Prace Naukowe Instytutu Cybernetyki Technicznej
Politechniki Wrocławskiej 104

Seria:
Monografie 28

Ewa Skubalska-Rafajłowicz

Krzywe wypełniające w rozwiązywaniu wielowymiarowych problemów decyzyjnych

Oficyna Wydawnicza Politechniki Wrocławskiej · Wrocław 2001

Recenzenci
Józef KORBICZ
Leszek RUTKOWSKI

Opracowanie redakcyjne
Aleksandra WAWRZYNKOWSKA

OFICYNA WYDAWNICZA POLITECHNIKI WROCŁAWSKIEJ
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław

ISSN 0324-9786

Drukarnia Oficyny Wydawniczej Politechniki Wrocławskiej. Zam. nr 795/2001.

Prace Naukowe Instytutu Cybernetyki Technicznej
Nr 104 Politechniki Wrocławskiej Nr 104
Monografie Nr 28 2001

krzywe wypełniające,
problemy decyzyjne, redukcja wymiaru,
rozpoznawanie, sterowanie jakością

Ewa SKUBALSKA-RAFAJŁOWICZ *

KRZYWE WYPEŁNIAJĄCE W ROZWIĄZYWANIU WIELOWYMIAROWYCH PROBLEMÓW DECYZYJNYCH

W monografii przedstawiono metodykę konstruowania i badania algorytmów decyzyjnych, która jest nowym podejściem do problemów przetwarzania i podejmowania decyzji na podstawie wielowymiarowych obserwacji. Polega ona na transformacji danych do postaci jednowymiarowej za pomocą quasi-odwrotności dobrze dobranej krzywej wypełniającej, a następnie na rozwiązaniu jednowymiarowego problemu decyzyjnego. W efekcie transformacji uzyskuje się redukcję wymiaru problemu i jego znaczącą kompresję, bez utraty istotnych informacji przestrzennych zawartych w danych wielowymiarowych. Prowadzi to do możliwości konstruowania szybkich algorytmów podejmowania decyzji, które mogą działać na bieżąco, na podstawie aktualnie uzyskiwanych obserwacji. Podejście to rozwija kierunki badań prowadzone obecnie w automatyce, polegające na projektowaniu elastycznych systemów, łączących konwencjonalne techniki z różnorodnymi metodami uczenia, które wykorzystują zgromadzone obserwacje pomiarowe i pozwalają na korygowanie pracy systemu.

Opracowano układy równań funkcyjnych i rekurencyjne metody wyznaczania odwzorowań quasi-odwrotnych do wielowymiarowych krzywych typu Peano, Hilberta i Sierpińskiego. Wykazano, że transformacje te zachowują istotne informacje statystyczne zawarte w wielowymiarowych danych. Wyprowadzono teoretyczną

*Instytut Cybernetyki Technicznej Politechniki Wrocławskiej, ul. Janiszewskiego 11/17, 50-372 Wrocław.

zależność między wymiarami fraktalnymi danych przed i po transformacji. Udowodniono, że badana klasa krzywych wypełniających zachowuje ryzyko Bayesa dla dowolnego rozkładu obserwacji o ograniczonym nośniku. Zdefiniowano nową klasę krzywych wypełniających, które zachowują zadaną miarę probabilistyczną i wykorzystano ją w problemach kwantyzacji. Zbadano asymptotyczną wartość dystorsji wektorowych kwantyzatorów otrzymanych poprzez redukcję wymiaru obserwacji. Zdefiniowano też pojęcie powierzchni wypełniającej oraz odpowiednie odwzorowanie quasi-odwrotne oraz zbadano ich podstawowe własności. Wprowadzono pojęcie multi-karty, która jest uogólnieniem tradycyjnej karty kontrolnej i pozwala oceniać stan wielowymiarowego procesu na podstawie przekształconych, skalarnych obserwacji, zbadano też jej własności. Otrzymane rezultaty teoretyczne zastosowano w odniesieniu do wielowymiarowych problemów decyzyjnych, dotyczących rozpoznawania i monitorowania stanu procesu, diagnostyki i problemów statystycznego wykrywania zmian w procesie oraz problemów kwantyzacji. Zaowocowało to powstaniem szczegółowych metod, spełniających nałożone wymagania teoretyczne. W wielu przypadkach zaproponowano także proste obliczeniowo algorytmy heurystyczne. Podstawową cechą wszystkich opracowanych metod jest mała złożoność obliczeniowa procesu podejmowania decyzji.

Rozdział 1

Wprowadzenie

1.1 Wstęp

Współczesne trendy w automatyce skupiają się obecnie głównie na projektowaniu elastycznych systemów, w których łączy się konwencjonalne, oparte na modelach techniki, z ogromną różnorodnością metod i podejść, wśród których najczęściej wymienia się sztuczne sieci neuronowe, rozmyte systemy wnioskowania, algorytmy genetyczne i ewolucyjne, a także systemy automatycznego uczenia, systemy ekspertowe, systemy diagnostyki, statystyczną kontrolę jakości i monitorowania procesów itd. [194], [205], [85], [86], [208], [30]. Tradycyjne metody koncentrują się głównie na dynamice systemu, wyborze zmiennych pomiarowych, ustalaniu struktury systemu i wyborze algorytmów sterowania, ignorując równocześnie ogromną ilość danych generowanych przez systemy pomiarowe. Dane te są niezastąpionym źródłem informacji pozwalającym na ulepszanie systemu [194].

W sytuacji, gdy złożoność problemu lub brak wiedzy a priori na jego temat uniemożliwiają otrzymanie satysfakcjonujących rozwiązań, pojawia się potrzeba wykorzystania metod uczenia opartych na bazie zebranych doświadczeń. Wymaga to gromadzenia i przetwarzania dużych ilości informacji. O ile w stosunku do wstępnego przetwarzania informacji nie stawia się nadmiernych ograniczeń czasowych, o tyle sam proces podejmowania bieżącej decyzji, jeśli ma być użytecznym elementem systemu automatyki, powinien trwać możliwie krótko.

W niniejszej monografii zaproponowano i zbadano nowe podejście do problemów przetwarzania i podejmowania decyzji [31], [57], [101] na podstawie wielowymiarowych obserwacji. Podejście to polega na transformacji wielowymiarowych danych do postaci jednowymiarowej za pomocą quasi-odwrotności dobrze dobranej krzywej wypełniającej. Transformacje te zachowują istotne informacje statystyczne zawarte w wielowymiarowych danych. W ostatecznym rezultacie, po procesie wstępnego przetwarzania, które może dotyczyć zarówno oryginalnych

wielowymiarowych danych, jak i danych przetransformowanych do postaci skalarnej, otrzymujemy szybkie algorytmy podejmowania decyzji, które mogą być wykonywane na bieżąco, z uwzględnieniem aktualnie uzyskiwanych obserwacji.

Monografia ta zawiera propozycję metodyki konstruowania i badania algorytmów decyzyjnych, która opiera się na transformacji obserwacji przez quasi-odwrotność krzywej wypełniającej, rozwiązaniu jednowymiarowego problemu decyzyjnego i opracowaniu szybkiej metody podjęcia decyzji. Metodyka ta oparta jest na publikacjach autorki, które zostały tu usystematyzowane i uzupełnione nowymi wynikami.

1.2 Notka historyczna na temat krzywych wypełniających

Rozważane tu krzywe wypełniające są ciągłym odwzorowaniem przekształcającym odcinek jednostkowy na d-wymiarową kostkę jednostkową $(I_d, d \leq \infty)$. Oznacza to, że krzywa wypełniająca przechodzi co najmniej raz przez każdy punkt kostki I_d.

Krzywe wypełniające zostały po raz pierwszy opisane przez G. Peano w roku 1890 [121], a następnie przez Hilberta [72] oraz W. Sierpińskiego [145]. Stanowią one dowód na możliwość istnienia ciągłego odwzorowania przestrzeni mniej wymiarowej w przestrzeń o większym wymiarze. Odwzorowanie takie nie może być jednak homeomorfizmem, co wynika ze znanego w topologii twierdzenia o niezmienniczości wymiaru przestrzeni [43]). W konsekwencji, także żadna krzywa wypełniająca nie może być odwzorowaniem wzajemnie jednoznacznym.

Krzywymi wypełniającymi płaszczyznę zaczęto się zajmować już pod koniec XIX wieku. Prace te zapoczątkował G. Peano [121]. Przez długi czas nie interesowano się w zasadzie krzywymi wypełniającymi przestrzeń więcej niż dwuwymiarową, mimo że wiedziano o istnieniu wielowymiarowych krzywych (popatrz np. praca Steinhausa z 1936 r. [137]). Sam G.Peano podał, oprócz konstrukcji krzywej wypełniającej kwadrat, także konstrukcję krzywej wypełniającej kostkę trójwymiarową.

Pod koniec lat 60. Butz zajmował się generowaniem wielowymiarowych krzywych Peano [23] i Hilberta [24], [25]. W 1980 roku S.C. Milne [110] pokazał, jak uogólnić dwuwymiarową krzywą Peano [121] do krzywej w przestrzeni d–wymiarowej. Milne udowodnił również, że zdefiniowana przez niego krzywa zachowuje miarę Lebesgue'a i spełnia warunek Höldera z wykładnikiem równym $1/d$, gdzie d oznacza wymiar wypełnianej kostki.

Krzywe wypełniające traktowano jako patologiczne obiekty podważające utarte wyobrażenia dotyczące pojęcia krzywej i wymiaru przestrzeni. Tematyka krzy-

wych wypełniających odżyła wraz z pojawieniem się i rozwojem pojęcia fraktala [106], [48], [7]. Obecnie wiemy, że krzywe wypełniające mogą być traktowane jako atraktory pewnych abstrakcyjnych systemów dynamicznych.

1.3 Obszary zastosowań krzywych wypełniających

Rola krzywych sprowadza się do redukcji wymiaru przestrzeni danych przy zachowaniu ich podstawowych własności statystycznych. Obszar zastosowań krzywych wypełniających jest dość szeroki. W latach trzydziestych krzywe Peano były stosowane w teorii całkowania w przestrzeniach wielowymiarowych [137], [110], [130]. Krzywe wypełniające, fraktalne w swej naturze [7], [48], mają własności statystyczne pozwalające na zastosowania w różnych dziedzinach obliczeń.

Obecnie krzywe wypełniające, traktowane raczej jako krzywe skanujące, są stosowane głównie w przetwarzaniu obrazów, w szczególności do ich skanowania, kodowania [1], [5], [119], [182], [97], [126] i przetwarzania [197],[69],[147], [28], [74], [95], [123], [129], kwantyzacji wektorowej [126], [164], kompresji obrazów w przestrzeni i kolorze [197], [119], [172], [89]. Należy zauważyć, że nie każda rodzina krzywych skanujących wielowymiarową przestrzeń jest zbieżna [122]. W związku z tym, nie każda taka rodzina definiuje krzywą wypełniającą. Przykładem tego typu rodzin krzywych mogą być rodziny krzywych alfa-gęstych [112].

Krzywe skanujące mogą być stosowane nie tylko w przetwarzaniu obrazów, ale także w rozwiązywaniu zadań optymalizacji [124]; ich efektywność jest jednak ograniczona do ustalonej struktury i dokładności danych. Krzywe skanujące nie wykazują wielu własności asymptotycznych, którymi cechują się krzywe wypełniające.

Kolejnym znanym obszarem zastosowań krzywych wypełniających jest optymalizacja. W przypadku optymalizacji kombinatorycznej najczęściej krzywe stosowane są do heurystycznego rozwiązywania problemu komiwojażera z odległościami euklidesowymi i problemów pokrewnych [8], [124], [178]. Wśród krzywych wypełniających kwadrat szczególne zastosowanie w rozwiązywaniu tego typu zadań znalazła krzywa Sierpińskiego [145]. Podejście to zostało uogólnione dla przypadku wielowymiarowego problemu komiwojażera w pracach autorki [156], [162].

O możliwości zastosowania krzywych wypełniających w odniesieniu do zadań wielowymiarowej optymalizacji ciągłej wspomina się w pracach [23], [184],[143]. Poza skrótowymi wzmiankami prace te nie zawierają jednak w tym aspekcie żadnych konkretnych wyników. Jest to o tyle zrozumiałe, że we wspomnianych problemach optymalizacyjnych krzywa wypełniająca proponowana jest jedynie jako formalne narzędzie zamiany zmiennych. W nurcie tym mieszczą się również prace współautorskie [175], [128]. Należy zauważyć, że po takiej zamianie zmiennych

nowa funkcja celu nie jest funkcją różniczkowalną (poza obszarami, gdzie jest funkcją stałą). Jak dotąd nie badano zastosowania do transformacji danych krzywych typu Lebesgue'a [137], które wprawdzie nie zachowują miary Lebesgue'a, lecz są prawie wszędzie różniczkowalne (w przestrzeni jednowymiarowej).

1.4 Omówienie tematyki niniejszej monografii

Kluczowym elementem opracowanej w niniejszej monografii metodologii rozwiązywania wielowymiarowych problemów decyzyjnych jest zastosowanie dobrze zdefiniowanej transformacji quasi-odwrotnej do odwzorowania w postaci krzywej wypełniającej w celu przekształcenia zbioru wielowymiarowych danych, które stanowią punkt wyjścia w procesie decyzyjnym, w ciąg danych jednowymiarowych zawartych w odcinku jednostkowym. Złożoność obliczeniowa takiej transformacji danych powinna być jak najmniejsza. W naszym przypadku jest ona liniowa ze względu na wymiar problemu d. Transformacja każdego elementu zbioru danych za pomocą quasi-odwrotności krzywej wypełniającej może odbywać się niezależnie, w dowolnym momencie czasowym i nie wymaga konstrukcji całej krzywej wypełniającej. Transformacja quasi-odwrotna pozwala uporządkować liniowo dane z przestrzeni wielowymiarowej. Jest to znacznie bardziej interesująca transformacja niż odpowiadająca jej pierwotna krzywa, gdyż może być wykorzystana w wielu bardzo różnorodnych zastosowaniach. W jej efekcie uzyskujemy redukcję wymiaru problemu, a w konsekwencji jego znaczącą kompresję, bez utraty istotnych informacji przestrzennych zawartych w danych wielowymiarowych. W odwzorowaniach bazujących na krzywej wypełniającej istotny jest fakt, że położone blisko siebie punkty z odcinka I_1 przekształcane są na punkty blisko siebie położone w I_d. W związku z tym, w transformacji quasi-odwrotnej, punkty leżące blisko siebie na odcinku są obrazami punktów blisko siebie położonych w przestrzeni wielowymiarowej.

W żadnym wypadku nie twierdzimy, iż odwzorowania bazujące na krzywych wypełniających stanowią łatwy środek przeciwdziałający ,,przekleństwu wielowymiarowości" Bellmana. Jednowymiarowe dane (dane po przetransformowaniu na odcinek) muszą być przechowywane z dostatecznie dużą dokładnością, która pozwoli na ich rozróżnianie. Należy dodać, że transformacja oparta na krzywej wypełniającej jest transformacją nieliniową i nie jest inwariantna ze względu na liniowe przekształcenia oryginalnej przestrzeni danych.

W rozdziale 2 przedstawiona została metoda reprezentacji krzywych wypełniających w postaci układu równań funkcyjnych [26], który ma jednoznaczne rozwiązanie. Przedstawiono rekurencyjne algorytmy wyznaczania wartości lokalnych współrzędnych danej krzywej, traktowanych jako wartość dwuwymiarowej

funkcji określonej dla danego argumentu $t \in [0,1]$. Algorytmy te nie wymagają tworzenia jakiegokolwiek przybliżenia całej krzywej, działają bowiem w sposób lokalny. Ponadto w rozdziale 2 przeanalizowano podstawowe własności trzech najbardziej znanych dwuwymiarowych krzywych wypełniających: krzywej Peano, Hilberta i Sierpińskiego. Autorka podała najmniejsze, optymalne, wartości stałej występujące w odpowiednim warunku Höldera (por. (1.1)) w przypadku krzywej Hilberta oraz krzywej Peano. Stała z warunku Höldera określa stopień zachowywania bliskości przez poszczególne krzywe, co ma istotne znaczenie dla korzystania z krzywych wypełniających w problemach decyzyjnych.

W rozdziale 3 przedstawiono przegląd różnych sposobów definiowania wielowymiarowych krzywych wypełniających. Samopodobną krzywą wypełniającą kostkę I_d można traktować jako obiekt fraktalny. Należy też zauważyć, że wszystkie znane krzywe wypełniające posiadają cechę samopodobieństwa. W związku z tym omówiono jedną z najbardziej popularnych technik definiowania obiektów fraktalnych, systemy iterowanych odwzorowań zwężających, tzw. IFS, jako narzędzie do generowania krzywych wypełniających. W szczególności, przedstawiono metodę konstrukcji wielowymiarowej krzywej Sierpińskiego za pomocą odpowiedniego układu IFS. Proponowana konstrukcja jest oryginalnym wkładem autorki.

Cechą charakterystyczną proponowanej metody generowania wielowymiarowej krzywej Sierpińskiego jest bardzo precyzyjny wybór zbioru początkowego, który podlega dalszemu iterowaniu poprzez IFS. Pozwala to na wyznaczanie w skończonej liczbie kroków dokładnych wartości atraktora IFS, którym jest krzywa wypełniająca.

W rozdziale 3 wskazano także na związki odpowiednich systemów iterowanych odwzorowań z układami równań funkcyjnych definiującymi te same krzywe wypełniające. Na koniec przedyskutowano metodę definiowania wielowymiarowych krzywych wypełniających, polegającą na składaniu odwzorowań dwuwymiarowych. W twierdzeniu 3.4.1 wskazano na słabą przydatność praktyczną tego typu metod w porównaniu do metod bezpośredniej konstrukcji wielowymiarowych krzywych, których przykładem mogą być konstrukcje wykorzystujące układy iterowanych odwzorowań zwężających lub rozwijane w następnym rozdziale metody opisu wielowymiarowych krzywych w postaci układu równań funkcyjnych.

W rozdziale 4 zaproponowano oryginalne opisy definicyjne wielowymiarowych krzywych wypełniających kostkę I_d, podane w postaci odpowiednich układów równań funkcyjnych. Rozwiązaniem tych układów równań są odwzorowania w postaci krzywej wypełniającej. Korzystając z omówionej metodyki definiowania krzywych wypełniających, podano oryginalne uogólnienia dwuwymiarowych krzywych Hilberta, Peano i Sierpińskiego do postaci wielowymiarowej oraz zbadano ich własności. W pracy posłużono się układami równań funkcyjnych, które oparte są bezpośrednio na postulowanych własnościach geometrycznych poszcze-

gólnych krzywych. W każdym z przypadków wyjściowy układ równań funkcyjnych został przekształcony do równoważnej postaci, umożliwiającej jego efektywne rozwiązanie za pomocą odpowiedniej procedury rekurencyjnej.

W rozdziale 4 zbadano również podstawowe własności zaproponowanych odwzorowań. Pokazane zostały też odpowiednie szczegółowe sformułowania warunku Höldera spełnianego przez poszczególne krzywe wielowymiarowe. Wskazano również, w jaki sposób należy wybrać przekształcenie układu równań funkcyjnych definiujących daną krzywą, tak by uzyskać postać pozwalającą efektywnie wyznaczać wartości odwzorowania quasi-odwrotnego Ψ. Szczegółowo omówiono proces definiowania quasi-odwrotności krzywej Sierpińskiego.

W rozdziale 4.5 zebrano własności krzywych wypełniających, które są zdaniem autorki istotne z punktu widzenia zastosowań w problemach decyzyjnych. W konsekwencji zdefiniowano klasę krzywych wypełniających, które umożliwiają efektywne użycie transformacji opartej na quasi-odwrotności krzywej wypełniającej w rozwiązywaniu wielowymiarowych problemów decyzyjnych.

W rozdziale 5 zbadano problem zmiany wymiaru wielowymiarowych danych po ich transformacji przy użyciu quasi-odwrotności krzywej wypełniającej. Badanym wymiarem był popularny wymiar fraktalny, nazywany wymiarem pudełkowym. W szczególności, wyprowadzono teoretyczną zależność między wymiarem pudełkowym wielowymiarowych danych a wymiarem pudełkowym danych przetransformowanych przy użyciu krzywej wypełniającej. Wyniki tych badań stały się podstawą do zaproponowania nowej, łatwiejszej obliczeniowo, metody oceny wymiaru fraktalnego obiektów wielowymiarowych.

W rozdziale 6 omówiono metody kwantyzacji, które łączą transformację wielowymiarowych danych za pomocą quasi-odwrotności wybranej krzywej wypełniającej ze skalarną kwantyzacją w odniesieniu do danych przetransformowanych na odcinek I_1. Rozwiązanie problemu kwantyzacji jest istotnym składnikiem wielu algorytmów decyzyjnych. Pokazano, że zastosowanie różnych wariantów algorytmów uczenia na bazie danych skalarnych umożliwia modyfikowanie kryterium kwantyzacji i w konsekwencji – własności asymptotycznych otrzymywanych kwantyzatorów (przy liczbie kwantyzatorów dążącej do nieskończoności). Pozwala to na kształtowanie rozkładu gęstości kwantyzatorów na odcinku I_1, a także, co wynika z własności odwzorowania F i jego quasi-odwrotności Ψ, umożliwia również wpływ na rozkład gęstości odpowiednich kwantyzatorów, wyznaczonych poprzez przetransformowanie za pomocą krzywej wypełniającej rozwiązania z odcinka I_1 do przestrzeni I_d. Zaproponowano nową klasę krzywych wypełniających, które zachowują zadaną miarę probabilistyczną. Może być ona teoretycznym narzędziem, mającym zastosowanie w kwantyzacji wielowymiarowych danych. W rozdziale 6 podano także algorytm aproksymacji tego typu krzywych, w sytuacji, gdy dane są tylko niezależne obserwacje wielowymiarowej zmiennej

losowej o nieznanym rozkładzie. Otrzymane odwzorowanie daje możliwość szybkiego wyznaczania zbioru punktów kwantyzacji, których gęstość rozkładu jest taka sama jak gęstość rozkładu danych wejściowych.

Podstawową klasą problemów decyzyjnych będących przedmiotem badań niniejszej monografii są problemy rozpoznawania. W rozdziale 7 zbadano własności nowych algorytmów rozpoznawania, których wspólną cechą było wykorzystanie krzywych wypełniających, a dokładnie ich quasi-odwrotności, jako narzędzia do transformacji wielowymiarowych danych do postaci jednowymiarowej. Podstawowym rezultatem zawartym w rozdziale 7 jest pokazanie, że krzywa wypełniająca, która zachowuje miarę Lebesgue'a i jest odwracalna prawie wszędzie (z dokładnością do zbioru miary zero) może być wykorzystana (a dokładnie jej quasi-odwrotność) jako odwzorowanie, które zachowuje ryzyko Bayesa dla dowolnego rozkładu danych o nośniku zawartym w wielowymiarowej kostce I_d. W rozdziale tym zdefiniowano także i zbadano własności separujące krzywych wypełniających w odniesieniu do różnych typów nośników rozkładów.

Ze względu na zachowywanie ryzyka Bayesa przez transformacje wykorzystujące krzywe wypełniające, możemy w odniesieniu do przetransformowanych danych użyć dowolnego, dostatecznie wrażliwego klasyfikatora, nie tracąc nic z informacji statystycznych zawartych w oryginalnych danych. Jeżeli reguła klasyfikacji w jednym wymiarze jest asymptotycznie optymalna (zbieżna do ryzyka Bayesa), to klasyfikator taki zastosowany w odniesieniu do przetransformowanych danych jest także asymptotycznie optymalny.

Podstawową cechą rozważanych algorytmów rozpoznawania jest szybkość samego procesu podejmowania decyzji. W końcowym efekcie konstrukcji klasyfikatora otrzymujemy dyskryminacyjny podział odcinka I_1 na m pododcinków odpowiadających poszczególnym klasom, przy czym m jest nie większe niż długość ciągu uczącego.

W rozdziale 7 zdefiniowano odwzorowanie analogiczne do krzywej wypełniającej, które przekształca kwadrat jednostkowy $I_2 = [0,1] \times [0,1]$ w wielowymiarową kostkę I_d oraz odpowiednie odwzorowanie quasi-odwrotne. Pokazano, że wprowadzone odwzorowanie spełnia warunek Höldera. Odwzorowanie to można wykorzystać bezpośrednio do reprezentacji wielowymiarowych danych na płaszczyźnie. Pokazano zastosowanie tego typu wizualizacji w rozwiązywaniu problemów rozpoznawania.

Przedmiotem badań w rozdziale 8 był problem szybkiego wykrywania zmian zachodzących w procesie stochastycznym [199] na podstawie sekwencji niezależnych obserwacji. Problem ten ma wiele istotnych zastosowań, poczynając od sterowania i kontroli jakości przemysłowych procesów produkcyjnych [114], [9], [104], [86], poprzez automatyczne wykrywanie uszkodzeń w systemie dynamicznym [195], [85], a skończywszy na uaktualnianiu współczynników w adaptacyjnych al-

gorytmach sterowania [140]. Podstawowym zadaniem umożliwiającym efektywne sterowanie procesem jest natomiast jak najszybsze wykrycie momentu zmiany.

W rozdziale 8 zaproponowano nową metodykę wykrywania zmian w monitorowanym procesie. Krzywe wypełniające stanowią wygodne narzędzie do transformacji wielowymiarowych danych do postaci jednowymiarowej. Transformacja ta jest niezmiennicza względem miary probabilistycznej w tym sensie, że zachowuje prawdopodobieństwa odpowiadających sobie zdarzeń w przestrzeni wielowymiarowej i po transformacji na odcinek I_1. W rozdziale tym zdefiniowano pojęcie multi-karty, która jest uogólnieniem tradycyjnej karty kontrolnej i pozwala oceniać stan wielowymiarowego procesu na podstawie przekształconych, skalarnych obserwacji.

Proponowane podejście można stosować nie tylko w odniesieniu do oryginalnych obserwacji wielowymiarowego procesu, ale także do danych przetworzonych (na przykład wielowymiarowych residuów itd.). Zbadano własności teoretyczne multi-karty wyznaczonej na podstawie histogramu. Jeśli gęstość rozkładu na odcinku estymowana jest za pomocą histogramu o szerokości przedziałów h_n, zależnej od liczby obserwacji n, w taki sposób, że $h_n \to 0$ w odpowiednim tempie, to przedziały ufności multi-karty empirycznej mają asymptotycznie te same prawdopodobieństwa i łączna ich długość jest taka sama, jak długość przedziałów wyznaczonych przy znajomości teoretycznego rozkładu.

Oprócz powyższego algorytmu, posiadającego pełną podbudowę teoretyczną, zaproponowano również dwa algorytmy heurystyczne konstrukcji multi-karty, wykorzystujące algorytmy uczenia stosowane w samoorganizujących sieciach neuronowych [83], [165].

Algorytm uczenia typu Kohonena pozwala na efektywne wyznaczenie granic decyzji w multi-karcie oraz na adaptacyjne, prowadzone na bieżąco, na podstawie aktualnych obserwacji, modyfikacje granic decyzji wyrażonych w postaci przedziałów multi-karty. Zbadany w niniejszym rozdziale algorytm (typu Kohonena) nie ma wprawdzie w pełni określonych własności asymptotycznych, jednakże działa na bieżąco i w konsekwencji nie wymaga przechowywania i łącznego przetwarzania dużej ilości danych. Ze względu na brak założeń dotyczących postaci rozkładów analizowane algorytmy mają charakter nieparametryczny. W konsekwencji, wymagania czynione w odniesieniu do monitorowanego procesu dotyczą tylko istnienia odpowiednich rozkładów prawdopodobieństwa oraz ewentualnie niezależności kolejnych obserwacji procesu.

Stosowane do tej pory metody statystyczne wymagają, by rozkład prawdopodobieństwa monitorowanego procesu był znany i na ogół założenia, że rozkład ten jest rozkładem normalnym. Zaproponowane tu podejście stanowi atrakcyjne narzędzie w przypadku, gdy wiadomo, że sterowany proces nie spełnia tego założenia. Transformacja wielowymiarowych danych do postaci jednowymiarowej

otwiera nowe możliwości w konstruowaniu algorytmów jakościowego diagnozowania stanu procesu.

Rezultaty przedstawione w rozdziałach 2 –8 niniejszej monografii są wynikami własnych badań autorki, usystematyzowanymi i pokazanymi na tle współczesnego stanu wiedzy. Część z tych rezultatów była wcześniej publikowana w pracach [173], [174], [163], [164], [165], [167], [168], [170], [171], [172] z lat 1995–2001. W tym samym czasie powstały także inne prace, w których przedstawiono zastosowania krzywych wypełniających [128], [175], [89].

Wcześniej autorka zajmowała się problemami optymalizacyjnymi i problemami decyzyjnymi, głównie w systemach transportowych [148], [149], [65], [66], [151], [150], [67], [152], [155], oraz problemami dokładności i szybkości algorytmów [153], [154], [127], [157]. W trakcie prowadzonych badań autorka zetknęła się z heurystyczną metodą rozwiązywania problemu komiwojażera i pokrewnych problemów transportowych za pomocą krzywych wypełniających [124], [8]. Własne prace na temat wielowymiarowego problemu komiwojażera [156], [162] spowodowały zainteresowanie wykorzystaniem własności krzywych wypełniających w rozwiązywaniu innych problemów decyzyjnych, szczególnie dotyczących rozpoznawania i monitorowania stanu procesu, diagnostyki oraz statystycznego sterowania jakością. Podsumowaniem prowadzonych w tej dziedzinie badań jest niniejsza monografia.

1.5 Podstawowe definicje i twierdzenia na temat krzywych wypełniających

Autorka ma świadomość, że tematyka krzywych wypełniających jest stosunkowo mało znana. W niniejszym podrozdziale zostały więc zebrane podstawowe klasyczne wyniki dotyczące krzywych wypełniających.

Definicja 1.1 *Krzywą wypełniającą F nazywamy ciągłe odwzorowanie odcinka w przestrzeń d-wymiarową $F : I \to R^d$, którego obraz $F(I)$ zawiera d-wymiarową kulę.*

Twierdzenie Hahna–Mazurkiewicza [42] określa rodzaj przestrzeni, którą można wypełnić krzywą.

Twierdzenie 1.5.1 *Przestrzeń topologiczna Hausdorffa może zostać w całości wypełniona przez ciągłą krzywą wtedy i tylko wtedy, gdy przestrzeń ta jest zwarta, spójna, lokalnie spójna i metryzowalna (czyli jest continuum Peano).*

W niniejszej pracy będziemy się zajmować tylko takimi krzywymi wypełniającymi, które odwzorowują odcinek w d-wymiarową kostkę I_d. Bez straty ogólności

można przyjąć, że rozpatrywany odcinek jest odcinkiem jednostkowym $I_1 = [0, 1]$, a d-wymiarowa kostka jest również kostką jednostkową

$$I_d = I_1 \times I_1 \times \ldots \times I_1 .$$

Krzywa wypełniająca $F : I_1 \rightarrow I_d$ jest surjekcją, czyli odwzorowaniem na cały obszar I_d.

Z twierdzenia Cantora [93], [43] wynika, że istnieje różnowartościowe odwzorowanie między kostkami o różnych wymiarach $I_n \rightarrow I_m \quad n \neq m$, przy n, m skończonych. W szczególności, istnieje różnowartościowe odwzorowanie odcinka I_1 w d-wymiarową kostkę I_d, odwzorowanie to nie może być jednak ciągłe [137]. Jest to konsekwencją twierdzenia Brouwera (por.[43]).

Twierdzenie 1.5.2 (*O niezmienniczości wymiaru*) *Nie istnieje homeomorfizm pomiędzy przestrzeniami R^m i R^n, jeśli tylko $n \neq m$.*

W rozpatrywanym tu przypadku odwzorowań z I_1 w I_d możemy z homeomorfizmem utożsamiać ciągłą bijekcję, gdyż ciągła bijekcja $f : X \rightarrow Y$ zwartej przestrzeni X na przestrzeń Hausdorffa Y jest homeomorfizmem.

Reasumując, w przypadku rozpatrywania różnych odwzorowań pomiędzy odcinkiem I_1 a wielowymiarową kostką I_d możemy mieć do czynienia z trzema następującymi przypadkami:

- odwzorowanie jest ciągłe i różnowartościowe, lecz nie jest surjekcją (nie jest odwzorowaniem na całą kostkę I_d), a obrazem tego odwzorowania jest krzywa (właściwa, czyli nieprzecinająca się) Jordana,
- odwzorowanie jest bijekcją, nie jest jednak odwzorowaniem ciągłym (twierdzenie Netto [137]),
- odwzorowanie jest krzywą wypełniającą, czyli ciągłą surjekcją, nie jest jednak różnowartościowe (nie jest injekcją). W konsekwencji nie jest to krzywa właściwa.

Wielowymiarowa kostka może więc posiadać przedstawienie parametryczne na odcinku, nie jest ono jednak różniczkowalne [137], w przeciwieństwie do klasycznych krzywych [93].

Istnienie krzywej wypełniającej można pokazać korzystając z następującej konstrukcji [93], którą przedstawimy w tym miejscu jedynie schematycznie.

Określmy ciąg nieskończony funkcji ciągłych $f_1, f_2, \ldots, f_n \ldots$ Każda z tych funkcji jest zwykłą krzywą Jordana i stanowi kolejną aproksymację idealnej krzywej wypełniającej. Załóżmy, że w n-tej iteracji kostka I_d została podzielona na 2^{dn}

pokrywających ją podkostek o boku 2^{-n}. Krzywa $f_n(I_1)$ powinna mieć punkty wspólne z każdą z tych podkostek. Łatwo dowieść w tym przypadku, że ciąg $f_1, f_2, \ldots, f_n \ldots$ jest jednostajnie zbieżny, a zatem jego granica jest funkcją ciągłą. Oznaczmy tę funkcję przez F. Każdy punkt kostki jest wartością funkcji F, czyli $F(I_1) = I_d$, gdyż $\overline{\bigcup_n f_n(I_1)} = I_d$.

Naszkicowany powyżej sposób konstrukcji krzywej wypełniającej wiąże się z sekwencyjnym podziałem wypełnianej przestrzeni wielowymiarowej na elementarne obszary, najczęściej tego samego kształtu i objętości. W konsekwencji, zdefiniowane zostaje wzajemnie jednoznaczne odwzorowanie pomiędzy wspomnianymi wcześniej elementarnymi obszarami a odpowiednimi pododcinkami odcinka jednostkowego I_1. W odwzorowaniu tym sąsiednie pododcinki są powiązane z sąsiadującymi ze sobą obszarami [111], [110]. W taki sposób mogą być definiowane klasyczne krzywe wypełniające: krzywa Peano [121], krzywa Hilberta [72] i krzywa Sierpińskiego [145].

Krzywe wypełniające nie są zwykłymi krzywymi Jordana, ponieważ przez pewne punkty wypełnianej przestrzeni przechodzą więcej niż jeden raz, krzywe te są samoprzecinające. By uniknąć nieporozumień terminologicznych, należy jednak zauważyć, że sam Jordan definiował krzywą jako obraz odcinka jednostkowego w ciągłym odwzorowaniu, które niekoniecznie musi być różnowartościowe [137]. W tym sensie określenia krzywa Jordana w odniesieniu do krzywej wypełniającej używał Sierpiński [145].

Innego typu dowód istnienia krzywej wypełniającej kostkę podał w 1912 roku Lebesgue (por. [137]). Krzywą wypełniającą można otrzymać korzystając z tego, iż dowolna (ze względu na wymiar d) kostka I_d jest ciągłym obrazem zbioru Cantora ([93] tw. 2, rozdz. XVI9). Dowód istnienia krzywej wypełniającej kostkę I_d wykorzystujący powyższy fakt można znaleźć także w książce [43].

Przypomnijmy, że zbiór Cantora jest zawarty w odcinku jednostkowym I_1 i powstaje z tego odcinka poprzez sukcesywne odrzucanie środkowej jednej trzeciej każdego pododcinka. Każdy element zbioru Cantora daje się przedstawić jako suma $2t_1/3+2t_2/3^2+\ldots+2t_j/3^j \ \ldots$, gdzie t_j jest równe 0 lub 1. W konsekwencji, do zbioru Cantora należą wszystkie liczby, których trójkowe rozwinięcia zawierają tylko cyfry 0 i 2. W skróconej notacji można liczby te zapisać w postaci: $\{O_3(2t_1)(2t_2)(2t_3)\ldots | t_j \in \{0,1\}\}$ [137].

Zbiór Cantora C jest homeomorficzny z potęgą nieskończoną zbioru złożonego z dwu elementów $B = \{0,1\} \times \{0,1\} \times \ldots$, natomiast kostka Cantora $C^n = C \times C \times C$ jest homeomorficzna ze zbiorem Cantora dla dowolnego n ([93] tw. 3, rozdz. XVI8).

Dowolnie wymiarowa kostka I_d, łącznie z kostką Hilberta I_∞, jest ciągłym obrazem zbioru Cantora ([93] tw. 2, rozdz. XVI9).

Wprowadzenie funkcji Cantora pozwala na zdefiniowanie odpowiadającej powyższemu odwzorowaniu krzywej wypełniającej.

Rozszerzając w sposób liniowy odwzorowanie ze zbioru Cantora na cały odcinek, otrzymujemy krzywą wypełniającą.

Funkcja Cantora $f : C \to I_1$ jest zdefiniowana w ten sposób, że jeśli $x = \{O_3(2t_1)(2t_2)(2t_3)\ldots|t_j \in \{0,1\}\}$, to $f(x) = \{O_2(t_1t_2t_3\ldots)\}$. Funkcja ta jest ciągłą surjekcją [43] i łatwo ją rozszerzyć na cały przedział I_1. Powstaje wtedy funkcja zwana ,,schodami diabła" ([93] s. 210), która jest odwzorowaniem odcinkami stałym w podprzedziałach I_1 nie należących do zbioru Cantora.

W ten sposób powstaje krzywa wypełniająca, która w przeciwieństwie do poprzedniej konstrukcji nie zachowuje miary Lebesgue'a. Odwzorowanie to jest prawie wszędzie stałe, a w związku z tym jest prawie wszędzie różniczkowalne (nie jest różniczkowalne na zbiorze miary zero). Jeśli jednak popatrzymy na własności tego typu krzywej przez pryzmat zbioru jej wartości (obszar w kostce I_d), to stwierdzamy, że obszar różniczkowalności krzywej jest zbiorem miary zero.

Krzywe wypełniające wielowymiarową przestrzeń są klasą odwzorowań ciągłych, które spełniają warunek Höldera postaci

$$||F(t_1) - F(t_2)|| \leq \alpha|t_1 - t_2|^{\beta},\ t_1, t_2 \in I_1, \tag{1.1}$$

gdzie zarówno β jak i α są stałymi zależnymi od rodzaju krzywej oraz od wymiaru topologicznego wypełnianej przestrzeni. Wiadomo przy tym, że istnieje ścisły związek pomiędzy wymiarem topologicznym d wypełnianej kostki I_d, a maksymalną wartością wykładnika β [110], [138]. Nie istnieje bowiem d-wymiarowa krzywa wypełniająca z wykładnikiem β większym od $1/d$.

Własności krzywych wypełniających, istotne z punktu widzenia zastosowań w problemach decyzyjnych, zostaną dalej szczegółowo zbadane i przedstawione w rozdziale 4.5.

Brak jest niestety w literaturze, nie tylko polskiej ale i światowej, opracowania tworzącego jednolite podejście do tematyki krzywych wypełniających. W wielokrotnie w niniejszej monografii cytowanej książce Sagana [137], która jest nieocenionym źródłem informacji na temat historii krzywych wypełniających, nie dostrzega się w ogóle problemu istnienia odwzorowań qausi-odwrotnych (odwzorowań pełniących rolę odwrotności odwzorowania w postaci krzywej), pomija się rolę wykładnika w warunku Höldera (1.1), a w konsekwencji zadowala się samym faktem istnienia wielowymiarowych krzywych wypełniających, bez uwzględniania ich własności w tym aspekcie. Problemy te, w różnych fragmentach, zostały podjęte w pracach [23], [24], [124], [110], [178].

Rozdział 2

Krzywe wypełniające kwadrat

Krzywymi wypełniającymi płaszczyznę zaczęto się zajmować już pod koniec XIX wieku. W roku 1890 G. Peano [121], a w 1891 D. Hilbert [72] zaprezentowali krzywe przechodzące przez każdy punkt kwadratu. Trzecią, najbardziej znaną konstrukcję krzywej wypełniającej kwadrat podał w 1912 r. W. Sierpiński. Krzywe te traktowano jako patologiczne obiekty podważające utarte wyobrażenia dotyczące pojęcia krzywej i wymiaru przestrzeni.

Przez długi czas nie interesowano się krzywymi wypełniającymi przestrzeń więcej niż dwuwymiarową, chociaż znane były metody definiowania wielowymiarowych krzywych wypełniających (popatrz np. praca Steinhausa z 1936 r. [181], [137]). Należy wspomnieć, że w swej oryginalnej pracy G. Peano podał nie tylko konstrukcję krzywej wypełniającej kwadrat, lecz także krzywej wypełniającej kostkę trójwymiarową [121].

Niekiedy wszystkie krzywe wypełniające określane są mianem krzywych Peano. W niniejszej pracy nazwy krzywa Peano będziemy używać tylko w odniesieniu do konkretnej konstrukcji podanej przez Peano, a nie jako ogólnej nazwy wszystkich krzywych wypełniających.

W rozdziale tym przedstawione zostanie jednolite podejście do opisu dwuwymiarowych krzywych wypełniających. Reprezentowanie krzywej wypełniającej w postaci układu równań funkcyjnych pozwala na jednoznaczne zdefiniowanie konkretnej krzywej. Równania funkcyjne opisują charakterystyczne własności geometryczne – różnego typu symetrie oraz samopodobieństwa – poszczególnych krzywych. Inspiracją do stworzenia tej klasy definicji krzywych wypełniających była dla autorki praca W. Sierpińskiego [145]. Poza krzywą Sierpińskiego przedstawione zostaną opisy pozostałych dwu klasycznych krzywych: krzywej Hilberta i krzywej Peano. We wszystkich przypadkach zaproponowane zostaną proste algorytmy rekurencyjne, pozwalające na wyznaczanie $(x(t), y(t))$ (z zadaną dokładnością) przy ustalonej wartości argumentu $t \in I_1$. Podane zostaną także wa-

runki wyznaczenia dokładnych wartości $x(t)$ oraz $y(t)$. Ponadto w rozdziale tym przeanalizowano podstawowe własności omawianych dwuwymiarowych krzywych wypełniających. W szczególności, podano najlepsze wartości stałej występującej w odpowiednich dla każdej z krzywych postaciach warunku Höldera. W przypadku krzywej Hilberta i krzywej Peano są to, o ile autorce wiadomo, nowe rezultaty. Stała w warunku Höldera określa stopień zachowywania bliskości przez poszczególne krzywe. W następnym rozdziale proponowany tu sposób definiowania krzywych wypełniających zostanie uogólniony w odniesieniu do krzywych wielowymiarowych o wymiarze większym niż 2.

2.1 Krzywa Sierpińskiego

W 1912 roku Sierpiński [145] zdefiniował krzywą wypełniającą kwadrat $I_2 = [-1,1] \times [-1,1]$ jako odwzorowanie:

$$x(t) = f(t), \quad y(t) = f(t - 1/4), \quad 0 \leq t \leq 1, \tag{2.1}$$

w którym $f(t)$ jest jednowymiarową funkcją jednoznacznie określoną przez warunki:

$$\begin{cases} f(t) = f(-t), \\ f(t) = -f(t + 1/2), \end{cases} \quad t \in R, \tag{2.2}$$

$$2 \cdot f(t/4) + f(t + 1/8) = 1, \quad 0 \leq t \leq 1. \tag{2.3}$$

Równania (2.1) oraz (2.2), (2.3) tworzą układ równań funkcyjnych, którego jednoznacznym rozwiązaniem jest funkcja będąca krzywą wypełniającą [145]. Warunek (2.3) można zapisać, korzystając z własności (2.2), w równoważnej, łatwiejszej do bezpośredniego zastosowania postaci jako:

$$f(t) = 1/2 - 1/2 \cdot f(1/8 + 4t), \qquad 0 \leq t \leq 1/4,$$

$$f(t) = -1/2 + 1/2 \cdot f(1/8 - 4t), \qquad 1/4 \leq t \leq 1/2,$$

$$f(t) = -1/2 + 1/2 \cdot f(1/8 + 4t), \qquad 1/2 \leq t \leq 3/4,$$

$$f(t) = 1/2 - 1/2 \cdot f(1/8 - 4t), \qquad 3/4 \leq t \leq 0.$$

Ponadto z równania (2.2) wynika, że $f(t)$ jest funkcją cykliczną z okresem 1, czyli $f(t) = f(t + 1)$ (por. [145]).

W niniejszej monografii przyjęto zasadę definiowania krzywych wypełniających kostkę jednostkową $[0,1] \times \ldots \times [0,1]$. W związku z tym pokażemy dalej, że stosując odpowiednią zamianę zmiennych, łatwo jest otrzymać równoważną postać opisu krzywej Sierpińskiego wypełniającej kwadrat jednostkowy $I_2 = [0,1] \times [0,1]$. Poprzez odpowiednie przesunięcie i przeskalowanie funkcji $f(t)$ otrzymamy, w postaci układu równań (2.4), równoważny opis dwuwymiarowej krzywej Sierpińskiego w I_2.

Niech $g(t)$ będzie funkcją spełniającą warunki:

$$\begin{aligned} &g(t) = g(t+1), && t \in R, \\ &g(t) = 1 - g(t - 1/2), && t \in R, \\ &g(t) = g(4\,t)/2, && 0 \le t \le 1/8, \;\; 7/8 \le t \le 1, \\ &g(t) = 1/2 + g(-4\,t + 1/2)/2, && 1/8 \le t \le 3/8. \end{aligned} \tag{2.4}$$

Krzywa wypełniająca, zdefiniowana jako $F_g(t) = (g(t), g(t-1/4))$, $t \in I_1$, jest dwuwymiarową krzywą Sierpińskiego wypełniającą kwadrat I_2. Łatwo sprawdzić, że $F_g(t)$ jest identyczna z krzywą $(f(t/4), f(t/4 - 1/4))$ dla $t \in I_1$.

Można podać wiele innych równoważnych definicji dwuwymiarowej krzywej Sierpińskiego. Poniżej przedstawiona zostanie definicja sformułowana w postaci układu równań funkcyjnych, w których korzystamy z geometrycznych własności krzywej Sierpińskiego. Tego typu definicja dwuwymiarowej krzywej Sierpińskiego będzie w następnym rozdziale punktem wyjścia do zdefiniowania wielowymiarowej krzywej.

Twierdzenie 2.1.1 *Odwzorowanie $F_S(t) = (x_S(t), y_S(t))$, $t \in I_1$, spełniające warunki:*

$$x_S(t) = x_S(4t)/2,\; y_S(t) = y_S(4t)/2, t \in [0, 1/8], \tag{2.5}$$

$$x_S(t) = 1/2 + y_S(t - 1/8),\; y_S(t) = 1/2 - x_S(t - 1/8), t \in [1/8, 1/4], \tag{2.6}$$

$$x_S(t) = 1 - y_S(t - 1/4),\; y_S(t) = x_S(t - 1/4), t \in [1/4, 1/2], \tag{2.7}$$

$$x_S(t) = y_S(1 - t),\; y_S(t) = x_S(1 - t), t \in [1/2, 1], \tag{2.8}$$

jednoznacznie definiuje dwuwymiarową krzywą Sierpińskiego wypełniającą kwadrat I_2.

Dalsza część niniejszego podrozdziału zmierzać będzie do wykazania słuszności twierdzenia 2.1.1. Dokładniej, udowodnimy dalej, że krzywa $F_S(t)$ jest identyczna z krzywą $F_g(t)$.

Przedtem wykażemy, że układ równań (2.5)–(2.8) można przekształcić do równoważnej postaci, która prowadzi bezpośrednio do otrzymania rekurencyjnego algorytmu wyznaczania położenia punktu z kwadratu odpowiadającego ustalonemu argumentowi $t \in I_1$ (parametrowi krzywej). Wyznaczenie dokładnej wartości $F_S(t)$ jest możliwe w przypadku takich argumentów t, które mają skończone czwórkowe rozwinięcia. Najpierw zauważmy, że z $F_S(t) = F_S(4t)/2$ dla $t \in [0, 1/8]$ wynika, iż $F_S(0) = (0, 0)$.

Ustalenie punktu początkowego odwzorowania $F_S(0) = (0, 0)$ wraz z równaniami (2.5)–(2.8) pozwala na sformułowanie konstruktywnego algorytmu wyznaczania wartości F_S dla ustalonego argumentu $t \in I_1$. Podstawą algorytmu są następujące równania:

S1) $F_S(0) = (0, 0)$,

S2) $x_S(t) = x_S(4t)/2,\ y_S(t) = y_S(4t)/2,\ t \in [0, 1/8)$,

S3) $x_S(t) = 1/2 + y_S(t - 1/8),\ y_S(t) = 1/2 - x_S(t - 1/8),\ t \in [1/8, 1/4)$,

S4) $x_S(t) = 1 - y_S(t - 1/4),\ y_S(t) = x_S(t - 1/4),\ t \in [1/4, 1/2]$,

S5) $x_S(t) = y_S(1 - t),\ y_S(t) = x_S(1 - t),\ t \in (1/2, 1]$.

Równanie (S5) oznacza, że krzywa Sierpińskiego jest symetryczna względem przekątnej kwadratu.

W celu wyznaczenia konkretnej wartości $F_S(t)$ należy użyć układu równań (S1)–(S5) w formie rekurencji wstecz. Równania (S2)–(S5) różnią się od układu równań (2.5)–(2.8) tylko w punktach granicznych $t = 1/8$, $t = 1/4$, $t = 1/2$. W każdym z tych przypadków wybrane zostało tylko jedno z równań z układu (2.5)–(2.8), które jest stosowane w odniesieniu do danej wartości t. Prowadzi to do ujednoznacznienia wyboru przekształcenia i zapewnia zbieżność rekurencyjnego algorytmu opartego na (S1)–(S5). Układ równań (S1)–(S5) można rozwiązać, korzystając z prostej procedury rekurencyjnej. W poniższym przykładzie pokażemy, w jaki sposób należy korzystać z przekształceń (S1)–(S5) w celu obliczenia określonej wartości $F_S(t)$.

Przykład zastosowania równań (S1)–(S5) do obliczania wartości $F_S(t)$ dla ustalonej wartości argumentu $t \in I_1$:

Niech $t = 7/64$. Ponieważ $t < 1/8$, korzystamy z (S2) i otrzymujemy $F_S(7/64) = F_S(7/16)/2$. W ten sposób wykonaliśmy pierwszą iterację algorytmu rekurencyjnego. Położenie krzywej w punkcie $t = 7/64$ zostało wyrażone poprzez wartość odwzorowania $F_S(7/16)$. W następnej iteracji stosujemy równanie (S4), gdyż $7/16 \in [1/4, 1/2]$ i w konsekwencji otrzymujemy $x_S(7/64) = x_S(7/16)/2 = 1/2 - y_S(3/16)/2$ oraz $y_S(7/64) = y_S(7/16)/2 = x_S(3/16)/2$. Dalej, z (S3) mamy $x_S(3/16) = 1/2 + y_S(1/16)$ i $y_S(3/16) = 1/2 - x_S(1/16)$. Następnie, z (S2) wynika $x_S(1/16) = x_S(1/4)/2$ oraz $y_S(1/16) = y_S(1/4)/2$, co kończy drugą iterację algorytmu. W ostatniej iteracji korzystamy z (S3) i otrzymujemy $x_S(1/4) = 1 - y_S(0) = 1$ oraz $y_S(1/4) = x_S(0) = 0$. Po odpowiednich podstawieniach, ostatecznie $F_S(7/64) = (1/2, 1/4)$. □

Zbieżność algorytmu opartego na układzie równań (S1)–(S5) jest zapewniona w przypadku, gdy wartości t mają skończone czwórkowe rozwinięcia, czyli $t = i_k\, 4^{-k}$, gdzie i_k jest liczbą całkowitą z przedziału $0 \le i_k \le 4^k$, a k jest pewną liczbą naturalną. Łatwo sprawdzić, że jednokrotne zastosowanie równań (S1)–(S5) prowadzi do wyrażenia wartości $F_S(t)$ w zależności od $F_S(0)$ bądź w zależności od argumentu $i_{k-1}\, 4^{-k+1}$, gdzie $0 \le i_{k-1} \le 4^{k-1}$.

Lemat 2.1 *Układ równań (S1)–(S5) jest równoważny z równaniami (2.5)–(2.8). Rozwiązaniem układu równań (S1)–(S5) jest krzywa wypełniająca $F_S(t) : I_1 \to I_2$.*

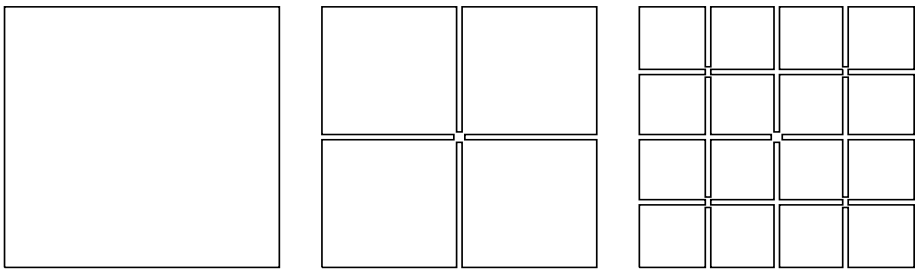
Dowód. Wykazanie równoważności (2.5)–(2.8) oraz (S1)–(S5) wymaga wyznaczenia na podstawie (S1)–(S5) wartości F_S w punktach $1/8$, $1/4$, $1/2$ oraz sprawdzenia, że spełniają je, odpowiednio, równania (2.5), (2.6) oraz (2.8). Z równań (S3) i (S1) wynika, że $F_S(1/8) = (1/2, 1/2)$, z (S4) i (S1) wynika, że $F_S(1/4) = (1, 0)$, skąd dalej otrzymujemy $F_S(1/2) = (1, 1)$. Sprawdzenie odpowiednich warunków (2.5)–(2.8) jest zupełnie elementarne.

Dalej przejdziemy do dowodu, że równania (S1)–(S5) definiują krzywą wypełniającą I_2. W pierwszym etapie określimy wartości $F_S(t)$ dla $t = 1/4, 2/4, 3/4, 1$. Są to, jak zobaczymy, współrzędne wierzchołków kwadratu I_2. Z równań (S5) i (S1) wynika, że $F_S(1) = (y_S(1-1), x_S(1-1)) = (y_S(0), x_S(0)) = (0, 0)$. Dalej, z (S4) i (S1) otrzymujemy $F_S(1/4) = (1 - y_S(1/4 - 1/4), x_S(1/4 - 1/4)) = (1, 0)$, a stąd $x_S(3/4) = y_S(1 - 3/4) = 0$ oraz $y_S(3/4) = x_S(1/4) = 1$. Dalej, z (S4) wynika, że $F_S(1/2) = (1 - y_S(1/2 - 1/4), x_S(1/2 - 1/4)) = (1, 1)$.

W następnym etapie wyznaczymy wartości $F_S(t)$ dla wszystkich nie ustalonych do tej pory wartości $t = i/16$, $i = 0, 1, 2, \ldots, 16$. Otrzymane punkty mają współrzędne będące wielokrotnością $1/2$, a mianowicie:

$F_S(1/16) = (x_S(4 \cdot 1/16)/2, y_S(4 \cdot 1/16)/2) = (1/2, 0)$,

$F_S(1/8) = (1/2 + y_S(1/8 - 1/8), 1/2 - x_S(1/8 - 1/8)) = (1/2, 1/2)$,



Rys. 2.1. Kolejne przybliżenia krzywej Sierpińskiego w 2-D
Fig. 2.1. Approximations of the Sierpiński space-filling curve in 2-D

$F_S(3/16) = (1/2 + y_S(3/16 - 1/8), 1/2 - x_S(3/16 - 1/8)) = (1/2, 0),$
$F_S(5/16) = (1 - y_S(5/16 - 1/4), x_S(5/16 - 1/4)) = (1, 1/2),$
$F_S(3/8) = (1 - y_S(3/8 - 1/4), x_S(3/8 - 1/4)) = (1/2, 1/2),$
$F_S(7/16) = (1 - y_S(7/16 - 1/4), x_S(7/16 - 1/4)) = (1, 1/2),$
$F_S(9/16) = (y_S(1 - 9/16), x_S(1 - 9/16)) = (1/2, 1),$
$F_S(10/16) = (y_S(1 - 10/16), x_S(1 - 10/16)) = (1/2, 1/2),$
$F_S(11/16) = (y_S(1 - 11/16), x_S(1 - 11/16)) = (1/2, 1),$
$F_S(13/16) = (y_S(1 - 13/16), x_S(1 - 13/16)) = (0, 1/2),$
$F_S(14/16) = (y_S(1 - 14/16), x_S(1 - 14/16)) = (1/2, 1/2),$
$F_S(15/16) = (y_S(1 - 15/16), x_S(1 - 15/16)) = (0, 1/2).$

Połączenie za pomocą odcinków kolejnych punktów odpowiadających wartościom $t = 0, 1/16, 2/16, \ldots 15/16, 1$ prowadzi do otrzymania ciągłego przybliżenia krzywej wypełniającej $F_S(t)$. Na rysunku 2.1 przedstawiono trzy kolejne przybliżenia krzywej Sierpińskiego. Należy zwrócić uwagę na fakt, iż jest to inna metoda aproksymacji krzywej Sierpińskiego, niż metoda oryginalnie proponowana przez samego Sierpińskiego. Łatwo sprawdzić, że przedstawiona tu metoda aproksymacji krzywej Sierpińskiego jest identyczna z generowaniem rodziny krzywych Sierpińskiego-Knoppa [137].

Niech $F_S^k(t)$ oznacza k-te przybliżenie krzywej, określone dokładnie w tym sensie, że $F_S^k(t) = F_S(t)$, w punktach $t \in T^k$, gdzie $T^k = \{t_i^k = i/4^k, \quad i = 0, 1, 2, \ldots, 4^k\}$, a k jest dowolną liczbą naturalną. W punktach pośrednich, czyli dla $t \notin T^k$, funkcja $F_S^k(t)$ jest zdefiniowana jako:

$$F_S^k(t) = [1 - 4^k\,(t - t_i^k)]\,F_S^k(t_i^k) + 4^k\,(t - t_i^k)\,F_S^k(t_{i+1}^k). \tag{2.9}$$

$F_S^k(t)$ jest funkcją ciągłą – odcinkami liniową. Łatwo wykazać (przez indukcję), że:

$$||F_S^k(t_i^k) - F_S^k(t_{i+1}^k)|| = ||F_S(t_i^k) - F_S(t_{i+1}^k)|| = \sqrt{2} \cdot 2^{-k}.$$

Stąd, dla dowolnego $t \in I_1$, mamy $||F_S^{k+1}(t) - F_S^k(t)|| \leq \sqrt{2} \cdot 2^{-k}$.

W konsekwencji dla każdego naturalnego k i m oraz dla każdego $t \in I_1$ zachodzi:

$$||F_S^{k+m}(t) - F_S^k(t)|| \leq ||F_S^{k+1}(t) - F_S^k(t)|| + \ldots + ||F_S^{k+m}(t) - F_S^{k+m-1}(t)|| \\ \leq \sqrt{2} \cdot 2^{-k}(1 + 2^{-1} + \ldots + 2^{-m}) < \sqrt{2} \cdot 2^{-k+1}.$$

Wnioskujemy stąd, że ciąg funkcji $F_S^k(t)$ jest jednostajnie zbieżny. W związku z tym jego granica jest także funkcją ciągłą. Przypomnijmy, że w przypadku funkcji ograniczonych ciągłość jest równoważna z jednostajną ciągłością. Podobnie łatwo można wykazać, że F_S jest funkcją jednostajnie ciągłą na zbiorze $\cup_k T^k$ $(k = 1, 2, \ldots)$. Funkcja $\lim_{k\to\infty} F_S^k(t)$ jest zgodna z $F_S(t)$ na zbiorze $\cup_k T^k$ $(k = 1, 2, \ldots)$, gęstym w I_1, a ponieważ jest funkcją ciągłą (jednostajnie), musi być więc identyczna z $F_S(t)$ na całym odcinku I_1 (jest to również alternatywny dowód ciągłości $F_S(t)$ na całym odcinku I_1). Funkcja F_S odwzorowuje zbiór $\cup_k T^k$, gęsty w I_1, w zbiór wszystkich punktów o współrzędnych posiadających skończone dwójkowe rozwinięcia, który jest zbiorem gęstym w I_2. Ponieważ I_1 jest zbiorem zwartym, stąd $F_S(I_1)$ jest także zbiorem zwartym (zawierającym zbiór gęsty w I_2). W związku z tym $F_S(I_1) = I_2$, co kończy dowód lematu. □

W ten sposób pokazaliśmy, że równania (S1)–(S5) jednoznacznie definiują krzywą wypełniającą kwadrat I_2.

Dalej przejdziemy do dowodu twierdzenia 2.1.1. Funkcję $x_S(t)$ możemy przedłużyć na całą prostą w taki sam sposób, jak funkcję $g(t)$, czyli: $x_S(t) = x_S(1+t)$ dla $t \leq 0$ oraz $x_S(t) = x_S(t-1)$ dla $t \geq 1$. W celu uniknięcia komplikacji zapisu, dla rozszerzonej funkcji użyto tego samego oznaczenia, $x_S(\cdot)$, co dla funkcji określonej na I_1. Dalej pokażemy, że $x_S(t)$ spełnia te same warunki, które spełnia funkcja $g(t)$, czyli warunki (2.4).

Lemat 2.2 *Dla każdego $t \in I_1$ zachodzi*

$$x_S(t) = 1 - x_S(t - 1/2). \tag{2.10}$$

Dowód. Wystarczy pokazać, że warunek (2.10) jest spełniony dla $1/2 \leq t \leq 1$, ponieważ, jeśli $t < 1/2$, to $x_S(t - 1/2) = x_S(t - 1/2 + 1) = x_S(t + 1/2)$.

Łatwo sprawdzić, że równanie (2.10) jest spełnione dla każdego $t \in T^1$. Pokażemy, że ze spełnienia (2.10) dla każdego $t \in T^k$ wynika spełnienie tego warunku dla $t \in T^{k+1}$. Wymaga to rozpatrzenia kolejnych szczególnych przypadków.

Niech $t \in T^{k+1}$ oraz $t \in [7/8, 1]$. Wtedy zachodzi $(1 - t) \in [0, 1/8]$. Z równań (S5) i (S2) otrzymujemy $x_S(t) = y_S(1-t) = y_S(4-4t)/2 = x_S(4t-3)/2$. Ponieważ (2.10) jest spełnione dla $t \in T^k$, a $(4t - 3) \in [1/2, 1]$ oraz $(4t - 3) \in T^k$, więc $x_S(t) = (1 - x_S(4t - 7/2))/2 = 1/2 - x_S(t - 7/8) = y_S(t - 7/8 + 1/8)$. Dalej mamy

$(t-3/4) \in [1/8, 1/4]$, zatem, korzystając z (S3), otrzymujemy $y_S(t-3/4) = 1 - x_S(t-3/4+1/4) = 1 - x_S(t-1/2)$.

