Algorytmy

Co ma Bézier do B-spline'a?*

Bogusław Jackowski, BOP s.c.

Wstęp. Każdy, kto korzystał z METAFONT-a, METAPOST-a czy PostScript-u wie, że podstawowym „budulcem" tych systemów są *krzywe Béziera*. Dość pobieżne wiadomości o krzywych Béziera można wyczytać z podręcznika Donalda E. Knutha *The* METAFONT*book*. Jeszcze mniej na ten temat mówi podstawowy podręcznik PostScript-u, tj. *PostScript Language Reference*, ale oba podręczniki zawierają przynajmniej definicję krzywych Béziera.

Natomiast zupełnie kuriozalna jest informacja o *krzywych giętych* (ang. *B-splines*), stanowiących podstawowy „budulec" fontów TrueType, zawarta w oficjalnej specyfikacji tego formatu:

> Kontury zbudowane są z linii prostych i krzywych. Krzywe definiowane są za pomocą ciągu punktów, opisujących krzywe gięte drugiego rzędu. Dla potrzeb formatu TrueType krzywe gięte określane są przez zadanie punktów leżących na krzywej i leżących poza krzywą. Dowolna kombinacja punktów „na" i „poza" jest dopuszczalna. Dwa występujące pod rząd punkty „na" określają linię prostą.

To wszystko. Ani słowa więcej o krzywych giętych.

Z kolei w podręcznikach metod numerycznych przedstawiana jest przeważnie „ogólna teoria wszystkiego", z której – nie będąc zawodowym matematykiem – ciężko zrozumieć, jaki jest związek tego, co w podręcznikach stoi napisane, z tym, co zostało zaimplementowane w komputerach. W szczególności żaden ze znanych mi podręczników nie pisze w sposób jawny, jaki jest związek między krzywymi Béziera a krzywymi giętymi drugiego rzędu, co z punktu widzenia zastosowań fontowych jest istotnym zagadnieniem.

Udzielenie sobie samemu odpowiedzi na to proste pytanie okazało się bynajmniej nie takie znów proste. Postanowiłem więc swoje notatki spisać w czytelnej formie, ilustrując je obficie przykładami, w nadziei, że dzięki temu może ktoś uniknie wyważania otwartych drzwi.

Interpolacja. Zadanie znalezienia krzywej interpolującej dany zbiór punktów sprowadza się w istocie do wyliczania pewnej kombinacji liniowej tych punktów, zależnej od jednego parametru przy interpolacji krzywą, od dwóch parametrów przy interpolacji powierzchnią, itd. W dalszym ciągu zajmować się będziemy interpolacją jednoparametrową na płaszczyźnie.

Niech zatem $P_0, P_1, \ldots, P_n$ będą punktami na płaszczyźnie. Przypuśćmy, że mamy dane funkcje $F_i(t)$, $i = 0, 1, \ldots, n$, określone na przedziale $[0, 1]$, takie, że dla każdego $0 \leqslant t \leqslant 1$ zachodzi:

$$0 \leqslant F_i(t) \leqslant 1, \quad \sum_{i=0}^{n} F_i(t) = 1\,. \tag{1}$$

O funkcji $F(t)$ określonej wzorem

$$F(t) = \sum_{i=0}^{n} F_i(t)\, P_i \tag{2}$$

będziemy mówić, że interpoluje punkty $P_0, P_1, \ldots, P_n$ lub że jest funkcją (krzywą) interpolującą te punkty; funkcje $F_0, F_1, \ldots, F_n$ nazywać będziemy funkcjami bazowymi. Sensowność wynikowej interpolacji zależy oczywiście od definicji funkcji bazowych. Niezależnie od sensowności, krzywa interpolująca *zawsze* leży wewnątrz wielokąta wypukłego opisanego na zbiorze punktów $P_0, P_1, \ldots, P_n$. Wynika to wprost definicji funkcji interpolujących (równanie 1).

Przykład 0. Jeśli $F_k(t) = 1$ dla pewnego $0 \leqslant k \leqslant n$ i dla każdego $0 \leqslant t \leqslant 1$ (wynika stąd, że pozostałe funkcje bazowe muszą być tożsamościowo równe zeru), to $F(t) = P_k$. Wypada się zgodzić, że przykład pokazuje nie najmądrzejszy dobór funkcji bazowych. □

$*$: Tekst był prezentowany po raz pierwszy na Krakowskich Warsztatach KrakoTEX 2000.

Przykład 1. Jeśli natomiast

$$F_i(t)= \begin{cases} 1 & \text{dla} \quad \frac{i}{n+1} \leqslant t < \frac{i+1}{n+1}, \quad i = 0, 1, \ldots, n-1, \\ 1 & \text{dla} \quad \frac{i}{n+1} \leqslant t \leqslant 1, \quad i = n, \\ 0 & \text{w pozostałych przypadkach,} \end{cases}$$

to otrzymujemy w wyniku aproksymację dyskretną: funkcja interpolująca $F(t)$ przyjmuje nad poszczególnymi przedziałami wartości P_0, P_1, ..., P_n. Zwróćmy uwagę, że wartość funkcji $F_n(t)$ należy określić nad przedziałem *obustronnie* domkniętym, gdyż w przeciwnym razie niezależnie od zbioru punktów P_0, P_1, ..., P_n zawsze zachodziłoby $F(1)= \mathbf{0}$, gdzie $\mathbf{0}$ oznacza punkt leżący w początku układu współrzędnych. Ten prosty przykład ilustruje pewien techniczny kłopot, który uporczywie pojawia się rozważaniach interpolacyjnych: należy starannie określić, który z końców rozważanego przedziału jest otwarty, a który domknięty, i co stąd wynika. □

