが読みやすい部分には使用する．$a, b, c \in \mathbb{F}$ に対して，$\mathrm{fl}(\mathrm{fl}(a+b)+c)$ は $\mathrm{fl}((a+b)+c)$ のように表記を簡略化する．$\mathrm{float}(\cdot)$ は括弧内を浮動小数点演算で計算した結果を意味し，括弧内の計算順序は任意とする．例えば $p \in \mathbb{F}^4$ に対して $\mathrm{float}(\sum_{i=1}^{4} p_i)$ は

$$\mathrm{fl}(((p_1+p_2)+p_3)+p_4), \quad \mathrm{fl}((p_1+p_2)+(p_3+p_4)), \quad \mathrm{fl}((p_1+p_3)+(p_2+p_4))$$

など，どれでも良い．本稿全体を通して，$\mathrm{fl}(\cdot)$，$\mathrm{float}(\cdot)$ の評価中にオーバーフローは発生しないと仮定する．$\mathbf{u}$ を相対丸めとする．IEEE 754 規格により $\mathbf{u} \in \{2^{-24}, 2^{-53}, 2^{-113}\}$ となり，倍精度浮動小数点数（binary64）に対しては $\mathbf{u} = 2^{-53}$ である．$a \in \mathbb{R}$ に対する unit in the first place（先頭桁の単位）を表す実関数 $\mathtt{ufp}(a)$ を下記のように導入する．

$$\mathtt{ufp}(a) := \begin{cases} 0, \text{ if } a = 0, \\ 2^{\lfloor \log_2 |a| \rfloor}, \text{ otherwise} \end{cases} \tag{1.1}$$

$a \neq 0$ の場合，$\mathtt{ufp}(a)$ は $|a|$ 以下の最大の 2 のべき乗数を表す．$\mathtt{ufp}$ に関して，

$$\mathtt{ufp}(a) \leq |a| < 2\mathtt{ufp}(a) \tag{1.2}$$

が成立する．この不等式は $\mathtt{ufp}(a) = |a|$ のときには自明であり，そうでない場合は

$$\mathtt{ufp}(a) = 2^{\lfloor \log_2 |a| \rfloor} < 2^{\log_2 |a|} = |a| < 2^{\lfloor \log_2 |a| \rfloor + 1} = 2\mathtt{ufp}(a)$$

から示される．

これから，成分がすべて浮動小数点数であるベクトルの総和を例に挙げながら，本論文で着目する誤差評価式の特徴を整理したい．浮動小数点演算によるベクトル $p \in \mathbb{F}^n$ の総和については，文献 [2] の第 3 章や第 4 章でよく紹介されており，

$$|\mathrm{float}(\sum_{i=1}^{n} p_i) - \sum_{i=1}^{n} p_i| \leq ((1+\mathbf{u})^{n-1} - 1) \sum_{i=1}^{n} |p_i| \tag{1.3}$$

や，$(n-1)\mathbf{u} < 1$ のとき

$$|\mathrm{float}(\sum_{i=1}^{n} p_i) - \sum_{i=1}^{n} p_i| \leq \gamma_{n-1} \sum_{i=1}^{n} |p_i|, \quad \gamma_n = \frac{n\mathbf{u}}{1-n\mathbf{u}} \tag{1.4}$$

が成立する．ここで，丸め誤差解析に関して得られる誤差評価式に対して着目する点を

ポイント (a) 任意の入力サイズ $n \in \mathbb{N}$ に対して成立するか？
ポイント (b) 誤差上限に関する係数に対する計算回数・計算コストは？
ポイント (c) 浮動小数点演算でそのまま誤差の上限が評価可能か？

ポイント (d) 誤差の上限は最適か？または過大評価を抑えているか？
ポイント (e) アンダーフローに対応できているか？

とする．ポイント (b) の誤差に関する係数とは，総和の場合は $\sum_{i=1}^{n}|p_i|$ に関する項に乗ずる $\mathbf{u}$ と n を含む値を意味する．すなわち，誤差評価式 (1.3) では $(1+\mathbf{u})^{n-1}-1$ が，また誤差評価式 (1.4) では γ_{n-1} が係数である．例えば誤差評価式 (1.3) はポイント (a) について達成されており，誤差評価式 (1.4) は n の大きさに制限がある．ポイント (b) を考えれば，$n-1$ 乗の計算を含む誤差評価式 (1.3) よりは誤差評価式 (1.4) のほうが大きな n に対して計算回数が少ない．ポイント (c) については，誤差の上限が浮動小数点演算の結果ではないため，誤差評価式 (1.3) と (1.4) のどちらも満たされていない．ポイント (d) については，どちらも係数は最適ではないものの，係数の大きさは誤差評価式 (1.3) のほうが誤差評価式 (1.4) よりも良い．浮動小数点数の総和については，非正規化数の範囲内における和の演算は正確に行われるため，ここではポイント (e) は議論しなくて良い．これらの特徴をすべて満たす誤差評価式を導出することは非常に難しく，目的によって優先する特徴を設定することになる．

近年，ベクトルの総和の丸め誤差解析に対して進展があった．Rump は浮動小数点数の総和について，

$$|\mathrm{float}(\sum_{i=1}^{n}p_i)-\sum_{i=1}^{n}p_i|\le(n-1)\mathbf{u}\cdot\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n}|p_i|)) \tag{1.5}$$

を示した [6,7][*1]．ここでは，誤差評価式 (1.5) の左辺と右辺にある float(·) 内の計算順序はともに同じとする．ベクトルを $p=(1,\mathbf{u},\mathbf{u},\ldots,\mathbf{u})^T\in\mathbb{F}^n$ とし，すべての n において前方から順に足しこめば $\mathrm{fl}(((p_1+p_2)+p_3+)+\ldots+p_n)=1$ となる．また誤差が $(n-1)\mathbf{u}$ となることから，零ベクトル以外の入力に対しても不等式における等号の例がすべての n に対して見つかる．よって，係数 $(n-1)\mathbf{u}$ はこれ以上小さくできないため，ポイント (d) において最適な評価と言える．また，$\alpha\in\mathbb{F}$ に対する $\mathtt{ufp}(\alpha)$ は 3 回の浮動小数点演算で計算可能であることが文献 [6] に示されているため，ufp の評価は簡単である．以上より，誤差評価式 (1.5) はポイント (a), (b), (d) が達成された優れた誤差評価式である．また，誤差評価式 (1.5) の右辺を浮動小数点演算ですべて評価可能にした

$$n\mathbf{u}\le 1\Longrightarrow|\mathrm{fl}(\sum_{i=1}^{n}p_i)-\sum_{i=1}^{n}p_i|\le\mathrm{fl}((n-1)\cdot(\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(\sum_{i=1}^{n}|p_i|)))) \tag{1.6}$$

も [6] にて示されており，誤差評価式 (1.6) に対してはポイント (b), (c), (d) が達成されている．ここで誤差評価式 (1.6) における $\mathrm{fl}(\sum_{i=1}^{n}p_i)$ は，前方から順に浮動小数点演算で足し

[*1] 計算順序を固定した誤差評価式が [6] にてまず示され，その後任意の計算順序について成立することが [7] にて示された．

こんだ総和の結果を表す．また，誤差評価式 (1.4) の改良として，

$$|\mathrm{float}(\sum_{i=1}^{n} p_i) - \sum_{i=1}^{n} p_i| \leq (n-1)\mathbf{u} \sum_{i=1}^{n} |p_i| \tag{1.7}$$

が，Jeannerod と Rump によって示された [4]．誤差評価式 (1.7) は n に制限がなく，かつ誤差上限の中に $\mathbf{u}^2$ 以降に相当する項がないことからポイント (a)，(b) を満たすきれいな評価式であり，誤差評価式 (1.3) と (1.4) に対してもポイント (d) の点で優れている．以上が総和に関する丸め誤差解析の紹介であった．

本稿の構成は以下の通りである．1 章において総和の丸め誤差解析について紹介し，本論文で着目する誤差評価式に関する特徴を紹介した．2 章では提案手法に用いられる丸め誤差解析のモデルと 2 分木を用いた帰納法の概要について紹介する．3 章では内積の丸め誤差解析に関する先行研究を紹介し，著者らによって改良できた誤差評価式について紹介を行う．また近年の CPU では標準的に使用できることも多い Fused Multiply-Add を用いた誤差評価式について紹介する．

2. 誤差解析のモデル

本章では，誤差評価式の導出に必要な浮動小数点演算に対する丸め誤差解析モデルを紹介する．

定理 2.1 (Jeannerod and Rump [5], 2014) $a, b \in \mathbb{F}$ に対して，

$$\begin{aligned} \mathrm{fl}(a \pm b) &= (a \pm b)(1+\delta_1), \quad |\delta_1| \leq \frac{\mathbf{u}}{1+\mathbf{u}} < \mathbf{u} \\ \mathrm{fl}(a \pm b) &= (a \pm b) + \delta_2, \quad |\delta_2| \leq \mathbf{u} \cdot \mathtt{ufp}(a \pm b) \end{aligned}$$

が成り立つ．さらに，一般に

$$\mathtt{ufp}(a \pm b) \leq \mathtt{ufp}(\mathrm{fl}(a \pm b)) \tag{2.1}$$

となるため，

$$|\delta_2| \leq \mathbf{u} \cdot \mathtt{ufp}(a \pm b) \leq \mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(a \pm b)) \tag{2.2}$$

が成立する．

定理 2.1 における $|\delta_1|$ の上限は，[2] を含む多くの文献では $|\delta_1| \leq \mathbf{u}$ で抑えられていた．その後，$|\delta_1| \leq \dfrac{\mathbf{u}}{1+\mathbf{u}}$ で抑えられ，これが最適であることが Jeannerod と Rump らによって示された [5].

次に，浮動小数点数の積に対する丸め誤差解析モデル [6] を紹介する．以後，$\mathbf{u}_S$ は非正規化数を含めた浮動小数点数における正の最小数とし，倍精度浮動小数点数であれば

$\mathbf{u}_S = 2^{-1074}$ である．また，$\mathbf{u}_N$ は，正の正規化数の最小値とする．倍精度浮動小数点数に対しては $\mathbf{u}_N = 2^{-1022}$ である．ここで，$\mathbf{u}, \mathbf{u}_N, \mathbf{u}_S$ の関係は

$$2\mathbf{u}_N\mathbf{u} = \mathbf{u}_S \tag{2.3}$$

である．

定理 2.2 (Rump [6], 2012) $a, b \in \mathbb{F}$ に対して，

$$\mathrm{fl}(a \cdot b) = ab + \delta + \eta, \quad |\delta| \le \mathbf{u} \cdot \mathtt{ufp}(ab), \quad |\eta| \le \frac{\mathbf{u}_S}{2}, \quad \delta\eta = 0$$

が成り立つ．

定理 2.2 においてアンダーフローが発生するときは $\delta = 0$ と，そうでない場合は $\eta = 0$ としてよい．

$a, b, c \in \mathbb{F}$ に対して，$a \cdot b + c$ に対する浮動小数点演算では，積と和の計算であわせて 2 回の丸め誤差が発生する可能性がある．実数 $ab + c$ を最近点の浮動小数点数に丸めた結果を返す FMA (Fused Multiply-Add) と言われる命令がある．$\mathrm{fl}_{\mathrm{F}}(a, b, c)$ を $a \cdot b + c$ を FMA により計算した結果とすると，次の定理 2.3 が成立する．

定理 2.3 $a, b, c \in \mathbb{F}$ に対して，

$$\mathrm{fl}_{\mathrm{F}}(a, b, c) = ab + c + \delta + \eta, \quad |\delta| \le \mathbf{u} \cdot \mathtt{ufp}(ab + c), \quad |\eta| \le \frac{\mathbf{u}_S}{2}, \quad \delta\eta = 0$$

が成り立つ．

また，ある実数 $a \in \mathbb{R}$ が浮動小数点数であるかを判定する定理を紹介する．

定理 2.4 $a \in \mathbb{R}$ がある．ただし，$|a|$ は $\mathbb{F}$ に属する浮動小数点数の最大値以下であり，$a \in \mathbf{u}_S\mathbb{Z}$ とする．このとき，$c \ge |a| \in \mathbf{u}c\mathbb{Z}$ となる 2 のべき乗数 $c = 2^k$, $k \in \mathbb{Z}$ が存在すれば $a \in \mathbb{F}$ である．

証明 仮定より，実数 a は絶対値の大きさの意味では浮動小数点数で表現するために適切な範囲にあり，問題は仮数部に情報が収まるか否かである．「$c = |a|$」と「$c > |a|$」の場合に分けて考える．$c = |a|$ の場合は a は 2 のべき乗数であるために $a \in \mathbb{F}$ となる．$c > |a|$ の場合，c が 2 のべき乗数であるから $\frac{1}{2}c \ge \mathtt{ufp}(a)$ となる．よって最悪の場合でも $\frac{1}{2}c$ を先頭ビットの単位とし，$\mathbf{u}c$ を最終ビットの単位とする仮数部の範囲に情報が収まるために $a \in \mathbb{F}$ となる．

