**PHÂN TÍCH HƯỚNG PHÁT TRIỂN CHAPTER 23**

Trước tiên chúng ta sẽ xem xét các số có dấu phẩy động z với p chữ số có nghĩa trong cơ số b, và chúng ta viết z= f (z).be(z) trong đó e (z) là phần mũ và f (z) là phần phân số. Nếu z 0, chúng ta giả sử rằng 1> |f(z)| 1/b, và f (z).bp là một số nguyên; ngược lại, 0 và tất cả các số như vậy là số dấu phẩy động.

Bây giờ chúng ta hãy xem xét chuỗi các số dấu phẩy động z, w, x', w', x ",…, trong đó w= x y, x'= wy , w '=x' y, x" = w 'y, v.v. Nếu x = x' thì w = w' ; nếu w = w ' thì x'= x "; và v.v. Để kiểm tra khả năng sai lệch, chúng ta sẽ xác định tất cả các giá trị dấu phẩy động x và y mà w w'.

Đối với mỗi phép cộng và phép trừ, chúng ta sẽ xem xét sự liên kết tương đối của các toán hạng và kết quả; và với mỗi liên kết, chúng ta cần xem xét sự đa dạng giá trị của lỗi làm tròn cho đến khi đạt đến trường hợp rõ ràng rằng w có thể hoặc không thể bằng w '.

Nếu bất kỳ hoạt động nào trong chuỗi x, w, x, x', w', x'',…, là chính xác thì tất cả các phép toán tiếp theo cũng vậy. Ví dụ: nếu w= x y = x + y thì x' = wy = round((x +y)-y), vì x là số thực có dấu phẩy động.

Chúng ta có thể giả định rằng x là dương, vì tất cả các đối số của chúng ta sẽ hợp lệ đối với phủ định x bằng cách thay đổi các dấu hiệu thích hợp; lưu ý rằng dấu phẩy động được xác định ở trên là đối xứng về không. Với sự hiểu biết này, chúng tôi đã chứng minh và phân tích cho các trường hợp.

*Trường hợp 1*, e(y) > e(x). Cho d= e(y) -e (x) và c = e (w) - e (y). Chúng ta có d1 và c1. Với mỗi kí tự nhỏ nhất z biểu thị tại số thực có dấu phẩy động, hãy đặt chữ cái in hoa tương ứng Z biểu thị số thực f (z).be(z)- e(x)+p. Khi đó X, Y, W đều là số nguyên và X + Y = W + R với |R| bd+c. Chúng ta có thể giả định rằng R0, hoặc phép cộng là chính xác, do đó d+c>0.

Với e(y) = 5; e(w) =2; d= 4; c=-3 → d + c = 1 > 0.

* Trường hợp 1A, với c0 thì R= X+Y –W X (modulo bd+c ), bởi vì Y0 (modulo bd ) and W 0 (modulo bd+c ). Vì thế, W –Y = X-R có nhiều nhất (p+1) –(d+c) = 10 chữ số có nghĩa → p = 10, và wy = w-y.
* Trường hợp 1B, với c=1 thì y> 0 và R X + Y0.bd (modulo bd+c ), trong đó Y0 là chữ số thứ p có nghĩa của f (y). Theo đó wy là chính xác, vì W - Y có nhiều nhất (p + 1) –d ≤ p <=> 7 ≤ 10 chữ số có nghĩa.

Trường hợp 2, e (x) e (y). Cho d= e(x) – e(y) và c = e(w) – e(x),; chúng ta có d và c≤1. Hãy xác định lại quy ước của *Trường hợp 1* sao cho Z = f (z). be(z)- e(x)+p. Khi đó X, Y, W đều là các số nguyên sao cho X + Y = W + R, với |R| bd+c như trước đó chúng ta đã giả sử d+c > 0.

