Siła rachunków interwałowych

Marek W. GUTOWSKI

Sporo już czasu minęło od przedstawienia Czytelnikom podstaw rachunków interwałowych (*Delta* 9 (232), str. 4–6, 1993), zaczniemy więc od ich przypomnienia. Zostały one wymyślone przede wszystkim dlatego, że – jak wiadomo – nie do końca można polegać na wynikach obliczeń wykonywanych za pomocą komputera. Dzieje się tak z kilku powodów, z których wymienimy dwa:

- każdy współczesny komputer posługuje się skończonymi przedstawieniami dwójkowymi potrzebnych liczb. Oznacza to, że komputer „nie zna" takich liczb, jak $\sqrt{2}$ czy π (bo, jako liczby niewymierne, mają przedstawienia nieskończonej długości), ale także ma kłopoty z pospolitymi ułamkami, takimi jak $\frac{1}{10}$ czy $\frac{1}{3}$. Nie jest też w stanie poprawnie przetwarzać liczb zbyt dużych lub zbyt bliskich zera.

  Tak naprawdę, ten cudowny wynalazek „zna" jedynie **niewielki podzbiór liczb wymiernych o skończonej liczbie elementów**.

- nawet jeśli liczby zostaną odwzorowane zupełnie dokładnie, to już wyniki działań arytmetycznych, zwłaszcza dzielenia, ale niekiedy także zwykłego dodawania, będą musiały być jakoś zaokrąglone.

Nietrudno sobie wyobrazić, że po długim cyklu obliczeń końcowy wynik może znacznie odbiegać od prawdziwego. Kiedy w grę wchodzi bezpieczeństwo ludzi albo sprzęt o ogromnej wartości – wyniki „mniej więcej dobre" nie mogą być akceptowane. Nawet jeśli nie są one dokładne, to przynajmniej powinniśmy wiedzieć, w jakim przedziale mieszczą się prawdziwe wartości.

Konieczność operowania wielkościami znanymi niedokładnie zainspirowała m.in. polskiego matematyka Mieczysława Warmusa, który w r. 1956 opublikował swoje pierwsze prace na ten temat. Za ojca rachunków interwałowych uznajemy obecnie Raymonda E. Moore'a, który w swojej rozprawie doktorskiej (1962), a później w formie książkowej podsumował swoje wcześniejsze badania w tej dziedzinie. Nic nowego pod słońcem, powiedzą ci, którzy wiedzą, że o wiele wcześniej Archimedes z Syrakuz pokazał, że $3\frac{10}{71} < \pi < 3\frac{10}{70}$, oraz podał sposób ulepszania tego oszacowania.

Nie byłoby jednak tego artykułu, gdyby rola rachunków przedziałowych ograniczała się do gwarantowanych oszacowań wyników obliczeń komputerowych.

Podstawowe przepisy

Będziemy odtąd oznaczać liczby znane niedokładnie tłustym drukiem: $\boldsymbol{x} := [\underline{x}, \overline{x}]$, gdzie $\underline{x}$ oraz $\overline{x}$ są liczbami rzeczywistymi, takimi że zachodzi podwójna nierówność $\underline{x} \leqslant x \leqslant \overline{x}$. Obiekt $\boldsymbol{x}$ jest zatem odcinkiem, który z całą pewnością zawiera naszą bliżej nieznaną liczbę x.

Zauważmy przy okazji, że zwyczajne liczby rzeczywiste można utożsamić z interwałami *cienkimi*, tj. mającymi postać: $x = \boldsymbol{x} = [x, x]$.

Na początek potrzebne nam będą przepisy na wykonywanie czterech podstawowych działań na takich obiektach. Oczywistym wymaganiem jest to, aby wynik takiego działania, sam będący przedziałem, *z całą pewnością* zawierał wynik prawdziwy.