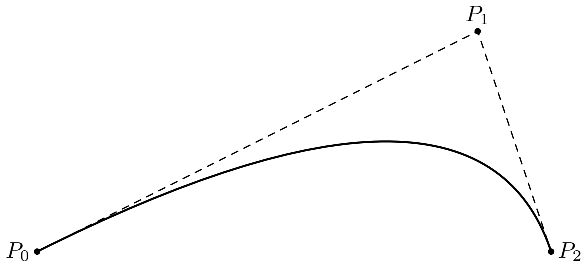
Przykład 2. Niech $n = 2$ i niech

$$F_0 = (1-t)^2, \quad F_1 = 2t(1-t), \quad F_2 = t^2.$$

Oczywiście $0 \leqslant F_i(t) \leqslant 1$ dla $0 \leqslant t \leqslant 1$, a ponadto

$$F_0(t)+F_1(t)+F_2(t)= (t + (1-t))^2 = 1,$$

czyli można te funkcje wykorzystać jako funkcje bazowe. Zauważmy, że krzywa interpolująca przechodzi przez skrajne punkty: $F(0)= P_0$, gdyż $F_0(0)= 1$, a $F_1(0)= F_2(0)= 0$; analogicznie $F(1)= P_2$ gdyż $F_0(1)= F_1(1)= 0$, $F_2(1)= 1$. Natomiast $F(\frac{1}{2})= P_1$ wtedy i tylko wtedy, gdy P_1 jest środkiem odcinka łączącego P_0 i P_2 (tzn. $P_1 = \frac{1}{2}P_0 + \frac{1}{2}P_2$). Z definicji funkcji interpolującej wynika, że jeśli punkty P_0, P_1 i P_2 leżą na jednej prostej, to wszystkie punkty krzywej $F(t)$ leżą na tej prostej (por. rys. 1). □



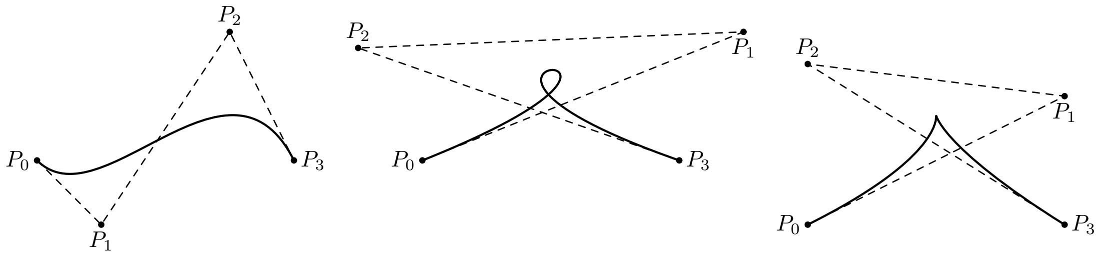
Rys. 1. Interpolacja za pomocą funkcji bazowych stopnia drugiego:
$F(t)= (1-t)^2\, P_0 + 2t(1-t)\, P_1 + t^2\, P_2$. W ogólnym przypadku P_1 nie należy do krzywej $F(t)$.
Zawsze jednak odcinek $P_0\ P_1$ jest styczny do krzywej w punkcie P_0, a odcinek
$P_1\ P_2$ – w punkcie P_2. Ponadto cała krzywa mieści się w trójkącie $P_0\ P_1\ P_2$.

Przykład 3. Niech $n = 3$ i niech

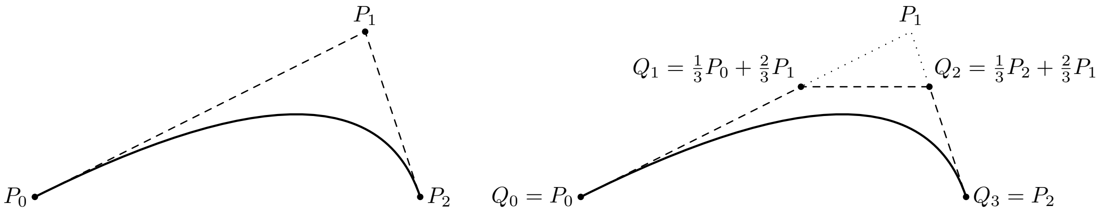
$$F_0 = (1-t)^3, \quad F_1 = 3t(1-t)^2, \quad F_2 = 3t^2(1-t), \quad F_3 = t^3.$$

Podobna argumentacja jak w przykładzie 2 pokazuje, że $F(0)= P_0$, $F(1)= P_3$; ponadto $F(\frac{1}{3})= P_1$ i $F(\frac{2}{3})= P_2$ jedynie wówczas, gdy $P_1 = \frac{2}{3}P_0 + \frac{1}{3}P_3$ i $P_2 = \frac{1}{3}P_0 + \frac{2}{3}P_3$. (Punkty P_1 i P_2 zwane są żargonowo „naciągami”.) Jeśli punkty P_0, P_1, P_2 i P_3 leżą na jednej prostej, to – analogicznie jak poprzednio – wszystkie punkty $F(t)$ leżą na tej prostej. Pełna analiza interpolacji za pomocą funkcji bazowych stopnia trzeciego jest znacznie bardziej skomplikowana, gdyż krzywe trzeciego stopnia mogą zawierać punkty przegięcia, mogą się też przecinać same z sobą (raz), może się pojawiać ostrze, tzn. punkt w którym styczna jest nieokreślona. Wymienione czynniki w istotny sposób utrudniają obróbkę numeryczną takich krzywych (p. rys. 2). □

Przykład 4. Pytanie o związek między krzywymi stopnia 2 i 3 nasuwa się samo. Wyrażenie funkcji stopnia 2 za pomocą funkcji stopnia 3 jest trywialne (rys. 3). Natomiast zamiana krzywej stopnia 3 na krzywą stopnia 2 jest w ogólnym przypadku niemożliwa – i to jest oczywiste. Nie jest natomiast oczywiste, w jaki sposób dobrze przybliżać krzywą stopnia 3 za pomocą krzywej lub kilku krzywych stopnia 2 (głównie z powodu trudności w formalnym określeniu słowa „dobrze”, acz nie tylko). Jest to zagadnienie zbyt obszerne, by przedstawić je w tak krótkim szkicu. □



Rys. 2. Interpolacja za pomocą funkcji bazowych stopnia trzeciego:
$F(t) = (1-t)^3 P_0 + 3t(1-t)^2 P_1 + t^2(1-t) P_2 + t^3 P_3$. Podobnie jak w przykładzie 2, w ogólnym przypadku „naciągi” P_1 i P_2 nie należą do krzywej $F(t)$. Zawsze jednak odcinek $P_0\,P_1$ jest styczny do krzywej w punkcie P_0, a odcinek $P_2\,P_3$ – w punkcie P_3. Ponadto cała krzywa mieści się w czworokącie wypukłym wyznaczonym przez punkty P_0, P_1, P_2 oraz P_3.