Với z = 0,1528535407.106 ; f(x) = 0,1528535407; b = 10; e(z) - e(y) + p = 6 → {p = 10; e(y) = 1; e(z) = -3}. Chọn e(x) = 5; e(w) = 2; d = 4; c= -3 → d+c = 1 > 0

* Trường hợp 2A, với c -2 thì d+c > 0 ngụ ý rằng d 3, do đó X + Y ≥ bp+d-1 - (b p - 1)> b p + d - 2 <=> 1010 + 4 - 1  - (1010 -1) > 1010 + 4 - 2, mâu thuẫn với c -2.
* Trường hợp 2B, c = -1. Khi đó d> 1 và Y <0; chúng ta có |R| ≤ bd-1 = 104 - 1 và RY (modulo bd-1).
  + Trường hợp 2B.1, e (x') = e (x) = e (w) + 1. Cho W - Y = X '+ S, trong đó |S| ≤bd = 104. Khi đó |X - X '| = |R + S | ≤bd-1 + bd <bd <=> 10 4 -1 + 104 < 104 → Đúng ,nên x=x'.
  + Trường hợp 2B.2, e (x') = e (x) - 1 = e (w)+1. Cho W - Y = X '+ S trong đó S -Y (modulobd-1) và |S| ≤bd-1 = 10 4 - 1
    - Trường hợp 2B.2.a, Y bd-1 = 10 4 - 1(modulo bd-1). Sau đó, có các lựa chọn duy nhất cho S và R, do đó S = -R và X - X '= R + S = 0.
    - Trường hợp 2B.2.b, Y bd-1 = 10 4 - 1(modulo bd - 1). Vì bp + d - 1 ≤ X = X '+ R + S ≤ (bp+d-1 –bd-1 ) +bd-1 + bd-1 nên ta có X = bp + d - 1  = 1013, X'=b p + d - 1 - bd - 1 = 1013 - 103 và R = S =bd-1 = 103. Có thể có w ' w, và trường hợp này sẽ được thảo luận thêm bên dưới.
* Trường hợp 2C, c = 0. Ta có d≥ 1, R Y (modulo bd), R≤ bd = 104.
* Trường hợp 2C.1, |R.| < bd = 104
* Trường hợp 2C.1.a, X> b p+ d - 1 = 1013 hoặc R <0. Khi đó chúng ta có round (W - Y) = round (X-R) = X, do đó x '= x.
* Trường hợp 2C.1.b, X = bp + d - 1 = 1013và R> 0. Khi đó Y> 0, e (x ') = e (x) -1, và W - Y = X' + S trong đó | S | ≤ bd-1 = 103
* Trường hợp 2C.1.b.1, W> b p+ d - 1 = 1013 hoặc S <bd-1 = 103. Khi đó round (X '+ Y) = round (W-S) = W, do đó w' = w.
* Trường hợp 2C.1.b.2, W = b p+ d - 1 = 1013và S = bd-1 = 103. Theo đó, w ≠ w 'là có thể; xem bên dưới .
* Trường hợp 2C.2, |R| = bd = 104. Sau đó, Theo đó, w ≠ w 'là có thể có trong nhiều trường hợp phụ được phân tích dưới đây.
* Trường hợp 2D, c= 1. Khi đó y> 0.
* Trường hợp 2D.1, e (x ') = e (w) = e (x) + 1. Khi đó W – Y= X' + S, với |S|≤ bd+1 = 105. Hơn thế nữa W – Y bp+d - bd  = 1014 - 104, và b p+d = 1014 ≤ X' = X – R – S ≤ X + bd+1  = 105. Từ đó, X' = bp+d và R + S ≤ -bd và S bd. Nó tuân theo quy luật rằng S < bd+1 = 105và W' = round(X'+Y) = round(W – S) = W.
* Trường hợp 2D.2, e (x') = e (w) -1 = e (x). Khi đó W - Y = X '+ S, trong đó S Y (modulo bd) và | S | ≤ bd = 104. Chúng ta có thể cho rằng d≥ 1.
* Trường hợp 2D.2.a, e (w ') = e (x'). Khi đó X '+ Y = W' + R ', trong đó |R'| = bd+1 = 105 (modulo bd) và |R'| ≤ bd = 104. Vì bp + d = 1014 ≤ W = W '+ S + R' ≤ (bp + d - bd) +bd + bd = b p+ d <=> (1014  - 104) + 104 + 104  = 1014 nên ta phải có W = b p+ d = 1014, S = R' = bd = 104, W '= bp + d- bd = 1014  - 104 .Trường hợp này được thảo luận thêm dưới đây.
* Trường hợp 2D.2.b, e (w ') = e (w). Khi đó X '+ Y = W' + R ', trong đó | R '| ≤ bd+1 = 105 và |W - W '| = | S + R '| <bd +1 = 105, do đó w '= w.