□

本論文中に使用される等式を紹介する．

補題 2.5 $n \in \mathbb{N}$ に対して，$n \leq \mathbf{u}^{-1}$ ならば $\mathrm{fl}(n\mathbf{u}) = n\mathbf{u}$ である．

この補題は定理 2.4 において $c = 1$ とすることにより直ちに証明される．次章より n の範囲に制約がある場合において，$\mathrm{fl}((n+2)\mathbf{u}) = (n+2)\mathbf{u}$ という関係が用いられる．

次に，内積の丸め誤差解析に必要な 2 分木を用いた帰納法の概要を説明する．$x, y \in \mathbb{F}^n$ に対する浮動小数点演算による内積計算は，$\tilde{p}_i = \mathrm{fl}(x_i \cdot y_i)$ と計算した後に $\sum_{i=1}^{n} \tilde{p}_i$ を浮動小数点演算で評価する．ここで，この総和に対する計算順序は 2 分木によって整理できる．例として $x, y \in \mathbb{F}^8$ における

$$\mathrm{fl}(((((((\tilde{p}_1 + \tilde{p}_2) + \tilde{p}_3) + \tilde{p}_4) + \tilde{p}_5) + \tilde{p}_6) + \tilde{p}_7) + \tilde{p}_8)$$

に対する 2 分木を Fig. 1 に示した．また，

$$\mathrm{fl}(((\tilde{p}_1 + \tilde{p}_2) + (\tilde{p}_3 + \tilde{p}_4)) + ((\tilde{p}_5 + \tilde{p}_6) + (\tilde{p}_7 + \tilde{p}_8)))$$

に対応する 2 分木を Fig. 2 に示した．

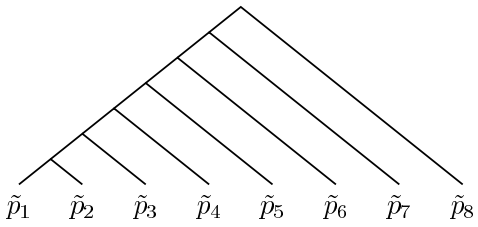

Fig. 1. Binary tree for the recursive order of the summation

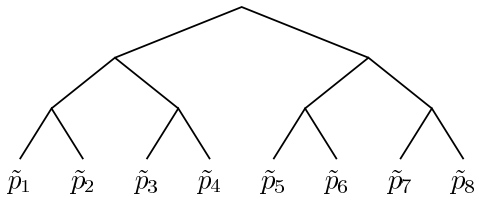

Fig. 2. Binary tree for the pairwise order of the summation

ここで，2 分木を用いた誤差評価式に関する帰納法の概要について解説する．まず 2 分木のすべての葉について誤差評価式が成立することを確認する．あるノードについて葉から左の子までに計算した $\tilde{p}_i$ のインデックスの集合を I_1 とし，葉から右の子までに計算した $\tilde{p}_i$ のインデックスの集合を I_2 とする（$I_1 \cap I_2 = \emptyset$）．Fig. 3 における丸印のノードに対しては

$$I_1 = \{1, 2, 3, 4\}, \quad I_2 = \{5, 6\}$$

となる．集合内の要素の個数を $|I_1| = n_1$, $|I_2| = n_2$ とする．2 分木の葉から左の子ノードと右の子ノードまでの計算について誤差評価式が成立すると仮定する．すなわち誤差評価式が I_1 と I_2 に対して成立すると仮定する．最後に $I_1 \cup I_2$ に対して誤差評価式が成立することを確かめれば，帰納的に証明が進み，最後は 2 分木のルートに関しての誤差評価式が成立する．詳細は文献 [4] を参照するとよい．

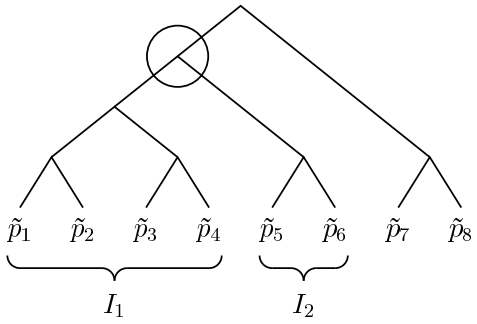

Fig. 3. I_1 and I_2 of the node in binary tree

3. 内積に関する誤差解析

本章のはじめに $x, y \in \mathbb{F}^n$ に対する内積 $x^T y$ の誤差評価式を紹介する．float($x^T y$) はすべての fl($x_i \cdot y_i$) を計算した後，総和を任意の計算順序にて計算した結果である．よって，float($x^T y$) は内積計算に対する Winograd の方法 [8] など，計算方法を根本的に変えて得られた結果を含まないとする．

これからベクトルの内積計算に関する誤差評価式について先行研究を紹介する．演算中にアンダーフローが発生しないと仮定すれば，

$$|\mathrm{float}(x^T y) - x^T y| \le ((1+\mathbf{u})^n - 1)|x|^T|y| \tag{3.1}$$

という誤差評価式が知られている（文献 [2] の 3 章を参照）．これは第 1 章で紹介したポイント (a) を満たしている．ここで，ベクトル $z \in \mathbb{R}^n$ に対して $|z| = (|z_1|, \cdots, |z_n|)^T$ とする．ベクトルの長さに関して $n\mathbf{u} < 1$ という制約をつければ

$$|\mathrm{float}(x^T y) - x^T y| \le \gamma_n |x|^T|y| \tag{3.2}$$

が知られている（文献 [2] の 3 章を参照）．誤差評価式 (3.2) は，ポイント (b) の意味では誤差評価式 (3.1) より優れている．近年，誤差評価式 (3.1) と (3.2) の改良として

$$|\mathrm{float}(x^T y) - x^T y| \le n\mathbf{u}|x|^T|y| \tag{3.3}$$

が示された [4]．誤差評価式 (3.3) はポイント (a) が達成されており，誤差評価式 (3.1) と (3.2) に対してポイント (b), (d) の点でも優れている．もし，浮動小数点演算の際にアンダーフローを考慮するならば

$$|\mathrm{float}(x^T y) - x^T y| \le n\mathbf{u}|x|^T|y| + \frac{n}{2}\mathbf{u}_S \tag{3.4}$$

が示される．

また，式 (1.1) にて定義された ufp を用いた内積の誤差評価についても研究が行われた．$(n+2)\mathbf{u} \le 1$ のとき

$$|\mathrm{fl}(x^T y) - x^T y| \le (n+2)\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(|x|^T|y|)) + \frac{n}{2}\mathbf{u}_S \tag{3.5}$$

が導出されている [6]．ここで，$\mathrm{fl}(x^T y)$ は $\mathrm{fl}(x_i \cdot y_i)$, $1 \le i \le n$ を計算した後，前から順に足しこんで得られた計算結果，すなわち

$$\mathrm{fl}((((x_1 \cdot y_1 + x_2 \cdot y_2) + x_3 \cdot y_3) + x_4 \cdot y_4)\ldots)$$

とする．ポイント (c) を満たす例として，$2(n+2)\mathbf{u} \le 1$ のとき，

$$|\mathrm{fl}(x^T y) - x^T y| \le \mathrm{fl}(((n+2)\mathbf{u}) \cdot \mathtt{ufp}(\mathrm{fl}(|x|^T|y|)) + \mathbf{u}_N) \tag{3.6}$$

と

$$|\mathrm{fl}(x^T y) - x^T y| \le \mathrm{fl}((n+2) \cdot (\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(|x|^T|y|))) + \mathbf{u}_N) \tag{3.7}$$

が文献 [6] に示されている．

ここで，先行研究 [6] における内積の誤差評価式について考察する．総和に関しては誤差評価式 (1.5) のように任意の計算順に対応している．ただし，内積については誤差評価式 (3.5) のように対応をしていない．また，誤差評価式 (3.5) に $n = 1$ を代入すれば，

$$|\mathrm{fl}(x_1 y_1) - x_1 y_1| \le 3\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(|x_1||y_1|)) + \frac{1}{2}\mathbf{u}_S$$

となり，定理 2.2 に比べて過大評価なことがすぐにわかる．また，ufp を用いた総和の誤差評価式 (1.5) に対しては n の制約はなく，ufp を用いた内積の誤差評価式 (3.5) については n の制約がある．一般にアンダーフローに対応するには，小さな定数を最後に足すことで対応をしており，他の方法は議論されていない．さらに FMA 命令を用いた場合の内積の丸め誤差解析に関しては先行研究が見当たらない．以上をふまえ，

- 誤差評価式 (3.5) と (3.6) の任意の計算順序化
- $n = 2$ のときの最適な誤差評価式，すなわちポイント (d) の追求
- n が小さいとき，より良い誤差上限の係数を与えられないか？すなわちポイント (d) の改善
- 任意の n に対応した誤差上限の係数の改善，すなわちポイント (a) の達成とポイント (d) の改善
- ポイント (e) に関する新しい対策
- FMA を用いた内積の誤差解析

について研究した成果を順次述べる．

3.1 任意の計算順序に対する ufp を用いたベクトルの内積の誤差評価

ここでは，誤差評価式 (3.5) と (3.6) に相当する誤差評価を任意の計算順序に拡張する．

定理 3.1 $x, y \in \mathbb{F}^n$ とする．$\mathrm{fl}(|x_i \cdot y_i|) \geq \mathbf{u}_N$ のとき $d_i = 1$ とし，そうでない場合には $d_i = 0$ とする．また $d = \sum_{i=1}^{n} d_i$ とする $(0 \leq d \leq n)$．このとき，次の不等式が成立する．

$$|\mathrm{float}(x^T y) - x^T y| \leq (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u}A + \frac{n-d}{2}\mathbf{u}_S, \quad A := \mathtt{ufp}(\mathrm{float}(|x|^T|y|)) \tag{3.8}$$

ここでは，不等式左辺と右辺の $\mathrm{float}(x^T y)$ の計算順は同じとする．さらに $n\mathbf{u} \leq 1$ ならば

$$|\mathrm{float}(x^T y) - x^T y| \leq (n + 2)\mathbf{u}A + \mathbf{u}_N \tag{3.9}$$

が成立する．さらに，$2(n + 1)\mathbf{u} \leq 1$ ならば

$$|\mathrm{float}(x^T y) - x^T y| \leq \mathrm{fl}(((n + 2)\mathbf{u}) \cdot A + \mathbf{u}_N) \tag{3.10}$$

と浮動小数点演算のみで評価できる誤差上限を得る．

証明 $p_i = x_i y_i$, $\tilde{p}_i = \mathrm{fl}(x_i \cdot y_i)$ とする．内積の誤差評価について，

$$\begin{aligned} |\mathrm{float}(x^T y) - x^T y| &= |\mathrm{float}(\sum_{i=1}^{n} \tilde{p}_i) - \sum_{i=1}^{n} \tilde{p}_i + \sum_{i=1}^{n} \tilde{p}_i - \sum_{i=1}^{n} p_i| \\ &\leq |\mathrm{float}(\sum_{i=1}^{n} \tilde{p}_i) - \sum_{i=1}^{n} \tilde{p}_i| + \sum_{i=1}^{n} |\tilde{p}_i - p_i| \end{aligned} \tag{3.11}$$

と，積に関する誤差と総和の誤差に分ける．さらに，総和の誤差ついては誤差評価式 (1.5) を適用すると

$$|\mathrm{float}(x^T y) - x^T y| \leq (n - 1)\mathbf{u}A + \sum_{i=1}^{n} |\tilde{p}_i - p_i| \tag{3.12}$$

を得る．ここで，$\tilde{p}_i$ に関する誤差の和 $\sum_{i=1}^{n} |\tilde{p}_i - p_i|$ の上限について

$$\begin{aligned} \sum_{i=1}^{n} |\tilde{p}_i - p_i| &\leq \mathbf{u} \sum_{i=1}^{n} |\tilde{p}_i| + \frac{n-d}{2}\mathbf{u}_S \leq \mathbf{u}(\mathrm{float}(\sum_{i=1}^{n} |\tilde{p}_i|) + (n - 1)\mathbf{u}A) + \frac{n-d}{2}\mathbf{u}_S \\ &\leq \mathbf{u}(2A + (n - 1)\mathbf{u}A) + \frac{n-d}{2}\mathbf{u}_S = (2 + n\mathbf{u} - \mathbf{u})\mathbf{u}A + \frac{n-d}{2}\mathbf{u}_S \end{aligned} \tag{3.13}$$

