Propagatie van afrondingsfouten door afleidbare functies

Johan Vervloet

14 maart 2003

Samenvatting

Dit kort tekstje beschrijft hoe afrondingsfouten zich propageren doorheen afleidbare-functie-evaluaties.

1 Een typisch argumentreductieprobleem

Wanneer we in de computeraritmetiek een functie f willen evalueren in een argument z, zullen we vaak argumentreductie toepassen. We brengen het argument z terug tot een beperkt deel van het domein, waar de functie gemakkelijker of sneller geëvalueerd kan worden.

Om bijvoorbeeld de arctan te berekenen van een argument $z > 1$, zullen we in praktijk $z' = z^{-1}$ nemen, en dan $\arctan(z)$ berekenen als

$$\arctan(z) = \frac{\pi}{2} - \arctan(z') \tag{1}$$

In praktijk is de argumentreductie van arctan nog iets verfijnder, maar dat is niet relevant in deze tekst. Meer algemeen zal de evaluatie van $f(z)$ doorgaans gebeuren als

$$f(z) = (p_z \circ g \circ a)(z) \tag{2}$$

waar a de argumentreductie is, g meestal terug de oorspronkelijke functie, en p_z het ongedaan maken van de argumentreductie. Toegepast op ons voorbeeld geeft dit :

$$a(x) = \begin{cases} x & \text{als } |x| \leq 1 \\ x^{-1} & \text{als } |x| > 1 \end{cases} \tag{3}$$

$$g(x) = \arctan(x) \tag{4}$$

$$p_z(x) = \begin{cases} x - \frac{\pi}{2} & \text{als } z < -1 \\ -x & \text{als } -1 \leq z \leq 0 \\ x & \text{als } 0 \leq z \leq 1 \\ \frac{\pi}{2} - x & \text{als } 1 < z \end{cases} \tag{5}$$

In praktijk staan we echter voor het probleem dat de implementaties van a, g en p afrondingsfouten zullen introduceren :

$$\tilde{f}(z) = (\tilde{p}_z \circ \tilde{g} \circ \tilde{a})(z) \tag{6}$$
$$= p_z(g(a(z)(1+\varepsilon_a))(1+\varepsilon_g))(1+\varepsilon_p) \tag{7}$$

Toch zal na de hele berekening moeten gelden dat

$$\tilde{f}(z) = f(z)(1+\varepsilon) \tag{8}$$

waar $|\varepsilon| \leq \overline{\varepsilon}$. Hier is $\overline{\varepsilon}$ de maximaal toegelaten afrondingsfout.

We moeten dus aan de hand van $\overline{\varepsilon}$ een bovengrens opstellen voor $|\varepsilon_a|$, $|\varepsilon_e|$ en $|\varepsilon_p|$, waarvoor we een relatie nodig hebben tussen $|\varepsilon_a|$, $|\varepsilon_g|$, $|\varepsilon_p|$ en $|\varepsilon|$. Zo'n relatie is niet altijd even gemakkelijk te vinden, aangezien bijv. de fout ε_a verder zal propageren doorheen de evaluatie van g en p.

2 Strategie

Gegeven een afleidbare functie f, een argument x en een relatieve fout ε, zullen we een uitdrukking $\delta_f(\varepsilon, x)$ bepalen zodat

$$f(x(1+\varepsilon)) = f(x)(1+\delta_f(\varepsilon, x)) \tag{9}$$

Wanneer we dat dan toepassen op uitdrukking (7) krijgen we

$$\tilde{f}(z) = p_z(g(a(z)(1+\varepsilon_a))(1+\varepsilon_g))(1+\varepsilon_p) \tag{10}$$
$$= p_z((g \circ a)(z)(1+\delta_g(\varepsilon_a, a(z)))(1+\varepsilon_g))(1+\varepsilon_p) \tag{11}$$
$$= p_z((g \circ a)(z)(1+\delta'(\varepsilon_a, \varepsilon_g, z)))(1+\varepsilon_p) \tag{12}$$
$$= (p_z \circ g \circ a)(z)(1+\delta_p(\delta'(\varepsilon_a, \varepsilon_g, z), (g \circ a)(z)))(1+\varepsilon_p) \tag{13}$$
$$= (p_z \circ g \circ a)(z)(1+\delta''(\varepsilon_a, \varepsilon_g, \varepsilon_p, z)) \tag{14}$$

waarbij

$$\delta'(\varepsilon_a, \varepsilon_g, z) = (1+\delta_g(\varepsilon_a, a(z)))(1+\varepsilon_g) - 1 \tag{15}$$
$$\delta''(\varepsilon_a, \varepsilon_g, \varepsilon_p, z) = (1+\delta_p(\delta'(\varepsilon_a, \varepsilon_g, z), (g \circ a)(z)))(1+\varepsilon_p) - 1 \tag{16}$$

Hierbij is $\delta''(\varepsilon_a, \varepsilon_g, \varepsilon_p, z)$ een uitdrukking voor de totale relatieve fout, die we zullen kunnen begrenzen door $\overline{\varepsilon}$.

3 Een uitdrukking voor $\delta_f(\varepsilon, x)$

We zullen gebruik maken van deze (volgens mij bekende) stelling, waarvan me de naam even ontglipt :

Theorema 1. *Als de reële functie f afleidbaar is op $[x-|\varepsilon|, x+|\varepsilon|]$, dan geldt dat*

$$f(x+\varepsilon) = f(x) + f'(x+\xi)\varepsilon \tag{17}$$

waarbij $|\xi| \leq |\varepsilon|$.