W następnym przypadku niech $t \in T^{k+1}$ oraz $t \in [6/8, 7/8]$. Wtedy zachodzi $(1-t) \in [1/8, 2/8]$. Z równań (S5), (S3) i (S2) otrzymujemy $x_S(t) = y_S(1-t) = 1/2 - x_S(7/8-t) = 1/2 - x_S(7/2-4t)/2$. Ponieważ (2.10) jest spełnione dla $t \in T^k$, a $(7/2-4t) \in [1/2, 1]$ oraz $(7/2-4t) \in T^k$, więc $x_S(t) = 1/2 - (1 - x_S(4-4t))/2 = x_S(4-4t)/2 = y_S(4t-3)/2$. Dalej, ponieważ $(4t-3) \in [0, 1/2]$, korzystamy z (S2) i otrzymujemy $y_S(4t-3)/2 = y_S(t-3/4)$. Z równania (S4) wynika, że dla dowolnego $s \in [0, 1/8]$ zachodzi $y_S(s) = 1 - x_S(s+1/4)$. W konsekwencji, $y_S(t-3/4) = 1 - x_S(t-3/4+1/4) = 1 - x_S(t-1/2)$, co kończy ten fragment dowodu.

Dalej, niech $t \in T^{k+1}$ oraz $t \in [5/8, 6/8]$, wtedy $(1-t) \in [2/8, 3/8]$. Z równań (S5), (S4) i (S2) otrzymujemy $x_S(t) = y_S(1-t) = x_S(6/8-t) = x_S(3-4t)/2$. Ponieważ (2.10) jest spełnione dla $t \in T^k$, a $(3-4t) \in [0, 1/2]$ oraz $(3-4t) \in T^k$, zatem $x_S(3-4t+1/2) = 1 - x_S(3-4t)$. W ten sposób otrzymujemy $x_S(t) = (1 - x_S(3-4t+1/2))/2 = 1/2 - y_S(4t-5/2)/2$. Ponieważ $(4t-5/2) \in [0, 1/2]$, możemy skorzystać z (S2), co prowadzi do równości $y_S(4t-5/2)/2 = y_S(t-5/8)$. Z równania (S3) wynika, że $y_S(t-5/8) = x_S(t-5/8+1/8) - 1/2$. Po podstawieniu otrzymujemy $1/2 - x_S(t-1/2) + 1/2 = 1 - x_S(t-1/2)$, co kończy ten fragment dowodu.

W końcu, w ostatnim przypadku, niech $t \in T^{k+1}$ oraz $t \in [1/2, 5/8]$. Możemy skorzystać z warunków (S5), (S4) i (S3), gdyż $(1-t) \in [3/8, 1/2]$. Otrzymujemy: $x_S(t) = y_S(1-t) = x_S(6/8-t) = 1/2 + y_S(5/8-t)$. Dalej, korzystając z (S2), mamy $1/2 + y_S(t-5/8) = 1/2 + y_S(5/2-4t)/2 = 1/2 + x_S(1-5/2+4t)/2$.

Ponieważ (2.10) jest spełnione dla $t \in T^k$, a $(4t-3/2) \in [1/2, 1]$ oraz $(4t-3/2) \in T^k$, więc $x_S(4t-3/2) = 1 - x_S(4t-2)$. W ten sposób otrzymujemy $x_S(t) = 1/2 + (1 - x_S(4t-2))/2 = 1 - x_S(t-1/2)$, co kończy ten fragment dowodu.

Funkcja $x_S(t)$ jest funkcją jednostajnie ciągłą, zatem warunek (2.10) jest spełniony dla $t \in T^k$, $k = 1, 2, \ldots$, gdzie zbiór $\cup_k T^k$ jest zbiorem gęstym w I_1, co pozwala na przedłużenie warunku (2.10) na cały odcinek jednostkowy. □

Lemat 2.3 *Dla każdego* $t \in [1/8, 3/8]$ *zachodzi*

$$x_S(t) = 1/2 + x_S(1/2 - 4t)/2. \tag{2.11}$$

Dowód. Niech $t \in [1/8, 1/4]$, wtedy z równania (S3) wynika, że $x_S(t) = 1/2 + y_S(t-1/8) = 1/2 + y_S(4t-1/2)/2$. Z równania (S5) otrzymujemy ostatecznie $x_S(t) = 1/2 + x_S(1-1/2-4t)/2 = 1/2 + x_S(1/2-4t)/2$. Podobnie, dla $t \in [1/4, 3/8]$ mamy, korzystając z (S4), $x_S(t) = 1 - y_S(t-1/4) = 1 - y_S(4t-1)/2 = 1 -$

$x_S(2-4t)/2$. Z lematu 2.2 otrzymujemy dalej $x_S(t) = 1-(1-x_S(3/2-4t))/2 = 1/2 + x_S(3/2-4t)/2$. W końcu, ponieważ, dla dowolnego argumentu $s \in I_1$, $x_S(s) = x_S(s-1)$, zatem $x_S(t) = 1/2 + x_S(1/2-4t)/2$, co kończy dowód. □

Pozostał jeszcze do udowodnienia następujący lemat:

Lemat 2.4 *Dla każdego $t \in [7/8, 1]$ zachodzi*

$$x_S(t) = x_S(4t)/2. \tag{2.12}$$

Dowód. Z prostych przekształceń wynika, że dla $s = 1-t \in [0, 1/8]$ mamy $x_S(t) = x_S(1-s) = y_S(s) = y_S(4s)/2 = x_S(1-4s)/2$. Dalej, korzystając z tego, że dla dowolnego s zachodzi $x_S(s) = x_S(s+3)$ otrzymujemy (po podstawieniu $s = 1-t$) $x_S(1-4s)/2 = x_S(4-4s)/2 = x_S(4t)/2$, co kończy dowód lematu. □

Z lematów 2.2, 2.3, 2.4 oraz z równania (2.5) wynika, że funkcja $x_S(t)$ spełnia równania definiujące funkcję $g(t)$. Ponieważ $g(t)$ jest określona jednoznacznie, a $x_S(t)$ jest funkcją ciągłą (jednostajnie), zatem są one identyczne, czyli $x_S(t) = g(t)$, $t \in R$. Wystarczy teraz wykazać, że $y_S(t) = x_S(t-1/4)$ dla każdego $t \in I_1$, by móc stwierdzić, że obie krzywe, $F_g(t)$ oraz $F_S(t)$, są identyczne.

Lemat 2.5 *Dla każdego $t \in I_1$ zachodzi*

$$y_S(t) = x_S(t-1/4). \tag{2.13}$$

Dowód. Dla $t \in [1/4, 1/2]$ warunek (2.13) otrzymujemy wprost z (S4). Dalej, dla tego samego przedziału przynależności t, zachodzi $x_S(t) = 1 - y_S(t-1/4)$. Stąd, dla $t \in [0, 1/4]$, wynika $y_S(t) = 1 - x_S(t+1/4)$. Z kolei z (2.11) otrzymujemy $y_S(t) = 1 - 1 + x_S(t-1/4) = x_S(t-1/4)$.

W przypadku, gdy $t \in (1/2, 1]$, z (S5) otrzymujemy $y_S(t) = x_S(1-t)$. Dalej, z (2.10) wynika, że $x_S(1-t) = 1 - x_S(1-t+1/2) = 1 - x_S(3/2-t)$. Korzystając ponownie z (S5), otrzymujemy

$$1 - x_S(3/2-t) = 1 - y_S(1+t-3/2) = 1 - y_S(t-1/2).$$

Ponieważ $t - 1/2 \leq 1/2$, możemy więc podstawić $y_S(t-1/2) = x_S(t-1/2-1/4)$. Jeżeli $t \in [3/4, 1]$, to z (2.10) wynika $1 - x_S(t-3/4) = x_S(t-1/4)$. Gdy natomiast $t \in (1/2, 1]$, wtedy $1 - x_S(t-3/4) = 1 - x_S(t-3/4+1) = 1 - x_S(t+1/4)$. Korzystając ponownie z równania (S5), otrzymujemy dalej $1 - x_S(t+1/4) = x_S(t-1/4)$, co kończy dowód lematu. □

W ten sposób w lematach 2.2–2.5 pokazaliśmy, że odwzorowanie $F_S(t) = (x_S(t), y_S(t))$ jest równoważne z odwzorowaniem $F_g(t) = (g(t), g(t-1/4))$, co kończy dowód twierdzenia 2.1.1.

Własności krzywej Sierpińskiego

Z układu równań (2.5)–(2.8) wynika bezpośrednio, że wartości funkcji $F_S(t)$ znajdują się w kwadracie $[0,1]\times[0,1]$, a w szczególności: dla $t \in [7/8, 1]$, $[0, 1/8]$ znajdują się w kwadracie $[0, 1/2]\times[0, 1/2]$; dla $t \in [1/8, 3/8]$ znajdują się w kwadracie $[1/2, 1] \times [0, 1/2]$; dla $t \in [3/8, 5/8]$ znajdują się w kwadracie $[1/2, 1] \times [1/2, 1]$; a dla $t \in [5/8, 7/8]$ przyjmują wartości w kwadracie $[0, 1/2]\times[1/2, 1]$. Ponadto (ze względu na (2.8)) wartości odwzorowania $F_S(t)$ dla $t \in [0, 1/2]$ znajdują się w trójkącie o wierzchołkach w punktach $(0,0),(1.0),(1,1)$, natomiast dla $t \in [1/2, 1]$ przyjmują wartości w trójkącie o wierzchołkach w punktach $(0,0),(0,1),(1,1)$.

Lemat 2.6 *Istnieje wzajemnie jednoznaczne odwzorowanie między pododcinkami* $A_i^k = [1/2 \cdot i/4^k, 1/2 \cdot (i+1)/4^k]$, $(i = 0, \ldots, 2 \cdot 4^k - 1$, *k naturalne), oraz odpowiednimi trójkątami prostokątnymi* $Q_i^k = F_S(A_i^k)$ *o przyprostokątnych o długości* 2^{-k} *pokrywającymi* I_2.

Dowód. Dowód lematu wynika ze spełnienia powyższej własności przez krzywe aproksymujące $F_S^k(t)$, $k = 1, 2, \ldots$ Porównaj też dowód podobnej własności w pracy [124]. □

Krzywa Sierpińskiego jest odwzorowaniem ciągłym, nie jest jednak różniczkowalna [137], spełnia natomiast warunek Höldera z wykładnikiem 1/2. Należy nadmienić, że stała występująca w tym warunku ma najmniejszą wartość w porównaniu z analogicznymi stałymi występującymi w odpowiednich warunkach Höldera sformułowanych dla innych dwuwymiarowych krzywych wypełniających. Warunek ten możemy zapisać w następującej postaci:

Lemat 2.7 *Dla każdego* $t_1, t_2 \in I_1$ *zachodzi*

$$||F_S(t_1) - F_S(t_2)||^2 \leq \alpha^2 \min |t_2 - t_1|, |1 - (t_2 - t_1)| \leq \alpha^2 |t_2 - t_1|, \qquad (2.14)$$

gdzie $\alpha = 2$. □

Należy zauważyć, że wartość $\alpha^2 = 4$ jest niepoprawialna, gdyż można wskazać takie pary punktów t_1, t_2, na przykład 0 i 1/8, dla których $||F_S(t_1) - F_S(t_2)||^2 = \alpha^2 |t_2 - t_1|$.

Dowód lematu 2.7 podano, korzystając z innego opisu krzywej Sierpińskiego, w pracy Platzmana i Bartholdiego [124]. Przedstawiono tam także inną wersję algorytmu generowania krzywej Sierpińskiego. Podstawą tego algorytmu jest równanie rekurencyjne, które pozwala na wyznaczanie punktów wierzchołkowych kolejnych aproksymacji krzywej na podstawie poprzedniej aproksymacji.

Dla porównania podamy tu wspomniany algorytm, korzystając jednak z wprowadzonych wcześniej oznaczeń i definicji.

Niech $F(t)$ oznacza krzywą wypełniającą. Dalej, niech $F(0) = F(1) = (0,0)$, $F(1/4) = (1,0)$, $F(1/2) = (1,1)$, $F(3/4) = (0,1)$. Poniższe równanie definiuje dwuwymiarową krzywą Sierpińskiego.

$$F(i/2^k) = \left(F(\lfloor i/4 \rfloor / 2^{k-2}) + F(\lceil i/4 \rceil / 2^{k-2})\right)/2 \qquad i = 1,3,5,\ldots,2^k-1,\ k = 3,4,\ldots \tag{2.15}$$

Funkcja $F(t)$ ma wartości określone dla wszystkich punktów odcinka jednostkowego o skończonych czwórkowych rozwinięciach, jednocześnie jest jednostajnie ciągła na wyżej wspomnianym zbiorze punktów, w związku z tym można ją przedłużyć na cały odcinek I_1.

Nietrudno pokazać (poprzez indukcję), że zdefiniowane odwzorowanie $F(t)$ jest identyczne z $F_S(t)$, czyli z dwuwymiarową krzywą Sierpińskiego.

Algorytm zaproponowany przez Platzmana i Bartholdiego jest algorytmem wygodnym w sytuacji, gdy generowane jest pewne przybliżenie całej krzywej Sierpińskiego. W przypadku ograniczenia się do dokładności 2^{-k}, a więc przy wyznaczaniu wszystkich wartości $F(i/2^k)$, $i = 0,1,\ldots,2^k$ – punktów wierzchołkowych krzywej wypełniającej, należy $O(2^k)$ razy wyznaczyć wartość funkcji F. Jeśli natomiast chcemy określić wartość $F(t)$ jedynie dla pewnego ustalonego t, podanego z tą samą co poprzednio dokładnością 2^{-k}, to musimy wykonać w najgorszym razie $O(2^{k/2})$ iteracji algorytmu Platzmana i Bartholdiego. Skorzystanie z równań (S1)–(S5) gwarantuje wtedy znacznie mniejsze nakłady obliczeniowe, rzędu $O(k)$. Niewątpliwie, dla naszych celów, czyli do przetwarzania pojedynczych obserwacji na bieżąco, bardziej efektywna będzie wersja algorytmu bazująca na układzie równań (S1)–(S5).

2.2 Krzywa Hilberta

W roku 1891 Hilbert [72] opisał kolejną, po krzywej Peano, krzywą wypełniającą kwadrat. Istnieje bardzo wiele różnych sposobów definiowania dwuwymiarowej krzywej Hilberta [137], [23], [111].

Proponujemy tutaj definicję krzywej Hilberta analogiczną do opisu (2.5)–(2.8) krzywej Sierpińskiego, w której korzystamy z podstawowych własności geometrycznych krzywej. Podobnie jak w przypadku krzywej Sierpińskiego, definicja tego typu prowadzi w prosty sposób do otrzymania konstruktywnego algorytmu generowania punktów krzywej Hilberta. Opis krzywej za pomocą układu równań funkcyjnych jednoznacznie definiuje odwzorowanie $I_1 \to I_2$ oraz pozwala na bardzo precyzyjne zbadanie jego własności. Dzięki temu opisowi będziemy w stanie wyznaczyć optymalną wartość stałej w warunku Höldera dotyczącym dwuwymiarowej krzywej Hilberta.

Niech $F_H(t) = (x_H(t), y_H(t))$, $t \in I_1$ oznacza odwzorowanie przeprowadzające punkty z odcinka w punkty kwadratu jednostkowego.

Twierdzenie 2.2.1 *Odwzorowanie $F_H(t) = (x_H(t), y_H(t))$, $t \in I_1$ spełniające warunki:*

$$x_H(t) = y_H(4t)/2, \;\; y_H(t) = x_H(4t)/2, \; 0 \le t \le 1/4, \tag{2.16}$$

$$x_H(t) = 1/2 - y_H(1/2 - t), \;\; y_H(t) = -1/2 + x_H(1/2 - t), \; 0 \le t \le 1/4, \tag{2.17}$$

$$x_H(t) = x_H(1 - t), \;\; y_H(1 - t) + y_H(t) = 1, \;\; 0 \le t \le 1, \tag{2.18}$$

jednoznacznie definiuje dwuwymiarową krzywą Hilberta.

Łatwo zauważyć, że z równania (2.16) wynika, iż $x_H(0) = y_H(0)$, natomiast z $y_H(1 - t) + y_H(t) = 1$ otrzymujemy $y_H(1/2) = 1/2$. Przepisując równanie (2.17) w postaci $x_H(t) = 1/2 + y_H(1/2 - t)$ oraz $y_H(t) = 1/2 - x_H(1/2 - t)$ dla $1/4 \le t \le 1/2$, a następnie podstawiając $t = 1/2$, dostajemy $x_H(0) = 0$. Stąd także wynika, że $y_H(0) = 0$.

Dalej, korzystając kolejno z (2.16) i (2.17), otrzymujemy przy $t \in [1/4, 1/2]$, $x_H(t) = 1/2 + y_H(1/2 - t) = 1/2 + x_H(2 - 4t)/2 = 1/2 + x_H(4t - 1)/2 = 1/2 + y_H(t - 1/4)$ i podobnie $y_H(t) = 1/2 - x_H(1/2 - t) = 1/2 - y_H(2 - 4t)/2 = 1/2 - 1/2 + y_H(4t - 1)/2 = x_H(t - 1/4)$. W ten sposób wykazaliśmy następującą własność:

Lemat 2.8 *Z układu równań (2.16)–(2.18) wynika, że równanie (2.17) jest równoważne z równaniami*

$$x_H(t) = 1/2 + y_H(t - 1/4), \;\; y_H(t) = x_H(t - 1/4), \;\; 1/4 \le t \le 1/2. \tag{2.19}$$

□

Własność wewnętrznego podobieństwa odwzorowania (2.17) można zastąpić równoważną własnością (2.19), mającą charakter przesunięcia względem argumentu t. Dalej będziemy często korzystać z tego właśnie równania.

Wyznaczenie wprost wartości $F_H(0) = (0, 0)$ pozwala na sformułowanie na podstawie (2.16)–(2.18) konstruktywnego algorytmu generowania krzywej, który umożliwia dokładne wyznaczenie współrzędnych punktów z kwadratu odpowiadających punktom z odcinka o skończonych dwójkowych rozwinięciach. Wyznaczymy jeszcze wartości F_H w punktach 1/3 i 2/3. Umożliwi to otrzymanie rekurencyjnego algorytmu pozwalającego na obliczenie dokładnych wartości krzywej Hilberta w punktach, którym odpowiada zbiór wartości odwzorowania o skończonych dwójkowych rozwinięciach obu współrzędnych. Z równań (2.18) wynika, że

$F_H(1) = F_H(1-0) = (x_H(0), 1-y_H(0)) = (0,1)$, natomiast z (2.19) otrzymujemy $F_H(1/3) = F_H(1/12 + 1/4) = (1/2 + y_H(1/12),\ x_H(1/12))$.

Dalej $F_H(1/12) = F_H(1/4 \cdot 1/3) = (y_H(1/3)/2, x_H(1/3)/2)$, stąd $x_H(1/3) = 1/2 + x_H(1/3)/2$ oraz $y_H(1/3) = y_H(1/3)/2$.

W konsekwencji otrzymujemy $F_H(1/3) = (1,0)$ oraz $F_H(2/3) = F_H(1-1/3) = (x_H(1/3), 1 - y_H(1/3)) = (1,1)$.

Poniższy układ równań jest równoważny z równaniami (2.16)–(2.18):

H1) $F_H(0) = (0,0), F_H(1/3) = (1,0)$,

H2) $x_H(t) = y_H(4t)/2,\ \ y_H(t) = x_H(4t)/2,\ \ t \in [0, 1/4)$,

H3) $x_H(t) = 1/2 + y_H(t - 1/4),\ \ y_H(t) = x_H(t - 1/4),\ \ t \in [1/4, 1/2]$,

H4) $x_H(t) = x_H(1 - t),\ \ y_H(1 - t) + y_H(t) = 1,\ \ t \in (1/2, 1]$.

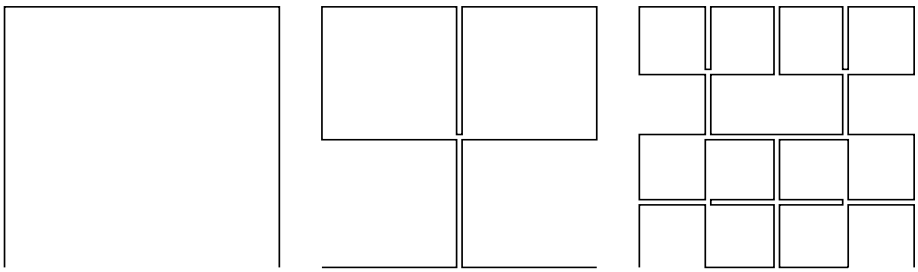
Lemat 2.9 *Układ równań funkcyjnych (H1)–(H4) jest równoważny z układem równań (2.16)–(2.18). Rozwiązaniem układu równań (H1)–(H4) jest krzywa wypełniająca $F_H(t) : I_1 \to I_2$.*

Dowód. Postępowanie dowodowe przebiegać będzie podobnie jak w przypadku analogicznego lematu sformułowanego dla krzywej Sierpińskiego. Wykazanie równoważności układu równań (H1)–(H4) z układem równań (2.16)–(2.18), jako elementarne, pominiemy. W pierwszym etapie określimy wartości $F_H(t)$ dla $t = 1/3, 2/3, 1$. Są to, jak widać, współrzędne wierzchołków kwadratu I_2.

W kolejnym etapie wyznaczymy wartości $F_H(t)$ dla wszystkich argumentów $t = i/12,\ i = 0, 1, 2, \ldots, 12$. Otrzymane punkty mają współrzędne będące wielokrotnością $1/2$, a mianowicie:

$F_H(1/4) = F_H(0 + 1/4) = (1/2 + y_H(0), x_H(0)) = (1/2, 0)$,
$F_H(1/2) = F_H(1/4 + 1/4) = (1/2 + y_H(1/4), x_H(1/4)) = (1/2, 1/2)$,
$F_H(3/4) = F_H(1 - 1/4) = (x_H(1/4), 1 - y_H(1/4)) = (1/2, 1)$,
$F_H(1/12) = (y_H(1/3)/2, x_H(1/3)/2) = (0, 1/2)$,
$F_H(1/6) = F_H(2/12) = F_H(1/4 \cdot 2/3) = (y_H(2/3)/2, x_H(2/3)/2) = (1/2, 1/2)$,
$F_H(5/12) = F_H(1/4 + 1/6) = (1/2 + y_H(1/6), x_H(1/6)) = (1, 1/2)$,
$F_H(7/12) = F_H(1 - 5/12) = (x_H(5/12), 1 - y_H(5/12)) = (1, 1/2)$,
$F_H(5/6) = F_H(1 - 1/6) = (x_H(1/6), 1 - y_H(1/6)) = (1/2, 1/2)$,
$F_H(11/12) = F_H(1 - 1/12) = (x_H(1/12), 1 - y_H(1/12)) = (0, 1/2)$.

Poprzez połączenie odcinkami punktów z I_2 odpowiadających kolejnym wartościom $t = i/12,\ (i = 0, 1, 2, \ldots, 12)$ na odcinku I_1, otrzymujemy ciągłą aproksymację krzywej wypełniającej. Na rysunku 2.2 przedstawiono trzy pierwsze przybliżenia krzywej Hilberta. Jest to inna metoda aproksymacji niż rozważane do tej pory (por. [137]). Jej cechą charakterystyczną jest określenie punktów węzłowych



Rys. 2.2. Kolejne przybliżenia krzywej Hilberta w 2-D
Fig. 2.2. Approximations of the Hilbert space-filling curve in 2-D

– punktów z kwadratu budujących dane przybliżenie – na podstawie dokładnych wartości odwzorowania. Zbiór punktów węzłowych tworzą, jak zobaczymy, wszystkie punkty I_2, których współrzędne mają skończone dwójkowe rozwinięcia.

Niech $F_H^k(t)$ oznacza k-tą aproksymację krzywej, określoną dokładnie w tym sensie, że $F_H^k(t) = F_H(t)$, w punktach $t \in T^k$, gdzie

$$T^k = \{t_i^k = i/(3 \cdot 4^k),\ \ i = 0, 1, 2, \ldots, 3 \cdot 4^k\},$$

a k jest dowolną liczbą naturalną.

Odcinkami liniowe przybliżenie dwuwymiarowej krzywej Hilberta jest wtedy postaci

$$F_H^k(t) = [1 - 3 \cdot 4^k\,(t - t_i^k)]\,F_H^k(t_i^k) + 3 \cdot 4^k\,(t - t_i^k)\,F_H^k(t_{i+1}^k).$$

Zauważmy, że $F_H^k(t)$ jest funkcją ciągłą. Łatwo pokazać (przez indukcję), że $||F_H^k(t_i^k) - F_H^k(t_{i+1}^k)|| = 2^{-k}$, przy czym różnica ta dotyczy zawsze tylko jednej współrzędnej.

Stąd $||F_H^{k+1}(t) - F_H^k(t)|| \le \sqrt{2} \cdot 2^{-k}$, a w konsekwencji dla każdego naturalnego k i m oraz $t \in I_1$ zachodzi

$$\begin{aligned} ||F_H^{k+m}(t) - F_H^k(t)|| &\le ||F_H^{k+1}(t) - F_H^k(t)|| + \ldots + ||F_H^{k+m}(t) - F_H^{k+m-1}(t)|| \\ &\le \sqrt{2} \cdot 2^{-k}(1 + 2^{-1} + \ldots + 2^{-m}) < \sqrt{2} \cdot 2^{-k+1}. \end{aligned}$$

Wynika stąd, że ciąg funkcji $F_H^k(t)$ jest jednostajnie zbieżny. Jego granica jest więc funkcją ciągłą. Dalszy ciąg dowodu przebiega analogicznie jak w przypadku dowodu lematu 2.1 sformułowanego dla krzywej Sierpińskiego. □

Własności krzywej Hilberta

Z równań (H2)–(H4) wynika bezpośrednio, że wartości funkcji $F_H(t)$ znajdują się w kostce $[0,1]\times[0,1]$, a w szczególności: dla $t \le 1/4$ znajdują się w kwadracie $[0,1/2]\times[0,1/2]$; dla $1/4 \le t \le 1/2$ znajdują się w kwadracie $[1/2,1]\times[0,1/2]$; dla $1/2 \le t \le 3/4$ znajdują się w kwadracie $[1/2,1]\times[1/2,1]$; natomiast dla $3/4 \le t \le 1$ znajdują się w kwadracie $[0,1/2]\times[1/2,1]$.

Lemat 2.10 *Istnieje wzajemnie jednoznaczne odwzorowanie między pododcinkami* $A_i^k = [i/4^k,(i+1)/4^k]$, $i = 0,\ldots,4^k-1$ *oraz odpowiednimi podkostkami* $Q_i^k = F_H(A_i^k)$ *o boku* 2^{-k} *pokrywającymi* I_2, *gdzie* k *jest dowolną liczbą naturalną.*

Dowód. Dowód lematu wynika ze spełnienia powyższej własności przez krzywe $F_H^k(t)$, $k = 1,2,\ldots$ □

Lemat 2.11 *Dla każdego* $t \in I_1$ *zachodzi*

$$||F_H(t)||^2 \le 3\,t. \tag{2.20}$$

Dowód. Najpierw pokażemy, że własność (2.20) jest spełniona przez wszystkie punkty wierzchołkowe kolejnych aproksymacji krzywej $F_H(t)$, czyli że

$$||F_H(t_i^k)||^2 \le 3\,t_i^k$$

zachodzi dla każdego $t_i^k \in T^k = \{i/(3\cdot 4^k)\ i = 0,1,\ldots,3\cdot 4^k\}$, dla dowolnego naturalnego k. Łatwo sprawdzić, że nierówność (2.20) jest spełniona dla $k=1$, a dokładniej dla wszystkich $t \in T^1 = \{0,1/12,2/12,\ldots,11/12,1\}$. Dalej pokażemy, że ze spełnienia warunku (2.20) dla pewnego $k \ge 1$ wynika spełnienie go dla $k+1$, co zakończy pierwszą część dowodu poprzez indukcję.

Niech $t \in T^{k+1}$ oraz niech $t < 1/4$. Z równania (H2) wynika, że

$$F_H(t) = (y_H(4t)/2, x_H(4t)/2).$$

Ponieważ $4t \in T^k$, więc $||F_H(t)||^2 = x_H^2(t) + y_H^2(t) = (y_H^2(4t) + x_H^2(4t))/4 \le 1/4 \cdot 3(4t) = 3t$. Dalej, niech $t \in T^{k+1}$ oraz $t \in [1/4,1/2]$. Z równania (H3) wynika, że $x_H(t) = 1/2 + y_H(t-1/4)$ oraz $y_H(t) = x_H(t-1/4)$. Ponieważ $t-1/4 \le 1/4$, mamy $y_H(t-1/4) \le 1/2$, otrzymujemy więc $||F_H(t)||^2 = x_H^2(t) + y_H^2(t) = 1/4 + y_H^2(t-1/4) + y_H(t-1/4) + x_H^2(t-1/4)) \le y_H(t-1/4) + 1/4 + 3(t-1/4) \le 3/4 + 3(t-1/4)$.

Z kolei niech $t \in T^{k+1}$ oraz niech $t \in [1/2,3/4]$. Z równań (H4) i (H3) wynika, że $x_H(t) = x_H(1-t) = 1/2 + y_H(3/4-t)$ oraz $y_H(t) = 1 - y_H(1-t) = 1 - x_H(3/4-t)$, natomiast z (H2) wynika dalej, że $x_H(t) = 1/2 + x_H(3-4t)$ oraz

$y_H(t) = 1 - y_H(3-4t)$. Łatwo sprawdzić, że jeśli $t \in T^{k+1}$, to $(3-4t) \in T^k$, a zatem

$$||F_H(t)||^2 = x_H^2(t) + y_H^2(t) = 1/4 + [y_H^2(3-4t) + x_H^2(3-4t)]/4$$
$$+x_H(3-4t)/2 - y_H(3-4t) \le 1/4 + 3/4 \cdot (3-4t) + 1/2 = 3\,(1-t) \le 3\,t.$$

Ostatnia nierówność wynika z prostego faktu, że dla $1/2 \le t \le 3/4$ zawsze musi być $1 - t \le t$.

W końcu, niech $t \in T^{k+1}$ oraz $t \in [3/4, 1]$. W tym przypadku punkty z krzywej leżą w kwadracie $[0, 1/2] \times [1/2, 1]$, stąd $||F_H(t)||^2 \le 1 + 1/4 = 5/4$. Dla $t \ge 3/4$ mamy więc $||F_H(t)||^2/t \le 5/4 \cdot 4/3 = 5/3 < 3$. Jeśli nierówność $||F_H(t_i^k)||^2 \le 3\,t_i^k$ zachodzi dla $t_i^k \in T^k$, $k = 0, 1, \ldots$, czyli na zbiorze gęstym w I_1, to nierówność ta jest spełniona dla każdego $t \in I_1$, co kończy dowód własności (2.11). Fakt ten wynika bezpośrednio z jednostajnej ciągłości $F_H(t)$. □

Sformułujemy twierdzenie, które poprawia wartość stałej w warunku Höldera sformułowanym w odniesieniu do dwuwymiarowej krzywej Hilberta.

Twierdzenie 2.2.2 *Dla każdego* $t_1,\ t_2 \in I_1$ *zachodzi*

$$||F_H(t_1) - F_H(t_2)||^2 \le \alpha^2\,|t_2 - t_1|, \tag{2.21}$$

gdzie $\alpha^2 = 6$.

Dowód. Postępowanie dowodowe przeprowadzimy podobnie jak w przypadku lematu 2.11. Najpierw pokażemy, że wartość α^2 w nierówności (2.21) jest nie mniejsza niż wymieniona w twierdzeniu wartość 6. Niech $t_1 = 23/48$ oraz $t_2 = 25/48$. Korzystając z równań (H1)–(H4), wyznaczamy $F_H(23/48) = (1/2, 1/4)$ oraz $F_H(25/48) = (1/2, 3/4)$. Stąd otrzymujemy równość $||(1/2, 1/4) - (1/2, 3/4)||^2 = 1/4 = 6 \cdot 2/48$.

Dalsza część dowodu będzie miała charakter indukcyjny. Na początku pokażemy, że własność (2.21) jest spełniona przez wszystkie punkty wierzchołkowe kolejnych aproksymacji krzywej $F_H(t)$, czyli pokażemy, że $||F_H(t1) - F_H(t_2)||^2 \le 6\,|t_1 - t_2|$ zachodzi dla wszystkich $t_1, t_2 \in T^k$, dla dowolnego naturalnego k.

T^k jest zdefiniowane tak samo jak w dowodzie poprzedniego twierdzenia, czyli $T^k = \{i/(3 \cdot 4^k),\ i = 0, 1, \ldots, 3 \cdot 4^k\}$.

Łatwo sprawdzić, że nierówność (2.21) jest spełniona dla $k = 1$, a dokładniej dla wszystkich $t_1, t_2 \in T^1 = \{0, 1/12, 2/12, \ldots, 11/12, 1\}$. Dalej pokażemy, że ze spełnienia warunku (2.21) dla pewnego $k \ge 1$ wynika spełnienie go dla $k + 1$, co zakończy pierwszą część dowodu przez indukcję.

Niech $t_1, t_2 \in T^{k+1}$ oraz $t_1, t_2 \le 1/4$.

Z równania (H2) (lub z (2.16)) wynika, że

$$F_H(t_i) = (y_H(4t_i)/2, x_H(4t_i)/2),\ i = 1, 2.$$

Ponieważ $4t_1,\ 4t_2 \in T^k$, stąd

$$\begin{aligned}||F_H(t_1) - F_H(t_2)||^2 &= [x_H(t_1) - x_H(t_2)]^2 + [y_H(t_1) - y_H(t_2)]^2 \\ &= 1/4 \cdot [y_H(4t_1) - y_H(4t_2)]^2 + 1/4 \cdot [x_H(4t_1) - x_H(4t_2)]^2 \\ &\leq 6/4 \cdot |4t_1 - 4t_2| = 6\,|t_1 - t_2|.\end{aligned}$$

Analogiczne rezultaty otrzymamy w przypadku, gdy $t_1, t_2 \in [1/4, 1/2]$ lub $t_1, t_2 \in [1/2, 3/4]$, lub $t_1, t_2 \in [3/4, 1]$.

Dalej, niech $t_1, t_2 \in T^{k+1}$ oraz niech $t_1 \in [0, 1/4]$, $t_2 \in [1/4, 1/2]$. Z równań (H2)–(H4) wynika, że

$$x_H(t_1) = y_H(4t_1)/2 = 1/2 - y_H(1 - 4t_1)/2 = 1/2 - x_H(1/4 - t_1)$$

oraz podobnie

$$y_H(t_1) = y_H(1/4 - t_1).$$

Korzystając z lematu 2.11, otrzymujemy

$$||F_H(t_1) - (1/2, 0)||^2 = x_H^2(1/4 - t_1) + y_H^2(1/4 - t_1) \leq 3\,(1/4 - t_1).$$

Dalej, podobnie jak dla t_1, dla t_2 otrzymujemy $x_H(t_2) = y_H(t_2 - 1/4) + 1/2$, $y_H(t_2) = x_H(t_2 - 1/4)$. Stąd wynika, że

$$||F_H(t_2) - (1/2, 0)||^2 = x_H^2(t_2 - 1/4) + y_H^2(t_2 - 1/4) \leq 3\,(t_2 - 1/4).$$

Z nierówności trójkąta otrzymujemy:

$$||F_H(t_1) - F_H(t_2)|| \leq 3^{1/2}\,(1/4 - t_1)^{1/2} + 3^{1/2}\,(t_2 - 1/4)^{1/2}.$$

Ponieważ dla dowolnego nieujemnego a i b zawsze zachodzi $a^{1/2} + b^{1/2} \leq (2a + 2b)^{1/2}$, zatem

$$||F_H(t_1) - F_H(t_2)|| \leq 6^{1/2}\,(t_2 - t_1)^{1/2}.$$

Dalej, niech $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [0, 1/4]$, $t_2 \in [1/2, 3/4]$. Z równań (H2)–(H4) wynika, że $x_H(t_1) = y_H(4t_1)/2$ oraz $y_H(t_1) = x_H(4t_1)/2$, natomiast

$$\begin{aligned}x_H(t_2) = x_H(1 - t_2) &= y_H(3/4 - t_2) + 1/2 = 1/2 + x_H(3 - 4t_2)/2 \\ &= 1/2 + x_H(1 - (3 - 4t_2)) = 1/2 + x_H(4(t_2 - 1/2))/2\end{aligned}$$

oraz podobnie $y_H(t_2) = 1/2 + y_H(4(t_2 - 1/2))/2$.

Wprowadźmy oznaczenia $s_1 = 4t_1$ oraz $s_2 = 4\,(t_2 - 1/2)$. Po odpowiednim podstawieniu otrzymujemy:

$$||F_H(t_2) - F_H(t_1)||^2 = 1/4 \cdot [1 + x_H(s_2) - y_H(s_1)]^2$$
$$+1/4 \cdot [1 + y_H(s_2) - x_H(s_1)]^2 = 1/4 \cdot [||F_H(s_2) - F_H(s_1)||^2$$
$$+2x_H(s_2)\,x_H(s_1) + 2y_H(s_2)\,y_H(s_1) + 1 + 2x_H(s_2) - 2(1 + x_H(s_2))\,y_H(s_1)$$
$$+1 + 2y_H(s_2) - 2(1 + y_H(s_2))\,x_H(s_1)].$$

Ponieważ zawsze zachodzi $x_H(t),\ y_H(t) \in I_1$, zatem

$$||F_H(t_2) - F_H(t_1)||^2 \leq 1/4 \cdot (||F_H(s_2) - F_H(s_1)||^2 + 10).$$

Zauważmy ponadto, że $s_i \in T^k,\ i = 1, 2$. W związku z tym

$$||F_H(s_2) - F_H(s_1)||^2 \leq 6\,(s_2 - s_1\,) = 24\,(t_2 - t_1) - 12.$$

Po podstawieniu do poprzedniej nierówności otrzymujemy

$$||F_H(t_2) - F_H(t_1)||^2 \leq 1/4 \cdot (24(t_2 - t_1) - 2) < 6\,(t_2 - t_1),$$

co kończy dowód dla $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [0, 1/4]$, $t_2 \in [1/4, 1/2]$.

Rozważmy dalej przypadek $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [1/4, 1/2]$, $t_2 \in [1/2, 3/4]$. Z równań (H2)–(H4) wynika, że $x_H(t_1) = 1/2 + y_H(t_1 - 1/4) = 1/2 + x_H(4t_1 - 1)/2 = 1/2 + x_H(2 - 4t_1)/2 = 1/2 + y_H(1/2 - t_1)$ oraz analogicznie $y_H(t_1) = 1/2 - x_H(1/2 - t_1)$. Podobnie, $x_H(t_2) = x_H(1 - t_2) = 1/2 + y_H(3/4 - t_2) = (1 + x_H(3 - 4t_2))/2 = 1/2 + x_H(1 - 3 + 4t_2)/2 = 1/2 + y_H(t_2 - 1/2)$. Przekształcając dalej $y_H(t_2)$, otrzymujemy $y_H(t_2) = 1 - y_H(1 - t_2) = 1 - x_H(3/4 - t_2) = 1 - y_H(3 - 4t_2)/2 = 1 - (1 - y_H(4t_2 - 2))/2 = 1/2 + x_H(t_2 - 1/2)$. Z lematu 2.11 wynika, że

$$||F_H(t_1) - (1/2, 1/2)||^2 = [x_H(t_1) - 1/2]^2 + [y_H(t_1) - 1/2]^2$$
$$= y_H^2(1/2 - t_1) + x_H^2(1/2 - t_1) \leq 3\,(1/2 - t_1)$$

oraz

$$||F_H(t_2) - (1/2, 1/2)||^2 = [x_H(t_2) - 1/2]^2 + [y_H(t_2) - 1/2]^2$$
$$+y_H^2(t_2 - 1/2) + x_H^2(t_2 - 1/2) \leq 3\,(t_2 - 1/2).$$

Stąd, korzystając z nierówności trójkąta, otrzymujemy:

$$||F_H(t_1) - F_H(t_2)|| \leq 3^{1/2}(1/2 - t_1)^{1/2} + 3^{1/2}(t_2 - 1/2)^{1/2}.$$

Ze znanej nierówności $a^{1/2} + b^{1/2} \leq (2a + 2b)^{1/2}$ wynika

$$||F_H(t_1) - F_H(t_2)|| \leq 6^{1/2}(|t_2 - t_1|)^{1/2}.$$

Rozważmy teraz przypadek $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [1/4, 1/2]$, $t_2 \in [3/4, 1]$. Z równań (H2)–(H4) wynika, że $x_H(t_1) = y_H(t_1 - 1/4) + 1/2 = x_H(4t_1 - 1)/2 + 1/2$ oraz $y_H(t_1) = y_H(4t_1 - 1)/2$, natomiast $x_H(t_2) = x_H(1 - t_2) = y_H(4(1 - t_2))/2 = (1 + x_H(4t_2 - 3))/2$; oraz podobnie, $y_H(t_2) = 1 - x_H(1 - 4t_2 + 4)/2 = 1 - x_H(4t_2 - 3)/2$. Niech $s_1 = 4t_1 - 1$ oraz $s_2 = 4(t_2 - 3/4)$. Po podstawieniu do poprzednich równości otrzymujemy:

$$||F_H(t_2) - F_H(t_1)||^2 = [1/2 \cdot x_H(s_1) + y_H(s_2)]^2 + [1 + y_H(s_1) - 1/2 x_H(s_2)]^2$$
$$= 1/4 \cdot [||F_H(s_2) - F_H(s_1)||^2 + 2\, x_H(s_2)\, x_H(s_1) + 2\, y_H(s_2)\, y_H(s_1)$$
$$+2\, x_H(s_2)\, y_H(s_1) + 2\, x_H(s_1)\, y_H(s_2)] + 1 - x_H(s_2) - y_H(s_1).$$

Ponieważ dla dowolnego $t \in I_1$ mamy $x_H(t), y_H(t) \in I_1$, więc

$$||F_H(t_2) - F_H(t_1)||^2 \leq 1/4 \cdot (||F_H(s_2) - F_H(s_1)||^2 + 8) + 1.$$

Zauważmy ponadto, że $s_i \in T^k$, $i = 1, 2$. W związku z tym

$$||F_H(s_2) - F_H(s_1)||^2 \leq 6(s_2 - s_1) = 24(t_2 - t_1) - 12.$$

Po podstawieniu do poprzedniej nierówności otrzymujemy

$$||F_H(t_2) - F_H(t_1)||^2 \leq 1/4 \cdot (24(t_2 - t_1) - 12 + 8) + 1 \leq 6(t_2 - t_1),$$

co kończy dowód dla $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [1/4, 1/2]$, $t_2 \in [3/4, 1]$.

Ostatni przypadek to $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [0, 1/4]$, $t_2 \in [3/4, 1]$. Tym razem możemy skorzystać z faktu, iż $F_H(t_1)$ przyjmuje wartości w kwadracie $[0, 1/2] \times [0, 1/2]$, natomiast $F_H(t_2)$ przyjmuje wartości w kwadracie $[1/2, 1] \times [1/2, 1]$. Stąd wynika, że $||F_H(t_2) - F_H(t_1)||^2 \leq 5/4$. Ponieważ $t_2 - t_1 \geq 1/2$, od razu możemy więc stwierdzić, że

$$||F_H(t_2) - F_H(t_1)||^2/(t_2 - t_1) \leq 10/4 < 6.$$

Funkcja $F_H(t)$ jest ciągła jednostajnie, stąd jeśli (2.21) jest spełnione dla $t_1, t_2 \in T^k = \{i/(3 \cdot 4^k)\}$, $(k = 0, 1, \ldots)$, to (2.21) jest spełnione także dla każdego t_1, $t_2 \in I_1$, co kończy dowód twierdzenia 2.2.2. □

W pracy [61] pokazano, że wartość $\alpha^2 \leq 6\frac{2}{3}$. Oszacowanie odpowiedniej stałej podane przez A.R. Butza [24] jest jeszcze słabsze i w przypadku dwuwymiarowym wynosi $\alpha^2 < 28$. Twierdzenie 2.2.2 podaje najlepszą, niepoprawialną wartość stałej $\alpha^2 = 6$.

2.3 Krzywa Peano

W roku 1890 Peano [121] skonstruował pierwszą krzywą wypełniającą kwadrat. W odwzorowaniu odcinka w kwadrat jednostkowy użyto trójkowych rozwinięć liczb. Obok przypadku dwuwymiarowego Peano przedstawił także przypadek trójwymiarowy.

Inne warianty krzywych wypełniających kwadrat, podobnych w konstrukcji do krzywej Peano, zaproponował w 1900 roku Moore [111]. Krzywe te bazują na nieparzystych, innych niż trójkowe, podziałach odcinka. Z kolei Wunderlich [201] przedstawił odmienne sposoby porządkowania przestrzeni dwuwymiarowej prowadzące w konsekwencji do różnych wariantów krzywej, bazujących nadal na podziałach trójkowych (krzywe typu switch-back). Istnieje dokładnie 272 różnych krzywych tego typu, wśród których występuje także oryginalna krzywa Peano [137].

W niniejszej pracy ograniczymy się do klasycznej krzywej podanej przez Peano [121]. Uogólnienie krzywej Peano na przypadek wielowymiarowy przedstawił Milne w pracy [110], a wcześniej, w mniej precyzyjny sposób, Butz [23].

Tutaj proponujemy inną, nową definicję dwuwymiarowej krzywej Peano, w której korzystamy z podstawowych własności geometrycznych odwzorowania i którą łatwo uogólnić na przypadek wielowymiarowy.

Definicja ta jest analogiczna do opisu (2.5)–(2.8) krzywej Sierpińskiego. Podobnie jak w przypadku krzywej Sierpińskiego, definicja w postaci układu równań funkcyjnych prowadzi w prosty sposób do otrzymania konstruktywnego algorytmu generowania punktów krzywej Peano. Podany dalej układ równań funkcyjnych jednoznacznie definiuje odwzorowanie $I_1 \rightarrow I_2$ oraz pozwala na precyzyjne zbadanie jego własności. Dzięki temu opisowi można wyznaczyć optymalną wartość stałej w warunku Höldera dotyczącym dwuwymiarowej krzywej Peano.

Niech $F_P(t) = (x_P(t), y_P(t))$, $t \in I_1$ oznacza odwzorowanie przeprowadzające punkty z odcinka I_1 w punkty kwadratu I_2.

Twierdzenie 2.3.1 *Odwzorowanie $F_P(t) = (x_P(t), y_P(t))$, $t \in I_1$ spełniające warunki:*

$$F_P(t) = F_P(9t)/3, \quad 0 \leq t \leq 1/9, \tag{2.22}$$

$$x_P(t) = 1/3 + x_P(t - 1/9), \quad y_P(t) = 1/3 - y_P(t - 1/9), \quad 1/9 \leq t \leq 1/3, \tag{2.23}$$

$$x_P(t) = 1 - x_P(t - 1/3), \quad y_P(t) = 1/3 + y_P(t - 1/3), \quad 1/3 \leq t \leq 1 \tag{2.24}$$

jednoznacznie definiuje dwuwymiarową krzywą Peano.

Łatwo zauważyć, że z $F_P(t) = F_P(9t)/3$ dla $t \in [0, 1/9]$ wynika $F_P(0) = (0,0)$. Ponadto z równania (2.24) otrzymujemy $F_P(1/3) = (1, 1/3)$, $F_P(2/3) = (0, 2/3)$, $F_P(1) = (1,1)$.

Ustalenie wartości $F_P(0)$ pozwala na sformułowanie rekurencyjnego algorytmu, który umożliwia dokładne wyznaczenie współrzędnych punktów z kwadratu odpowiadających punktom z odcinka (argumentom odwzorowania) o skończonych trójkowych rozwinięciach. Podstawą algorytmu są następujące równania:

P1) $F_P(0) = (0,0)$,

P2) $x_P(t) = x_P(9t)/3$, $y_P(t) = y_P(9t)/3$, $0 \le t < 1/9$,

P3) $x_P(t) = 1/3 + x_P(t - 1/9)$, $y_P(t) = 1/3 - y_P(t - 1/9)$, $1/9 \le t < 1/3$,

P4) $x_P(t) = 1 - x_P(t - 1/3)$, $y_P(t) = 1/3 + y_P(t - 1/3)$, $1/3 \le t \le 1$.

Równania (P2)–(P4) różnią się od układu równań (2.22)–(2.24) tylko zastąpieniem pewnych ograniczeń typu $\le$ przez ostre nierówności, co pozwala na jednoznaczne zastosowanie układu równań (P1)–(P4) do wyznaczania wartości $F_P(t)$. Dla granicznych wartości $t = 1/9$ oraz $t = 1/3$ spełniony jest zarówno warunek (P2), jak i (P3) oraz odpowiednio (P3) i (P4), lecz w obliczeniach stosujemy zawsze tylko jeden z nich.

Lemat 2.12 *Układ równań funkcyjnych (P1)–(P4) jest równoważny z równaniami (2.22)–(2.24). Rozwiązaniem układu równań (P1)–(P4) jest krzywa wypełniająca* $F_P(t) : I_1 \to I_2$.

Dowód. Dowód powyższego lematu przeprowadzimy podobnie, jak w przypadku krzywych Hilberta i Sierpińskiego (por. podrozdziały 2.1, 2.2). W pierwszym etapie aproksymacji krzywej najpierw określimy wartości $F_P(t)$ dla $t = 0, 1/3, 2/3, 1$. Dalej, dla $t = 1/9, 2/9, \ldots 7/9$, z równania (P3) otrzymujemy:

$F_P(1/9) = (1/3, 1/3)$, $F_P(2/9) = (2/3, 0)$.

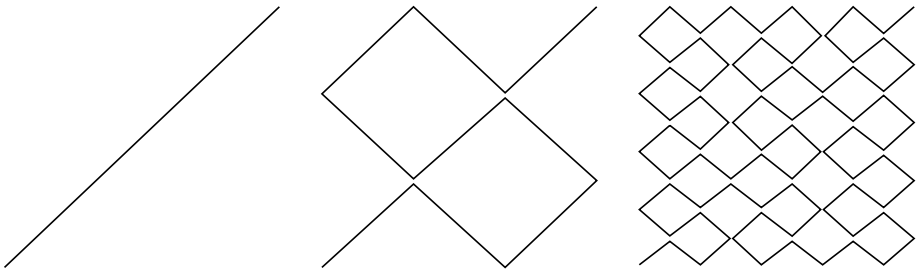
Następnie z (P4) uzyskujemy:

$F_P(4/9) = (2/3, 2/3)$, $F_P(5/9) = (1/3, 1/3)$ oraz

$F_P(7/9) = (1/3, 1)$, $F_P(8/9) = (2/3, 2/3)$.

W kolejnym etapie aproksymacji wyznaczymy wartości $F_P(t)$ dla wszystkich nie ustalonych do tej pory wartości $t = i/81$, $i = 0, 1, 2, \ldots 81$. Otrzymane w ten sposób punkty z I_2 mają współrzędne będące wielokrotnością $1/9$.

Poprzez połączenie odcinkami punktów z I_2, odpowiadających kolejnym wartościom $t = i/81$, $i = 0, 1, 2, \ldots, 81$, otrzymujemy następne ciągłe przybliżenie



Rys. 2.3. Kolejne przybliżenia krzywej Peano w 2-D
Fig. 2.3. Approximations of the Peano space-filling curve in 2-D

krzywej Peano. Łatwo pokazać poprzez indukcję, że układ równań (P1)–(P4) pozwala wygenerować wartości odwzorowania $F_P(t)$ dla wszystkich wartości $t \in T^k$, gdzie

$$T^k = \{t_i^k = i/9^k, \;\; i = 0, 1, 2, \ldots, 9^k\},$$

a k jest liczbą naturalną. Na rysunku 2.3 przedstawiono trzy kolejne przybliżenia krzywej Peano, przy czym pierwszą ($k = 0$) aproksymację krzywej stanowi po prostu odcinek łączący punkt początkowy $(0, 0)$ z punktem końcowym krzywej wypełniającej, czyli punktem $(1, 1)$. Kolejne przybliżenia stanowią krzywe odcinkami liniowe, łączące punkty wierzchołkowe odpowiadające wartościom $t = 0, 1/9^k, \ldots, (9^k - 1)/9, 1, \;\; k = 1, 2$.

Niech $F_P^k(t)$ oznacza k–tą aproksymację krzywej określoną dokładnie, w tym sensie, że $F_H^k(t) = F_H(t)$, w punktach $t \in T^k$, $k = 1, 2, \ldots$ Funkcja ta przyjmuje wartości

$$F_P^k(t) = (1 - 9^k(t - t_i^k))F_P^k(t_i^k) + 9^k(t - t_i^k)F_P^k(t_{i+1}^k)$$

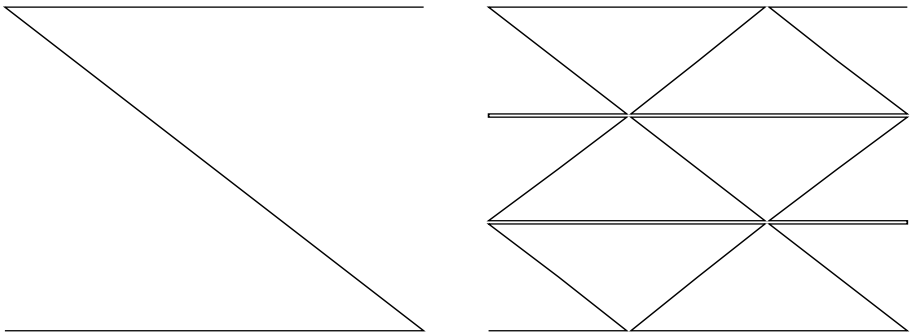
dla t nie należących do T^k. $F_P^k(t)$ jest funkcją ciągłą, odcinkami liniową. Łatwo pokazać (przez indukcję), że $||F_P^k(t_i^k) - F_P^k(t_{i+1}^k)|| = \sqrt{2}\,3^{-k}$. Obie współrzędne $F_P^k(t_i^k)$ i $F_P^k(t_{i+1}^k)$ różnią się dokładnie o 3^{-k}. Stąd otrzymujemy

$$||F_P^{k+1}(t) - F_P^k(t)|| \leq \sqrt{2}\,3^{-k}.$$

W konsekwencji dla każdego naturalnego k i m i dla każdego $t \in I_1$ zachodzi:

$$\begin{aligned} ||F_P^{k+m}(t) - F_P^k(t)|| &\leq ||F_P^{k+1}(t) - F_P^k(t)|| + \ldots + ||F_P^{k+m}(t) - F_P^{k+m-1}(t)|| \\ &\leq \sqrt{2} \cdot 3^{-k}(1 + 3^{-1} + \ldots + 3^{-m}) < \sqrt{2} \cdot 3^{-k} \cdot 3/2. \end{aligned}$$

Wnioskujemy stąd, że ciąg funkcji $F_P^k(t)$ jest jednostajnie zbieżny. Jego granica jest więc funkcją ciągłą. Dalszy ciąg dowodu jest analogiczny jak w przypadku



Rys. 2.4. Kolejne przybliżenia krzywej Peano w 2-D
Fig. 2.4. Approximations of the Peano space-filling curve in 2-D

odpowiednich lematów 2.1, 2.9, które dotyczą krzywej Sierpińskiego i krzywej Hilberta. □

Punkty wierzchołkowe kolejnych aproksymacji krzywej Peano nie obejmują wszystkich punktów z kwadratu o współrzędnych mających skończone trójkowe rozwinięcia. Na przykład, punkty $(1/3, 0)$ lub $(0, 1/3)$ nie pojawią się wśród punktów wierzchołkowych kolejnych krzywych aproksymujących, niezależnie od wartości k.

Na podstawie równań funkcyjnych (2.22)–(2.24) (bądź równoważnie równań (P1)–(P4)) można w prosty sposób wyznaczyć dodatkowe punkty wierzchołkowe, które będą stanowić podstawę innych sposobów aproksymacji krzywej $F_P(t)$.

Na rysunku 2.4 przedstawiono dwa kolejne przybliżenia krzywej Peano, przy czym pierwszą (dla $k = 0$) aproksymację krzywej stanowią odcinki łączące kolejno punkty kwadratu odpowiadające argumentom $t = 0, 1/4, 3/4, 1$. Następnym przybliżeniem jest funkcja odcinkami liniowa, łącząca punkty wierzchołkowe odpowiadające wartościom argumentów $t = i/(4 \cdot 9)$, $i = 0, 1, \ldots, 36$. Kolejne przybliżenia prowadzą do wyznaczania wartości odwzorowania dla $t = i/(4 \cdot 9^k)$, $i = 0, 1, \ldots, 4 \cdot 9^k$, $k = 0, 1, \ldots$

Zauważmy, że z (2.22) wynika, że $F_P(1/12) = F_P(3/4)/3$, natomiast konsekwencją równania (2.24) jest $F_P(3/4) = (1 - x_P(5/12), 1/3 + y_P(5/12))$. Dalej, z (2.24) wynika $F_P(5/12) = (1 - x_P(1/12), 1/3 + y_P(1/12))$. Po elementarnych przekształceniach otrzymujemy $F_P(1/12) = (0, 1/3)$, $F_P(3/4) = (0, 1)$ oraz $F_P(5/12) = (1, 2/3)$.

Podobnie, z (2.22) wynika $F_P(1/36) = F_P(1/4)/3$, natomiast z (2.23) wynika, że $F_P(1/4) = (1/3 + x_P(5/36), 1/3 - y_P(5/36))$. Dalej, z (2.23) otrzymujemy $F_P(5/36) = (1/3 + x_P(1/36), 1/3 - y_P(1/36))$. I znowu, po elemen-

tarnych przekształceniach mamy $F_P(1/36) = (1/3, 0)$, $F_P(1/4) = (1, 0)$ oraz $F_P(5/36) = (2/3, 1/3)$.

Analogicznie można wyznaczyć:
$F_P(7/36) = (1/3, 0)$,
$F_P(11/36) = F_P(13/36) = F_P(19/36)(2/3, 1/3)$,
$F_P(5/12) = F_P(11/12) = (1, 2/3)$,
$F_P(17/36) = F_P(23/36) = F_P(25/36) = F_P(31/36) = (1/3, 2/3)$,
$F_P(21/36) = (0, 1/3)$,
$F_P(29/36) = F_P(35/36) = (2/3, 1)$.

Zauważmy, że z (2.24) wynika, iż $F_P(1/2) = (1 - x_P(1/6), 1/3 + y_P(1/6))$, natomiast z (2.22) wynika, że $F_P(1/2) = 3\,F_P(1/18)$, a z (2.23) otrzymujemy $F_P(1/6) = (1/3 + x_P(1/18), 1/3 - y_P(1/18))$. Po elementarnych przekształceniach prowadzi to do wyznaczenia $F_P(1/2) = (1/2, 1/2)$, $F_P(1/18) = (1/6, 1/6)$ oraz $F_P(1/6) = (1/2, 1/6)$. Podsumowując, do wyznaczenia dokładnej wartości odwzorowania $F_P(t)$ w punktach $t = i/(4 \cdot 9^k)$, $i = 0, 1, \ldots, 4 \cdot 9^k$, $k = 0, 1, \ldots$ wystarczy znajomość wartości $F_P(1/4) = (1, 0)$, $F_P(1/2) = (1/2, 1/2)$ oraz $F_P(3/4) = (0, 1)$.

Lemat 2.13 *Układ równań (P2)–(P4) wraz z warunkami:*
$F_P(0) = (0, 0)$,
$F_P(1/4) = (1, 0)$,
$F_P(1/2) = (1/2, 1/2)$,
$F_P(3/4) = (0, 1)$
jest równoważny z układem równań (2.22)–(2.24) definiujących krzywą Peano $F_P : I_1 \to I_2$. Układ równań (P2)–(P4), wraz z podanymi warunkami początkowymi, pozwala w skończonej liczbie rekurencji dokładnie wyznaczyć wszystkie punkty odwzorowania, których współrzędne mają skończone trójkowe rozwinięcia, a dokładniej, pozwala wyznaczyć w skończonej liczbie rekurencji wartości $F_P(t)$ dla wszystkich $t \in \{i/(4 \cdot 9^k),\ i = 0, 1, \ldots, 4 \cdot 9^k\}$, dla dowolnego naturalnego k.

Formalny dowód powyższego lematu można otrzymać poprzez indukcję. □

Podstawowe własności dwuwymiarowej krzywej Peano

Z równań (2.22)–(2.24) wynika bezpośrednio, że wartości funkcji $F_P(t)$ znajdują się w kwadracie $[0, 1] \times [0, 1]$, a w szczególności: dla $t \le 1/9$ przyjmują wartości w kwadracie $[0, 1/3] \times [0, 1/3]$; dla $1/9 \le t \le 2/9$ – w kwadracie $[1/3, 2/3] \times [0, 1/3]$; itd. ..., a dla $8/9 \le t \le 1$ przyjmują wartości w kwadracie $[2/3, 1] \times [2/3, 1]$.

Lemat 2.14 *Istnieje wzajemnie jednoznaczne odwzorowanie między pododcinkami $A_i^k = [i/9^k, (i + 1)/9^k]$, $i = 0, \ldots, 9^k - 1$ oraz odpowiednimi podkostkami $Q_i^k = F_P(A_i^k)$ o boku 3^{-k} pokrywającymi I_2, gdzie k jest dowolną liczbą naturalną.*

Dowód lematu wynika ze spełnienia powyższej własności przez krzywe aproksymujące $F_P^k(t)$, $k = 1, 2, \ldots$ □

Bezpośrednio z równań (2.22)–(2.24) wynika następująca własność krzywej Peano:

Lemat 2.15 *Dla każdego* $t_1, t_2 \leq 1/9$ *oraz* $a = 1/9, 2/9, \ldots, 8/9$ *zachodzi*

$$||F_P(t_1) - F_P(t_2)|| = ||F_P(t_1 + a) - F_P(t_2 + a)||. \tag{2.25}$$

Dalej pokażemy, że krzywa Peano spełnia następujący warunek symetrii:

Lemat 2.16 *Dla każdego* $t \in I_1$ *zachodzi*

$$\begin{aligned} x_P(t) + x_P(1-t) &= 1, \\ y_P(t) + y_P(1-t) &= 1. \end{aligned} \tag{2.26}$$

Dowód. Rozpoczniemy od pokazania, że warunek (2.26) jest spełniony przez wszystkie punkty wierzchołkowe kolejnych przybliżeń krzywej $F_P(t)$, czyli punkty wyznaczone dla $T^k = \{i/9^k,\ i = 0, 1, 2, \ldots, 9^k\}$, dla dowolnego naturalnego k. Łatwo sprawdzić, że równość (2.26) jest spełniona dla $k = 1$, a dokładniej dla wszystkich $t \in T^1$.

Dalej pokażemy, że ze spełnienia warunku (2.26) dla pewnego $k \geq 1$ wynika spełnienie go dla $k + 1$, co zakończy pierwszą część dowodu poprzez indukcję.

W dowodzie skorzystamy z układu równań (2.22)–(2.24) definiujących krzywą Peano. Ze względu na symetrię wystarczy pokazać, że równość (2.26) spełniona jest dla $t \leq 1/2$.

Niech $t \in T^{k+1}$ oraz $t \in [4/9, 1/2]$. Wtedy $1 - t \in [1/2, 5/9]$.