が導かれる．この不等式は，はじめに定理 2.2 をすべての i に使用し，次に $\sum_{i=1}^{n}|\tilde{p}_i|$ の上限について誤差評価式 (1.5) を $\sum_{i=1}^{n}|\tilde{p}_i|$ について解いた式を用いる．さらに不等式 (1.2) の関係を用いて float($\sum_{i=1}^{n}|\tilde{p}_i|$) の上限をとることにより得られる．

式 (3.13) における上限を式 (3.12) の $\sum_{i=1}^{n}|\tilde{p}_i - p_i|$ に代入し，整理すると誤差評価式 (3.8) を得る．さらに，$n\mathbf{u} \le 1$ ならば $n\mathbf{u} - \mathbf{u} < 1$ であり，さらに $n\mathbf{u} \le 1$ と等式 (2.3) の関係より $\frac{n-d}{2}\mathbf{u}_S \le \mathbf{u}_N$ となるため，誤差評価式 (3.9) が得られる．

ここから誤差評価式 (3.10) の証明を行う．$2(n+1)\mathbf{u} \le 1$ を仮定する．このとき，補題 2.5 より $(n+2)\mathbf{u} = \mathrm{fl}((n+2)\mathbf{u})$ となる．これより「$(n+2)\mathbf{u}A < \mathbf{u}_N$」,「$\mathbf{u}_N \le (n+2)\mathbf{u}A < \frac{1}{2}\mathbf{u}^{-1}\mathbf{u}_N$」,「$\frac{1}{2}\mathbf{u}^{-1}\mathbf{u}_N \le (n+2)\mathbf{u}A$」の場合に分けて証明を行う．

1. まず $(n+2)\mathbf{u}A < \mathbf{u}_N$ の場合を考える．このとき，定理 2.2 より

$$(n+2)\mathbf{u}A = \mathrm{fl}((n+2)\mathbf{u})A = \mathrm{fl}(((n+2)\mathbf{u})\cdot A) + \eta, \quad |\eta| \le \frac{1}{2}\mathbf{u}_S \tag{3.14}$$

が成立する．また，仮定の $(n+2)\mathbf{u}A < \mathbf{u}_N$ より

$$\mathrm{fl}(((n+2)\mathbf{u})\cdot A) \le \mathbf{u}_N \tag{3.15}$$

が成立する．n の制限から $n\mathbf{u} - \mathbf{u} < 1$ であり，さらに等式 (2.3) の関係から $\frac{n+1}{2}\mathbf{u}_S < \frac{3}{4}\mathbf{u}_N$ が得られる．不等式 (3.15) より，$\mathrm{fl}(((n+2)\mathbf{u})\cdot A)$ は最大でも $\mathbf{u}_N$ までとなる．最大で $\mathbf{u}_N$ までの数と非正規化数の和は，定理 2.4 において $\mathbf{u}c = \mathbf{u}_S$ と置くことにより無誤差であることが示される．よって

$$\mathrm{fl}(((n+2)\mathbf{u})\cdot A) + \frac{3}{4}\mathbf{u}_N = \mathrm{fl}(((n+2)\mathbf{u})\cdot A + \frac{3}{4}\mathbf{u}_N) \tag{3.16}$$

となる．誤差評価式 (3.8) の上限は，

$$(n+1+n\mathbf{u}-\mathbf{u})\mathbf{u}A + \frac{n-d}{2}\mathbf{u}_S \tag{3.17}$$

$$\le (n+2)\mathbf{u}A + \frac{n}{2}\mathbf{u}_S = \mathrm{fl}((n+2)\mathbf{u})A + \frac{n}{2}\mathbf{u}_S \tag{3.18}$$

$$= \mathrm{fl}(((n+2)\mathbf{u})\cdot A) + \eta + \frac{n}{2}\mathbf{u}_S \le \mathrm{fl}(((n+2)\mathbf{u})\cdot A) + \frac{n+1}{2}\mathbf{u}_S \tag{3.19}$$

$$< \mathrm{fl}(((n+2)\mathbf{u})\cdot A) + \frac{3}{4}\mathbf{u}_N \tag{3.20}$$

$$= \mathrm{fl}(((n+2)\mathbf{u})\cdot A + \frac{3}{4}\mathbf{u}_N) < \mathrm{fl}(((n+2)\mathbf{u})\cdot A + \mathbf{u}_N) \tag{3.21}$$

となる．上記の変形は，(3.17) から (3.18) の変形は $n\mathbf{u}-\mathbf{u}<1$ を，(3.18) から (3.19) の変形は等式 (3.14) を，(3.19) から (3.20) の変形は $\frac{n+1}{2}\mathbf{u}_S < \frac{3}{4}\mathbf{u}_N$ を用いている．最後に，(3.20) から (3.21) の等式変形には等式 (3.16) を用いた．

2. 次に $\mathbf{u}_N \le (n+2)\mathbf{u}A < \frac{1}{2}\mathbf{u}^{-1}\mathbf{u}_N$ の場合を考える．定理 2.2 と n の制限により得られる $(n+2)\mathbf{u}<1$ より

$$(n+2)\mathbf{u}A = \mathrm{fl}((n+2)\mathbf{u})A = \mathrm{fl}(((n+2)\mathbf{u})\cdot A)+\delta \tag{3.22}$$

$$|\delta| \le \mathbf{u}\cdot \mathtt{ufp}((n+2)\mathbf{u}\cdot A) \le \frac{1}{2}\mathbf{u}A \tag{3.23}$$

を得る．また，n の制限により $\frac{n}{2}\mathbf{u}_S < \frac{1}{2}\mathbf{u}_N$，$n\mathbf{u}-\mathbf{u}-\frac{1}{2}<0$ である．誤差評価式 (3.8) の上限は，

$$\begin{aligned}
&(n+1+n\mathbf{u}-\mathbf{u})\mathbf{u}A+\frac{n-d}{2}\mathbf{u}_S \\
&= (n+2)\mathbf{u}A+(n\mathbf{u}-\mathbf{u}-1)\mathbf{u}A+\frac{n-d}{2}\mathbf{u}_S && (3.24)\\
&\le \mathrm{fl}(((n+2)\mathbf{u})\cdot A)+\delta+(n\mathbf{u}-\mathbf{u}-1)\mathbf{u}A+\frac{n}{2}\mathbf{u}_S && (3.25)\\
&< \mathrm{fl}(((n+2)\mathbf{u})\cdot A)+(n\mathbf{u}-\mathbf{u}-\frac{1}{2})\mathbf{u}A+\frac{1}{2}\mathbf{u}_N && (3.26)\\
&= \mathrm{fl}(((n+2)\mathbf{u})\cdot A)+\mathbf{u}_N+(n\mathbf{u}-\mathbf{u}-\frac{1}{2})\mathbf{u}A-\frac{1}{2}\mathbf{u}_N && (3.27)\\
&\le \mathrm{fl}(((n+2)\mathbf{u})\cdot A)+\mathbf{u}_N-\frac{1}{2}\mathbf{u}_N && (3.28)\\
&\le \mathrm{fl}(((n+2)\mathbf{u})\cdot A+\mathbf{u}_N) && (3.29)
\end{aligned}$$

となる．これより，上式の導出について説明する．(3.24) から (3.25) の変形には不等式 (3.22) を，(3.25) から (3.26) の変形には不等式 (3.23) と $\frac{n}{2}\mathbf{u}_S < \frac{1}{2}\mathbf{u}_N$ を用いた．(3.27) から (3.28) は $n\mathbf{u}-\mathbf{u}-\frac{1}{2}$ が負であることから，この項をなくすことにより上限を得た．最後の (3.28) から (3.29) の変形では，$(n+2)\mathbf{u}A$ に対する仮定から得られる $\mathrm{fl}(((n+2)\mathbf{u})\cdot A) \le \frac{1}{2}\mathbf{u}^{-1}\mathbf{u}_N$ の条件下で成立する

$$\mathrm{fl}(((n+2)\mathbf{u})\cdot A)+\mathbf{u}_N = \mathrm{fl}(((n+2)\mathbf{u})\cdot A+\mathbf{u}_N)+\delta_1, \quad |\delta_1| \le \frac{1}{2}\mathbf{u}_N$$

を使用した．

3. 最後に $(n+2)\mathbf{u}A \ge \frac{1}{2}\mathbf{u}^{-1}\mathbf{u}_N$ の場合を考える．ここで，n に対すると制約から得る n の最大値と，仮定 $(n+2)\mathbf{u}A \ge \frac{1}{2}\mathbf{u}^{-1}\mathbf{u}_N$ から得る $\mathbf{u}A$ の最小値を考えれば，式 (3.26) において

$$(n\mathbf{u}-\mathbf{u}-1)\mathbf{u}A+\frac{1}{2}\mathbf{u}_N \le 0$$

を得るため，誤差評価式 (3.10) を得る．

□

3.2 長さ 2 のベクトルの内積に関する最適な誤差評価

本節では，長さ 2 のベクトルの内積に関する最適な誤差評価について述べる．ここで最適な誤差評価とは誤差上限がこれ以上小さくできないものとし，誤差上限を表す式が唯一であることは保証しない．この長さ 2 のベクトルの内積は 2 次の正方行列に対する行列式の計算や，複素数の積において現れる．

補題 3.2 $0 \le a, b \in \mathbb{F}$ に対して

$$\mathtt{ufp}(a) + \mathtt{ufp}(b) \le 1.5\mathtt{ufp}(a+b) \tag{3.30}$$

が成立する．$a \neq 0$ のとき $a = m_a \cdot 2^{e_a}$，$b \neq 0$ のとき $b = m_b \cdot 2^{e_b}$ と表す．ここでは $1 \le m_a, m_b < 2$, $m_a, m_b \in \mathbb{F}$, $e_a, e_b \in \mathbb{Z}$ である．$e_a = e_b$ または $ab = 0$ のとき，

$$\mathtt{ufp}(a) + \mathtt{ufp}(b) = \mathtt{ufp}(a+b) \tag{3.31}$$

となる．また，$a \neq 0$ かつ $b \neq 0$ の場合は

$$\mathtt{ufp}(a) + \mathtt{ufp}(b) = \frac{1 + 2^{e_b - e_a}}{\mathtt{ufp}(m_a + m_b 2^{e_b - e_a})} \mathtt{ufp}(a+b) \tag{3.32}$$

が成立する．

証明　a と b のどちらか一方が少なくとも 0 の場合は，不等式 (3.30) と不等式 (3.31) の成立は自明であるため，それ以外の場合を証明する．まず

$$\mathtt{ufp}(a) + \mathtt{ufp}(b) = 2^{e_a} + 2^{e_b} = 2^{e_a}(1 + 2^{e_b - e_a}) \tag{3.33}$$

であり，

$$\mathtt{ufp}(a+b) = \mathtt{ufp}(m_a 2^{e_a} + m_b 2^{e_b}) = 2^{e_a}\mathtt{ufp}(m_a + m_b 2^{e_b - e_a}) \tag{3.34}$$

となるため，式 (3.32) を得る．$e_a = e_b$ のとき，$e_b - e_a = 0$ を式 (3.33) と式 (3.34) に代入し，$2 \le m_a + m_b < 4$ から $\mathtt{ufp}(m_a + m_b) = 2$ を式 (3.34) に代入すると式 (3.31) を得る．

最後に式 (3.30) の証明を 2 通りの場合に分けて考える．ここでは $|a| \ge |b|$ を仮定しても一般性を失わないために $e_a \ge e_b$ とし，$e_a = e_b$ のときには式 (3.31) が示されているために $e_a > e_b$ の場合を考える．式 (3.34) から，$2 \le m_a + m_b 2^{e_b - e_a}$ のときは $\mathtt{ufp}(a+b) = 2^{e_a+1}$ となり，$2 > m_a + m_b 2^{e_b - e_a}$ のときは $\mathtt{ufp}(a+b) = 2^{e_a}$ となる．以上より $\mathtt{ufp}(a+b)$ の最小値は 2^{e_a} である．また，$\mathtt{ufp}(a) + \mathtt{ufp}(b)$ の最大値は $e_b - e_a = -1$ のときで $1.5 \cdot 2^{e_a}$ となるため，不等式 (3.30) が成立する．□

上述の補題を用いて，長さが 2 であるベクトルの内積に対する最適な誤差評価を以下のように得た．

定理 3.3 $a, b, c, d \in \mathbb{F}$ に対して，

$$|\mathrm{fl}(a \cdot b + c \cdot d) - (ab + cd)| \leq (2.5 - \mathbf{u})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(|a \cdot b| + |c \cdot d|)) \tag{3.35}$$

が成立する．この不等式において，$ab \neq 0$ かつ $cd \neq 0$ のときに等号が成り立つ例が存在するため，この誤差評価式は最適である．ただし，$\mathrm{fl}(\cdot)$ 内でアンダーフローが発生しないと仮定する．