Rys. 3. Krzywą stopnia 2 (lewa strona rysunku) można automatycznie przekształcić w krzywą stopnia 3 (prawa strona) – wystarczy „naciągi” umieścić na bokach trójkąta $P_0\,P_1\,P_2$ w dwóch trzecich odległości od skrajnych węzłów. Aproksymacja funkcji stopnia 3 za pomocą funkcji stopnia 2 wymaga na ogół zaawansowanych technik analizy numerycznej.

Dygresja historyczna. Funkcje bazowe stopnia drugiego i trzeciego, występujące w przykładach 2 i 3, dają się w sposób naturalny uogólnić:

$$F_i(t) = \binom{n}{i}\, t^i\, (1-t)^{n-i}, \quad i = 0, 1, \ldots, n.$$

Z definicji symbolu Newtona $\binom{n}{i}$ jako współczynnika przy członie $a^i\, b^{n-i}$ rozwinięcia dwumianu $(a+b)^n$ wynika, że $\sum_{i=0}^{n} F_i(t) = 1$, a tym samym spełniony jest podstawowy warunek, jaki nałożony został na funkcje bazowe (równanie 1).

Z wielomianami wysokiego stopnia jest jednak pewien problem. Trudno mianowicie zapanować nad przebiegiem krzywej interpolującej – kłopoty wspomniane w przykładzie 3 rosną wraz ze wzrostem stopnia wielomianu.

Tak zdefiniowane funkcje bazowe zwane są czasem wielomianami Bernsztejna, gdyż Sergej N. Bernsztejn (Сергей Н. Бернштейн; w literaturze anglosaskiej: Sergeĭ N. Bernsteĭn) wprowadził je w swoich pracach dotyczących teorii aproksymacji już w 1912 roku.

Krzywe konstruowane za pomocą wielomianów Bernsztejna stopnia trzeciego znane są powszechnie jako krzywe Béziera. Pierre Bézier, pracujący w latach sześćdziesiątych dla firmy Renault, jako jeden z pierwszych uzmysłowił sobie, że krzywe te można wykorzystać w grafice komputerowej. Nie był on jedynym, który wykorzystywał krzywe, składające się z połączonych ze sobą kawałków krzywych opisywanych wielomianami niskiego stopnia. W tym samym czasie podobne podejście zastosowali między innymi Paul de Casteljau (Citroen), Garret Birkhoff, Henry L. Garabedian i Carl de Boor (General Motors), James C. Ferguson (Boeing), Malcolm A. Sabin (British Aircraft) i inni.

Najwidoczniej w latach sześćdziesiątych sytuacja dojrzała i w wielu miejscach na świecie równocześnie pojawiły się podobne rozwiązania. Nieprzypadkowo prym wiodły firmy konstruujące samochody i samoloty.

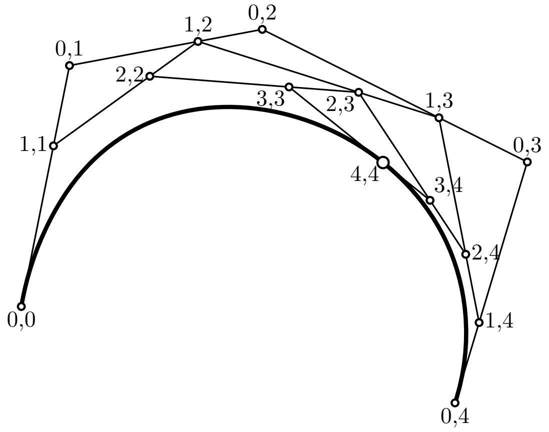
Dygresja algorytmiczna. Paul de Casteljau jest autorem pięknego – bo prostego – algorytmu, pozwalającego w efektywny sposób wyznaczać wartość wielomianu Bernsztejna w danym punkcie:

dane: $0 \leqslant t \leqslant 1$ oraz punkty na płaszczyźnie $P_0, P_1, \ldots, P_n$
wynik: $Q_{n,n} = F(t)$ – wartość wielomianu Bernsztejna stopnia n w punkcie t

dla $j = 0, 1, \ldots, n : Q_{0,j} \leftarrow P_j$ **ald**
dla $i = 1, \ldots, n :$
  dla $j = i, \ldots, n :$
    $Q_{i,j} \leftarrow (1 - t)\, Q_{i-1,j-1} + t\, Q_{i-1,j}$
  ald
ald

Algorytm ten ma prostą interpretację geometryczną: punkty są łączone linią łamaną, następnie każdy z odcinków jest dzielony w proporcji $t : (1 - t)$; tak wyznaczone punkty ponownie są łączone linią łamaną i znów odcinki dzielone są w proporcji $t : (1 - t)$; proces powtarza się dopóty, dopóki nie otrzyma się w wyniku pojedynczego punktu (por. rys. 4).

W istocie algorytm de Casteljau określa równocześnie podział krzywej danego stopnia na dwa fragmenty (tego samego stopnia). Pierwszy fragment wyznaczają punkty $Q_{0,0}$, $Q_{1,1}$, …, $Q_{n,n}$, drugi – punkty $Q_{n,n}$, $Q_{n-1,n}$, …, $Q_{0,n}$.



Rys. 4. Geometryczna interpretacja algorytmu de Casteljau: interpolacja wielomianami stopnia 4; $t = \frac{2}{3}$; węzły $Q_{i,j}$, wyznaczane w trakcie obliczeń, na rysunku opatrzone są etykietami i, j. Interpolacja punktów $Q_{0,0}$, $Q_{1,1}$, $Q_{2,2}$, $Q_{3,3}$ i $Q_{4,4}$ daje w wyniku początkowy fragment krzywej (łuk $Q_{0,0}$ — $Q_{4,4}$), interpolacja punktów $Q_{4,4}$, $Q_{3,4}$, $Q_{2,4}$, $Q_{1,4}$ i $Q_{0,4}$ – końcowy fragment (łuk $Q_{4,4}$ — $Q_{0,4}$).