Hieruit leiden we het volgende af :

Gevolg 2. *Als de reële functie f afleidbaar is op $x[1-|\varepsilon|, 1+|\varepsilon|]$ met $|\varepsilon| \leq 1$, dan geldt dat*

$$f(x(1+\varepsilon)) = f(x)(1+\delta_f(\varepsilon, x)) \tag{18}$$

waarbij

$$|\delta_f(\varepsilon, x)| \leq \max_{|\eta|\leq|\varepsilon|} \frac{|f'(x(1+\eta))x\varepsilon|}{|f(x)|} \tag{19}$$

Bewijs. Als we $f(x(1+\varepsilon))$ herschrijven als $f(x+x\varepsilon)$, volgt uit theorema 1 dat

$$f(x(1+\varepsilon)) = f(x) + f'(x+\xi)x\varepsilon \tag{20}$$

met $|\xi| \leq |x\varepsilon|$. Een beetje herschrijfwerk levert nu

$$f(x(1+\varepsilon)) = f(x)\left(1 + \frac{f'(x(1+\eta))x\varepsilon}{f(x)}\right) \tag{21}$$

waarbij $|\eta| \leq |\varepsilon|$. Als we nu

$$f(x(1+\varepsilon)) = f(x)(1+\delta_f(\varepsilon, x)) \tag{22}$$

stellen, dan zal

$$|\delta_f(\varepsilon, x)| \leq \max_{|\eta|\leq|\varepsilon|} \frac{|f'(x(1+\eta))x\varepsilon|}{|f(x)|} \tag{23}$$

□

4 Toepassing op het boogtangensvoorbeeld

Ik geloof dat ik in deze sectie wel wat foutjes scoor :-).

4.1 In theorie

Laat ons eens kijken wat het effect is als we de boogtangens loslaten op een gereduceerd argument. (Het ongedaan maken van de reductie laten we nog even buiten beschouwing).

We nemen bij wijze van voorbeeld een argument $z > 1$ (kwestie van een interessante reductie te hebben), en we willen een $\delta = \delta_g(\varepsilon_a, a(z))$ vinden zodat

$$g(a(z)(1+\varepsilon_a)) = (g \circ a)(z)(1+\delta) \tag{24}$$

De functies $a(z)$ en $g(z)$ zijn hierbij gegeven in (3) en (4). Aangezien

$$g'(x) = \arctan'(x) = \frac{1}{1+x^2} \tag{25}$$

geeft gevolg 2 nu deze bovengrens voor δ :

$$|\delta| \leq \max_{|\eta|\leq|\varepsilon_a|} \left| \frac{\frac{\frac{1}{x}\varepsilon_a}{1+\left(\frac{1}{x}(1+\eta)\right)^2}}{\arctan\frac{1}{x}} \right| \tag{26}$$

$$= \max_{|\eta|\leq|\varepsilon_a|} \left| \frac{\varepsilon_a}{\left(\arctan\frac{1}{x}\right)\left(x+\frac{1}{x}(1+\eta)^2\right)} \right| \tag{27}$$

Aangezien $x > 1$ en $0 \leq |\eta| \leq |\varepsilon_a| \leq 1$, zal het rechterlid maximaal zijn voor $\eta = -\varepsilon_a$:

$$|\delta| \leq \left| \frac{\varepsilon_a}{\left(\arctan\frac{1}{x}\right)\left(x+\frac{1}{x}(1-\varepsilon_a)^2\right)} \right| \tag{28}$$

En omdat $0 \leq |\varepsilon_a| \leq 1$, zal ook $0 \leq (1-|\varepsilon_a|)^2 \leq 1$:

$$|\delta| \leq \left| \frac{1}{\left(\arctan\frac{1}{x}\right)\left(x+\frac{1}{x}\right)} \right| \frac{|\varepsilon_a|}{(1-|\varepsilon_a|)^2} \tag{29}$$

Aangezien de linkerterm stijgend is, en lijkt te convergeren naar 1, blijft er over :

$$|\delta| \leq \frac{|\varepsilon_a|}{(1-|\varepsilon_a|)^2} \tag{30}$$

$$= \sum_{n=1}^{\infty} n|\varepsilon_a|^n \tag{31}$$

Een beetje rekenwerk leert dat wanneer $|\varepsilon_a| \leq \frac{2-\sqrt{2}}{2}$ er zal gelden dat $\delta \leq 2\varepsilon_a$. Aan deze voorwaarde is voldaan als de precisie minstens drie is.

We weten nu dat na de argumentreductie de totale relatieve fout begrensd wordt door $(1+\delta)(1+\varepsilon_g) \leq (1+2\varepsilon_a)(1+\varepsilon_g)$. De waarde ε_g is de combinatie van truncatie- en afrondingsfout bij de kettingbreukontwikkeling. Op een gelijkaardige manier moet het effect van deze fout bekeken worden op de 'postargumentreductie' p. (Gelukkig is dit een eenvoudige functie met een nog eenvoudigere afgeleide, dus vermoedelijk is dit niet zo veel werk.)

4.2 Controle in praktijk

Het zou natuurlijk tof zijn moesten een aantal experimentjes deze resultaten bevestigen.