Phân tích trường hợp đã hoàn tất và nó cho thấy rằng một quy tắc làm tròn chung như được mô tả ở trên sẽ tạo ra w = w 'ngoại trừ trong bốn trường hợp. Không có trường hợp nào có thể xảy ra khi cơ số b là lẻ, do đó chúng ta có thể giả sử rằng b là chẵn. Ba trong số bốn trường hợp còn lại là "hiếm", nhưng Trường hợp 2C.2 là đủ phổ biến để đáng được quan tâm đặc biệt.

Trường hợp 2C.2 xảy ra nếu và chỉ nếu e (w) = e (x) = e (y) + d, trong đó d≥ 1 và Y = Y₁ + R, R = ± bd = ± 102 , W = X + Y₁bd. (Ví dụ ở đầu ghi chú này có p = 3, b = 10, d = 2, Y₁ = 3.) Để tránh bị sai lệch trong trường hợp này, chúng ta làm cho hàm round (Wbd) = round(W + bd). Do đó, chúng ta hãy chỉ rõ thêm về hàm làm tròn để trong trường hợp không rõ ràng, hàm này luôn chọn giá trị mà phần nhân của phân số bp là chẵn, hoặc sao cho hàm này luôn chọn giá trị lẻ. Hai quy tắc này sẽ được gọi là round- to – even (làm tròn đến chẵn ) hoặc round-to-odd (làm tròn đến lẻ). Làm tròn ổn định là quy tắc làm tròn bất kỳ thỏa mãn z = round (z - be(z)-p ) nếu và chỉ nếu z= round (z - be(z)-p ) với mọi số thực có dấu phẩy động với |f (z) |# 1 / b. (Khi f (z) = 1 / b, chúng ta luôn có hàm round (z - be(z)-p) = z - be(z)-p ≠ z, nhưng chúng ta có thể có round(z - be(z)-p ) = z; nhận xét tương tự áp dụng khi f (a) = - 1 / b.) Lưu ý rằng làm tròn ổn định như được định nghĩa ở đây không nhất thiết phải làm tròn đến chẵn hoặc làm tròn thành lẻ; nó chỉ phải đưa ra lựa chọn nhất quán là chẵn hoặc lẻ cho mỗi tập hợp số thực có dấu phẩy động có số mũ và dấu cho trước. Để bảo toàn đồng dạng xy = - (yx) chúng ta nên quy định thêm rằng round (-x) = - round (x); nhưng thảo luận sau đây không yêu cầu quy định này.

Với việc làm tròn ổn định, chúng ta có thể chứng minh rằng hầu như tất cả trường hợp 2C.2 đều thỏa mãn w = w ':

Trường hợp 2C.2.a, R = +bd = +104. Khi đó x' ≤ x và w ' ≤ w theo tính đơn điệu của hàm làm tròn.

Trường hợp 2C.2.a.1, e(x') = e(x) -1. [Ví dụ: x= 100, y= 10.5, ư = 110, x'= 99.5] Khi đó wy là chính xác.

Trường hợp 2C.2.a.2, e(x') = e(x). W – Y = X' +S khi đó S = ± bd ± 104. Bởi vì X' = X – R – S, chúng ta có thể giả định rằng S=R và X'= X – bd.

Trường hợp 2C.2.a.2.a, W > bp+d-1 = 1013. Bởi vì Ư = round(W + bd), chúng ta có W' = round(X' + Y) = round (W -bd) = Ư, bằng cách làm tròn ổn định.

Trường hợp 2C. 2.a.2.b, Ư = bp+d-1 = 1013. Khi đó w≠w' là có thể, xem bên dưới.

Trường hợp 2C.2.b, R = -bd = -104thì x' ≥ x và w' ≥ w.