W stronę krzywych giętych. Matematykom nie odpowiadało to, że stopień funkcji bazowych miałby zależeć od liczebności zbioru interpolowanych punktów. Wymyślili więc dość skomplikowane twory, znane w literaturze anglosaskiej pod nazwą *B-splines*. „B” pochodzi od słowa „basis” (*baza*), zaś słowo „spline” jest technicznym określeniem, oznaczającym m.in. giętki przedmiot służący do kreślenia krzywych (krzywik nastawny, giętka traserska). W literaturze polskiej można natrafić beztroskie używanie nazwy *B-spline*, czasem spotyka się określenie *funkcje B-sklejane*, rzadziej *funkcje składane*; w niniejszym opracowaniu stosowane jest najzgrabniejsze zdaniem autora tłumaczenie *krzywe* (lub *funkcje*) *gięte*, pochodzące od gdańskiego matematyka Zbigniewa Ciesielskiego, uznanego autorytetu w dziedzinie teorii aproksymacji i probabilistyki.

Zanim całkiem przestawimy się na aparat pojęciowy krzywych giętych, wróćmy na chwilę do tego, co już wiemy, tzn. do interpolacji wielomianami stopnia 2 i 3.

Przykład 5. Przypuśćmy, że mamy pięć funkcji bazowych:

$$F_0^0(t) = F_1^0(t) = F_3^0(t) = F_4^0(t) = 0, \quad F_2^0(t) = 1, \quad \text{dla} \quad 0 \leqslant t \leqslant 1.$$

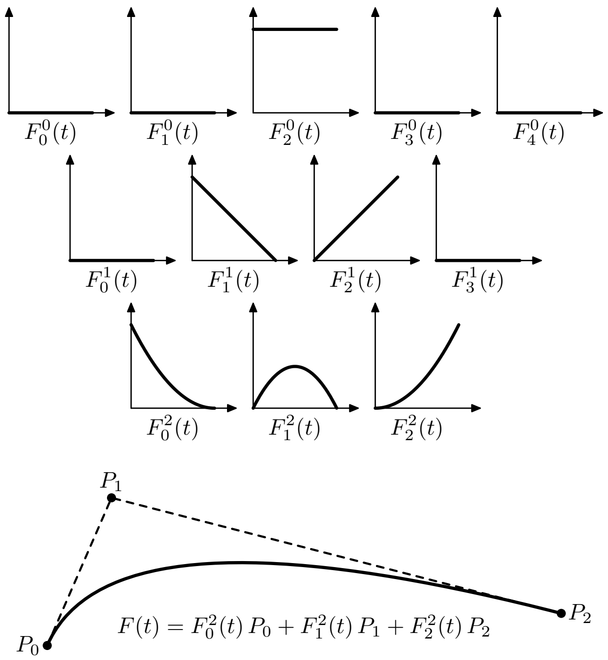
Taki zestaw funkcji bazowych określony został w przykładzie 0 mianem „nie najmądrzejszego”. Dlatego nie użyjemy tego zestawu do interpolacji zbioru pięciu punktów, tylko zdefiniujemy sobie na jego podstawie inny zestaw:

$$F_0^1(t) = t\,F_0^0(t) + (1-t)\,F_1^0(t),$$
$$F_1^1(t) = t\,F_1^0(t) + (1-t)\,F_2^0(t),$$
$$F_2^1(t) = t\,F_2^0(t) + (1-t)\,F_3^0(t),$$
$$F_4^1(t) = t\,F_3^0(t) + (1-t)\,F_4^0(t).$$

Ten zestaw także jest niezadowalający, gdyż funkcje $F_0^1(t)$ i $F_4^1(t)$ są tożsamościowo równe zeru. Nie poprzestaniemy więc na tym i zdefiniujemy w analogiczny sposób zestaw trzech funkcji bazowych:

$$F_0^2(t) = t\,F_0^1(t) + (1-t)\,F_1^1(t),$$
$$F_1^2(t) = t\,F_1^1(t) + (1-t)\,F_2^1(t),$$
$$F_2^2(t) = t\,F_2^1(t) + (1-t)\,F_3^1(t).$$

Teraz już widać do czego zmierzaliśmy: funkcje $F_0^2(t)$ $F_1^2(t)$ i $F_2^2(t)$ to po prostu funkcje $F_0(t)$, $F_1(t)$ i $F_2(t)$ z przykładu 2, czyli wielomiany $(1-t)^2$, $2(1-t)\,t$, t^2 (por. rys. 5). $\square$



Rys. 5. Interpolacja punktów P_0, P_1, P_2 za pomocą krzywych giętych drugiego rzędu (wyznaczona przez podział odcinka $[0,1]$ punktami $t_0 = t_1 = t_2 = 0$, $t_3 = t_4 = t_5 = 1$; por. punkt *Krzywe gięte* niżej), równoważna interpolacji za pomocą wielomianów Bernsztejna stopnia 2.

Przykład 6. Rozważmy nieco inny przypadek, bardziej rodem ze świata krzywych giętych. Niech będzie danych $n+1$ funkcji bazowych ($n > 2$) zdefiniowanych następująco:

$$F_0^0(t) = F_n^0(t) = 0 \quad \text{dla} \quad 0 \leqslant t \leqslant 1,$$

$$F_i^0(t) = \begin{cases} 1 & \text{dla} \quad \frac{i-1}{n-1} \leqslant t < \frac{i}{n-1}, \quad i = 1, \ldots, n-2, \\ 1 & \text{dla} \quad \frac{i-1}{n-1} \leqslant t \leqslant 1, \quad i = n-1, \\ 0 & \text{w pozostałych przypadkach.} \end{cases}$$