証明 a, b, c, d のうち 1 つでも 0 があれば，定理 2.2 から誤差評価式 (3.35) が成立する．よって，a, b, c, d はすべて 0 でないと仮定し，

$$\mathrm{fl}(a \cdot b) = \mathtt{sign}(ab) \cdot m_1 \cdot 2^{e_1}, \quad \mathrm{fl}(c \cdot d) = \mathtt{sign}(cd) \cdot m_2 \cdot 2^{e_2},$$
$$e_1, e_2 \in \mathbb{Z}, \quad 1 \leq m_1, m_2 < 2, \quad m_1, m_2 \in \mathbb{F}$$

とする．ここで sign は括弧内の実数の符号を返す関数とする．$ab + cd$ の計算には対称性があるため，$|\mathrm{fl}(a \cdot b)| \geq |\mathrm{fl}(c \cdot d)|$ の場合を証明すれば良い ($e_1 \geq e_2$)．アンダーフローが起きない仮定から

$$\mathrm{fl}(a \cdot b) = ab + \delta_1, \quad \mathrm{fl}(c \cdot d) = cd + \delta_2, \quad \mathrm{fl}(a \cdot b + c \cdot d) = \mathrm{fl}(a \cdot b) + \mathrm{fl}(c \cdot d) + \delta_3$$

とすれば

$$\mathrm{fl}(a \cdot b + c \cdot d) = \mathrm{fl}(a \cdot b) + \mathrm{fl}(c \cdot d) + \delta_3 = ab + \delta_1 + cd + \delta_2 + \delta_3$$

となる．δ_1, δ_2 は定理 2.2 より, また δ_3 は定理 2.1 により

$$|\delta_1| = f_1\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(a \cdot b)), \quad |\delta_2| = f_2\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(c \cdot d)),$$
$$|\delta_3| = f_3\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(a \cdot b) + \mathrm{fl}(c \cdot d)), \quad 0 \leq f_1, f_2, f_3 \leq 1$$

とおける．ここで，

$$\begin{aligned} |\mathrm{fl}(a \cdot b + c \cdot d) - (ab + cd)| &\leq |\delta_1| + |\delta_2| + |\delta_3| \\ &= f_1\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(a \cdot b)) + f_2\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(c \cdot d)) \\ &\quad + f_3\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(a \cdot b) + \mathrm{fl}(c \cdot d)) \\ &\leq f\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}(|a \cdot b| + |c \cdot d|)) \end{aligned} \tag{3.36}$$

となる $f \in \mathbb{R}$ を考える．以下「$e_1 = e_2$」,「$e_1 - e_2 = 1$ かつ $2 \leq m_1 + m_2 2^{e_2 - e_1}$」,「$e_1 - e_2 = 1$ かつ $2 > m_1 + m_2 2^{e_2 - e_1}$」,「$e_1 - e_2 \geq 2$」の場合について考えていく．

1. $e_1 = e_2$ のとき，式 (3.36) において $f_1 = f_2 = f_3 = 1$ とする．さらに補題 3.2 の式 (3.31) より

$$\mathtt{ufp}(\mathrm{fl}(|a \cdot b|)) + \mathtt{ufp}(\mathrm{fl}(|c \cdot d|)) = \mathtt{ufp}(\mathrm{fl}(|a \cdot b|) + \mathrm{fl}(|c \cdot d|))$$

を用いて式 (3.36) を整理すると，

$$|\mathrm{fl}(a\cdot b+c\cdot d)-(ab+cd)|\le 2\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(|a\cdot b|)+\mathrm{fl}(|c\cdot d|))$$

となる．さらに不等式 (2.1) を用いれば

$$|\mathrm{fl}(a\cdot b+c\cdot d)-(ab+cd)|\le 2\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(|a\cdot b|+|c\cdot d|))$$

となるため，$f=2$ を得る．

2. $e_1-e_2=1$ かつ $2\le m_1+m_2 2^{e_2-e_1}$ の場合，式 (3.36) において $f_1=f_2=f_3=1$ とする．さらに補題 3.2 の式 (3.32) より

$$\mathtt{ufp}(\mathrm{fl}(|a\cdot b|))+\mathtt{ufp}(\mathrm{fl}(|c\cdot d|))=0.75\mathtt{ufp}(\mathrm{fl}(|a\cdot b|)+\mathrm{fl}(|c\cdot d|))$$

を用いて式 (3.36) を整理し，最後に不等式 (2.1) を用いることにより

$$\begin{aligned}|\mathrm{fl}(a\cdot b+c\cdot d)-(ab+cd)|&\le 1.75\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(|a\cdot b|)+\mathrm{fl}(|c\cdot d|))\\&\le 1.75\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(|a\cdot b|+|c\cdot d|))\end{aligned}$$

を導け，$f=1.75$ を得る．

3. $e_1-e_2=1$ かつ $2>m_1+m_2 2^{e_2-e_1}$ の場合を考える．ここではさらに「$\mathrm{fl}(c\cdot d)$ の仮数部の最下位ビットが 0 の場合」と「$\mathrm{fl}(c\cdot d)$ の仮数部の最下位ビットが 1 の場合」の 2 通りに場合分けを行う．以下，$a\in\mathbb{F}$ に対して $a\in 2\mathbf{u}\cdot\mathtt{ufp}(a)\mathbb{Z}$ が成り立つことを利用する（浮動小数点数は仮数部の最下位ビットの位の整数倍で表現されるという意味である）．

（a）$\mathrm{fl}(c\cdot d)$ の最下位ビットが 0 の場合，$\mathrm{fl}(c\cdot d)\in 4\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(c\cdot d))\mathbb{Z}=2\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(a\cdot b))\mathbb{Z}$ である．また，$\mathrm{fl}(a\cdot b)+\mathrm{fl}(c\cdot d)\in 2\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(a\cdot b))\mathbb{Z}$ であり，仮定より $|\mathrm{fl}(a\cdot b)+\mathrm{fl}(c\cdot d)|<2\mathtt{ufp}(\mathrm{fl}(a\cdot b))$ であるから，定理 2.4 より

$$\mathrm{fl}(a\cdot b)+\mathrm{fl}(c\cdot d)=\mathrm{fl}(a\cdot b+c\cdot d)$$

となり，誤差が発生しないために $f_3=0$ である．よって式 (3.36) において $f_1=f_2=1$ とし，補題 3.2 の式 (3.32) を用い，

$$\begin{aligned}|\mathrm{fl}(a\cdot b+c\cdot d)-(ab+cd)|&\le \mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(a\cdot b))+\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(c\cdot d))\\&\le 1.5\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(|a\cdot b|)+\mathrm{fl}(|c\cdot d|))\end{aligned}$$

を得て，最後に不等式 (2.1) を用いれば $f=1.5$ となる．

（b）次に $\mathrm{fl}(c\cdot d)$ の最下位ビットが 1 の場合を考えるが，これは偶数丸めが発生しなかったことを表すために $f_2\neq 1$ である．よって，$\mathrm{fl}(c\cdot d)$ の誤差が $\mathbf{u}\cdot\mathtt{ufp}(c\cdot d)$ となることはなく，$\mathbf{u}\cdot\mathtt{ufp}(c\cdot d)$ 未満の最大の誤差を考える．アンダーフローがない場合，浮動小数点数の積の誤差は浮動小数点数とな

る [*1]．また，$\mathrm{fl}(c\cdot d)$ の誤差が大きいときは $2\mathtt{ufp}(c)\mathtt{ufp}(d) = \mathtt{ufp}(cd)$ となる場合である．浮動小数点数の基本より，$c \in 2\mathbf{u}\cdot\mathtt{ufp}(c)\mathbb{Z}$, $d \in 2\mathbf{u}\cdot\mathtt{ufp}(d)\mathbb{Z}$ であるから，$cd \in 4\mathbf{u}^2\cdot\mathtt{ufp}(c)\mathtt{ufp}(d)\mathbb{Z}$ となる．$\mathbf{u}\cdot\mathtt{ufp}(cd)$ より小さい最大の誤差は $\mathbf{u}\cdot\mathtt{ufp}(cd) - 4\mathbf{u}^2\cdot\mathtt{ufp}(c)\mathtt{ufp}(d) = (\mathbf{u}-2\mathbf{u}^2)\cdot\mathtt{ufp}(cd)$ となるため，$|\delta_2| \le (1-2\mathbf{u})\mathbf{u}\cdot\mathtt{ufp}(cd)$ となり，$f_2 \le 1-2\mathbf{u}$ を得る．さらに，仮定である $e_1 - e_2 = 1$ かつ $2 > m_1 + m_2 2^{e_2-e_1}$ から

$$\mathtt{ufp}(\mathrm{fl}(a\cdot b)+\mathrm{fl}(c\cdot d)) = \mathtt{ufp}(\mathrm{fl}(a\cdot b)) = 2\mathtt{ufp}(\mathrm{fl}(c\cdot d)) \tag{3.37}$$

を得る．式 (3.36) において $f_1 = 1$, $f_2 = 1-2\mathbf{u}$, $f_3 = 1$ を代入し，等式 (3.37) から $\mathtt{ufp}(\mathrm{fl}(a\cdot b)) = \mathtt{ufp}(\mathrm{fl}(a\cdot b)+\mathrm{fl}(c\cdot d))$ と $\mathtt{ufp}(\mathrm{fl}(c\cdot d)) = \dfrac{1}{2}\mathtt{ufp}(\mathrm{fl}(a\cdot b)+\mathrm{fl}(c\cdot d))$ を使用し，さらに不等式 (2.1) を用いることにより

$$\begin{aligned}
&|\mathrm{fl}(a\cdot b + c\cdot d) - (ab+cd)| \\
&\le \mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(a\cdot b)) + (1-2\mathbf{u})\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(c\cdot d)) + \mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(a\cdot b)+\mathrm{fl}(c\cdot d)) \\
&= (2.5-\mathbf{u})\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(|a\cdot b|)+\mathrm{fl}(|c\cdot d|)) \le (2.5\mathbf{u}-\mathbf{u}^2)\mathtt{ufp}(\mathrm{fl}(|a\cdot b|+|c\cdot d|))
\end{aligned}$$

を得る．よって $f = 2.5-\mathbf{u}$ となる．

4. $e_1 - e_2 \ge 2$ の場合，式 (3.36) において $f_1 = f_2 = f_3 = 1$ とする．さらに補題 3.2 の式 (3.32) より

$$\mathtt{ufp}(\mathrm{fl}(|a\cdot b|)) + \mathtt{ufp}(\mathrm{fl}(|c\cdot d|)) \le 1.25\mathtt{ufp}(\mathrm{fl}(|a\cdot b|)+\mathrm{fl}(|c\cdot d|))$$

を用いて整理し，最後に不等式 (2.1) を用いることにより $f = 2.25$ を得る．

以上より，$f = 2.5-\mathbf{u}$ が最大のため，誤差評価式 (3.35) は成立する．最後に最適性について例を挙げて説明する．

$$a = 2^2\cdot 1.25, \quad b = 1+12\mathbf{u}, \quad c = 1.5-2\mathbf{u}, \quad d = 1.5-2\mathbf{u}$$

のときに

$$\begin{aligned}
\mathrm{fl}(a\cdot b + c\cdot d) &= \mathrm{fl}((5+64\mathbf{u})+(2.25-4\mathbf{u})) = 4(1.5+2^{-2}+2^{-4}+16\mathbf{u}) \\
ab+cd &= 4(1.5+2^{-2}+2^{-4}+13.5\mathbf{u}+\mathbf{u}^2)
\end{aligned}$$

となり

$$|\mathrm{fl}(a\cdot b + c\cdot d) - (ab+cd)| = 4(2.5\mathbf{u}-\mathbf{u}^2) = (2.5-\mathbf{u})\mathbf{u}\cdot\mathtt{ufp}(\mathrm{fl}(a\cdot b + c\cdot d))$$