Po użyciu równań (2.22)–(2.24) otrzymujemy

$$\begin{aligned} x_P(1-t) &= 1 - x_P(2/3 - t) = 1 - 1/3 - x_P(5/9 - t) = 2/3 - x_P(5 - 9t)/3, \\ y_P(1-t) &= 1/3 + y_P(2/3 - t) = 2/3 - y_P(5/9 - t) = 2/3 - y_P(5 - 9t)/3. \end{aligned}$$

Ponieważ $5/9 - t < 1/9$ oraz $(5/9 - t) \in T^{k+1}$, otrzymujemy $(5 - 9t) \in T^k$. Możemy zatem przyjąć, że:

$$\begin{aligned} x_P(5 - 9t) &= 1 - x_P(9t - 4) = 1 - x_P(t - 4/9) \cdot 3, \\ y_P(5 - 9t) &= 1 - y_P(9t - 4) = 1 - y_P(t - 4/9) \cdot 3. \end{aligned}$$

Po dalszych przekształceniach

$$\begin{aligned} &x_P(1-t) = 1/3 + x_P(t - 4/9) = 1/3 + x_P(t - 1/3) - 1/3 \\ &= 1/3 + 1 - x_P(t) - 1/3 = 1 - x_P(t) \end{aligned}$$

oraz

$$y_P(1-t) = 2/3 - 1/3 + y_P(t-4/9) = 1/3 - (1/3 - y_P(t-1/3))$$
$$= 2/3 + (1/3 - y_P(t)) = 1 - y_P(t).$$

Analogiczne rozważania można przeprowadzić dla pozostałych wartości t, kolejno zakładając, że $t \in [1/3, 4/9], \ldots, [0, 1/9]$. Ponieważ $F_P(t)$ jest funkcją ciągłą (jednostajnie) w I_1, warunek (2.26) można przedłużyć na cały odcinek I_1. □

Z lematu 2.16 wynika następująca własność, z której dalej będziemy wielokrotnie korzystać. W celu uniknięcia niepotrzebnych komplikacji poniższy lemat sformułowany jest w uproszczonej postaci. W rzeczywistości lemat 2.17 jest prawdziwy również dla innych par liczb $0 \leq s, t \leq 1$.

Lemat 2.17 *Dla każdego $t \in I_1$ oraz dla każdego $s = 1/9, 2/9, \ldots 1$ takiego, że $0 \leq s, t \leq 1/3$ lub $1/3 \leq s, t \leq 2/3$, lub $2/3 \leq s, t \leq 1$, spełniona jest jedna z równości:*

$$\begin{aligned} |x_P(s) - x_P(t)| &= x_P(|s-t|), \\ |y_P(t) - y_P(s)| &= y_P(|t-s|). \end{aligned} \tag{2.27}$$

Dowód. Niech $s = 1/9$ oraz $t \leq s$. Wtedy $F_P(s) = (1/3, 1/3)$. Z równań (2.23) wynika, że $1/3 - x_P(t)) = 1/3 - x_P(9t)/3$ oraz $1/3 - y_P(t) = 1/3 - y_P(9t)/3$.

Z lematu 2.16 mamy $x_P(9t) = 1 - x_P(1-9t)$ oraz $y_P(9t) = 1 - y_P(1-9t)$. Po podstawieniu otrzymujemy, odpowiednio, $1/3 - x_P(t) = x_P(9(1/9-t))/3 = x_P(1/9-t)$ i $1/3 - y_P(t) = y_P(9(1/9-t))/3 = y_P(1/9-t)$. Podobnie, dla $t > s = 1/9$. Z (2.23) wynika, że $x_P(t) - 1/3 = 1/3 + x_P(t-1/9) - 1/3 = x_P(t-1/9)$ oraz $y_P(t) - 1/3 = 1/3 - y_P(t-1/9) - 1/3 = -y_P(t-1/9)$.

Dalej, dla $s = 2/9$ oraz $t \leq 1/9$ mamy $2/3 - x_P(t) = 2/3 - x_P(9t)/3 = 2/3 - 1/3(1 - x_P(9t-1)) = 1/3 + x_P(2/9-t) = x_P(2/9-t)$ oraz $y_P(t) = y_P(9t)/3 = 1/3 - y_P(1-9t)/3 = 1/3 - y_P(1/9-t) = y_P(2/9-t)$. Podobnie, dla $s = 2/9$ oraz $t \in [1/9, 2/9]$ mamy $2/3 - x_P(t) = 2/3 - x_P(t-1/9) - 1/3 = 1/3 - x_P(9t-1)/3 = 1/3 - 1/3 + x_P(2-9t)/3 = x_P(2/9-t)$ oraz $y_P(t) = 1/3 - y_P(t-1/9) = 1/3 - y_P(9t-1)/3 = y_P(2-9t)/3 = y_P(2/9-t)$.

Z kolei dla $s = 2/9$ i $t \in [2/9, 1/3]$ otrzymujemy, po podwójnym zastosowaniu własności (2.23), $x(t) - 2/3 = 2/3 + x_P(t-2/9) - 2/3 = x_P(t-2/9)$ oraz $y_P(t) = 1/3 - y_P(t-1/9) = 1/3 - 1/3 + y_P(t-2/9) = y_P(t-2/9)$. Analogiczne postępowanie prowadzi do wykazania słuszności lematu dla pozostałych par liczb s, t. □

Lemat 2.18 *Dla każdego $t \in I_1$ zachodzi*

$$||F_P(t)||^2 \leq 4\,t. \tag{2.28}$$

Dowód. Najpierw pokażemy, że nierówność (2.28) jest spełniona dla wszystkich punktów wierzchołkowych kolejnych przybliżeń krzywej $F_P(t)$ w postaci punktów kwadratu o współrzędnych mających skończone trójkowe rozwinięcia, czyli dla $F_P(t),\ t \in T^k$, gdzie

$$T^k = \{i/(4 \cdot 9^k),\ i = 0, 1, 3, 4, 5, 7, 8, \ldots, 4 \cdot 9^k\}$$

oraz dowolnego naturalnego k. Zauważmy, że żaden zbiór T^k nie zawiera punktów postaci $(i - 1/2)\,9^{-k}$, $i = 1, 2, \ldots, 9^k$. Mimo to zbiór $\cup_k T^k$ $(k = 1, 2, \ldots)$ jest zbiorem gęstym w I_1.

Łatwo sprawdzić, że nierówność (2.28) jest spełniona dla $k = 1$, a dokładniej dla wszystkich $t \in T^1$. Dalej pokażemy, że ze spełnienia warunku (2.28) dla pewnego $k \geq 1$ wynika spełnienie go dla $k+1$, co zakończy pierwszą część dowodu poprzez indukcję.

Niech $t \in T^{k+1}$ oraz $t \leq 1/9$.

Z (2.22) wynika, że $F_P(t) = (x_P(9t)/3, y_P(9t)/3)$. Ponieważ $9t \in T^k$, zatem

$$||F_P(t)||^2 = x_P^2(t) + y_P^2(t) = [x_P^2(9t) + y_P^2(9t)]/9 \leq 1/9 \cdot 4\,(9t) = 4\,t.$$

Dalej, niech $t \in T^{k+1}$ oraz $t \in [1/9, 2/9]$. Z (2.23) wynika, że $x_P(t) = 1/3 + x_P(t - 1/9)$ oraz $y_P(t) = 1/3 - y_P(t - 1/9)$. Ponieważ $t - 1/9 \leq 1/9$, po elementarnych przekształceniach wynikających z pierwszej części dowodu i tego, że dla $t - 1/9 \leq 1/9$ zachodzi $x_P(t - 1/9) \geq 0$, $y_P(t - 1/9) \leq 1/3$, otrzymujemy

$$\begin{aligned} &||F_P(t)||^2 = x_P^2(t) + y_P^2(t) = [1/3 + x_P(t - 1/9)]^2 + [1/3 - y_P(t - 1/9)]^2 \\ &\leq 1/9 + 2/3 \cdot x_P(t - 1/9) + 1/9 + 4\,(t - 1/9) \leq 4/9 + 4\,(t - 1/9) = 4\,t. \end{aligned}$$

Analogiczne rozważania można przeprowadzić dla kolejnych przedziałów I_1. Dla $t \in [2/9, 1/3]$ należy dwukrotnie zastosować własność (2.23). Otrzymamy wtedy $F_P(t) = (2/3 + x_P(t - 2/9), y_P(t - 2/9))$. Stąd wynika, że

$$\begin{aligned} ||F_P(t)||^2 = x_P^2(t) + y_P^2(t) = 4/9 + x_P^2(t - 2/9) + 4/3 \cdot x_P(t - 2/9) \\ + y_P^2(t - 2/9) \leq 8/9 + 4\,(t - 2/9) = 4\,t. \end{aligned}$$

Natomiast dla $t \in [1/3, 4/9]$ korzystamy z własności (2.24). Wtedy zachodzi $F_P(t) = (1 - x_P(t - 1/3), 1/3 + y_P(t - 2/9))$. W konsekwencji otrzymujemy

$$\begin{aligned} &||F_P(t)||^2 = x_P^2(t) + y_P^2(t) \\ &= 1 + x_P^2(t - 1/3) - 2\,x_P(t - 1/3) + 1/9 + 2/3 \cdot y_P(t - 1/3) + y_P^2(t - 2/9) \\ &\leq 1 + 1/9 + 2/3 \cdot 1/3 + 4\,(t - 1/3) = 4\,t. \end{aligned}$$

Analogiczne wnioskowanie możemy przeprowadzić dla $t \in [4/9, 5/9]$.

W końcu, niech $t \in T^{k+1}$ oraz $t \in [5/9, 1]$. Ponieważ zawsze $||F_P(t)||^2 \leq 2$, zatem dla $t \geq 5/9$ otrzymujemy

$$||F_P(t)||^2/t \leq 2 \cdot 9/5 = 18/5 < 4,$$

co kończy pierwszą część dowodu. W celu przedłużenia nierówności (2.28) na cały odcinek I_1 korzystamy z jednostajnej ciągłości $F_P(t)$, co kończy dowód lematu. □

Twierdzenie 2.3.2 *Dla każdego $t_1, t_2 \in I_1$ zachodzi*

$$||F_P(t_1) - F_P(t_2)||^2 \leq \alpha^2 \, |t_2 - t_1|, \tag{2.29}$$

gdzie $\alpha^2 = 8$.

Dowód. Postępowanie dowodowe przeprowadzimy podobnie jak w przypadku twierdzenia 2.2.2. Najpierw pokażemy, że własność (2.29) jest spełniona przez wszystkie punkty wierzchołkowe zdefiniowane tak samo jak w lemacie 2.18, czyli pokażemy, że (2.29) zachodzi dla wszystkich $t_1, t_2 \in T^k$, gdzie

$$T^k = \{i/(4 \cdot 9^k),\ i = 0, 1, 3, 4, 5, 7, 8, \ldots, 4 \cdot 9^k\},\ \ k = 1, 2, \ldots$$

Łatwo sprawdzić, że nierówność (2.29) jest spełniona dla $k = 1$, a dokładniej dla wszystkich $t_1, t_2 \in T^1 = \{0, 1/12, 2/12, \ldots, 11/12, 1\}$. Dalej pokażemy, że ze spełnienia warunku (2.29) dla pewnego $k \geq 1$ wynika spełnienie go dla $k + 1$, co zakończy pierwszą część dowodu przeprowadzaną poprzez indukcję.

Niech $t_1, t_2 \in T^{k+1}$ oraz $t_1, t_2 \leq 1/9$. Z równania (2.22) wynika, że $F_P(t_i) = F_P(9t_i)/3$, $i = 1, 2$. Ponieważ $9t_1, 9t_2 \in T^k$, otrzymujemy

$$\begin{aligned} ||F_P(t_1) - F_P(t_2)||^2 &= [x_P(t_1) - x_P(t_2)]^2 + [y_P(t_1) - y_P(t_2)]^2 \\ &= 1/9 \cdot [x_P(9t_1) - x_P(9t_2)]^2 + 1/9 \cdot [y_P(9t_1) - y_P(9t_2)]^2 \\ &\leq 8/9 \cdot |9t_1 - 9t_2| = 8\,|t_1 - t_2|. \end{aligned}$$

Analogiczny wynik uzyskamy dla $t_1, t_2 \in [1/9, 2/9]$, ..., $t_1, t_2 \in [8/9, 1]$.

Kolejna klasa przypadków będzie dotyczyła sytuacji, gdy oba punkty $F_P(t_1)$ i $F_P(t_2)$ leżą w sąsiednich podkwadratach w I_2, a więc gdy t_1 oraz t_2 należą do sąsiednich podprzedziałów $[i/9, (i+1)/9]$, $[(i+1)/9, (i+2)/9]$, $i = 0, 1, \ldots, 7$.

Bez straty ogólności możemy przyjąć, że $t_1 < t_2$. Niech $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [0, 1/9]$, $t_2 \in [1/9, 2/9]$. Z nierówności trójkąta wnioskujemy, że

$$||F_P(t_1) - F_P(t_2)|| \leq ||(1/3, 1/3) - F_P(t_1)|| + ||F_P(t_2) - (1/3, 1/3)||.$$

Z lematu 2.17 otrzymujemy

$$|1/3 - x_P(t_1)| = x_P(1/9 - t_1),$$
$$|1/3 - y_P(t_1)| = y_P(1/9 - t_1),$$

natomiast z lematu 2.18 wynika, że jeśli

$$D1 = ||(1/3, 1/3) - F_P(t_1)|| = ||F_P(1/9 - t_1||,$$

to $D1 \leq 2\,(1/9 - t_1)^{1/2}$.

Podobnie, niech $D2 = ||F_P(t_2) - (1/3, 1/3)||$. Z lematu 2.17 wynika, że

$$|x_P(t_2) - 1/3| = x_P(t_2 - 1/9),$$
$$|y_P(t_2) - 1/3| = y_P(t_2 - 1/9),$$

natomiast z lematu 2.18 otrzymujemy $D2 \leq 2\,(t_2 - 1/9)^{1/2}$. Ponieważ zawsze zachodzi $a^{1/2} + b^{1/2} \leq (2a + 2b)^{1/2}$, zatem

$$||F_P(t_1) - F_P(t_2)|| \leq [8(t_2 - t_1)]^{1/2}.$$

Dla kolejnej pary podkwadratów, czyli dla $t_1 \in [1/9, 2/9]$ oraz $t_2 \in [2/9, 1/3]$ mamy

$$||F_P(t_1) - F_P(t_2)|| \leq ||(2/3, 0) - F_P(t_1)|| + ||F_P(t_2) - (2/3, 0)||.$$

Z układu równań (2.22)–(2.24) oraz lematu 2.17 wynika, że

$$2/3 - x_P(t_1) = 2/3 - 1/3 - x_P(t_1 - 1/9)) = 1/3 - x_P(9t_1 - 1)/3$$
$$= 1/3 - 1/3(1 - x_P(2 - 9t_1)) = x_P(2/9 - t_1)$$

oraz

$$y_P(t_1) = 1/3 - y_P(t_1 - 1/9) = 1/3 - y_P(9t_1 - 1)/3$$
$$= y_P(2 - 9t_1)/3 = y_P(2/9 - t_1).$$

Dalej, z lematu 2.18 wynika, że jeśli

$$D1 = ||(2/3, 0) - F_P(t_1)|| = ||F_P(2/9 - t_1||,$$

to $D1 \leq 2\,(2/9 - t_1)^{1/2}$.

Podobnie, niech $D2 = ||F_P(t_2) - (2/3, 0)||$. Z równań (2.23) wynika, że

$$x_P(t_2) - 2/3 = 1/3 + x_P(t_2 - 1/9) - 2/3 = x_P(t_2 - 2/9),$$
$$y_P(t_2) = 1/3 - y_P(t_2 - 1/9) = 1/3 - 1/3 + y_P(t_2 - 2/9).$$

Następnie, z lematu 2.18 otrzymujemy $D2 = ||F_P(t_2 - 2/9)|| \leq 2\,(t_2 - 2/9)^{1/2}$. Z ogólnej nierówności $a^{1/2} + b^{1/2} \leq (2a + 2b)^{1/2}$ wynika dalej, że

$$||F_P(t_1) - F_P(t_2)|| \leq [8(t_2 - t_1)]^{1/2}.$$

Analogiczne rozumowanie można przeprowadzić dla kolejnych par sąsiednich podkwadratów I_2.

Niech $t_1, t_2 \in T^{k+1}$ oraz $t_1 \in [0, 1/9]$, $t_2 \in [2/9, 1/3]$. Z nierówności trójkąta otrzymujemy

$$||F_P(t_1) - F_P(t_2)|| \leq ||(1/3, 0) - F_P(t_1)|| + ||F_P(t_2) - (1/3, 0)||.$$

Z lematu 2.17 oraz układu równań (2.22)–(2.24) wynika, że

$$\begin{array}{ll} |1/3 - x_P(t_1)| = x_P(1/9 - t_1), & |1/3 - y_P(t_1)| = y_P(1/9 - t_1), \\ |1/3 - x_P(t_2)| = x_P(t_2 - 1/9), & |1/3 - y_P(t_1)| = y_P(t_2 - 1/9). \end{array}$$

Korzystając z lematu 2.18 oraz ze znanej nierówności $a^{1/2} + b^{1/2} \leq (2a + 2b)^{1/2}$, otrzymujemy dalej

$$||F_P(t_1) - F_P(t_2)|| \leq 2\,(1/9 - t_1)^{1/2} + 2\,(t_2 - 1/9)^{1/2} \leq [8\,(t_2 - t_1)]^{1/2}.$$

Podobne rozumowanie można przeprowadzić również dla przypadków $t_1 \in [1/3, 4/9]$ i $t_2 \in [5/9, 2/3]$ oraz $t_1 \in [2/3, 7/9]$ i $t_2 \in [8/9, 1]$. Pozostałe warianty można sprawdzić, szacując wprost maksymalną wartość ilorazu $||F_P(t_1) - F_P(t_2)||^2/|t_2 - t_1|$. Na przykład, dla $t_1 \in [1/9, 2/9]$ oraz $t_2 \in [1/3, 4/9]$ minimalna wartość $|t_2 - t_1|$ jest równa $1/9$, natomiast maksymalna wartość $||F_P(t_1) - F_P(t_2)||^2$ wynosi $8/9$. Stąd wynika dalej, że

$$||F_P(t_1) - F_P(t_2)||^2/|t_2 - t_1| \leq 8.$$

Po sprawdzeniu wszystkich wariantów możemy wnioskować, że nierówność (2.29) jest spełniona dla każdej pary t_1, t_2, której elementy należą do $\cup_k T^k$ $(k = 1, 2, \ldots)$ – zbioru gęstego w I_1. Z ciągłości $F_P(t)$ wynika dalej spełnienie nierówności (2.29) dla wszystkich $t_1, t_2 \in I_1$, co kończy dowód twierdzenia. □

Należy zauważyć, że stała w warunku Höldera (2.29) nie może mieć wartości mniejszej niż $8^{1/2}$, gdyż istnieją takie pary punktów $t_1, t_2 \in I_1$, dla których warunek (2.29) jest spełniony w postaci równości. Na przykład, dla $t_1 = 1/12$ i $t_2 = 5/36$ otrzymujemy $F_P(1/12) = (0, 1/3)$ oraz $F_P(5/36) = (2/3, 1/3)$, co prowadzi wprost do równości w warunku Höldera.

2.4 Podsumowanie

W niniejszym rozdziale przedstawiona została metoda reprezentacji krzywych wypełniających w postaci układu równań funkcyjnych mającego jednoznaczne rozwiązanie. Przedstawiono rekurencyjne algorytmy wyznaczania współrzędnych danej krzywej, traktowanych jako wartość dwuwymiarowej funkcji określonej dla danego argumentu $t \in I_1$. Należy zwrócić uwagę na fakt, iż algorytmy te nie wymagają tworzenia jakiegokolwiek przybliżenia całej krzywej, działają bowiem w sposób lokalny. Złożoność obliczeniowa algorytmów wyznaczania współrzędnych danej krzywej zależy liniowo od dokładności reprezentacji wartości argumentu t. Ponadto w powyższym rozdziale przeanalizowano podstawowe własności omawianych dwuwymiarowych krzywych wypełniających. Podano także najlepsze wartości stałej występującej w odpowiednich dla każdej z krzywych postaciach warunku Höldera. Stała z warunku Höldera określa stopień zachowywania bliskości przez poszczególne krzywe, co ma istotne znaczenie dla korzystania z krzywych wypełniających w problemach decyzyjnych.

W przypadku krzywej Sierpińskiego wartość stałej jest od dawna znana [124] i równa się 2. W powyższym rozdziale przeanalizowano także własności krzywej Hilberta oraz krzywej Peano. Udowodniono, że najmniejsza wartość stałej występującej w odpowiednim warunku Höldera jest w przypadku krzywej Hilberta równa $6^{1/2}$, natomiast w przypadku krzywej Peano jest równa $8^{1/2}$.

Rozdział 3

Metody konstruowania wielowymiarowych krzywych wypełniających

W rozdziale tym przedstawiono różne, stosowane obecnie, metody definiowania wielowymiarowych krzywych wypełniających. Są one głównie związane ze spojrzeniem na krzywą jako na obiekt geometryczny, który charakteryzuje samopodobieństwo. Wprawdzie w ogólnej definicji krzywej wypełniającej nie ma wymagania samopodobieństwa, jednakże wszystkie znane krzywe wypełniające tę właśnie cechę posiadają. W związku z tym trudno wprost utożsamiać krzywe wypełniające z pewną klasą obiektów fraktalnych, niemniej jednak nie są znane krzywe wypełniające, które miałyby inną, niefraktalną naturę. W przypadku definiowania konkretnej krzywej szczególnie użyteczne są systemy odwzorowań zwężających, które mogą służyć jako metoda definiowania krzywych o dowolnym wymiarze.

W niniejszym rozdziale pokażemy związki odpowiednich systemów iterowanych odwzorowań z układami równań funkcyjnych definiującymi krzywe wypełniające. Wyniki te są oryginalnym wkładem autorki. Także metoda konstrukcji wielowymiarowej krzywej Sierpińskiego ma charakter oryginalnego wyniku.

Na koniec przedstawiono również całkiem inne podejście do problemu definiowania wielowymiarowych krzywych wypełniających, polegające na składaniu odwzorowań dwuwymiarowych. Twierdzenie 3.4.1 wskazuje na słabą przydatność praktyczną tego typu metod, szczególnie w kontekście własności zachowywania przez krzywą bliskości, w porównaniu do metod bezpośredniej konstrukcji wielowymiarowych krzywych, których przykładem mogą być konstrukcje, w których zastosowano układy iterowanych odwzorowań zwężających lub układy równań funkcyjnych zaproponowane w następnym rozdziale. Własności krzywych wskazane w twierdzeniu 3.4.1 są nowym rezultatem sformułowanym przez autorkę.

3.1 Krzywe fraktalne

Pojęcie fraktala zostało wprowadzone przez Mandelbrota i w ciągu ostatnich dwudziestu lat stało się nieodłącznym składnikiem prawie wszystkich dziedzin nauki i techniki [26], [122], [7], [189].

Fraktalem jest obiekt geometryczny, definiowany iteracyjnie, posiadający cechy samopodobieństwa – fragmenty fraktala są pomniejszonymi kopiami całości, ponadto wymiar takiego obiektu nie jest liczbą całkowitą. W 1975 roku Mandelbrot [106] podał jako wyróżnik fraktala jego wymiar. Zgodnie z tą definicją fraktalem jest obiekt, którego wymiar Hausdorffa [48], [90] przekracza wymiar topologiczny. Sam Mandelbrot nie traktował jednak tej definicji zbyt dogmatycznie, gdyż niewątpliwie istnieją obiekty o fraktalnej naturze, to znaczy cechujące się samopodobieństwem, które tej ostatniej definicji nie spełniają. Przykładem takich obiektów są właśnie krzywe wypełniające, traktowane jako ciągłe odwzorowanie F_d odcinka $I = [0, 1]$ w $F_d(I)$, zwarty i ograniczony zbiór w przestrzeni metrycznej np. E^n. W przypadku krzywych wypełniających omawianych w niniejszej monografii, zbiorem $F_d(I_1)$ jest kostka wielowymiarowa I_d. Istnieją jednak także krzywe wypełniające inne obiekty geometryczne poza kostką wielowymiarową, na przykład krzywa zwana smokiem (dragon) Heighway'a [137].

Klasyczna krzywa wypełniająca, traktowana jako obiekt topologiczny, jest jednowymiarowa, natomiast zbiór jej wartości ma wymiar d. Pojęcie wymiaru w topologii ma długą i skomplikowaną historię. Istnieje kilka różnych definicji wymiaru topologicznego, które jednak pokrywają się w przypadku przestrzeni metrycznych ośrodkowych (takich, jak na przykład E^n) [47], [93]. W odniesieniu do potrzeb niniejszej pracy wystarczy utożsamienie wymiaru topologicznego z wymiarem pokryciowym dim wprowadzonym przez Lebesgue'a.

Definicja 3.1 *Przestrzeń X ma wymiar pokryciowy d, jeśli w każde skończone, otwarte pokrycie tej przestrzeni można wpisać pokrycie otwarte rzędu $d+1$ i nie można wpisać pokrycia rzędu d. Przez rząd pokrycia rozumiemy maksymalną liczbę elementów pokrycia, których przekrój jest niepusty (każdy układ złożony z większej liczby elementów jest zbiorem pustym).*

Zgodnie z tą definicją pojedynczy punkt, przeliczalny zbiór punktów, a także zbiór Cantora mają wymiar 0; prosta i krzywe Jordana (właściwe, nieprzecinające się) mają wymiar 1, natomiast E^d oraz kostka I_d mają wymiar d. Wymiar topologiczny zostaje zachowany przy transformacjach, które są homeomorficzne. To właśnie z tej własności – twierdzenia o niezmienniczości wymiaru topologicznego [43] – wynika niemożność istnienia przekształcenia homeomorficznego między odcinkiem i kostką.

Zarówno wymiar Hausdorffa, jak i inne wymiary fraktalne, przyjmują w przypadku krzywej wypełniającej te same całkowitoliczbowe wartości. Mimo to, krzywe te przytaczane są jako interesująca klasa obiektów fraktalnych [106], [7], [122] ze względu na iteracyjne metody ich konstruowania oraz ze względu na samopodobieństwo, które ujawnia się właśnie na etapie konstruowania (definiowania) konkretnej krzywej.

Przy opisie fraktali często używa się pojęcia wymiaru Hausdorffa (Hausdorffa –Besicovitcha) [48], [90]. Wymiar Hausdorffa jest raczej pojęciem teoretycznym i nie bywa używany w praktyce ze względu na duży stopień komplikacji obliczeń [189], [26]. Często bywa on niesłusznie utożsamiany z innymi, prostszymi definicyjnie wymiarami fraktalnymi, takimi jak wymiar samopodobieństwa, wymiar pojemnościowy, wymiar pudełkowy (tzw. *box-couting*). Znane są jeszcze inne wymiary związane z obiektami fraktalnymi, takie jak wymiar informacyjny, wymiar korelacyjny, wymiar Minkowskiego–Bouligand'a czy wymiar Lapunova [48], [7], [26], [189], [90]. Faktem jest, iż mimo różnych definicji, w wielu konkretnych przypadkach wartości tych wymiarów są takie same bądź bezpośrednio wzajemnie przeliczalne.

Reasumując, samopodobna krzywa F_d wypełniająca kostkę I_d jest dość zdegenerowanym obiektem fraktalnym, gdyż wymiar topologiczny $F_d(I_1) = I_d$ jest równy d, podobnie jak wymiar pudełkowy, wymiar Hausdorffa czy wymiar samopodobieństwa. Pozornym paradoksem jest również to, że każdą z klasycznych krzywych wypełniających można przedstawić w postaci granicy ciągu krzywych Jordana (odwzorowań różnowartościowych) o wymiarze topologicznym równym 1, mniejszym od wymiaru fraktalnego (ale i topologicznego) atraktora, czyli kostki I_d.

Wymiar topologiczny krzywej wypełniającej jest równy wymiarowi przestrzeni, w której jest ona zanurzona, w związku z tym jej wymiary fraktalne (nawet wymiar samopodobieństwa) są równe wymiarowi topologicznemu $F_d(I_1)$.

W istocie rzeczy, znacznie ciekawszym problemem niż problem wymiaru samej krzywej jest określenie związku wymiaru fraktalnego (na przykład pudełkowego) zbiorów $E \subset I$ oraz wymiaru fraktalnego obiektów powstałych po przekształceniu ich przez krzywą wypełniającą F_d, czyli wymiaru $F_d(E) \subset I_d$. Problem ten będzie dokładniej zbadany w rozdziale 5.

3.2 Iterowane systemy przekształceń zwężających

Jedną z najbardziej popularnych technik definiowania obiektów fraktalnych są systemy iterowanych odwzorowań zwężających, popularnie nazywanych IFS [48], [7], [90] od używanego w języku angielskim terminu *iterated function system.*

Formalnie IFS zostały wprowadzone przez Hutchinsona [73], niemniej jednak już w latach pięćdziesiątych Wunderlich zastosował tego typu metody do opisu różnego typu krzywych wypełniających podobnych do krzywej Peano [200], [137].

Definicja 3.2 [7] *Odwzorowanie* $S : X \to X$, *gdzie* X *jest zupełną i zwartą przestrzenią metryczną z metryką* ρ, *jest nazywane odwzorowaniem zwężającym w* X, *jeśli istnieje taka liczba* $0 < s < 1$, *że dla każdego* $x, y \in X$

$$\rho(S(x), S(y)) \leq s\,\rho(x, y),$$

przy czym s *jest nazywane współczynnikiem kontrakcji odwzorowania* S.

Dalej, niech $S_1, \ldots, S_n$ będą odwzorowaniami zwężającymi w X ze współczynnikami kontrakcji, odpowiednio $s_1, \ldots, s_n$. Dowodzi się (por. [48], [7]), że istnieje niepusty zbiór zwarty $Z \subset X$ taki, że

$$Z = \cup_{i=1}^{n} S_i(Z),$$

który jest atraktorem systemu $S_1, \ldots, S_n$. Zbiór Z nazywany jest fraktalem generowanym przez system iterowanych odwzorowań $S_1, \ldots, S_n$.

Niech $\mathcal{H}(X)$ oznacza rodzinę wszystkich niepustych, zwartych podzbiorów X. Transformacja $\mathcal{S} = \cup_{i=1}^{n} S_i(E)$, gdzie $E \in \mathcal{H}(X)$ spełnia warunek

$$Z = \cap_{k=1}^{\infty} \mathcal{S}^k(E),$$

dla każdego zbioru $E \in \mathcal{H}(X)$ takiego, że $S_i(E) \subset E$, $i = 1, \ldots, n$.

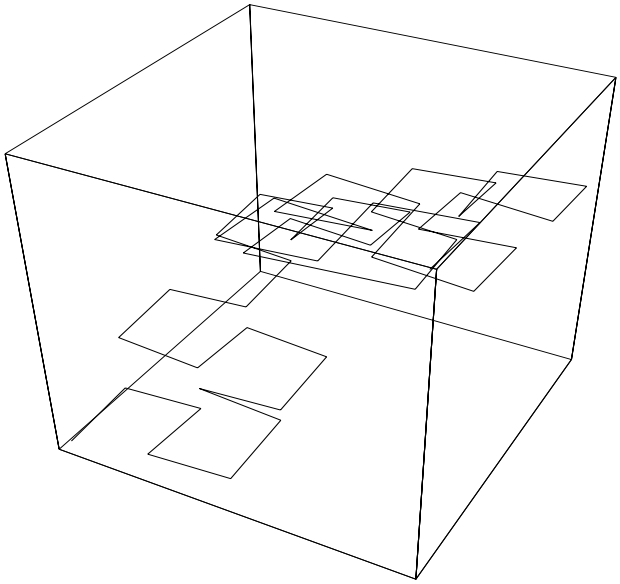
$\mathcal{S}^k$ oznacza tu złożenie k odwzorowań $\mathcal{S}$, czyli $\mathcal{S}^1 = \mathcal{S}$, $\mathcal{S}^k = \mathcal{S} \circ \mathcal{S}^{k-1}$ (dowód por. [48], [7]).

Hutchinson [73], [7] wykazał, że system iterowanych odwzorowań (zwężających): $S_1, S_2, \ldots, S_n$, spełniających warunki

$$S_i(x_n) = S_{i+1}(x_1), \quad i = 1, 2, \ldots, n-1, \tag{3.1}$$

gdzie x_1 oraz x_n są punktami stałymi odwzorowań S_1 oraz S_n, definiuje krzywą (fraktalną).

Jeśli spełniony jest warunek (3.1), to istnieje ciągłe odwzorowanie odcinka jednostkowego w atraktor systemu $S_1, S_2, \ldots, S_n$. Jednakże, w takim ujęciu, IFS nie definiuje bezpośrednio krzywej, lecz zbiór osiąganych przez nią wartości, czyli obraz odcinka jednostkowego $F_d(I_1)$. Opis definicyjny przy użyciu IFS nie pozwala więc na bezpośrednie wskazanie zależności pomiędzy argumentem ($t \in I_1$) oraz wartością odwzorowania ($F_d(t) \in I_d$). Innymi słowy, IFS nie daje przedstawienia parametrycznego krzywej, które jest niezbędne w przypadku stosowania krzywych



Rys. 3.1. Aproksymacja dwuwymiarowej krzywej Hilberta w przestrzeni 3-D
Fig. 3.1. Approximation of the 2-D Hilbert space-filling curve in 3-D

wypełniających do rozwiązywania rozmaitych problemów obliczeniowych, w tym także problemów decyzyjnych.

Trudność tę pozwala rozwiązać metoda definiowania krzywych wypełniających wprowadzona przez Barnsleya [7]. Metoda ta polega na traktowaniu krzywej wypełniającej jako wykresu funkcji $(t, F_d(t)),\ t \in [0,1]$ w przestrzeni $d+1$–wymiarowej. Układ odwzorowań zwężających, spełniający warunek Hutchinsona (3.1), dodatkowo zostaje uzupełniony o system liniowych transformacji (zwężających) działających na odcinku jednostkowym, którego atraktorem jest cały odcinek $[0,1]$. Atraktorem obu systemów odwzorowań, traktowanych łącznie, jest natomiast w tym przypadku pewna krzywa Jordana o nieskończonej długości. Na rysunku 3.1 przedstawiono otrzymaną w ten sposób aproksymację dwuwymiarowej krzywej Hilberta, zanurzonej w kostce I_3.

3.2.1 Krzywa Hilberta jako atraktor systemu odwzorowań zwężających

Równania (2.16)–(2.18) w definicji dwuwymiarowej krzywej Hilberta (por. rozdział 2.2) można zastąpić równoważnymi równaniami postaci:

Z1) $x_H(t) = y_H(4 \cdot t)/2,\ \ y_H(t) = x_H(4 \cdot t)/2,\ \ 0 \le t \le 1/4.$

Z2) $x_H(t) = 1/2 + x_H(4 \cdot t)/2, \;\; y_H(t) = y_H(4 \cdot t)/2, \;\; 1/4 \leq t \leq 1/2.$

Z3) $x_H(t) = 1/2 + x_H(4 \cdot t)/2, \;\; y_H(t) = 1 - y_H(4 \cdot t)/2, \;\; 0 \leq t \leq 1/4.$

Z4) $x_H(t) = y_H(4 \cdot t)/2, \;\; y_H(t) = 1 - x_H(4 \cdot t)/2, \;\; 0 \leq t \leq 1/4.$

Z1) możemy dalej przedstawić w równoważnej postaci:

$$x_H(t/4) = y_H(t)/2, \;\; y_H(t/4) = x_H(t)/2, \;\; t \in I_1.$$

Przekształcając w podobny sposób także (Z2), (Z3) i (Z4), możemy sformułować wniosek, że układowi równań (Z1)–(Z4) odpowiadają cztery pary odwzorowań zwężających $(g_i(x,y), w_i(t))$ postaci:

O1) $g_0(x,y) = (y/2, x/2), \;\; (x,y) \in I_2,$
$w_0(t) = t/4, \; t \in I_1.$

O2) $g_1(x,y) = (1/2 + x/2, y/2), \;\; (x,y) \in I_2,$
$w_1(t) = 1/4 + t/4, \; t \in I_1.$

O3) $g_2(x,y) = (1/2 + x/2, 1 - y/2), \;\; (x,y) \in I_2 \;,$
$w_2(t) = 1/4 + t/4, \; t \in I_1.$

O4) $g_3(x,y) = (y/2, 1 - x/2), \;\; (x,y) \in I_2,$
$w_3(t) = 3/4 + t/4, \;\; t \in I_1,$

które definiują dwuwymiarową krzywą Hilberta za pomocą zbioru iterowanych odwzorowań.

System iterowanych odwzorowań $(g_0, \ldots, g_3)$ spełnia warunek Hutchinsona (3.1), a jego atraktorem jest, co łatwo sprawdzić, kwadrat I_2. System odwzorowań $(w_0, \ldots, w_3)$ działa na odcinku jednostkowym, a jego atraktorem jest cały odcinek I_1. System (O1)–(O4) jest równoważny z opisem dwuwymiarowej krzywej Hilberta przedstawionym w pracy [7]. Podobną równoważność odpowiedniego układu równań funkcyjnych oraz IFS można łatwo pokazać w odniesieniu do wszystkich krzywych omawianych w niniejszej monografii.

Korzystając z IFS, łatwo uogólnić na przypadek wielowymiarowy zarówno dwuwymiarową krzywą Hilberta, jak i dwuwymiarową krzywą Peano. Nowy system odwzorowań musi jednak zawierać 2^d (w przypadku krzywej Hilberta) lub 3^d (w przypadku krzywej Peano) elementów, gdzie d jest wymiarem przestrzeni [137].

3.2.2 Krzywa Sierpińskiego jako atraktor systemu iterowanych odwzorowań

Jak wynika z rozdziału 2.1, istnieje wiele równoważnych reprezentacji dwuwymiarowej krzywej Sierpińskiego. Poniżej przedstawimy kolejną z nich. Ma ona bezpośredni związek z traktowaniem krzywej wypełniającej jako specjalnego rodzaju krzywej fraktalnej.

Lemat 3.1 *Odwzorowanie* $F(t) = (x(t), y(t))$ *określone poprzez równania postaci:*

$$\begin{cases} x(t) = 1/2 - x(4(t+1/8))/2 \\ y(t) = 1/2 - y(4(t+1/8))/2 \end{cases}, \ 0 \le t \le 1/8,$$

$$\begin{cases} x(t) = 1/2 - x(4(t-7/8))/2 \\ y(t) = 1/2 - y(4(t-7/8))/2 \end{cases}, \ 7/8 \le t \le 1,$$

$$\begin{cases} x(t) = 1/2 + x(1-4(t-1/8))/2 \\ y(t) = 1/2 - y(1-4(t-1/8))/2 \end{cases}, \ 1/8 \le t \le 3/8,$$

$$\begin{cases} x(t) = 1/2 + x(4(t-3/8))/2 \\ y(t) = 1/2 + y(4(t-3/8))/2 \end{cases}, \ 3/8 \le t \le 5/8, \tag{3.2}$$

$$\begin{cases} x(t) = 1/2 - x(1-4(t-5/8))/2 \\ y(t) = 1/2 + y(1-4(t-5/8))/2 \end{cases}, \ 5/8 \le t \le 7/8,$$

$$\begin{cases} x(t) = x(1+t) \\ y(t) = y(1+t) \end{cases}, \ t < 0$$

jest równoważne z krzywą Sierpińskiego $F_S(t)$ *w przedziale* $[0, 1]$.

Zauważmy, że ze względu na symetrię omawianej krzywej względem prostej $y = 0,5$, równania odnoszące się do przedziałów argumentu: $t \in [3/8, 5/8]$ i $t \in [5/8, 7/8]$ można zastąpić jedną parą równań postaci

$$\begin{cases} x(t) = x(3/4 - t) \\ y(t) = 1 - y(3/4 - t) \end{cases}, \ 3/8 \le t \le 7/8. \tag{3.3}$$

Opis krzywej Sierpińskiego, sformułowany w postaci układu równań funkcyjnych (3.2), jest ściśle związany z metodą generowania krzywych za pomocą systemu iterowanych odwzorowań zwężających (IFS).

Każdemu z układów równań w (3.2), podobnie jak w przypadku krzywej Hilberta, można przyporządkować wprost parę odwzorowań: z odcinka jednostkowego I_1 i z kostki I_2, odpowiednio, w I_1 oraz w I_2.

Uogólnienie krzywej Sierpińskiego na przypadek wielowymiarowy za pomocą IFS jest znacznie bardziej skomplikowane niż w przypadku krzywej Hilberta czy Peano. W następnym podrozdziale przedstawiono pewne uogólnienie tego typu, którego użyjemy do dokładniejszego zbadania własności krzywych wypełniających.

3.2.3 Zastosowanie IFS do konstrukcji wielowymiarowej krzywej Sierpińskiego

W niniejszym podrozdziale pokażemy oryginalną konstrukcję zbioru par odwzorowań typu IFS, które prowadzą do równoległego podziału d–wymiarowej kostki I_d oraz odcinka jednostkowego I_1. Są one podstawą do zdefiniowania wielowymiarowej krzywej wypełniającej, będącej wielowymiarowym uogólnieniem krzywej Sierpińskiego. Otrzymana krzywa jest identyczna z krzywą F_{Md}, która zostanie zdefiniowana w następnym rozdziale za pomocą układu równań funkcyjnych: (4.21), (4.22), (4.33), (4.24)–(4.25).

Zastosowanie IFS do konstrukcji wielowymiarowych krzywych typu Hilberta czy Peano jest względnie łatwe [7], [137], choć niewątpliwie wymaga wprowadzenia wielu dodatkowych pojęć, które nie są potrzebne w rozwijanym w pracy podejściu, polegającym na tworzeniu układu równań funkcyjnych opisujących geometryczne własności krzywej.

W przypadku krzywej typu Sierpińskiego zadanie to jest jeszcze trudniejsze, wymaga bowiem poszerzenia klasy stosowanych odwzorowań o transformacje, które nie są sensu stricto odwzorowaniami zwężającymi.

Zdefiniujmy najpierw system odwzorowań działających w kostce I_d. Niech W oznacza rodzinę odwzorowań $w_i : R^d \to R^d$ następującej postaci:

$$w_i(x_1, x_2, \ldots, x_d) = \begin{cases} \frac{1}{2} - (\frac{1}{2} - \beta_{1,i}) \cdot x_1 \\ \frac{1}{2} - (\frac{1}{2} - \beta_{2,i}) \cdot x_2 \\ \qquad \ldots \\ \frac{1}{2} - (\frac{1}{2} - \beta_{d,i}) \cdot x_d \end{cases} \tag{3.4}$$

gdzie $\beta_{j,i} \in \{0, 1\}$, $j = 1, 2, \ldots, d$, $i = 0, 1, \ldots, 2^d - 1$.

Rodzina W ma 2^d elementów, które zostały ponumerowane w taki sposób, że wektory $g_{d,i} = (\beta_{1,i}, \beta_{2,i}, \ldots, \beta_{d,i})$, definiujące jednoznacznie dane w_i, są uporządkowane w ten sam sposób jak wierzchołki kostki jednostkowej w schemacie binarnych kodów Graya (por. na przykład [59]). Oznacza to, że $g_{d,i}$ $(i = 0, 1, 2, \ldots, 2^d - 1)$ tworzą listę d–wymiarowych wektorów, zawierających tylko

0 i 1, w której to liście sąsiednie wektory różnią się między sobą dokładnie na jednej pozycji. W sensie geometrycznym taka lista opisuje zamkniętą drogę (hamiltonowską) przechodzącą przez wszystkie wierzchołki d–wymiarowej kostki. Spośród wielu różnych możliwości ograniczymy się tu do porządku definiowanego przez klasyczny, refleksywny (odbity) binarny kod Graya, w którym kod $d+1$–wymiarowy powstaje z d–wymiarowego $G_d = (g_{d,0}, g_{d,1}, \ldots, g_{d,2^d-1})$ w następujący sposób:

$$G_{d+1} = ((g_{d,0},0),(g_{d,1},0),\ldots,(g_{d,2^d-1},0),(g_{d,2^d-1},1),(g_{d,2^d-2},1),\ldots,(g_{d,0},0)).$$

W przypadku d–wymiarowym wprowadza to następującą kolejność uporządkowania wierzchołków kostki I_d:

$$\begin{aligned} &g_{d,0} = (0,0,\ldots,0)\\ &g_{d,1} = (1,0,\ldots,0)\\ &\vdots\\ &g_{d,2^d-b_d} = (1,1,\ldots,1)\\ &\vdots\\ &g_{d,2^d-1} = (0,0,\ldots,0,1) \end{aligned}$$

gdzie b_d jest określone przez następujący schemat obliczeniowy: $b_1 = 1$, $b_k = 2^{k-1} - b_{k-1} + 1$, $k = 2,3,\ldots,d$. Dokładniej, $g_{d,i}$ $(i = 0,1,\ldots,2^d-1)$ jest definiowane rekurencyjnie, jak następuje:

Definicja 3.3 *Niech* $g_{1,0} = (0)$, $g_{1,1} = (1)$.

$$\begin{aligned} &g_{d,i} = (\beta_{1,i},\ldots,\beta_{d-1,i},0), && gdy && i = 0,1,\ldots,2^{d-1}-1,\\ &g_{d,i} = (\beta_{1,2^d-i-1},\ldots,\beta_{d-1,2^d-i-1},1), && gdy && i = 2^{d-1},2^{d-1}+1,\ldots,2^d-1, \end{aligned}$$

gdzie $g_{d-1,i} = (\beta_{1,i},\ldots,\beta_{d-1,i})$.

Lemat 3.2 *$b_d \to \infty$, gdy $d \to \infty$. Wyrażenie $2^{-d}b_d$ zależy od wymiaru d i jego wartość zmienia się od $\frac{1}{2}$ do $\frac{1}{3}$ wraz ze wzrostem d od 2 do nieskończoności.* □

Zauważmy ponadto, że $g_{d,2^d-b_d}$ ma wszystkie współrzędne równe 1. W związku z tym można sformułować następujący prosty lemat:

Lemat 3.3 *Punktem stałym odwzorowania $w_{2^d-b_d}$ jest punkt* $(1,1,\ldots,1)$. □

Każde odwzorowanie $w_i(x_1,\ldots,x_d)$, $(i = 0,1,\ldots,2^d-1)$ transformuje kostkę I_d w podkostkę $\{(x_1,\ldots,x_d) : a_j \le x_j \le e_j, j = 1,\ldots,d\}$, której wszystkie krawędzie mają długość równą $e_j - a_j = 1/2$.

Jednym z wierzchołków $w_i(I_d)$, niezależnie od numeru odwzorowania i, jest punkt $(1/2, 1/2, \ldots, 1/2)$, natomiast wierzchołkiem charakterystycznym (punktem stałym danego odwzorowania) jest zawsze wierzchołek o współrzędnych równych $g_{d,i} = (\beta_{1,i}, \ldots, \beta_{d,i})$. Jak widać, każde odwzorowanie w_i jest transformacją zwężającą określoną na I_d, ze współczynnikiem kontrakcji $1/2$.

Każdy punkt kostki I_d może być rozpatrywany jako obraz innego punktu I_d względem pewnego odwzorowania z rodziny W, czyli

$$\bigcup_{i=0}^{2^d-1} w_i(I_d) = I_d, \quad \text{gdzie} \quad w_i(S) = \{y : \ y = w_i(x), \ x \in S\}, \quad S \subseteq I_d. \tag{3.5}$$

Złożenie $w_{i_1} \circ w_{i_2} \circ \ldots \circ w_{i_n}$ dowolnego n–elementowego ciągu odwzorowań z W, działających na I_d, definiuje podkostkę o krawędziach długości 2^{-n}. Suma wszystkich otrzymanych w ten sposób podkostek pokrywa kostkę I_d. Oznacza to, że dla każdego naturalnego n zachodzi

$$\bigcup_{i_1=0}^{2^d-1} \bigcup_{i_2=0}^{2^d-1} \ldots \bigcup_{i_n=0}^{2^d-1} w_{i_1} \circ w_{i_2} \circ \ldots \circ w_{i_n}(I_d) = I_d. \tag{3.6}$$

Innymi słowy, system odwzorowań:

$$W = \{w_1, w_2, \ldots, w_{2^d-1} : I_d \to I_d\}$$

jest hiperbolicznym systemem iterowanych funkcji (IFS) (por. rozdział 4 w [7] lub [90]). Z zależności (3.5) wynika bezpośrednio, że atraktorem tego systemu odwzorowań jest cała kostka I_d.

Zauważmy, że dla każdego $x \in I_d$ można wyznaczyć (niekoniecznie jednoznacznie) półnieskończony ciąg numerów odwzorowań z rodziny W postaci

$$\sigma = i_1, i_2, \ldots, i_n, \ldots, \quad 0 \le i_j \le 2^d - 1, \ j = 1, \ldots$$

takich, że

$$x \in w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_n}(I_d), \quad n = 1, 2, \ldots.$$

$W_\sigma^n = w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_n}(I_d)$ $(n = 1, 2, \ldots)$ jest malejącym ciągiem niepustych zbiorów domkniętych (w przestrzeni zwartej), w związku z tym ich iloczyn nie jest zbiorem pustym [43].

Ponieważ każde odwzorowanie w_i $(i = 0, \ldots, 2^d - 1)$ jest odwzorowaniem zwężającym, możemy zatem wnioskować, że istnieje dokładnie jeden taki punkt $x \in I_d$, że

$$x = \bigcap_{n=1}^{\infty} W_\sigma^n(I_d). \tag{3.7}$$

Każdy ciąg $\sigma = i_1, i_2, \ldots, i_n, \ldots$, którego elementy i_j są numerami odwzorowań rodziny W, $i_j \in \{0, 1, \ldots, 2^d - 1\}$, definiuje adres jakiegoś punktu z I_d w przestrzeni kodowej $\boldsymbol{\Sigma}$ (por. [7], [90]).

Ponadto, zgodnie z (3.7), każdy adres $\sigma \in \boldsymbol{\Sigma}$ jednoznacznie wyznacza pewien punkt $x \in I_d$.

Zdefiniujmy na odcinku I_1 rodzinę odwzorowań afinicznych

$$H = \{h_0, h_1, \ldots h_{2^d-1} : \ I_1 \to \text{rodzina podzbiorów } I_1\},$$

w której

$$h_i(t) = \frac{b_d}{2^{2d}} + 2^{-d}(i-1) + 2^{-d}t, \ 0 \le t \le 1, i = 1, 2, \ldots, 2^d - 1 \tag{3.8}$$

oraz

$$h_0(t) = \begin{cases} 1 - 2^{-d} + b_d \cdot 2^{-2d} + 2^{-d} \cdot t, \ 0 \le t \le 1 - b_d \cdot 2^{-d}, \\ \\ -2^{-d} + b_d \cdot 2^{-2d} + 2^{-d} \cdot t, \ 1 - b_d \cdot 2^{-d} \le t \le 1. \end{cases} \tag{3.9}$$

Odwzorowania h_i $(i = 1, 2, \ldots, 2^d - 1)$ są niewątpliwie afinicznymi odwzorowaniami zwężającymi I_1 ze względu na zwykłą metrykę euklidesową. Sytuacja komplikuje się w przypadku metryki $\Delta(t_1, t_2) = \min(|t_1 - t_2|, 1 - |t_1 - t_2|)$ rozumianej jako odległość po zamkniętym okręgu o długości jednostkowej.

Dodatkowo, h_0 nie jest funkcją, a raczej jest odwzorowaniem punktów z odcinka I_1 w zbiory punktów z odcinka I_1, gdyż dla $t_h = 2^{-d} - b_d \cdot 2^{-2d}$ wartość $h_0(t_h)$ nie jest określona jednoznacznie. A mianowicie, h_0 odwzorowuje punkt t_h w zbiór dwuelementowy $\{0, 1\}$.

Każde odwzorowanie h_i $(i = 0, 1, \ldots, 2^d - 1)$ przekształca I_1 w domknięty przedział zawarty w I_1 bądź w sumę dwu rozłącznych, domkniętych podprzedziałów I_1 o łącznej długości 2^{-d}.

Z formalnego punktu widzenia H nie jest systemem odwzorowań zwężających, ale, podobnie jak w przypadku rodziny W, rodzina H pokrywa cały odcinek I_1 w tym sensie, że:

$$I_1 = \bigcup_{i_1=0}^{2^d-1} h_{i_1}(I_1). \tag{3.10}$$

Ciąg iterowanych odwzorowań $h_0^{\circ n}(t)$, $t \in I_1$, dla którego $h_0^{\circ n}(t) = h_0 \circ h_0 \circ \ldots h_0(n \text{ razy})(t)$, zbiega się do orbity (z okresem 2) punktów $\{f_1, f_2\}$, czyli $h_0(f_1) = f_2$ i $h_0(f_2) = f_1$. Łatwo sprawdzić, że

$$f_1 = \frac{2^d}{2^{2d} - 1} - \frac{2^d - b_d}{2^d} \cdot \frac{1 + 2^d}{2^{2d} - 1},$$

$$f_2 = \frac{2^{2d}}{2^{2d}-1} - \frac{2^d - b_d}{2^d} \cdot \frac{1+2^d}{2^{2d}-1}.$$

Zbiór $\{f_1, f_2\}$ jest punktem stałym odwzorowania h_0, to znaczy $h_0(\{f_1, f_2\}) = \{f_1, f_2\}$.

Zauważmy, że $h_0 \circ h_0(f_1) = f_1$ i $h_0 \circ h_0$, ograniczone do przedziału $[0, t_h]$, są ciągłymi odwzorowaniami zwężającymi ze współczynnikiem kontrakcji 2^{-2d}, i podobnie $h_0 \circ h_0(f_2) = f_2$ oraz $h_0 \circ h_0$, obcięte do przedziału $[t_h, 1]$, są ciągłymi odwzorowaniami zwężającymi ze współczynnikiem kontrakcji 2^{-2d}.

Każde n–elementowe złożenie odwzorowań ze zbioru $H = \{h_0, h_1, \ldots, h_{2^d-1}\}$ zastosowane w odniesieniu do dowolnego podzbioru I_1 prowadzi nadal do otrzymania podzbioru I_1.

Rozpatrzmy wszystkie możliwe n–elementowe złożenia odwzorowań z H. Z definicji (3.8), (3.9) wynika natychmiast, że suma wszystkich obrazów I_1 powstałych w wyniku działania takich złożonych odwzorowań pokrywa całe I_1, czyli

$$I_1 = \bigcup_{i_1=0}^{2^d-1} h_{i_1}(I_1) = \bigcup_{i_1=0}^{2^d-1} \bigcup_{i_2=0}^{2^d-1} h_{i_1} \circ h_{i_2}(I_1) = \bigcup_{i_1=0}^{2^d-1} \cdots \bigcup_{i_n=0}^{2^d-1} h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(I_1).$$

Stąd możemy wnioskować, że każdy punkt $t \in I_1$ jest skojarzony z półnieskończonym ciągiem liczb $\sigma = i_1, i_2, \ldots, i_n, \ldots$, będących numerami odwzorowań z rodziny H, czyli liczbami całkowitymi ze zbioru $\{0, 1, \ldots, 2^d - 1\}$ i może być przedstawiony w postaci:

$$t \in h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(I_1), \quad n = 1, 2, \ldots$$

Lemat 3.4 *Dla każdego $t \in I_1$ istnieje malejący ciąg niepustych zbiorów zwartych*

$$H_\sigma^n = h_{i_1} \circ h_{i_2} \circ h_{i_3} \circ \ldots \circ h_{i_n}(I_1), \quad H_\sigma^{n+1} \subset H_\sigma^n$$

taki, że

$$t \in \bigcap_{n=1}^{\infty} H_\sigma^n.$$

□

Naturalnie zbiór $\bigcap_{n=1}^{\infty} H_\sigma^n$ nie jest zbiorem pustym, a ze względu na własności h_0, niekoniecznie jest to pojedynczy punkt z I_1.

Ciąg liczb postaci $\sigma = i_1, i_2, \ldots, i_n, \ldots$ tworzy adres $t \in I_1$ w przestrzeni adresów $\boldsymbol{\Sigma}$. W tym momencie należy zauważyć, że istnieją punkty należące do I_1, które mają wiele adresów. Można także znaleźć pary punktów, które mają ten sam adres w przestrzeni $\boldsymbol{\Sigma}$. To ostatnie zjawisko jest konsekwencją istnienia odwzorowania h_0, które w ścisłym sensie nie jest zwężające.

Obie rodziny odwzorowań W oraz H mają tę samą przestrzeń adresów $\boldsymbol{\Sigma}$. Istotne dla dalszych rozważań związanych z konstrukcją krzywej wypełniającej jest traktowanie odwzorowań $h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n} \circ \ldots$ oraz $w_{i_1} \circ w_{i_2} \circ \ldots \circ w_{i_n} \circ \ldots$ związanych z tym samym adresem $\sigma = i_1, i_2, \ldots, i_n, \ldots$ jako odwzorowań działających równolegle, odpowiednio na I_1 oraz I_d.

Niech

$$\boldsymbol{\Sigma}_n = \{(i_1, i_2, \ldots, i_n) : \ \ i_1, \ldots, i_n \in \{0, 1, \ldots, 2^d - 1\}\}$$

oznacza zbiór wszystkich n–elementowych ciągów $(i_1, i_2, \ldots, i_n)$, przy czym $i_j \in \{0, 1, \ldots, 2^d - 1\}\}$, $j = 1, \ldots, n$.

Wszystkie punkty $t \in h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(I_1)$ mają te same adresy w podprzestrzeni adresów $\boldsymbol{\Sigma}_n$ oraz adresy z tą samą n–elementową sekwencją początkową w $\boldsymbol{\Sigma}$.

Podobnie, wszystkie $x \in w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_{n_1}}(I_d)$ z odpowiedniej podkostki I_d mają identyczne adresy w $\boldsymbol{\Sigma}_n$.

Niech

$$R_{nd} = \{B_n : B_n = w_{i_1} \circ w_{i_2} \circ \ldots \circ w_{i_n}(I_d), \ \ i_1, i_2, \ldots, i_n \in \{0, 1, \ldots, 2^d - 1\}\}$$

będzie zbiorem wszystkich podkostek, które są obrazami kostki I_d we wszystkich możliwych n–krotnych złożeniach transformacji pochodzących z rodziny W. Z własności (3.6) wynika bezpośrednio, że dla każdego naturalnego n, zbiór R_{nd} pokrywa I_d.

Podobnie, niech

$$\tilde{P}_{nd} = \{A_n : A_n = h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(I_1), \ \ i_1, i_2, \ldots, i_n \in \{0, 1, \ldots, 2^d - 1\}\}$$

będzie zbiorem wszystkich obrazów I_1, otrzymanych w wyniku działania wszystkich przekształceń będących n–krotnym złożeniem transformacji z rodziny H.

Zauważmy, że

$$\tilde{P}_{1d} = \{h_i(I_1), \ \ i = 0, 1, \ldots, 2^d - 1\} = \{\left[0, b_d 2^{-2d}\right] \cup \left[1 - (2^d - b_d)2^{-2d}, 1\right], \\ \left[b_d 2^{-2d}, 2^{-d} + b_d 2^{-2d}\right], \ldots, \left[b_d 2^{-2d} + (2^d - 2)2^{-d}, b_d 2^{-2d} + (2^d - 1)2^{-d}\right]\}.$$

Definicja 3.4 *Podzbiór I_1, który jest domkniętym odcinkiem lub sumą dwu rozłącznych domkniętych odcinków, będziemy nazywać p-odcinkiem.*

Lemat 3.5 *Każdy element zbioru*

$$\tilde{P}_{nd} = \{h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(I_1), \ \ i_1, i_2, \ldots, i_n = 0, 1, \ldots, 2^d - 1\}$$

jest p-odcinkiem.

Dowód (przez indukcję). Zauważmy, że $t_h = 1 - b_d \cdot 2^{-d}$ jest punktem stałym odwzorowania $h_{2^d - b_d}$. Ponieważ $\tilde{P}_{1d}$ jest zbiorem p-odcinków: $h_i([0,t_h]) \cup h_i([t_h,1])$, gdzie $h_i([0,t_h]) \subset [0,t_h]$ lub $h_i([0,t_h]) \subset [t_h,1]$ oraz $h_i([t_h,1]) \subset [0,t_h]$ lub $h_i([t_h,1]) \subset [t_h,1]$, $i = 0,1,\ldots 2^d - 1$. Załóżmy, że A_n jest dowolnym elementem $\tilde{P}_{nd}$, takim że:

W1. A_n jest p–odcinkiem.

W2. Jeśli A_n składa się z dwu rozłącznych odcinków, to punkt t_h nie należy do A_n.

Łatwo sprawdzić, że dla każdego $A_1 \in \tilde{P}_{1d}$ spełnione są warunki **W1** i **W2**. Należy wykazać, że dowolny $A_{n+1} \in \tilde{P}_{n+1,d}$ także spełnia **W1** i **W2**. Bezpośrednio z definicji wynika, że $\tilde{P}_{n+1,d} = \{h_i(A_n), \quad A_n \in \tilde{P}_{nd}, \quad i = 0,1,2,\ldots,2^d - 1\}$. Jeżeli A_n składa się z pojedynczego przedziału, to także jego obraz $h_i(A_n)$, $(i = 1,2,\ldots,2^d - 1)$ jest pojedynczym przedziałem.

Punkt t_h należy do A_{n+1} tylko wtedy, gdy $A_{n+1} = h_{2^d - b_d}(A_n)$ oraz $t_h \in A_n$, $A_n \in \tilde{P}_{nd}$. W związku z tym $t_h \notin h_0(A_n)$.

$h_0(A_n)$ składa się z dwu rozłącznych przedziałów jedynie w sytuacji, gdy $t_h \in A_n$. W ten sposób pokazaliśmy, że gdy A_n jest pojedynczym domkniętym przedziałem, wówczas warunki **W1** i **W2** są spełnione dla dowolnego $h_i(A_n)$.

Jeśli A_n jest parą przedziałów, to także $h_i(A_n)$ $(i = 1,2,\ldots,2^d - 1)$ jest sumą dwu rozłącznych przedziałów. Zgodnie z **W2**, $t_h \notin A_n$. Transformacja h_0 odwzorowuje dowolny przedział $S \subseteq I_1$ w sumę dwu rozłącznych przedziałów tylko wtedy, gdy t_h jest punktem wewnętrznym S. Z warunku **W2** wynika, że $h_0(A_n)$ jest także sumą dwu rozłącznych domkniętych przedziałów. W związku z tym $h_0(A_n)$ nadal spełnia warunek **W2**, co kończy dowód lematu. □

Na podstawie lematu 3.5 możemy do lematu 3.4 dodać następującą obserwację.

Dla dowolnego $\sigma \in \boldsymbol{\Sigma}$ zbiór $\bigcap_{n=1}^{\infty} H_\sigma^n$ składa się z jednego bądź dwu punktów z I_1. Z powyższej uwagi wynika natychmiast, że każdy adres z przestrzeni adresów $\boldsymbol{\Sigma}$ jest skojarzony z co najwyżej dwoma punktami odcinka jednostkowego I_1.

Lemat 3.6 *Dla każdej pary liczb naturalnych n i $n_1 \leq n$ oraz dowolnego zbioru $A_n = h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_{n_1}} \circ \cdots \circ h_{i_n}(I_1) \in \tilde{P}_{nd}$ można wskazać zbiór $A_{n_1} = h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_{n_1}}(I_1) \in \tilde{P}_{n_1 d}$ taki, że $A_n \subseteq A_{n_1}$.*

Podobnie, dla każdego $B_n = w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_{n_1}} \circ \cdots w_{i_n}(I_d) \in R_{nd}$ istnieje $B_{n_1} = w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_{n_1}}(I_d) \in R_{n_1 d}$, przy czym $B_n \subseteq B_{n_1}$. □

W konsekwencji

$$A_{n_1} = \bigcup_{A_n \in \tilde{P}_{nd} \ \wedge \ A_n \subseteq A_{n_1}} A_n,$$

czyli dowolny p-odcinek A_{n_1} ma pokrycie, którego elementy są p-odcinkami $A_n \in \tilde{P}_{nd}$, $n_1 \leq n$.

Analogicznie,

$$B_{n_1} = \bigcup_{B_n \in R_{nd} \ \wedge \ B_n \subseteq B_{n_1}} B_n,$$

co oznacza, że każda podkostka B_{n_1} może być pokryta przez mniejsze podkostki $B_n \in R_{nd}$, $n_1 \leq n$.

W tym momencie możemy sformułować następujące twierdzenie.

