5.3. Reprezentacje zmiennopozycyjne

W przypadku reprezentacji stałopozycyjnej słowo binarne reprezentujące liczbę nie zawierało informacji o pozycji przecinka. Pozycja ta była ustalona zewnętrznie przez parametr *m* określający współczynnik skalujący. Taki sposób reprezentacji liczb wiąże się ze stosunkowo wąskim zakresem reprezentowanych liczb przy użyciu słowa binarnego o ustalonej długości. W wielu wypadkach wygodniej jest operować większym zakresem liczb kosztem obniżenia dokładności ich reprezentacji. W tym celu stosuje się tzw. reprezentacje zmiennopozycyjne zwane również reprezentacjami zmiennoprzecinkowymi, gdzie w przeciwieństwie do reprezentacji stałopozycyjnych informacja o pozycji przecinka jest przechowywana w samym słowie binarnym. Kodowanie zmiennopozycyjne opiera się na następującym twierdzeniu:

Twierdzenie 5.3.1. Dla dowolnych liczb rzeczywistych $p>1$ i $x\neq 0$ istnieje dokładnie jedna liczba całkowita r spełniająca nierówność:

(5.3.2) $$\frac{1}{p} \leq \frac{|x|}{p^r} < 1.$$

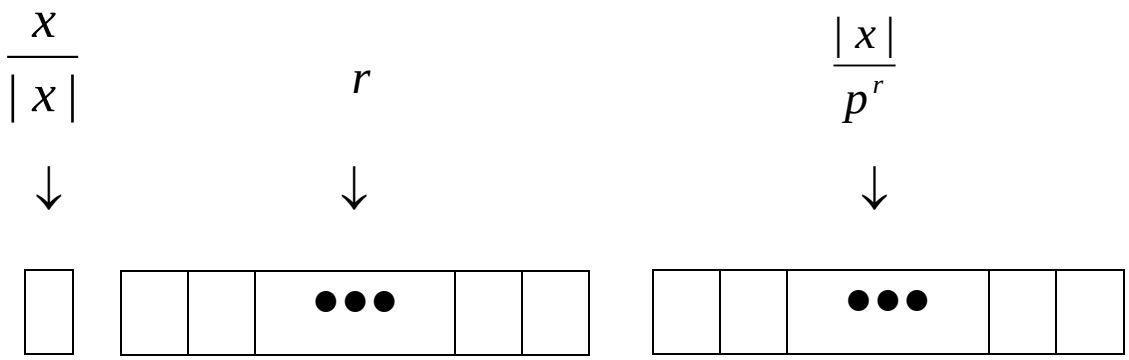
Z twierdzenia 5.3.1 wynika, że przy ustalonej liczbie rzeczywistej $p>1$, dowolną liczbę $x \in R\setminus\{0\}$ można jednoznacznie przedstawić w postaci:

(5.3.3) $$x = \frac{x}{|x|} \cdot \frac{|x|}{p^r} \cdot p^r,$$

gdzie r jest jedyną liczbą całkowitą spełniającą nierówności (5.3.2). Liczbę całkowitą r oraz liczbę dodatnią $\frac{|x|}{p^r}$ nazywa się odpowiednio cechą i mantysą liczby x przy podstawie p. Załóżmy, np. że $p=10$. Z (5.3.2) wynika, że liczba $\frac{|x|}{p^r}$ jest liczbą ułamkową o części całkowitej 0 i pierwszej cyfrze po przecinku różnej od 0. Liczba r jest całkowita i pokazuje o ile w reprezentacji dziesiętnej liczby $\frac{|x|}{p^r}$ należy przesunąć przecinek w lewo bądź w prawo, w zależności od znaku liczby r. Zatem liczba r wskazuje położenie przecinka. Czynnik $\frac{x}{|x|}$ jest równy 1, gdy $x>0$ i -1, gdy $x<0$, a więc wyraża znak liczby x. Powyższe uwagi odnoszą się do dowolnego systemu pozycyjnego przy podstawie $p \geq 2$.

Z uwagi na równość (5.3.3) liczbę x można zakodować z zadaną dokładnością wyrażoną przez ilość cyfr mantysy za

pomocą trzech słów binarnych kodujących każdy z czynników iloczynu (5.3.3).



1 bit n_1 bitów n_2 bitów

Dla zakodowania znaku liczby x wystarczy słowo jednobitowe, gdzie standardowo bit ”0” oznacza liczbę dodatnią, zaś bit ”1” oznacza liczbę ujemną. Mantysa $\frac{|x|}{p^r}$ liczby x jest zawsze liczbą z przedziału [1/p;1) i można ją reprezentować z zadaną dokładnością za pomocą kodu stałopozycyjnego NKB z odpowiednio dobranym parametrem m. Wreszcie liczba całkowita r może być reprezentowana przy użyciu dowolnego kodu stałopozycyjnego uwzględniającego liczby ujemne.