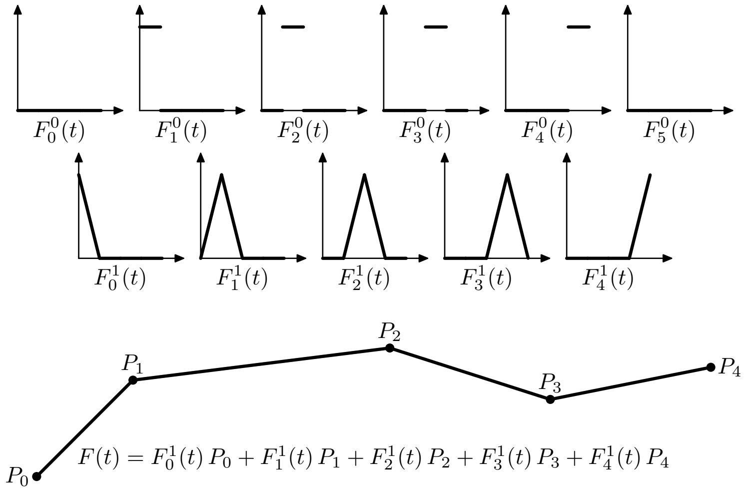
W powyższej definicji pojawia się ten sam techniczny kłopot, o którym była już mowa w przykładzie 1. Konsekwentne użycie przedziałów obustronnie domkniętych w definicji funkcji $F_i^0(t)$ byłoby niepoprawne, gdyż wówczas $\sum_{i=0}^{n} F_i^0(t) = 2$ dla $t = \frac{1}{n-1}, \frac{2}{n-1}, \ldots, \frac{n-2}{n-1}$, wbrew podstawowej własności funkcji bazowych (równanie 1).

Zdefiniujemy teraz, podobnie jak w poprzednim przykładzie, funkcje $F_i^1(t)$. Definicja będzie jednak bardziej skomplikowana, aby zatem uprościć notację, nadajmy „nazwy” charakterystycznym punktom przedziału $[0,1]$: $t_0 = t_1 = 0$, $t_2 = \frac{1}{n-1}$, $t_{n-1} = \frac{n-2}{n-1}$, $t_n = t_{n+1} = 1$. Zdefiniujmy teraz za pomocą tak określonych punktów nowe funkcje bazowe:

$$F_i^1(t) = \frac{t - t_i}{t_{i+1} - t_i} F_i^0(t) + \frac{t_{i+2} - t}{t_{i+2} - t_{i+1}} F_{i+1}^0(t), \quad t_i \leqslant t \leqslant t_{i+2}, \quad i = 0, 1, \ldots, n-1.$$

Czujny czytelnik na pewno zauważył, że może tu dojść do dzielenia zera przez zero, np. współczynnik przed pierwszym składnikiem w definicji funkcji $F_0^1(t)$ ma taką postać; tradycja związana z teorią krzywych giętych nakazuje przyjąć, że $\frac{0}{0} \stackrel{\text{def}}{=} 0$ i to leczy chorobę. Na pierwszy rzut oka na błąd wygląda również użycie przedziałów obustronnie domkniętych w definicji funkcji $F_i^1(t)$. Staranniejsze przyjrzenie się formułom pokazuje, że dzięki ciągłości funkcji wynikowych definicja jest poprawna.

Trudno się połapać, o co chodzi w tych wzorach, ale co nieco powinien wyjaśniać rysunek 6. Okazuje się, że w dość zawiły sposób otrzymaliśmy zestaw funkcji do interpolacji liniowej. Gdyby zaprezentowana tu technika miała służyć tylko definiowaniu interpolacji liniowej, byłoby to niepotrzebne uogólniactwo. Oczywiście tak źle nie jest. Podejście to okazało się bardzo owocne i współcześnie znajduje wiele zastosowań, m.in. wykorzystane zostało – jak zostało na wstępie odnotowane – w specyfikacji fontów TrueType. □



Rys. 6. Interpolacja za pomocą krzywych giętych pierwszego rzędu czyli interpolacja liniowa, odpowiadająca podziałowi odcinka $[0,1]$ punktami $t_0 = t_1 = 0$, $t_2 = \frac{1}{4}$, $t_3 = \frac{1}{2}$, $t_4 = \frac{3}{4}$, $t_5 = t_6 = 0$ (por. przykład 6).

Przykład 7. W podobny sposób jak w przykładzie 5 można uzyskać krzywe Béziera. Należy rozpocząć od siedmiu funkcji bazowych:

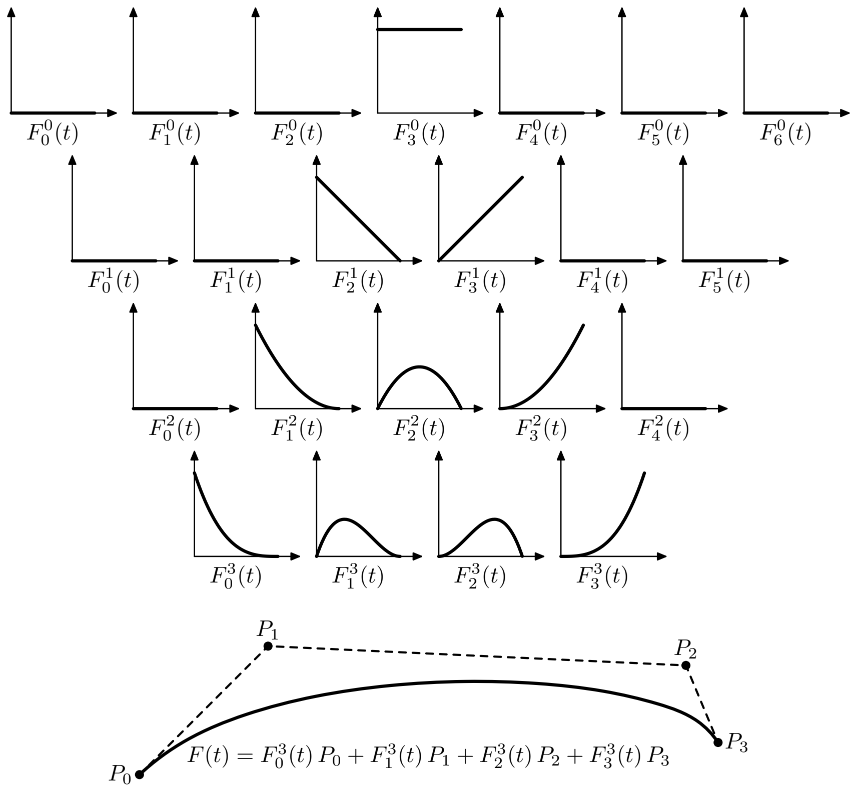
$$F_0^0(t) = F_1^0(t) = F_2^0(t) = F_4^0(t) = F_5^0(t) = F_6^0(t) = 0, \quad F_3^0(t) = 1, \quad \text{dla} \quad 0 \leqslant t \leqslant 1$$