Twierdzenie 3.2.1 *Istnieje wzajemnie jednoznaczne odwzorowanie f_{nd}, które jednoznacznie kojarzy elementy ze zbioru p-odcinków $\tilde{P}_{nd}$ z elementami zbioru R_{nd}, w taki sposób, że*

$$f_{nd}(A_n) = w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_n}(I_d) \in R_{nd}, \ \ gdzie \ \ A_n = h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(I_1).$$

Innymi słowy, f_{nd} przyporządkowuje jednoznacznie p–odcinki $A_n \in \tilde{P}_{nd}$ podkostkom $f_{nd}(A_{nd})$. Ponadto, dla każdego naturalnego $n_1 \leq n$, f_{nd} przyporządkowuje $A_{n_1} \in \tilde{P}_{n_1}$ do kostki $f_{nd}(A_{n_1}) = w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_{n_1}}(I_d) \in R_{n_1 d}$.

Dowód. Wynika bezpośrednio z lematów 3.5 i 3.6. □

Dwa różne p–odcinki A_n i $A_n^{'} \in \tilde{P}_n$ będziemy nazywać przyległymi, jeśli $A_n \cap A_n^{'} \neq \emptyset$.

Lemat 3.7 *$A_n \in \tilde{P}_{nd}$ oraz $A_n^{'} \in \tilde{P}_{nd}$ są przyległe, jeśli n–elementowe adresy punktów ze zbiorów A_n i $A_n^{'}$ różnią się na dokładnie jednej pozycji i różnica ta wynosi 1 modulo 2^d. Ponadto, jeżeli k ($k \leq n$) oznacza numer pozycji, na której oba adresy się różnią, to pozostała część wektora kodującego adres jest postaci $i_{k+1} = 0, i_{k+2} = 2^d - b_d, \ldots i_n = 2^d - b_d$.* □

Z lematu 3.7 wynika, że A_n i $A_n^{'}$ są przyległe w jednym z następujących przypadków:

- gdy n–elementowe adresy punktów z p–odcinków A_n i $A_n^{'}$ różnią się jedynie na ostatniej pozycji, czyli mają postać odpowiednio: $i_1, i_2, \ldots, i_n$ oraz $i_1, i_2, \ldots, i_n + 1 (\mathrm{mod}\, 2^d)$;

- adresy różnią się na przedostatniej pozycji, natomiast ostatnia pozycja kodowa jest w obu przypadkach równa zeru, czyli adresy mają postać $i_1, \ldots, i_{n-1}, 0$ oraz $i_1, \ldots, i_{n-1} + 1 (\mathrm{mod}\, 2^d), 0$;

- adresy różnią się na jednej z wcześniejszych pozycji i pozostała część obu adresów jest równa odpowiednio: $\ldots, i_k, 0, 2^d - b_d, \ldots, 2^d - b_d$ oraz $\ldots, i_k + 1 (\mathrm{mod}\, 2^d), 0, 2^d - b_d, \ldots, 2^d - b_d$.

Kostki B_n i $B_n^{'} \in R_{nd}$ będą nazywane przyległymi, jeśli mają wspólną $(d-1)$–wymiarową ścianę boczną. Innymi słowy, warunek $B_n \cap B_n^{'} \neq \emptyset$ nie wystarcza do zagwarantowania przyległości dwu różnych podkostek B_n i $B_n^{'}$. Przedstawioną poniżej definicję przyległości zaproponował Milne w pracy [110] dotyczącej wielowymiarowej krzywej Peano.

Definicja 3.5 *Dwie podkostki* $B_n = \{x : a_i \leq x_i \leq d_i\}$ *i* $B_n^{'} = \{x : e_i \leq x_i \leq f_i\}$, *gdzie* $d_1 - a_1 = d_2 - a_2 = \cdots = d_d - a_d = f_1 - e_1 = \cdots = f_d - e_d = 2^{-n}$, *są przyległe wtedy i tylko wtedy, gdy istnieje* $i_0 \in \{1, \ldots, d\}$ *takie, że suma* $B_n \bigcup B_n^{'}$ *może być wyrażona w postaci:*

$$B_n \bigcup B_n^{'} = \{a_i \leq x_i \leq d_i, \;\; i = 1, \ldots, d, \;\; i \neq i_0 \;\; h \leq x_{i_0} \leq l\},$$

gdzie $[h, l]$ *równa się* $[a_{i_0}, f_{i_0}]$ *lub* $[e_{i_0}, d_{i_0}]$.

Jest oczywiste, że współrzędne środków dwu przyległych podkostek różnią się tylko na jednej pozycji. Ponadto, jeśli B_n i $B_n^{'}$ są przyległe, to dla każdej pary punktów $x \in B_n$ i $y \in B_n^{'}$ odległość euklidesowa między x i y jest nie większa niż $\sqrt{(2c)^2 + (d-1)c^2} = c\sqrt{d+3}$, gdzie $c = 2^{-n}$.

Ponadto, jeśli B_n i $B_n^{'}$ nie są przyległe, ale istnieje $B_n^{"}$, która jest przyległa zarówno do B_n jak i do $B_n^{'}$, to największa możliwa odległość pomiędzy $x \in B_n$ i $y \in B_n^{'}$ równa się $c\sqrt{d+8}$.

Zauważmy, że jeśli B_n i $B_n^{''}$ są przyległe oraz gdy $B_n^{''}$ i $B_n^{'}$ są także przyległe względem siebie, przy czym środki podkostek B_n, $B_n^{'}$ i $B_n^{"}$ nie są współliniowe, to największa możliwa odległość między $x \in B_n$ i $y \in B_n^{'}$ nie przekracza wartości $c\sqrt{d+6}$.

Zauważmy również, że:

Lemat 3.8 *Każde odwzorowanie* $w_i \in W$ *transformuje przyległe podkostki* B_n *i* $B_n^{'} \in R_{nd}$ *w dwie przyległe podkostki* B_{n+1} *i* $B_{n+1}^{'} \in R_{n+1,d}$. □

Twierdzenie 3.2.2 f_{nd} *przekształca przyległe p-odcinki* A_n, $A_n^{'} \in \tilde{P}_{nd}$, *odpowiednio, w przyległe podkostki* $B_n, B_n^{'} \in R_{nd}$. *Ponadto dla każdego naturalnego* $n_1 \leq n$, *przyległe p–odcinki* $A_{n_1}, A_{n_1}^{'} \in \tilde{P}_{n_1 d}$ *są odwzorowywane, odpowiednio, w przyległe podkostki* $B_{n_1}, B_{n_1}^{'} \in R_{n_1 d}$.

Dowód. Dowód wynika bezpośrednio z faktu, że przyległe p–odcinki A_n i $A_n^{'} \in \tilde{P}_{nd}$ mogą być zdefiniowane, zgodnie z lematem (3.7), jako

$$h_{i_1} \circ \cdots \circ h_{i_k} \circ h_0 \circ (h_{2^d - b_d})^{\circ(n-k-1)}(I_1)$$

oraz

$$h_{i_1} \circ \cdots \circ h_{i_k+1(mod\,2^d)} \circ h_0 \circ (h_{2^d-b_d})^{\circ(n-k-1)}(I_1).$$

Z definicji f_{nd} wynika, że

$$B_n = f_{nd}(A_n) = w_{i_1} \circ \cdots \circ w_{i_k} \circ w_0 \circ (w_{2^d-b_d})^{\circ(n-k-1)}(I_d),$$

$$B_n^{'} = f_{nd}(A_n^{'}) = w_{i_1} \circ \cdots \circ w_{i_k+1(mod\,2^d)} \circ w_0 \circ (w_{2^d-b_d})^{\circ(n-k-1)}(I_d).$$

Różnica między odwzorowaniami w_{i_k} i $w_{i_k+1(mod\,2^d)}$ dotyczy tylko jednej współrzędnej. Odwzorowania te transformują I_d w przyległe względem siebie podkostki, a $w_0 \circ (w_{2^d-b_d})^{\circ(n-k-1)}(I_d) = \left[0, 2^{-n+k}\right] \times \cdots \times \left[0, 2^{-n+k}\right] = [0, 2^{-n+k}]_d$ transformują I_d w dwie przyległe podkostki z R_{n-k+1}.

Dowód kończy obserwacja, że dowolna transformacja $w_{i_1} \circ \cdots \circ w_{i_{k-1}}$ zachowuje przyległość podkostek (por. lemat 3.8). □

Z lematów 3.7, 3.8 oraz twierdzenia 3.2.2 można wywnioskować, że jeśli A_n, $A_n^{''}$ są p–odcinkami równocześnie przyległymi do $A_n^{'}$, to f_{nd} transformuje je w podkostki, odpowiednio: B_n, $B_n^{''}$ i $B_n^{''}$, których środki nie są współliniowe.

Wybierzmy punkt $t_h = 1 - b_d \cdot 2^{-d}$, który jest punktem stałym przekształcenia $h_{2^d-b_d}$, jako punkt charakterystyczny, który będzie podstawą konstrukcji krzywej. W kostce I_d odpowiada mu $(1, 1, \ldots, 1)$ – punkt stały przekształcenia $w_{2^d-b_d}$.

Określmy rekurencyjnie zbiór $P_{nd} \subset I_1$ w taki sposób, że

$$P_{1d} = \bigcup_{i=0}^{2^d-1} h_i(t_h) = \left\{ i \cdot 2^{-d},\ i = 0, 1, \ldots, 2^d \right\}.$$

Dalej, niech

$$P_{2d} = \bigcup_{i=0}^{2^d-1} h_i(P_{1d}) = \{ \bigcup h_{i_1} \circ h_{i_2}(t_h),\ \ i_1, i_2 \in \{0, 1, \ldots, 2^d - 1\} \}$$

oraz analogicznie

$$P_{nd} = \bigcup_{i=0}^{2^d-1} h_i(P_{(n-1)d}) = \left\{ h_{i_1} \circ \cdots \circ h_{i_n}(t_h),\ \ i_1, \ldots, i_n \in \{0, \ldots, 2^d - 1\} \right\}.$$

Można w prosty sposób, przez indukcję, udowodnić następujący rezultat:

Lemat 3.9 *Dla każdego naturalnego n*

$$P_{nd} = \{ i \ \cdot 2^{-nd},\ \ i = 0, 1, \ldots, 2^{nd} \}$$

□

Teraz możemy już przystąpić do zdefiniowania rodziny odwzorowań aproksymujących krzywą wypełniającą I_d. Powiążmy punkty P_{nd} z obrazami $(1, 1, \ldots, 1)$ w n–elementowym złożeniu odwzorowań z rodziny W. Każdemu punktowi $t \in h_{i_1} \circ \cdots \circ h_{i_n}(t_h)$ odpowiada punkt $w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_n}(1, 1, \ldots, 1)$. W konsekwencji otrzymamy odwzorowanie

$$F_{nd}(t) = w_{i_1} \circ w_{i_2} \circ \cdots \circ w_{i_n}(1, 1, \ldots, 1), \quad t \in h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(t_h), \quad t \in P_{nd}. \tag{3.11}$$

Odwzorowanie to można rozszerzyć na cały odcinek I_1:

$$F_{nd}(t) = F_{nd}(t_1) + \frac{t - t_1}{c_{nd}} \frac{F_{nd}(t_2) - F_{nd}(t_1)}{2}, \quad t \in [t_1, t_1 + c_{nd}], \tag{3.12}$$

$$F_{nd}(t) = \frac{F_{nd}(t_1) + F_{nd}(t_2)}{2} + \frac{t - t_1 - c_{nd}}{2^{-nd} - c_{nd}} \frac{F_{nd}(t_2) - F_{nd}(t_1)}{2}, \quad t \in [t_1 + c_{nd}, t_2],$$

gdzie $c_{nd} = b_d\, 2^{-(n+1)d}$ oraz

$$t_1 = k\, 2^{-nd}, \quad t_2 = (k+1)\, 2^{-nd}, \quad k = 0, 1, \ldots, 2^{nd} - 1.$$

Funkcje $F_{nd}(I_1) \subset I_d$ są krzywymi, które mają postać linii łamanych.

Pokażemy dalej, że punkty węzłowe w F_{nd} są także punktami węzłowymi $F_{n+1,d}$. Naszkicowana tu konstrukcja krzywej wypełniającej bazuje w istocie na geometrycznym podejściu Moore'a [111], [110].

Zauważmy, że w definicji odwzorowania F_{nd} występują dwa różne współczynniki skalujące. Każdy przedział $[k\, 2^{-nd}, (k+1)\, 2^{-nd}]$, $k = 0, \ldots, 2^{nd} - 1$, podzielony jest na dwa podprzedziały $[k\, 2^{-nd}, b_d/2^{(n+1)d} + k\, 2^{-nd}]$ i $[b_d/2^{(n+1)d} + k\, 2^{-nd}, (k+1)\, 2^{-nd}]$, które są podzbiorami przyległych p-odcinków z $\tilde{P}_{nd}$. Wprowadzenie dwóch różnych współczynników skalujących pozwala na uzyskanie odwzorowania aproksymującego krzywą wypełniającą, które przekształca przyległe p–odcinki $\tilde{P_{nd}}$ w zbiory zawarte odpowiednio w przyległych podkostkach z R_{nd}.

Definicja F_{nd} może budzić pewne wątpliwości co do jej jednoznaczności, ponieważ istnieją punkty $t \in P_{nd}$, które mają podwójne adresy w $\boldsymbol{\Sigma}_n$. Zauważmy, że dowolny zbiór postaci $h_{i_1} \circ \cdots \circ h_{i_n}(t_h)$ zawiera jeden lub dwa punkty z I_1. W związku z tym można zadać pytanie, jak zdefiniować $F_{nd}(\cdot)$ w wybranym punkcie t_0 w przypadku, gdy $t_0 \in P_{nd}$ i równocześnie $t_0 \in h_{i_1} \circ h_{i_2} \circ \cdots \circ h_{i_n}(t_h)$ oraz $t_0 \in h_{i_1'} \circ h_{i_2'} \circ \cdots \circ h_{i_n'}(t_h)$, gdzie adresy $i_1, \ldots, i_n$ oraz $i_1', \ldots, i_n'$ różnią się na co najmniej jednej pozycji.

Niech $A_n, A_n' \in \tilde{P}_{nd}$ będą przyległymi p–odcinkami, które (oba) zawierają t_0. Z lematu 3.7 wynika, że adresy punktu t_0 mają następującą postać:

$$i_1, \ldots i_k, 0, 2^d - b_d, \ldots 2^d - b_d$$

oraz

$$i_1, \ldots i_k + 1(\mathrm{mod}\, 2^d), 0, 2^d - b_d, \ldots 2^d - b_d.$$

Z drugiej strony, n–elementowe adresy t_0 nie mogą różnić się tylko na ostatniej pozycji, gdyż odwzorowania związane z takimi adresami przekształcają t_h w dwa różne punkty z I_1.

Przypomnijmy, że wierzchołek $(1, 1, \ldots, 1)$ jest punktem stałym odwzorowania $w_{2^d-b_d}$, podczas gdy $w_0(1, \ldots, 1) = (0, 0, \ldots, 0)$. Ponadto dowolne odwzorowanie z rodziny W transformuje $(0, 0, \ldots, 0)$ do postaci $(1/2, 1/2, \ldots, 1/2)$.

Lemat 3.10 *Transformacje* $w_{i_1} \circ \ldots \circ w_{i_k} \circ w_0 \circ w_{2^d-b_d}^{\circ n-k-1}$ *oraz* $w_{i_1} \circ \ldots \circ w_{i_{k+1(mod\, 2^d)}} \circ w_0 \circ w_{2^d-b_d}^{\circ n-k-1}$, $(n = 2, 3, \ldots;\ k = 1, 2, \ldots, n-1)$ *odwzorowują wierzchołek kostki* $(1, 1, \ldots, 1)$ *w ten sam punkt* $(1, 1, \ldots, 1)$. □

Z powyższych rozważań wynika, że F_{nd} jest jednoznacznie zdefiniowane.

Niech $Q_{nd} = F_{nd}(P_{nd})$. Łatwo zauważyć, że

$$Q_{nd} = \{(x_1, \ldots, x_d) \in I_d : x_i = k\, 2^{-n+1},\ k = 0, 1, \ldots, 2^{n-1},\ i = 1, \ldots d\}.$$

Innymi słowy, Q_{nd} jest zbiorem punktów z I_d, których współrzędne mają skończone (n–elementowe) dwójkowe rozwinięcia. Ponadto dla dowolnego naturalnego k, zachodzi $F_{n+k,d}(t) = F_{n,d}(t)$, $t \in P_{nd}$. Stąd, $F_{n+k,d}(P_{nd}) = Q_{nd}$.

F_{nd} ma następującą własność, analogiczną do własności odwzorowania f_{nd}, a mianowicie:

Lemat 3.11 *F_{nd} transformuje każdy p-odcinek $A_n \in \tilde{P}_{nd}$ w $F_{nd}(A_n) \subseteq f_{nd}(A_n) = B_n \in R_{nd}$. Ponadto dowolny p–odcinek $A_{n_1} \in \tilde{P}_{nd}$ jest odwzorowywany w postaci $F_{nd}(A_{n_1}) \subseteq f_{nd}(A_{n_1}) = B_{n_1}$, przy czym $n_1 \leq n$. Jeśli p–odcinki $A_{n_1}, A'_{n_1} \in \tilde{P}_{n_1 d}$ są przyległe, to F_{nd} odwzorowuje je w przyległe podkostki $F_{nd}(A_{n_1}) \subseteq B_{n_1}$, $B_{n_1} \in R_{n_1 d}$ oraz $F_{nd}(A'_{n_1}) \subseteq B'_{n_1}$, $B'_{n_1} \in R_{n_1 d}$, gdzie B_{n_1} i B'_{n_1} są przyległe.* □

Dalej pokażemy, że każda funkcja aproksymująca F_{nd} spełnia warunek Höldera z wykładnikiem $1/d$:

Twierdzenie 3.2.3 *Niech F_{nd} będzie odcinkami liniowym odwzorowaniem I_1 w I_d, zdefiniowanym przez (3.11), (3.12). Wtedy, dla dowolnych $t_1, t_2 \in I_1$ i dowolnego naturalnego n zachodzi:*

$$\| F_{nd}(t_1) - F_{nd}(t_2) \| \leq \alpha_d (| t_1 - t_2 |)^{\frac{1}{d}},\ \ t_1, t_2 \in I_1$$

oraz

$$\| F_{nd}(t_1) - F_{nd}(t_2) \| \leq \alpha_d\, \Delta(t_1, t_2)^{\frac{1}{d}},\ \ t_1, t_2 \in I_1,$$

gdzie $\Delta(t_1, t_2) = \min\{1 - | t_1 - t_2 |, | t_1 - t_2 |\}$ *oraz α_d jest pewną stałą zależną od d.*

Dowód. Dla dowolnej pary punktów $t_1, t_2 \in I_1$ można znaleźć liczbę całkowitą $N \geq 0$, taką że $2^{-(N+1)d} \leq \mid t_1 - t_2 \mid \leq 2^{-Nd}$. Stąd t_1, t_2 spełniają jeden z następujących warunków:

C1) t_1, t_2 należą do tego samego p–odcinka $A_N \in \tilde{P}_{Nd}$,

C2) t_1, t_2 należą do przyległych p–odcinków, oznaczmy je $A_N, A_N^{'} \in \tilde{P}_{Nd}$,

C3) t_1, t_2 należą do dwu różnych p–odcinków $A_N, A_N^{'}$, które są przyległe do trzeciego, oznaczmy go $A_N^{"}$ oraz $A_N, A_N^{'}, A_N^{"} \in \tilde{P}_{Nd}$.

Trzeci przypadek jest możliwy tylko wtedy, gdy $A_N^{"}$ jest sumą rozłącznych przedziałów i zbiory $A_N, A_N^{'}$ mają odpowiednio wspólne punkty końcowe z każdym z nich.

Załóżmy dalej, że $n \geq N$ i $n \geq 1$ przy $N \geq 0$. Zgodnie z lematem 3.11, odwzorowanie F_{nd} transformuje zbiory $A_N, A_N^{'}, A_N^{"}$ (dla $N \leq n$) odpowiednio w przyległe podkostki I_d: $B_N, B_N^{'}, B_N^{"} \in R_{Nd}$. Środki tych podkostek nie są współliniowe, stąd

$$\| F_{nd}(t_1) - F_{nd}(t_2) \| \leq \sqrt{d+6}\, 2^{-N}, \quad 0 \leq N \leq n. \tag{3.13}$$

Biorąc pod uwagę, że $2^{-(N+1)d} \leq |t_1 - t_2|$, otrzymujemy $2^{-N} \leq 2(\mid t_1 - t_2 \mid)^{1/d}$.

Zastosowanie powyższej nierówności w (3.13) kończy dowód pierwszej części twierdzenia dla $\alpha_d = 2\sqrt{d+6}$ i $n \geq N$. Analogiczne rozumowanie można przeprowadzić w odniesieniu do metryki $\Delta(t_1, t_2)$.

Załóżmy teraz, że $n < N$. Rozpatrzmy najpierw przypadek, w którym t_1, t_2 znajdują się bardzo blisko siebie w porównaniu z dokładnością aproksymacji krzywej n. Zauważmy, że $\mid t_1 - t_2 \mid \leq 2^{-Nd} < b_d\, 2^{-(n+1)d} < 2^{-nd}$. Oba punkty t_1 i t_2 znajdują się równocześnie w pewnym p–odcinku $A_n \in \tilde{P}_{nd}$ lub należą odpowiednio do przyległych p–odcinków $A_n, A_n^{'} \in \tilde{P}_{nd}$. Przypomnijmy, że F_{nd} transformuje przyległe p–odcinki $A_n, A_n^{'}$ w przyległe podkostki $B_n, B_n^{'} \in R_{nd}$. Ponieważ F_{nd} jest odcinkami liniowa i transformuje

$$\left[i\, 2^{-nd}, (i + b_d/2^d)\, 2^{-nd}\right] \quad \text{oraz} \quad \left[(i + b_d/2^d)\, 2^{-nd}, (i+1)\, 2^{-nd}\right],$$

odpowiednio w następujące odcinki zawarte w I_d:

$$\left[F_{nd}(i\, 2^{-nd}), (F_{nd}(i\, 2^{-nd}) + F_{nd}((i+1)\, 2^{-nd}))/2\right]$$

oraz

$$\left[(F_{nd}(i\, 2^{-nd}) + F_{nd}((i+1)\, 2^{-nd}))/2, F_{nd}((i+1)\, 2^{-nd})\right],$$

zatem, korzystając z nierówności $b_d\,2^{-d} \geq 1/3$ (por. lemat 3.2), otrzymujemy

$$\| F_{nd}(t_1) - F_{nd}(t_2) \| \leq | t_1 - t_2 | \frac{2^{-n}}{b_d\,2^{-(n+1)d}} \leq 3 \cdot 2^{n(d-1)} \; | t_1 - t_2 |,$$

przy czym

$$| t_1 - t_2 | \leq 2^{-Nd}, \;\; N > n.$$

Dalej, z $| t_1 - t_2 | \leq 2^{-Nd}$ oraz $2^{-N} \leq 2(| t_1 - t_2 |)^{1/d}$ otrzymujemy

$$3 \cdot 2^{n(d-1)} \; | t_1 - t_2 | \leq 6 \cdot 2^{-(N-n)(d-1)} (| t_1 - t_2 |)^{1/d}, \;\; | t_1 - t_2 | \leq 2^{-Nd}, \;\; n < N.$$

Rozpatrzymy teraz przypadek, gdy nadal $n > N$ oraz punkty t_1 oraz t_2 są bliskie sobie w metryce Δ, lecz odległość między nimi w metryce euklidesowej jest duża. Wtedy zachodzi $| t_1 - t_2 | \geq 1 - 2^{-Nd}$ i $\Delta(t_1, t_2) \leq 2^{-Nd}$. Jest to przypadek, kiedy każdy z punktów znajduje się w pobliżu innego końca odcinka jednostkowego. W tej sytuacji t_1 i t_2 należą do tego samego p–odcinka $A_n = h_0 \circ (h_{b_d - 2^d})^{\circ(n-1)}(I_1) = \left[0, b_d\,2^{-(n+1)d}\right] \cup \left[1 - (2^d - b_d)\,2^{-(n+1)d}, 1\right]$. Bez straty ogólności możemy przyjąć, że $t_1 < t_2$. Łącząc równości

$$\| F_{nd}(t_1) - (0, \ldots, 0)) \| = t_1 \frac{2^{-n}}{b_d}\, 2^{(n+1)d}$$

oraz

$$\| F_{nd}(t_2) - (0, \ldots, 0)) \| = t_2 \frac{2^{-n}}{2^d - b_d}\, 2^{(n+1)d}$$

oraz korzystając z nierówności trójkąta, otrzymujemy tę samą nierówność co w poprzednim przypadku. W ten sposób zakończyliśmy dowód twierdzenia. □

Z powyższych rozważań wynika, że F_{nd} jest funkcją hölderowską (czasami nazywaną także funkcją Lipschitza) rzędu $1/d$ ze stałą nie większą niż

$$\max\left\{6 \cdot 2^{-(d-1)}, 2\sqrt{d+6}\right\} = 2\sqrt{d+6} \;\text{ dla }\; d = 2, 3, \ldots$$

W tym momencie możemy już zdefiniować krzywą wypełniającą $F_d(t) : I_1 \to I_d$ jako granicę ciągu odwzorowań F_{nd}.

$$F_d = \lim_{n \to \infty} F_{nd}. \tag{3.14}$$

Lemat 3.12 *Ciąg odwzorowań $\{F_{nd}\}_{n=1}^{\infty}$ jest jednostajnie zbieżny w I_1 przy $n \to \infty$ do ciągłego odwzorowania F_d transformującego I_1 na kostkę I_d, to znaczy $F_d(I_1) = I_d$. Ponadto F_d transformuje każdy p–odcinek $A_n \in \tilde{P}_{nd}$ długości 2^{-nd} na podkostkę $F_{nd}(A_n) = B_n \in R_{nd}$ o objętości 2^{-nd}.*

Dowód. Rozpoczniemy od pokazania, że dla każdego $t \in I_1$ ciąg wartości funkcji $\{F_{nd}(t)\}$ jest ciągiem Cauchy'ego. Wynika to z lematu 3.11, ponieważ dla każdego $\varepsilon > 0$ można znaleźć $N > \log_2(\sqrt{d}/\varepsilon)$ takie, że dla wszystkich $n_1, n_2 \geq N$

$$\| F_{n_1 d}(t) - F_{n_2 d}(t) \| < \varepsilon, \quad t \in I_1.$$

Zauważmy, że N jest niezależne od t. Przypomnijmy również, że funkcje F_{nd} $(n = 1, 2, \ldots)$ są funkcjami ciągłymi w I_1. Stąd, wobec jednostajnej zbieżności F_{nd}, także F_d jest ciągła.

Niech $\hat{P}$ oznacza zbiór wszystkich liczb o skończonych dwójkowych rozwinięciach postaci $k\,2^{-nd}$, $k = 0, 1, \ldots, 2^{nd}$, gdzie n jest dowolną skończoną liczbą naturalną. Podobnie, niech $\hat{Q}$ oznacza wszystkie takie punkty w I_d, które mają skończone dwójkowe rozwinięcia. Zbiór $\hat{Q}$ jest gęsty w I_d. Zgodnie z definicją F_{nd}, F_d odwzorowuje zbiór $\hat{P}$ z przedziału jednostkowego I_1 w zbiór $\hat{Q}$, gęsty w I_d. Zwartość I_1 oraz ciągłość F_d gwarantuje zwartość $F_d(I_1)$. Stąd obraz $F_d(I_1)$ będący domkniętym i gęstym podzbiorem I_d jest równy całej kostce I_d.

Dowód drugiej części twierdzenia wynika bezpośrednio z lematu 3.11 i definicji odwzorowania f_{nd}. □

Z twierdzenia 3.2.3 i lematu 3.12 bezpośrednio wynika:

Twierdzenie 3.2.4 *Niech $F_d : I_1 \to I_d$ będzie zdefiniowane przez (3.14). Wtedy F_d spełnia warunek Höldera z wykładnikiem $1/d$, czyli*

$$\| F_d(t_1) - F_d(t_2) \| \leq 2\sqrt{d+6} \mid t_1 - t_2 \mid^{1/d}, \quad t_1, t_2 \in I_1 \tag{3.15}$$

i równocześnie

$$\| F_d(t_1) - F_d(t_2) \| \leq 2\sqrt{d+6}\,\Delta(t_1, t_2)^{1/d}, \quad t_1, t_2 \in I_1. \tag{3.16}$$

Dowód. Dowód twierdzenia wynika z udowodnionych w twierdzeniu 3.2.3 własności funkcji F_{nd} aproksymujących krzywą F_d. □

Jak wiadomo (por. rozdział 1.5), jeśli krzywa F_d jest odwzorowaniem odcinka I_1 w kostkę I_d i odwzorowanie to jest ciągłe z wykładnikiem Höldera równym β, to $\beta \leq 1/d$. W związku z tym wykładnik $1/d$ z twierdzenia 3.2.4 osiąga największą możliwą wartość.

Zauważmy, że dla każdego skończonego n zachodzi $F_d(t) = F_{nd}(t)$, $t \in P_{nd}$. Stąd, zgodnie z lematem 3.10, $F_d(t) = \lim_{n\to\infty} W_\sigma^n$, $t \in \hat{P}$, gdzie $\sigma \in \boldsymbol{\Sigma}$ jest dowolnym adresem $t \in \hat{P}$.

Lemat 3.13 *Dla każdego $t \in I_1 - \hat{P}$ istnieje dokładnie jeden adres σ związany z t, czyli $t \in \lim_{n\to\infty} H_\sigma^n$.*

Dowód. Niech $t \in \lim_{n\to\infty} H^n_\sigma$ i równocześnie $t \in \lim_{n\to\infty} H^n_{\sigma'}$, gdzie $\sigma = i_1, i_2, \ldots, i_n, i_{n+1}, \ldots$ oraz $\sigma' = i_1, i_2, \ldots, i'_n, i'_{n+1}, \ldots$, czyli σ i odpowiednio σ' różnią się po raz pierwszy na n–tej pozycji. Stąd $t \in h_{i_1} \circ h_{i_2} \circ \ldots \circ h_{i_n}(I_1) = A_n$ oraz $t \in h_{i_1} \circ h_{i_2} \circ \ldots \circ h_{i'_n}(I_1) = A'_n$, przy czym $A_n,\ A'_n \in \tilde{P}_{nd}$. Ponieważ $A_n \cap A'_n \neq \emptyset$, zatem A_n i A'_n muszą być przyległymi p–odcinkami oraz t wspólnym punktem końcowym A_n i A'_n. Naturalnie $t \in h_{i_1} \circ h_{i_2} \circ \ldots \circ h_{i_n} \circ h_0(t_h)$. W konsekwencji, $t \in P_{n+1,d}$, co jest sprzeczne z wyjściowym założeniem. □

Niech Q oznacza zbiór punktów takich, że co najmniej jedna współrzędna punktu ma skończone dwójkowe rozwinięcie. Naturalnie, $\hat{Q} \subset Q$. Jeżeli $x \in Q$, to zawsze istnieją liczba naturalna n i zbiór $B_n \in R_{nd}$, takie że $x \in \partial B_n$, gdzie ∂B_n oznacza brzeg B_n, stąd Q jest zbiorem miary zero.

Lemat 3.14 *Dla każdego $x \in I_d - Q$ istnieje dokładnie jeden adres σ punktu x w przestrzeni adresów $\boldsymbol{\Sigma}$, czyli $x \in \lim_{n\to\infty} W^n_\sigma$.* □

Zgodnie z lematem 3.13, możemy jednoznacznie zdefiniować odwzorowanie $\tilde{F} : I_1 \to I_d$ za pomocą adresów punktów z I_1 oraz I_d.

Definicja 3.6 *Niech $\tilde{F}_d(t) = \lim_{n\to\infty} W^n_\sigma$, gdzie $\sigma \in \boldsymbol{\Sigma}$ jest adresem $t \in I_1$, czyli $t \in lim_{n\to\infty} H^n_\sigma$.*

Lemat 3.15 *Odwzorowanie $\tilde{F}_d : I_1 \to I_d$ jest identyczne z odwzorowaniem $F_d : I_1 \to I_d$ zdefiniowanym w (3.14), a mianowicie $\tilde{F}_d(t) = F_d(t),\ \ t \in I_1$.*

Dowód. Łatwo pokazać, że $\tilde{F}_d(t) = F_d(t)$ dla $t \in \hat{P}$. Przypomnijmy także, iż zbiór $\hat{P}$ jest gęsty w I_1. Zgodnie z twierdzeniem 3.2.4, odwzorowanie F_d obcięte do $\hat{P}$ jest jednostajnie ciągłe. Stąd istnieje jednoznaczne rozszerzenie $F : I_1 \to I_d$ odwzorowania $F_d : \hat{P} \to I_d$. □

Z powyższego lematu wynika, iż jesteśmy w stanie wyznaczyć dokładne wartości odwzorowania F_d we wszystkich punktach odcinka I_1, które mają skończone dwójkowe rozwinięcia. Co więcej, wartości te możemy wyznaczyć w skończonej liczbie operacji.

Niech $\hat{R}_{nd}$ oznacza klasę wszystkich półotwartych podkostek zawartych w I_d, których domknięcia tworzą zbiór R_{nd}. Jeśli $B_n = \{a_1 \le x_1 \le c_1, \ldots, a_d \le x_d \le c_d\} \in R_{nd}$, wtedy odpowiadająca B_n półotwarta podkostka B^o_n z $\hat{R}_{nd}$ jest postaci $B^o_n = \{a_1 \le x_1 < c_1, \ldots, a_d \le x_d < c_d\} \in \hat{R}_{nd}$.

Zauważmy, że brzeg każdej podkostki $B_n \in R_{nd}$ składa się z $2d$ $(d-1)$–wymiarowych hiperkostek. Oznaczmy brzeg B_n przez ∂B_n.

Twierdzenie 3.2.5 *Odwzorowanie $F_d : I_1 \to I_d$ zdefiniowane przez (3.14) zachowuje miarę Lebesgue'a, to znaczy każdy zbiór borelowski $B \subseteq I_d$ spełnia warunek*

$$\mu_d(B) = \mu_1(F_d^{-1}(B)),$$

gdzie μ_d jest miarą produktową w I_d.

Dowód. Ponieważ F_d jest ciągłe, jest także mierzalne. Łatwo sprawdzić, że minimalne σ–ciało generowane przez $\hat{R}_d$ składa się ze wszystkich podzbiorów borelowskich I_d. Korzystając z twierdzenia Caratheodory [15] o przedłużaniu miary, wystarczy pokazać, że dla każdego zbioru

$$B_n^o \in \hat{R}_d = \bigcup_{n=1}^{\infty} \hat{R}_{nd}$$

zachodzi

$$\mu_d(B_n^o) = \mu_1(F_d^{-1}(B_n^o)). \tag{3.17}$$

Zauważmy, że z $B_n^o \subset B_n$ wynika $\mu_d(B_n^o) \leq \mu_d(B_n) = 2^{-nd}$. Ponadto $B_n - \partial B_n \subset B_n^o$ oraz każda kostka B_n ma $2d$ ścian, kostek o wymiarze $(d-1)$. Każda z tych ścian ma punkt wspólny z $(2^k)^{(d-1)}$ różnymi podkostkami zawartymi w B_n i należącymi do $R_{n+k,d}$. Stąd liczba podkostek z $R_{n+k,d}$, które są zawarte w B_n i dodatkowo mają punkty wspólne z brzegiem ∂B_n jest nie większa niż $2\,d\,(2^k)^{(d-1)}$. Objętość dowolnej podkostki z $R_{n+k,d}$ jest równa $[2^{-(n+k)}]^d$. W związku z tym otrzymujemy

$$\mu_d(B_n^o) \geq \mu_d(B_n) - 2\,d\,(2^k)^{(d-1)}\,2^{-(n+k)d} \geq 2^{-nd}(1 - 2d\,2^{-k}),$$

gdzie k jest dowolnie dużą liczbą naturalną. Stąd wynika $\mu_d(\delta B_n) = 0$.

Przypomnijmy, że zgodnie z lematem 3.11, F_d odwzorowuje wzajemnie jednoznacznie elementy ze zbioru $\tilde{P}_{nd}$ w podkostki ze zbioru R_{nd}. Wybierzmy $A_n \in \tilde{P}_{nd}$ tak, że $F_d(A_n) = B_n$. Oznaczmy następnie przez C_n przeciwobraz brzegu ∂B_n. Dalej, niech $D_k \subset I_1$ będzie sumą wnętrz zbiorów A_{n+k}, których odwzorowania $F_d(A_{n+k}) \in R_{n+k,d}$ oraz które nie mają punktów wspólnych z δB_n. Zbiór $I_1 - D_k$ jest zbiorem domkniętym (D_k jest zbiorem otwartym) i dla każdego naturalnego k mamy $C_n \subset I_1 - D_k$. Teraz wystarczy zauważyć, że dla dowolnego naturalnego k zachodzi:

$$0 \leq \mu_1(C_n) \leq \mu_1\left(\bigcap_{k=1}^{\infty}(I_1 - D_k)\right) \leq \mu_1(I_1 - D_k) \leq 2d\,(2^k)^{(d-1)}\,2^{-(n+k)d}.$$

Stąd wynika natychmiast, że $\mu_1(C_n) = 0$.

Niech C oznacza przeciwobraz brzegu wszystkich $3^d - 1$ podkostek z R_{nd}, które mają wspólne pewne punkty brzegowe z kostką B_n. Musi wtedy zachodzić

$$0 \leq \mu_1(C) \leq (3^d - 1)\, 2d\, (2^k)^{(d-1)}\, 2^{-(n+k)d} = (3^d - 1)\, 2d\, 2^{-nd-k}.$$

Ponadto wiadomo, że

$$A_n - C_n \subset F_d^{-1}(B_n^o) \subset F_d^{-1}(B_n) \subset A_n \cup C.$$

Stąd otrzymujemy

$$2^{-nd} = \mu_1(A_n \cup C) \geq \mu_1(F_d^{-1}(B_n^o) \geq \mu_1(A_n - C_n) = 2^{-nd}.$$

Powyższe stwierdzenie jest równoważne z (3.17), co kończy dowód twierdzenia. □

Jak wiadomo, F_d nie jest odwzorowaniem wzajemnie jednoznacznym. Przypomnijmy ponadto, że $\hat{P}$ jest zbiorem punktów takich, że dla każdego $t \in \hat{P}$ istnieje $t' \in \hat{P}$, $t \neq t'$ oraz $F_d(t) = F_d(t')$.

Niech $D_n \subset I_1$ oznacza zbiór punktów, które mają wspólne n–elementowe adresy z innymi punktami z I_1. Dokładniej, jeśli $t \in D_n$, to istnieje $t' \in I_1$, $t \neq t'$ o tym samym n–elementowym adresie co t. W konsekwencji zachodzi $2^{-(n+1)d} < |t - t'| \leq 2^{-nd}$, gdzie $n = 0, 1, \ldots$, gdyż t i t' należą do tego samego p–odcinka $A_n \in \tilde{P}_{nd}$. Ponadto t i t' muszą należeć do tego samego p–odcinka $A_{n+1} \in \tilde{P}_{n+1,d}$, $A_{n+2} \in \tilde{P}_{n+2,d}$, ... W związku z tym adres t i t' składa się z 0 lub $2^d - b_d$ na n-tej i dalszych pozycjach. Stąd wynika, że

$$\mu_1(D_n) < \frac{2^{nd}}{2^{(d-1)k}\, 2^{nd}},\ k = 1, 2, \ldots$$

Oznaczmy przez D zbiór wszystkich punktów w I_1, które nie są jednoznacznie wskazywane przez ich adres, $D = \bigcup_{n=1}^{N} D_n$, $N = 1, 2, \ldots$

Łatwo pokazać, że D jest zbiorem miary zero, ponieważ

$$\mu_1(D) < \lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{2^{(d-1)N}} = \lim_{N\to\infty} \frac{N}{2^{(d-1)N}} = 0.$$

Przypomnijmy, że $\hat{P}$ jest zbiorem wszystkich punktów odcinka I_1, które mają skończone dwójkowe rozwinięcia. Z definicji tej wynika, że $\hat{P} \subset D$. Dalej, niech S_o będzie główną przekątną w kostce I_d, czyli odcinkiem łączącym punkty $(0, \ldots, 0)$ oraz $(1, \ldots, 1)$]. Oznaczmy przez

$$S_{nd} = \bigcup_{i_1=1}^{2^d-1} \cdots \bigcup_{i_n=1}^{2^d-1} w_{i_1} \circ \cdots \circ w_{i_n}(S_o)$$

zbiór punktów, które należą do głównych przekątnych wszystkich podkostek R_{nd}. Dalej, niech S oznacza zbiór wszystkich takich punktów, to znaczy $S = \cup_{i=1}^{n} S_{nd}$ dla dowolnego skończonego n.

Zauważmy, że w_0 jest postaci $x_i := 1/2 - 1/2\, x_i$, $i = 1, 2, \ldots, d$ i odpowiednio $w_{2^d - b_d}$ ma postać $x_i := 1/2 + 1/2 \cdot x_i$, $i = 1, 2, \ldots, d$, stąd, dla każdego naturalnego m, $w_{i_1} \circ \cdots \circ w_{i_m}(I_d) = [a,b] \times [a,b] \times \cdots \times [a,b]$, gdzie $i_1, i_2, \ldots, i_m \in \{0, 2^d - b_d\}$. Naturalnie, $|a - b| = 2^{-m}$. Niech $t \in D$, wtedy adres punktu t jest postaci $i_1, i_2, \ldots, i_n, \ldots$, gdzie dla pewnego naturalnego n: $i_{n+1}, i_{n+2}, \ldots, \in \{0, 2^d - b_d\}$. Z postaci w_0 oraz w_{2^d-1} wynika, że $F_d(D) \subset S$. Ponieważ jednak S_o jest przekątną I_d, zatem $F_d^{-1}(S_o) \subset D$. W konsekwencji zachodzi także $F_d^{-1}(S) \subset D$. Ponieważ $F_d^{-1} F_d(D) \supset D$ oraz $F_d(D) \subset S$, więc $F_d^{-1}(S) \supset F_d^{-1} F_d(D) \supset D$. Jeśli $F_d^{-1}(S) \subset D$ oraz $D \subset F_d^{-1}(S)$, możemy wnioskować, że $F_d^{-1}(S) = D$.

Podobnie otrzymamy równość $F_d(D) = S$. Ponieważ dla dowolnego $A \subset I_1$ zachodzi $F_d F_d^{-1}(A) = A \cap F_d(I_1)$, zatem $F_d F_d^{-1}(S) = S \cap F_d(I_1) = S$. Ponadto mamy $F_d^{-1}(S) = D$.

Niech I_1^o oznacza zbiór punktów z odcinka jednostkowego: $I_1 - D - F_d^{-1}(Q)$. Przypomnijmy, że D oraz $F_d^{-1}(Q)$ są zbiorami miary zero (gdyż F_d jest odwzorowaniem zachowującym miarę Lebesgue'a).

Lemat 3.16 *Odwzorowanie F_d obcięte do I_1^o jest odwzorowaniem wzajemnie jednoznacznym $I_1^o \to I_d$.*

Dowód. Zgodnie z lematem 3.13, istnieje odwzorowanie wzajemnie jednoznaczne między punktami z I_1^o oraz $\boldsymbol{\Sigma}$. Podobnie istnieje odwzorowanie wzajemnie jednoznaczne pomiędzy punktami z $I_d - Q$ oraz adresami z przestrzeni $\boldsymbol{\Sigma}$. Zauważmy, że $F_d(I_1^o) = I_d - F_d(D) - F_d F_d^{-1}(Q) = I_d - S - Q$. Z lematu 3.15 wynika bezpośrednio, że $F_d(t_1) \neq F_d(t_2)$, jeśli $t_1 \neq t_2$ oraz $t_1, t_2 \in I_1^o$. □

Niech $\boldsymbol{\Psi}_d(x) : I_d \to I_1$ będzie odwzorowaniem quasi-odwrotnym w stosunku do F_d, czyli niech $\boldsymbol{\Psi}_d(x) \in F_d^{-1}(x)$, $x \in I_d$. W konsekwencji zawsze $F_d(\boldsymbol{\Psi}_d(x)) = x$, $x \in I_d$. Z poprzednich obserwacji wynika, że $\boldsymbol{\Psi}_d(x)$ może być jednoznacznie określone na zbiorze $I_d - S - Q$ oraz, ponieważ S i Q są zbiorami miary zero, jest odwzorowaniem prawie wszędzie odwrotnym w stosunku do F_d.

Lemat 3.17 *Odwzorowanie $\boldsymbol{\Psi}_d : I_d \to I_1$, spełniające warunek $\boldsymbol{\Psi}_d(x) \in F_d^{-1}(x)$, $x \in I_d$, zachowuje miarę Lebesgue'a.*

Dowód. Najpierw pokażemy, że $\boldsymbol{\Psi}_d$ jest funkcją mierzalną. Funkcja jest mierzalna, gdy istnieje ciąg funkcji prostych, który jest do niej jednostajnie zbieżny. Dotyczy to funkcji ograniczonych, ale z takimi mamy tu do czynienia. Niech $\boldsymbol{\Phi}(x) : I_d \to \boldsymbol{\Sigma}$ będzie odwzorowaniem, które kojarzy z każdym punktem kostki I_d

dokładnie jeden adres σ z $\boldsymbol{\Sigma}$, przy czym $x \in \lim_{n\to\infty} W^n_\sigma$. Zdefiniujmy ciąg funkcji prostych $\boldsymbol{\Psi}_{nd}(x) = \min(h_{i_1} \circ \cdots \circ h_{i_n}(t_h))$ lub $\boldsymbol{\Psi}_{nd}(x) = \max(h_{i_1} \circ \cdots \circ h_{i_n}(t_h))$, $x \in W^n_\sigma$, $\sigma = i_1, \ldots, i_n, \ldots = \boldsymbol{\Phi}(x)$, $x \in I_d$.

Łatwo zauważyć, że ciąg $\boldsymbol{\Psi}_{nd}(x)$ $(n = 1, 2, \ldots)$ jest jednostajnie zbieżny i zgodnie z lematem 3.15, zachodzi $\lim_{n\to\infty} \boldsymbol{\Psi}_{nd}(x) \in F_d^{-1}(x)$.

Pozostało do pokazania, że każdy zbiór borelowski $A \in I_1$ spełnia warunek

$$\mu_1(A) = \mu_d[\boldsymbol{\Psi}_d^{-1}(A)].$$

Zauważmy, że $\boldsymbol{\Psi}_d^{-1}(A) = \{x : \boldsymbol{\Psi}_d(x) \in A,\ x \in I_d\} = F_d(A \cap \boldsymbol{\Psi}_d(I_d))$. Ponadto $\boldsymbol{\Psi}_d^{-1}(A)$ różni się od $F_d(A)$ na zbiorze miary zero, gdyż $\boldsymbol{\Psi}_d(I_d)$ różni się od I_1 na zbiorze miary zero. Z lematu 3.2.5 wynika, że $\mu_1[F_d^{-1}\boldsymbol{\Psi}_d^{-1}(A)] = \mu_d[\boldsymbol{\Psi}_d^{-1}(A)]$. Ponieważ $\boldsymbol{\Psi}_d(x) \in F_d^{-1}(x)$, zatem $\boldsymbol{\Psi}_d\boldsymbol{\Psi}_d^{-1}(A) \subset F_d^{-1}\boldsymbol{\Psi}_d^{-1}(A)$ oraz $\mu_1[\boldsymbol{\Psi}_d\boldsymbol{\Psi}_d^{-1}(A)] \leq \mu_1[F_d^{-1}\boldsymbol{\Psi}_d^{-1}(A)]$. Dalej, $\boldsymbol{\Psi}_d\boldsymbol{\Psi}_d^{-1}(A) = A \cap \boldsymbol{\Psi}_d(I_d) \supset A \cap (I_1 - D - F_d^{-1}(Q))$, tak więc zachodzi również $\mu_1[\boldsymbol{\Psi}_d\boldsymbol{\Psi}_d^{-1}(A)] \geq \mu_1(A)$. W konsekwencji otrzymujemy $\mu_1(A) \leq \mu_d[\boldsymbol{\Psi}_d^{-1}(A)]$.

Z drugiej strony, ponieważ $\boldsymbol{\Psi}_d^{-1}(A) \subset F_d(A)$, mamy także $\mu_d[\boldsymbol{\Psi}_d^{-1}(A)] \leq \mu_d[F_d(A)]$. Ponadto zachodzi $A \subset F_d^{-1}F_d(A) \subset A \cup D \cup F_d^{-1}(Q)$, stąd możemy wnioskować, że $\mu_1(A) = \mu_1[F_d^{-1}F_d(A)]$. Dalej, jeśli $F_d(A)$ jest mierzalnym, to otrzymujemy $\mu_1[F_d^{-1}F_d(A)] = \mu_d[F_d(A)]$, stąd $\mu_d[\boldsymbol{\Psi}_d^{-1}(A)] \leq \mu_1(A)$, co kończy dowód. □

W następnym rozdziale (patrz 4) sformułowany zostanie, między innymi, opis takiej samej wielowymiarowej krzywej Sierpińskiego, w którym zastosowano odpowiedni układ równań funkcyjnych. Opis w postaci układu równań funkcyjnych prowadzi do podobnego algorytmu wyznaczania wartości odwzorowania F_d, mimo że przy definiowaniu krzywej nie korzystano z pojęcia IFS.

3.3 Systemy Lindenmayera

W roku 1968 biolog Aristid Lindenmayer zaproponował matematyczne modele rozwoju biologicznych systemów komórkowych. Modele te stały się podstawą stworzenia języka formalnego, służącego do czasowo-przestrzennego opisu obiektów wykazujących cechy samopodobieństwa. Od nazwiska twórcy nazywane są one systemami Lindenmayera bądź też L-systemami [125], [108]. Obiektami opisywanymi za pomocą L-systemów mogą być nie tylko żywe organizmy bądź ich fragmenty, ale i abstrakcyjne twory geometryczne, czy też ogólniej symboliczne systemy dynamiczne. L-systemy okazały się także bardzo wygodnym narzędziem opisu struktur fraktalnych, w tym szczególnie krzywych wypełniających. W przypadku L-systemu otrzymujemy prosty przepis przetwarzania krzywej aproksymującej w celu uzyskania kolejnego przybliżenia krzywej wypełniającej. W ogólnym

przypadku proces generowania krzywej przebiega iteracyjnie i, z natury rzeczy, może być kontynuowany dowolnie długo, natomiast jego asymptotyczne własności nie są brane pod uwagę. W związku z tym opis krzywej wypełniającej, utworzony za pomocą L-systemu, jest przydatny raczej w przypadku generowania graficznych przedstawień krzywej niż w algorytmach przetwarzania danych [26].

3.4 Iterowane krzywe dwuwymiarowe

Krzywą wypełniającą wielowymiarową kostkę można otrzymać poprzez odpowiednie iterowanie krzywej wypełniającej kwadrat jednostkowy [146], [180], [179], [137].

Pokażemy tutaj, że krzywe wypełniające uzyskane tą drogą tylko w bardzo szczególnych przypadkach charakteryzują się optymalną wartością wykładnika w warunku Höldera, pożądaną ze względu na stopień zachowywania bliskości odwzorowania.

Niech $F(t) = (x(t), y(t))$ będzie pewną dwuwymiarową krzywą wypełniającą kwadrat, na przykład dwuwymiarową krzywą Hilberta, Peano czy Sierpińskiego. Odwzorowanie to spełnia warunek Höldera

$$||F(t_2) - F(t_1)|| \leq \alpha^{1/2}|t_2 - t_1|^{1/2}, \quad t_1, t_2 \in [0,1]. \tag{3.18}$$

Ponadto $F(t)$ spełnia równanie

$$||F(t)|| = p||F(t/p^2)||, \; t \in [0,1], \tag{3.19}$$

gdzie $p > 0$ jest stałą zależną od rodzaju krzywej. Łatwo sprawdzić, że wszystkie omawiane w poprzednim rozdziale krzywe spełniają warunek (3.19) odpowiednio dla $p = 2$ (krzywa Hilberta lub Sierpińskiego) lub $p = 3$ (krzywa Peano).

Odwzorowania zdefiniowane na przykład przez złożenia typu:

$$G_3(t) = (x(t), x(y(t)), y(y(t))),$$

$$G_4(t) = (x(x(t)), y(x(t)), x(y(t)), y(y(t))),$$

$$G_d(t) = (x(t), x(y(t)), x(y(y(t))), \ldots, x(y^{\circ d-1}(t))),$$

są krzywymi wypełniającymi odpowiednio I_3, I_4, I_d, przy czym $y^{\circ k}$ oznacza k–krotne złożenie funkcji $y(t)$. Warunkiem otrzymania krzywej wypełniającej jest stochastyczna niezależność funkcji opisujących jej współrzędne [181], [137]. W szczególności, odwzorowania $G_3(t)$, $G_4(t)$, $G_d(t)$, zdefiniowane powyżej, spełniają właśnie ten warunek [146], [180], [179], [137].

Twierdzenie 3.4.1 *Niech* $G(t) = (v_{n1,1} \circ \ldots \circ v_{1,1}(t), v_{n2,1} \circ \ldots \circ v_{1,2}(t), \ldots, v_{nd,1} \circ \ldots \circ v_{1,d}(t))$ *będzie* d*–wymiarową krzywą wypełniającą, przy czym dla każdego* i, j*, przekształcenie* $v_{i,j}(t)$ *jest tożsamościowo równe jednemu z odwzorowań składowych krzywej wypełniającej* $F(t) = (x(t), y(t))$*, czyli* $x(t)$ *lub* $y(t)$*. Ponadto załóżmy, że* $F(t)$ *spełnia warunki (3.18) oraz (3.19). Krzywa wypełniająca* $G(t)$ *spełnia nierówność*

$$||G(t_2) - G(t_1)|| \leq \hat{\alpha}|t_2 - t_1|^r, \quad t_1, t_2 \in [0,1], \tag{3.20}$$

gdzie $r \leq 2^{-K}$ *,* $K = \max_j n_j$*, natomiast* $\hat{\alpha}$ *jest pewną stałą zależną typu krzywej dwuwymiarowej* $F(t) = (x(t), y(t))$ *oraz wymiaru* d*.*

Dowód. Przypomnijmy, że dwuwymiarowa krzywa wypełniająca $(x(t), y(t))$ spełnia warunek (3.19), który, korzystając z oznaczeń z twierdzenia 3.4.1, możemy zapisać w ogólnej postaci jako

$$v(t/p^2) = v(t)/p. \tag{3.21}$$

Dalej, niech $t_i = t_0 p^{-2^K i}$, $i = 0, 1, 2, \ldots$, będzie nieskończonym ciągiem punktów z odcinka I_1. Korzystając z własności (3.21), otrzymujemy dla $j = 1, 2, \ldots, d$:

$$\begin{aligned} &v_{n_j,j} \circ \ldots \circ v_{1,j}(t_i) = v_{n_j,j} \circ \ldots v_{2,j}(v_{1,j}(t_0)/p^{2^{K-1} \cdot i}) \\ &= v_{n_j,j} \circ \ldots \circ v_{1,j}(t_0)/p^{2^{K-n_j} i} \leq v_{n_j,j} \circ \ldots \circ v_{1,j}(t_0)/p^i, \end{aligned} \tag{3.22}$$

stąd $||G(t_i)|| \leq ||G(t_0)||/p^i = t_i^{1/2^K} ||G(t_0)||/t_0^{1/2^K}$. Ponadto zachodzi

$$||G(t_i)|| \geq v_{K,j} \circ \ldots \circ v_{1,j}(t_0)/p^i, \quad i = 0, 1, \ldots, \tag{3.23}$$

gdzie $v_{K,j} \circ \ldots \circ v_{1,j}(t_0)/p^i$ jest wartością dowolnej współrzędnej $G(t_i)$, dla której $n_j = K$.

Z równania (3.19) wynika, że $F(0) = (0,0)$, a w konsekwencji

$$G(0) = (0, 0, \ldots, 0).$$

Załóżmy teraz, że istnieje taka stała $\beta < \infty$, że dla każdego zdefiniowanego wcześniej t_i zachodzi $||G(t_i)|| = ||G(t_i) - G(0)|| \leq \beta\, t_i^r$, przy czym $r > 2^{-K}$. Gdyby taka stała β istniała, musiałaby spełniać warunek

$$\begin{aligned} &\beta \geq ||G(t_i)||/t_i^r \geq v_{K,j} \circ \ldots \circ v_{1,j}(t_0)/t_0^r\, p^{-i}\, p^{2^K i\, r} \\ &= v_{K,j} \circ \ldots \circ v_{1,j}(t_0)/t_0^r \cdot p^{(2^K r - 1)\, i}. \end{aligned} \tag{3.24}$$

Wartość $v_{K,j} \circ \ldots \circ v_{1,j}(t_0)/t_0^r$ nie ulega zmianom, natomiast jeśli $r > 2^{-K}$, to przy i dążącym do nieskończoności, także wyrażenie $p^{(2^K r-1)\,i}$ dąży do nieskończoności.

W związku z tym wartość β nie może być ograniczona, co przeczy wcześniejszemu założeniu. W konsekwencji, wykładnik r w nierówności (3.20) nie może być większy niż 2^{-K}, co kończy dowód twierdzenia. □

Jak wiadomo [110], maksymalna wartość wykładnika r w warunku Höldera jest równa odwrotności wymiaru kostki, którą krzywa wypełnia, czyli $r \leq 1/d$. W powyższym twierdzeniu K oznacza maksymalną liczbę złożeń elementarnych funkcji $x(t)$, $y(t)$, występujących w definicji krzywej iterowanej $G(t)$. W przypadku d-wymiarowym minimalna wartość K nie może być mniejsza niż $\lceil \log_2 d \rceil$, gdyż tylko wtedy każda z d współrzędnych opisana jest przez inną funkcję złożoną – kombinację $x(t)$ oraz $y(t)$.

Twierdzenie 3.4.1 mówi o tym, że wartość wykładnika $r = 2^{-K}$ nie może być zwiększona. W konsekwencji, poza szczególnymi przypadkami, gdy K jest dokładnie równe $\log_2 d$, krzywe wypełniające zdefiniowane metodą iterowania dwuwymiarowych odwzorowań nie osiągają maksymalnej wartości wykładnika w warunku Höldera równej $1/d$. Tylko w sytuacji, gdy $d = 4,\ 8,\ \ldots,2^K,\ K \geq 2$, można zdefiniować w ten sposób krzywą wypełniającą kostkę I_d, która należy do klasy funkcji hölderowskich rzędu $1/d$. Innymi słowy, technika składania odwzorowań dwuwymiarowych prowadzi wprawdzie do konstruowania wielowymiarowych krzywych wypełniających, lecz są to krzywe o na ogół gorszych własnościach zachowywania bliskości.

3.5 Podsumowanie

W rozdziale przedstawiono przegląd różnych sposobów definiowania wielowymiarowych krzywych wypełniających. Samopodobną krzywą F_d wypełniającą kostkę I_d można traktować jako obiekt fraktalny. W związku z tym omówiono jedną z najbardziej popularnych technik definiowania obiektów fraktalnych, systemy iterowanych odwzorowań zwężających, tzw. IFS, jako narzędzie do generowania krzywych wypełniających. W szczególności, przedstawiono metodę konstrukcji wielowymiarowej krzywej Sierpińskiego za pomocą odpowiedniego układu IFS. Proponowana konstrukcja jest oryginalnym wkładem autorki w dziedzinę. Cechą charakterystyczną proponowanej metody jest bardzo precyzyjny wybór zbioru początkowego, który podlega dalszemu iterowaniu. Zbiór ten składa się z punktów stałych dwóch skojarzonych ze sobą odwzorowań (zdefiniowanych w kostce I_d oraz na odcinku I_1). Wybór ten pozwala na wyznaczanie w skończonej liczbie kroków dokładnych wartości atraktora IFS, którym jest krzywa wypełniająca.

W powyższym rozdziale wskazano także na związki odpowiednich systemów iterowanych odwzorowań z układami równań funkcyjnych definiującymi te same krzywe wypełniające. Wyniki te są również oryginalnym wkładem autorki.

Na koniec przedyskutowano metodę konstruowania wielowymiarowych krzywych wypełniających, polegającą na składaniu odwzorowań dwuwymiarowych. W twierdzeniu 3.4.1 wskazano na słabą przydatność praktyczną tego typu metod, szczególnie w kontekście własności zachowywania przez krzywą bliskości, w porównaniu do metod bezpośredniej konstrukcji wielowymiarowych krzywych, których przykładem mogą być konstrukcje stosujące układy iterowanych odwzorowań zwężających lub rozwijane w następnym rozdziale metody opisu wielowymiarowych krzywych w postaci układu równań funkcyjnych.

Rozdział 4

Wielowymiarowe krzywe wypełniające

W rozdziale tym zaproponowane zostaną oryginalne opisy definicyjne wielowymiarowych krzywych wypełniających kostkę I_d, podane w postaci odpowiednich układów równań funkcyjnych, których rozwiązaniem jest odwzorowanie w postaci konkretnej krzywej wypełniającej. Krzywe te są uogólnieniami dwuwymiarowych krzywych Hilberta, Peano i Sierpińskiego przedstawionych w rozdziale 2.

Wspomniane układy równań funkcyjnych oparte są bezpośrednio na postulowanych własnościach geometrycznych poszczególnych krzywych. Niektóre z tych własności są od dawna znane [111], [137], nie były one jednak używane do definiowania krzywych. Jedynym wyjątkiem jest wspomniana w rozdziale 2 praca Sierpińskiego [145], w której rozważana jest dwuwymiarowa krzywa wypełniająca, nazwana potem krzywą Sierpińskiego.

W każdym z przypadków wyjściowy układ równań funkcyjnych zostanie przekształcony do równoważnej postaci, umożliwiającej jego efektywne rozwiązanie dzięki odpowiedniej procedurze rekurencyjnej. Procedury te pozwalają na wyznaczenie dokładnych wartości odwzorowań na zbiorach gęstych w I_1. Naturalnie, zbiory te są różne dla różnych typów krzywych. Ponadto odpowiedni zbiór wartości odwzorowania jest za każdym razem pewnym zbiorem gęstym w I_d.

Z jednostajnej ciągłości badanych odwzorowań (spełnienie odpowiednich warunków Höldera na zbiorze gęstym w I_1) wynika możliwość jednoznacznego przedłużenia każdego z odwzorowań na cały odcinek I_1.

W rozdziale tym zbadano również podstawowe własności zaproponowanych odwzorowań, w szczególności odpowiednie szczegółowe sformułowania warunku Höldera spełnianego przez poszczególne krzywe wielowymiarowe.

4.1 Wielowymiarowa krzywa Hilberta

W przypadku definiowania krzywej Hilberta wypełniającej wielowymiarową kostkę I_d należy raczej mówić o całej rodzinie wielowymiarowych krzywych Hilberta, której reprezentanci charakteryzują się podobnymi, lecz nie identycznymi, własnościami geometrycznymi.

Ogólną koncepcję definiowania wielowymiarowej krzywej Hilberta, w oparciu o różnego typu kody Graya, przedstawił Gilbert [60]. W tym miejscu należy odwołać się do ogólnego opisu konstrukcji krzywej wypełniającej przytoczonego w podrozdziale 1.5 oraz szczegółów konstrukcji krzywej Sierpińskiego z rozdziału 3. Użycie różnych typów kodów Graya do porządkowania podziału kostki wielowymiarowej prowadzi do otrzymania różnych wariantów wielowymiarowych krzywych wypełniających [60], [59]. W niniejszej pracy ograniczymy się do porządku definiowanego przez klasyczny, refleksywny (odbity) binarny kod Graya, którego opis znajduje się w podrozdziale 3.2.3.