Określamy zatem

$$\boldsymbol{x} + \boldsymbol{y} = [\underline{x} + \underline{y}, \overline{x} + \overline{y}], \tag{1}$$

$$\boldsymbol{x} - \boldsymbol{y} = [\underline{x} - \overline{y}, \overline{x} - \underline{y}], \tag{2}$$

$$\boldsymbol{x} \cdot \boldsymbol{y} = [\min(\underline{x} \cdot \underline{y},\ \underline{x} \cdot \overline{y},\ \overline{x} \cdot \underline{y},\ \overline{x} \cdot \overline{y}),\ \max(\underline{x} \cdot \underline{y},\ \underline{x} \cdot \overline{y}, \overline{x} \cdot \underline{y},\ \overline{x} \cdot \overline{y})]. \tag{3}$$

Przepis na dzielenie otrzymujemy, zamieniając w przepisie na mnożenie znak „·" na znak dzielenia „/" oraz pamiętając, żeby interwał $\boldsymbol{y}$, ten w mianowniku, nie zawierał zera. Można się przekonać, że tak otrzymane wyniki są *ciasne* (dokładne?), w tym sensie, że zawierają wszystkie możliwe wyniki obliczeń, dla dowolnych $x \in \boldsymbol{x}$ oraz $y \in \boldsymbol{y}$, a jednocześnie *tylko te* wyniki – nic tu nie da się ulepszyć.

Pracując z komputerem powinniśmy do powyższych przepisów dodać jeszcze jedną, żelazną regułę: *każdy* wynik pośredni, oraz końcowy, trzeba koniecznie *zaokrąglać na zewnątrz* – to znaczy lewy kraniec wyniku w dół, a prawy – w górę. Jedynie tym sposobem uzyskamy gwarancję poprawnego wyniku, niezależnie od wspomnianych wcześniej niedostatków arytmetyki komputera.

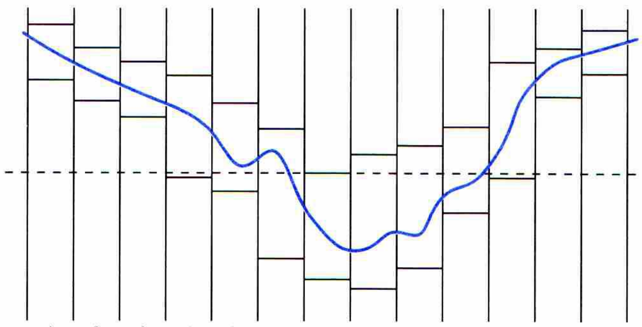
Niewątpliwie Czytelnik będzie zaskoczony faktem, że choć dodawanie i mnożenie są, jak należy, łączne i przemienne, to jednak prawo rozdzielności mnożenia względem dodawania nie zawsze jest spełnione. Zawsze natomiast mamy:

$$\boldsymbol{x} \cdot (\boldsymbol{y} + \boldsymbol{z}) \subseteq \boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{x} \cdot \boldsymbol{z}.$$

Wyposażeni w podstawowe umiejętności będziemy odtąd potrafili obliczać wartości wszelkich wyrażeń wymiernych. Jak jednak pokazuje ostatni przykład, nasz wynik końcowy ma prawo, przynajmniej od czasu do czasu, być „zbyt szeroki". Możemy się spodziewać tego niepożądanego efektu zawsze wtedy, gdy w obliczanym wyrażeniu przynajmniej jedna zmienna pojawia się więcej niż jeden raz albo w potędze innej niż pierwsza.