i wykonać trzy kroki analogiczne jak w przykładzie 5. W wyniku otrzymuje się znane nam z przykładu 3 wielomiany Bernsztejna stopnia 3: $F_0^3(t) = (1-t)^3$, $F_1^3(t) = 3(1-t)^2 t$, $F_3^3(t) = 3(1-t)t^2$ oraz $F_0^3(t) = t^3$ (por. rys 7). □

Krzywe gięte. W opisie rysunków 5 i 7 pojawiły się „podziały odcinka” $[0,1]$, o których nie było mowy w odpowiadających tym rysunkom przykładach 5 i 7, chociaż analogii z przykładem 6 nietrudno się dopatrzeć. Otóż podział odcinka $[0,1]$ to kluczowe pojęcie związane z krzywymi giętymi: funkcje bazowe wyznaczone są jednoznacznie przez podział odcinka $0 = t_0 \leqslant t_1 \leqslant \ldots \leqslant t_n \leqslant t_{n+1} = 1$. Proszę zauważyć, że nie żąda się, aby punkty podziału były parami różne. Teoretycy mówią w takiej sytuacji o punktach lub węzłach wielokrotnych.

Określenie „węzły” jest dwuznaczne: może dotyczyć punktów podziału odcinka albo interpolowanych punktów; w dalszym ciągu używane będzie tylko w tym drugim znaczeniu, a do punktów wyznaczających podział odcinka będzie stosowane określenie „punkty charakterystyczne”.

Jak z przykładów wynika, w praktyce jedynie na krańcach odcinka $[0,1]$ pojawiają się punkty wielokrotne, i to tej samej krotności. Ma przeto sens ograniczenie się tylko do takiej klasy krzywych giętych.



Rys. 7. Interpolacja punktów P_0, P_1, P_2, P_3 za pomocą krzywych giętych trzeciego rzędu (wyznaczona przez podział odcinka $[0,1]$ punktami $t_0 = t_1 = t_2 = t_3 = 0$, $t_4 = t_5 = t_6 = t_7 = 1$), równoważna interpolacji za pomocą wielomianów Bernsztejna stopnia 3.

Niech zatem będzie dany następujący podział odcinka $[0,1]$:

$$0 = t_0 = t_1 = \ldots = t_k < t_{k+1} < \ldots < t_{n-k} < t_{n-k+1} = \ldots = t_n = t_{n+1} = 1, \tag{3}$$

z ewentualnymi punktami wielokrotnymi znajdującymi się na skraju odcinka. Bazowe funkcje gięte rzędu 0 definiuje się na podstawie takiego podziału w następujący sposób:

$$F_0^0(t) = \ldots F_{k-1}^0(t) = F_{n-k+1}^0(t) = \ldots = F_n^0(t) = 0 \quad \text{dla} \quad 0 \leqslant t \leqslant 1$$
$$F_i^0(t) = \begin{cases} 1 & \text{dla} \quad t_i \leqslant t < t_{i+1}, \quad i = k, \ldots, n-k-1, \\ 1 & \text{dla} \quad t_i \leqslant t \leqslant 1, \quad i = n-k, \\ 0 & \text{w pozostałych przypadkach.} \end{cases} \tag{4a}$$

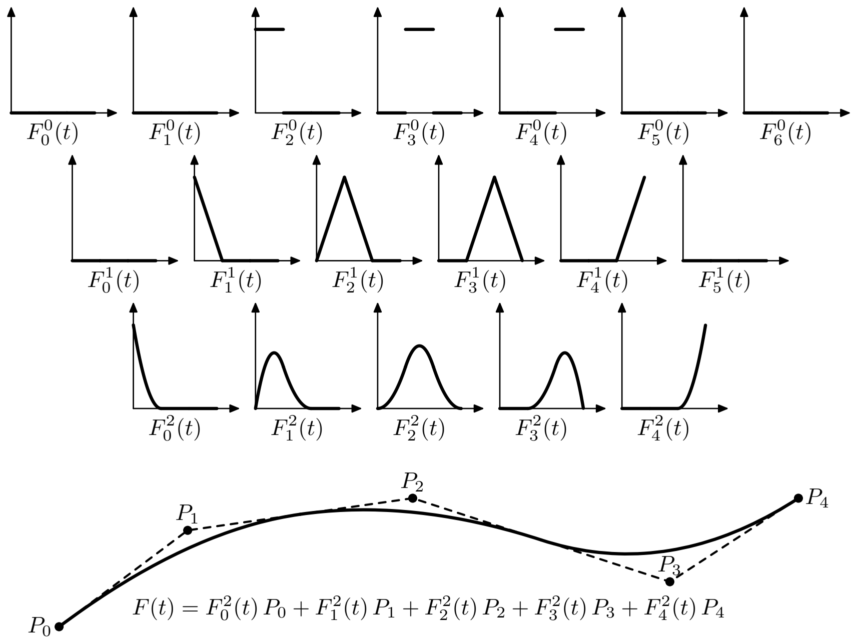
Ten piętrowy wzór mówi raptem tyle, że jeśli $t_i = t_{i+1}$, to $F_i^0(t) = 0$ nad całym odcinkiem $[0,1]$, a jeśli $t_i < t_{i+1}$, to $F_i^0(t) = 1$ nad odcinkiem prawostronnie otwartym $[t_i, t_{i+1})$, a $F_i^0(t) = 0$ poza nim, z wyjątkiem prawego skrajnego odcinka (niezdegenerowanego), gdzie $F_{n-k}^0(t) = 1$ nad odcinkiem obustronnie domkniętym $[t_{n-k}, t_{n-k+1}]$ (i oczywiście $F_{n-k}^0(t) = 0$ poza nim).