Przedstawimy dalej nową definicję wielowymiarowej krzywej Hilberta, która zawiera, ze względu na własności symetrii, z których korzystamy, jedynie $d+1$ wielowymiarowych równań funkcyjnych. Niech $F_{Hd}(t) = (x_1(t), x_2(t), \ldots, x_d(t))$, $t \in [0,1]$ oznacza odwzorowanie przeprowadzające punkty z odcinka w punkty hyperkostki $I_d = [0,1]\times[0,1]\times\ldots\times[0,1]$. W szczególnym przypadku, gdy wymiar kostki (d) jest równy trzy, równania definiujące krzywą wypełniającą mają postać:

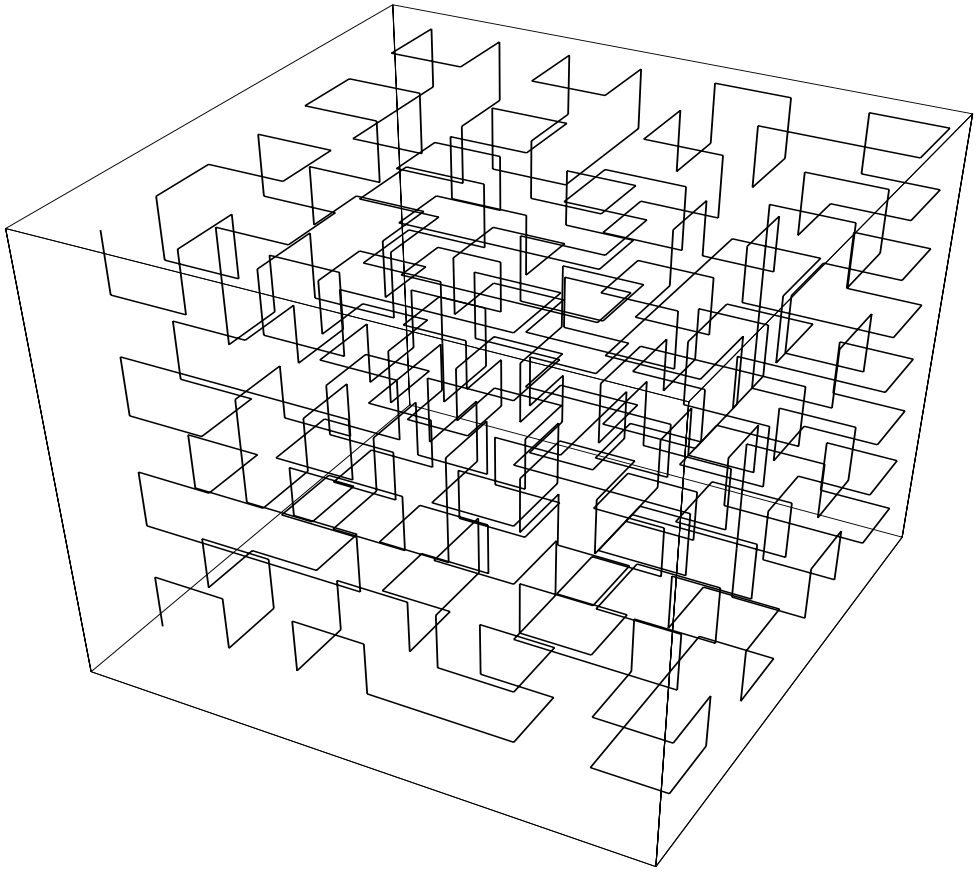
$$\left.\begin{array}{l} x_1(t) = x_3(8t)/2 \\ x_2(t) = x_1(8t)/2 \\ x_3(t) = x_2(8t)/2 \end{array}\right. , \quad 0 \le t \le 1/8, \tag{4.1}$$

$$\left.\begin{array}{l} x_1(t) = 1/2 + x_3(1/4 - t) \\ x_2(t) = 1/2 - x_1(1/4 - t) \\ x_3(t) = x_2(1/4 - t) \end{array}\right. , \quad 1/8 \le t \le 1/4, \tag{4.2}$$

$$\left.\begin{array}{l} x_1(t) = x_1(1/2 - t) \\ x_2(t) = 1/2 + x_3(1/2 - t) \\ x_3(t) = 1/2 - x_2(1/2 - t) \end{array}\right. , \quad 1/4 \le t \le 1/2, \tag{4.3}$$

$$\left.\begin{array}{l} x_1(t) = x_1(1 - t) \\ x_2(t) = x_2(1 - t) \\ x_3(t) = 1 - x_3(1 - t) \end{array}\right. , \quad 1/2 \le t \le 1. \tag{4.4}$$

Łatwo sprawdzić, że $F_{H3}(0) = (0,0,0)$, a $F_{H3}(1) = (0,0,1)$. Wartość odwzorowania w punkcie $1/2$, czyli $F_{H3}(1/2) = (0,1/2,1/2)$ jest charakterystyczna dla tej wersji krzywej. Na rysunku 4.1 przedstawiono aproksymację krzywej F_{H3}



Rys. 4.1. Aproksymacja krzywej Hilberta w 3-D
Fig. 4.1. Approximation of the Hilbert space-filling curve in 3-D

powstałą przez połączenie punktów z kostki odpowiadających wartościom argumentów $t = i/8^3 + 3/8^4$, $i = 0, \ldots, 8^3 - 1$. Środkiem kostki I_3 jest punkt odpowiadający argumentowi $t = 3/8$.

Istnieje kilka różnych algorytmów generowania wielowymiarowych krzywych Hilberta. Niektóre z nich, stosowane w przetwarzaniu obrazów [147], [74], dotyczą nie tyle idealnych krzywych, ile dyskretnych odwzorowań, które nie posiadają wymaganych własności (porównaj rozdział 4.5). Ciągłą wersję odwzorowania pozwala uzyskać algorytm Butza [24], [25]. Jest to bitowo zorientowany, rekurencyjny algorytm, który ma podobne własności jak przedstawiony w niniejszym podrozdziale algorytm będący rozwiązaniem odpowiedniego układu równań funkcyjnych. Podobnie użyteczne, naturalnie w odpowiedniej implementacji, mogą być algorytmy, w których zastosowano IFS [137], [7], [48].

Trójwymiarowa krzywa Hilberta, zdefiniowana za pomocą IFS przez Sagana [137], oraz krzywa generowana przez algorytm Butza [25] są różnymi wersjami wielowymiarowej krzywej Hilberta. Różnią się one także od krzywej zdefiniowanej przez układy równań (4.1)–(4.4). Jak widać, już w przypadku trójwymiarowym można otrzymać niewątpliwie różne krzywe, choć ich geometria jest bardzo podobna i we wszystkich stosowany jest ten sam kod Graya do uporządkowania wierzchołków kostki I_d.

Wielowymiarową krzywą Hilberta wypełniającą kostkę I_d możemy zdefiniować jako odwzorowanie $F_{Hd}(t) = (x_1(t), \ldots, x_d(t)) \;:\; I_1 \rightarrow I_d$ spełniające następujące warunki:

$$\left.\begin{array}{r} x_1(t) = x_d(2^d\, t)/2 \\ x_2(t) = x_1(2^d\, t)/2 \\ \cdots \\ x_d(t) = x_{d-1}(2^d\, t)/2 \end{array}\right\} \quad , \quad t \in [0, 2^{-d}], \tag{4.5}$$

$$\left.\begin{array}{r} x_1(t) = 1/2 + x_d(1/2^{d-1} - t) \\ x_2(t) = 1/2 - x_1(1/2^{d-1} - t) \\ x_3(t) = x_2(1/2^{d-1} - t) \\ \cdots \\ x_d(t) = x_{d-1}(1/2^{d-1} - t) \\ \cdots \end{array}\right\} \quad , \quad t \in [1/2^d, 1/2^{d-1}], \tag{4.6}$$

$$\left.\begin{array}{r} x_1(t) = x_1(1/2^{d-k} - t) \\ x_2(t) = x_2(1/2^{d-k} - t) \\ \cdots \\ x_{k-1}(t) = x_{k-1}(1/2^{d-k} - t) \\ x_k(t) = 1/2 + x_d(1/2^{d-k} - t) \\ x_{k+1}(t) = 1/2 - x_k(1/2^{d-k} - t) \\ x_{k+2}(t) = x_{k+1}(1/2^{d-k} - t) \\ \cdots \\ x_d(t) = x_{d-1}(1/2^{d-k} - t) \\ \cdots \end{array}\right\} \quad , \quad t \in [1/2^{d-k+1}, 1/2^{d-k}], \tag{4.7}$$

$$\left.\begin{array}{r} x_1(t) = x_1(1/2 - t) \\ x_2(t) = x_2(1/2 - t) \\ \cdots \\ x_{d-2}(t) = x_{d-2}(1/2 - t) \\ x_{d-1}(t) = 1/2 + x_d(1/2 - t) \\ x_d(t) = 1/2 - x_{d-1}(1/2 - t) \end{array}\right\} \quad , \quad t \in [1/4, 1/2], \tag{4.8}$$

$$\begin{array}{r} x_1(t) = x_1(1-t) \\ x_2(t) = x_2(1-t) \\ \cdots \\ x_{d-1}(t) = x_{d-1}(1-t) \\ x_d(t) = 1 - x_d(1-t) \end{array} \quad , \quad t \in [1/2, 1]. \tag{4.9}$$

Układ równań (4.5) opisuje rodzaj samopodobieństwa definiowanej krzywej. Ostatni z układów równań (4.9) opisuje symetrię krzywej ze względu na zmienną x_d. Pozostałe równania (4.6)–(4.8) dotyczą wewnętrznych symetrii krzywej F_{Hd}.

Z równania (4.5) wynika wprost, że $F_{H,d}(0) = (0, \ldots, 0)$. Przekształcimy układ równań (4.5)–(4.9) do postaci:

$$\begin{array}{r} x_1(t) = x_d(2^d t)/2 \\ x_2(t) = x_1(2^d t)/2 \\ \cdots \\ x_d(t) = x_{d-1}(2^d t)/2 \end{array} \quad , \quad t \in [0, 2^{-d}], \tag{4.10}$$

$$\begin{array}{r} x_1(t) = 1/2 + x_d(1/2^{d-1} - t) \\ x_2(t) = 1/2 - x_1(1/2^{d-1} - t) \\ x_3(t) = x_2(1/2^{d-1} - t) \\ \cdots \\ x_d(t) = x_{d-1}(1/2^{d-1} - t) \\ \cdots \end{array} \quad , \quad t \in (1/2^d, 1/2^{d-1}], \tag{4.11}$$

$$\begin{array}{r} x_1(t) = x_1(1/2^{d-k} - t) \\ x_2(t) = x_2(1/2^{d-k} - t) \\ \cdots \\ x_{k-1}(t) = x_{k-1}(1/2^{d-k} - t) \\ x_k(t) = 1/2 + x_d(1/2^{d-k} - t) \\ x_{k+1}(t) = 1/2 - x_k(1/2^{d-k} - t) \\ x_{k+2}(t) = x_{k+1}(1/2^{d-k} - t) \\ \cdots \\ x_d(t) = x_{d-1}(1/2^{d-k} - t) \\ \cdots \end{array} \quad , \quad t \in (1/2^{d-k+1}, 1/2^{d-k}], \tag{4.12}$$

$$\begin{array}{r} x_1(t) = x_1(1/2 - t) \\ x_2(t) = x_2(1/2 - t) \\ \cdots \\ x_{d-2}(t) = x_{d-2}(1/2 - t) \\ x_{d-1}(t) = 1/2 + x_d(1/2 - t) \\ x_d(t) = 1/2 - x_{d-1}(1/2 - t) \end{array} \quad , \quad t \in (1/4, 1/2], \tag{4.13}$$

$$\begin{array}{c} x_1(t) = x_1(1-t) \\ x_2(t) = x_2(1-t) \\ \cdots \\ x_{d-1}(t) = x_{d-1}(1-t) \\ x_d(t) = 1 - x_d(1-t) \end{array} \quad , \quad t \in (1/2, 1]. \tag{4.14}$$

Twierdzenie 4.1.1 *Układ równań (4.10)–(4.14), wraz z warunkiem* $F_{Hd}(0) = (0, \ldots, 0)$, *jest równoważny z układem równań (4.5)–(4.9). Rozwiązaniem układu równań (4.10)–(4.14) jest krzywa wypełniająca kostkę* I_d.

Dowód. Dowód powyższego twierdzenia można przeprowadzić analogicznie jak dowody lematów 2.1, 2.9, 2.12 dotyczących krzywych dwuwymiarowych. □

Użycie układu równań (4.10)–(4.14), wraz z warunkiem $F_{Hd}(0) = (0, \ldots, 0)$, pozwala na wyznaczenie dokładnych wartości odwzorowania we wszystkich punktach $t = i/2^{kd}$, $i = 0, 1, \ldots, 2^{kd}$, $k = 1, 2, \ldots$ W przypadku dowolnego, naturalnego, skończonego k, liczba rekurencji potrzebnych przy wyznaczaniu wartości $F_{Hd}(i/2^{kd})$ nie przekracza $(d+1)\,k$. Skorzystanie z własności poszczególnych równań prowadzi do wyznaczenia dokładnej wartości $F_{Hd}(i/2^{kd})$ przy nakładzie obliczeń rzędu $O(k\,d)$. Podobną złożoność obliczeniową ma wspomniany wcześniej algorytm Butza [25].

Dalsze własności krzywej Hilberta

Z własności definiujących wielowymiarową krzywą Hilberta (4.5)–(4.9) wynika bezpośrednio, że wartości funkcji $F_{Hd}(t)$ znajdują się w kostce I_d, a w szczególności: dla $0 \le t \le 1/2^d$ znajdują się w kostce $[0, 1/2] \times [0, 1/2] \times \ldots \times [0, 1/2]$; dla $1/2^d \le t \le 1/2^{d-1}$ znajdują się w kostce $[1/2, 1] \times [0, 1/2] \times \ldots \times [0, 1/2]$; itd. ..., a dla $(2^d - 1)/2^d \le t \le 1$ należą do kostki $[0, 1/2] \times \ldots \times [0, 1/2] \times [1/2, 1]$.

Zauważmy, że z (4.5)–(4.9) wynika, iż

$F_{Hd}(0) = (0, 0, \ldots, 0)$,

$F_{Hd}(1/2) = (0, 0, \ldots, 0, 1/2, 1/2)$,

$F_{Hd}(1/4) = (0, 0, \ldots, 0, 1/2, 1/2, 0)$,

itd.,

$F_{Hd}(1/2^d) = (1/2, 1/2, 0, 0, \ldots, 0)$,

$F_{Hd}(1) = (0, 0, \ldots, 0, 1)$.

Korzystając z równań (4.5)–(4.9), jesteśmy w stanie określić pozostałe współrzędne na odcinku, które są odwzorowywane w punkty wierzchołkowe punktów wierzchołkowych kostki I_d, a w konsekwencji ich kolejność na krzywej wypełniającej.

Przykładowo, dla wymiaru $d = 3$, wierzchołkowi $(1, 0, 0)$ odpowiada argument $t = t_0$, który musi się znajdować w przedziale $[1/8, 1/4]$, gdyż tylko wtedy $x_1(t) \geq 1/2$ oraz $x_2(t),\ x_3(t) \leq 1/2$. Stąd

$$\begin{gathered} x_1(t_0) = 1 = 1/2 + x_3(1/4 - t_0), \\ x_2(t_0) = 0 = 1/2 - x_1(1/4 - t_0), \\ x_3(t_0) = 0 = x_2(1/4 - t_0) \end{gathered}$$

i dalej stosując równanie (4.5):

$$\begin{gathered} x_3(2 - 8t_0)/2 = 1/2, \\ x_1(2 - 8t_0)/2 = 0, \\ x_2(2 - 8t_0)/2 = 1/2. \end{gathered}$$

W konsekwencji otrzymujemy:

$$\begin{gathered} x_1(2 - 8t_0) = 0, \\ x_2(2 - 8t_0) = 1, \\ x_3(2 - 8t_0) = 1 \end{gathered}$$

oraz z (4.9) otrzymujemy:

$$\begin{gathered} x_1(8t_0 - 1) = 0, \\ x_2(8t_0 - 1) = 1, \\ x_3(8t_0 - 1) = 0. \end{gathered}$$

Z powyższych równości wynika, że $8t_0 - 1 \in [1/4, 1/2]$, co w konsekwencji prowadzi do

$$\begin{gathered} x_1(8t_0 - 1) = 0 = x_1(3/2 - 8t_0), \\ x_2(8t_0 - 1) = 1 = 1/2 + x_3(3/2 - 8t_0), \\ x_3(8t_0 - 1) = 0 = 1/2 - x_2(3/2 - 8t_0) \end{gathered}$$

oraz (przy użyciu (4.5))

$$\begin{gathered} x_1(12 - 64t_0) = 1, \\ x_2(12 - 64t_0) = 1, \\ x_3(12 - 64t_0) = 0. \end{gathered}$$

Oznaczmy, dla skrócenia zapisu, $t_1 = 12 - 64t_0$. Korzystając kolejno z równań (4.8), (4.6) i (4.5), otrzymujemy:

$$\begin{gathered} x_1(8t_1 - 2) = 1, \\ x_2(8t_1 - 2) = 1, \\ x_3(8t_1 - 2) = 0, \end{gathered}$$

stąd wynika, że $t_1 = 8t_1 - 2$, czyli $t_1 = 2/7$. Automatycznie z równania (4.9) wynika, że punktowi $(1,1,1)$ musi odpowiadać argument $1-2/7 = 5/7$. Korzystając z oprzednich wyników, możemy obliczyć wartość t_0 jako $(12 - t_1)/64 = 41/234$ i wyznaczyć $F_{H3}(1-t_0) = F_{H3}(193/234) = (1,0,1)$, $F_{H3}(2-8t_0) = F_{H3}(15/28) = (0,1,1)$, $F_{H3}(8t_0 - 1) = F_{H3}(13/28) = (0,1,0)$. W ten sposób określiliśmy przyporządkowanie wierzchołkom kostki I_d odpowiednich punktów z odcinka I_1.

Należy zwrócić uwagę na fakt, iż punktom wierzchołkowym kostki I_d odpowiadają pojedyncze wartości argumentu $t \in I_1$, odwzorowanie $F_{Hd}(t)$ jest w tych punktach wzajemnie jednoznaczne.

Lemat 4.1 *Dla każdego* $t_1, t_2 \leq 1/2^d$ *oraz* $a = 1/2^d, \ldots, (2^d - 1)/2^d$ *zachodzi*

$$||F_{Hd}(t_1) - F_{Hd}(t_2)|| = ||F_{Hd}(t_1 + a) - F_{Hd}(t_2 + a)||. \tag{4.15}$$

Dowód. Dowód lematu wynika bezpośrednio z własności układu równań definiujących odwzorowanie, czyli (4.5)–(4.9). □

Twierdzenie 4.1.2 *Wielowymiarowa krzywa Hilberta spełnia warunek Höldera:*

$$||F_{Hd}(t_1) - F_{Hd}(t_2)|| \leq \alpha_d \, |t_2 - t_1|^{1/d}, \;\; t_1, \, t_2 \in I_1, \tag{4.16}$$

gdzie $\alpha_d = 2\,(d+3)^{1/2}$.

Dowód. Dla dowolnej pary $t_1, \, t_2 \in I_1$ istnieje takie całkowite $N \geq 0$, że

$$2^{-(N+1)d} \leq |t_1 - t_2| \leq 2^{-Nd}.$$

Zauważmy, że bez straty ogólności możemy przyjąć $t_1 \leq t_2$. Jeśli $N = 0$, to zachodzi $||F_{Hd}(t_1) - F_{Hd}(t_2)|| \leq \sqrt{d}$ oraz $|t_1 - t_2| \geq 2^{-d}$, co od razu prowadzi do spełnienia nierówności (4.16), gdyż

$$||F_{Hd}(t_1) - F_{Hd}(t_2)||/|t_1 - t_2|^{1/d} \leq 2\sqrt{d} < 2\sqrt{d+3}.$$

Jeżeli $N = 1$, to z lematu 4.1 oraz z równania (4.5) wynika, że

$$||F_{Hd}(t_1) - F_{Hd}(t_2)|| = ||F_{Hd}(\hat{t}_1) - F_{Hd}(\hat{t}_2)||,$$

gdzie $|t_1 - t_2| = |\hat{t}_1 - \hat{t}_2|$, przy czym $\hat{t}_1, \; \hat{t}_2 \in [0, 2^{-d}]$ bądź $\hat{t}_1 \in [0, 2^{-d}]$, a $\hat{t}_2 \in [2^{-d}, 2^{-d+1}]$. W konsekwencji $||F_{Hd}(t_1) - F_{Hd}(t_2)|| \leq 1/2\sqrt{d+3}$, gdyż w najgorszym przypadku $F_{Hd}(\hat{t}_1) \in [0,1/2] \times [0,1/2] \times \ldots \times [0,1/2]$, natomiast $F_{Hd}(\hat{t}_2) \in [1/2,1] \times [0,1/2] \times \ldots \times [0,1/2]$. Ponieważ $|t_1 - t_2| \geq 2^{-2d}$, czyli $|t_1 - t_2|^{1/d} \geq 2^{-2}$, otrzymujemy po podzieleniu $||F_{Hd}(t_1) - F_{Hd}(t_2)||/|t_1 - t_2|^{1/d}$ – dokładnie wartość stałej α_d. Jeśli $N > 1$, to postępujemy podobnie, korzystając uprzednio z własności (4.5). □

Łatwo zauważyć, że podana w twierdzeniu wartość α_d zależy od wymiaru przestrzeni i jest jedynie zgrubnym oszacowaniem najlepszej stałej. Na przykład dla $d = 2$ wartość $2(d+3)^{1/2}$ wynosi $20^{1/2}$. Jak wykazano w podrozdziale 2.2, optymalna wartość stałej wynosi w tym przypadku $6^{1/2}$.

Istnieje możliwość wyznaczenia optymalnej stałej także dla wymiaru większego od $d = 2$. Ze względu na złożoność, pokażemy metodę wyznaczania takiej stałej na przykładzie trójwymiarowej krzywej Hilberta zdefiniowanej przez (4.5)–(4.9).

W szczególnym przypadku, gdy $t_1 = 0$ lub $t_2 = 0$, wartość stałej w nierówności Höldera można jeszcze bardziej ograniczyć. Mówi o tym następujący lemat:

Lemat 4.2 *Trójwymiarowa krzywa Hilberta zdefiniowana przez (4.5)–(4.9) spełnia warunek*

$$||F_{H3}(t)||^3 \leq \hat{\alpha}_3\, t,\ 0 \leq t \leq 1, \tag{4.17}$$

gdzie $\hat{\alpha}_3 = \sqrt{33}\,231/131$ *jest najmniejszą wartością, dla której zachodzi nierówność (4.17).*

Dowód. Dowód ten ma podobną konstrukcję jak dowód analogicznego lematu (por. lemat 2.11 i twierdzenie 2.2.2) w przypadku dwuwymiarowym.

Niech zbiór

$$\begin{aligned} T^0 &= \{\, 0, 41/234, 2/7, 13/28, 15/28, 5/7, 193/234, 1\,\} \\ &= \{\, F_{H3}^{-1}(0,0,0), F_{H3}^{-1}(1,0,0), F_{H3}^{-1}(1,1,0), F_{H3}^{-1}(0,1,0), \\ &\quad F_{H3}^{-1}(0,1,1), F_{H3}^{-1}(1,1,1), F_{H3}^{-1}(1,0,1), F_{H3}^{-1}(0,0,1)\,\} \end{aligned}$$

będzie zbiorem punktów na odcinku I_1, którym odpowiadają w transformacji F_{H3} wierzchołki kostki jednostkowej I_d.

Wartości ze zbioru $T^k = \{(t_0 + i)/8^k, i = 0, 1, \ldots, 8^k - 1,\ t_0 \in T^0\}$ będą punktami węzłowymi w k-tej aproksymacji krzywej Hilberta. Zawsze spełniony jest warunek $T^k \subset T^{k+1}$. Ponadto łatwo pokazać, korzystając z (4.5)–(4.9), że punkty $F_{H3}(t)$, przy $t \in T^k$, mają współrzędne będące wielokrotnością 2^{-k}.

Wyznaczmy $\alpha_{0k} = \max_{t \in T^k} ||F_{H3}(t)||/t,\ t \neq 0$, dla kolejnych wartości $k = 0, 1, 2, \ldots$ Począwszy od $k = 2$ otrzymujemy wartość $\alpha_{02} = \alpha_{03} = \sqrt{33}\,231/131$. Maksimum osiągane jest w punkcie $(1, 1, 1/4)$ dla argumentu $t_g = 131/448 = 18/64 + (5/7)/64$. Proces indukcji niezupełnej należy w tym miejscu uzupełnić poprzez pokazanie, iż ze spełnienia (4.17) dla $t \in T^k$ wynika, że (4.17) zachodzi także dla $t \in T^{k+1}$, przy czym $k = 3, \ldots$

Niech $t \in T^{k+1}$ oraz niech $t \leq 1/2^d = 1/8$. Wtedy, korzystając z (4.5) oraz faktu, iż $8t \in T^k$, otrzymujemy $||F_{H3}(t)|| = ||F_{H3}(8t)||/2 \leq 0,5\,\hat{\alpha}_3^{1/3}(8t)^{1/3} = \hat{\alpha}_3^{1/3} t^{1/3}$, co kończy dowód dla $t \leq 1/8$.

Dla pozostałych wartości argumentu t, poza przedziałem $[18/64, 19/64]$, spełnienie nierówności (4.17) wynika z analizy najgorszego przypadku w odpowiednich podprzedziałach. I tak, dla $t \geq 5/8$ otrzymujemy $\max ||F_{H3}(t)|| = 3^{1/2}$, stąd $||F_{H3}(t)||^3/t \leq 8/5 \cdot \sqrt{27} < \hat{\alpha}_3$.

Podobne rozumowanie można przeprowadzić w pozostałych przypadkach, dotyczących nie sprawdzonych do tej pory wartości argumentu t, przy czym podziału odcinka jednostkowego wystarczy dokonać z dokładnością do, co najwyżej, $1/64$.

Przy $t \in [18/64, 19/64]$ mamy $\max ||F_{H3}(t)|| = ||(1, 1, 1/4)||$, stąd nierówność (4.17) jest spełniona dla $t \in [t_g, 19/64]$.

Pozostała część dowodu wynika z faktu, że w odpowiednich przedziałach $[t_g - 5/7 \cdot 2^{-3(s-1)}, t_g - 5/7 \cdot 2^{-3s}]$, $s = 3, 4, \ldots, k+2$, minimalna wartość $t \in T^{k+1}$ jest równa $t_{sm} = t_g - 5/7 \cdot 8^{-(s-1)}$. Ponadto przy $t \in [t_g - 5/7 \cdot 2^{-3(s-1)}, t_g - 5/7 \cdot 2^{-3s}]$ zachodzi

$$||F_{H3}(t))||^2 \leq \max[1 + 1/16 + (1 - 2^{-s})^2, 2 + (1/4 - 2^{-s})^2] = 2 + (1/4 - 2^{-s})^2,$$

stąd otrzymujemy dla $s \geq 4$ i przy założeniu, że $k \geq 2$, iż

$$\max ||F_{H3}(t)||^3/t \leq [2 + (1/4 - 2^{-s})^2]^{3/2}/t_{sm} < 7 \cdot 33^{3/2}/131 = \hat{\alpha}_3. \qquad (4.18)$$

Przedział $[t_g - 5/7 \cdot 2^{-3(k+2)}, t_g]$ nie zawiera żadnego punktu należącego do T^{k+1}, dlatego możemy go pominąć. W przypadku $s = 3$ mamy $t_g - 5/7 \cdot 8^{-2} = 18/64$. Przedział $[18/64, t_g - 5/7 \cdot 8^{-3}] = [144/512, 149/512]$ trzeba podzielić na dwa podprzedziały: $[18/64, 18/64 + 8^{-3}] = [144/512, 145/512]$ oraz $145/512, 149/512]$. W przedziale $[18/64, 145/512]$ zachodzi $\max ||F_{H3}(t)||^3 \cdot 64/18 \leq 64/18 \cdot (2(7/8)^2 + (1/8)^2)^{3/2} < \hat{\alpha}_3$. Jeśli $s = 3$, to $t_{sm} = 145/512$ i nadal $\max ||F_{H3}(t)||^3 \leq [2 + (1/4 - 2^{-s})^2]^{3/2}$. Gdy $t \in [145/512, 149/512]$, wtedy otrzymujemy

$$\max ||F_{H3}(t)||^3/t \leq \frac{(2 + (1/4 - 2^{-s})^2)^{3/2}}{t_{sm}} = \hat{\alpha}_3, \qquad (4.19)$$

co kończy dowód poprzez indukcję. Ponieważ T^k $(k = 2, 3, \ldots)$ jest zbiorem gęstym w $[0, 1]$, a $F_{H3}(t)$ jest odwzorowaniem ciągłym, własność (4.17) możemy więc przedłużyć na cały odcinek I_1. □

Twierdzenie 4.1.3 *Trójwymiarowa krzywa Hilberta, zdefiniowana przez układ równań (4.1)–(4.4), spełnia nierówność Höldera:*

$$||F_{H3}(t_1) - F_{H3}(t_2)|| \leq \alpha_3 \, |t_2 - t_1|^{1/3}, \;\; t_1, \, t_2 \in I_1, \qquad (4.20)$$

gdzie $\alpha_3^3 = \sqrt{2} \cdot 1512/79$.

Dowód. Dowód twierdzenia można przeprowadzić w analogiczny sposób jak dowód lematu 4.2. □

Schemat dowodowy, dotyczący wyznaczania najlepszej stałej w warunku Höldera, można zastosować także w przypadku wymiaru $d > 3$. Jednakże liczba szczególnych przypadków, którą należy uwzględnić, rośnie wykładniczo wraz ze wzrostem wymiaru krzywej.

W pracy Butza [24] podano inne oszacowanie stałej w warunku Höldera. Mimo że odnosi się ono do nieco innego wariantu wielowymiarowej krzywej Hilberta, to, ze względu na własności, z których korzystamy w dowodzie, może być zastosowane również w odniesieniu do pozostałych typów wielowymiarowej krzywej Hilberta. Jeżeli $d = 3$, to wartość stałej wyznaczona na podstawie twierdzenia 4.1.2 wynosi $117,57^{1/3}$, odpowiednia wartość wynikająca z oszacowania Butza jest równa $116,19^{1/3}$, natomiast optymalna wartość stałej wyznaczona na podstawie twierdzenia 4.1.3 wynosi nieco mniej niż $27,067^{1/3}$.

4.2 Wielowymiarowe krzywe Sierpińskiego

Opis dwuwymiarowej krzywej Sierpińskiego w postaci układu równań (3.2) jest ściśle związany z metodą generowania krzywych z użyciem iterowanych odwzorowań zwężających (IFS) (por. rozdział 3). Na podstawie każdej z par równań (3.2) można podać wprost parę odwzorowań: odcinka jednostkowego I_1 i jednostkowej kostki I_2, odpowiednio, w I_1 oraz I_2.

Uogólnianie dwuwymiarowej krzywej Sierpińskiego na przypadek wielowymiarowy wymaga odpowiedzi na szereg pytań wiążących się z wyborem postulowanych własności konstruowanej krzywej. Oczywiście, należy dalej zweryfikować, czy w ogóle istnieje wielowymiarowa krzywa wypełniająca, która spełnia postulowane warunki. W przypadku dwuwymiarowym, kwadrat I_2 podzielony jest na dwa elementarne trójkąty i krzywa wypełniająca I_2 faktycznie kolejno wypełnia najpierw jeden, a potem, symetrycznie, drugi trójkąt. Na pytanie, co ma być odpowiednikiem trójkąta w przypadku kostki I_d, trudno jest odpowiedzieć wprost. Jak zobaczymy dalej, sposób podziału kostki I_d na dwie części, które można pokryć samopodobnymi kopiami, zostanie zdefiniowany pośrednio, jako efekt działania odwzorowania będącego odpowiednią krzywą wypełniającą.

Opis złożony z $2^d + 1$ transformacji odwzorowujących całą kostkę I_d na odpowiednią podkostkę lub jej część (trójkątną w przypadku dwuwymiarowym) można zastąpić $d+2$ układami równań funkcyjnych, w których korzystamy z wybranych własności geometrycznych konstruowanej krzywej. Porządek, w którym krzywa wypełniająca przechodzi przez poszczególne podkostki o boku 1/2, zostanie dalej zdefiniowany z użyciem odpowiednich kodów Graya.

Wielowymiarową krzywą Sierpińskiego, dla wymiaru d będącego liczbą parzystą, możemy zdefiniować jako odwzorowanie $F_{Sd}(t) = (x_1(t), \ldots, x_d(t)) \; : \; I_1 \rightarrow I_d$ spełniające następujące warunki:

$$\begin{gathered} x_1(t) = 1/2 - x_1(2^d\,[t + c_d/2^{2d}])/2, \\ x_2(t) = 1/2 - x_2(2^d\,[t + c_d/2^{2d}])/2, \\ \cdots \\ x_d(t) = 1/2 - x_d(2^d\,[t + c_d/2^{2d}])/2, \\ t \in [0, b_d/2^{2d}], \end{gathered} \tag{4.21}$$

$$\begin{gathered} x_1(t) = 1/2 - x_1(2^d\,[t + c_d/2^{2d} - 1])/2, \\ x_2(t) = 1/2 - x_2(2^d\,[t + c_d/2^{2d} - 1])/2, \\ \cdots \\ x_d(t) = 1/2 - x_d(2^d\,[t + c_d/2^{2d} - 1])/2, \\ t \in [1 - c_d/2^{2d}, 1], \end{gathered} \tag{4.22}$$

$$\begin{gathered} x_1(t) = 1/2 + x_1(1 - 2^d[t - b_d/2^{2d}])/2, \\ x_2(t) = 1/2 - x_2(1 - 2^d[t - b_d/2^{2d}])/2, \\ \cdots \\ x_d(t) = 1/2 - x_d(1 - 2^d[t - b_d/2^{2d}])/2, \\ t \in [b_d/2^{2d}, t_1], \;\; t_1 = b_d\,2^{-2d} + 2^{-d}, \end{gathered} \tag{4.23}$$

$$\begin{gathered} x_1(t) = x_1(2t_1 - t), \\ x_2(t) = 1 - x_2(2t_1 - t), \\ x_3(t) = x_3(2t_1 - t), \\ \cdots \\ x_d(t) = x_d(2t_1 - t), \\ t \in [t_1, t_2], \;\; t_2 = t_1 + 2^{-d+1}, \end{gathered} \tag{4.24}$$

$$\cdots$$

$$\begin{gathered} x_1(t) = x_1(2t_{d-1} - t), \\ x_2(t) = x_2(2t_{d-1} - t), \\ \cdots \\ x_d(t) = 1 - x_d(2t_{d-1} - t), \\ t \in [t_{d-1}, t_d], \;\; t_d = t_{d-1} + 1/2 = 1 - c_d\,2^{-2d}. \end{gathered} \tag{4.25}$$

Jeśli w którymś z powyższych wyrażeń wartość argumentu jest mniejsza od zera, to należy przyjąć $x_i(t) = x_i(t+1)$ dla $t < 0$, $i = 1, 2, \ldots, d$, gdy natomiast $t > 1$, przyjmujemy $x_i(t) = x_i(t-1)$, $i = 1, 2, \ldots, d$. W powyższych układach równań c_d jest stałą zależną od wymiaru. Stała ta określa względne położenie $(1, 1, 1, \ldots, 1)$ w ciągu binarnych, d–elementowych kodów Graya [59], natomiast

$b_d/2^d = 1 - c_d/2^d$. Łatwo sprawdzić, że dla $d = 2$ zachodzi $c_d/2^d = b_d/2^d = 1/2$ oraz w ogólnym przypadku $c_{2k} = 2/3 \cdot (1 - 2^{-2k})\, 2^{2k}$, $c_{2k-1} = c_{2k}/2$, $k = 1, 2, \ldots$

Z układu równań (4.21)–(4.25) wynika, że $F_{Sd}(c_d/2^d) = (1, \ldots, 1)$, a w konsekwencji $F_{Sd}(0) = (0, \ldots, 0)$ oraz $F_{Sd}(1) = (0, \ldots, 0)$.

Przekształcimy układ równań (4.21)–(4.25) do postaci:

$$\begin{gathered} x_1(t) = 1/2 - x_1(2^d\,[t + c_d/2^{2d}])/2, \\ x_2(t) = 1/2 - x_2(2^d\,[t + c_d/2^{2d}])/2, \\ \cdots \\ x_d(t) = 1/2 - x_d(2^d\,[t + c_d/2^{2d}])/2, \\ t \in [0, b_d/2^{2d}], \end{gathered} \tag{4.26}$$

$$\begin{gathered} x_1(t) = 1/2 - x_1(2^d\,[t + c_d/2^{2d} - 1])/2, \\ x_2(t) = 1/2 - x_2(2^d\,[t + c_d/2^{2d} - 1])/2, \\ \cdots \\ x_d(t) = 1/2 - x_d(2^d\,[t + c_d/2^{2d} - 1])/2, \\ t \in [1 - c_d/2^{2d}, 1], \end{gathered} \tag{4.27}$$

$$\begin{gathered} x_1(t) = 1/2 + x_1(1 - 2^d[t - b_d/2^{2d}])/2, \\ x_2(t) = 1/2 - x_2(1 - 2^d[t - b_d/2^{2d}])/2, \\ \cdots \\ x_d(t) = 1/2 - x_d(1 - 2^d[t - b_d/2^{2d}])/2, \\ t \in (b_d/2^{2d}, t_1], \;\; t_1 = b_d\, 2^{-2d} + 2^{-d}, \end{gathered} \tag{4.28}$$

$$\begin{gathered} x_1(t) = x_1(2t_1 - t), \\ x_2(t) = 1 - x_2(2t_1 - t), \\ x_3(t) = x_3(2t_1 - t), \\ \cdots \\ x_d(t) = x_d(2t_1 - t), \\ t \in (t_1, t_2], \;\; t_2 = t_1 + 1/2^{d-1}, \\ \cdots \end{gathered} \tag{4.29}$$

$$\begin{gathered} x_1(t) = x_1(2t_{d-1} - t), \\ x_2(t) = x_2(2t_{d-1} - t), \\ \cdots \\ x_d(t) = 1 - x_d(2t_{d-1} - t), \\ t \in (t_{d-1}, t_d), \;\; t_d = t_{d-1} + 1/2 = 1 - c_d\, 2^{-2d}. \end{gathered} \tag{4.30}$$

Jeśli d jest liczbą nieparzystą, to układ równań (4.21)–(4.25) jest sprzeczny. Jak łatwo sprawdzić, nie ma na przykład jednoznacznego rozwiązania układu równań (4.21)–(4.25) dla $t = c_d/2^d$. Oznacza to, że gdy wymiar d jest liczbą nieparzystą, nie można w postaci układu równań (4.21)–(4.25) zdefiniować krzywej

wypełniającej, a jedynie odwzorowanie nieciągłe w przeliczalnej liczbie punktów z odcinka I_1.

Podstawienie równań (4.21), (4.22) i (4.23) do pozostałych $d-1$ równań (4.24)–(4.25) prowadzi do otrzymania 2^d-2 nowych układów równań postaci:

$$\begin{gathered} x_1(t) = 1/2 - (1/2 - \beta_{1l}) \cdot x_1(2^d(t - m_l)), \\ x_2(t) = 1/2 - (1/2 - \beta_{2l}) \cdot x_2(2^d(t - m_l)), \\ \ldots \\ x_d(t) = 1/2 - (1/2 - \beta_{dl}) \cdot x_d(2^d(t - m_l)), \\ m_l = b_d\, 2^{-2d} + (l-1)\, 2^{-d},\ t \in [m_l, m_{l+1}], \\ l = 2, 4, \ldots, 2^d - 2 \end{gathered} \tag{4.31}$$

oraz

$$\begin{gathered} x_1(t) = 1/2 - (1/2 - \beta_{1l}) \cdot x_1(2^d(m_l - t)), \\ x_2(t) = 1/2 - (1/2 - \beta_{2l}) \cdot x_2(2^d(m_l - t)), \\ \ldots \\ x_d(t) = 1/2 - (1/2 - \beta_{dl}) \cdot x_d(2^d(m_l - t)), \\ m_l = b_d\, 2^{-2d} + l\, 2^{-d}, t \in [m_{l-1}, m_l], \\ l = 3, 5, \ldots, 2^d - 1, \end{gathered} \tag{4.32}$$

gdzie $\beta_{il} = 0$ lub 1 oraz $(\beta_{1l}, \beta_{2l}, \ldots, \beta_{dl}) = g_{d,l}$ jest l–tym elementem ciągu kodów Graya, odpowiadającym l–temu wierzchołkowi kostki I_d (por. rozdział 3.2.2).

Przypomnijmy, że z (4.21)–(4.25) wynika: $F_{Sd}(c_d/2^d) = (1, 1, \ldots, 1)$ oraz $F_{Sd}(0) = F_{Sd}(1) = (0, 0, \ldots, 0)$. Korzystając z własności (4.21)–(4.25), jesteśmy w stanie określić pozostałe współrzędne na odcinku, które są odwzorowywane w punkty wierzchołkowe punktów wierzchołkowych kostki I_d, a w konsekwencji kolejność wierzchołków na krzywej wypełniającej. W przypadku dowolnego, parzystego d, argument $t = t_0$, który odpowiada wierzchołkowi $(1, 0, \ldots, 0, 0)$, musi znajdować się w przedziale $[b_d\, 2^{-2d}, 2^{-d} + b_d\, 2^{-2d})]$. Tylko wtedy bowiem zachodzi $x_1(t_0) \geq 1/2$ oraz $x_2(t_0), \ldots, x_d(t_0) \leq 1/2$. Stąd, korzystając z (4.23) lub (4.28) otrzymujemy

$$\begin{gathered} x_1(t_0) = 1 = 1/2 + x_1(1 + b_d/2^d - 2^d t_0)/2, \\ x_2(t_0) = 0 = 1/2 - x_2(1 + b_d/2^d - 2^d t_0)/2, \\ x_3(t_0) = 0 = 1/2 - x_3(1 + b_d/2^d - 2^d t_0)/2, \\ \ldots \\ x_d(t_0) = 0 = 1/2 - x_d(1 + b_d/2^d - 2^d t_0)/2. \end{gathered}$$

Z powyższego wynika, że $x_i(1 + b_d/2^d - 2^d t_0) = 1$ dla $i = 1, 2, \ldots, d$. W konsekwencji mamy $1 + b_d/2^d - 2^d t_0 = c_d/2^d$ i $t_0 = 2\, b_d\, 2^{-2d}$.

W analogiczny sposób możemy określić współrzędne na odcinku I_1, odpowiadające dalszym wierzchołkom kostki I_d. W ten sposób otrzymamy

$$g_{d,i} = F_{Sd}(i\,2^{-d}) \qquad \text{dla} \quad i = 0, 2, 4, \ldots, 2^d - 2,$$

$$g_{d,i} = F_{Sd}((i-1)\,2^{-d} + b_d\,2^{-2d+1}) \quad \text{dla} \quad i = 1, 3, \ldots, 2^d - 1.$$

Korzystając ze znanego schematu geometrycznej konstrukcji krzywej wypełniającej, możemy teraz wykazać, że odwzorowanie F_{Sd} jest d–wymiarową krzywą wypełniającą.

Twierdzenie 4.2.1 *Załóżmy, że d jest liczbą parzystą. Układ równań funkcyjnych (4.26)–(4.30), uzupełniony o warunek $F_{Sd}(c_d/2^d) = (1, 1, \ldots, 1)$, jest równoważny z równaniami (4.21)–(4.25). Rozwiązaniem układu równań (4.26)–(4.30), wraz z warunkiem $F_{Sd}(c_d/2^d) = (1, 1, \ldots, 1)$ jest krzywa wypełniająca kostkę I_d.*

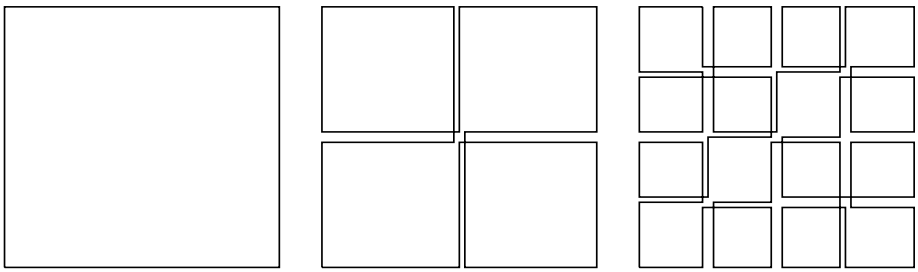
Dowód. Dowód powyższego twierdzenia można przeprowadzić analogicznie jak dowody lematów 2.1, 2.9, 2.12 dotyczących krzywych dwuwymiarowych. □

Użycie równań (4.26)–(4.30), wraz z warunkiem $F_{Sd}(c_d/2^d) = (1, 1, \ldots, 1)$, pozwala na wyznaczenie dokładnych wartości odwzorowania F_{Sd} we wszystkich punktach $t = i/2^{kd}$, $i = 0, 1, \ldots, 2^{kd}$, $k = 1, 2, \ldots$ W przypadku dowolnego, naturalnego, skończonego k, liczba rekurencji potrzebnych przy wyznaczaniu odpowiednich wartości F_{Sd} nie przekracza $(d+1)(k+1)$. Ponadto zbiór wartości $F_{Sd}(i/2^{kd})$, $i = 0, 1, \ldots, 2^{kd}$, $k = 1, 2, \ldots$ jest zbiorem wszystkich punktów z I_d, których współrzędne mają skończone dwójkowe rozwinięcia. Podobnie jak w przypadku uprzednio analizowanej wielowymiarowej krzywej Hilberta, skorzystanie z własności poszczególnych równań umożliwia wyznaczenie dokładnej wartości $F_{Sd}(i/2^{kd})$ przy nakładzie obliczeń rzędu $O(k\,d)$ [161].

Zaproponujemy teraz inną wersję wielowymiarowej krzywej Sierpińskiego. Dla odróżnienia od poprzedniej krzywej, będziemy ją dalej oznaczać przez F_{Md}. W celu uproszczenia zapisu, oznaczenia poszczególnych współrzędnych $(x_1, \ldots, x_d)$ pozostaną bez zmiany (bez dodatkowego indeksu wskazującego, którego odwzorowania dotyczą).

Krzywa F_{Md} jest rozwiązaniem zmodyfikowanego układu równań funkcyjnych (4.21)–(4.25). Modyfikacja ta ma na celu zachowanie ciągłości transformacji odcinka jednostkowego na kostkę wielowymiarową I_d, niezależnie od tego, czy d jest liczbą parzystą, czy też nie. Zastąpienie równań (4.23) przez

$$\begin{aligned} x_1(t) &= 1/2 + x_1\,(2^d(t - b_d/2^{2d}))/2, \\ x_2(t) &= 1/2 - x_2\,(2^d(t - b_d/2^{2d}))/2, \\ &\cdots \\ x_d(t) &= 1/2 - x_d\,(2^d(t - b_d/2^{2d}))/2, \\ b_d\,2^{-2d} \le t \le t_1 &= b_d\,2^{-2d} + 2^{-d} \end{aligned} \tag{4.33}$$



Rys. 4.2. Kolejne aproksymacje zmodyfikowanej krzywej Sierpińskiego w 2-D
Fig. 4.2. Approximations of the modified Sierpiński space-filling curve in 2-D

prowadzi do powstania nowej krzywej wypełniającej. W tym przypadku układ równań (4.21), (4.22), (4.33), (4.24)–(4.25) ma rozwiązanie dla dowolnego wymiaru przestrzeni $d \geq 2$. Jednakże, gdy $d = 2$, otrzymujemy krzywą wypełniającą, która nie jest klasyczną krzywą Sierpińskiego (por. rysunek (4.2)), a jedynie krzywą podobną do niej.

Podstawienie równań (4.21), (4.22) oraz (4.33) do pozostałego układu równań (4.24)–(4.25) prowadzi do otrzymania $2^d - 2$ przekształceń postaci:

$$\begin{gathered} x_1(t) = 1/2 - (1/2 - \beta_{1l})\, x_1(2^d(t - m_l)), \\ x_2(t) = 1/2 - (1/2 - \beta_{2l})\, x_2(2^d(t - m_l)), \\ \cdots \\ x_d(t) = 1/2 - (1/2 - \beta_{dl})\, x_d(2^d(t - m_l)), \\ m_l = b_d\, 2^{-2d} + (l-1)\, 2^{-d},\ t \in [m_l, m_{l+1}], \\ l = 2, \ldots, 2^d - 1, \end{gathered} \tag{4.34}$$

gdzie $\beta_{il} = 0$ lub 1 oraz $(\beta_{1l}, \beta_{2l}, \ldots, \beta_{dl}) = g_{d,l}$, jest l–tym elementem ciągu kodów Graya odpowiadającym l–temu wierzchołkowi kostki I_d.

Układy równań (4.34) wraz równaniami (4.21), (4.22) i (4.33) tworzą równoważny opis krzywej $F_{Md}(t) = (x_1(t), x_2(t), \ldots, x_d(t)$, dla którego można wskazać odpowiedni, równoważny zestaw IFS (por. (3.4) z rozdziału 3), którego atraktorem jest zdefiniowana w podrozdziale 3.2.2 krzywa F_d. W pracy autorki [161] zamieszczono implementację algorytmu generowania wartości F_{Md} o złożoności obliczeniowej $O(d)$.

4.3 Wielowymiarowa krzywa Peano

G.Peano nie poprzestał na podaniu opisu krzywej dwuwymiarowej wypełniającej kwadrat jednostkowy, ale sformułował również sposób konstrukcji krzywej trój-

wymiarowej wypełniającej kostkę 3-D [121], [137]. W odwzorowaniu podanym przez Peano zastosowano trójkowe rozwinięcia liczb, trudne do bezpośredniego uogólnienia. Geometryczną konstrukcję dwuwymiarowej krzywej Peano sformułował w 1900 roku Moore [111]. Konstrukcję tę w roku 1980 uogólnił na przypadek wielowymiarowy Milne [110]. W pracy tej wskazano optymalną wartość wykładnika w odpowiednim warunku Höldera charakteryzującym daną krzywą wypełniającą. Krzywa zdefiniowana przez Milne'a jest jedną z pierwszych (poza pracami Butza [23], [24]) w pełni formalnych konstrukcji wielowymiarowej krzywej wypełniającej, która spełnia warunek Höldera z wykładnikiem $1/d$, gdzie d jak zwykle oznacza wymiar wypełnianej kostki. Konstrukcja Milne'a jest trudna do bezpośredniego przekształcenia w efektywny algorytm obliczeniowy. Dalej zaproponujemy opis definiujący za pomocą układu równań funkcyjnych, tę samą (z dokładnością do kolejności zmiennych) wielowymiarową krzywą Peano, której geometryczną konstrukcję podał Milne [110]. Definicja wielowymiarowej krzywej Peano, zaproponowana przez autorkę, jest kontynuacją schematu przedstawionego w rozdziale poświęconym krzywym dwuwymiarowym.

Niech $F_P^d(t) = (x_1(t), x_2(t), \ldots, x_d(t))$, $t \in I_1$ oznacza ciągłe odwzorowanie przeprowadzające punkty z odcinka I_1 w punkty d–wymiarowej kostki I_d.

Do opisania odwzorowania $F_P^d(t)$ $t \in I_1$ wystarczy $d+1$ wielowymiarowych równań funkcyjnych. W przypadku trójwymiarowym ($d = 3$) mają one postać:

$$\left.\begin{array}{l} x_1(t) = x_1(27t)/3 \\ x_2(t) = x_2(27t)/3 \\ x_3(t) = x_3(27t)/3 \end{array}\right\} \quad t \in [0, 1/27], \tag{4.35}$$

$$\left.\begin{array}{l} x_1(t) = 1/3 + x_1(t - 1/27) \\ x_2(t) = 1/3 - x_2(t - 1/27) \\ x_3(t) = 1/3 - x_3(t - 1/27) \end{array}\right\} \quad t \in [1/27, 1/9], \tag{4.36}$$

$$\left.\begin{array}{l} x_1(t) = 1 - x_1(t - 1/9) \\ x_2(t) = 1/3 + x_2(t - 1/9) \\ x_3(t) = 1/3 - x_3(t - 1/9) \end{array}\right\} \quad t \in [1/9, 1/3], \tag{4.37}$$

$$\left.\begin{array}{l} x_1(t) = 1 - x_1(t - 1/3) \\ x_2(t) = 1 - x_2(t - 1/3) \\ x_3(t) = 1/3 + x_3(t - 1/3) \end{array}\right\} \quad t \in [1/3, 1]. \tag{4.38}$$

W ogólnym przypadku, zestaw równań definiujących d–wymiarową krzywą Peano można otrzymać poprzez odpowiednie rozszerzenie układu równań definiującego krzywą $(d-1)$–wymiarową. Przedstawimy tu jedynie nieformalny opis tego typu przekształcenia. We wszystkich równaniach odpowiednie wartości współczynników równe $3^{-k}, k = 1, 2, \ldots, d-1$ zastępowane są przez wartości

podzielone przez 3, czyli przez 3^{-k-1}. Równania dotyczące pierwszych $d-1$ zmiennych pozostają bez zmian. Dodatkowe równania, dotyczące zmiennej $x_d(t)$, mają w każdym przypadku postać $x_d(t) = 1/3 - x_d(t - 3^{-d+i-1})$ $t \in [3^{-d+i-1}, 3^{-d+i}]$, gdzie $i = 1, 2, \ldots d-1$. W ostatnim $(d+1)$ wielowymiarowym równaniu, odnoszącym się do przedziału $t \in [1/3, 1]$, mamy $x_j(t) = 1-x_j(t-1/3)$, $j = 1, 2, \ldots, d-1$ oraz $x_d(t) = 1/3 + x_d(t - 1/3)$. W ten sposób otrzymujemy następujący układ równań funkcyjnych, którego rozwiązaniem jest wielowymiarowa (d–wymiarowa) krzywa Peano:

$$\begin{array}{l} x_1(t) = x_1(3^d\, t)/3 \\ x_2(t) = x_2(3^d\, t)/3 \\ \qquad \cdots \\ x_d(t) = x_d(3^d\, t)/3 \end{array} \quad , \quad t \in [0, 3^{-d}], \tag{4.39}$$

$$\begin{array}{l} x_1(t) = 1/3 + x_1(t - 1/3^d) \\ x_2(t) = 1/3 - x_2(t - 1/3^d) \\ \qquad \cdots \\ x_d(t) = 1/3 - x_d(t - 1/3^d) \end{array} \quad , \quad t \in [1/3^d, 1/3^{d-1}], \tag{4.40}$$

$$\begin{array}{l} x_1(t) = 1 - x_1(t - 1/3^{d-1}) \\ x_2(t) = 1/3 + x_2(t - 1/3^{d-1}) \\ x_3(t) = 1/3 - x_3(t - 1/3^{d-1}) \\ \qquad \cdots \\ x_d(t) = 1/3 - x_d(t - 1/3^{d-1}) \end{array} \quad , \quad t \in [1/3^{d-1}, 1/3^{d-2}], \tag{4.41}$$

$$\cdots$$

$$\begin{array}{l} x_1(t) = 1 - x_1(t - 1/9) \\ x_2(t) = 1 - x_2(t - 1/9) \\ \qquad \cdots \\ x_{d-1}(t) = 1/3 + x_{d-1}(t - 1/9) \\ x_d(t) = 1/3 - x_d(t - 1/9) \end{array} \quad , \quad t \in [1/9, 1/3], \tag{4.42}$$

$$\begin{array}{l} x_1(t) = 1 - x_1(t - 1/3) \\ x_2(t) = 1 - x_2(t - 1/3) \\ \qquad \cdots \\ x_{d-1}(t) = 1 - x_{d-1}(t - 1/3) \\ x_d(t) = 1/3 + x_d(t - 1/3) \end{array} \quad , \quad t \in [1/3, 1]. \tag{4.43}$$

Z równania (4.39) wynika wprost, że $F_{P,d}(0) = (0, \ldots, 0)$. Przekształcimy układ równań (4.39)–(4.43) do postaci:

$$\left.\begin{array}{l} x_1(t) = x_1(3^d t)/3 \\ x_2(t) = x_2(3^d t)/3 \\ \cdots \\ x_d(t) = x_d(3^d t)/3 \end{array}\right., \quad t \in [0, 3^{-d}), \tag{4.44}$$

$$\left.\begin{array}{l} x_1(t) = 1/3 + x_1(t - 1/3^d) \\ x_2(t) = 1/3 - x_2(t - 1/3^d) \\ \cdots \\ x_d(t) = 1/3 - x_d(t - 1/3^d) \end{array}\right., \quad t \in [1/3^d, 1/3^{d-1}), \tag{4.45}$$

$$\left.\begin{array}{l} x_1(t) = 1 - x_1(t - 1/3^{d-1}) \\ x_2(t) = 1/3 + x_2(t - 1/3^{d-1}) \\ x_3(t) = 1/3 - x_3(t - 1/3^{d-1}) \\ \cdots \\ x_d(t) = 1/3 - x_d(t - 1/3^{d-1}) \end{array}\right., \quad t \in [1/3^{d-1}, 1/3^{d-2}), \tag{4.46}$$

$$\cdots$$

$$\left.\begin{array}{l} x_1(t) = 1 - x_1(t - 1/9) \\ x_2(t) = 1 - x_2(t - 1/9) \\ \cdots \\ x_{d-1}(t) = 1/3 + x_{d-1}(t - 1/9) \\ x_d(t) = 1/3 - x_d(t - 1/9) \end{array}\right., \quad t \in [1/9, 1/3), \tag{4.47}$$

$$\left.\begin{array}{l} x_1(t) = 1 - x_1(t - 1/3) \\ x_2(t) = 1 - x_2(t - 1/3) \\ \cdots \\ x_{d-1}(t) = 1 - x_{d-1}(t - 1/3) \\ x_d(t) = 1/3 + x_d(t - 1/3) \end{array}\right., \quad t \in [1/3, 1]. \tag{4.48}$$

Twierdzenie 4.3.1 *Układ równań funkcyjnych (4.44)–(4.48), wraz z warunkiem $F_{Pd}(0) = (0, \ldots, 0)$, jest równoważny z układem równań (4.39)–(4.43). Rozwiązaniem układu równań (4.44)–(4.48), wraz z warunkiem $F_{Pd}(0) = (0, \ldots, 0)$, jest krzywa wypełniająca kostkę I_d.*

Dowód. Dowód powyższego twierdzenia można przeprowadzić analogicznie jak dowody lematów 2.1, 2.9, 2.12 dotyczących krzywych dwuwymiarowych. □

Użycie układu równań (4.44)–(4.48), wraz z warunkiem $F_{Pd}(0) = (0, \ldots, 0)$, pozwala na wyznaczenie dokładnych wartości odwzorowania F_{Pd} we wszystkich punktach $t = i/3^{kd}$, $i = 0, 1, \ldots, 3^{kd}$, $k = 1, 2, \ldots$ W przypadku dowolnego, naturalnego, skończonego k, liczba rekurencji potrzebna przy wyznaczaniu wartości

$F_{Pd}(i/3^{kd})$ nie przekracza $(d+1)k$. Podobnie jak w przypadku uprzednio omawianych krzywych Hilberta i Sierpińskiego, skorzystanie z własności poszczególnych równań umożliwia wyznaczenie dokładnej wartości $F_{Pd}(i/3^{kd})$ przy nakładzie obliczeń rzędu $O(k\,d)$. Implementację algorytmu wyznaczania wartości F_{Pd} podano w pracy autorki [161].

Dalsze własności krzywej Peano

Z układu równań (4.39)–(4.43) wynika bezpośrednio, że wartości funkcji $F_{Pd}(t)$ znajdują się w kostce I_d, a w szczególności: dla $t \le 1/3^d$ znajdują się w kostce $[0,1/3] \times [0,1/3] \times \ldots \times [0,1/3]$; dla $1/3^d \le t \le 2/3^d$ znajdują się w kostce $[1/3,2/3] \times [0,1/3] \times \ldots \times [0,1/3]$; itd. ..., a dla $(3^d-1)/3^d \le t \le 1$ znajdują się w kostce $[2/3,1] \times \ldots \times [2/3,1]$.

W szczególności, z równań (4.39)–(4.43) wynika, że

$$\begin{array}{lcr} x_1(t), x_2(t), \ldots, x_d(t) \le 1/3 & \Leftrightarrow & t \in [0, 1/3^d] \\ x_1(t) \ge 1/3,\ x_2(t), \ldots, x_d(t) \le 1/3 & \Leftrightarrow & t \in [1/3^d, 1/3^{d-1}] \\ x_2(t) \ge 1/3,\ \ x_3(t), \ldots, x_d(t) \le 1/3 & \Leftrightarrow & t \in [1/3^{d-1}, 1/3^{d-2}] \\ & \ldots & \\ x_{d-1}(t) \ge 1/3,\ x_d(t) \le 1/3 & \Leftrightarrow & t \in [1/9, 1/3] \\ x_d(t) \ge 1/3 & \Leftrightarrow & t \in [1/3, 1] \end{array} \tag{4.49}$$

Korzystając z powyższych własności, jesteśmy w stanie wyznaczyć wartości współrzędnych na odcinku, które są odwzorowywane w punkty wierzchołkowe odpowiadające wszystkim punktom wierzchołkowym kostki I_d i w związku z tym ustalić ich porządek na krzywej wypełniającej. Jest to porządek typu leksykograficznego, co oznacza, że wierzchołek $(\ldots, 0, a, b, c, \ldots, x)$ zawsze poprzedza wierzchołek $(\ldots, 1, a, b, c, \ldots, x)$.

W szczególności, wierzchołek $(1,0,0,\ldots,0)$ odpowiada na odcinku punktowi $t_0 = 2/(3^d-1)$, czyli $F_{Pd}(t_0) = (1,0,0,\ldots,0)$. Z warunków (4.49) wynika, że $t_0 \le 1/3^{d-1}$. W konsekwencji otrzymujemy:

$$\begin{array}{c} x_1(t_0 - 2/3^d) = 1/3, \\ x_2(t_0 - 2/3^d) = 0, \\ \ldots \\ x_d(t_0 - 2/3^d) = 0 \end{array}$$

i dalej

$$\begin{array}{c} x_1(3^d(t_0 - 2/3^d)) = 1, \\ x_2(3^d(t_0 - 2/3^d)) = 0, \\ \ldots \\ x_d(3^d(t_0 - 2/3^d)) = 0. \end{array}$$

Stąd wynika, że $t_0 = 3^d(t_0 - 2/3^d)$, co w konsekwencji daje $t_0 = 2/(3^d - 1)$. Należy zwrócić uwagę na fakt, iż w punktach wierzchołkowych kostki I_d odwzorowanie $F_{Pd}(t)$ jest wzajemnie jednoznaczne.

W ogólnym przypadku można podać następujący iteracyjny algorytm wyznaczania argumentu $t \in I_1$ odpowiadającego wierzchołkom kostki I_d w odwzorowaniu F_{Pd}.

Lemat 4.3 *Niech t_{d-1} będzie argumentem odpowiadającym pewnemu wierzchołkowi $(e_1, e_2, \ldots, e_{d-1})$, $e_i \in \{0, 1\}$ kostki I_{d-1} w odwzorowaniu F_{Pd-1}, wtedy*

$$\begin{aligned} F_{Pd}^{-1}(e_1, e_2, \ldots, e_{d-1}, 0) &= t_d \tfrac{3^{d-1}-1}{3^d-1}, \\ F_{Pd}^{-1}(e_1, e_2, \ldots, e_{d-1}, 1) &= 1 - (1 - t_d)\tfrac{3^{d-1}-1}{3^d-1}. \end{aligned} \tag{4.50}$$

Dowód. Dowód powyższego lematu łatwo przeprowadzić poprzez indukcję. □

Bezpośrednio z układu równań (4.39)–(4.43) wynika następująca własność krzywej Peano:

Lemat 4.4 *Dla każdego $t_1, t_2 \leq 1/3^d$ oraz $a = 1/3^d, \ldots, (3^d - 1)/3^d$ zachodzi*

$$||F_{Pd}(t_1) - F_{Pd}(t_2)|| = ||F_{Pd}(t_1 + a) - F_{Pd}(t_2 + a)||. \tag{4.51}$$

Pokażemy dalej, że krzywa Peano jest symetryczna, a mianowicie:

Lemat 4.5 *Dla każdego $t \in I_1$ zachodzi*

$$x_i(t) + x_i(1 - t) = 1, \quad i = 1, 2, \ldots, d. \tag{4.52}$$

Dowód. Dowód jest analogiczny jak w przypadku dwuwymiarowym (por. dowód lematu 2.16). □

Z lematu 4.5 wynika następna własność F_{Pd}, sformułowana w kolejnym lemacie.