を得る．よって誤差評価式 (3.35) の右辺をこれ以上小さくできない例が存在するために最適な誤差評価式と言える． □

[*1] 文献 [1] により，2 つの浮動小数点数の積は 2 つの浮動小数点数の和に変換できることからもわかる．

3.3 n が小さいときの係数の改良

$n \le 1-\log_2 \mathbf{u}$ の範囲（単精度浮動小数点数では $n \le 25$, 倍精度浮動小数点数では $n \le 54$）における誤差評価式 (3.5) の改良を議論する．集合を $\mathbb{F}' = \mathbf{u}_S\mathbb{Z}$ と定める．$0 \le p \in \mathbb{F}'$ は，ある $l,\ h \in \mathbb{Z},\ \log_2 \mathbf{u}_S \le l \le h$ を用いて

$$p = \sum_{i=l}^{h} d_i \cdot 2^i, \quad d_i \in \{0, 1\}$$

と表せる．p に対する d_i の和を

$$\underline{p} = \sum_{i=l}^{h} d_i, \quad d_i \in \{0, 1\}$$

と下線を引いた表記を用いる [*1]．これより誤差評価式を改良するために，いくつかの補題を示す．

補題 3.4 $0 \le a, b \in \mathbb{F}',\ p = a + b$ に対して

$$\underline{p} \le \underline{a} + \underline{b} \tag{3.38}$$

が成立する．

証明 $a = 0$ または $b = 0$ の場合は自明に成立するため，それ以外の場合を考える．a, b を

$$a = \sum_{i=l_a}^{h_a} d_i \cdot 2^i, \quad d_i = \begin{cases} 1 & (i = l_a, h_a) \\ 0 \text{ or } 1 & (l_a + 1 \le i \le h_a - 1) \\ 0 & (\text{otherwise}) \end{cases}$$

$$b = \sum_{i=l_b}^{h_b} e_i \cdot 2^i, \quad e_i = \begin{cases} 1 & (i = l_b, h_b) \\ 0 \text{ or } 1 & (l_b + 1 \le i \le h_b - 1) \\ 0 & (\text{otherwise}) \end{cases}$$

とおくと，

$$p = \sum_{i=\min(l_a,l_b)}^{\max(h_a,h_b)+1} f_i \cdot 2^i, \quad f_i = \begin{cases} 0 \text{ or } 1 & (\min(l_a, l_b) \le i \le \max(h_a, h_b) + 1) \\ 0 & (\text{otherwise}) \end{cases}$$

と表現できる．ここで，$a_j,\ b_j,\ p_j = a_j + b_j,\ \min(l_a, l_b) \le j$ に対して

$$a_j = \sum_{i=\min(l_a,l_b)}^{j} d_i \cdot 2^i, \quad b_j = \sum_{i=\min(l_a,l_b)}^{j} e_i \cdot 2^i, \quad p_j = \sum_{i=\min(l_a,l_b)}^{j+1} f_{i,j} \cdot 2^i, \quad f_{i,j} \in \{0, 1\}$$

[*1] 正規化数においては暗黙の 1 (an implicit hidden bit) も含めた仮数部の 1 の個数と同じである．これは暗黙の 1 を保存しない浮動小数点数のビット表現における仮数部の 1 の個数とは異なる．

を定義する．ここで，

(3.39) $$\underline{p_j} \le \underline{a_j} + \underline{b_j}$$

を j についての帰納法で証明できれば，

$$a_{\max(h_a,h_b)+1} = a, \quad b_{\max(h_a,h_b)+1} = b, \quad p_{\max(h_a,h_b)+1} = p$$

となるため不等式 (3.38) は証明される．

1. $j = \min(l_a, l_b) =: l$ のとき，以下のように場合分けで考える．
   （a）$l_a \neq l_b$ のとき $d_l + e_l = 1$, $(f_{l+1,l}, f_{l,l}) = (0,1)$ となり，$\underline{p_l} = \underline{a_l} + \underline{b_l} = 1$ から不等式 (3.39) は成り立つ．
   （b）$l_a = l_b$ のとき $d_l = 1, e_l = 1$, $(f_{l+1,l}, f_{l,l}) = (1,0)$ となり，$\underline{p_l} = 1 < \underline{a_l} + \underline{b_l} = 2$ から不等式 (3.39) は成り立つ．
2. $j = k$ のとき $\underline{p_k} \le \underline{a_k} + \underline{b_k}$ が成り立つと仮定する．$j = k+1$ のとき

$$\underline{a_{k+1}} + \underline{b_{k+1}} = d_{k+1} + \underline{a_k} + e_{k+1} + \underline{b_k}$$
$$\underline{p_{k+1}} = f_{k+2,k+1} + f_{k+1,k+1} - f_{k+1,k} + \underline{p_k}$$

となる．ここで $f_{k+1,k}$ について場合分けを行う．
（a）$f_{k+1,k} = 0$ のとき

$$(f_{k+2,k+1}, f_{k+1,k+1}) = \begin{cases} (1,0) & (d_{k+1} + e_{k+1} = 2) \\ (0,1) & (d_{k+1} + e_{k+1} = 1) \\ (0,0) & (d_{k+1} + e_{k+1} = 0) \end{cases}$$

となり，$f_{k+2,k+1} + f_{k+1,k+1} - f_{k+1,k} \le d_{k+1} + e_{k+1}$ を得る．よって仮定より $\underline{p_{k+1}} \le \underline{a_{k+1}} + \underline{b_{k+1}}$ となる．
（b）$f_{k+1,k} = 1$ のとき

$$(f_{k+2,k+1}, f_{k+1,k+1}) = \begin{cases} (1,1) & (d_{k+1} + e_{k+1} = 2) \\ (1,0) & (d_{k+1} + e_{k+1} = 1) \\ (0,1) & (d_{k+1} + e_{k+1} = 0) \end{cases}$$

となり，$f_{k+2,k+1} + f_{k+1,k+1} - f_{k+1,k} \le d_{k+1} + e_{k+1}$ を得る．よって仮定より $\underline{p_{k+1}} \le \underline{a_{k+1}} + \underline{b_{k+1}}$ となる．

以上より $\min(l_a, l_b) \le j$ に対して常に成り立つ． □

補題 3.5 $p \in \mathbb{F}'$ に対して

$$\mathrm{fl}(p) \neq p \Leftrightarrow \underline{\mathrm{fl}(p)} < \underline{p}, \quad \mathrm{fl}(p) = p \Leftrightarrow \underline{\mathrm{fl}(p)} = \underline{p}, \quad \underline{\mathrm{fl}(p)} \le \underline{p}$$

が成立する．$p \in \mathbb{F}'$ に対して $\mathrm{fl}(p)$ は最近偶数の丸めにより p を $\mathbb{F}$ の要素に対応させる表記とする．

証明　$p \geq 0$ を仮定する．

1. $\mathrm{fl}(p) = p \Rightarrow \underline{\mathrm{fl}(p)} = \underline{p}$ は明らかに成り立つ．
2. $\mathrm{fl}(p) \neq p \Rightarrow \underline{\mathrm{fl}(p)} < \underline{p}$ を示す．p を $\mathbb{F}$ の要素に原点方向に丸めた結果（切り捨て）を a とし，b は $p = a + b$ を満たすものとする（図 4 参照）．ここでは，$a > 2\mathbf{u}_N$ となり，a は非正規化数や 0 にならないことに留意する．

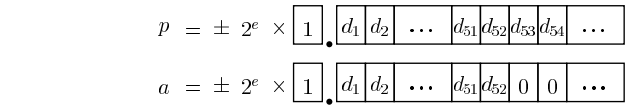

Fig. 4.　The relation between $p \in \mathbb{F}'$ and $a \in \mathbb{F}$ $(d_i \in \{0, 1\},\ \mathbf{u} = 2^{-53})$

このとき $\underline{p} = \underline{a} + \underline{b}$ であり，$\mathbf{u} \cdot \mathtt{ufp}(p) \geq \mathtt{ufp}(b) > 0$ が成立する．また，仮定 $\mathrm{fl}(p) \neq p$ より $\underline{b} \geq 1$ となる．ここで，以下 4 つの場合について考える．

（a）$\mathbf{u} \cdot \mathtt{ufp}(p) > \mathtt{ufp}(b) > 0$ の場合を考える．これは p の仮数部の下位ビットがそのまま切り捨てられて $\mathrm{fl}(p)$ が得られた場合に相当する．$\mathrm{fl}(p) = a$ となるため，$\underline{p} = \underline{a} + \underline{b} > \underline{a} = \underline{\mathrm{fl}(p)}$ となる．

（b）$\mathbf{u} \cdot \mathtt{ufp}(p) = \mathtt{ufp}(b)$, $\underline{b} \geq 2$ のときを考える．これは p を $\mathrm{fl}(p)$ に丸める際に切り上げが起きたことを意味する．$\mathrm{fl}(p) = a + 2\mathbf{u} \cdot \mathtt{ufp}(a)$ となるため補題 3.4 より $\underline{\mathrm{fl}(p)} \leq \underline{a} + 1$ となる．よって $\underline{p} = \underline{a} + \underline{b} > \underline{a} + 1 \geq \underline{\mathrm{fl}(p)}$ となる．

（c）$\mathbf{u} \cdot \mathtt{ufp}(p) = \mathtt{ufp}(b)$, $\underline{b} = 1$, a の仮数部最終ビットが 0 のときを考える．これは $\mathrm{fl}(p)$ が偶数丸め（切り捨て）の結果であることを意味する．$\mathrm{fl}(p) = a$ となるため (a) と同様に $\underline{p} > \underline{\mathrm{fl}(p)}$ となる．

（d）$\mathbf{u} \cdot \mathtt{ufp}(p) = \mathtt{ufp}(b)$, $\underline{b} = 1$, a の仮数部最終ビットが 1 のときを考える．これは $\mathrm{fl}(p)$ が偶数丸め（切り上げ）の結果であることを意味する．$\mathrm{fl}(p) = a + 2\mathbf{u} \cdot \mathtt{ufp}(a)$ となり，a の仮数部最終ビットに 1 を加えるため繰り上がりが必ず発生する．ここで $w = -\log_2 \mathbf{u}$ とし，

$$p = 2^e \sum_{i=1}^{w+1} d_i 2^{1-i}, \quad d_i = \begin{cases} 0 \text{ or } 1 & (2 \leq i \leq w-1) \\ 1 & (i = 1, w, w+1) \end{cases}, \quad e \in \mathbb{Z}$$

とおく．$\underline{\mathrm{fl}(p)}$ は $\underline{p}$ に対してまず $i = w, w+1$ の 2 つの 1 を失い，繰り上げのための 1 を 1 つ追加するため $\underline{\mathrm{fl}(p)} < \underline{a} + 1$ となり，$\underline{p} = \underline{a} + \underline{b} > \underline{a} \geq \underline{\mathrm{fl}(p)}$ となる．

以上より $\mathrm{fl}(p) \neq p \Rightarrow \underline{\mathrm{fl}(p)} < \underline{p}$ が成り立つ．

3. $\mathrm{fl}(p) = p \Leftarrow \underline{\mathrm{fl}(p)} = \underline{p}$ の対偶は 2 から証明される．
4. $\mathrm{fl}(p) \neq p \Leftarrow \underline{\mathrm{fl}(p)} < \underline{p}$ は，1 の対偶から証明される．
5. $\underline{\mathrm{fl}(p)} \leq \underline{p}$ は 1 と 2 より成り立つ．

p に対する $\mathrm{fl}(p)$ と，$-p$ に対する $\mathrm{fl}(-p)$ では同様の証明ができるため，$p < 0$ についても補題は成立する．　□

成分が 2 のべき乗数または 0 であるベクトルの総和について，次の補題が成立する．

補題 3.6 $p \in \mathbb{F}^n$, $\mathtt{ufp}(p_i) = p_i$, $1 \le i \le n$, $\alpha := \mathrm{float}(\sum_{i=1}^{n} p_i)$ に対して，

$$\sum_{i=1}^{n} p_i = \mathrm{float}(\sum_{i=1}^{n} p_i) + \Delta, \quad |\Delta| \le (n - \underline{\alpha})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} p_i))$$

が成立する．ただし，上式にある 2 つの float(·) 内の計算順序は同じとする．

証明 2 章で説明をした 2 分木（ただし葉は p_i）を用いた帰納法で証明する．まず，2 分木の葉に関しては計算が存在しないために自明に成立する．$j = 1, 2$ に対して，$\alpha_j := \mathrm{float}(\sum_{i \in I_j} p_i)$ としたとき

$$\sum_{i \in I_j} p_i = \mathrm{float}(\sum_{i \in I_j} p_i) + \Delta_j = \alpha_j + \Delta_j,$$
$$|\Delta_j| \le (n_j - \underline{\alpha_j})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i \in I_j} p_i)) = (n_j - \underline{\alpha_j})\mathbf{u} \cdot \mathtt{ufp}(\alpha_j)$$

を仮定する．このとき，$I_3 = I_1 \cup I_2$ に関する計算に関して，定理 2.1 より

$$\alpha_3 := \mathrm{fl}(\alpha_1 + \alpha_2) = \alpha_1 + \alpha_2 + \delta, \quad |\delta| \le \mathbf{u} \cdot \mathtt{ufp}(\alpha_1 + \alpha_2) \le \mathbf{u} \cdot \mathtt{ufp}(\alpha_3)$$

を得る．この α_3 に対して

$$\sum_{i \in I_3} p_i = \alpha_3 + \Delta_3, \quad |\Delta_3| \le (n_3 - \underline{\alpha_3})\mathbf{u} \cdot \mathtt{ufp}(\alpha_3), \quad n_3 = n_1 + n_2$$

