Foutenanalyse voor arctan

Johan Vervloet

27 februari 2003

Samenvatting

Deze tekst beschrijft een foutenanalyse voor de evaluatie van de arctan functie in floating-pointaritmetiek.

Inhoudsopgave

1 De verschillende gevallen

De evaluatie van de boogtangens in een positief argument z valt uiteen in vier gevallen [1] :

$$0 \leq z \leq 2-\sqrt{3} \quad z' = z \quad \arctan(z) = \arctan(z') \tag{1}$$

$$2-\sqrt{3} \leq z \leq 1 \quad z' = \frac{z\sqrt{3}-1}{\sqrt{3}+z} \quad \arctan(z) = \arctan(z') + \frac{\pi}{6} \tag{2}$$

$$1 \leq z \leq 2+\sqrt{3} \quad z' = \frac{\sqrt{3}-z}{z\sqrt{3}+1} \quad \arctan(z) = \frac{\pi}{3} - \arctan(z') \tag{3}$$

$$2+\sqrt{3} \leq z \quad z' = \frac{1}{z} \quad \arctan(z) = \frac{\pi}{2} - \arctan(z') \tag{4}$$

- Als $z \leq 0$, maken we gebruik van $\arctan z = -\arctan -z$
- De werkwijze hierboven reduceert het argument naar $\left[\sqrt{3}-2, 2-\sqrt{3}\right]$. Dus eventueel gaan we nog eens voor $\arctan z = -\arctan -z$.
- In het kader van de propagatie van de rounding error in de 'post-argumentreductie', is het nuttig om weten dat $\arctan\left[0, 2-\sqrt{3}\right] = \left[0, \frac{\pi}{12}\right]$.

2 'Gereedschap'

Wanneer we de evaluatie van arctan opsplitsen zoals in [2], krijgen we deze functies a, g en p_z :

$$a(x) = \begin{cases} x & \text{als } 0 \leq x < 2-\sqrt{3} \\ \frac{x\sqrt{3}-1}{\sqrt{3}+x} & \text{als } 2-\sqrt{3} \leq x < 1 \\ \frac{\sqrt{3}-x}{x\sqrt{3}+1} & \text{als } 1 \leq x < 2+\sqrt{3} \\ \frac{1}{x} & \text{als } 2+\sqrt{3} \leq x \end{cases} \tag{5}$$

$$g(x) = \arctan(x) \tag{6}$$

$$p_z(x) = \begin{cases} x & \text{als } 0 \leq z < 2-\sqrt{3} \\ x + \frac{\pi}{6} & \text{als } 2-\sqrt{3} \leq z < 1 \\ \frac{\pi}{3} - x & \text{als } 1 \leq z < 2+\sqrt{3} \\ \frac{\pi}{2} - x & \text{als } 2+\sqrt{3} \leq z \end{cases} \tag{7}$$

Passen we de resultaten toe uit [2], dan hebben we[1] :

$$g(x(1+\delta)) = g(x)(1+\varepsilon) \quad \text{met } |\varepsilon| < 2|\delta| \tag{8}$$

$$p_z(x(1+\delta)) = p_z(x)(1+\varepsilon) \quad \text{met } \begin{cases} |\varepsilon| \leq |\delta| & \text{als } 0 \leq z < 2-\sqrt{3} \\ |\varepsilon| \leq \frac{|\delta|}{3} & \text{als } 2-\sqrt{3} \leq z < 1 \\ |\varepsilon| \leq \frac{|\delta|}{5} & \text{als } 1 \leq z < 2+\sqrt{3} \\ |\varepsilon| \leq \frac{|\delta|}{7} & \text{als } 2+\sqrt{3} \leq z \end{cases} \tag{9}$$

3 Foutenanalyse

3.1 Geval (1)

Wanneer ons argument in $\left[0, 2-\sqrt{3}\right[$ valt, dan wordt geen argumentreductie toegepast. De enige fout waarmee we moeten afrekenen, is die van de kettingbreukontwikkeling, en die hebben we volledig onder controle.

3.2 Geval (4)

In dit geval zal de argumentreductie (4) een relatieve fout δ_a introduceren, die gemakkelijk begrensd kan worden tot een halve ulp.

Wanneer we uitdrukkingen (8) en (9) hierop loslaten, vinden we

$$\tilde{\arctan}\, x = (\tilde{p} \circ \tilde{g} \circ \tilde{a})(x) \tag{10}$$

$$= (p \circ g \circ a)(x)(1+\varepsilon) \tag{11}$$

$$= (\arctan x)(1+\varepsilon) \tag{12}$$

waarbij

$$\varepsilon = \varepsilon_2 + \delta_p + \varepsilon_2\delta_p \tag{13}$$

$$|\varepsilon_2| \leq \frac{1}{3}|\varepsilon_1 + \delta_1 + \varepsilon_1\delta_1| \tag{14}$$

$$|\varepsilon_1| \leq 2|\delta_g| \tag{15}$$

en δ_a, δ_g en δ_p de relatieve fouten zijn, geïntroduceerd door a, g en p.