Lemat 4.6 *Dla każdego $t \in I_1$ oraz dla każdego $s = 1/3^d, 2/3^d, \ldots 1$ takiego, że $s, t \in [(l-1)/3^d, l/3^d]$, $l = 1, 2, \ldots, 3^{d-1}$ spełniona jest równość:*

$$|x_i(s) - x_i(t)| = x_i(|s - t|), \; i = 1, 2, \ldots, d. \tag{4.53}$$

Dowód. Dowód jest podobny jak w przypadku dwuwymiarowym (por. dowód lematu 2.17). □

Milne [110] pokazał, że wielowymiarowa krzywa Peano spełnia warunek Höldera (nazywany w pracy warunkiem Lipschitza):

Twierdzenie 4.3.2 *Dla każdego* $t_1, t_2 \in I_1$ *zachodzi*

$$||F_P(t_1) - F_P(t_2)|| \leq \alpha_d \, |t_2 - t_1|^{1/d}, \tag{4.54}$$

gdzie $\alpha_d = 3\sqrt{d+3}$.

Dowód. Dowód twierdzenia podany w pracy [110] bazuje na geometrycznej konstrukcji krzywej Peano. Dowód ten można także przeprowadzić analogicznie jak dowód twierdzenia 4.1.2 dotyczącego krzywej Hilberta. □

Inne oszacowanie stałej w nierówności Höldera podano w pracy [23]. Łatwo zauważyć, że podana w twierdzeniu wartość α_d jest jedynie zgrubnym oszacowaniem najlepszej stałej. Na przykład, dla $d = 2$ wartość ta wynosi $\sqrt{45}$. Jak wykazano w rozdziale 2.3, optymalna, niepoprawialna wartość stałej, jaką możemy wpisać w nierówności (4.54) wynosi, w przypadku $d = 2$, jedynie $\sqrt{8}$. Istnieje możliwość wyznaczenia optymalnej stałej także dla wymiaru większego od $d = 2$. Jednakże uogólnienie metody, zastosowanej w rozdziale 2 w odniesieniu do krzywej dwuwymiarowej, wymaga odrębnego wyznaczania stałej dla każdej wartości d i jest zadaniem wymagającym ogromnych nakładów obliczeniowych nawet dla niedużych wartości d.

4.4 Transformacje quasi-odwrotne do krzywej wypełniającej

Przypomnijmy, że omawiane tu krzywe wypełniające, oznaczmy taką krzywa ogólnie przez F, są ciągłymi krzywymi, które przechodzą przez wszystkie punkty jednostkowej wielowymiarowej kostki I_d. Krzywa wypełniająca $F : I_1 \rightarrow I_d$ jest ciągłym odwzorowaniem odcinka I_1 w kostkę d–wymiarową.

Wiemy również, że odwzorowanie w postaci krzywej wypełniającej nie jest, bo nie może być, odwzorowaniem różnowartościowym. Przeciwobraz dowolnego punktu z I_d względem krzywej wypełniającej $F^{-1}(x)$, $x \in I_d$ może zawierać więcej niż jeden element, przy czym $F^{-1}(x) = \{t : F(t) = x, \; t \in I_1\}$.

Chociaż odwzorowanie F nie jest wzajemnie jednoznaczne, to jednak łatwo można zdefiniować odwzorowanie quasi-odwrotne względem F, oznaczmy je $\boldsymbol{\Psi}$, jako transformację z I_d w odcinek I_1, [124], [158], taką że $\boldsymbol{\Psi}(x) \in F^{-1}(x)$, $x \in I_d$, przy czym oczywiście $F(\boldsymbol{\Psi}(x)) = x$.

W efekcie działania transformacji $\boldsymbol{\Psi}$ dane wielowymiarowe pochodzące z kostki I_d stają się danymi jednowymiarowymi, które znacznie łatwiej można przetwarzać i kompresować, nie tracąc przy tym, jak zobaczymy w dalszych rozdziałach niniejszej monografii, istotnych informacji przestrzennych i statystycznych zawartych w danych wielowymiarowych.

Każda z wielowymiarowych krzywych wypełniających zdefiniowanych w poprzednich podrozdziałach opisana jest za pomocą odpowiedniego układu równań funkcyjnych. Z układów tych równań można wyznaczyć (z dowolną dokładnością) wartość $F(t)$ dla dowolnego $t \in I_1$, gdyż wszystkie omawiane krzywe są nie tylko ciągłe, ale również jednostajnie ciągłe. Korzystając z tych samych układów równań, można wyznaczyć także wartości $F^{-1}(x)$.

Dalej pokażemy, jak wybrać przekształcenie układu równań funkcyjnych definiujących daną krzywą, aby uzyskać postać pozwalającą efektywnie wyznaczać wartości pewnego odwzorowania Ψ. Szczegółowo omówimy quasi-odwrotność krzywej Sierpińskiego F_{Sd}. W przypadku pozostałych krzywych podamy jedynie równania funkcyjne definiujące odpowiednie quasi-odwrotności.

4.4.1 Odwzorowania quasi-odwrotne do krzywych Sierpińskiego

Rozpatrzmy najpierw uogólnienie krzywej Sierpińskiego opisane przez układ równań (4.21)–(4.25), czyli $F_{Sd}(t)$. Przypomnijmy, że $F_{Sd}(t)$ jest krzywą wypełniającą tylko w przypadku, gdy wymiar d jest liczbą parzystą.

Z równań (4.21)–(4.25) wynika bezpośrednio, że wartości odwzorowania $F_{Sd}(t)$ znajdują się w kostce I_d. W szczególności, dla $t \in [1 - c_d\, 2^{-2d}, 1]$ i $t \in [0, b_d\, 2^{-2d}$ znajdują się one w kostce $[0, 1/2] \times [0, 1/2] \times \ldots \times [0, 1/2]$; dla $t \in [b_d\, 2^{-2d}, 2^{-d} + b_d\, 2^{-2d} = t_1]$ znajdują się w kostce $[1/2, 1] \times [0, 1/2] \times \ldots \times [0, 1/2]$; dla $t \in [t_1, t_1 + 2^{-d+1} = t_2]$ wartości odwzorowania znajdują się w kostce $[0, 1] \times [1/2, 1] \times [0, 1/2] \times \ldots \times [0, 1/2]$; ... , a dla $t \in [t_{d-1}, t_{d-1} + 1/2 = t_d = 1 - c_d\, 2^{-2d}]$ znajdują się w kostce $[0, 1] \times \ldots \times [0, 1] \times [1/2, 1]$.

Łatwo zauważyć, że punkty z kostki I_d, których współrzędne mają skończone dwójkowe rozwinięcia mają na ogół, poza punktami leżącymi na brzegu kostki I_d, przeciwobrazy złożone z 2^d punktów na odcinku jednostkowym. Na przykład, punkt $x = (1/2, 1/2, \ldots, 1/2)$ ma przeciwobraz na odcinku w postaci $\{b_d/2^{2d} + i/2^d\}_{i=0,\ldots,2^d-1}$.

W szczególności, z równań (4.21) wynika, że $x_i(t) = 1/2$ $(i = 1, \ldots d)$, gdy $2^d t + c_d/2^d = 1$, czyli $t = 1 - c_d/2^{2d} = b_d/2^{2d}$. Kolejne wartości argumentu t można otrzymać w analogiczny sposób, korzystając z pozostałych równań (4.22)–(4.25). Odwzorowanie $\Psi_{Sd}(x)$ powinno w każdym z tego typu przypadków wskazywać tylko jeden z elementów przeciwobrazu $F_{Sd}^{-1}(x)$.

Lemat 4.7 *Odwzorowanie $\Psi_{Sd} : I_d \to I_1$, gdzie d jest liczbą parzystą, spełniające następujące warunki:*

$$\Psi_{Sd}(x) = -c_d/2^{2d} + \Psi_{Sd}(\hat{x})/2^d,$$
$$lub,\ gdy\ -c_d/2^{2d} + \Psi_{Sd}(\hat{x})/2^d \leq 0,$$

$$\begin{gathered}\boldsymbol{\Psi}_{Sd}(x) = 1 - c_d/2^{2d} + \boldsymbol{\Psi}_{Sd}(\hat{x})/2^d, \\ \hat{x} = (1-2x_1, 1-2x_2, \ldots, 1-2x_d), \\ x = (x_1, \ldots, x_d) \\ x_i \in [0, 1/2], \;\; i = 1, \ldots, , d,\end{gathered} \tag{4.55}$$

$$\begin{gathered}\boldsymbol{\Psi}_{Sd}(x) = (2^d + b_d)/2^{2d} - \boldsymbol{\Psi}_{Sd}(\hat{x})/2^d, \\ \hat{x} = (2x_1 - 1, 1-2x_2, \ldots, 1-2x_d), \\ x = (x_1, \ldots, x_d), \\ x_1 \in (1/2, 1], \;\; x_2, \ldots, x_d \in [0, 1/2],\end{gathered} \tag{4.56}$$

$$\begin{gathered}\boldsymbol{\Psi}_{Sd}(x) = 2t_1 - \boldsymbol{\Psi}_{Sd}(\hat{x}), \\ \hat{x} = (x_1, 1-x_2, x_3, \ldots, x_d), \\ x = (x_1, \ldots, x_d), \\ x_1 \in [0, 1], \;\; x_2 \in (1/2, 1], \;\; x_3, \ldots, x_d \in [0, 1/2], \\ \ldots\end{gathered} \tag{4.57}$$

$$\begin{gathered}\boldsymbol{\Psi}_{Sd}(x) = 2t_{d-2} - \boldsymbol{\Psi}_{Sd}(\hat{x}), \\ \hat{x} = (x_1, \ldots, x_{d-1}, 1-x_{d-1}, x_d), \\ x = (x_1, \ldots, x_d), \\ x_1, \ldots, x_{d-2} \in [0, 1], \;\; x_{d-1} \in (1/2, 1], \;\; x_d \in [0, 1/2],\end{gathered} \tag{4.58}$$

$$\begin{gathered}\boldsymbol{\Psi}_{Sd}(x) = 2t_{d-1} - \boldsymbol{\Psi}_{Sd}(\hat{x}), \\ \hat{x} = (x_1, x_2, \ldots, x_{d-1}, 1-x_d), \\ x = (x_1, \ldots, x_d), \\ x_1, \ldots, x_{d-1} \in [0, 1], \;\; x_d \in (1/2, 1],\end{gathered} \tag{4.59}$$

dla którego zachodzi ponadto $\boldsymbol{\Psi}_{Sd}(1, \ldots, 1) = c_d/2^d$, *jest odwzorowaniem quasi-odwrotnym w stosunku do odwzorowania* F_{Sd}, *opisującego wielowymiarową krzywą Sierpińskiego w postaci (4.21)–(4.25). Podobnie jak w (4.21)–(4.25):* $t_{d-1} = 1 - c_d/2^{2d} - 1/2$ *oraz* $t_l = t_{l+1} - 1/2^{d-l}$, $l = 1 \ldots, d-2$. □

Równanie (4.55) jest równoważnym zapisem równań (4.21), (4.22). Łatwo sprawdzić, że $\boldsymbol{\Psi}_{Sd}(x) = t$, gdzie $x = F_{Sd}(t)$, natomiast $\boldsymbol{\Psi}_{Sd}(\hat{x}) = 2^d(t + c_d/2^{2d})$, gdzie $\hat{x} = F_{Sd}(2^d(t + c_d/2^{2d}))$. Spełnienie warunku $\boldsymbol{\Psi}_{Sd}(x) \in [0, b_d/2^{2d}]$ lub $\boldsymbol{\Psi}_{Sd}(x) \in [1 - c_d/2^{2d}, 1]$ jest zapewnione przez warunek $0 \le x_i \le 1/2, \;\; i = 1, \ldots, d$. Podobnie, równania (4.56)–(4.59) są równoważne (z dokładnością do granicznych wartości współrzędnych $x_i = 1/2$), odpowiednio (4.23)–(4.25). Ustalenie, do którego fragmentu kostki I_d należy punkt $(1/2, 1/2, \ldots, 1/2)$ jest arbitralne i jednoznacznie wiąże się z wyborem wartości $\boldsymbol{\Psi}_{Sd}(1/2, \ldots, 1/2)$ ze zbioru $F_{Sd}^{-1}((1/2, \ldots, 1/2)$. Istnieje 2^d różnych tego typu wyborów. W tym konkretnym przypadku otrzymujemy $\boldsymbol{\Psi}_{Sd}(1/2, \ldots, 1/2) = b_d/2^{2d}$. Wprowadzony w powyższym lemacie warunek $\boldsymbol{\Psi}_{Sd}(1, \ldots, 1)) = c_d/2^d$ nie jest niezbędny, wynika bowiem

z układu równań (4.55)–(4.59). Jest to analogiczna sytuacja do tej, która występuje przy wyznaczaniu krzywej F_{Sd}. Uzupełnienie układu równań (4.55)–(4.59) o wyznaczoną wartość $\Psi_{Sd}(1,\ldots,1)) = c_d/2^d$ pozwala na obliczenie w sposób rekurencyjny, w skończonej liczbie kroków, dokładnej wartości $\Psi_{Sd}(x)$ dla każdego punktu kostki I_d, który ma skończone dwójkowe rozwinięcia wszystkich współrzędnych. Łatwo sprawdzić, że nakład obliczeniowy tego algorytmu jest rzędu $O(d)$ [161].

W odniesieniu do krzywej wypełniającej $F_{Md}(t)$, zdefiniowanej poprzez układ równań (4.21), (4.22),(4.33)–(4.25), odwzorowanie quasi-odwrotne do F_{Md} wyznaczamy podobnie jak w przypadku oryginalnej krzywej Sierpińskiego F_{Sd}.

Lemat 4.8 *Odwzorowanie $\Psi_{Md}(x_1, x_2, \ldots, x_d) : I_d \to I$, które spełnia następujące warunki:*

$$\begin{gathered}\Psi_{Md}(x) = -c_d/2^{2d} + \Psi_{Md}(\hat{x})/2^d,\\ lub,\ gdy\ -c_d/2^{2d} + \Psi_{Md}(\hat{x})/2^d < 0\\ \Psi_{Md}(x) = 1 - c_d/2^{2d} + \Psi_{Md}(\hat{x})/2^d,\\ \hat{x} = (1-2x_1, 1-2x_2, \ldots, 1-2x_d),\\ x = (x_1, \ldots, x_d),\\ x_i \in [0,1/2],\ \ i = 1, \ldots, , d,\end{gathered} \tag{4.60}$$

$$\begin{gathered}\Psi_{Md}(x) = b_d/2^{2d} + \Psi_{Md}(\hat{x})/2^d,\\ \hat{x} = (2x_1 - 1, 1-2x_2, \ldots, 1-2x_d),\\ x = (x_1, \ldots, x_d),\\ x_1 \in (1/2,1],\ \ x_2, \ldots, x_d \in [0,1/2],\end{gathered} \tag{4.61}$$

$$\begin{gathered}\Psi_{Md}(x) = 2t_1 - \Psi_{Md}(\hat{x}),\\ \hat{x} = (x_1, 1-x_2, x_3, \ldots, x_d),\\ x = (x_1, \ldots, x_d),\\ x_1 \in [0,1],\ \ x_2 \in (1/2,1],\ \ x_3, \ldots, x_d \in [0,1/2],\\ \ldots\end{gathered} \tag{4.62}$$

$$\begin{gathered}\Psi_{Md}(x) = 2t_{d-2} - \Psi_{Md}(\hat{x}),\\ \hat{x} = (x_1, \ldots, x_{d-1}, 1-x_{d-1}, x_d),\\ x = (x_1, \ldots, x_d),\\ x_1, \ldots, x_{d-2} \in [0,1],\ \ x_{d-1} \in (1/2,1],\ \ x_d \in [0,1/2],\end{gathered} \tag{4.63}$$

$$\begin{gathered}\Psi_{Md}(x) = 2t_{d-1} - \Psi_{Md}(\hat{x}),\\ \hat{x} = (x_1, x_2, \ldots, x_{d-1}, 1-x_d),\\ x = (x_1, \ldots, x_d),\\ x_1, \ldots, x_{d-1} \in [0,1],\ \ x_d \in (1/2,1]\end{gathered} \tag{4.64}$$

przy $t_{d-1} = 1 - c_d/2^{2d} - 1/2$ oraz $t_l = t_{l+1} - 1/2^{d-l}$, $l = 1 \ldots, d-2$, jest odwzorowaniem quasi-odwrotnym do F_{Md}. □

Również w przypadku odwzorowania Ψ_{Md} dołożenie warunku $\Psi_{Md}(1,\ldots,1) = c_d/2^d$ prowadzi wprost do rekurencyjnego algorytmu wyznaczania wartości funkcji $\Psi_{Md}(x)$. Algorytm ten ma złożoność obliczeniową rzędu $O(d)$ [161] .

4.4.2 Odwzorowanie quasi-odwrotne do krzywej Hilberta

Lemat 4.9 *Odwzorowanie $\Psi_{Hd}(x_1, x_2, \ldots, x_d) : I_d \to I$, które spełnia następujące warunki:*

$$\begin{gathered}\Psi_{Hd}(x) = \Psi_{Hd}(\hat{x})/2^d,\\ \hat{x} = (2x_2, 2x_3, \ldots, 2x_d, 2x_1),\\ x = (x_1, \ldots, x_d),\\ x_i \in [0, 1/2],\ \ i = 1, \ldots, , d,\end{gathered} \tag{4.65}$$

$$\begin{gathered}\Psi_{Hd}(x) = 1/2^d - \Psi_{Hd}(\hat{x}),\\ \hat{x} = (1/2 - x_2, x_3, \ldots, x_d, x_1 - 1/2),\\ x = (x_1, \ldots, x_d),\\ x_1 \in (1/2, 1],\ \ x_2, \ldots, x_d \in [0, 1/2],\\ \ldots\end{gathered} \tag{4.66}$$

$$\begin{gathered}\Psi_{Hd}(x) = 1/2^{d-k} - \Psi_{Hd}(\hat{x}),\\ \hat{x} = (x_1, \ldots, x_{k-1}, 1/2 - x_{k+1}, x_{k+2}, \ldots, x_d, x_k - 1/2),\\ x = (x_1, \ldots, x_d),\\ x_1, \ldots, x_{k-1} \in [0, 1],\ \ x_k \in (1/2, 1],\ \ x_{k+1}, \ldots, x_d \in [0, 1/2],\\ \ldots\end{gathered} \tag{4.67}$$

$$\begin{gathered}\Psi_{Hd}(x) = 1/2 - \Psi_{Hd}(\hat{x}),\\ \hat{x} = (x_1, \ldots, x_{d-2}, 1/2 - x_d, x_{d-1} - 1/2),\\ x = (x_1, \ldots, x_d),\\ x_1, \ldots, x_{d-2} \in [0, 1],\ \ x_{d-1} \in (1/2, 1],\ \ x_d \in [0, 1/2],\end{gathered} \tag{4.68}$$

$$\begin{gathered}\Psi_{Hd}(x) = 1 - \Psi_{Hd}(\hat{x}),\\ \hat{x} = (x_1, \ldots, x_{d-1}, 1 - x_d),\\ x = (x_1, \ldots, x_d),\\ x_1, \ldots, x_{d-1} \in [0, 1],\ \ x_d \in (1/2, 1]\end{gathered} \tag{4.69}$$

jest odwzorowaniem quasi-odwrotnym względem wielowymiarowej krzywej Hilberta F_{Hd}. □

Dołożenie warunku $\Psi_{Hd}(0,\ldots,0) = 0$ prowadzi do rekurencyjnego algorytmu wyznaczania wartości odwzorowania Ψ_{Hd}, który pozwala na obliczenie $\Psi_{Hd}(x)$,

w skończonej liczbie kroków, dla każdego $x \in I_d$, który dodatkowo należy do zbioru $\cup_k \cup_{i=0}^{2^{dk}} \{F_{Hd}(i/2^{dk})\}$. Zbiór ten jest gęsty w I_d. Złożoność obliczeniowa algorytmu jest rzędu $O(d)$. Tę samą złożoność obliczeniową ma również algorytm obliczania quasi-odwrotności innego wariantu wielowymiarowej krzywej Hilberta podany przez Butza [25].

4.4.3 Odwzorowanie quasi-odwrotne do krzywej Peano

Lemat 4.10 *Odwzorowanie $\boldsymbol{\Psi}_{Pd}(x_1, x_2, \ldots, x_d) : I_d \to I$, które spełnia następujące warunki:*

$$\begin{gathered} \boldsymbol{\Psi}_{Pd}(x) = \boldsymbol{\Psi}_{Pd}(\hat{x})/3^d, \\ \hat{x} = (3x_1, 3x_2, \ldots, 3x_d), \\ x = (x_1, \ldots, x_d), \\ x_i \in [0, 1/3], \;\; i = 1, \ldots, , d, \end{gathered} \tag{4.70}$$

$$\begin{gathered} \boldsymbol{\Psi}_{Pd}(x) = \boldsymbol{\Psi}_{Pd}(\hat{x}) - 1/3^d, \\ \hat{x} = (x_1 - 1/3, 1/3 - x_2, \ldots, 1/3 - x_d), \\ x = (x_1, \ldots, x_d), \\ x_1 \in (1/3, 1], \;\; x_2, \ldots, x_d \in [0, 1/3], \end{gathered} \tag{4.71}$$

$$\ldots$$

$$\begin{gathered} \boldsymbol{\Psi}_{Pd}(x) = 1/3^{d-k+1} - \boldsymbol{\Psi}_{Pd}(\hat{x}), \\ \hat{x} = (1 - x_1, \ldots, 1 - x_{k-1}, x_k - 1/3, 1/3 - x_{k+1}, \ldots, 1/3 - x_d), \\ x = (x_1, \ldots, x_d), \\ x_1, \ldots, x_{k-1} \in [0, 1], \;\; x_k \in (1/3, 1], \;\; x_{k+1}, \ldots, x_d \in [0, 1/3], \end{gathered} \tag{4.72}$$

$$\ldots$$

$$\begin{gathered} \boldsymbol{\Psi}_{Pd}(x) = \boldsymbol{\Psi}_{Pd}(\hat{x}) - 1/3, \\ \hat{x} = (1 - x_1, x_2, \ldots, 1 - x_{d-1}, x_d - 1/3), \\ x = (x_1, \ldots, x_d), \\ x_1, \ldots, x_{d-1} \in [0, 1], \;\; x_d \in (1/3, 1] \end{gathered} \tag{4.73}$$

jest odwzorowaniem quasi-odwrotnym względem wielowymiarowej krzywej Peano F_{Pd}. □

Dołożenie warunku $\boldsymbol{\Psi}_{Pd}(0, \ldots, 0) = 0$ prowadzi do rekurencyjnego algorytmu wyznaczania wartości $\boldsymbol{\Psi}_{Pd}(x)$, który pozwala na obliczenie $\boldsymbol{\Psi}_{Pd}(x)$, w skończonej liczbie kroków dla każdego x należącego do zbioru $\cup_k \cup_{i=0}^{3^{dk}} \{F_{Pd}(i/3^{dk})\}$, który jest gęsty w I_d. Złożoność obliczeniowa tego algorytmu jest rzędu $O(d)$, a jego implementację w środowisku *Mathematica* podano w pracy autorki [161].

4.5 Ogólne własności krzywych wypełniających i ich quasi-odwrotności

Ważnym etapem badań prowadzonych przez autorkę był wybór własności krzywych wypełniających, które umożliwiają efektywne użycie transformacji opartej na quasi-odwrotności krzywej wypełniającej w rozwiązywaniu wielowymiarowych problemów decyzyjnych. Pozwolił on zdefiniować klasę krzywych wypełniających efektywnych z tego punktu widzenia jako rodzinę krzywych spełniających sformułowane dalej warunki **C1**–**C3**.

C1. Krzywa wypełniająca F jest ciągłym odwzorowaniem spełniającym warunek Höldera, czyli

$$||F(t_1) - F(t_2)|| \leq \alpha |t_1 - t_2|^{\beta}, \;\; t_1, \, t_2 \in I_1, \tag{4.74}$$

przy czym $\beta = 1/d$. Symbol $||.||$ oznacza normę euklidesową w I_d, natomiast $\alpha > 0$ jest pewną stałą zależną od wymiaru d i typu krzywej.

Należy zauważyć, że postać odwzorowania F zależy od wymiaru d, nie jest to tutaj jednak bezpośrednio uwzględnione w notacji. Konwencja ta będzie zachowana dalej także w stosunku do odwzorowania quasi-odwrotnego $\boldsymbol{\Psi}$.

W przypadku klasycznych krzywych dwuwymiarowych jesteśmy w stanie podać dokładne wartości stałej α (por. odpowiednie twierdzenia z rozdziału 2), natomiast w odniesieniu do krzywych wielowymiarowych znane są w ogólnym przypadku tylko górne oszacowania wartości stałej Höldera.

W sformułowaniu własności **C1** może, zamiast odległości euklidesowej, pojawić się inna metryka. Z ogólnej własności przestrzeni R^d wynika, że przy zmianie metryki warunek Höldera **C1** zostanie zachowany z dokładnością do wartości stałej α.

Przypomnijmy dalej, że w ogólnym przypadku wielowymiarowej krzywej wypełniającej, wykładnik Höldera, $\beta > 0$ (nazywany czasami także wykładnikiem Lipschitza) jest nie większy niż $1/d$, czyli dla $d \geq 2$ jest mniejszy od jedności.

Istnieje ścisły związek pomiędzy wymiarem topologicznym d wypełnianej kostki I_d, a maksymalną wartością wykładnika β [110], [138].

Lemat 4.11 [*110*] *Nie istnieje d–wymiarowa krzywa wypełniająca spełniająca warunek Höldera (4.74) z wykładnikiem β większym od* $1/d$.

Dowód. Dowód powyższej własności podano w pracy Milne [110]. □

Ponieważ wykładnik w warunku Höldera jest co najwyżej równy $1/d$, stąd wprawdzie F jest ciągła, lecz nie jest różniczkowalna. Dowody tego faktu można

znaleźć w [137]. W warunku **C1** domagamy się, by wykładnik β był dokładnie równy $1/d$. Warunek Höldera możemy interpretować jako własność zachowywania bliskości w takim sensie, że punkty bliskie sobie na odcinku jednostkowym są transformowane przez krzywą wypełniającą F w punkty bliskie sobie na kostce I_d. Mniejsza od $1/d$ wartość wykładnika β powoduje pogorszenie własności krzywej w tym względzie. Nie wszystkie wielowymiarowe krzywe wypełniające spełniają warunek **C1**. Przykładem są krzywe powstałe poprzez iterowanie odwzorowań dwuwymiarowych (patrz rozdział 3).

Kolejną własnością omawianej tu klasy krzywych wypełniających jest to, że krzywe wypełniające są odwzorowaniami wzajemnie jednoznacznymi prawie wszędzie, czyli z dokładnością do zbiorów miary Lebesgue'a zero. Na własność tę w odniesieniu do dwuwymiarowej krzywej Sierpińskiego zwrócili uwagę Bartholdi i Platzman w pracy [124].

C2. Krzywa wypełniająca F jest prawie wszędzie (z dokładnością do zbiorów miary Lebesgue'a zero) odwzorowaniem wzajemnie jednoznacznym.

Wszystkie omawiane w niniejszej monografii krzywe posiadają tę własność. Formalny dowód własności **C2** wymaga w każdym przypadku dokładnej analizy odwzorowania. Dowód taki w odniesieniu do wielowymiarowej krzywej Sierpińskiego przedstawiono w rozdziale 3.2.2 (porównaj także dowód w pracy autorki [158]). Rozumowanie to łatwo powtórzyć w odniesieniu do krzywych Hilberta i Peano.

Kolejną istotną własnością omawianej klasy krzywych wypełniających jest zachowywanie przez nie miary Lebesgue'a.

Będziemy definiować przeciwobraz zbioru B względem odwzorowania F jako $F^{-1}(B) = \{t \in I_1 : F(t) \in B\}$ dla $B \subseteq I_d$. Niech μ_d oznacza miarę Lebesgue'a w I_d, która jest tu traktowana jako wielkość bezwymiarowa.

C3. Krzywa wypełniająca $F : I_1 \to I_d$ zachowuje miarę Lebesgue'a, to znaczy, dla każdego zbioru borelowskiego $B \subseteq I_d$ zachodzi

$$\mu_d(B) = \mu_1\left(F^{-1}(B)\right). \tag{4.75}$$

Zachowywanie miary Lebesgue'a może być interpretowane jako jednostajnie gęste wypełnianie kostki I_d przez krzywą wypełniającą F, co oznacza, iż fragmenty krzywej o tej samej długości wypełniają wielowymiarowe obszary o tych samych objętościach.

Prawie wszystkie klasyczne krzywe wypełniające zachowują miarę Lebesgue'a (wyjątkiem jest między innymi krzywa Lebesgue'a) [137]. Formalny dowód własności **C3** w odniesieniu do wielowymiarowej krzywej Peano podał Milne [110].

Analogiczny dowód dla wielowymiarowej krzywej Sierpińskiego przedstawiono w rozdziale 3.2.2 (porównaj także dowód w pracy autorki [158]). Dowód taki łatwo powtórzyć także w odniesieniu do krzywej Hilberta.

Kolejne lematy wynikają z własności **C1**–**C3**. Dalej będziemy stale milcząco zakładać, że warunki **C1**–**C3** są zawsze spełnione.

Niech Ψ oznacza funkcję, która dla każdego $x \in I_d$, wybiera dokładnie jeden element z jego przeciwobrazu $F^{-1}(x)$, czyli Ψ spełnia warunek $\Psi(x) \in F^{-1}(x)$, a zatem $F(\Psi(x)) = x$.

Lemat 4.12 *Jeżeli krzywa wypełniająca F jest odwzorowaniem wzajemnie jednoznacznym prawie wszędzie, z dokładnością do zbiorów miary Lebesgue'a zero, to istnieje funkcja $\Psi : I_d \rightarrow I_1$, która jest odwzorowaniem prawie wszędzie odwrotnym do F.* □

Wiadomo, iż dowolna krzywa wypełniająca F nie może być odwzorowaniem wzajemnie jednoznacznym (porównaj na przykład [137], [178]). W związku z tym nie jest to odwzorowanie odwracalne. Żadna krzywa wypełniająca nie jest krzywą Jordana w ścisłym sensie, a zatem miejscom geometrycznym przecinania się krzywej samej ze sobą odpowiadają różne punkty z odcinka $[0, 1]$.

Możliwość szybkiego wyznaczenia dla każdego punktu $x \in I_d$ z przestrzeni wielowymiarowej odpowiadającego mu punktu z odcinka $t \in I_1$ takiego, że $F(t) = x$, jest najważniejszą cechą odwzorowania w postaci krzywej wypełniającej. Transformacja $\Psi : I_d \rightarrow I_1$, taka że $\Psi(t) \in F^{-1}(x)$, gdzie $F^{-1}(x)$ oznacza przeciwobraz x względem krzywej wypełniającej F, jest nazywana tu quasi-odwrotnością krzywej wypełniającej.

Odwzorowanie to pozwala uporządkować liniowo dane z przestrzeni wielowymiarowej. Jest to, z punktu widzenia celów tej monografii, transformacja znacznie ważniejsza niż odpowiadające jej odwzorowanie w postaci krzywej wypełniającej. Zastosowanie transformacji Ψ w odniesieniu do wielowymiarowych danych prowadzi do formalnej redukcji wymiaru problemu, a w konsekwencji ułatwia kompresję danych, bez utraty istotnych informacji przestrzennych zawartych w danych wielowymiarowych.

W odwzorowaniach bazujących na krzywej wypełniającej istotne jest to, że położone blisko siebie punkty z odcinka I_1 przekształcane są w punkty położone blisko siebie w I_d. W związku z tym punkty leżące blisko siebie na odcinku są w transformacji quasi-odwrotnej obrazami punktów położonych blisko siebie w przestrzeni wielowymiarowej.

W ogólnym przypadku można, bazując jedynie na założeniu **C3** i korzystając z pewnika wyboru [93], zagwarantować istnienie odwzorowania quasi-odwrotnego

do określonej krzywej wypełniającej. Jednakże, w praktyce potrzebujemy konstruktywnej metody definiowania quasi-odwrotności $\boldsymbol{\Psi}$. Takie metody wyznaczania quasi-odwrotności dla klasy krzywych wypełniających, którymi zajmujemy się w niniejszej monografii, przedstawiono w podrozdziale 4.4. W przypadku omawianych krzywych, $F^{-1}(\{x\})$ jest zbiorem jednoelementowym (prawie wszędzie) lub jest zbiorem liczbowym (zawartym w I_1) zawierającym skończoną liczbę elementów, co pozwala na konstruktywne definiowanie quasi-odwrotności $\boldsymbol{\Psi}$.

Kolejną istotną własnością odwzorowania $\boldsymbol{\Psi}$ jest własność sformułowana w następującym lemacie:

Lemat 4.13 *$\boldsymbol{\Psi}$ zachowuje miarę Lebesgue'a w tym sensie, że dla każdego zbioru borelowskiego $A \subset I_1$ zachodzi*

$$\mu_1 \left(A \cap \boldsymbol{\Psi}(I_d) \right) = \mu_d \left(\boldsymbol{\Psi}^{-1} \left(A \cap \boldsymbol{\Psi}(I_d) \right) \right). \tag{4.76}$$

□

Na własność sformułowaną w lemacie 4.13 w odniesieniu do dwuwymiarowej krzywej Sierpińskiego zwrócili uwagę Bartholdi i Platzman w pracy [124]. Formalny dowód powyższej własności w odniesieniu do wielowymiarowej krzywej Sierpińskiego podano jako dowód lematu 3.17 z rozdziału 3.2.2.

Z warunku **C1** wynika także następująca własność:

Lemat 4.14 *Jeśli $f(x)$ jest funkcją ciągłą, określoną w I_d, to $g(t) \stackrel{def}{=} f(F(t))$, $t \in I_1$ jest także funkcją ciągłą. Ponadto, jeśli $f(x)$ spełnia warunek Höldera z wykładnikiem ν, $0 < \nu \leq 1$, to znaczy*

$$||f(x) - f(y)|| \leq c\, ||x - y||^{\nu},\ x, y \in I_d,$$

gdzie c jest pewną stałą, to $g(t)$ jest także funkcją ciągłą i spełnia warunek Höldera z wykładnikiem ν/d.

Dowód. Dla każdego x, $y \in I_d$ zachodzi $||f(x) - f(y)|| \leq c\, ||x - y||^{\nu}$. Z własności **C1** wynika więc, że dla każdego t_1, $t_2 \in I_1$ spełnione jest $||g(t_1) - g(t_2)|| = ||f(F(t_1)) - f(F(t_2))|| \leq c\, ||F(t_1) - F(t_2)||^{\nu} \leq c\, \alpha(|t_1 - t_2|^{1/d})^{\nu}$, co kończy dowód. □

Korzystać będziemy także z następującego lematu:

Lemat 4.15 *Dla każdej funkcji mierzalnej $f : I_d \rightarrow R$, przy czym f należy do klasy funkcji całkowalnych w I_d, $f \in L(I_d)$, następujące całki Lebesgue'a, określone odpowiednio w I_d i w I_1, są sobie równe*

$$\int_{I_d} f(x)dx = \int_{I_1} f(F(t))dt. \tag{4.77}$$

Dowód. Dowód lematu wynika bezpośrednio z własności **C3** oraz twierdzenia o przenoszeniu miary i zamianie zmiennych (patrz twierdzenie 16.12 z [15]). □

Ponieważ Ψ jest funkcją mierzalną, zatem jeśli X jest zmienną losową przyjmującą wartości w I_d (wektorem losowym), to $\Psi(X)$ jest także zmienną losową przyjmującą wartości w I_1. W szczególności spełniona jest następująca własność:

Lemat 4.16 *Jeśli zmienne losowe $X_1, X_2, \ldots, X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie określonym w I_d i posiadającymi gęstość $f(x)$, to zmienne losowe $t_i = \Psi(X_i)$ są również niezależnymi zmiennymi losowymi o tym samym rozkładzie w I_1, przy czym gęstość tego rozkładu istnieje i ma postać $f(F(t))$.* □

Własność tego typu została sformułowana w pracy [124] dla przypadku dwuwymiarowej krzywej Sierpińskiego, łatwo ją jednak uogólnić dla dowolnej krzywej wypełniającej spełniającej warunki **C1**–**C3**.

Konsekwencje lematów 4.16 oraz 4.15 są daleko idące. Pokazują one bowiem, że transformacja Ψ nie zachowuje wartości momentów rozkładów czy odległości Mahalanobisa, zachowuje natomiast wszelkie kryteria całkowe będące funkcjonałami gęstości rozkładów prawdopodobieństwa, takie jak entropia rozkładu czy odległość Kullbacka-Leiblera [91].

Z formalnego punktu widzenia wszystkie krzywe wypełniające, które spełniają sformułowane powyżej warunki **C1**–**C3**, mogą być stosowane w omawianych w niniejszej monografii problemach decyzyjnych. Spełniają je wszystkie krzywe wypełniające zdefiniowane w niniejszej pracy. Należy jednak stwierdzić, że istnieją krzywe wypełniające, które nie spełniają wszystkich wymienionych dalej własności. Przykładem jest tu krzywa Lebesgue'a [137], która nie spełnia warunku **C3**.

4.6 Podsumowanie

W rozdziale tym zaproponowano oryginalne opisy definicyjne wielowymiarowych krzywych wypełniających kostkę I_d, podane w postaci odpowiednich układów równań funkcyjnych. Rozwiązaniem tych układów równań są odpowiednie odwzorowania w postaci krzywej wypełniającej.

Podano uogólnienia dwuwymiarowych krzywych Hilberta, Peano i Sierpińskiego do postaci wielowymiarowej oraz zbadano ich własności. Użyte układy równań funkcyjnych oparte są bezpośrednio na postulowanych własnościach geometrycznych poszczególnych krzywych.

W każdym z przypadków wyjściowy układ równań funkcyjnych został przekształcony do równoważnej postaci, umożliwiającej jego efektywne rozwiązanie za pomocą odpowiedniej procedury rekurencyjnej. Procedury te pozwalają na

wyznaczenie dokładnych wartości odwzorowań na zbiorach gęstych w I_1. W rozdziale tym zbadano również podstawowe własności zaproponowanych odwzorowań, w szczególności podano odpowiednie szczegółowe sformułowania warunku Höldera spełnianego przez poszczególne krzywe wielowymiarowe.

Pokazano, że korzystając z tych samych układów równań funkcyjnych, które definiują daną krzywą wypełniającą, można także wyznaczać zbiory wartości $F^{-1}(x)$.

Wskazano również, w jaki sposób należy wybrać przekształcenie układu równań funkcyjnych definiujących daną krzywą, tak by uzyskać postać pozwalającą efektywnie wyznaczać wartości odwzorowania quasi-odwrotnego $\boldsymbol{\Psi}$. Szczegółowo omówiono proces definiowania quasi-odwrotności krzywej Sierpińskiego.

Zdefiniowano klasę krzywych wypełniających, które umożliwiają efektywne użycie transformacji opartej na quasi-odwrotności krzywej wypełniającej w rozwiązywaniu wielowymiarowych problemów decyzyjnych.

Rozdział 5

Ocena wymiaru fraktalnego obiektów wielowymiarowych z użyciem krzywych wypełniających

Ocena wymiaru fraktalnego na podstawie danych pochodzących z rzeczywistych obiektów jest coraz częściej stosowaną procedurą pomiarową w wielu dziedzinach nauki i techniki. Dokonanie prawidłowej oceny wymiaru fraktalnego nie jest jednak łatwym zadaniem. Istnieje bowiem bardzo wiele różnych definicji wymiaru związanych z obiektami o strukturze fraktalnej [7], [48]. Wymiar pudełkowy (*box-counting dimension*) jest najczęściej stosowany w przypadku szacowania wymiaru obiektów na podstawie danych empirycznych. W większości przypadków fraktalnych obiektów geometrycznych – zbiorów zwartych w przestrzeni metrycznej – wartość wymiaru pudełkowego jest równa wymiarowi Hausdorffa [48], dlatego czasami, aczkolwiek niesłusznie, bywają one ze sobą utożsamiane. Prostota koncepcyjna najbardziej znanej metody, polegającej na bezpośrednim szacowaniu wymiaru pudełkowego, często przysłania fakt, że metoda ta ma bardzo dużą złożoność obliczeniową, rosnącą szybko z wymiarem przestrzeni, w której zanurzony jest mierzony obiekt (por. [141], gdzie metodę tę ocenia się jako zupełnie nieefektywną w przestrzeniach wielowymiarowych). Duża złożoność obliczeniowa z kolei prowadzi do utraty dokładności, jeśli próbuje się szacować wymiary w przestrzeniach wielowymiarowych, ograniczając nakłady obliczeniowe poprzez redukcję liczby zagłębień ,,pudełek", które służą do zliczania punktów fraktalnego obiektu. Z tego powodu opracowuje się inne metody, takie jak pomiar wymiaru korelacyjnego. Są one jednak oparte na innych, nie zawsze równoważnych definicjach wymiaru fraktalnego, a ponadto mają ograniczony obszar zastosowań, na przykład metoda pomiaru wymiaru korelacyjnego wymaga spełnienia specyficznych założeń (por. [141]).

Fakty te motywują podjęcie próby opracowania nowego podejścia do szacowania wymiaru fraktalnego i zbadania jego dokładności na przykładach obiektów o znanych wymiarach, tym bardziej że monitorowanie na bieżąco wymiaru fraktalnego aktualnych obserwacji może stanowić cenne źródło informacji o stanie systemu.

Istotą tego podejścia jest dokonanie transformacji danych wielowymiarowych na odcinek I_1 przez quasi-odwrotność krzywej wypełniającej, dokonanie tu oceny wymiaru fraktalnego za pomocą szacowania wymiaru pudełkowego i bardzo prostego przetransformowania wyniku z powrotem do przestrzeni wielowymiarowej na podstawie ścisłej zależności pomiędzy wymiarami pudełkowymi danych przed i po transformacji. Zaletą proponowanego podejścia jest znaczna redukcja nakładów obliczeniowych, a co za tym idzie – także potencjalny wzrost dokładności, gdyż zliczanie wymiaru pudełkowego na odcinku jest bardzo proste, a transformacja danych na odcinek również ma liniową względem wymiaru d złożoność obliczeniową. Powyżej użyto określenia „potencjalny wzrost dokładności", gdyż dokładność jakiejkolwiek metody empirycznej oceny wymiaru fraktalnego zależy od liczby danych. Jednakże dopiero posiadanie narzędzia do efektywnej obróbki dużej liczby danych pozwala uzyskiwać spodziewaną dokładność oszacowań. Głównym wynikiem niniejszego rozdziału jest wyprowadzenie teoretycznej zależności między wymiarem pudełkowym wielowymiarowych danych a wymiarem pudełkowym danych przetransformowanych przy użyciu krzywej wypełniającej.

5.1 Uzasadnienie i opis metody

Zastosujemy tu krzywe wypełniające, a raczej ich quasi-odwrotności, które spełniają omówione w rozdziale 4.5 warunki **C1**–**C3**.

Szczególnie ważną dla naszych celów własnością krzywych wypełniających jest ich zdolność do zachowywania miary Lebesgue'a, czyli własność **C3**. Przypomnijmy, że w tym przypadku własność zachowywania miary można zapisać następująco:

$$\forall_{B\in\mathcal{B}} \qquad \mu_d(B) = \mu_1(F^{-1}(B)), \tag{5.1}$$

gdzie μ_d oznacza miarę Lebesgue'a w przestrzeni I_d, μ_1 jest miarą na odcinku, $\mathcal{B}$ oznacza klasę wszystkich zbiorów borelowskich w kostce I_d, natomiast przeciwobrazem zbioru $B \in \mathcal{B}$ względem krzywej F jest $F^{-1}(B) = \{t \in [0,1] : F(t) \in B\}$. Zauważmy, że własność **C3** oznacza **liczbową** równość miar odpowiadających sobie zbiorów z kostki i odcinka. Jeśli na przykład wyróżnimy w I_d kostkę o boku $\epsilon \in (0,1)$, to jej przeciwobraz na odcinku będzie miał długość ϵ^d, gdyż tyle właśnie wynosi objętość tej kostki. Wiadomo również, że krzywe wypełniające nie mogą mieć odwzorowania odwrotnego w zwykłym sensie. Zgodnie z **C2**, zbiór $\Psi(B)$

różni się od zbioru $F^{-1}(B)$ tylko przeliczalną liczbą punktów, stąd

$$\mu_1(\Psi(B)) = \mu_1(F^{-1}(B)).$$

Jeśli geometryczna konstrukcja krzywej wypełniającej zawiera podział kostki I_d na podkostki o bokach γ^k, to podkostkom tym odpowiadają, wzajemnie jednoznacznie, pododcinki odcinka I_1 o długości $\gamma^{k\,d}$ (por. rozdział 1.5 oraz [93], [111]). Dodajmy, że dla krzywych Hilberta i Sierpińskiego $\gamma = 1/2$, a dla krzywej Peano $\gamma = 1/3$, niezależnie od wymiaru krzywej. W niniejszej monografii nie korzystamy przy definiowaniu krzywych z takiej geometrycznej konstrukcji, jednakże wspomniana własność jest cechą charakterystyczną krzywych wypełniających, które zachowują miarę Lebesgue'a. W podrozdziale 3.2.2 wykazano, że taka wzajemnie jednoznaczna zależność istnieje w odniesieniu do wielowymiarowej krzywej Sierpińskiego. Dowód podobnej własności w przypadku wielowymiarowej krzywej Peano podano w pracy Milne [110].

5.1.1 Teoretyczna zależność między wymiarami fraktalnymi zbioru i jego obrazu na odcinku

Oznaczmy przez $X \subset I_d$ zbiór, którego wymiar fraktalny mamy zdefiniować, a w następnych podrozdziałach także empirycznie zmierzyć. Jak wiadomo [48], tak zwany wymiar pudełkowy zbioru $X \subset R^d$, oznaczany dalej przez $\mathrm{Dim}(X)$, jest zdefiniowany następująco:

Definicja 5.1 *Dolny wymiar pudełkowy jest równy*

$$\underline{\mathrm{Dim}}(X) = \liminf_{\epsilon \to 0} - \frac{\log N(\epsilon)}{\log \epsilon}. \tag{5.2}$$

Górny wymiar pudełkowy jest równy

$$\overline{\mathrm{Dim}}(X) = \limsup_{\epsilon \to 0} - \frac{\log N(\epsilon)}{\log \epsilon}. \tag{5.3}$$

Wymiar pudełkowy jest równy

$$\mathrm{Dim}(X) = \lim_{\epsilon \to 0} - \frac{\log N(\epsilon)}{\log \epsilon}, \tag{5.4}$$

jeśli powyższa granica istnieje. $N(\epsilon)$ jest liczbą określoną w jeden z następujących sposobów:

(i) *jest najmniejszą liczbą kul domkniętych o promieniu ϵ, które pokrywają X,*

(ii) *jest najmniejszą liczbą kostek domkniętych o boku ϵ, które pokrywają X,*

(iii) *jest liczbą kostek domkniętych o boku ϵ, które pokrywają I_d i które równocześnie mają punkty wspólne z X,*

(iv) *jest najmniejszą liczbą zbiorów (każdy z nich zawarty w kuli o średnicy ϵ), które pokrywają X,*

(v) *jest największą liczbą rozłącznych kul domkniętych o promieniu ϵ, których środki zawarte są w X.*

Zauważmy, że $\boldsymbol{\Psi}(X) \in F^{-1}(X)$ oraz zbiory te różnią się tylko przeliczalną liczbą punktów. Niestety, wymiar pudełkowy zbioru przeliczalnego, w przeciwieństwie do jego wymiaru Hausdorffa, niekoniecznie jest równy zeru [48]. Niemniej jednak dla każdego $x \in I_d$ zbiór $F^{-1}(x)$ zawiera co najwyżej 2^d elementów, punktów z I_1. Stąd liczba ϵ–odcinków pokrywających zbiór $F^{-1}(X)$ jest co najwyżej 2^d razy większa od liczby ϵ-odcinków pokrywających zbiór $\boldsymbol{\Psi}(X)$, niezależnie od wartości $\epsilon > 0$. W związku z tym odpowiednie wymiary pudełkowe zbiorów $F^{-1}(X)$ oraz $\boldsymbol{\Psi}(X)$ są sobie równe, czyli:

Lemat 5.1

$$\underline{\mathrm{Dim}}(F^{-1}(X)) = \underline{\mathrm{Dim}}(\boldsymbol{\Psi}(X)), \tag{5.5}$$

$$\overline{\mathrm{Dim}}(F^{-1}(X)) = \overline{\mathrm{Dim}}(\boldsymbol{\Psi}(X)), \tag{5.6}$$

$$\mathrm{Dim}(F^{-1}(X)) = \mathrm{Dim}(\boldsymbol{\Psi}(X)), \tag{5.7}$$

jeśli powyższy wymiar istnieje.

Skorzystajmy z definicji (iii) i podzielmy kostkę I_d na podkostki o jednakowych, wynoszących $\epsilon > 0$, długościach boków. Dokładniej, kostki te są postaci:

$$[(i_1 - 1)\epsilon, i_1\epsilon] \times \ldots \times [(i_d - 1)\epsilon, i_d\epsilon],$$

gdzie $i_1, \ldots, i_d = 1, 2, \ldots, \lceil 1/\epsilon \rceil$. Niech $N(\epsilon)$ oznacza liczbę tych ϵ-podkostek, które mają punkty wspólne ze zbiorem X. Zależność $N(\epsilon)$ użyta w (5.4) określona jest operacyjnie w sposób opisany powyżej. Jeśli wymiar pudełkowy danego zbioru istnieje, to granicę w (5.4) można obliczać wybierając dowolne podciągi $\epsilon_k \to 0$, gdy $k \to \infty$. W naszym przypadku wygodnie będzie posłużyć się podciągami postaci $\epsilon_k = \gamma^k$, $0 < \gamma < 1$, które odpowiadają podziałom kostki I_d przy konstrukcji krzywych Hilberta i Sierpińskiego (podział dwójkowy, $\gamma = 1/2$) oraz krzywej Peano (podział trójkowy, $\gamma = 1/3$). Wówczas

$$\mathrm{Dim}(X) = \lim_{k \to \infty} - \frac{\log N(\gamma^k)}{k \log(\gamma)}. \tag{5.8}$$

Zbadajmy teraz wymiar Dim($\Psi(X)$) zbioru X po przekształceniu go na odcinek $[0,1]$ za pomocą Ψ. Podkostki o bokach γ^k zostają wówczas przetransformowane na pododcinki o długościach $\gamma^{k\,d}$, zatem

$$\mathrm{Dim}(\ \Psi(X)) = \lim_{k\to\infty} \ - \frac{\log L(\gamma^{k\,d})}{k\,d\,\log(\gamma)}, \tag{5.9}$$

gdzie $L(\gamma^{k\,d})$ oznacza liczbę pododcinków, które mają punkty wspólne z $\Psi(X)$. Jednakże mamy wzajemnie jednoznaczną odpowiedniość między podkostkami w I_d (o boku γ^k) i odpowiednimi podocinkami w I_1. Zatem $N(\gamma^k) = L(\gamma^{k\,d})$, co poprzez porównanie (5.8) oraz (5.9) prowadzi do następującego twierdzenia:

Twierdzenie 5.1.1 *Niech $X \subset I_d$. Jeśli wymiar pudełkowy* Dim(X) *istnieje oraz istnieje* Dim($\Psi(X)$), *to*

$$\mathrm{Dim}(X) \ = \ d\,\mathrm{Dim}(\ \Psi(X)). \tag{5.10}$$

□

W istocie rzeczy można też udowodnić mocniejszą wersję powyższego twierdzenia, która pozwala równość (5.10) zastosować także w odniesieniu do $\underline{\mathrm{Dim}}(X)$ oraz $\overline{\mathrm{Dim}}(X)$.

Dość łatwo jest pokazać, korzystając z definicji (i) oraz własności **C1** krzywej, że zachodzi

$$\begin{aligned} \underline{\mathrm{Dim}}(X) \ &\leq \ d\,\underline{\mathrm{Dim}}(\ \Psi(X)), \\ \overline{\mathrm{Dim}}(X) \ &\leq \ d\,\overline{\mathrm{Dim}}(\ \Psi(X)). \end{aligned} \tag{5.11}$$

W konsekwencji, jeśli Dim(X) istnieje, to otrzymujemy

$$\mathrm{Dim}(X) \ \leq \ d\,\mathrm{Dim}(\ \Psi(X)). \tag{5.12}$$

Podzielmy odcinek I_1 na przedziały o długości ϵ, mające postać:

$$C_k = [(k-1)\,\epsilon, k\,\epsilon], \ \ k = 1, 2, \ldots, \lceil 1/\epsilon \rceil.$$

Niech $L(\epsilon)$ oznacza liczbę takich przedziałów, które mają punkty wspólne z $\Psi(X)$. Zgodnie z własnością **C1**, zbiór $F(C_k)$ jest zawarty w pewnej kuli o średnicy $\alpha\epsilon^{1/d}$. W konsekwencji najmniejsza liczba kul pokrywających zbiór X, oznaczmy ją przez $N_d(\epsilon)$, jest mniejsza, co najwyżej równa liczbie $L(\epsilon)$. Stąd otrzymujemy

$$\frac{\log N_d(\epsilon)}{-\log\left(\alpha\epsilon^{1/d}\right)} \leq d\frac{\log L(\epsilon)}{-d\log\alpha - \log\epsilon},$$

co prowadzi do nierówności (5.11). Pozostaje jeszcze wykazanie nierówności przeciwnych do nierówności (5.11). Wybierzmy taką liczbę naturalną k_ϵ, że

$$\gamma^{k_\epsilon+1} \leq \alpha\epsilon^{1/d} \leq \gamma^{k_\epsilon}.$$

Zauważmy, że każda kula o średnicy $\alpha\epsilon^{1/d}$ ma punkty wspólne z co najwyżej 2^d kostkami o boku γ^{k_ϵ}, a w konsekwencji ma punkty wspólne z co najwyżej $(2/\gamma)^d$ kostkami o boku $\gamma^{k_\epsilon+1}$. Liczba takich kostek, które mają punkty wspólne z X, czyli $N(\gamma^{k_\epsilon+1})$, jest zatem ograniczona od góry przez $(2/\gamma)^d N_d(\epsilon)$. Zgodnie z opisanymi wcześniej własnościami krzywej wypełniającej, mamy $L(\gamma^{(k_\epsilon+1)d}) = N(\gamma^{k_\epsilon+1})$. Ponieważ $\gamma^{(k_\epsilon+1)d} \leq \alpha^d\epsilon$, więc $L(\gamma^{(k_\epsilon+1)d}) \geq L(\alpha^d\epsilon)$. W konsekwencji otrzymujemy nierówność

$$\frac{\log N_d(\epsilon)}{-\log(\alpha\epsilon^{1/d})} \geq \frac{\log L(\alpha^d\epsilon) - d\log 2/\gamma}{-\log(\alpha\epsilon^{1/d})} = \frac{d\log L(\alpha^d\epsilon) - d^2\log 2/\gamma}{-\log(\alpha^d\epsilon)},$$

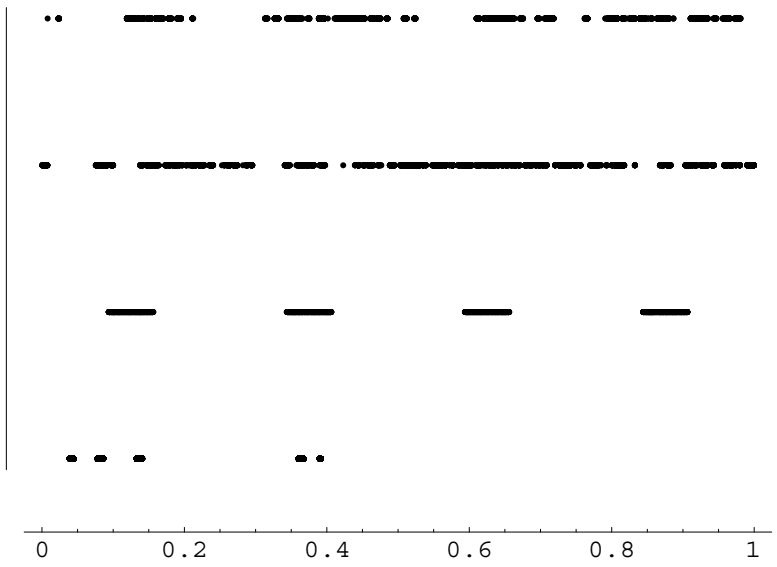
co prowadzi do nierówności przeciwnych nierównościom (5.11), czyli

$$\begin{aligned} \underline{\mathrm{Dim}}(X) &\geq d\,\underline{\mathrm{Dim}}(\boldsymbol{\Psi}(X)), \\ \overline{\mathrm{Dim}}(X) &\geq d\,\overline{\mathrm{Dim}}(\boldsymbol{\Psi}(X)). \end{aligned} \tag{5.13}$$

Przy istnieniu odpowiednich granic otrzymujemy oczywiście równość (5.10). Zależność (5.10) stanowi podstawę teoretyczną algorytmu pomiaru wymiaru fraktalnego wielowymiarowego zbioru $X \subset I_d$ na podstawie pomiaru wymiaru zbioru $\boldsymbol{\Psi}(X) \subset I_1$. Jej prostota nie powinna przesłaniać tego, że całe bogactwo wielowymiarowych tworów geometrycznych zostaje tu zredukowane do jednego wymiaru bez straty informacji o ich wymiarach fraktalnych. Różnorodność kształtów zbiorów wielowymiarowych zostaje w istocie odwzorowana w lokalne zagęszczenia punktów na odcinku. Na rysunku 5.1 pokazano równocześnie kilka obrazów na odcinku zbiorów o znanych wymiarach, pochodzących z przestrzeni o różnych wymiarach topologicznych. Dla zachowania czytelności, zbiory te (a raczej ich przedstawienia graficzne) rozsunięto w pionie.

5.1.2 Opis algorytmu empirycznej oceny wymiaru fraktalnego

Niech $x_i \in I_d$ $(i = 1, 2, \ldots, n)$ będzie ciągiem punktów pochodzących ze zbioru X, którego wymiar fraktalny ma zostać zmierzony. Milcząco zakładamy przy tym, że n jest dostatecznie duże, nie precyzując przy tym rzędu wielkości n, gdyż – jak to widać z przytoczonych dalej badań symulacyjnych – rząd ten musi zależeć od wymiaru przestrzeni, w której zanurzony jest mierzony zbiór. Należy zauważyć, że teoretyczny wymiar skończonego zbioru punktów jest zawsze równy zeru.



Rys. 5.1. Obrazy tworów geometrycznych o różnych wymiarach, po przetransformowaniu ich na odcinek $[0,1]$ za pomocą quasi-odwrotności krzywej Sierpińskiego. Licząc od dołu rysunku pokazano obrazy: a) funkcji kwadratowej o wymiarze 1, b) kwadratu – wymiar = 2, c) powierzchni $|\sin(\pi(x_1^2+x_2^2))|$ dwuwymiarowej w R^3 – wymiar = 2, d) sygnału fraktalnego (traktowanego jako funkcja na płaszczyźnie) o wymiarze 1,5. We wszystkich przypadkach obrazy uzyskano dla 5000 punktów, a fakt, że nie są one widoczne na wykresach wynika nie tylko z zastosowanej rozdzielczości, ale głównie z geometrycznej natury transformowanych obiektów
Fig. 5.1. Images of different geometric objects after transforming them to $[0,1]$ via quasi-inverse of the Sierpiński curve. From below: a) quadratic function with dimension 1, b) square – dimension = 2, c) two–dimensional surface $|\sin(\pi(x_1^2+x_2^2))|$ in R^3 – dimension = 2, d) signal with fractal structure (treated as the graph of a function) dimension= 1.5. Each plot consists of 5000 points. Points are not visible due to the geometric structure of the transformed objects

W przypadku szacowania wymiaru fraktalnego zakładamy, iż rzeczywisty obiekt zawiera w istocie nieprzeliczalną liczbę punktów, których struktura jest samopodobna i ściśle określona przez strukturę pobranego ciągu punktów. Możemy teraz opisać proponowany algorytm pomiaru wymiaru fraktalnego.

ALGORYTM

Krok 1. Obliczyć współrzędne t_i, $i = 1,2, \ldots, n$ punktów na odcinku $[0,1]$, będące wynikiem transformacji danych przez quasi-odwrotność krzywej, to znaczy, $t_i = \Psi(x_i)$, $i = 1,2, \ldots, n$.

Krok 2. Wybrać malejący ciąg $0 < \epsilon_k < 1$, $k = 1, 2, \ldots, M$ długości przedziałów, na które dzielony jest odcinek $[0, 1]$ (standardowo wybierany jest ciąg $\epsilon_k = 1/2^k$).

Krok 3. Zliczyć liczbę $N(\epsilon_k)$ pododcinków długości ϵ_k, które zawierają co najmniej po jednym punkcie ciągu t_i, $i = 1, 2, \ldots, n$. Powtórzyć te obliczenia dla $k = 1, 2, \ldots, M$.

Krok 4. Metodą najmniejszych kwadratów obliczyć współczynnik $\hat{\delta}$, będący empirycznym oszacowaniem wymiaru pudełkowego δ zbioru $\Psi(X) \subset I_1$. Dokładniej, znaleźć $\hat{\delta}$ i $\hat{c}$, której minimalizują sumę kwadratów błędów

$$\sum_{k=1}^{M} \left[\log(N(\epsilon_k)) + \delta \log(\epsilon_k) - c\right]^2, \tag{5.14}$$

względem c i δ.

Krok 5. Przeliczyć oszacowanie wymiaru na odcinku $\hat{\delta}$, uzyskane w kroku 4, na oszacowanie $\hat{D}$ wymiaru zbioru $X \subset I_d$ według wzoru: $\hat{D} = d\,\hat{\delta}$.

Łatwo zauważyć, że kroki 2, 3 i 4 są w istocie znanym algorytmem szacowania wymiaru pudełkowego dla fraktali na odcinku. Istotą propozycji jest połączenie tych kroków z krokiem 1, w którym zachodzi redukcja wymiaru oraz krokiem 5, w którym zachodzi przeliczenie oszacowania wymiaru z jednego do d-wymiarów.

5.2 Pomiar wymiaru atraktora Lorenza i inne wyniki badań symulacyjnych

5.2.1 Wymiary prostych tworów geometrycznych

W tabeli 5.1 zestawiono wyniki (uzyskane proponowaną metodą) oszacowań wymiarów fraktalnych różnych zbiorów, których wymiar teoretyczny jest znany. Porównania kolumn pozwalają stwierdzić, że proponowana metoda zapewnia dostateczną dokładność (błąd względny nie przekracza $0,06$). Do obliczeń używano 5000 punktów losowanych z mierzonego obiektu z rozkładem równomiernym. Skrót FBM w ostatnim wierszu tabeli odnosi się do procesu stochastycznego, zwanego ruchem Browna o ułamkowym wymiarze.