の成立を示す．δ は定理 2.1 から求まる和の誤差上限であり，$\Delta_3 = \Delta_1 + \Delta_2 - \delta$ の関係が成立している．左の子ノードまでの計算結果と右の子ノードまでの計算結果の和を

$$\alpha_1 + \alpha_2 =: b$$

とすると，補題 3.5 と補題 3.4 より

$$\underline{\alpha_3} \le \underline{b} \le \underline{\alpha_1} + \underline{\alpha_2} \tag{3.40}$$

となる．ここで不等式 (3.40) に関して次の場合を考える．

1. $\underline{\alpha_3} = \underline{b} \le \underline{\alpha_1} + \underline{\alpha_2}$ のとき，補題 3.5 より $\underline{\alpha_3} = \underline{b} \Rightarrow \alpha_3 = b$ となるため，

$$\alpha_3 = \alpha_1 + \alpha_2$$

となるため $\delta = 0$ である．よって

$$\begin{aligned}\Delta_3 &= \Delta_1 + \Delta_2 \\ &\le (n_1 - \underline{\alpha_1})\mathbf{u} \cdot \mathtt{ufp}(\alpha_1) + (n_2 - \underline{\alpha_2})\mathbf{u} \cdot \mathtt{ufp}(\alpha_2) \\ &\le (n_1 + n_2 - (\underline{\alpha_1} + \underline{\alpha_2}))\mathbf{u} \cdot \mathtt{ufp}(\alpha_3) \le (n_3 - \underline{\alpha_3})\mathbf{u} \cdot \mathtt{ufp}(\alpha_3)\end{aligned}$$

となる．

2. $\underline{\alpha_3} < \underline{b} \le \underline{\alpha_1} + \underline{\alpha_2}$ のとき $\underline{\alpha_3} \le \underline{\alpha_1} + \underline{\alpha_2} - 1$ となるため，

$$\begin{aligned}\Delta_3 &= \Delta_1 + \Delta_2 - \delta \\ &\le (n_1 - \underline{\alpha_1})\mathbf{u}\cdot\mathtt{ufp}(\alpha_1) + (n_2 - \underline{\alpha_2})\mathbf{u}\cdot\mathtt{ufp}(\alpha_2) + \mathbf{u}\cdot\mathtt{ufp}(\alpha_3) \\ &\le (n_1 + n_2 - (\underline{\alpha_1} + \underline{\alpha_2} - 1))\mathbf{u}\cdot\mathtt{ufp}(\alpha_3) \le (n_3 - \underline{\alpha_3})\mathbf{u}\cdot\mathtt{ufp}(\alpha_3)\end{aligned}$$

を得る．

以上より補題は証明された． □

補題 3.7 $1 \le m < 2$, $m \in \mathbb{F}'$ に対して

$$m \le \sum_{i=1}^{\underline{m}} 2^{1-i} \tag{3.41}$$

が成立する．さらに $p \in \mathbb{F}^n$, $p_i = \mathtt{ufp}(p_i)$ が与えられ，$\sum_{i=1}^{n} p_i \neq 0$, $\sum_{i=1}^{n} p_i = m \cdot 2^e$, $e \in \mathbb{Z}$ と置く．このとき

$$\sum_{i=1}^{\underline{m}} 2^{1-i} \le \sum_{i=1}^{n} 2^{1-i} \tag{3.42}$$

が成立する．

証明 不等式 (3.41) に関しては自明である．また，不等式 (3.42) に関しては，補題 3.4 を再帰的に使用すれば $\underline{m} \le n$ を得ることから証明される． □

補題 3.8 $p \in \mathbb{F}^n$, $p_i = \mathtt{ufp}(p_i)$ が与えられ，$1 \le m_1, m_2 < 2$, $m_1 \in \mathbb{F}'$, $m_2 \in \mathbb{F}$, $e_1, e_2 \in \mathbb{Z}$ とする．$\sum_{i=1}^{n} p_i \neq 0$ かつ $\mathrm{float}(\sum_{i=1}^{n} p_i) \neq 0$ を仮定し，

$$\sum_{i=1}^{n} p_i = m_1 \cdot 2^{e_1}, \quad \mathrm{float}(\sum_{i=1}^{n} p_i) = m_2 \cdot 2^{e_2}$$

と置く．このとき $n \le 1 - \log_2 \mathbf{u}$ ならば $e_1 \le e_2$ が成立する．

証明 $e_1 > e_2$ が成り立つと仮定する．このとき，

$$m_1 2^{e_1} = \sum_{i=1}^{n} p_i > \mathrm{float}(\sum_{i=1}^{n} p_i) = m_2 2^{e_2}$$

となる．ある実数 $\Delta > 0$ と $\alpha \in \mathbb{F}$ を

$$m_1 2^{e_1} = m_2 2^{e_2} + \Delta, \quad \alpha := \mathrm{float}(\sum_{i=1}^{n} p_i)$$

と置く．補題 3.6 から

$$\Delta \le (n-\underline{\alpha})\mathbf{u}\cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} p_i)) = (n-\underline{\alpha})\mathbf{u}2^{e_2}$$

となる．補題 3.7 の不等式 (3.41) より $m_2 \le \sum_{i=1}^{\underline{\alpha}} 2^{1-i}$ であるから

$$\begin{aligned} m_1 2^{e_1} &\le m_2 2^{e_2} + (n-\underline{\alpha})\mathbf{u}2^{e_2} \le 2^{e_2}\sum_{i=1}^{\underline{\alpha}} 2^{1-i} + (n-\underline{\alpha})\mathbf{u}2^{e_2} = 2^{e_2}(2-2^{-\underline{\alpha}+1} + (n-\underline{\alpha})\mathbf{u}) \\ &= 2^{e_2}(2 + n\mathbf{u} - (2^{-\underline{\alpha}+1} + \underline{\alpha}\mathbf{u})) \end{aligned}$$

を得る．ここで $1 \le \underline{\alpha} \le -\log_2 \mathbf{u}$ であり，$\underline{\alpha} = -\log_2 \mathbf{u}$ のときに $2^{-\underline{\alpha}+1} + \underline{\alpha}\mathbf{u}$ は最小になる．さらに $n \le 1 - \log_2 \mathbf{u}$ ならば

$$m_1 2^{e_1} \le 2^{e_2}(2 + (n - 2 + \log_2 \mathbf{u})\mathbf{u}) < 2\cdot 2^{e_2}$$

となり，$e_1 \le e_2$ が得られ，仮定した $e_1 > e_2$ に矛盾する． □

補題 3.9 $p \in \mathbb{F}^n$ に対して，$n \le 1 - \log_2 \mathbf{u}$ ならば

$$\sum_{i=1}^{n} \mathtt{ufp}(p_i) \le (2-2^{1-n})\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |p_i|)) \tag{3.43}$$

が成立する．

証明 すべての i において $p_i = 0$ ならば不等式 (3.43) は自明に成立する．よって，$\sum_{i=1}^{n} \mathtt{ufp}(p_i) \ne 0$ かつ $\mathrm{float}(\sum_{i=1}^{n} \mathtt{ufp}(p_i)) \ne 0$ の場合を考え，下記の表記を導入する．

$$\sum_{i=1}^{n} \mathtt{ufp}(p_i) = m_1 \cdot 2^{e_1}, \quad \mathrm{float}(\sum_{i=1}^{n} \mathtt{ufp}(p_i)) = m_2 \cdot 2^{e_2},$$
$$e_1, e_2 \in \mathbb{Z}, \quad 1 \le m_1, m_2 < 2, \quad m_1 \in \mathbb{F}', \ m_2 \in \mathbb{F}.$$

$a \in \mathbb{R}$ に対して不等式 (1.2) より $|a| \ge \mathtt{ufp}(a)$ であるから

$$\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |p_i|)) \ge \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} \mathtt{ufp}(p_i))) = \mathtt{ufp}(m_2 \cdot 2^{e_2}) = 2^{e_2}$$

となる．ここで，$\sum_{i=1}^{n} \mathtt{ufp}(p_i)$ と $\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |p_i|))$ の比を考える．補題 3.8 より $e_1 \le e_2$ で

あるから

$$\frac{\sum_{i=1}^{n} \mathtt{ufp}(p_i)}{\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |p_i|))} \leq \frac{m_1 \cdot 2^{e_1}}{2^{e_2}} \leq m_1 \tag{3.44}$$

を得る．また，補題 3.7 より $m_1 \leq \sum_{i=1}^{n} 2^{1-i} = 2 - 2^{1-n}$ であるため，この m_1 の上限を式 (3.44) に代入し，分母を払うことにより不等式 (3.43) が得られる． □

定理 3.10 $x, y \in \mathbb{F}^n$ に対して，$n \leq 1 - \log_2 \mathbf{u}$ のとき

$$|\mathrm{float}(x^T y) - x^T y| \leq (n + 1 - 2^{1-n})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(|x|^T |y|))$$

が成り立つ．ただし，float(·) 内でアンダーフローが発生しないとする．さらに不等式の左辺と右辺にある float(·) の計算順は同じとする．

証明 $p_i = x_i y_i$，$\tilde{p}_i = \mathrm{fl}(x_i \cdot y_i)$ とする．内積における誤差の上限を計算するため，不等式 (3.11) のように総和の計算の誤差と 2 項の積に関する誤差に分ける．

$$|\mathrm{float}(x^T y) - x^T y| \leq |\mathrm{float}(\sum_{i=1}^{n} \tilde{p}_i) - \sum_{i=1}^{n} \tilde{p}_i| + \sum_{i=1}^{n} |\tilde{p}_i - p_i| \tag{3.45}$$

ここで，定理 2.2 と補題 3.9 より

$$\sum_{i=1}^{n} |\tilde{p}_i - p_i| \leq \mathbf{u} \sum_{i=1}^{n} \mathtt{ufp}(\tilde{p}_i) \leq (2 - 2^{1-n})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |\tilde{p}_i|)) \tag{3.46}$$

を得る．よって，誤差評価式 (1.5) と不等式 (3.46) をそれぞれ不等式 (3.45) の右辺に使用すれば

$$\begin{aligned} |\mathrm{float}(x^T y) - x^T y| &\leq (n-1)\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |\tilde{p}_i|)) + (2 - 2^{1-n})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |\tilde{p}_i|)) \\ &= (n + 1 - 2^{1-n})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |\tilde{p}_i|)) \end{aligned}$$

となる． □

誤差に関する係数の改良により，定理 3.10 の誤差評価式はポイント (b) の観点では悪化したが，$n \leq 1 - \log_2 \mathbf{u}$ のとき

$$|\mathrm{float}(x^T y) - x^T y| \leq (n + 1)\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(|x|^T |y|))$$

とポイント (b) を達成する簡単な係数で評価してもよい．

3.4 一般の n に対する係数の改良

ここでは，ufp を用いた内積の誤差評価について，ポイント (a) を満たす誤差評価式を導出する．

補題 3.11 $p \in \mathbb{F}^n$ に対して，

$$\sum_{i=1}^{n} \mathtt{ufp}(p_i) \le (2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u})\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |p_i|)) \tag{3.47}$$

が成り立つ．

証明 すべての i に対して $p_i = 0$ ならば不等式 (3.47) は自明に成立する．以下，少なくともある i において $p_i \neq 0$ とし，下記の表記を導入する．

$$\sum_{i=1}^{n} \mathtt{ufp}(p_i) = m_1 2^{e_1},\ \mathrm{float}(\sum_{i=1}^{n} \mathtt{ufp}(p_i)) = m_2 2^{e_2},$$
$$e_1, e_2 \in \mathbb{Z},\ 1 \le m_1, m_2 < 2,\ m_1 \in \mathbb{F}',\ m_2 \in \mathbb{F}$$

まず

$$m_1 2^{e_1} \le 2^{e_2}(2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u}) \tag{3.48}$$