Trường hợp 2C.2.b.1, e (x') = e (x) + 1. [Ví dụ: x = 999, y=-11.5, w= 998, x'= 1000]. Khi đó X' = round (X + bd) = round (X + 104) = X + bd = X + 104 và W' = round (W + bd) = round (W + 104) = W vì round (W-bd) = round (W -104) =W.

Trường hợp 2C.2.b.2, e (x')= e (x). Sau đó, như trong Trường hợp 2C.2.a.2, chúng ta có thể giả sử rằng X'= X + bd =X + 104; và như trong Trường hợp 2C.2.b.1, điều này ngụ ý W '=W.

Nếu áp dụng quy tắc làm tròn ổn định cho Trường hợp 2B.2.b, chúng ta luôn có w'=w, vì W' = round(X' + Y) = round(W - bd-1) = round (W -104) = round (W + bd-1) = round (W + 104) = W.

Do đó, làm tròn ổn định đảm bảo rằng w'= w ngoại trừ ba trường hợp "hiếm" (và phủ định của chúng) có thể được đặc trưng hoàn toàn bằng các công thức sau: tuy nhiên sẽ có trường hợp xảy ra với w’ = w bất kể quy tắc làm tròn ổn định nào được chọn. Vì vậy, chúng tôi đã chứng minh các kết quả sau đây:

Trường hợp 2C.1.b.2, X = b2p-1  = 1019, Y = Y1bb-1+ bp-1 = Y1109 + 109, khi đó 0 <Y1< b, Y1 là số chẵn, và hàm làm tròn đến lẻ được sử dụng cho x'. [Ví dụ: x=100, y= 0.250, w = 100, x'= 99.7, w' = 99.9] Khi đó chúng ta có W = b2p-1  = 1019, X'= (bp-Y1-1).bp-1 = (1010 - Y1 - 1)109, W' = ( bp -1 )bd-1 = (1010 - 1)109, X' '= X'

Trường hợp 2C.2.a.2.b, [Ví dụ: x=169, y= -68.5, w = 100, x'= 168, w' = 99.5]. X= bp+d-1+Y1bd = 1013 + Y1104, Y = bd -Y1bd  = 104  - Y1104, 1≤d≤p <=> 1≤4<10 → Hợp lý, khi đó bp+d-1 ≤ Y1 ≤ bp-d <=> 105 ≤ Y1 ≤ 106, Y1 là số lẻ, quy tắc làm tròn đến chẵn sử dụng cho w. Sau đó W = bp+d-1 = 1015, X'= X-bd  = X - 104, W' = ( bp -b )bd-1 = (1010 - 10)103, X''= X'

Trường hợp 2D.2.a, [Ví dụ: x=936, y= 68.5, w = 1000, x'= 931, w' = 999]. Khi đó chúng ta có X= (bp-Y1+k)bd  = (104 - Y1 + k)104, Y = Y1bd+bd = 104  + Y1104 , 1≤d<p <=> 1≤4<10 → Hợp lý, khi đó 0 ≤ k ≤ b <=> 0 ≤ k ≤5, bp-d-1 ≤ Y1 ≤ bp-d <=> 105 ≤ Y1 ≤ 106, Y1 là số chẵn, Y1 > k, và làm tròn đến lẻ sử dụng cho x'. Sau đó W = bp+d = 1014, X'= (bp-Y1-1).bd = (1010 - Y1 - 1)104, W' = ( bp -1 )bd = (1010 - 1)104, X''= X'

Trong mỗi trường hợp trên, x'' =x' nên sự chênh lệch bị loại bỏ. Tuy nhiên lưu ý rằng sẽ có trường hợp xảy ra với w≠w' bất kể quy tắc làm tròn ổn định nào được chọn. Vì vậy, chúng tôi đã chứng minh các kết quả sau đây.

Định lý 1. Nếu sử dụng làm tròn ổn định, (((x y)y) y)y = (x y)y với mọi số thực có dấu phẩy động x và y như các phép toán đã được xác định

Định lý 2. Tồn tại x và y sao cho (x y)y y ≠ x y bất kể những quy tắc làm tròn nào được áp dụng, khi cơ số là chẵn.