Tabela 5.1. Wyniki obliczeń wymiaru fraktalnego proponowaną metodą dla różnych obiektów o znanych wymiarach

Obiekt	Wymiar	
	Teor.	Empir.
Kwadrat w R^2	2	2,0
Funkcja $y = x^2$ w R^2	1	1,035
Powierzchnia $y = \lvert \sin(\pi(x_1^2 + y_1^2))\rvert$	2	2,12
FBM w R^2	1,5	1,521

5.2.2 Empiryczny wymiar atraktora Lorenza

Układ Lorenza określony jest równaniami (por. [48], gdzie opisano również interpretację zmiennych stanu tego systemu):

$$x' = \sigma(y - x), \tag{5.15}$$
$$y' = rx - yxz, \tag{5.16}$$
$$z' = xy - bz. \tag{5.17}$$

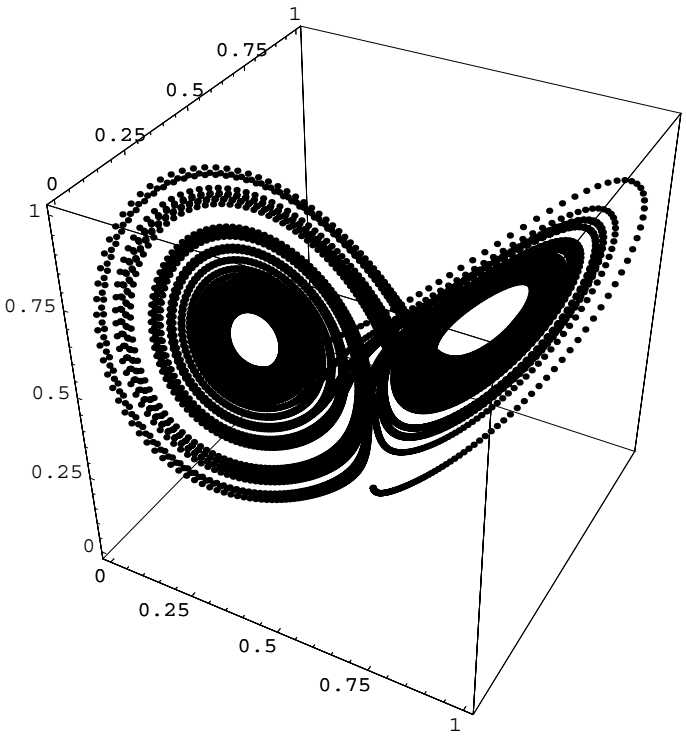
Jak wiadomo, układ ten posiada atraktor, zwany dziwnym, którego teoretyczny wymiar fraktalny nie jest znany. Na rysunku 5.2 pokazano 10 000 próbek pobranych z numerycznie obliczonego atraktora układu Lorenza.

W pracy [48] podano numeryczne oszacowanie wymiaru atraktora Lorenza wynoszące 2,06 (dla $\sigma = 10$, $b = 8/3$ i $r = 28$), nie określając jednak dokładności tego oszacowania. W celu oszacowania dokładności estymacji wymiaru atraktora przeprowadzono intensywne badania symulacyjne według następującego schematu.

Krok 1. Losowano warunki początkowe dla układu (5.15)–(5.17) z rozkładem równomiernym w kostce jednostkowej.

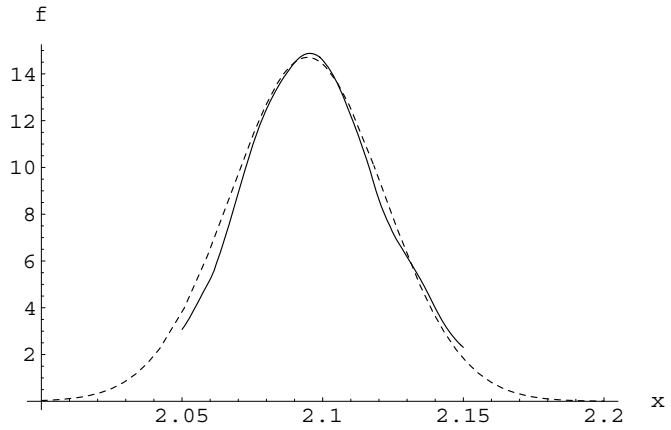
Krok 2. Metodą Runge-Kutty czwartego rzędu (z modyfikacjami zastosowanymi w środowisku *Mathematica*) rozwiązywano układ (5.15)–(5.17) i pobierano 10^4 próbek $(x(t_i), y(t_i), z(t_i))$.

Krok 3. Na ich podstawie szacowano wymiar fraktalny atraktora proponowaną tu metodą, a wynik zapamiętywano.



Rys. 5.2. Przykładowy zestaw 10 000 próbek z atraktora Lorenza używanych do oceny dokładności proponowanej metody. Inne zestawy uzyskiwane były przez losowe zmiany warunków początkowych trajektorii
Fig. 5.2. Example 10 000 samples taken from the Lorentz attractor and used for the evaluation of the accuracy of the proposed method. Other sample sets were obtained by random changes of the initial conditions of the trajectories

Kroki 1–3 powtarzano 100 krotnie i na tej podstawie estymowano gęstość rozkładu prawdopodobieństwa ocen wymiaru fraktalnego. Jako estymatora gęstości użyto nieparametrycznego estymatora Parzena–Rosenblata [35] z jądrem Epanechnikova i współczynnikiem wygładzania $h = 0,0295$ dobranym empirycznie. Na rysunku 5.3 gęstość tę zestawiono z gęstością prawdopodobieństwa rozkładu normalnego o średniej i wariancji estymowanych z tych samych danych. Uzyskana empiryczna ocena dyspersji, wynosząca $\hat{\sigma} = 0,0271306$ pozwala rekomendować proponowaną metodę do oceniania wymiarów fraktalnych wielowymiarowych układów dynamicznych.



Rys. 5.3. Oszacowanie gęstości rozkładu pomiarów wymiaru fraktalnego atraktora Lorenza (linia ciągła) i porównanie jej z gęstością rozkładu normalnego (linia przerywana) o średniej $\bar{x} = 2,09462$ i dyspersji $\hat{\sigma} = 0,0271306$. Wartości $\bar{x}$ i $\hat{\sigma}$ otrzymano w wyniku uśrednienia 100 pomiarów wymiaru fraktalnego, z których każdy uzyskano na podstawie 10^4 obserwacji przebiegu trajektorii na atraktorze
Fig. 5.3. Estimate of the probability density of the fractal dimension measurements of the Lorentz attractor (solid line) with comparison to the standard normal probability density (dashed line) with the mean $\bar{x} = 2.09462$ and the dispersion $\hat{\sigma} = 0.0271306$, estimated as the average from 100 fractal dimension measurements, each obtained from 10^4 trajectory samples

5.3 Podsumowanie

W powyższym rozdziale zbadano problem zmiany wymiaru wielowymiarowych danych po ich transformacji przy użyciu quasi-odwrotności krzywej wypełniającej. Badanym wymiarem był popularny wymiar fraktalny, nazywany wymiarem pudełkowym.

Wyprowadzono teoretyczną zależność między wymiarem pudełkowym wielowymiarowych danych a wymiarem pudełkowym danych przetransformowanych przy użyciu krzywej wypełniającej. Wyniki tych badań stały się podstawą do zaproponowania nowej, łatwiejszej obliczeniowo, a przez to dokładniejszej, metody oceny wymiaru fraktalnego obiektów wielowymiarowych. Metoda wymaga przekształcenia wielowymiarowych danych na odcinek I_1 przy użyciu quasi-odwrotności krzywej wypełniającej.

Ocena wymiaru fraktalnego jednowymiarowych danych na podstawie szacowania ich wymiaru pudełkowego jest problemem obliczeniowo znacznie prostszym, niż analogiczne zadanie dokonywane w odniesieniu do danych jednowymiarowych.

Rozdział 6

Kwantyzacja wektorowa, krzywe wypełniające i samoorganizujące sieci Kohonena

Główną ideą metod opisanych w niniejszym rozdziale jest zastosowanie algorytmu uczenia SOM oraz innych efektywnych algorytmów kwantyzacji do przetwarzania jednowymiarowych danych, powstałych w wyniku transformacji danych wielowymiarowych przy użyciu odwzorowania quasi-odwrotnego do krzywej wypełniającej [164].

Połączenie obu transformacji prowadzi do powstania nowych efektywnych algorytmów przetwarzania wielowymiarowych danych o dobrze określonych własnościach asymptotycznych [163], [164], [168].

Przeglądając literaturę dotyczącą zastosowań krzywych wypełniających oraz samoorganizujących odwzorowań Kohonena (*self-organizing maps* – SOM) [83], łatwo zauważyć, że pozycje z obu dziedzin często dotyczą podobnych problemów związanych w ogólnym sensie z problemami kwantyzacji wielowymiarowych danych [4], [52], [55], [58], [63], [100], [132], [202], [206] oraz zadaniami odtwarzania przestrzennej organizacji danych wejściowych w ograniczonej strukturze topologicznej sieci SOM (odwzorowanie topograficzne prowadzące do redukcji wymiaru, wizualizacja, [2], [96], [190], [192]).

Z punktu widzenia każdego z tych zadań rozpatrywanych odrębnie nie jest to rozwiązanie idealne, stąd wiele pomysłów łączenia algorytmów uczenia konkurencyjnego z różnymi transformacjami danych wejściowych [165], [163], [164].

Sieci SOM z topologią jednowymiarową (w postaci łańcucha neuronów) znajdują też praktyczne zastosowanie w rozwiązywaniu problemu komiwojażera [44], [50], [22], [71] oraz problemu sortowania liczb [21]. Omówienie literatury dotyczącej sieci samoorganizujących można znaleźć w pracy autorki [165].

Sieci SOM [81], [82], [83] są specjalnym typem sieci neuronowych [185], [117], [85], w których uczenie odbywa się bez nadzoru (bez nauczyciela).

W sieciach Kohonena mamy do czynienia z tak zwanym uczeniem konkurencyjnym. W przeciwieństwie do typowych sieci, na przykład sieci jednokierunkowych, w których modyfikacji podlegają wszystkie neurony, w sieciach z uczeniem konkurencyjnym po prezentacji wzorca wejściowego następuje określenie neuronu wygrywającego i tylko ten neuron, ewentualnie zbiór sąsiadujących z nim neuronów, aktualizuje swoje wagi tak, by zbliżyć je do aktualnego wzorca, przy czym obowiązujące konkretne zasady rywalizacji neuronów mogą być różne [202], [54], [83], [107].

Generalnie rzecz ujmując, reguły miękkiej konkurencji, w której równocześnie uczy się większa liczba neuronów (wygrywających neuronów), przyspieszają proces uczenia oraz zwiększają jego odporność na zakłócenia. Każdy neuron, a dokładniej jego wagi, staje się pewnym wzorcem (prototypem) grupy bliskich sobie sygnałów wejściowych, przy czym neurony sąsiadujące ze sobą reprezentują bliskie sobie, podobne, obszary przestrzeni danych wejściowych. W ten sposób sieć SOM tworzy obraz przestrzeni sygnałów wejściowych przeniesiony na wewnętrzną strukturę topologiczną sieci.

Kohonen [82] dostrzegał związki pomiędzy krzywymi wypełniającymi przestrzeń i sieciami SOM z jednowymiarową topologią w postaci łańcucha neuronów, a więc taką, którą ustala liniowy porządek w zbiorze neuronów.

W przypadku jednowymiarowej, liniowej topologii sieci neuronowej mechanizm uczenia SOM prowadzi do liniowego uporządkowania w przestrzeni wektorów wejściowych. Ponadto położenia neuronów w wielowymiarowej przestrzeni wejść koncentrują się w obszarach, z których pochodzą dane wejściowe. Jeśli założymy, że istnieje gęstość rozkładu prawdopodobieństwa wektorów wejść, to wraz ze wzrostem rozmiaru sieci SOM koncentracja neuronów staje się tym większa, im większa jest gęstość rozkładu wejść. Ciąg punktów o współrzędnych odpowiadających wagom neuronów, połączonych ze sobą zgodnie z przyjętą numeracją (topologią sieci), tworzy łamaną skanującą przestrzeń wejść.

Badanie procesu polegającego na zwiększaniu liczby neuronów prowadzi do obserwacji, że w wyniku działania algorytmu uczenia SOM otrzymujemy krzywą coraz gęściej skanującą wielowymiarową przestrzeń wejść. Proces taki nie jest jednak w jakimkolwiek sensie zbieżny do krzywej wypełniającej (traktowanej jako ciągłe odwzorowanie $I_1 \rightarrow I_d$).

Możemy przyjąć jedynie, że jednowymiarowa (w sensie topologii) sieć SOM aproksymuje pewną krzywą wypełniającą przestrzeń wejść, zachowując ponadto w pewnym stopniu związane z tą przestrzenią miary probabilistyczne. Sama procedura uczenia sieci przebiega w przypadku jednowymiarowym znacznie szybciej niż w odniesieniu do wielowymiarowych danych.

6.1 Problem wektorowej kwantyzacji

Zadanie kwantyzacji wektorowej [63], [58], [206] polega na kodowaniu wielowymiarowych danych za pomocą skończonego zbioru wektorów odniesienia (kwantyzerów) $w = (w_1, \ldots, w_n)$. Każdy punkt z przestrzeni wielowymiarowej jest jednoznacznie przyporządkowany do określonego, najbliższego w sensie pewnej metryki wektora odniesienia $w_{s(x)}$, dla którego różnica $||x - w_{s(x)}||^\beta$, $\beta > 0$, nazywana błędem kwantyzacji, jest najmniejsza. Zwykle $||\cdot||$ jest normą euklidesową, natomiast $\beta = 2$.

Jeśli rozkład prawdopodobieństwa danych wejściowych ma gęstość $f(x)$, to średni błąd kwantyzacji (wartość oczekiwana dystorsji) określony jest przez

$$1/d \int ||x - w_{s(x)}||^\beta f(x) dx \tag{6.1}$$

i zależy od wyboru wektorów odniesienia w.

Zaprojektowanie dobrego zbioru wektorów odniesienia jest podstawowym problemem wektorowej kwantyzacji i wymaga rozwiązania zagadnienia wielowymiarowej minimalizacji. Problem ten cechuje istnienie wielu minimów lokalnych [202].

Znanych jest wiele algorytmów wektorowej kwantyzacji: [206],[202], [107], [204]. Najczęściej używanym jest uogólniony algorytm Lloyda, znany też jako schemat LBG [100], [132], [202]. Tego typu algorytm był rozpatrywany również w zastosowaniu do klasteryzacji wielowymiarowych danych w postaci algorytmu K–średnich [105], [41], [63], [17], [27].

Podobnie algorytm uczenia SOM (z topologią jednowymiarową w postaci łańcucha neuronów) [81], [82], [83] może być traktowany jako uogólnienie metody LBG w wersji, w której metodę aproksymacji stochastycznej [136], [12], [102] stosuje się do optymalizacji dokonywanej adaptacyjnie, na podstawie pojedynczych obserwacji (por. [132], [165]).

Algorytm ten ma postać [83], [132]:

$$w_i^{\text{nowy}} = w_i^{\text{stary}} + h_{i,s(x)} (x - w_i^{\text{stary}}), \tag{6.2}$$

gdzie $h_{i,s(x)}$ jest tak zwaną funkcją sąsiedztwa, która przyjmuje wartość 1 dla $i = s(x)$, w pozostałych przypadkach natomiast jej wartość jest nierosnącą funkcją $|i - s(x)|$. Najczęściej $h_{i,s(x)} = 1$, gdy $|i - s(x)| \leq m$ oraz $h_{i,s(x)} = 0$, gdy $|i - s(x)| > m$. Taką funkcję sąsiedztwa nazywa się sąsiedztwem prostokątnym o szerokości m. Analiza algorytmu uczenia (6.2) jest w ogólnym przypadku bardzo skomplikowana (por. przegląd znanych wyników badań zawarty w pracy autorki [165] oraz w pracach [203], [99], [56], [98], [188], [83], [187], [133], [134], [135]). Niestety, zbieżność z prawdopodobieństwem 1 algorytmu uczenia SOM do ustalonego stanu równowagi oraz istnienie i jednoznaczność takiego stanu równowagi

(istnienie i jednoznaczność rozwiązania asymptotycznego) wykazano tylko przy założeniu równomiernego rozkładu danych wejściowych [83] lub przy założeniu, że $f(x) > 0, x \in [0,1]$ oraz że logarytm gęstości rozkładu jest funkcją wklęsłą na odcinku $[0,1]$ [18], [11]. Równocześnie jednak nie jest znany żaden przykład świadczący o niejednoznaczności rozwiązania asymptotycznego algorytmu (6.2) przy $f(x) > 0$ oraz przy dowolnej (ze względu na liczbę sąsiadów) prostokątnej funkcji sąsiedztwa [18]. Ponadto wiadomo, że algorytm (6.2) w przypadku jednowymiarowym zapewnia osiągnięcie i utrzymanie stanu uporządkowania neuronów polegającego na spełnieniu warunku $w_1 \leq \ldots \leq w_i \leq w_{i+1} \ldots \leq w_N$ [83].

Dalej będziemy korzystać tylko z wyników dotyczących przypadku jednowymiarowej sieci SOM i jednowymiarowych danych pochodzących z odcinka I_1. Z pracy [132] wynika, że w sieci SOM, w sytuacji gdy liczba neuronów dąży do nieskończoności, funkcja gęstości rozkładu prawdopodobieństwa położenia neuronu, $Q(x)$, rozpatrywana w przestrzeni danych wejściowych (jednowymiarowych), jest proporcjonalna do funkcji gęstości rozkładu wejść $f(x)$ w potędze a, gdzie

$$a = 2/3 - 1/3[m^2 + (m+1)^2], \tag{6.3}$$

natomiast m oznacza liczbę sąsiadów w jednowymiarowym algorytmie Kohonena z prostokątną funkcją sąsiedztwa, stąd

$$Q(x) \propto f(x)^a. \tag{6.4}$$

Formalny dowód powyższej własności wymaga założenia, że $f(x) > 0$. Ponadto w trakcie przeprowadzania dowodu w pracy [132] założono, że funkcja gęstości jest różniczkowalna. Założenie to jest niepotrzebnie zbyt restrykcyjne. W rzeczywistości wystarczy przyjąć, że funkcja gęstości daje się aproksymować z dowolną dokładnością za pomocą funkcji odcinkami liniowej. Warunek ten spełnia każda, ograniczona od góry, funkcja gęstości.

Należy zauważyć, że zależność (6.4) udowodniono w zasadzie tylko w przypadku jednowymiarowej przestrzeni wejść i przy strukturze sieci w postaci łańcucha neuronów. Nic natomiast nie wiadomo o istnieniu tego typu asymptotycznych rezultatów (poza pewnymi bardzo szczególnymi przypadkami [133]), gdy dane wejściowe są wielowymiarowe).

Przy liczbie sąsiadów $m \to \infty$, równanie (6.3) prowadzi do otrzymania wartości $a = 2/3$. W przypadku, gdy $m = 0$, ze wzoru (6.3) otrzymujemy $a = 1/3$. Wynik ten pokrywa się ze znanymi wynikami dotyczącymi wektorowej kwantyzacji. Jak wiadomo [206], jeśli liczba kwantyzatorów dąży do nieskończoności, to gęstość ich rozkładu w d–wymiarowej przestrzeni wejść, minimalizująca kryterium (6.1), jest proporcjonalna do $f(x)^a$, gdzie $a = d/(d+\beta)$. Łatwo sprawdzić, że gdy $d = 1$ i wartość wykładnika w kryterium (6.1) wynosi $\beta = 2$, otrzymujemy tę samą wartość $a = 1/3$.

Oprócz algorytmu uczenia (6.2) stosowany bywa także algorytm uczenia nazywany ,,gazem elektronowym" [107], [117]. W algorytmie tym nie ma ustalonej topologii sieci (stałego uporządkowania neuronów). Zbiór sąsiadów wyznaczany jest na bieżąco, jako aktualny zbiór k najbliższych sąsiadów. Wszystkie neurony są porządkowane zgodnie z ich odległością od prezentowanego wzorca, czyli w postaci: $\{w_{i_1}, w_{i_2}, \ldots, w_{i_N}\}$, przy czym

$$||w_{i_k} - x|| \leq ||w_{i_{k+1}} - x||,\ k = 1, 2, \ldots, N-1.$$

Reguła uczenia ,,gazu elektronowego" ma postać:

$$w_i^{\text{nowy}} = w_i^{\text{stary}} + h_\lambda(k_i(x, w))\,(x - w_i^{\text{stary}}),\ i = 1, 2, \ldots, N, \tag{6.5}$$

gdzie $k_i(x, w)$ jest liczbą naturalną, która oznacza, którym z kolei najbliższym sąsiadem obserwacji x jest neuron w_i. Funkcja

$$h_\lambda(k_i(x, w)) = e^{-(k_i(x,w)-1)/\lambda}$$

jest specyficzną funkcją sąsiedztwa z parametrem $\lambda \geq 0$. Przy $\lambda \to 0$, tylko położenie najbliższego sąsiada podlega aktualizacji. Gdy $\lambda > 0$, wtedy aktualizowane są wagi wszystkich neuronów, lecz współczynnik uczenia odległych neuronów (dalekich sąsiadów) szybko dąży do zera. Jest to sytuacja podobna do tej, z którą mamy do czynienia w klasycznym algorytmie uczenia SOM z gaussowską funkcją sąsiedztwa [83].

W przypadku asymptotycznym, gdy liczba neuronów N dąży do nieskończoności, gęstość rozkładu neuronów w przestrzeni wejść opisywana jest równaniem różniczkowym o postaci takiej, jak równanie charakteryzujące dyfuzję gazu – stąd nazwa sieci ,,gaz neuronowy". Stacjonarne rozwiązanie tego równania przy $\lambda \to \infty$ jest postaci:

$$Q(x) \propto f(x)^a. \tag{6.6}$$

Wykładnik $a = d/(d+2)$ ma wartość optymalną ze względu na zadanie wektorowej kwantyzacji z kwadratową dystorsją ($\beta = 2$) [206]. Algorytm ,,gazu elektronowego" jest szczególnie prosty obliczeniowo w przypadku jednowymiarowym, co pozwala na jego efektywne współdziałanie z transformacją wielowymiarowych danych za pomocą krzywych wypełniających.

Wracając do spojrzenia na algorytm Kohonena jako na narzędzie służące do wektorowej kwantyzacji, możemy w prostych przypadkach sformułować asymptotyczną postać kryterium, zależną od przyjętej szerokości sąsiedztwa. Zmiana szerokości sąsiedztwa m prowadzi do uzyskania różnych wartości wykładnika a (por. (6.3)). Także inne modyfikacje algorytmu Kohonena [132], [204], [10] pozwalają wpływać na wartość wykładnika a w zależności (6.4).

Korzystając ze wspomnianych już wyników Zadora [206], dotyczących wektorowej kwantyzacji, otrzymujemy, poprzez zmianę a, różne postaci kryterium (6.1), w których wykładnik β zmienia się w zależności od sposobu modyfikacji algorytmu Kohonena. Postać kryterium odnosi się naturalnie tylko do przypadku jednowymiarowego i sytuacji asymptotycznej, gdy liczba neuronów (kwantyzatorów) dąży do nieskończoności. Między innymi nieliniowa metoda aktualizacji wag, zaproponowana w [204], pozwala na uzyskanie (w jednym wymiarze) różnych wartości wykładnika a. Do podobnych efektów prowadzi także adaptacyjna zmiana współczynnika uczenia określanego indywidualnie dla każdego z neuronów [10].

W kolejnym rozdziale pracy zaproponowana zostanie metoda kwantyzacji, która łączy transformację wielowymiarowych danych za pomocą quasi-odwrotności wybranej krzywej wypełniającej ze skalarną kwantyzacją przy użyciu algorytmu (6.2) w odniesieniu do danych przetransformowanych na odcinek I_1.

Zastosowanie różnych wariantów algorytmu uczenia (6.2) umożliwia w pewnym stopniu kształtowanie gęstości rozkładu kwantyzatorów na odcinku I_1, a także – jeśli skorzystamy z własności zachowywania miary przez odwzorowania F i jego quasi-odwrotność Ψ – odpowiednich kwantyzatorów w przestrzeni I_d.

6.2 Kwantyzacja wektorowa na odcinku

W pracy autorki [164] zaproponowano nową metodę kwantyzacji wielowymiarowych danych, która polega na połączeniu transformacji wielowymiarowych danych za pomocą krzywych wypełniających ze skalarną kwantyzacją wykorzystującą jednowymiarowy algorytm SOM na danych z odcinka. Zamiast SOM można w tym miejscu użyć innego klasycznego algorytmu kwantyzacji typu LBG [55] lub algorytmu ,,gazu elektronowego". Eksperymenty obliczeniowe pokazują jednak, że algorytm SOM jest w wielu przypadkach bardziej efektywny i prowadzi do mniejszego błędu kwantyzacji. Końcowym etapem procesu jest wyznaczenie wektorów odniesienia poprzez przekształcenie skalarnych kwantyzatorów przy użyciu krzywej wypełniającej do postaci wektorowej w I_d.

Podobną ideę zastosowania krzywej Hilberta do kompresji i kwantyzacji obrazów można znaleźć w pracach [97], [126]. W pracach tych nie skorzystano jednak z własności krzywych wypełniających, a w szczególności własności zachowywania miary.

W przedstawianym tu podejściu, zmieniając algorytm skalarnej kwantyzacji, mamy możliwość wyboru pożądanej zależności pomiędzy rozkładami wejść a rozkładami kwantyzatorów w przestrzeni wielowymiarowej.

6.2.1 Algorytm wektorowej kwantyzacji uzyskany za pomocą krzywych wypełniających

Niech $\{X_i\}$ będzie nieskończonym ciągiem realizacji niezależnych, wielowymiarowych zmiennych losowych (wektorów losowych). Zakładamy, że X_i są identycznymi zmiennymi losowymi o rozkładzie opisanym gęstością f, które przyjmują wartości z kostki I_d. Przyjmujemy też, że współrzędne punktów z kostki są liczbami rzeczywistymi, a ich pozycje na krzywej wypełniającej kostkę, czyli $\Psi(X)$, mogą być wyznaczone z dowolną dokładnością.

Ponieważ Ψ jest funkcją mierzalną, stąd także $t_i = \Psi(X_i)$ są niezależnymi zmiennymi losowymi. Ich rozkład opisuje funkcja gęstości $g(t) = f(F(t))$ (por. lemat 4.16).

Zmienne losowe t_i przyjmują wartości z odcinka $[0,1]$. Kwantyzacji danych o nieznanym rozkładzie $g(t)$ dokonujemy korzystając ze zmodyfikowanego algorytmu (6.2) (por. praca autorki [164]).

W celu uniknięcia nieporozumień będziemy dalej oznaczać literami v pozycje kwantyzatorów na odcinku I_1, natomiast literami w kwantyzatory w przestrzeni wielowymiarowej.

Podobieństwo między sygnałem wejściowym $t = \Psi(x) \in [0,1]$ a skalarnym kwantyzatorem v_i jest mierzona poprzez ich odległość na odcinku $[0,1]$. Wygrywający współzawodnictwo neuron v_c wyznaczany jest zatem przez formułę:

$$|t - v_{s(t)}| = \min_{1 \le i \le N} \{|t - v_i|\}. \tag{6.7}$$

Obszary oddziaływania poszczególnych kwantyzatorów

$$O_c = \{t : |t - v_{c)}| = \min_{1 \le i \le N} \{|t - v_i|\}\},\ c = 1, \ldots, N \tag{6.8}$$

tworzą podział odcinka I_1, zwany w przypadku wielowymiarowym diagramem Voronoia [58]. Zauważmy, że – w przeciwieństwie do problemu wielowymiarowego – podział ten jest bardzo łatwo wyznaczyć w stanie uporządkowania sieci, gdy $v_i \le v_{i+1}$, a mianowicie $O_i = [(v_i - v_{i-1})/2, (v_{i+1} - v_i)/2]$ dla $i = 2, \ldots, N-1$ oraz $O_1 = [0, (v_2 - v_1)/2]$, $O_N = [(v_N - v_{N-1})/2, 1]$.

A oto zaproponowany algorytm, nazywany dalej algorytmem SFCVQ.

ALGORYTM SFCVQ

Przyjmij początkową wartość współczynnika uczenia η oraz wartość $\kappa > 0$.

Krok 1. Rozpocznij od losowo wybranych pozycji $v_1, v_2, \ldots, v_N$, $v_i \in [0,1]$, $i = 1, \ldots, N$.

Krok 2. Przetransformuj nową obserwację $x \in I_d$ do postaci $t = \Psi(x) \in I_1$.

Krok 3. Oblicz odległość t w stosunku do aktualnych wartości $v_1, v_2, \ldots, v_N$:

$$|t - v_i|, \quad i = 1, \ldots, N$$

i znajdź numer wygrywającego neuronu $s = s(t)$, czyli oblicz $\arg\min_i |t - v_i|$.

Krok 4. Zmień wagę v_s oraz wagi sąsiadujących $2\,m$ neuronów zgodnie z regułą:

$$v_j^{\text{nowe}} = v_j^{\text{stare}} + \eta\, h_{sj}\, \text{sgn}(t - v_j^{\text{stare}}) \cdot |t - t_j|^{\kappa} \; j = 1, \ldots, N,$$

gdzie $h_{sj} = 1$, jeśli $|s - j| \leq m$ oraz $h_{sj} = 0$ w przeciwnym przypadku.

Krok 5. Zmień η i ewentualnie szerokość sąsiedztwa m. Idź do kroku 2.

Sąsiedztwo neuronów m–tego rzędu obejmuje, poza neuronem wygrywającym o numerze c, także m jego poprzedników i m następników:

$$N_c = \{i : |i - c| \leq m,\ 1 \leq i \leq N\}. \tag{6.9}$$

W powyższym algorytmie uwzględniono w uogólnionej postaci metodę uczenia zaproponowaną w pracy [204]. W przypadku, gdy $\kappa = 1$, algorytm ten jest równoważny ze znaną regułą uczenia (6.2), jeśli natomiast $\kappa > 0$ jest dowolną liczbą, to przy $m = N$, $N \to \infty$

$$Q(x) \propto f(x)^{\frac{2}{2+\kappa}}. \tag{6.10}$$

Umożliwia to zmianę wykładnika a w (6.4) w zakresie od wartości bliskich 0 do wartości bliskich 1. Podobnie, ustalanie różnych szerokości sąsiedztwa m, które jest dalej stałe w trakcie całego procesu uczenia, przy $\kappa = 1$, pozwala na uzyskanie wartości a z przedziału $[1/3, 2/3]$ (por.[132]).

Podsumowując, jeśli $N \to \infty$, to dystorsja zapewniana przez algorytm SFCVQ dąży do minimalnej dystorsji postaci:

$$1/d \int |t - v_{s(t)}|^{\beta} g(t) dt,$$

gdzie $\beta = a^{-1} - 1$ zależy od parametrów algorytmu SFCVQ. Uzasadnienie powyższego stwierdzenia wynika bezpośrednio ze znanej zależności [206] pomiędzy wartościami wykładnika β w wartości oczekiwanej dystorsji (6.1), wykładnika a w optymalnej, dla tego kryterium, gęstości kwantyzatorów oraz wymiarem przestrzeni d. Zależność ma postać $a = d/(\beta + d)$. Po przekształceniu i podstawieniu $d = 1$ otrzymujemy wartość $\beta = a^{-1} - 1$.

Zauważmy, że skoro F jest funkcją wzajemnie jednoznaczną z dokładnością do zbioru miary Lebesgue'a zero (por. własność **C2** z rozdziału 4), to zachodzi następująca własność:

Lemat 6.1 *Niech F będzie krzywą wypełniającą I_d, która spełnia warunki* **C1–C3**. *Jeśli zmienna losowa Y przyjmuje wartości z odcinka I_1, przy czym jej rozkład opisuje funkcja gęstości $g(t)$, $t \in I_1$, to $F(Y)$ jest zmienną losową, która przyjmuje wartości w I_d, a jej gęstość jest równa $f(x) = g(\Psi(x)), x \in I_d$ prawie wszędzie (z dokładnością do zbioru miary Lebesgue'a zero).*

Na podstawie powyższego lematu możemy dalej wnioskować o asymptotycznym, przy $N \to \infty$, rozkładzie $Q_d(x)$ kwantyzatorów otrzymanych poprzez przetransformowanie do kostki I_d zbioru kwantyzatorów $\{v_1, v_2 \ldots\}$, wyznaczonych w wyniku działania algorytmu SFCVQ na podstawie danych o rozkładzie $g(t) = f(F(t))$, gdzie $f(x) > 0$ jest funkcją gęstości rozkładu oryginalnych danych pochodzących z kostki I_d. Jeśli kwantyzatory na odcinku mają asymptotyczną gęstość rozkładu $\propto f(F(t))^a$, to $Q_d(x) \propto f(F(\Psi(x)))^a = f(x)^a$ prawie wszędzie w I_d (z dokładnością do zbioru miary Lebesgue'a zero), por.[164].

6.2.2 Asymptotyczny błąd kwantyzacji

W podrozdziale tym oszacujemy błąd dystorsji (6.1) uzyskiwany w przestrzeni wielowymiarowej, gdy zbiór kwantyzatorów zostanie ustalony poprzez przetransformowanie, za pomocą krzywej wypełniającej F, zbioru skalarnych kwantyzatorów $\{v_1, v_2 \ldots\}$ obliczonych dla danych z odcinka I_1. Funkcja gęstości rozkładu prawdopodobieństwa danych jednowymiarowych jest postaci $g(t) = f(F(t))$, gdzie $f(x) > 0$ jest funkcją gęstości rozkładu oryginalnych danych pochodzących z kostki I_d.

W ogólnym przypadku błąd dystorsji (6.1) w przestrzeni wielowymiarowej ma wartość

$$E_\beta(N, F(V)) = \frac{1}{d} \sum_{i=1}^{N} \int_{S_i} ||F(t) - F(v_i)||^\beta g(t) dt, \tag{6.11}$$

gdzie $S_i = [s_i, s_{i+1}] = [(v_{i-1} + v_i)/2, (v_i + v_{i+1})/2]$, $s_1 = 0$, $s_{N+1} = 1$. Zauważmy, że wartość $E_\beta(N, F(V))$ można przedstawić w równoważnej postaci jako

$$E_\beta(N, F(V)) = \frac{1}{d} \sum_{i=1}^{N} \int_{C_i} ||x - F(v_i)||^\beta f(x) dx, \tag{6.12}$$

gdzie $C_i = F(S_i), i = 1, \ldots, N$.

Dalej, wartość $E_\beta(N, F(V))$ można oszacować od góry, korzystając z własności **C1** krzywej wypełniającej (patrz rozdział 4.5), w następujący sposób:

$$\frac{\alpha^\beta}{d} \sum_{i=1}^{N} \int_{s_i}^{s_{i+1}} |t - v_i|^{\beta/d} g(t) dt, \tag{6.13}$$

gdzie α jest stałą w odpowiednim warunku Höldera (4.74), której wartość zależy od typu krzywej wypełniającej i wymiaru przestrzeni d.

Jeśli liczba poziomów kwantyzacji N jest bardzo duża, a w konsekwencji szerokość odpowiadających im na odcinku przedziałów bardzo mała, to możemy przyjąć, że $g(t)$ jest w przybliżeniu stałe w przedziałach $[s_i, s_{i+1}]$. Możemy zatem przyjąć, że lokalne zachowanie kwantyzatorów w każdym z tych przedziałów jest bliskie optymalnemu (równomiernemu) zachowaniu kwantyzatorów przy założeniu jednostajnego rozkładu sygnałów wejściowych (por. [58]).

Stąd otrzymujemy dalej, że (6.13) jest w przybliżeniu równe

$$\frac{\alpha^\beta}{d} \sum_{i=1}^{N} g(v_i) \int_{s_i}^{s_{i+1}} |t - v_i|^{\beta/d} dt. \tag{6.14}$$

Przy równomiernym rozkładzie sygnałów wejściowych kwantyzator v_i powinien znajdować się dokładnie w połowie przedziału $[s_i, s_{i+1}]$. W związku z tym, wykonując standardowe obliczenia, otrzymujemy

$$\int_{s_i}^{s_{i+1}} |t - v_i|^{\beta/d} dt = 2 \int_0^{\Delta_i/2} t^{\beta/d} dt = \frac{2d}{\beta + d} \left(\frac{\Delta_i}{2}\right)^{\frac{\beta+d}{d}}, \tag{6.15}$$

gdzie $\Delta_i = s_{i+1} - s_i$. Po podstawieniu do (6.14) otrzymujemy

$$\frac{\alpha^\beta}{\beta + d} \sum_{i=1}^{N} g(v_i) \, (\Delta_i/2)^{\beta/d} \, \Delta_i. \tag{6.16}$$

Niech $q(t)$ będzie asymptotyczną (przy $N \to \infty$) gęstością rozkładu kwantyzatorów, wtedy $\Delta_i \simeq \frac{1}{N\, q(t)}$. Przyjmijmy, że $q(t) = C^{-1}\, g(t)^a$, gdzie $C = \int_0^1 g(t)^a$. Wartość a zależy od użytej metody kwantyzacji na odcinku I_1, stąd otrzymujemy

$$E_\beta(N, F(V)) \leq \frac{\alpha^\beta}{d + \beta} \, (2N/C)^{-\beta/d} \sum_{i=1}^{N} g(v_i)^{1 - a\beta/d} \, \Delta_i. \tag{6.17}$$

Przy $N \to \infty$ zachodzi

$$\sum_{i=1}^{N} g(v_i)^{1 - \frac{a\,\beta}{d}} \, \Delta_i \simeq \int_0^1 g(t)^{1 - \frac{a\,\beta}{d}} dt.$$

Załóżmy, że $g(t)$ spełnia warunki:

$$g(t) > 0, \; t \in I_1, \tag{6.18}$$

$$C = \int_0^1 g(t)^a dt < \infty, \tag{6.19}$$

$$D = \int_0^1 g(t)^{\frac{d - a\,\beta}{d}} dt < \infty. \tag{6.20}$$

Twierdzenie 6.2.1 *Jeżeli $g(t)$, gęstość rozkładu prawdopodobieństwa sygnałów wejściowych rozpatrywanych na odcinku I_1, spełnia warunki (6.18)–(6.20), to wtedy asymptotyczny błąd kwantyzacji $E_\beta(N, F(V))$ jest ograniczony z góry przez wyrażenie*

$$C(\beta, d, \alpha, a)\, N^{-\beta/d}, \tag{6.21}$$

gdzie $C(\beta, d, \alpha, a) = \frac{\alpha^\beta}{d+\beta} \left(\frac{C}{2}\right)^{\beta/d} D$.

Z powyższego twierdzenia wynika, że asymptotyczna wartość dystorsji (6.11) maleje wraz z N w sposób typowy dla wielowymiarowych kwantyzatorów, czyli jak $O(N^{-\beta/d})$ (por.[64], [206]).

6.3 Krzywe odtwarzające rozkład danych wejściowych

Za pracą autorki [163] zaproponujemy konstrukcję krzywej wypełniającej, która zachowując geometrię klasycznych, omawianych w niniejszej monografii krzywych wypełniających, równocześnie dokładnie odtwarza rozkład danych wejściowych, tak jak to postulował Kohonen [82] w odniesieniu do samoorganizujących odwzorowań SOM.

Innymi słowy, celem przedstawionego w tym rozdziale algorytmu jest otrzymanie krzywej, która wypełnia podobszary o równej objętości fragmentami krzywej, której odpowiednie długości na odcinku są proporcjonalne do miary probabilistycznej sygnałów wejściowych odpowiadających tym podobszarom.

Mówiąc bardziej precyzyjnie, proponowana krzywa wypełniająca powinna spełniać następujący warunek:

C Krzywa $F_P : I_1 \to I_d$ zachowuje miarę probabilistyczną P, to znaczy, dla każdego zbioru borelowskiego $B \subseteq I_d$ zachodzi

$$P(B) = \mu\left(F_P^{-1}(B)\right), \tag{6.22}$$

gdzie μ oznacza miarę Lebesgue'a na odcinku I_1.

Zakładamy, że dane wejściowe przyjmują wartości w kostce I_d. Ich zachowanie opisuje zmienna losowa X o rozkładzie absolutnie ciągłym (względem miary Lebesgue'a) w I_d i danym za pomocą gęstości $f(x)$, $x \in I_d$.

Jak wiadomo, jeśli U jest zmienną losową o rozkładzie równomiernym na odcinku I_1, to zmienna losowa $\mathcal{W}^{-1}(U)$ ma rozkład o dystrybuancie $\mathcal{W}$. $\mathcal{W}$ musi być dystrybuantą, niemalejącą funkcją, która zmienia się od wartości 0 do wartości 1. Dla naszych celów dodatkowo zakładamy, że $\mathcal{W}$ jest określona na przedziale I_1,

w tym sensie, że $\mathcal{W}(t) = 0,\ t \leq 0$ oraz $\mathcal{W}(t) = 1,\ t \geq 1$. Jeśli dodatkowo założymy, że $f(x) > 0,\ x \in I_d$, to w konsekwencji $\mathcal{W}$ jest ściśle monotoniczna. Stąd dla rozkładów oddzielonych od zera (w kostce I_d) $\mathcal{W}$ jest odwzorowaniem wzajemnie jednoznacznym odcinka I_1 w I_1 i jego odwrotność $\mathcal{W}^{-1}$ jest także funkcją ciągłą i ściśle monotoniczną.

Twierdzenie 6.3.1 *Niech funkcja $F_P : I_1 \to I_d$ będzie zdefiniowana w następujący sposób:*

$$F_P(t) = F(\mathcal{W}^{-1}(t)), \quad t \in I_1,$$

gdzie F jest d–wymiarową krzywą wypełniającą zachowującą miarę Lebesgue'a, a $\mathcal{W}(t) = \int_0^t f(F(s))ds$, gdzie $f(x)$, $x \in I_d$ jest funkcją gęstości rozkładu prawdopodobieństwa w I_d, której nośnikiem jest cała kostka I_d.

Wtedy F_P jest ciągłym odwzorowaniem odcinka I_1 na kostkę I_d, które zachowuje miarę probabilistyczną określoną przez gęstość rozkładu prawdopodobieństwa $f(x) > 0,\ x \in I_d$, to znaczy, dla każdego zbioru borelowskiego $B \subseteq I_d$ zachodzi

$$P(B) = \mu\left(F_P^{-1}(B)\right). \tag{6.23}$$

Dowód. Z własności krzywej wypełniającej F (patrz lemat 4.15) wynika, że

$$P(B) = \int_B f(x)dx = \int_{F^{-1}(B)} f(F(s))ds.$$

Korzystając z zamiany zmiennych $s = \mathcal{W}^{-1}(t)$, możemy zapisać

$$ds = d\mathcal{W}^{-1}(t) = f[F(\mathcal{W}^{-1}(t))]^{-1}dt$$

i dalej

$$\int_{F^{-1}(B)} f(F(s))ds = \int_{\mathcal{W}(F^{-1}(B))} f[F(\mathcal{W}^{-1}(t))]\, f[F(\mathcal{W}^{-1}(t))]^{-1}dt$$

$$= \int_{\mathcal{W}(F^{-1}(B))} dt = \mu[\mathcal{W}(F_P^{-1}(B))].$$

Zauważmy, że

$$F_P^{-1}(x) = \mathcal{W}(F^{-1}(x)),\ x \in I_d,$$

co kończy dowód własności **C**. Ponieważ $f(x) > 0$, stąd także $f(F(t)) > 0$, a w konsekwencji $\mathcal{W}(t)$ jest funkcją ciągłą, ściśle rosnącą w I_1. Złożenie ciągłej bijekcji $\mathcal{W}^{-1} : I_1 \to I_1$ oraz ciągłej surjekcji $F : I_1 \to I_d$ daje w wyniku ciągłe i surjektywne odwzorowanie odcinka I_1 na kostkę I_d, co kończy dowód. □

Powyższe twierdzenie jest podstawą proponowanego dalej algorytmu generowania krzywych wypełniających zależnych od danych.

Zakładamy dalej, że dany jest ciąg niezależnych zmiennych losowych X_1, X_2, ... o tym samym rozkładzie oraz istnieje funkcja gęstości tego rozkładu $f(x), x \in I_d$. Z lematu 4.16 wynika, że zmienne losowe $t_i = \Psi(X_i)$ są niezależnymi zmiennymi losowymi przyjmującymi wartości z odcinka I_1 i z funkcją gęstości prawdopodobieństwa $g(t) = f(F(t))$. Niech $\mathcal{W}$ oznacza jak poprzednio ciągłą dystrybuantę rozkładu prawdopodobieństwa na odcinku I_1, odpowiadającą gęstości $g(t)$.

Algorytm tworzenia aproksymacji krzywej F_P, bazujący tylko na obserwacjach $X_1, X_2, \ldots$, będzie składał się z dwóch części. Pierwsza polega na przekształceniu wielowymiarowych danych pochodzących z kostki I_d do postaci jednowymiarowych danych zawartych w odcinku I_1 (odwzorowanie Ψ). Druga część algorytmu, wiążąca się z największym nakładem obliczeniowym, polega na estymowaniu dystrybuanty rozkładu danych jednowymiarowych $\mathcal{W}$.

Etapem pośrednim będzie estymacja gęstości rozkładu za pomocą histogramu. Szerokości przedziałów w histogramie będą zmieniane hierarchicznie w sposób dopasowany do geometrycznej struktury kolejnych przybliżeń krzywej wypełniającej F.

ALGORYTM część I

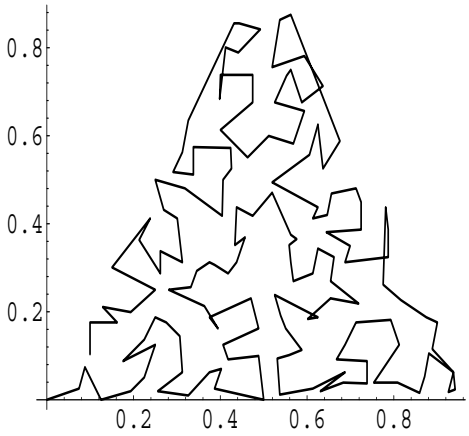
Przyjmij $g_{0j} = 0$, $j = 1, \ldots, m$. Oblicz:

$$g_{ij} = g_{i-1\,j}\,(i-1)/i + i^{-1}\,h_{ij}, \quad j = 1, \ldots, m,$$

gdzie $h_{ij} = 1$, jeśli $t_i = \Psi(X_i) \in [(j-1)/m, j/m]$ oraz $h_{ij} = 0$ w przeciwnym przypadku; $m = 2^{d\,k}$, $k = 1, 2, 3, \ldots$, $i = 1, 2, \ldots, n$.

Powyższy algorytm jest w istocie rzeczy wersją rekurencyjną histogramowego estymatora funkcji gęstości $g(t)$ [35], [37], naturalnie z dokładnością do współczynnika skalującego $m = 2^{dk}$. Zmieniające się liczby $k = 1, 2, \ldots$ określają pośrednio szerokość przedziałów histogramu, a w konsekwencji dokładność aproksymacji funkcji gęstości $g(t)$. Szerokość przedziałów histogramu wynosi odpowiednio 2^{-dk}, gdzie d, jak dotychczas, oznacza wymiar przestrzeni I_d, przy czym wartość k powinna być dopasowana do ilości dostępnych obserwacji. Proces zwiększania liczby k można kontynuować lokalnie, tylko w odniesieniu do wybranych przedziałów $[(j-1)/m, j/m]$.

Zamiast dystrybuanty $\mathcal{W}$, która nie jest znana, użyjemy jej estymaty uzyskanej poprzez obliczenie całki (zsumowanie) z histogramu.



Rys. 6.1. Zależna od danych krzywa Sierpińskiego ze 150 punktami wierzchołkowymi
Fig. 6.1. Data-driven Sierpiński space-filling curve with 150 nodal points

ALGORYTM część II

$$\hat{\mathcal{W}}_n(t) = \sum_{l=1}^{j(t)-1} g_{nl} + g_{nj(t)}(mt - j(t) + 1), \quad \text{gdzie} \quad j(t) = \lceil t\, m \rceil,$$

$$m = 2^{d\,k}, \quad k = 1, 2, 3, \ldots$$

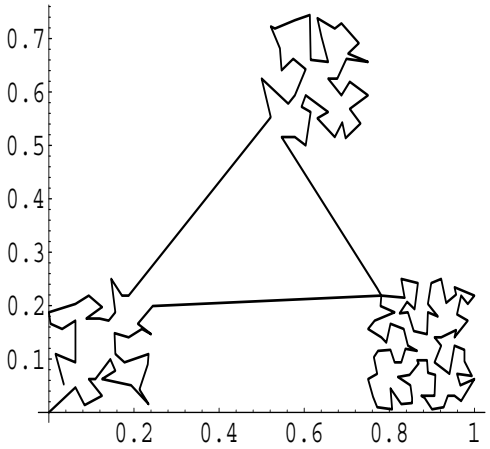
$$\hat{F}_P(t) = F(\hat{\mathcal{W}}_n^{-1}(t)),\ t \in I_1,$$

a krzywa F jest tą samą krzywą wypełniającą, której quasi-odwrotność $\boldsymbol{\Psi}$ była używana przy wyznaczaniu histogramu g_{nj}, $j = 1, \ldots, m$.

Otrzymana w powyższy sposób funkcja $\hat{\mathcal{W}}(t)$ jest funkcją ciągłą, odcinkami liniową. W związku z tym wyznaczenie funkcji odwrotnej do $\hat{\mathcal{W}}(t)$ nie stwarza większych problemów.

6.3.1 Przykłady obliczeniowe

Rysunek 6.1 pokazuje zależną od danych krzywą o geometrii krzywej Sierpińskiego, aproksymowaną za pomocą 150 punktów rozmieszczonych równomiernie na odcinku I_1. Oryginalne dane stanowiło 1000 punktów o rozkładzie równomiernym na trójkącie. Analogiczną krzywą skonstruowaną dla danych pochodzących z trzech kwadratów pokazano na rysunku 6.2, przy czym gęstość rozkładu była



Rys. 6.2. Zależna od danych krzywa typu Sierpińskiego ze 160 punktami wierzchołkowymi
Fig. 6.2. Data-driven Sierpiński space-filling curve with 160 nodal points

równomierna. W jednym z kwadratów (prawy dolny róg) gęstość rozkładu jest dwa razy większa niż w dwu pozostałych kwadratach.

Podobnie jak poprzednio, długość ciągu uczącego wynosiła 1000. Krzywą aproksymowano za pomocą 160 punktów wierzchołkowych.

Należy zwrócić uwagę na to, że pokazane tu przykłady nie spełniają warunku $f(x) > 0, x \in I_d$. W związku z tym dystrybuanta odpowiedniego rozkładu $\mathcal{W}$ nie jest funkcją ściśle rosnącą. W obszarach, gdzie mamy $g(t) = 0$, $\mathcal{W}(t)$ jest funkcją stałą. W konsekwencji nie można jednoznacznie określić $\mathcal{W}^{-1}$.

Nie stanowi to jednak w praktyce większego problemu, wystarczy wybrać jedną z granicznych wartości przedziału stałości $\mathcal{W}$ jako odpowiednią wartość funkcji odwrotnej (dokładniej quasi-odwrotnej) do $\mathcal{W}$. Otrzymane w takiej sytuacji odwzorowanie F_P nie jest jednak funkcją ciągłą, a więc formalnie nie jest krzywą wypełniającą, a jedynie sumą krzywych wypełniających pewne podzbiory I_d.

Jest to szczególnie widoczne w przykładzie drugim (dane z trzech kwadratów). Idealna krzywa F_P nie jest w tym przypadku ciągła w punktach $t = 1/8, 4/8, 6/8$.

Przedstawiony algorytm można zastosować do generowania danych zgodnie z rozkładem zdefiniowanym przez ciąg uczący. Przede wszystkim umożliwia on jednak szybkie wyznaczanie nowych punktów kwantyzacji, których rozkład jest proporcjonalny do gęstości rozkładu danych wejściowych.

Niewątpliwą trudnością proponowanej metody jest stosunkowo duży nakład obliczeniowy potrzebny przy estymowaniu dystrybuanty rozkładu danych na odcinku. Obliczenia te mogą być jednak wykonane wcześniej, w charakterze obliczeń wstępnych.

Ze względu na ustaloną z góry geometrię krzywej wypełniającej F, generowanie kolejnych przybliżeń krzywej F_P, zachowującej własności statystyczne danych wejściowych, jest zadaniem, w którym czas obliczeń zależy liniowo od liczby punktów wierzchołkowych aproksymacji i od dokładności wyznaczania pomocniczej krzywej F.

6.4 Podsumowanie

Omówiono metody kwantyzacji, które łączą transformację wielowymiarowych danych za pomocą quasi-odwrotności wybranej krzywej wypełniającej ze skalarną kwantyzacją w odniesieniu do danych przetransformowanych na odcinek I_1.

Zamiast algorytmu uczenia SOM, jako narzędzia kwantyzacji, można w tym miejscu użyć innego algorytmu kwantyzacji, na przykład algorytmu LBG [100].

Zastosowanie różnych wariantów algorytmu uczenia umożliwia, w pewnym stopniu, modyfikowanie kryterium kwantyzacji i w konsekwencji – własności asymptotycznych otrzymywanych kwantyzatorów (przy liczbie kwantyzatorów dążącej do nieskończoności). Pozwala to na kształtowanie gęstości rozkładu kwantyzatorów na odcinku I_1, a także, co wynika z własności odwzorowania F i jego quasi-odwrotności $\boldsymbol{\Psi}$, umożliwia również wpływ na gęstość rozkładu odpowiednich kwantyzatorów, przetransformowanych za pomocą krzywej F z odcinka I_1 do przestrzeni I_d.

Innym, nowym teoretycznym narzędziem, mogącym mieć także zastosowanie w kwantyzacji jest klasa krzywych F_P, które zachowują zadaną miarę probabilistyczną P.

Zaproponowano także algorytm aproksymacji F_P, wówczas gdy miara probabilistyczna P nie jest znana, lecz mamy dane obserwacje wielowymiarowej zmiennej losowej o rozkładzie P. Otrzymane odwzorowanie $\hat{F}_P$ daje możliwość szybkiego wyznaczania zbioru punktów kwantyzacji, których gęstość rozkładu jest taka sama jak gęstość rozkładu danych wejściowych.

Rozdział 7

Krzywe wypełniające w statystycznych problemach rozpoznawania

W ciągu ostatnich 30 lat zaproponowano bardzo wiele różnych algorytmów rozpoznawania. Część z nich cechuje prostota i ograniczony zakres zastosowania, inne są bardziej skomplikowane, nie wymagają jednak prawie żadnych dodatkowych założeń poza ogólnymi, dotyczącymi istnienia rozkładów cech w klasach. Szeroki przegląd literatury z tej dziedziny można znaleźć na przykład w pracach [19], [36], [57], [92], [131], [94], [186].

Klasyfikatory o dobrych własnościach teoretycznych wymagają na ogół dużych nakładów obliczeniowych w trakcie procesu klasyfikacji nowego obiektu. Większość klasyfikatorów wymaga ponadto przechowywania i przetwarzania całego ciągu uczącego. Jest to dużym utrudnieniem, szczególnie w przypadku problemów wielowymiarowych. Problemy związane z wizualizacją wielowymiarowych wzorców prowadzą także do braku możliwości włączenia człowieka w proces podejmowania decyzji.

Idea użycia krzywych wypełniających do redukcji wymiaru problemów sięga końca lat sześćdziesiątych. Wtedy to Patrick i in. [119] zaproponowali użycie odwzorowania kolumnowego, pewnego uogólnienia krzywej Peano, oraz odwzorowania Dovetaila (niestety nieciągłego) jako metody graficznego przedstawienia przybliżonych gęstości rozkładów prawdopodobieństwa w klasach w celu wizualnego określenia stopnia ich separowalności.

O ile autorce wiadomo, żadne wyniki związane z proponowanym podejściem nie zostały opublikowane. Idea użycia odwzorowań, które nie są wzajemnie jednoznaczne, a takim odwzorowaniem są krzywe wypełniające, została odrzucona w książce Patricka (por. [120] s. 355).

Pokażemy, że mimo iż krzywe wypełniające nie są odwzorowaniami wzajemnie jednoznacznymi, to głębsze zbadanie i wykorzystanie ich własności prowadzi do uzyskania efektywnych klasyfikatorów. Naszym celem jest podanie podstaw teoretycznych, które pozwalają na uzyskanie uniwersalnych, zgodnych klasyfikatorów operujących na danych przetransformowanych za pomocą dobrze dobranych krzywych wypełniających. Wskażemy także własności krzywych, które to zapewniają. Należy bowiem zauważyć, że nie wszystkie krzywe skanujące stosowane w przetwarzaniu obrazów są efektywne w przypadku problemów statystycznego rozpoznawania obrazów. Na przykład, wspomniane wcześniej odwzorowanie Dovetaila [119] nie powinno być zalecane, gdyż jest odwzorowaniem nieciągłym.

Główna idea proponowanego tutaj podejścia do konstrukcji klasyfikatorów w przypadku problemów wielowymiarowych polega na zastosowaniu dobrze zdefiniowanej transformacji quasi–odwrotnej do krzywej wypełniającej i przetransformowaniu wielowymiarowego ciągu uczącego w ciąg liczb z odcinka $[0,1]$. Złożoność obliczeniowa takiej transformacji jest liniowa ze względu na wymiar przestrzeni. Ponadto każdy element ciągu uczącego może być przetransformowany niezależnie od pozostałych.

Podstawową cechą proponowanych w tym rozdziale algorytmów rozpoznawania jest szybkość samego procesu podejmowania decyzji. W końcowym efekcie konstrukcji klasyfikatora otrzymujemy dyskryminacyjny podział odcinka I_1 na m pododcinków odpowiadających poszczególnym klasom, przy czym m jest nie większe, a zwykle nawet znacznie mniejsze niż długość ciągu uczącego. Nakład obliczeń potrzebny do przetransformowania za pomocą krzywej wypełniającej d–wymiarowego punktu na odcinek I_1 jest rzędu $O(d)$. W związku z tym nakład pracy potrzebny do dokonania samego procesu klasyfikacji nowego obiektu x jest rzędu $O(d) + \log_2 m$.

Kluczowym rezultatem tego rozdziału jest twierdzenie o zachowywaniu ryzyka bayesowskiego przez transformację bazującą na krzywych wypełniających o określonych własnościach, spełnianych przez klasę krzywych opisanych w niniejszej monografii.

Jedynym ograniczeniem, które w związku z tym się pojawia jest wymaganie, by odpowiednie rozkłady miały ograniczony nośnik (dane powinny pochodzić z kostki I_d). Ograniczenie to nie jest w istocie rzeczy zbyt restrykcyjne, gdyż ryzyko Bayesa jest niezmiennicze ze względu na dowolne transformacje współrzędnych, które są ciągłe i ściśle monotoniczne. Niezmienniczość ta pozwala przekształcić R^d w kostkę jednostkową I_d za pomocą odpowiednio wyskalowanej funkcji logistycznej (lub innej funkcji ściśle monotonicznej, która przekształca $R \to I_1$), co pozwala objąć teorią także rozkłady o nośnikach nieograniczonych.

Niezależnie od użytej metody klasyfikacji danych na odcinku, jednowymiarowe wzorce ciągu uczącego (po transformacji) powinny być przechowywane z dosta-

tecznie dużą dokładnością. Formalnie wielowymiarowy wektor, którego współrzędne przedstawiono z dokładnością do k cyfr znaczących, zastępowany jest przez jedną liczbę zawierającą $k\,d$ cyfr znaczących. W praktyce dokładność tę możemy znacznie zmniejszyć, dopasowując lokalnie dokładność na odcinku tak, by dane były rozróżnialne (separowalne na odcinku).

W tym kontekście proponowana metodologia nie jest na pewno łatwym ,,lekarstwem" na ,,przekleństwo wielowymiarowości", pozwala jednak w znaczym stopniu skompresować przechowywane dane bez straty informacji o zadaniu rozpoznawania, którą ze sobą niosą.

Po przetransformowaniu danych do $[0,1]$ można zastosować dowolną efektywną metodę klasyfikacji. Niezależnie od zastosowanej metody konstrukcji klasyfikatora, w końcowym efekcie otrzymujemy podział I_1 na pododcinki odpowiadające poszczególnym klasom. W tym kontekście algorytm k najbliższych sąsiadów zastosowany do przetransformowanych danych okazał się dość efektywnym podejściem, zarówno ze względu na uzyskiwane błędy klasyfikacji, jak i kompresję danych wejściowych.

Także połączenie metody LVQ Kohonena [83] z transformacją danych za pomocą krzywej prowadzi do uzyskania efektywnych algorytmów rozpoznawania. W niniejszym rozdziale szczegółowo analizujemy również metodę szeregów ortogonalnych, ze szczególnym uwzględnieniem układu Haara.

W następnym podrozdziale sformułowano statystyczny problem rozpoznawania, a w rozdziale 7.2 – podstawowe twierdzenia i lematy dotyczące możliwości rozdzielania klas i przenoszenia ryzyka Bayesa po transformacji danych za pomocą krzywej wypełniającej.

7.1 Sformułowanie problemu rozpoznawania

Rozpatrywany jest problem klasyfikacji z M klasami i ciągiem uczącym L_n złożonym z n, d–wymiarowych wektorów cech o znanej przynależności do jednej z klas. Decyzję o zakwalifikowaniu nieznanego obiektu, opisanego przez wektor cech $x = (x_1, \ldots, x_d)$, do jednej z klas podejmuje się na podstawie informacji uzyskanych podczas analizy statystycznej ciągu uczącego L_n [57], [92].

Ciąg uczący $L_n = \{(X_k, Y_k),\ k = 1, 2, \ldots, n\}$ składa się z niezależnych par (X_k, Y_k) obserwacji o tych samych rozkładach (niezależnych od k). $X_k \in R^d$ jest wektorem cech k–tego wzorca, a $Y_k \in \{1, 2, \ldots, M\}$ wskazuje prawdziwy numer klasy, do której ten wzorzec należy. Dalej, (X, Y) jest parą zmiennych losowych niezależną od L_n, którą cechuje ten sam łączny rozkład prawdopodobieństwa co (X_k, Y_k). Rozkład prawdopodobieństwa (X, Y) istnieje, lecz nie jest znany. Należy zadecydować, do której klasy należy $X = x$, a dokładniej, podać funkcję

decyzyjną, która każdemu wektorowi x przypisze numer klasy $1, 2, \ldots, M$, w taki sposób, by minimalizować prawdopodobieństwo popełnienia błędu.

Podanie optymalnego rozwiązania (optymalnej funkcji decyzyjnej) na podstawie samego ciągu uczącego L_n nie jest możliwe. Możliwe jest natomiast podanie reguł konstrukcji takiej funkcji decyzyjnej, która jest bliska, w podanym dalej sensie, optymalnej funkcji decyzyjnej.

Niech $p_i = P\{Y = i\}$, $i = 1, \ldots, M$ oznacza prawdopodobieństwo a priori pojawienia się obserwacji z klasy i, oraz niech $p_i(x) = P\{Y = i/X = x\}$, $x \in R^d$, $i = 1, 2, \ldots, M$ będzie prawdopodobieństwem, że zaobserwowany wektor x należy do i–tej klasy (prawdopodobieństwo a posteriori). Dalej, niech $f_i(x)$ oznacza gęstość rozkładu cech w klasie i–tej, a $f(x) = \sum_{i=1}^{m} p_i f_i(x)$ łączną gęstość wektora cech X. Gęstości te istnieją tylko wtedy, gdy odpowiednie rozkłady są absolutnie ciągłe względem miary Lebesgue'a.

Znajomość powyższych rozkładów pozwala skonstruować teoretycznie najlepszą regułę decyzyjną, która minimalizuje prawdopodobieństwo błędnej klasyfikacji (reguła Bayesa) bądź w ogólniejszym przypadku minimalizuje z góry określoną funkcję strat [41], [57], [94].