となることを以下の 2 通りに場合分けをして示す．

1. $m_1 2^{e_1} \le m_2 2^{e_2}$ のとき $e_1 \le e_2$ となる．ここで $f(n) = (n + \log_2 2^{-E} - 2)2^{-E} + 2^{1-n}$ と置く（ただし E は自然数）．$n < E$ のとき $f(n)$ は単調減少，$f(E) = f(E+1) = 0$，$E + 1 < n$ のとき単調増加であるため，自然数 n に対しては $f(n) \ge 0$ が証明される．以上と補題 3.7 により
$$m_1 2^{e_1} \le 2^{e_1}(2 - 2^{1-n}) \le 2^{e_2}(2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u})$$
を得る．よって式 (3.48) は成り立つ．
2. $m_1 2^{e_1} > m_2 2^{e_2}$ のときを考える．$\alpha := m_2 2^{e_2}$ とすると，補題 3.6 より
$$m_1 2^{e_1} = m_2 2^{e_2} + \Delta, \quad \Delta \le (n - \underline{\alpha})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} \mathtt{ufp}(p_i))) = (n - \underline{\alpha})\mathbf{u}2^{e_2}$$
が成り立つ．さらに補題 3.7 を用いて
$$\begin{aligned} m_1 2^{e_1} &\le m_2 2^{e_2} + (n - \underline{\alpha})\mathbf{u}2^{e_2} \le 2^{e_2} \sum_{i=1}^{\underline{\alpha}} 2^{1-i} + (n - \underline{\alpha})\mathbf{u}2^{e_2} \\ &= 2^{e_2}(2 - 2^{-\underline{\alpha}+1} + (n - \underline{\alpha})\mathbf{u}) = 2^{e_2}(2 - \mathbf{u}2^{-\underline{\alpha}+1-\log_2 \mathbf{u}} + (n - \underline{\alpha})\mathbf{u}) \\ &= 2^{e_2}(2 + n\mathbf{u} - \mathbf{u}(2^{-\underline{\alpha}+1-\log_2 \mathbf{u}} + \underline{\alpha})) \end{aligned}$$

を得る．このとき，α が浮動小数点数であることから $1 \leq \underline{\alpha} \leq -\log_2 \mathbf{u}$ であり，$\underline{\alpha} = -\log_2 \mathbf{u}$ のとき $2^{-\underline{\alpha}+1-\log_2 \mathbf{u}} + \underline{\alpha}$ は最小となるため

$$m_1 2^{e_1} \leq 2^{e_2}(2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u})$$

となる．よって式 (3.48) は成り立つ．

次に式 (3.48) より

$$\frac{\sum_{i=1}^{n} \mathtt{ufp}(p_i)}{\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} \mathtt{ufp}(p_i)))} \leq \frac{2^{e_2}(2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u})}{2^{e_2}} = 2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u}$$

となり，上式の分母を払い，不等式 (1.2) を用いて

$$\begin{aligned}\sum_{i=1}^{n} \mathtt{ufp}(p_i) &\leq (2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u}) \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} \mathtt{ufp}(p_i))) \\ &\leq (2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u}) \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |p_i|))\end{aligned}$$

を得る． □

定理 3.12 $x, y \in \mathbb{F}^n$ に対して

$$|\mathrm{float}(x^T y) - x^T y| \leq (n + 1 + (n + \log_2 \mathbf{u} - 2)\mathbf{u})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(|x|^T |y|))$$

が成立する．ただし，浮動小数点演算の際にアンダーフローが発生しないことを仮定する．また，不等式の左辺と右辺で float(·) 内の演算順序は同じとする．

証明 $p_i = x_i y_i$, $\tilde{p}_i = \mathrm{fl}(x_i \cdot y_i)$ とする．内積における誤差の上限を計算するため，式 (3.11) のように総和の計算の誤差と積に関する誤差に分ける．

$$|\mathrm{float}(x^T y) - x^T y| \leq |\mathrm{float}(\sum_{i=1}^{n} \tilde{p}_i) - \sum_{i=1}^{n} \tilde{p}_i| + \sum_{i=1}^{n} |\tilde{p}_i - p_i| \tag{3.49}$$

ここで定理 2.2 と補題 3.11 より

$$\sum_{i=1}^{n} |\tilde{p}_i - p_i| \leq \sum_{i=1}^{n} \mathbf{u} \cdot \mathtt{ufp}(\tilde{p}_i) \leq (2 + (n + \log_2 \mathbf{u} - 2)\mathbf{u})\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |\tilde{p}_i|)) \tag{3.50}$$

を得る．よって，誤差評価式 (1.5) と不等式 (3.50) をそれぞれ不等式 (3.49) の右辺にあてはめれば

$$|\mathrm{float}(x^T y) - x^T y| \leq (n - 1)\mathbf{u} \cdot \mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n} |\tilde{p}_i|))$$

$$+(2+(n+\log_2\mathbf{u}-2)\mathbf{u})\mathbf{u}\cdot\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n}|\tilde{p}_i|))$$
$$=(n+1+(n+\log_2\mathbf{u}-2)\mathbf{u})\mathbf{u}\cdot\mathtt{ufp}(\mathrm{float}(\sum_{i=1}^{n}|\tilde{p}_i|))$$

と上限をとれる． □

定理 3.12 の結果は任意の n に対して成立するうえに，誤差評価式の係数は $(n+2)\mathbf{u}<1$ の範囲で誤差評価式 (3.5) の係数よりも小さい．

3.5 内積におけるアンダーフローの取り扱い

誤差評価式 (3.6) と (3.7) では最後に $\mathbf{u}_N$ を足すことでアンダーフローへの対策を施している．これは $n<\mathbf{u}^{-1}$ のときは過大評価であり，また右辺の最小値は $\mathbf{u}_N$ よりも小さくはならない．ただし誤差評価式 (3.4) のように非正規化数である $\frac{n}{2}\mathbf{u}_S$ を足しては，計算のパフォーマンスを下げることがある．本節ではポイント (e) に関して誤差評価式を改良する．

定理 3.13 $x,y\in\mathbb{F}^n$, $2(n+1)\mathbf{u}\le 1$ に対して

$$|\mathrm{float}(x^Ty)-x^Ty|\le \mathrm{fl}((n+2)\mathbf{u}\cdot(\mathtt{ufp}(\mathrm{float}(|x|^T|y|))+\mathbf{u}_N))$$

が成立する．ここでは，不等式の左辺と右辺にある float(·) の計算順は同じとする．

証明 まず $\mathrm{fl}(|x_i\cdot y_i|)\ge\mathbf{u}_N$ ならば $d_i=1$ であり，$\mathrm{fl}(|x_i\cdot y_i|)<\mathbf{u}_N$ ならば $d_i=0$ とし，$d=\sum_{i=1}^{n}d_i$ とする．表記の簡略化のために $A=\mathtt{ufp}(\mathrm{float}(|x|^T|y|))$ と置く．$d=n$ であれば，定理 3.1 の誤差評価式 (3.8) と n に関する仮定から得る $\mathrm{fl}((n+2)\mathbf{u})=(n+2)\mathbf{u}$ から

$$|\mathrm{float}(x^Ty)-x^Ty|\le\mathrm{fl}((n+2)\mathbf{u}\cdot A)\le\mathrm{fl}((n+2)\mathbf{u}\cdot(A+\mathbf{u}_N))$$

と証明できるため，以後は $0\le d\le n-1$ について考える．以降，A について 2 通りに場合を分けて考える．

1. $A\ge\mathbf{u}^{-1}\mathbf{u}_N$ のとき，$(n+2)\mathbf{u}$ と 2 のべき乗数 A との積においてアンダーフローが発生しないため

   (3.51) $$(n+2)\mathbf{u}A=\mathrm{fl}(((n+2)\mathbf{u})\cdot A)\le\mathrm{fl}(((n+2)\mathbf{u})\cdot(A+\mathbf{u}_N))$$

   となる．定理 3.1 の誤差評価式 (3.8) より

   $$|\mathrm{float}(x^Ty)-x^Ty|\le(n+1+n\mathbf{u}-\mathbf{u})\mathbf{u}A+\frac{n-d}{2}\mathbf{u}_S$$
   $$=(n+2+n\mathbf{u}-\mathbf{u}-1)\mathbf{u}A+\frac{n-d}{2}\mathbf{u}_S$$

を得る．ここで，仮定による n の制限から得る $n\mathbf{u} - \mathbf{u} - 1 < 0$ と，仮定である $A \geq \mathbf{u}^{-1}\mathbf{u}_N$ から得られる $\mathbf{u}A \geq \mathbf{u}_N$ より

$$(n\mathbf{u} - \mathbf{u} - 1)\mathbf{u}A + \frac{n-d}{2}\mathbf{u}_S \leq (n\mathbf{u} - \mathbf{u} - 1)\mathbf{u}_N + \frac{n}{2}\mathbf{u}_S < 0$$

となり [*1]，

$$|\mathrm{float}(x^T y) - x^T y| \leq (n+2)\mathbf{u}A$$

が得られ，この式と不等式 (3.51) をつなげればよい．

2. $A < \mathbf{u}^{-1}\mathbf{u}_N$ のときを考える．A が 2 のべき乗数であるから定理 2.4 より

$$\mathrm{fl}(A + \mathbf{u}_N) = A + \mathbf{u}_N \tag{3.52}$$

を導出できる．これは定理 2.4 において $A \leq \mathbf{u}_N$ のとき，$\mathbf{u}c = \mathbf{u}_S$ として，それ以外のときには $\mathbf{u}c = \mathbf{u}_N$ と置くことにより証明される．等式 (2.3) より $\mathbf{u} \cdot \mathbf{u}_N = \frac{1}{2}\mathbf{u}_S$ であることから，

$$-(n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u} \cdot \mathbf{u}_N + \frac{n}{2}\mathbf{u}_S \leq -\frac{1}{2}\mathbf{u}_S \tag{3.53}$$

が得られる．ここで，定理 3.1 の誤差評価式 (3.8) からはじめて

$$\begin{aligned}
|\mathrm{float}(x^T y) - x^T y| &\leq (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u}A + \frac{n-d}{2}\mathbf{u}_S \\
&\leq (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u}A + \frac{n}{2}\mathbf{u}_S \\
&= (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u}(A + \mathbf{u}_N - \mathbf{u}_N) + \frac{n}{2}\mathbf{u}_S \\
&= (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u}(A + \mathbf{u}_N) - (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u} \cdot \mathbf{u}_N + \frac{n}{2}\mathbf{u}_S \qquad (3.54) \\
&\leq (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u}(A + \mathbf{u}_N) - \frac{1}{2}\mathbf{u}_S \qquad (3.55) \\
&< (n + 1 + n\mathbf{u} - \mathbf{u})\mathbf{u}(A + \mathbf{u}_N) \qquad (3.56)
\end{aligned}$$

が導出される．(3.54) から (3.55) の変形は不等式 (3.53) による．ここで，定理 2.2 より $|\delta| < \mathbf{u}$, $|\eta| \leq \frac{1}{2}\mathbf{u}_S, \delta\eta = 0$，さらに等式 (3.52) を用いて

$$\begin{aligned}
\mathrm{fl}((n+2)\mathbf{u} \cdot (A + \mathbf{u}_N)) &= (1+\delta)\mathrm{fl}((n+2)\mathbf{u})\mathrm{fl}(A + \mathbf{u}_N) + \eta \\
&= (1+\delta)(n+2)\mathbf{u}(A + \mathbf{u}_N) + \eta \qquad (3.57)
\end{aligned}$$

が成立する．$\delta\eta = 0$ より，ここから「$\eta = 0$」と「$\delta = 0$」の場合に分ける．

[*1] 最後の不等式は，等式 (2.3) より $\mathbf{u}_N = \frac{1}{2}\mathbf{u}^{-1} \cdot \mathbf{u}_S$ であることから，すぐに成立が確認できる．

（a）式 (3.57) において $\eta = 0$ を考える．n の制限から $1 - n\mathbf{u} - 2\mathbf{u} > n\mathbf{u} - \mathbf{u}$ が成立するため，

$$\begin{aligned}\mathrm{fl}((n+2)\mathbf{u}\cdot(A+\mathbf{u}_N)) &= (1+\delta)(n+2)\mathbf{u}(A+\mathbf{u}_N) \geq (1-\mathbf{u})(n+2)\mathbf{u}(A+\mathbf{u}_N)\\ &= (n+2-n\mathbf{u}-2\mathbf{u})\mathbf{u}(A+\mathbf{u}_N)\\ &\geq (n+1+n\mathbf{u}-\mathbf{u})\mathbf{u}(A+\mathbf{u}_N)\end{aligned}$$

となり，これを式 (3.56) へと不等式をつなげればよい．

（b）式 (3.57) において $\delta = 0$ の場合は

$$\begin{aligned}\mathrm{fl}((n+2)\mathbf{u}\cdot(A+\mathbf{u}_N)) &= (n+2)\mathbf{u}(A+\mathbf{u}_N)+\eta \geq (n+2)\mathbf{u}(A+\mathbf{u}_N) - \frac{1}{2}\mathbf{u}_S\\ &\geq (n+1+n\mathbf{u}-\mathbf{u})\mathbf{u}(A+\mathbf{u}_N) - \frac{1}{2}\mathbf{u}_S\end{aligned}$$