Funkcje interwałowe

Wprawdzie cztery działania arytmetyczne powinny nam wystarczyć do wielu zastosowań, ale tak naprawdę to chcielibyśmy czegoś więcej. Chcielibyśmy umieć rzetelnie odpowiadać na pytania w rodzaju: *jeśli x jest dowolnie wybraną liczbą z pewnego ustalonego przedziału $\boldsymbol{x}$, to w jakim przedziale jest x^2? albo* $\log x$*?*. Okazuje się, że konstrukcja funkcji, które byłyby ścisłymi odpowiednikami zwykłych funkcji rzeczywistych, tyle że określonymi na zbiorze interwałów (funkcje takie nazywane są po angielsku *range functions*, czyli funkcje-zakresy), bywa niełatwym zadaniem, nawet dla wielomianów. Niełatwym, gdyż opisujące je wyrażenia niekoniecznie dadzą się przedstawić jako funkcje samych tylko krańców interwałowych argumentów. Jedynym wyjątkiem są funkcje monotoniczne (czyli wszędzie rosnące albo wszędzie malejące) względem każdej zmiennej.



Fragmenty obwoluty interwałowej funkcji narysowanej kolorem. Choć nie mamy pewności, jak przebiega jej wykres, to jednak możemy twierdzić, że na pewno nie ma ona minimum w obszarach z numerami 1, 2, 3, 12 i 13. Linia przerywana przebiega po górnej granicy obwoluty wyliczonej dla obszaru z numerem 7.

Z konieczności musimy się więc posługiwać innymi funkcjami, łatwiejszymi do skonstruowania i obliczeń, które nazwiemy obwolutami interwałowymi (ang. *interval enclosures*) funkcji-zakresów. W tym jednakże przypadku musimy się pogodzić z nieuchronnymi nadmiarami, wynikami o zawyżonej szerokości. Nawet w tak trudnych warunkach bardzo często uda nam się uzyskać wartościowe wyniki, choćby częściowe, jak to można prześledzić na rysunku. Widzimy tam właśnie taką niedoskonałą obwolutę, dzięki której możemy się podjąć zadania wykrycia, gdzie oryginalna funkcja osiąga wartość najmniejszą.

Pomocne nam będą przede wszystkim interwałowe odpowiedniki zwykłych funkcji, które noszą nazwę *monotonicznie inkluzywnych*. Są to obwoluty interwałowe o niesłychanie cennej właściwości. Mówiąc obrazowo, generują one tym mniejsze nadmiary wyników, im węższe są ich argumenty. W granicznym przypadku, kiedy $\underline{x} = \overline{x}$ (interwał zdegenerowany, czyli punktowy, innymi słowy najzwyklejsza liczba rzeczywista), nadmiar ten jest dokładnie równy zeru (w komputerze: z dokładnością do zaokrągleń). Dobra wiadomość jest taka, że wyrażenia obliczane według podanych przepisów są dobrymi, monotonicznie inkluzywnymi obwolutami interwałowymi wyrażeń i funkcji rzeczywistych, które zastępują. Oj, przydałaby się obwoluta o takich właściwościach do rozwiązania zadania zilustrowanego rysunkiem. Łatwo można by wtedy zawężać granice, w których „siedzi" nasze rozwiązanie, gdyż obszar opisany numerami od 4 do 11 wydaje się niewiele mniejszy niż ten, od którego rozpoczęliśmy.

Algorytmy interwałowe

Przykład rysunkowy jest typowy dla przedziałowych metod rozwiązywania rozmaitych problemów. Ich charakterystyczną cechą jest podział podejrzanego obszaru na mniejsze części, z których te na pewno „niedobre" zostają sukcesywnie eliminowane. Te, o których nie potrafimy powiedzieć, czy są dobre czy nie, muszą być rozdrobnione jeszcze bardziej, i tak aż do skutku. A zatem, jeśli rozwiązanie w ogóle istnieje, to musi być elementem zbioru obszarów jeszcze nie przebadanych, nigdzie indziej.

Po co to wszystko fizykom albo inżynierom?