W praktyce trzy słowa binarne łączy się w jedno $1+n_1+n_2$-bitowe słowo, gdzie domyślnie pierwszy bit odpowiada za znak, kolejne n_1 bitów za cechę i ostatnie n_2 bitów za mantysę liczby x. Stąd podział n–bitowego słowa na trzy pola (podsłowa) jest jednoznacznie wyznaczony przez liczbę n_2. Wówczas $n_1 = n-1-n_2$. Powyższe uwagi prowadzą w naturalny sposób do ogólnej postaci funkcji kodu zmiennopozycyjnego KZP:

$$KZP[n,m]("a_{n-1}a_{n-2}...a_1a_0")[P,F] := (-1)^{a_{n-1}} \cdot$$

$$\cdot \left(\frac{1}{P} + \frac{P-1}{P} NKB[m]("a_{m-1}a_{m-2}...a_1a_0")[m]\right) \cdot$$

$$\cdot P^{F[n-m-1]("a_{n-2}...a_{m+1}a_m")[0]}$$

gdzie odpowiednio:

- n i m są parametrami opcjonalnymi wyrażającymi długość słowa binarnego "$a_{n-1}a_{n-2}...a_1a_0$" i długość jego podsłowa "$a_{m-1}a_{m-2}...a_1a_0$" reprezentującego mantysę kodowanej liczby, przy czym $1 \le m \le n$-3;
- F jest opcjonalną nazwą jednej ze stałopozycyjnych funkcji kodujących rozważanych w rozdziale 5.3;
- P jest opcjonalnym parametrem wyrażającym liczbę bazową kodu zmiennopozycyjnego.

Pierwszy czynnik w definicji kodu KZP wyraża znak kodowanej liczby, drugi jej mantysę, a trzeci czynnik skalujący. Wartością domyślną argumentu *P* jest liczba 2. W praktyce najczęściej bywa stosowana jedna z trzech wartości bazowych: 2, 10 albo 16. Wartością domyślną argumentu *F* jest funkcja kodu przesuniętego KP. Domyślnie wartości parametrów *n* i *m* zależą na ogół od przyjętego standardu. Najpowszechniej stosowany jest standard amerykański IEEE 754 lub IEEE 854 (ten ostatni dopuszcza możliwość różnych liczb bazowych).

Standard IEEE 754 dopuszcza dwa podstawowe formaty liczb: pojedynczej precyzji (32 bity) i podwójnej precyzji (64 bity) z możliwościami ich rozszerzenia. Format pojedynczej precyzji zakłada bit znaku, 23 – bitową mantysę i 8 – bitową cechę. Zakres wykładnika [-126; +127] i zakres formatu ok. 2^{128}, czyli ok. $3{,}8\times10^{38}$. Dokładność tego formatu wynosi ok. 2^{-23}, czyli ok. 10^{-7}. Format podwójnej precyzji zakłada bit znaku, 52 – bitową mantysę i 11 – bitową cechę. Zakres wykładnika [-1022, +1023] i zakres formatu ok. 2^{1024}, czyli ok. 9×10^{307}. Dokładność tego formatu wynosi ok. 2^{-52}, czyli ok. 10^{-15}.

Możliwość rozszerzenia wspomnianych formatów jest przewidziana w celu zwiększenia dokładności obliczeń pośrednich w celu zminimalizowania błędów zaokrągleń na ostateczny wynik reprezentowany przez formaty podstawowe. Nieco mniejszy zakres tych formatów w porównaniu z teoretycznym zakresem wynikającym z definicji funkcji KZP bierze się stąd, że pewne ciągi binarne są zarezerwowane do reprezentowania wartości specjalnych.

Dowód twierdzenia 5.3.1: Ustalmy dowolnie $p>1$ i $x \in \mathrm{R}\setminus\{0\}$.

Wówczas $\lim_{k\to\infty} \frac{x}{p^k} = 0$, a to oznacza istnienie wskaźnika $k_0 \in N$ spełniającego nierówność $0 < \frac{|x|}{p^{k_0}} < 1$. Wówczas

$$A_p(x) := \{k \in Z : \frac{|x|}{p^k} < 1\}$$

jest niepustym zbiorem, gdyż $k_0 \in A_p(x)$. Ponieważ

$$\lim_{k\to\infty} \frac{|x|}{p^{-k}} = \lim_{k\to\infty} |x| \cdot p^k = +\infty,$$

więc $\frac{|x|}{p^{-k}} \geq 1$ dla $k \geq k_1$, gdzie $k_1 \in N$ jest pewną liczbą.

Stąd $\frac{|x|}{p^k} \geq 1$ dla $k \leq -k_1$ i w konsekwencji $k \notin A_p(x)$ dla $k \leq -k_1$.

Zatem liczba $-k_1$ ogranicza zbiór $A_p(x)$ od dołu. Reasumując, zbiór $A_p(x)$ jest niepustym i ograniczonym z dołu podzbiorem zbioru liczb całkowitych Z. Z zasady minimum wynika zatem, że zbiór $A_p(x)$ ma element najmniejszy. Możemy zatem określić liczbę $r := \min A_p(x)$. Stąd na podstawie nierówności

$$0 < \frac{|x|}{p^r} < 1 \quad \text{i} \quad \frac{|x|}{p^{r-1}} \geq 1$$

dostajemy nierówności (5.3.2). Pozostaje wykazać jedyność liczby r spełniającej nierówności (5.3.2). Załóżmy więc, że zachodzą nierówności

$$\frac{1}{p} \leq \frac{|x|}{p^{r'}} < 1$$

dla dowolnie zadanej liczby całkowitej r'. Powyższe nierówności łącznie z nierównościami (5.3.2) dają przez podzielenie stronami nierówności

$$p^{-1} < p^{r'-r} < p^1.$$

Stąd $-1 < r'-r < 1$. Ponieważ liczba $r'-r$ jest całkowita więc $r'-r=0$. Zatem $r'=r$, co kończy dowód jedyności liczby r i tym samym dowód twierdzenia.