となり，この不等式と式 (3.55) をつなげればよい．

□

定理 3.13 の誤差上限は $n = 1$ かつ $\mathrm{float}(|x|^T|y|) = 0$ のとき最小値 $2\mathbf{u}_S$ となり，誤差評価式 (3.6) と (3.7) では最小値が $\mathbf{u}_N$ である点が異なる．ここで $\mathbf{u}_N$ を足すことにより誤差上限の変化を考える．誤差評価式 (3.6) と (3.7) では $(n+2)\mathbf{u}$ をかけて A より小さくなった値に $\mathbf{u}_N$ を足し，定理 3.13 の誤差上限は A そのものに $\mathbf{u}_N$ を足すことから，$\mathrm{fl}(A+\mathbf{u}_N) = A$ になる場合は多くなるため，定理 3.13 のほうが誤差上限が良い場合が多いと考える．

3.6 FMA を用いた場合の内積に対する誤差評価

最後に Fused Multiply-Add を用いた $x, y \in \mathbb{F}^n$ に対する内積の最適な誤差評価を与える．$\mathrm{fl}_{\mathrm{Fd}}(x, y)$ は

$t = \mathrm{fl}(x_1 \cdot y_1)$
`for` $i = 2 : n$
　$t = \mathrm{fl}_{\mathrm{F}}(x_i, y_i, t)$
`end`

と計算する表記とする．一般に内積計算の丸め誤差解析は，任意の計算順序も考慮に入れた式である $\mathrm{float}(\cdot)$ を考えてきた．ただし，FMA を使用する場合は任意の計算順序を考えず，逐次的な計算方法についてのみ結果を与える．$x, y \in \mathbb{F}^4$ に対する内積 $x^T y$ に対して，前から順に計算する

$$\mathrm{fl}_{\mathrm{F}}(x_4, y_4, \mathrm{fl}_{\mathrm{F}}(x_3, y_3, \mathrm{fl}_{\mathrm{F}}(x_2, y_2, \mathrm{fl}(x_1 \cdot y_1))))$$

においては FMA は 3 回使用可能である．また，2 項どうしの計算結果を最後に足し合わせる

$$\mathrm{fl}(\mathrm{fl}_{\mathrm{F}}(x_2, y_2, \mathrm{fl}(x_1 \cdot y_1))) + \mathrm{fl}_{\mathrm{F}}(x_4, y_4, \mathrm{fl}(x_3 \cdot y_3)))$$

では FMA を 2 回しか使用できない．このため，FMA を使用する回数が計算順により異なるため，本節では計算順を固定する．FMA を用いた内積計算に対して，次の定理が成立する．

定理 3.14 $x, y \in \mathbb{F}^n$ に対して，FMA による $\mathrm{fl}_{\mathrm{F}}(x_i, y_i, \mathrm{fl}_{\mathrm{Fd}}(x_{1:i-1}, y_{1:i-1})) \geq \mathbf{u}_N$ のとき $d_i = 1$ とし，それ以外のときに $d_i = 0$ とする．ただし，d_1 については $\mathrm{fl}(x_1 y_1) \geq \mathbf{u}_N$ のとき $d_1 = 1$，そうでないときに $d_1 = 0$ とする．$d = \sum_{i=1}^{n} d_i$ として

$$\begin{aligned} |\mathrm{fl}_{\mathrm{Fd}}(x, y) - (x^T y)| &\leq d\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|, |y|)) + \frac{n-d}{2}\mathbf{u}_S \\ &\leq n\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|, |y|)) + \frac{n}{2}\mathbf{u}_S \end{aligned}$$

が成り立つ．アンダーフローが発生しないとき，この誤差評価式は最適である．

証明　ベクトル $x_{1:k} = (x_1, x_2, \ldots, x_k)^T \in \mathbb{F}^k$ という表記を導入し（y についても同様とする），帰納法により証明する．

$n = 1$ のとき，定理 2.2 より自明に成立する．$n = k$ のとき，

$$|\mathrm{fl}_{\mathrm{Fd}}(x_{1:k}, y_{1:k}) - x^T y| = \sum_{i=1}^{k} d_i \cdot \mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x_{1:k}|, |y_{1:k}|)) + \sum_{i=1}^{k} \frac{1-d_i}{2}\mathbf{u}_S$$

が成り立つと仮定する．

$n = k + 1$ のとき，定理 2.3 を用いて

$$\begin{aligned} &|\mathrm{fl}_{\mathrm{Fd}}(x_{1:k+1}, y_{1:k+1}) - \sum_{i=1}^{k+1} x_i y_i| \\ &= |\mathrm{fl}_{\mathrm{F}}(x_{k+1}, y_{k+1}, \mathrm{fl}_{\mathrm{Fd}}(x_{1:k}, y_{1:k})) - (\sum_{i=1}^{k} x_i y_i + x_{k+1} y_{k+1})| \\ &= |\mathrm{fl}_{\mathrm{Fd}}(x_{1:k}, y_{1:k}) + x_{k+1} y_{k+1} + \delta + \eta - (\sum_{i=1}^{k} x_i y_i + x_{k+1} y_{k+1})| \\ &\quad \left(|\delta| \leq d_{k+1} \cdot \mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x_{1:k}|, |y_{1:k}|) + |x_{k+1}||y_{k+1}|), \quad |\eta| \leq \frac{1-d_{k+1}}{2}\mathbf{u}_S \right) \\ &\leq |\mathrm{fl}_{\mathrm{Fd}}(x_{1:k}, y_{1:k}) - \sum_{i=1}^{k} x_i y_i| + |\delta| + |\eta| \\ &\leq \sum_{i=1}^{k} d_i \cdot \mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x_{1:k}|, |y_{1:k}|)) + \sum_{i=1}^{k} \frac{1-d_i}{2}\mathbf{u}_S \\ &\quad + d_{k+1}\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x_{1:k+1}|, |y_{1:k+1}|)) + \frac{1-d_{k+1}}{2}\mathbf{u}_S \end{aligned}$$

$$\leq \sum_{i=1}^{k+1} d_i \cdot \mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x_{1:k+1}|, |y_{1:k+1}|)) + \sum_{i=1}^{k+1} \frac{1-d_i}{2}\mathbf{u}_S$$

となり，$n = k+1$ のときも成り立つ．以上より，すべての自然数に対して成立する．ここで $0 \leq d \leq n$ であるから

$$|\mathrm{fl}_{\mathrm{Fd}}(x, y) - (x^T y)| \leq n\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|, |y|)) + \frac{n}{2}\mathbf{u}_S$$

が成立する．

もし，アンダーフローがすべてに発生しない場合には

$$|\mathrm{fl}_{\mathrm{Fd}}(x, y) - (x^T y)| \leq n\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|, |y|)) \tag{3.58}$$

となる．この不等式において $x = (1 + 2^{-2}, 1, \ldots, 1)^T$, $y = (1 + 2^2\mathbf{u}, \mathbf{u}, \ldots, \mathbf{u})^T \in \mathbb{F}^n$ の場合を考える．ベクトルの内積は

$$\begin{aligned} x^T y &= (1 + 2^{-2}, 1, \ldots, 1)(1 + 2^2\mathbf{u}, \mathbf{u}, \ldots, \mathbf{u})^T \\ &= (1 + 2^{-2})(1 + 2^2\mathbf{u}) + 1 \cdot \mathbf{u} + \cdots + 1 \cdot \mathbf{u} \\ &= (1 + 2^{-2} + 2^2\mathbf{u} + \mathbf{u}) + \mathbf{u} + \cdots + \mathbf{u} = 1 + 2^{-2} + 2^2\mathbf{u} + n\mathbf{u} \end{aligned}$$

となる．また，

$$\mathrm{fl}_{\mathrm{F}}(1, \mathbf{u}, \mathrm{fl}((1 + 2^{-2})(1 + 2^2\mathbf{u}))) = \mathrm{fl}_{\mathrm{F}}(1, \mathbf{u}, 1 + 2^{-2} + 2^2\mathbf{u}) = 1 + 2^{-2} + 2^2\mathbf{u}$$

であるから

$$\mathrm{fl}_{\mathrm{Fd}}(x, y) \;=\; 1 + 2^{-2} + 2^2\mathbf{u}$$

を得る．また，

$$\mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|, |y|)) \;=\; \mathtt{ufp}(1 + 2^{-2} + 2^2\mathbf{u}) \;=\; 1$$

となる．よって

$$\begin{aligned} |\mathrm{fl}_{\mathrm{Fd}}(x, y) - x^T y| &= |(1 + 2^{-2} + 2^2\mathbf{u}) - (1 + 2^{-2} + 2^2\mathbf{u} + n\mathbf{u})| \\ &= n\mathbf{u} = n\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|, |y|)) \end{aligned}$$

となるため，$|x|^T|y| \neq 0$ に対して不等式 (3.58) の等号が成り立つ例がすべての n に存在するため，不等式は最適である． □

定理 3.14 の結果については，ポイント (a)，(b)，(d)，(e) を達成できている．ポイント (c) を達成するならば，$n\mathbf{u} \leq 1$ のときに

$$|\mathrm{fl}_{\mathrm{Fd}}(x, y) - (x^T y)| \leq \mathrm{fl}(n\mathbf{u} \cdot \mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|, |y|)))$$

と右辺全体を浮動小数点演算にて評価できる．

FMA を用いた内積のアンダーフローにも対応する誤差評価について，定理 3.13 のように $\mathbf{u}_N$ を加える位置を工夫できたらよいが，それが不可能なことを以下に示す．

注意 1 $x, y \in \mathbb{F}^n$ に対して，$n\mathbf{u} \le 1$ のとき

$$|\mathrm{fl}_{\mathrm{Fd}}(x,y) - x^T y| \le \mathrm{fl}(n\mathbf{u} \cdot (\mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|,|y|)) + \mathbf{u}_N)) \tag{3.59}$$

は少なくとも $n = 4k-3$, $k \in \mathbb{N}$ のときに成立しない．

証明 $x = (0.5, \ldots, 0.5)^T$, $y = (\mathbf{u}_S, \ldots, \mathbf{u}_S)^T$ を考える．このとき，$x^T y = \dfrac{n}{2}\mathbf{u}_S$ であり，また $\mathrm{fl}_{\mathrm{Fd}}(x,y) = 0$ となる．ただし，$n = 4k-3$, $k \in \mathbb{N}$ のとき，$\mathrm{fl}(n\mathbf{u} \cdot (\mathtt{ufp}(\mathrm{fl}_{\mathrm{Fd}}(|x|,|y|)) + \mathbf{u}_N)) = \dfrac{n-1}{2}\mathbf{u}_S$ であることから不等式 (3.59) は成立しない． □

4. おわりに

本稿では，浮動小数点演算における内積の丸め誤差の評価について，先行研究を改良または拡張した誤差評価式を導出できた．具体的には，長さが 2 のベクトルの内積における最適な誤差評価式の導出，長さが小さいときの内積の誤差評価の改良，長さに制約のない誤差評価式の導出である．さらに，丸め誤差解析におけるアンダーフローに関する対策法や，Fused Multiply-Add を用いた場合の内積の丸め誤差解析について提案した．

謝辞 終始丁寧に査読をいただき，有益なコメントをいただきました 2 名の査読者の方に深く感謝をいたします．本研究は，科学研究費補助金・若手研究 B，誤差解析の意味で高精度な数値線形計算の基盤の創出（研究課題番号：25730076）の支援を受けた．

参考文献

[1] Dekker, T.J., A floating-point technique for extending the available precision, Numerische Mathematik, **18** (1971), 224–242.

[2] Higham, N.J., Accuracy and Stability of Numerical Algorithms, second edition, SIAM Publications, Philadelphia, 2002.

[3] IEEE Standard for Floating-Point Arithmetic, Std 754–2008, 2008.

[4] Jeannerod, C.-P., Rump, S.M., Improved error bounds for inner products in floating-point arithmetic, SIAM. J. Matrix Anal. & Appl. (SIMAX), **34** (2013), 338–344.

[5] Jeannerod, C-P., Rump, S.M., On relative errors of floating-point operations: optimal bounds and applications, Research Report, 2014, pp.15, hal-00934443.

[6] Rump, S.M., Error estimation of floating-point summation and dot product, BIT Numer. Math., **52** (2012), 201-220.

[7] Rump, S.M., Computable backward error bounds for basic algorithms in linear algebra. Nonlinear Theory and Its Applications, IEICE, **6** (2015), 360–363.

[8] Winograd, S., A new algorithm for inner product, IEEE Trans. Comput., C-18 (1968), 693–694.

樋口 裕幸 (非会員) 〒337-8570 さいたま市見沼区深作 307

2014 年芝浦工業大学システム理工学部数理科学科卒業．現在，芝浦工業大学大学院理工学研究科システム理工学専攻修士課程．応用数学，特に浮動小数点演算の誤差解析の理論に興味を持つ．

尾崎 克久 (正会員) 〒337-8570 さいたま市見沼区深作 307

2007 年早稲田大学大学院理工学研究科博士課程修了，博士 (工学)．現在，芝浦工業大学システム理工学部数理科学科准教授．精度保証付き数値計算，誤差解析，高精度計算の研究に従事．