Krzywe gięte rzędu $r > 0$ definiuje się rekurencyjnie:

$$F_i^r(t) = \frac{t - t_i}{t_{i+1} - t_i}\, F_i^{r-1}(t) + \frac{t_{i+r+1} - t}{t_{i+r+1} - t_{i+1}}\, F_{i+1}^{r-1}(t), \quad t_i \leqslant t \leqslant t_{i+r+1}, \quad i = 0, 1, \ldots, n-r. \tag{4b}$$

przy czym obowiązuje konwencja, że $\frac{0}{0} = 0$; ze względu na ciągłość funkcji bazowych rzędu większego niż 0 nie ma potrzeby rozważania przedziałów półotwartych (por. przykład 6).

W praktyce interesujące są jedynie funkcje bazowe odpowiedniego rzędu, mianowicie rzędu o jeden niższego niż krotność skrajnych węzłów, czyli rzędu k (por. równanie 4a): zarówno funkcje rzędu niższego niż k, jak i wyższego, dają funkcje interpolujące *niezależne* od skrajnych węzłów. Szczegóły są w obu przypadkach nieco inne (przypadek zbyt małego rzędu przedstawiony jest na rys. 9a), ale w sumie efekt jest podobny: takie funkcje bazowe rozsądnie jest odrzucić. Można zatem powiedzieć, że w praktycznych zastosowaniach krotność skrajnych węzłów wyznacza rząd funkcji bazowych.



Rys. 8. Interpolacja za pomocą krzywych giętych drugiego rzędu: $n = 6$, $k = 2$, $t_0 = t_1 = t_2 = 0$, $t_3 = \frac{1}{3}$, $t_4 = \frac{2}{3}$, $t_5 = t_6 = t_7 = 1$. Wynikowa krzywa jest styczna do łamanej łączącej „naciągi".

Zauważmy, że aproksymacja za pomocą wielomianów Bernsztejna to szczególny przypadek krzywych giętych, określonych przez następujący rozkład punktów charakterystycznych: $0 = t_0 = t_1 = \ldots = t_k$, $t_{k+1} = \ldots = t_{2k} = t_{2k+1} = 1$.

Krzywe gięte w fontach TrueType. Jeśli chodzi o formalizm, to tyle nam wystarczy. Pozostało nam jedynie spróbować zrozumieć, w jaki sposób definiowane są za pomocą pomocą krzywych giętych kontury znaków w formacie TrueType.

Krzywą interpolującą stopnia drugiego można – przy założeniu równomierności rozkładu punktów charakterystycznych – w trywialny sposób wyrazić za pomocą omawianych wcześniej krzywych stopnia drugiego. Wystarczy wyznaczyć środki „wewnętrznych odcinków"; niezależna interpolacja kolejnych trójek punktów daje w wyniku tę samą krzywą (por. rys. 10). Rozkład nierównomierny wymaga trochę więcej zachodu, ale w dalszym ciągu zamiana na interpolowane niezależne trójki punktów jest możliwa.

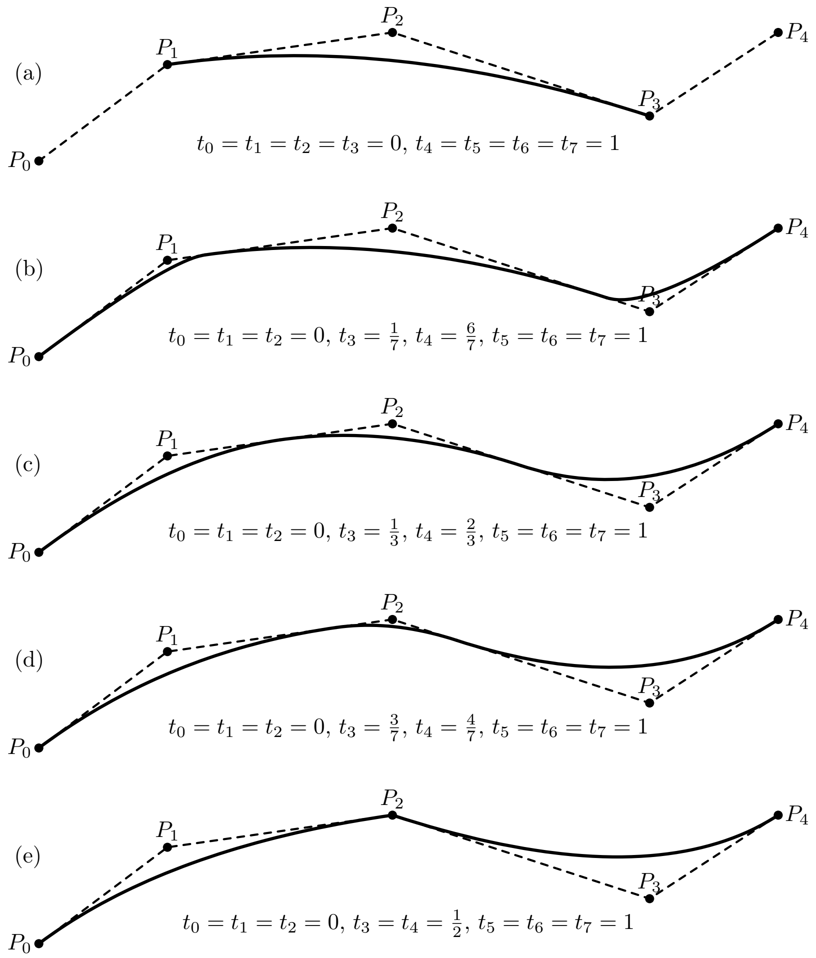
W przypadku fontów TrueType wolno przypuszczać, choć nie jest to nigdzie *explicite* napisane, że w grę wchodzi jedynie równomierny rozkład punktów charakterystycznych. Przyjąwszy takie założenie można nadać sens przytoczonej na wstępie bełkotliwej frazie ze specyfikacji formatu TrueType.