Zacznijmy od potrzeb „inżynierskich". Konstruktor nowego urządzenia musi się liczyć z tym, że elementy, z których ono ma się składać, produkowane są z określoną dokładnością, albo jak kto woli – tolerancją. Przekonanie się, czy telewizor zbudowany z podzespołów o znanych rozrzutach parametrów rzeczywiście będzie działał, wymaga wykonania tzw. analizy wrażliwości, jeszcze na etapie projektowania. Analiza interwałowa, w swojej najprostszej wersji, nadaje się tutaj znakomicie. Otrzymamy wynik obejmujący każdy, nawet najgorszy możliwy przypadek. Jeśli nasza konstrukcja przejdzie ten test, to wadliwy wyrób końcowy może być jedynie efektem wadliwych podzespołów (spoza zakresu tolerancji) lub innych uszkodzeń, których zwykle gwarancja nie obejmuje, np. spowodowanych przez powódź. Standardowa statystyczna analiza wrażliwości bada raczej „przypadek przeciętny", co dopuszcza możliwość, że wyrób złożony wyłącznie z dobrych elementów nie zadziała jako całość.

A co dla fizyków?

Jak dotąd – jeszcze niewiele. Pierwsze zastosowania rachunków interwałowych pojawiły się – gdzieżby indziej? – w fizyce doświadczalnej. Przydały się one w opisie sił działających pomiędzy częściami aparatury do precyzyjnego wyznaczania stałej grawitacji. Chodziło o to, że niektóre części były wykonane z mosiądzu, który jest stopem nie do końca jednorodnym. W obrębie kilkudziesięciokilogramowego mosiężnego walca należy się liczyć z lokalnymi odstępstwami gęstości od wartości średniej. Z tego powodu dokładne wyliczenie sił grawitacyjnych, w tym także ich kierunku, wytwarzanych przez taki obiekt, nie może ograniczyć się do wykonania kilku całkowań. Rachunki interwałowe pozwoliły ponadto uwzględnić możliwe odstępstwa od idealnego kształtu walca, spowodowane obróbką mechaniczną.

Co dalej?

Poprzedni przykład może wydać się nieco egzotyczny. Porozmawiajmy zatem o zastosowaniach, które, przynajmniej potencjalnie, są bardziej uniwersalne.

Częsta w praktyce badawczej jest czynność zwana gwarowo *fitowaniem* danych. Chodzi o odgadywanie wartości parametrów fizycznych, których nie można zmierzyć bezpośrednio, ale których wartości można wywnioskować z innych obserwacji. Możemy np., nie posiadając omomierza, próbować „zmierzyć" opór elektryczny poprzez pomiar napięcia na próbce podczas przepływu przez nią prądu elektrycznego o znanym natężeniu. Pojedynczy pomiar może dać nam wynik obarczony znaczną niepewnością. Powtarzamy więc doświadczenie kilkakrotnie, zmieniając za każdym razem natężenie prądu. Otrzymane wyniki możemy przedstawić na wykresie. W przypadku przewodnika metalicznego prostokąty niepewności powinny układać się na linii prostej, przechodzącej przez początek układu współrzędnych. Możemy, oczywiście, znaleźć szukane nachylenie nieznanej linii prostej łatwą metodą „graficzną", poprzez odpowiednie przyłożenie linijki. Jednakże liczba położeń, przy których linia prosta przetnie wszystkie prostokąty niepewności, będzie zwykle spora. A to oznacza, że szukany opór można wyznaczyć jedynie z ograniczoną dokładnością. Trudno się zdecydować, które położenie jest najlepsze. Rutynowym trikiem jest użycie metody najmniejszych kwadratów, która pozwala na wyliczenie szukanego parametru i jego niepewności.

Naiwne podejście interwałowe sugeruje takie obracanie linijki wokół punktu $(I = 0, U = 0)$, aż znajdziemy dwa skrajne położenia: w jednym z nich wszystkie prostokąty niepewności znajdą się powyżej linijki, w drugim – wszystkie będą poniżej. Ale czy to jest najlepsze rozwiązanie? Mamy lepszy sposób: rozpoczynając od położenia, w którym linia prosta przecina wszystkie prostokąty, obracamy linijkę najpierw w jedną, a potem w drugą stronę, aż do momentu, kiedy *nie wszystkie* prostokąty będą miały punkty wspólne z próbną prostą. I to są poprawne granice poszukiwanego rozwiązania. Widać, że „naciąganie" wyników pomiarów, poprzez rozmyślne zaniżanie niepewności pomiarowych, może się skończyć bardzo nieprzyjemnie: może się okazać, że szukana prosta zwyczajnie nie istnieje.