[1] aangenomen dat $|\varepsilon| \leq \frac{2-\sqrt{2}}{2}$

Zij nu $\delta = \max(|\delta_a|, |\delta_g|, |\delta_p|)$, dan wordt $|\varepsilon|$ begrensd door

$$|\varepsilon| \leq \frac{9}{7}\delta + \frac{3}{7}\delta^2 + \frac{1}{7}\delta^3 \tag{16}$$

Een beetje rekenwerk leert dat voor $\delta \leq 1$ (wat meestal het geval is), dit zich vereenvoudigt tot

$$|\varepsilon| \leq 2\delta \tag{17}$$

3.3 Geval (2)

3.3.1 Niet te schatten

Moeilijker wordt het in geval (2). Een floating-pointevaluatie van de teller van het gereduceerde argument, $z\sqrt{3} - 1$, levert namelijk

$$z\sqrt{3}(1 + \delta_1) - 1 = \left(z\sqrt{3} - 1\right)\left(1 + \frac{z\sqrt{3}\delta_1}{z\sqrt{3} - 1}\right) \tag{18}$$

De waarde van δ_1 is de relatieve fout die optrad bij het berekenen van $z\sqrt{3}$. Wanneer z nu in de buurt komt van $\frac{\sqrt{3}}{3}$, dan zal de gepropageerde fout $\frac{z\sqrt{3}\delta_1}{z\sqrt{3}-1}$ zeer grote proporties aannemen, die op het eerste zicht niet meer zullen verdwijnen in (8) en (9).

De relatieve fout in de noemer kan wel begrensd worden, omdat z in het interval $\left[2 - \sqrt{3}, 1\right[$ ligt.

3.3.2 Maar blijkbaar ook niet erg

Het feit dat de relatieve fout voor de argumentreductie oneindig groot kan worden, is gelukkig niet rampzalig voor de implementatie. Veel wordt goedgemaakt door de $p_z(x)$ functie, die de reductie terug moet compenseren.

$$p_z(x) = x + \frac{\pi}{6} \tag{19}$$

Als er nu op x een relatieve fout δ zit, dan zal deze als volgt propageren :

$$p_z(x(1 + \delta)) = \left(x + \frac{\pi}{6}\right)\left(1 + \frac{\delta x}{x + \frac{\pi}{6}}\right) \tag{20}$$

We gaan er voor de eenvoud even onterecht van uit dat $\frac{\pi}{6}$ exact voorstelbaar is.

Het argument waarop p_z zal losgelaten worden, ziet er ongeveer als volgt uit :

$$x = \arctan\left(\frac{z\sqrt{3} - 1}{\sqrt{3} + z}\right)\left(1 + \frac{z\sqrt{3}\delta_1}{z\sqrt{3} - 1}\right)(1 + \delta_2) \tag{21}$$

De fout δ_2 is de combinatie van de andere fouten, die we wel kunnen afschatten. Opnieuw voor de eenvoud gaan we er even van uit dat we deze zodanig klein kunnen maken dat we ze kunnen negeren.

En nu we toch allerlei louche vereenvoudigingen aan het maken zijn... Het probleem van de fout buiten proportie doet zich voor wanneer z in de buurt ligt van $\frac{\sqrt{3}}{3}$. In dit geval is het argument van arctan ongeveer 0, en dus kunnen we arctan benaderen door de identiteit. (Het kleine foutje dat we hier maken zal wellicht niet opwegen tegen onze afrondingsfout die oneindig benadert.)

Als we na al deze vereenvoudigingen de substituties

$$x = \frac{z\sqrt{3} - 1}{\sqrt{3} + z} \tag{22}$$

$$\delta = \frac{z\sqrt{3}\delta_1}{z\sqrt{3} - 1} \tag{23}$$

loslaten op uitdrukking (20), blijft van de verschikkelijk grote fout δ het volgende over :

$$\delta = \frac{z\sqrt{3}\delta_1}{z\sqrt{3} - 1 + \frac{\pi}{6}} \tag{24}$$

Aangezien

$$\lim_{z \to \frac{\sqrt{3}}{3}} \frac{z\sqrt{3}\delta_1}{z\sqrt{3} - 1 + \frac{\pi}{6}} = \frac{6}{\pi}\delta_1 \tag{25}$$

herleidt de fout in het punt $\frac{\sqrt{3}}{3}$ zich tot $\frac{6}{\pi}\delta_1$.

3.3.3 En hoe groot is de fout nu ?

Geen idee. De berekening van de exacte fout lijkt me behoorlijk omslachtig. Ik zou gefundeerd gokken dat een werkprecisie die 5 hoger ligt dan de gevraagde resultaatprecisie volstaat, maar 't zou tof zijn moest dit exact wiskundig bewezen kunnen worden natuurlijk.

3.4 Geval (3)

Voor geval (3) gaat een analoge redenering op als voor geval (2).

Referenties

[1] W.J. Cody and W. Waite. *Software manual for the elementary functions.* Prentice Hall, New Jersey, 1980.

[2] Johan Vervloet. Propagatie van afrondingsfouten door afleidbare functies. http://www.gewestpallieter.be/~johan/onderzoek/properr.pdf.