Otóż należy ją rozumieć następująco. Przypuśćmy, że mamy dany ciąg punktów P_0, P_1, P_2, ..., P_n. Każdy z tych punktów ma przypisany atrybut n lub p odpowiadający żądaniu, by punkt leżał *na* krzywej lub *poza* nią, co zapisywać będziemy odpowiednio $s(P_i)= n$ lub $s(P_i)= p$. Chociaż specyfikacja twierdzi, że *dowolna kombinacja punktów „na" i „poza" jest dopuszczalna*, należy raczej założyć, że skrajne punkty powinny leżeć na krzywej, tzn. $s(P_0)= s(P_n)= n$.

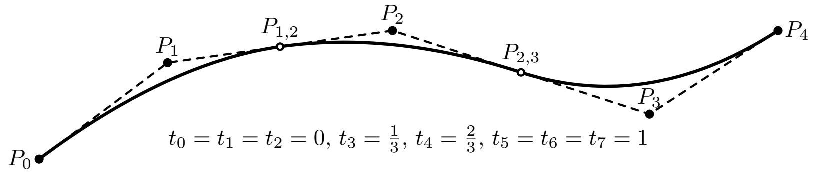
Podzielmy wyjściowy ciąg na podciągi takie, że ich skrajne punkty znajdują się na krzywej, a wewnętrzne – poza. Niech jednym z takich ciągów będzie P_i, P_{i+1}, ..., P_{j-1}, P_j. Z założenia $s(P_i)= s(P_j)= n$, $s(P_{i+1})= \ldots = s(P_{j-1})= p$. Każdy z takich ciągów interpolujemy niezależnie krzywą giętą drugiego rzędu, przyjmując stosowny dlań rozkład punktów charakterystycznych:

$$t_i = \begin{cases} 0 & \text{dla} \quad i = 0,\, 1,\, 2 \\ \frac{i-2}{n-2} & \text{dla} \quad i = 3, \ldots, n-2 \\ 1 & \text{dla} \quad i = n-1,\, n,\, n+1 \end{cases}$$

gdzie $n = j - i + 4$. Z rozważań przeprowadzonych w poprzednim punkcie wynika, że każdy podciąg P_i, $P_{i+1}, \ldots, P_{j-1}$, P_j można jeszcze bardziej rozdrobnić, wstawiając między punktami P_k i P_{k+1} (dla $k = i + 1$,



Rys. 9. Położenie punktów styczności zależy od doboru punktów charakterystycznych: przy podziale równomiernym punkty styczności wypadają na środku odcinków łączących poszczególne węzły (wariant c). Zbliżenie środkowych punktów charakterystycznych ku sobie spowodowałoby przesunięcie punktów styczności ku węzłowi P_2, oddalenie zaś – rozsunięcie ku punktom P_1 i P_3 (warianty b i d). Zmniejszenie odległości między węzłami środkowymi do zera (wariant e) prowadzi do przypadku równoważnego niezależnej interpolacji trójek punktów $P_0\,P_1\,P_2$ oraz $P_2\,P_3\,P_4$. Natomiast rozsunięcie węzłów środkowych do skraju odcinka (wariant a) prowadzi do nieciągłego zachowania funkcji interpolującej: krzywa nagle przestaje zależeć od węzłów skrajnych.



Rys. 10. Równomierny rozkład punktów charakterystycznych (z wyjątkiem skrajnych punktów wielokrotnych) powoduje, że krzywą giętą drugiego rzędu łatwo można rozbić na kawałki, dające się wyrazić za pomocą wielomianów Bernsztejna stopnia 2: niech $P_{1,2}$ i $P_{2,3}$ oznaczają odpowiednio środki odcinków $P_1\,P_2$ i $P_2\,P_3$; niezależna interpolacja trójek węzłów $P_0\,P_1\,P_{1,2}$, $P_{1,2}\,P_2\,P_{2,3}$ oraz $P_{2,3}\,P_3\,P_4$ za pomocą krzywych stopnia drugiego w bazie wielomianów Bernsztejna da dokładnie ten sam wynik.

$i+2, \ldots, j-1$) punkty $P_{k,k+1} = \frac{1}{2}(P_k + P_{k+1})$ i przyjmując, że $s(P_{k,k+1}) = n$ (por. rys. 10). Można też pójść w przeciwnym kierunku i dla dowolnego ciągu, którego skrajne węzły leżą na krzywej skonstruować odpowiedni ciąg punktów charakterystycznych, aczkolwiek jest mało prawdopodobne, żeby w ten właśnie sposób działała maszyneria truetype'owa.

Reasumując, z praktycznego punktu widzenia odpowiedź na postawione w tytule pytanie sprowadza się do konstatacji, że krzywe gięte drugiego dają się przedstawić w bazie wielomianów Bernsztejna drugiego stopnia, a zatem dają się wyrazić za pomocą odpowiednio „przyrządzonych” krzywych Béziera (rys. 3). Konwersja w drugą stronę nie jest już taka prosta: wyrażenie krzywych Béziera za pomocą krzywych giętych drugiego rzędu to zaawansowany problem numeryczny, nie posiadający jednoznacznego rozwiązania.

Tematy pominięte. Lista tematów wiążących się mniej lub bardziej z problemem interpolacji, a pominiętych w niniejszym opracowaniu, jest całkiem spora. Obejmuje m.in. wspomnianą aproksymację krzywych wyższego rzędu krzywymi niższego rzędu, badanie przebiegu krzywych Béziera (np. punktów osobliwych, takich jak samoprzecięcia czy ostrza), przedłużanie funkcji interpolującej poza odcinek $[0, 1]$, dyskretyzacja krzywych (tzn. zamiana na postać pikselową), znajdowanie wnętrza krzywej zamkniętej, znajdowanie „estetycznie zadowalającej” krzywej przechodzącej przez zadane punkty, czy algorytm de Boora, będący uogólnieniem algorytmu de Casteljau na krzywe gięte. Choćby pobieżne omówienie tych spraw spowodowałoby, że tekst rozrósłby się nieprzyzwoicie, a i tak jest dłuższy, niż można by się spodziewać.

◇ Bogusław Jackowski, BOP s.c.
B.Jackowski@GUST.org.pl