Już w przypadku jednej niewiadomej (współczynnik kierunkowy prostej) widzimy, że otrzymany przedział zawiera w sobie poszukiwany wynik, ale także wiele innych liczb, które *nie* są rozwiązaniami naszego problemu. Przy dwóch niewiadomych wynikiem będzie fragment płaszczyzny o kształcie prostokąta, przy trzech – fragment przestrzeni trójwymiarowej o kształcie prostopadłościanu, itd. Ważnym zadaniem analizy interwałowej jest poszukiwanie takich sposobów rozwiązania, aby te obszary były możliwie małe. Obszary, których rozmiarów nie da się już dalej zmniejszyć bez jednoczesnej straty zawartych w nich rozwiązań, nazywamy naukowo *powłokami interwałowymi* zbiorów rozwiązań.

Rozmiary powłoki interwałowej są wobec tego ściśle, bez żadnych przybliżeń, związane z niepewnościami danych wejściowych, i w sposób jednoznaczny ograniczają od góry niepewności wyników.

Siła

Siła metod interwałowych przejawia się w kilku miejscach. Po pierwsze, w każdym kroku prawie dowolnego algorytmu przedziałowego przetworzeniu ulega nie jedna, dwie czy n liczb, lecz od razu nieskończona ich ilość – podobnie jak wtedy, gdy „ręcznie" rozwiązujemy układ równań. Po drugie, jak dowiódł R.E. Moore, można wskazać przypadki, w których wraz z powłoką zbioru rozwiązań jednocześnie otrzymujemy gwarancję, że poszukiwane rozwiązanie znajduje się na pewno w tym zbiorze i jest w nim jedyne. Inne metody mogą nas zwodzić, wskazując jedynie kandydatów na rozwiązania, niekoniecznie prawdziwe rozwiązania.

Niewątpliwym przejawem siły metod przedziałowych jest możliwość rozwiązania problemów, dla których – przynajmniej dziś – nie znamy innych metod.

Częściej pewnie spotkamy się z innym, pozornie destruktywnym, przejawem mocy algorytmów interwałowych. W prowadzonych przez autora badaniach absorpcji mikrofal przez bardzo cienkie druty żelazne okazało się, że przypuszczenie, iż – jak się wszystkim wydawało – obserwujemy zwykły rezonans ferromagnetyczny, jest błędne. Zbiór poszukiwanych parametrów opisujących to zjawisko okazał się pusty, mimo drastycznego zawyżenia niepewności pomiarowych. Wynik ten, choć negatywny, okazał się nie do podważenia, właśnie dlatego, że został uzyskany metodami interwałowymi. Trzeba było poszukać innego wyjaśnienia, co się, na szczęście, udało.

Ciąg dalszy nastąpi?

Doceniając znaczenie rachunków, które produkują zawsze niezawodne wyniki, amerykańska rządowa agencja DARPA, ta sama, dzięki której powstał internet, przydzieliła w lipcu 2003 r. grant o wartości blisko 50 mln dolarów na kontynuację prac nad przyszłym komputerem. Będzie on 50 razy szybszy niż najszybsze maszyny współczesne, a obliczenia będzie wykonywać zgodnie z regułami arytmetyki interwałowej. Firma *Sun*, beneficjent tego grantu, już od lat oferuje kompilatory (fortran, C), które „znają się" na rachunkach przedziałowych. W niektórych krajach metody interwałowe są wykładane studentom wyższych lat studiów. Warto, abyśmy i my zapoznali się z tym potężnym narzędziem.