Bayesowska reguła klasyfikacji, oznaczana przez $g^* : R^d \to \{1, 2, \ldots, M\}$, taka że

$$g^*(X) = i, \quad \text{gdy} \quad p_i(X) = \max_{1 \le j \le M} p_j(X)$$

minimalizuje prawdopodobieństwo błędnej klasyfikacji.

Ryzyko bayesowskie J^* jest definiowane przez

$$P\{g^*(X) \neq Y\} = \inf_g P\{g(X) \neq Y\},$$

gdzie kres dolny jest wyznaczany na zbiorze wszystkich mierzalnych odwzorowań $g : R^d \to \{1, 2, \ldots, M\}$. Równoważnie,

$$g^*(X) = i, \quad \text{gdy} \quad \beta_i(X) = \max_{1 \le j \le M} \beta_j(X), \tag{7.1}$$

gdzie $\beta_j(x) \stackrel{\triangle}{=} p_j\, f_j(x), x \in R^d$ (w przypadku, gdy maksimum w (7.1) osiągane jest dla kilku klas równocześnie, wówczas wybierana jest klasa o najniższym numerze). Natomiast lokalne ryzyko Bayesa, oznaczane jako $J^*(x)$, jest równe $P\{g^*(X) \neq Y | X = x\}$.

Zastosowanie transformacji $\boldsymbol{\Psi}$ wymaga ograniczenia się do przypadku wzorców, których cechy pochodzą ze zbioru zawartego w kostce I_d. Z formalnego punktu widzenia, jak już wspominaliśmy, założenie to nie jest zbyt restrykcyjne, gdyż błąd Bayesa jest niezmienniczy ze względu na dowolne ciągłe i ściśle monotoniczne transformacje współrzędnych.

Naszym zadaniem jest skonstruowanie empirycznej (bazującej na ciągu uczącym L_n) reguły decyzyjnej $g_n : I_d \to \{1, 2, \ldots, M\}$, która jest asymptotycznie optymalna w tym sensie, że

$$J_n = P\{g_n(X) \neq Y | L_n\} \to J^\star, \text{ gdy } n \to \infty \text{ z prawdopodobieństwem 1.} \quad (7.2)$$

Ciąg reguł decyzyjnych g_n, $n \to \infty$, spełniających powyższy warunek nazywamy zgodnym z prawdopodobieństwem 1 (mocno zgodnym) z regułą Bayesa $g^\star$ [36].

Dalej będziemy rozpatrywać tylko przypadek, gdy istnieją jedynie dwie klasy ($M = 2$) o numerach (etykietach) 0 oraz 1. Przypadek ogólny ($M > 2$), po przetransformowaniu problemu wielowymiarowego do zadania w jednym wymiarze (na odcinku I_1), łatwo redukuje się do ciągu dychotomii.

W niniejszym rozdziale posługujemy się terminologią dotyczącą statystycznych problemów rozpoznawania zgodnie z [36], [57].

7.2 Ryzyko bayesowskie i rozdzielanie przetransformowanych wzorców

Rozpatrzmy najpierw podstawowe własności wzorców (a raczej wektorów ich cech) po przetransformowaniu do odcinka $[0, 1]$. Wybierzmy krzywą wypełniającą $F : I_1 \to I_d$ spełniającą warunki **C1–C3** oraz jej quasi-odwrotność $\boldsymbol{\Psi} : I_d \to I_1$. W dalszym ciągu zakładać będziemy, że znana jest dokładna wartość $\boldsymbol{\Psi}(x)$, $x \in I_d$. Przypomnijmy, że $F(\boldsymbol{\Psi}(x)) = x$, $x \in I_d$. Ponadto zauważmy, że jeśli X jest zmienną losową przyjmującą wartości w I_d, to $\boldsymbol{\Psi}(X)$ jest także zmienną losową (funkcja $\boldsymbol{\Psi}$ jest mierzalna) przyjmującą wartości w I_1. W przypadku ogólnym trudno jest podać wprost zależność między rozkładem X i $\boldsymbol{\Psi}(X)$.

W przypadku istnienia funkcji f – gęstości rozkładu X – istnieje również gęstość rozkładu $\boldsymbol{\Psi}(X)$. Z lematu 4.16 wynika bezpośrednio, że gęstość ta ma postać $f(F(t))$. Pokażemy dalej, że krzywe wypełniające, które spełniają warunki **C1–C3** z rozdziału 4.5, prowadzą do transformacji, które nie zmieniają ryzyka Bayesa.

Niech S_i, $i = 0, 1$ oznacza nośnik (z prawdopodobieństwem 1) rozkładu cech w klasach. Jeśli istnieją gęstości rozkładu w klasach $f_i(x)$ określone w I_d, i znikające poza I_d, to zbiory $S_i = \{x \in I_d, f_i(x) > 0\}$, $i = 0, 1$ są nośnikami gęstości rozkładów w klasach.

Definicja 7.1 *Mówimy, że klasy* 0 *i* 1 *są:*

a) **ściśle separowalne** *(ze względu na metrykę $D(x, x')$ w R^d), wtedy i tylko wtedy, gdy istnieje $\varepsilon > 0$ takie, że* $\operatorname{dist}(S_0, S_1) = \inf_{x \in S_0, x' \in S_1} D(x, x') \geq \varepsilon$,

b) **przemieszane**, *wtedy i tylko wtedy, gdy* $\mu_d(S_0 \cap S_1) > 0$,

c) **słabo separowalne**, *wtedy i tylko wtedy, gdy* $\mu_d(S_0 \cap S_1) = 0$.

Następujące własności są łatwe do udowodnienia.

Lemat 7.1 1) *S_0, S_1 są ściśle separowalne ze względu na metrykę D, wtedy i tylko wtedy, gdy $\bar{S}_0 \cap \bar{S}_1 = \emptyset$, gdzie $\bar{S}_0$, $\bar{S}_1$ oznacza domknięcia zbiorów S_0, S_1 (odpowiednio) w topologii wprowadzonej przez metrykę D.*
2) *Jeśli $\mu_d(S_0 \cap S_1) = 0$, to nośniki S_0, S_1 są słabo separowalne, a nie są ściśle separowalne wtedy i tylko wtedy, gdy* $\operatorname{dist}(S_0, S_1) = 0$ □.

Z punktu widzenia problemów rozpoznawania najistotniejsze jest to, że transformacja nośników S_0 i S_1 za pomocą quasi–odwrotności krzywej wypełniającej zachowuje podstawowe relacje między nimi. Bardziej precyzyjnie formułuje ten fakt poniższe twierdzenie:

Twierdzenie 7.2.1 *Przeciwobrazy nośników S_0, S_1 względem krzywej wypełniającej F określają, odpowiednio, zbiory $A_i = \{t \in I_1 : F(t) \in S_i\}$, $i \in \{0,1\}$. Ponadto oznaczmy $\tilde{A}_i = \boldsymbol{\Psi}(S_i)$, $i \in \{0,1\}$.*

Jeśli krzywa F spełnia warunki **C1**–**C3**, *to:*

a) *jeśli S_0, S_1 są ściśle separowalne w I_d, to A_0, A_1 ($\tilde{A}_0$, $\tilde{A}_1$) są także ściśle separowalne w I_1, ze względu na tę samą metrykę, którą wybrano w* **C1**,

b) *jeśli S_0, S_1 są słabo separowalne w I_d, to wtedy także A_0, A_1 ($\tilde{A}_0$, $\tilde{A}_1$) są słabo separowalne w I_1,*

c) *jeśli S_0, S_1 są przemieszane w I_d, to A_0, A_1 ($\tilde{A}_0$, $\tilde{A}_1$) są przemieszane.*

Dowód. Ponieważ $\boldsymbol{\Psi}(x) \in F^{-1}(x)$, zatem $\tilde{A}_i \subset A_i$, $i = 0, 1$. Załóżmy, że zbiory S_0 i S_1 są ściśle rozdzielone. Wtedy dla każdego $x \in S_0$, $x' \in S_1$ zachodzi

$$\varepsilon \leq \| x - x' \| = \| F(t) - F(t') \| \leq \alpha_d |t - t'|^{1/d}, \tag{7.3}$$

gdzie $t \in A_0$, $t' \in A_1$ są przeciwobrazami x oraz x', odpowiednio, natomiast ostatnia nierówność w (7.3) wynika z własności **C1** krzywej wypełniającej F. Stąd A_0 oraz A_1 są ściśle rozdzielone, a ich odległość jest nie mniejsza niż $(\varepsilon/\alpha_d)^d$, co kończy dowód własności a). By udowodnić część b), wystarczy zauważyć, korzystając z własności **C2** krzywej, że z $\mu_d(S_0 \cap S_1) = 0$ wynika, iż $\mu_1(A_0 \cap A_1) = 0$. Z własności a) i b) wynika naturalnie c), co kończy dowód twierdzenia. □

Należy zwrócić uwagę na fakt, iż stwierdzenie odwrotne do twierdzenia 7.2.1 niekoniecznie musi być prawdziwe, to znaczy klasy, które są ściśle rozdzielone w I_1 mogą po transformacji przez krzywą F stać się jedynie słabo rozdzielonymi w I_d. Z drugiej strony, jeżeli klasy są ściśle rozdzielone w I_d, możemy znacznie więcej wnioskować o nośnikach klas po transformacji $\boldsymbol{\Psi}$, czyli o zbiorach A_i.

Twierdzenie 7.2.2 *Jeśli zbiory S_0, S_1 są ściśle rozdzielone w I_d, to istnieje skończony podział odcinka I_1, który rozdziela obie klasy w ten sposób, że wnętrze żadnego z pododcinków nie zawiera równocześnie punktów z A_0 i A_1. Ponadto liczba punktów konieczna do rozdzielenia punktów ze zbioru A_0 od punktów ze zbioru A_1 jest nie większa niż $\lceil(\alpha_d/\varepsilon)^d\rceil - 1$.*

Dowód. Istnieje skończone pokrycie I_1 zbiorem domkniętych odcinków o długości $\delta = (\varepsilon/\alpha_d)^d)$, gdzie $\varepsilon = \mathrm{dist}(S_0, S_1)$ oraz α_d jest stałą z warunku Höldera **C1**, który spełnia krzywa F. Stąd $I_1 \subset [0,\delta]\cup\ldots\cup[(\lceil 1/\delta\rceil - 1)\,\delta, 1]$. Z warunku Höldera **C1** wynika, że wnętrze żadnego z odcinków $[i\delta, (i+1)\delta]$, $i = 0, 1, \ldots$ nie może zawierać równocześnie punktów z A_0 i z A_1. W konsekwencji liczba punktów rozdzielających A_0 od A_1 na pewno nie jest większa niż $\lceil(L_d/\varepsilon)^d\rceil - 1$. □

Kluczowym wnioskiem wynikającym z powyższego twierdzenia jest stwierdzenie, że możliwość rozdzielenia zbiorów ściśle rozdzielonych po przetransformowaniu ich na odcinek I_1 za pomocą skończonej liczby punktów dyskryminujących jest własnością konstruktywną, gdyż można podać górne oszacowanie liczby punktów dyskryminujących.

Poniższe twierdzenie uzasadnia poprawność zastosowania transformacji $\boldsymbol{\Psi}$ w problemach rozpoznawania, mimo iż nie jest to odwzorowanie wzajemnie jednoznaczne.

Twierdzenie 7.2.3 *Niech $g^\star(X)$ będzie bayesowską regułą klasyfikacji dla problemu opisanego rozkładami (X,Y), $X \in I_d$, a $J_X^\star$ ryzykiem Bayesa. Niech $T = \boldsymbol{\Psi}(X)$, gdzie $\boldsymbol{\Psi}$ jest odwzorowaniem quasi-odwrotnym krzywej F spełniającej warunki* **C1**–**C3**. *Wtedy reguła klasyfikacji postaci: $G(T) \stackrel{def}{=} g^\star(F(T))$ jest regułą Bayesa dla problemu klasyfikacji o rozkładach (T,Y), $T \in I_1$. Ponadto ryzyko Bayesa $J_T^\star$ dla problemu (T,Y) jest także równe $J_X^\star$.*

Dowód. Zauważmy, że $F(T)$ jest zmienną losową, a ponadto $F(T) = F(\boldsymbol{\Psi}(X)) = X$. Dalej, niech $G^\star(T)$ będzie regułą Bayesa dla problemu przetransformowanego (T,Y). Łatwo zauważyć, że $g^\star(F(T))$ jest pewną regułą klasyfikacji w problemie (T,Y), stąd $J_T^\star = P\{G^\star(T) \neq Y\} \leq P\{g^\star(F(T)) \neq Y\} = P\{g^\star(X) \neq Y\} = J_X^\star$. Z drugiej strony, $G^\star(\boldsymbol{\Psi}(X))$ jest pewną regułą klasyfikacji oryginalnego problemu (X,Y). Stąd $P\{g^\star(X) \neq Y\} \leq P\{G^\star(\boldsymbol{\Psi}(X)) \neq Y\} = P\{G^\star(T) \neq Y\} = J_T^\star$. W konsekwencji $J_X^\star = J_T^\star$ i $g^\star(F(T))$ musi być optymalną regułą klasyfikującą dla problemu (T,Y). □

W twierdzeniu 7.2.3 nie zakładaliśmy żadnych ograniczeń na rozkład X (poza wstępnymi założeniami, że X przyjmuje wartości z ograniczonego obszaru I_d, których spełnienie jest łatwo zagwarantować, dokonując odpowiedniej wstępnej transformacji zmiennych). Ponadto, gdy Z jest zmienną losową przyjmującą wartości w I_1, błąd Bayesa $J_{F(Z)}^\star$ dla problemu przetransformowanego za pomocą

krzywej, czyli problemu opisanego przez zmienne losowe $(F(Z), Y)$, może być większy niż odpowiedni błąd w problemie przed transformacją (Z, Y). Różnica ta wynika stąd, że w przypadku którego dotyczy twierdzenie 7.2.3, transformacji podlega nie dowolna zmienna losowa, lecz zmienna losowa, która przyjmuje wartości w $\Psi(I_d)$, a nie w całym odcinku I_1.

Z twierdzenia 7.2.3 wynika w szczególności, że jeśli istnieją gęstości rozkładów w klasach f_0 i f_1, to reguła klasyfikacji Bayesa

$$g^\star(x) = \begin{cases} 0, & \text{gdy } p_1 f_1(x) - p_0 f_0(x) \leq 0 \\ 1, & \text{w przeciwnym przypadku} \end{cases}$$

prowadzi do reguły $G^\star(t) \stackrel{def}{=} g^\star(F(t))$, $t \in I_1$, która jest regułą Bayesa problemu klasyfikacji z tymi samymi prawdopodobieństwami a priori p_0, p_1 oraz rozkładami w klasach $f_0(F(t))$ i $f_1(F(t))$, $t \in I_1$.

Dalej koncentrować się będziemy na estymacji reguł Bayesa, w sytuacji, gdy S_0 i S_1 pokrywają się choćby częściowo na zbiorze o niezerowej mierze Lebesgue'a, a ryzyko Bayesa jest większe od zera. W tym przypadku nie jesteśmy w stanie zagwarantować, że istnieje skończona liczba punktów na odcinku, które rozdzielają obszary należące do różnych klas (ze względu na optymalną regułę klasyfikacji). Zauważmy bowiem, że każdą regułę decyzyjną na odcinku możemy jednoznacznie zdefiniować, podając położenie punktów, w których następuje zmiana decyzji o przynależności do danej klasy (z klasy 0 na klasę 1 lub odwrotnie) oraz numeru klasy, do której należy przyporządkować punkty z pierwszego podprzedziału I_1. Niestety, w ogólnym przypadku liczba takich punktów może być nie tylko nieskończona, ale i nieprzeliczalna. Możemy jednak pokazać, jak wybierając skończony podział odcinka jednostkowego na odpowiednie pododcinki związane z różnymi klasami możemy aproksymować regułę decyzyjną Bayesa z dowolną wymaganą dokładnością $\delta > 0$. W związku z tym rozpatrzmy następującą regułę klasyfikacji, która dopuszcza przydzielenie danej obserwacji x etykiety ,,niesklasyfikowana" (por. [36], [57]). Reguła ta jest postaci:

$$g^\star_\delta(x) = \begin{cases} 0, & p_1(x) - p_0(x) \leq -\delta \\ 1, & p_1(x) - p_0(x) \geq \delta \\ \text{niesklasyfikowany}, & |p_1(x) - p_0(x)| < \delta. \end{cases}$$

W tym przypadku zbiory $C_i^\delta \stackrel{def}{=} \{x \in I_d : g^\star_\delta(x) = i\}$, $i = 0, 1$ są ściśle rozdzielone i możemy do nich zastosować wnioski wynikające z twierdzenia 7.2.2.

7.3 Klasyfikatory w postaci szeregów ortogonalnych

W niniejszym rozdziale będziemy używać przetransformowanego na odcinek $[0,1]$ ciągu uczącego do estymowania współczynników w rozwinięciu ortogonalnym optymalnej reguły decyzyjnej. Postępowanie to prowadzi do uzyskania klasyfikatora, który:

1. Jest asymptotycznie optymalny w tym sensie, że prawdopodobieństwo popełnienia błędu jest zbieżne do ryzyka Bayesa z prawdopodobieństwem jeden (prawie na pewno).
2. Cechuje się znacznym stopniem kompresji danych, gdyż nie wymaga przechowywania całego ciągu uczącego, wystarczy jedynie zapamiętać współczynniki w rozwinięciu ortogonalnym.
3. Pozwala na szybką klasyfikację nowego obiektu do rozpoznania.
4. Umożliwia łatwą graficzną interpretację ciągu uczącego i otrzymanej na odcinku $[0,1]$ reguły decyzyjnej.
5. Pozwala na łatwą aktualizację klasyfikatora w przypadku poszerzenia ciągu uczącego.

Klasyfikatory w postaci szeregu ortogonalnego należą do grupy klasyfikatorów, które są konstruowane na zasadzie ,,wstawiania" (*plug-in*) w optymalną regułę klasyfikacyjną odpowiedniej estymaty [36]. Ciąg uczący zostanie tu użyty bezpośrednio do estymacji współczynników rozwinięcia optymalnej funkcji decyzyjnej w szereg o ortogonalnej bazie. Użycie klasyfikatorów w postaci szeregów ortogonalnych zostało zapoczątkowane w pracach [68], [177].

Zastosowanie szeregów ortogonalnych wymaga, w ogólnym przypadku, założenia, że istnieją gęstości rozkładów w klasach f_0 i f_1, a ponadto, że gęstości te należą do klasy funkcji całkowalnych z kwadratem, czyli

$$f_1,\ f_2 \in L_2(I_d),\ \int_{I_d} f_i^2(x)dx < \infty,\ i = 0, 1.$$

Przypomnijmy, że bayesowska reguła decyzyjna jest określona przez

$$g^\star(x) = \begin{cases} 0, & \text{gdy} \quad \alpha(x) \leq 0, \\ 1, & \text{w przeciwnym przypadku,} \end{cases}$$

gdzie

$$\alpha(x) = p_1 f_1(x) - p_0 f_0(x),\ \ p_0 + p_1 = 1$$

lub – równoważnie,

$$\alpha(x) = [2y(x) - 1]\, f(x), \;\; y(x) \stackrel{\triangle}{=} E(Y|X = x) = P\{Y = 1|X = x\} = p_1(x).$$

Zauważmy, że w przypadku istnienia gęstości

$$y(x) = p_1 f_1(x)/f(x), \;\; \text{gdy} \;\; f(x) = p_1 f_1(x) + p_0 f_0(x) > 0$$

oraz $y(x) = 0$, gdy $f(x) = 0$.

Aby estymować $\alpha(x)$, potrzebny jest system funkcji $v_1(t), v_2(t), \ldots, t \in I_1$, który tworzy bazę ortonormalną w $L_2(I_1)$, przestrzeni funkcji określonych i całkowalnych z kwadratem na odcinku $I_1 = [0, 1]$. Zakładamy ponadto, że istnieje niemalejący ciąg liczb V_j, czyli taki że

$$|v_j(t)| \leq V_j, \; t \in I_1, \;\; j = 1, 2, \ldots$$

Funkcję $\alpha(x)$ możemy przedstawić w postaci następującego rozwinięcia:

$$\alpha(x) = \sum_{j=1}^{\infty} b_j \, v_j \left(\boldsymbol{\Psi}(x)\right), \tag{7.4}$$

gdzie $\boldsymbol{\Psi}$ jest quasi-odwrotnością wybranej krzywej wypełniającej F, podczas gdy

$$\begin{aligned} b_j &= E\{(2Y - 1) v_j \left(\boldsymbol{\Psi}(X)\right)\} \\ &= \textstyle\int_{I_d} [2y(x) - 1]\, v_j \left(\boldsymbol{\Psi}(x)\right) f(x) dx, \quad j = 1, 2, \ldots \end{aligned} \tag{7.5}$$

Naturalnym estymatorem b_j jest wyrażenie

$$\begin{aligned} \hat{b}_j &= E_n\{(2Y - 1) v_j \left(\boldsymbol{\Psi}(X)\right)\} \\ &= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (2Y_i - 1)\, v_j(t_i), \quad j = 1, 2, \ldots, \end{aligned} \tag{7.6}$$

gdzie $t_i \stackrel{\triangle}{=} \boldsymbol{\Psi}(X_i)$, $i = 1, 2, \ldots, n$. Łatwo zauważyć, że

$$\begin{aligned} E\{\hat{b}_j\} &= b_j, \\ \text{var}\left(\hat{b}_j\right) &\leq V_j^2/n, \quad j = 1, 2, \ldots \end{aligned} \tag{7.7}$$

Ponadto, jeśli potrzeba dokładniejszego oszacowania, to:

$$\begin{aligned} &\textstyle\sum_{j=1}^{k} \text{var}\left(\hat{b}_j\right) \\ &\leq n^{-1} E\left[\textstyle\sum_{j=1}^{k} v_j^2 \left(\boldsymbol{\Psi}(X)\right)(2y(X) - 1)^2\right] \\ &\leq \; n^{-1} \sup_{t \in I_1} \left(\textstyle\sum_{j=1}^{k} v_j^2(t)\right). \end{aligned}$$

Naturalnym estymatorem $\alpha(x)$ jest

$$\hat{\alpha}_n(x) = \sum_{j=1}^{k(n)} \hat{b}_j v_j\left(\boldsymbol{\Psi}(x)\right), \tag{7.8}$$

gdzie $k(n)$ oznacza odpowiednio wolno rosnący ciąg liczb dodatnich (dalsze wymagania dotyczące $k(n)$ podamy dalej).

W konsekwencji możemy sformułować następującą regułę klasyfikacji, która jest tylko pewnym ogólnym schematem postępowania:

ALGORYTM TYPU SZEREG ORTOGONALNY

Krok 1. Przetransformuj ciąg uczący (X_i, Y_i), $i = 1, 2, \ldots, n$ w ciąg (t_i, Y_i), $t_i = \boldsymbol{\Psi}(X_i)$, $i = 1, 2, \ldots, n$.

Krok.2 Wyznacz wartości $\hat{b}_j$, $j = 1, 2, \ldots, k(n)$ zgodnie ze wzorem (7.6).

Krok 3. Aby sklasyfikować nowy wzorzec x, oblicz $t = \boldsymbol{\Psi}(x)$ oraz

$$\hat{\alpha}_n = \sum_{j=1}^{k(n)} \hat{b}_j v_j(t).$$

Następnie przydziel x do klasy 0, jeśli $\hat{\alpha}_n \leq 0$, lub do klasy 1, w przeciwnym przypadku.

Efektywność powyższego algorytmu zależy od wybranej transformacji $\boldsymbol{\Psi}$, od postaci wybranego szeregu ortogonalnego oraz od $k(n)$, jak to zostanie dokładnie przeanalizowane dalej.

Przed przejściem do szczegółów algorytmu należy zaznaczyć, że szybkość zbieżności (7.4) będziemy rozpatrywać względem normy L_2. Rozwinięcie (7.4) rozumiane jest w następującym sensie:

$$\lim_{k\to\infty} \int_{I_d} \left(\alpha(x) - \sum_{j=1}^{k} b_j v_j(\boldsymbol{\Psi}(x)) \right)^2 dx = 0$$

lub równoważnie (korzystając z lematu 4.15)

$$\lim_{k\to\infty} \int_{I_1} \left(\alpha\left(F(t)\right) - \sum_{j=1}^{k} b_j v_j(t) \right)^2 dt = 0. \tag{7.9}$$

Biorąc pod uwagę zupełność i ortonormalność układu funkcji $\{v_j\}_{j=1}^{\infty}$ w $L_2(I_1)$, wystarczy założyć, że funkcja $\alpha(F(t))$ jest całkowalna z kwadratem, by zapewnić spełnienie warunku (7.9). Z lematu 4.15 wynika także równość całek $\int_{I_d} \alpha^2(x)dx = \int_{I_1} \alpha^2(F(t))dt$. Podsumowując, wystarczy założyć, że gęstości w klasach $f_0(x)$ i $f_1(x)$ są całkowalne z kwadratem w I_d, z czego wynika, że $\alpha(x) = p_0 f_0(x) + p_1 f_1(x) \in L_2(I_d)$, a to z kolei implikuje równość (7.9).

7.3.1 Zgodność klasyfikatorów w postaci szeregów ortogonalnych

Obecnie pokażemy zgodność klasyfikatora określonego przez (7.8), przy stosunkowo słabych założeniach dodatkowych.

W dużej mierze korzystać będziemy w tym miejscu ze znanych rezultatów dotyczących zgodności klasyfikatorów w postaci szeregów ortogonalnych (patrz [36] strona 284) oraz zebranych w rozdziale 4 własności odwzorowań F i $\boldsymbol{\Psi}$.

Będzie nam również potrzebne uogólnienie dotyczące szeregów, które nie są wspólnie ograniczone od góry przez pewną stałą, lecz $\{v_j\}$ rosną w sposób nieograniczony.

Oznaczmy całkę z kwadratu błędu $\hat{\alpha}_n$ przez $\mathrm{ISE}(n)$.

$$\begin{aligned} \mathrm{ISE}(n) &= \int_{I_d} \left[\alpha(x) - \sum_{j=1}^{k(n)} \hat{b}_j v_j\left(\boldsymbol{\Psi}(x)\right)\right]^2 dx \\ &= \int_{I_1} \left[\alpha\left(F(t)\right) - \sum_{j=1}^{k(n)} \hat{b}_j v_j(t)\right]^2 dt \\ &= \sum_{j=1}^{k(n)} \left(b_j - \hat{b}_j\right)^2 + \sum_{j=k(n)+1}^{\infty} b_j^2. \end{aligned} \tag{7.10}$$

Ostatnia równość w (7.10) wynika z zupełności i ortonormalności szeregu $\{v_j\}$ w $L_2(I_1)$, oraz z (7.5) i lematu 4.15.

Bezpośrednio z własności (7.7) i (7.10) wynika dalej

$$E\left[\mathrm{ISE}(n)\right] \leq W_{k(n)}/n + R_{k(n)} \leq \frac{k(n)V_{k(n)}^2}{n} + R_{k(n)}, \tag{7.11}$$

gdzie

$$W_{k(n)} \stackrel{def}{=} \sup_{t \in I_1} \left(\sum_{j=1}^{k(n)} v_j^2(t) \right) \tag{7.12}$$

i

$$R_{k(n)} \stackrel{def}{=} \sum_{j=k(n)+1}^{\infty} b_j^2 = \int_{I_1} \left[\alpha(F(t)) - \sum_{j=1}^{k(n)} b_j v_j(t) \right]^2 dt. \tag{7.13}$$

Możemy także nałożyć dodatkowe warunki na postać układu funkcji $\{v_j\}$, z których to warunków wynika, że $\{v_j\}$ nie są wspólnie ograniczone, lecz

$$W_{k(n)} \leq V^2\, k(n), \tag{7.14}$$

gdzie V jest pewną stałą.

W podrozdziale 7.3.3 pokażemy, że warunek (7.14) jest spełniony przez układ Haara, który jak wiadomo nie jest ograniczony.

W tym momencie możemy sformułować następujące twierdzenie:

Twierdzenie 7.3.1 *Niech v_j, $j = 1, 2, \ldots$ będzie zupełną i ortonormalną bazą funkcji w przestrzeni $L_2(0,1)$. Dalej, niech istnieje niemalejący ciąg liczb V_j takich, że $|v_j(t)| \leq V_j$, $t \in I_1$, $j = 1, 2, \ldots$*

Załóżmy ponadto, iż gęstości rozkładów w klasach f_0 i f_1 są określone w I_d oraz pochodzą z przestrzeni $L_2(I_d)$.

1) *Jeśli ciąg $k(n)$ wybierzemy tak, by*

$$k(n) \to \infty \quad oraz \quad \frac{V_{k(n)}^2\, k(n)}{n} \to 0, \quad gdy \quad n \to \infty, \tag{7.15}$$

to reguła klasyfikacji:

$$\hat{g}_n(x) = \begin{cases} 0, & jeśli \quad \hat{\alpha}_n(x) \leq 0, \\ 1, & w\ przeciwnym\ przypadku, \end{cases} \tag{7.16}$$

gdzie $\hat{\alpha}_n(x)$ określono w (7.6), jest zgodna, czyli

$$\lim_{n\to\infty} E\, J(\hat{g}_n) = J^\star. \tag{7.17}$$

2) *Jeżeli*

$$k(n) \to \infty \quad oraz \quad V_{k(n)}^2\, \frac{k(n)\log(n)}{n} \to 0, \quad gdy \quad n \to \infty, \tag{7.18}$$

to wtedy reguła $\hat{g}_n$ jest mocno zgodna, czyli z prawdopodobieństwem jeden zachodzi $\lim_{n\to\infty} J(\hat{g}_n) = J^\star$.

Dowód. Aby wykazać (7.17), wystarczy zauważyć, że z (7.11), (7.15) oraz (7.9) wynika, iż $E[\mathrm{ISE}(n)] \to 0$, gdy $n \to \infty$. Dalej, korzystając z twierdzenia 3 z pracy [198] otrzymujemy (7.17).

Udowodnienie drugiej części twierdzenia wymaga stwierdzenia, że istnieje liczba $\varepsilon > 0$ oraz n (naturalne), dla których $R_{k(n)} < \varepsilon/2$. Wtedy

$$\begin{aligned} &P\left\{\mathrm{ISE}(n) > \varepsilon\right\} \leq P\left\{\sum_{j=1}^{k(n)} \left(\hat{b}_j - b_j\right)^2 > \varepsilon/2\right\} \\ &\leq \sum_{j=1}^{k(n)} P\left\{\left|\hat{b}_j - b_j\right| > (\varepsilon/2k(n))^{1/2}\right\}. \end{aligned} \tag{7.19}$$

Dalej, z własności (7.6) i (7.7) wynika, że

$$\hat{b}_j - b_j = \frac{1}{n}\sum_{i=1}^{n}\left[(2Y_i - 1)\, v_j(t_i) - E\left((2Y_i - 1)\, v_j(t_i)\right)\right]. \tag{7.20}$$

Korzystając z (7.19), (7.20) oraz nierówności Hoeffdinga ([36] twierdzenie 8.1), otrzymujemy

$$P\left\{\left|\hat{b}_j - b_j\right| > (\varepsilon/2k(n))^{1/2}\right\} \leq 2\exp\left[-\frac{n\varepsilon}{2k(n)V_j^2}\right]. \tag{7.21}$$

Z monotoniczności ciągu liczb V_j^2 oraz warunków (7.19)–(7.21) wynika dalej

$$P\left\{\mathrm{ISE}(n) > \varepsilon\right\} \leq 2k(n)\exp\left[-\frac{n\varepsilon\,\log(n)}{2k(n)V_{k(n)}^2\log(n)}\right] \tag{7.22}$$

$$= 2\,\frac{k(n)}{n}n^{1-\varepsilon/(2\delta_n)},$$

gdzie $\delta_n \stackrel{def}{=} k(n)V_{k(n)}^2\log(n)/n$. Stąd, korzystając z (7.18) oraz ograniczoności $k(n)/n$, możemy wnioskować, iż szereg

$$\sum_{n=1}^{\infty} P\{\mathrm{ISE}(n) > \varepsilon\}$$

jest zbieżny dla każdego $\varepsilon > 0$.

Z lematu Borela–Cantellego [142] wynika, że $\mathrm{ISE}(n) \to 0$ z prawdopodobieństwem jeden. Dowód kończy zastosowanie wspomnianego wcześniej twierdzenia 3 z pracy [198]. □

Dodatkowo, jeśli poza założeniami twierdzenia 7.3.1, układ funkcji bazowych spełnia także warunek (7.14), to w (7.15) oraz (7.18) zamiast $V_{k(n)}^2$ można wpisać stałe V^2 niezależne od $k(n)$ i twierdzenie 7.3.1 może mieć taką samą postać jak w przypadku użycia ograniczonego szeregu funkcyjnego.

Zauważmy, że założenia dotyczące rozkładów nie są nadmiernie restrykcyjne, a mianowicie f_0, f_1 powinny należeć do $L_2(I_d)$, co jest typowym wymaganiem w przypadku klasyfikatorów typu szeregu ortogonalnego. Nie jest to niestety wynik o charakterze uniwersalnym. Pokażemy jednak dalej, iż istnieją takie układy funkcji, które pozwalają w zasadzie nie nakładać żadnych ograniczeń na postacie rozkładów cech w klasach.

7.3.2 Szybkość zbieżności klasyfikatorów opartych na szeregach ortogonalnych

Dwa ciągi zmiennych losowych, A_n i B_n, spełniają warunek $A_n = O(B_n)$ z prawdopodobieństwem 1, jeśli dla dowolnego ciągu a_n zbieżnego do zera także ciąg $a_n A_n / B_n \to 0$, gdy $n \to \infty$ z prawdopodobieństwem 1.

By zaobserwować, jakie czynniki mają wpływ na tempo zbieżności reguł klasyfikujących opartych na szeregach ortogonalnych, zbadajmy najpierw, jak szybko dąży do zera (z prawdopodobieństwem 1) wyrażenie $\mathrm{ISE}(n) = \int_{I_d} (\alpha(x) - \hat{\alpha}_n(x))^2 dx$.

Oszacowanie szybkości zbieżności $L_n - L^\star$ wymaga nałożenia na postać $\alpha(x)$ pewnych dodatkowych wymagań. Na początek wyznaczmy ciąg B_n taki, że dla dowolnie małego $\varepsilon > 0$ oraz dla każdego zbieżnego do 0 ciągu a_n szereg a_n^2 / B_n^2 spełnia warunek

$$\sum_{n=1}^{\infty} P\left\{ \frac{a_n^2}{B_n^2} \mathrm{ISE}(n) > \varepsilon \right\} < \infty. \tag{7.23}$$

Oznaczmy $\mathrm{ISB}(n) \stackrel{\triangle}{=} \sum_{j=1}^{k(n)} \left(b_j - \hat{b}_j \right)^2$. Jeśli n jest wystarczająco duże, by zachodziło $B_n^2 \varepsilon / a_n^2 - R_{k(n)} > 0$, to korzystając dodatkowo z (7.10) oraz (7.13), otrzymujemy

$$P\{\mathrm{ISE}(n) > \frac{B_n^2 \varepsilon}{a_n^2}\} = P\left\{ \mathrm{ISB}(n) > \frac{B_n^2 \varepsilon}{a_n^2} - R_{k(n)} \right\} \tag{7.24}$$

$$\leq 2k(n) \exp\left[- \frac{n\left(\varepsilon B_n^2 / a_n^2 - R_{k(n)} \right)}{k(n) V^2} \right],$$

gdzie ostatnia nierówność wynika z nierówności Hoeffdinga.

Lemat 7.2 *Załóżmy, że funkcje gęstości f_0 i f_1 należą do przestrzeni* $\mathrm{Lip}(\nu)$, *to znaczy spełniają warunek*

$$|f_i(x) - f_i(y)| \leq \mathrm{const} ||x - y||^\nu, \quad x, y \in I_d,\ \nu \in (0, 1], \quad i = 0, 1.$$

Wtedy funkcja $h(t) = \alpha(F(t)) = p_1 f_1(F(t)) - p_0 f_0(F(t))$, $t \in I_1$ należy do klasy $\mathrm{Lip}(\nu/d)$ *na odcinku I_1.*

Dowód. Dowód wynika z warunku Höldera **C2** spełnianego przez krzywą wypełniającą F. □

Lemat 7.3 *Niech układ funkcji ortogonalnych* $\{v_j\}_{j=1}^{\infty}$ *spełnia dodatkowo warunek*

$$||h - \sum_{j=1}^{k} \langle h, v_j \rangle v_j(.)||_{L_2}^2 = O\left(k^{-2\gamma}\right), \quad \gamma \in (0,1],$$

przy czym $h(t)$ *jest dowolną funkcją z klasy* $\mathrm{Lip}(\gamma)$ *na odcinku* I_1. *Jeśli wielowymiarowe gęstości* f_0, f_1 *należą do klasy* $\mathrm{Lip}(\nu)$, $\nu \in (0,1]$, *to błąd* $R_{k(n)}$, *zdefiniowany w (7.13), jest rzędu* $O\left(k^{-2\nu/d}(n)\right)$. □

Należy zwrócić uwagę na to, że założenia dotyczące układu funkcji $\{v_j\}_{j=1}^{\infty}$ w lemacie 7.3 spełnia między innymi szereg Haara–Fouriera (patrz twierdzenie 1 w pracy [45]), a także układ funkcji trygonometrycznych, jeśli, $h(t)$ jest okresowe. Wymaganie okresowości $h(t)$ może być łatwo spełnione poprzez zastosowanie krzywej wypełniającej zdefiniowanej na okręgu jednostkowym, a nie tylko w I_1. Taką krzywą jest każda krzywa zamknięta (na przykład krzywa Sierpińskiego). Wtedy $h(t) = \alpha(F(t))$ może być traktowana jako funkcja okresowa, niezależnie od tego, czy α jest okresowa czy też nie.

Wybierzmy ciągi $k(n)$ i B_n w następujący sposób:

$$k(n) = k_0\, n^{d/(2\nu+d)}, \quad k_0 > 0, \tag{7.25}$$

$$B_n = b_0 \sqrt{\log\log(n)}\, n^{-\nu/(2\nu+d)}, \quad b_0 > 0. \tag{7.26}$$

Z (7.24) otrzymujemy

$$\begin{aligned} &P\left\{\mathrm{ISE}(n) > B_n^2 \varepsilon / a_n^2\right\} \\ &\leq 2c' k(n) \exp\left[-c'' \varepsilon a_n^{-2} \log\log(n)\right] = 2c' \frac{k(n)}{n} n^{1+\gamma_n}, \end{aligned} \tag{7.27}$$

gdzie $c' > 0$, $c'' > 0$ są pewnymi stałymi, natomiast $\gamma_n \stackrel{\triangle}{=} -c''\varepsilon \log(n)/a_n^2$.

Ponieważ $a_n \to 0$, gdy $n \to \infty$, stąd wynika, że $\gamma_n \to -\infty$ i dla dostatecznie dużego n zachodzi $1 + \gamma_n < -2$. Dalej, korzystając z (7.27) oraz ograniczoności $k(n)/n$, możemy wnioskować, że szereg (7.23) jest zbieżny dla dowolnego $\varepsilon > 0$. W ten sposób udowodniliśmy następujący lemat:

Lemat 7.4 *Przy założeniach twierdzenia 7.3.1 spełniony jest lemat 7.3. Niech dodatkowo szereg funkcji bazowych spełnia (7.14). Jeżeli* $k(n)$ *jest takie, że*

$$c_1\, n^{d/(2\nu+d)} \leq k(n) \leq c_2\, n^{d/(2\nu+d)},$$

gdzie c_1, c_2 *są pewnymi stałymi oraz* $0 < c_1 \leq c_2 < \infty$, *to*

$$\sqrt{\mathrm{ISE}(n)} = O\left(\sqrt{\log\log(n)}\, n^{-\nu/(2\nu+d)}\right) \tag{7.28}$$

z prawdopodobieństwem 1. □

Wiadomo (patrz np. [36]), że wartość $J(\hat{g}_n) - J^*$ może być oszacowana od góry przez $\sqrt{\mathrm{ISE}(n)}$, toteż korzystając z lematu 7.3, otrzymujemy:

Twierdzenie 7.3.2 *Przy tych samych założeniach jak w lemacie 7.4 w przypadku reguły klasyfikacyjnej (7.16) zachodzi*

$$J(\hat{g}_n) - J^* = O\left(\sqrt{\log\log(n)}\, n^{-\nu/(2\nu+d)}\right) \tag{7.29}$$

z prawdopodobieństwem 1. □

Oszacowanie szybkości zbieżności dla funkcji regresji należących do klasy $\mathrm{Lip}(\nu)$ różni się od typowych wyników, uzyskanych na przykład w pracy [87], o mnożnik $\sqrt{\log\log(n)}$. Oszacowanie tempa zbieżności dla (7.26) i (7.28) przez $n^{-\nu/(2\nu+d)}$ jest najlepszym wynikiem otrzymywanym w zadaniach nieparametrycznej estymacji regresji w przypadku funkcji należących do klasy $\mathrm{Lip}(\nu)$ [183]. W szczególności, tempo zbieżności jest (asymptotycznie) tym większe, im większa jest wartość ν, co oznacza że odpowiednia funkcja regresji jest bardziej ,,gładka". Zasadnicza różnica pomiędzy typowym problemem estymacji regresji a rozważanymi tu funkcjami regresji transformowanymi na odcinek I_1 polega na tym, że w naszym przypadku wartość ν nie może być większa niż 1, niezależnie od tego, jak gładka jest α w przestrzeni I_d, gdzie gładkość wyrażana jest liczbą istniejących ciągłych pochodnych.

Nawet gdy α jest funkcją bardzo gładką (wiele razy różniczkowalną w I_d), to jednowymiarowa funkcja $h(t) = \alpha(F(t)) \in Lip(\gamma)$, $\gamma \leq 1/d$ nie jest różniczkowalna, mimo iż jest ciągła, ponieważ krzywa wypełniająca F nie jest różniczkowalna w całym swoim obszarze (patrz własność **C1**). Dokładniej mówiąc, złożenie funkcji $f(F(t))$ może być gładkie tylko wówczas, gdy $f(x)$ jest funkcją stałą. W ogólnym przypadku szybkość zbieżności nie zależy zatem od tego, ile razy α jest różniczkowalna i jest co najwyżej rzędu $O(n^{-1/(2+d)})$.

7.3.3 Rozpoznawanie za pomocą układu Haara

Obecnie zbadamy algorytm rozpoznawania 7.3 w przypadku użycia jako szeregu ortonormalnego układu funkcji Haara na odcinku I_1.

Układ funkcji Haara ma postać: dla $k = 1, 2, \ldots$, $m = \lceil \log_2 k \rceil - 1$, $j = k - 2^m$ i dla $t \in [0,1)$, $v_k(t) = v_m^{(j)}(t)$, $v_1(t) = v_0^0(t) \equiv 1$,

$$v_m^{(j)}(t) \stackrel{\triangle}{=} \begin{cases} \sqrt{2^m}, & t \in [(2j-2)/2^{m+1}, (2j-1)/2^{m+1}) \\ -\sqrt{2^m}, & t \in [(2j-1)/2^{m+1}, (2j)/2^{m+1}) \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

oraz dla $t = 1$

$$v_m^{(j)}(1) = \begin{cases} -\sqrt{2^m}, & j = 2^m,\ m = 0, 1, 2, \ldots, \\ 0, & j = 1, \ldots, 2^m - 1,\ m = 0, 1, 2, \ldots \end{cases}$$

Wiadomo (patrz [45], [46] i bibliografia cytowana tamże), że jądro układu Haara, czyli $K_k(s,t) = \sum_{r=1}^{k} v_r(t) v_r(s)$ może być przedstawione w postaci

$$K_k(s,t) = \begin{cases} \frac{1}{t^{(l)} - t^{(l-1)}}, & \text{gdy } s,t \in [t^{(l-1)}, t^{(l)}), \\ & l = 1, 2, \ldots, k-1, \\ \frac{1}{t^{(k)} - t^{(k-1)}}, & \text{gdy } s,t \in [t^{(k-1)}, t^{(k)}], \\ \\ 0, & \text{w przeciwnym przypadku,} \end{cases} \tag{7.30}$$

gdzie $t^{(l)}$ oznacza:

$$t^{(l)} = \begin{cases} l/2^{m+1} & \text{dla} \quad l = 0, 1, \ldots, 2j, \\ (l-j)/2^m & \text{dla} \quad l = 2j+1, \ldots, k. \end{cases} \tag{7.31}$$

Lemat 7.5 *Układ Haara spełnia warunek (7.14), czyli*

$$W_{k(n)} = \sup_{t \in I_1} \sum_{j=1}^{k(n)} v_j^2(t) \le V^2\, k(n).$$

□

Zauważmy, że dla dowolnego $t \in I_1$ oraz dla $2^m \le k(n) \le 2^{m+1}$ zachodzi $\sum_{j=1}^{k(n)} v_j^2(t) = K_{k(n)}(t,t) \le 2^{m+1} \le 2\, k(n)$. Stąd $W_{k(n)} \le 2\, k(n)$ i warunek (7.14) jest spełniony, przy czym $V^2 = 2$.

Z lematu 7.5 oraz z twierdzenia 7.3.1 wynika następujący wniosek:

Lemat 7.6 *Jeśli spełnione są wszystkie założenia twierdzenia 7.3.1 oraz $k(n)$ zostanie wybrany tak, że*

$$k(n) \to \infty \quad \textit{oraz} \quad \frac{k(n)\log(n)}{n} \to 0, \ \textit{gdy}\ n \to \infty, \tag{7.32}$$

to ciąg reguł klasyfikacji $\hat{g}_n$ zdefiniowany w (7.16) jest mocno zgodny, przy założeniu, że $\{v_k\}_{k=1}^{\infty}$ jest układem Haara. □

Korzystając z jądra układu Haara zapisanego w postaci (7.30) oraz z (7.6) i (7.8), możemy regułę decyzyjną przedstawić następująco:

$$\hat{\alpha}_n(x) = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)K_{k(n)}(\boldsymbol{\Psi}(X_i), \boldsymbol{\Psi}(x)). \tag{7.33}$$

Ze wzoru (7.30) wynika, że $K_{k(n)}$ jest różne od zera tylko w przypadku, gdy przeciwobrazy aktualnie rozpoznawanego wzorca $\boldsymbol{\Psi}(x)$ i co najmniej jednego wzorca pochodzącego z ciągu uczącego $\boldsymbol{\Psi}(X_i)$ należą do tego samego podprzedziału $[t^{(l-1)}, t^{(l)})$ (lub $[t^{(k(n)-1)}, t^{(k(n))}]$). Niech $n_0^{(l)}$ oraz $n_1^{(l)}$ oznacza odpowiednio liczbę obserwacji $\boldsymbol{\Psi}(X_i)$, $i = 1, 2, \ldots, n$ z klasy 0 oraz 1, które należą do przedziałów $\left[t^{(l-1)}, t^{(l)}\right)$, $l = 1, 2, \ldots, k(n) - 1$ lub $\left[t^{(k(n)-1)}, t^{(k(n))}\right]$. W konsekwencji otrzymujemy następujący prosty algorytm:

ALGORYTM ROZPOZNAWANIA ZA POMOCĄ UKŁADU HAARA NA ODCINKU I_1

Dla uproszczenia, algorytm podany poniżej będziemy dalej nazywać algorytmem Haara.

Krok 1. Przekształć ciąg uczący (X_i, Y_i) do postaci (t_i, Y_i), gdzie $t_i = \boldsymbol{\Psi}(X_i)$, $i = 1, 2, \ldots, n$.

Krok 2. Oblicz $n_0^{(l)}$, $n_1^{(l)}$, $l = 1, 2, \ldots, k(n)$.

Krok 3. Aby zaklasyfikować obserwację x, wyznacz przedział $S_x = \left[t^{(l_x-1)}, t^{(l_x)}\right)$ (lub $\left[t^{(l_x-1)}, t^{(l_x)}\right]$, gdy $l_x = k(n)$), którego końce określają punkty (7.31), przy czym $t = \boldsymbol{\Psi}(x) \in S_x$. Przydziel x do klasy 0, gdy $n_0^{(l_x)} \geq n_1^{(l_x)}$, w przeciwnym przypadku przydziel x do klasy 1. Jeśli obie liczby: $n_0^{(l_x)}$ i $n_1^{(l_x)}$ są sobie równe, to przydziel x do klasy 0 lub 1 losowo, z prawdopodobieństwami odpowiednio równymi oszacowaniom empirycznym prawdopodobieństw a priori: $\hat{p}_0 = n^{-1}\sum_{l=1}^{k(n)} n_0^l$ oraz $\hat{p}_1 = n^{-1}\sum_{l=1}^{k(n)} n_1^l$. □

Korzystając z (7.33) i (7.30), możemy regułę klasyfikacji (7.16) zapisać w równoważnej postaci

$$\hat{g}_n(x) = \begin{cases} 0, & \text{gdy} \quad \sum_{i=1}^{n} Y_i I(\boldsymbol{\Psi}(X_i) \in S_x) \\ & \quad \leq \sum_{i=1}^{n}(1 - Y_i) I(\boldsymbol{\Psi}(X_i) \in S_x), \\ 1, & \text{w przeciwnym przypadku,} \end{cases} \tag{7.34}$$

gdzie $I(A)$ funkcją wskaźnikową zbioru A.

Powyższy algorytm możemy zakwalifikować jako algorytm typu podziałowego (por. [36]).

Zauważmy, że $t^{(l)} - t^{(l-1)} = 2^{-m+1}$ lub 2^{-m}, gdzie $m = \lceil \log_2 k(n) \rceil - 1$ oraz $l = 1, \ldots, k(n)$. Łatwo zauważyć, że klasyfikator typu szereg Haara–Fouriera sprowadza się do algorytmu opartego na histogramie obserwacji pochodzących z ciągu uczącego, przy czym punkty $t^{(1)} = 0, \ldots, t^{(k(n))} = 1$ wyznaczają podprzedziały histogramu. Własności histogramów bazujące na falkach Haara (w zastosowaniu do estymacji gęstości) przeanalizowano w [46].

Korzystając z dowodu twierdzenia 9.4 w [36], możemy otrzymać następujące twierdzenie:

Twierdzenie 7.3.3 *Jeśli $k(n)$ zostało wybrane w taki sposób, że*

$$2/2^{\lceil \log_2 k(n) \rceil} \to 0 \quad oraz \quad \frac{n}{2^{\lceil \log_2 k(n) \rceil}} \to \infty, \quad gdy \ n \to \infty, \tag{7.35}$$

to ciąg reguł klasyfikacji $\hat{g}_n$ (7.16), w których $\{v_k\}_{k=1}^{\infty}$ jest układem Haara, jest zgodny, z prawdopodobieństwem 1, z regułą Bayesa przy dowolnym rozkładzie (X, Y) z nośnikiem w $I_d \times \{0, 1\}$. □

Zauważmy, że $2^{\lceil \log_2 k(n) \rceil} \geq 2^{\log_2 k(n)} = k(n)$ oraz $2^{\lceil \log_2 k(n) \rceil} \leq 2^{\lfloor \log_2 k(n) \rfloor + 1}$ $\leq 2^{\log_2 k(n)+1} = 2k(n)$. Warunki (7.35) są więc równoważne z warunkami

$$k(n) \to \infty \quad \text{oraz} \quad \frac{k(n)}{n} \to 0, \ \text{gdy} \ n \to \infty \tag{7.36}$$

otrzymanymi w [36](por. twierdzenie 17.2) w odniesieniu do uogólnionego liniowego klasyfikatora, ze współczynnikami wybranymi w taki sposób, że minimalizują one empiryczne ryzyko $1/n \sum_{i=1}^{n} I(\hat{g}_n(X_i) \neq Y_i)$. W przypadku dowolnego wyboru współczynników b_j, $j = 1, \ldots, k(n)$, jeśli $\{v_k\}_{k=1}^{\infty}$ jest szeregiem Haara, to wartość $\sum_{j=1}^{k(n)} b_j v_j(t)$ jest stała w przedziałach $\left[t^{(l-1)}, t^{(l)}\right)$ lub $\left[t^{(l-1)}, t^{(l)}\right]$, $l = 1, 2, \ldots, k(n)$. Nie jest więc niczym zaskakującym, że współczynniki $\hat{b_j}$, $j = 1, \ldots, k(n)$ określone zgodnie z (7.6) minimalizują błąd empiryczny.

Rezultat uzyskany w twierdzeniu 7.3.3 zasługuje na uwagę, gdyż wykazano w nim zbieżność, z prawdopodobieństwem 1, błędu klasyfikacji ciągu reguł (7.16) z układem Haara do minimalnej wartości ryzyka, bez żadnych dodatkowych założeń dotyczących rozkładów (X, Y), $(X, Y) \in I_d \times \{0, 1\}$.

7.3.4 Rezultaty badań symulacyjnych

Zastosowanie algorytmu Haara w odniesieniu do danego ciągu uczącego L_n prowadzi do podziału odcinka I_1 na pewną liczbę pododcinków, z którymi skojarzona jest decyzja o przynależności do określonej klasy. Oznaczmy liczbę tych

podprzedziałów przez c. Wartość $c \leq n$ jest sumą wybranych lokalnie wartości $k(n)$. Ten krok, konstruowanie klasyfikatora, choć obliczeniowo dość złożony, jest wykonywany tylko raz na etapie wstępnego przetwarzania danych. Sklasyfikowanie napływającej obserwacji x wymaga obliczenia $\Psi(x)$. Przy dokładności p cyfr znaczących jest to operacja o złożoności $O(p\,d)$. Znalezienie podprzedziału, który zawiera $\Psi(x)$ wymaga co najwyżej $O(\log(c))$ porównań (przy czym zwykle $c << n$). Ponadto operację wyznaczania $\Psi(x)$ można wykonywać częściowo równolegle z wyznaczaniem podprzedziału w I_1, który zawiera $\Psi(x)$. Takie postępowanie pozwala dopasować dokładność p do istniejącego podziału odcinka I_1. Innymi słowy, dokładność transformacji można ustalać lokalnie w zależności od gęstości danych jednowymiarowych (po transformacji na odcinek). Punkt obcięcia szeregu $k(n)$ też może być dobierany lokalnie w sposób zależny od liczby elementów ciągu uczącego, pojawiających się w różnych pododcinkach I_1.

W przedstawionych dalej przykładach pokażemy dużą efektywność algorytmu rozpoznawania za pomocą układu Haara na odcinku I_1 w zastosowaniu do standardowych problemów testowych stosowanych w rozpoznawaniu.

Przykłady *Iris* oraz *Muscular dystrophy* zostały zaczerpnięte z książki [3]. Jakość algorytmu rozpoznawania w odniesieniu do zbiorów danych rzeczywistych była oceniana za pomocą procedury typu cross-validation, w której każdy z wzorców pochodzących ze zbioru danych jest rozpoznawany przy użyciu klasyfikatora skonstruowanego na podstawie pozostałych danych użytych jako ciąg uczący [57]. Procedurę tę będziemy w skrócie oznaczać jako $CV1$. Ponieważ w przypadku skończonego zbioru danych nie znamy ryzyka Bayesa, jako podstawę do porównań przyjęto więc błąd rozpoznawania uzyskiwany metodą k–NN (k najbliższych sąsiadów) [29], [41].

W pierwszych dwu przykładach dane były symulowane z wielowymiarowych rozkładów normalnych, które mają nieograniczone nośniki. Nie korzystano jednak z nieliniowej transformacji danych (na przykład za pomocą funkcji logistycznej), a jedynie dokonano liniowej transformacji wygenerowanych wcześniej danych, tak by po przekształceniu zawierały się w kostce I_d. Jeśli obserwacja x pochodzi (po liniowej transformacji) spoza ograniczonego obszaru I_d, to $\Psi(x)$ nie istnieje i x nie może zostać sklasyfikowane. Można również interpretować tę sytuację jako badanie przykładu, w którym dane pochodzą nie z nieograniczonego rozkładu normalnego, lecz rozkładu normalnego obciętego.

Przykład 7.1 *Klasy o rozkładach normalnych [57]*

Przykład ten został zaczerpnięty z książki Fukunagi [57]. Dane wygenerowano z dwu rozkładów normalnych w przestrzeni 8-wymiarowej. Prawdopodobieństwa a priori obu klas były równe 1/2. Rozkład cech w klasie pierwszej charakteryzował wektor wartości średnich: $[0, 0, \ldots, 0]$ i macierz kowariancji diag$[1, \ldots, 1]$.

Dane z drugiej klasy miały także rozkład normalny z wektorem wartości średnich równym

$$[0,01,\ 0,26,\ 1,08,\ 1,64,\ 0,84,\ 0,84,\ 3,1,\ 3,86]$$

i macierzą kowariancji

$$\mathrm{diag}[1,65^2,\ 0,56^2,\ 1,33^2,\ 1,22^2,\ 0,469^2,\ 0,346^2,\ 3,47^2,\ 2,9^2].$$

Wartość ryzyka Bayesa wyliczona teoretycznie jest równa $1,9\%$ [57].

Testowanie algorytmu Haara składało się z 10 przebiegów algorytmu uczenia (konstruowania klasyfikatora) wykonanych dla 10 różnych, niezależnie wygenerowanych ciągów uczących o długości $n = 100,\ n = 200,\ , \ldots , n = 500$. Każdy z otrzymanych klasyfikatorów był testowany za pomocą odrębnego ciągu testującego zawierającego 1000 (lub 2000) elementów. Wyniki testowania przedstawiono na rysunkach 7.1 i 7.2. Pojedynczy punkt na tym rysunku odpowiada wyrażonej w procentach wartości błędnych klasyfikacji obliczonych (i uśrednionych) dla 10 różnych realizacji.

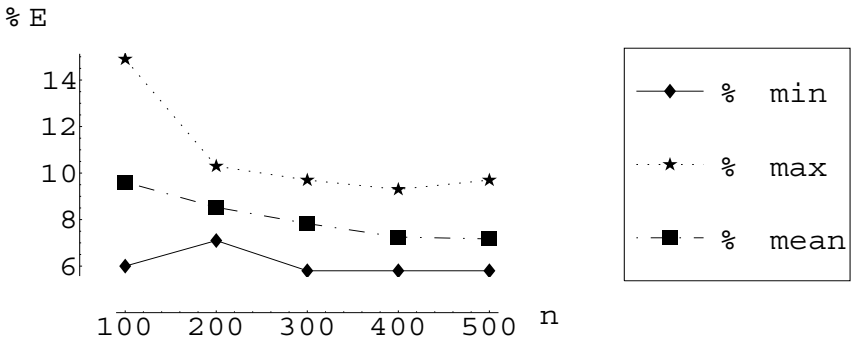
Na rysunkach tych zaznaczono też minimalną i maksymalną wartość błędów uzyskanych dla 10 różnych ciągów uczących. W każdym z przypadków jako narzędzia transformacji do 1–D użyto quasi-odwrotności krzywej Sierpińskiego obliczanej z dokładnością 2^{-96}. Błędy te były zauważalnie większe w przypadku użycia krzywej Peano i krzywej Hilberta.

Rysunek 7.1 przedstawia zależność empirycznego błędu klasyfikacji od długości ciągu uczącego. Na rysunku 7.2 zilustrowano wpływ dokładności transformowania danych za pomocą krzywej (liczby cyfr znaczących w rozwinięciu dwójkowym) na wielkość błędu klasyfikacji. Z rysunku tego wynika, że algorytmy klasyfikacji bazujące na transformacji danych za pomocą krzywych wypełniających nie są zbyt wrażliwe na wybór dokładności przekształcenia. Kiedy dokładność osiągnie pewien poziom, związany z gęstością wypełnienia przestrzeni przez dane z ciągu uczącego, dalszy wzrost dokładności (zwiększanie parametru p) nie zmniejsza liczby błędnych klasyfikacji.

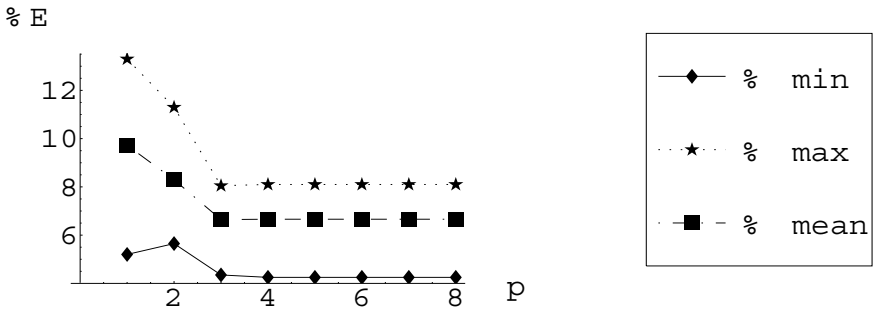
Średnie błędy klasyfikacji są nieco większe niż odpowiednie wyniki otrzymane metodą k najbliższych sąsiadów (k–NN), która pozwala na uzyskanie błędów rzędu 5% (patrz [57] s. 312).

Przykład 7.2 *Dane o wielowymiarowym rozkładzie normalnym – klasy rozdzielone*

Przykład ten został zaczerpnięty z pracy [53]. Występują w nim wzorce z przestrzeni o $d = 10$ wymiarach, należące do $M = 2$ klas. Dane te mają łączny rozkład będący rozkładem standardowym normalnym. Element x należy do klasy



Rys. 7.1. Maksymalne, minimalne i średnie (dla 10 różnych ciągów uczących) wartości procentowych błędów klasyfikacji otrzymane w wyniku działania algorytmu Haara (zbiór testujący składał się z 1000 elementów) w zależności od długości ciągu uczącego n wylosowanego dla dwu klas o rozkładach normalnych (por. przykład 7.1)
Fig. 7.1. Minimal, maximal and averaged (over 10 learning sequences) percentage of misclassifications obtained by Haar Algorithm (1000 testing patterns) versus the length n of the learning sequence simulated from two normally distributed classes (see Example 7.1 for details)



Rys. 7.2. Maksymalne, minimalne i średnie (dla 10 różnych ciągów uczących) wartości procentowych błędów klasyfikacji otrzymane w wyniku działania algorytmu Haara (zbiór testujący składał się z 2000 elementów, $n = 1000$) w zależności od wartości parametru p, który odpowiada dokładności aproksymacji krzywej Sierpińskiego. Dane były losowane z dwu rozkładów normalnych (patrz szczegóły opisane w przykładzie 7.1)
Fig. 7.2. Minimal, maximal and averaged (over 10 learning sequences) percentage of misclassifications obtained by Haar Algorithm (2000 testing patterns, $n = 1000$) versus parameter p, which reflects the degree of approximation of the Sierpiński curve. Data were simulated from two normally distributed classes (see Example 7.1 for details)

o numerze 0, gdy $\sum_{i=1}^{d} x_i^2 \leq 9,8$, a w przeciwnym przypadku x należy do klasy o numerze 1.

Błąd Bayesa jest w tym przykładzie równy zeru. Z drugiej strony, wiadomo [53], że jest to problem „trudny" z punktu widzenia budowy klasyfikatora na podstawie ciągu uczącego, niezależnie od obranej metody. Przy założeniu, że ciąg uczący składa się z $N = 500$ elementów, klasyczna metoda k najbliższych sąsiadów prowadzi do błędów rozpoznawania rzędu 34%[53].

Podobne wyniki daje także badany algorytm Haara zastosowany w odniesieniu do 500–elementowego ciągu uczącego przetransformowanego na odcinek za pomocą quasi-odwrotności krzywej Peano.

Przykład 7.3 *Zbiór danych Iris*

Zbiór danych *Iris* jest bardzo popularnym zbiorem testowym [3], [51]. Zawiera 150 wzorców scharakteryzowanych przez cztery pomiary (szerokości i długości dwu typów płatków) trzech różnych gatunków irysów (*setosa, versicolor* i *virginica*), po 50 przykładów dla każdego z gatunków.

Dane dotyczące jednego z gatunków (*Iris setosa*) są wyraźnie odseparowane od pozostałych 100 wzorców i dlatego często bywają odrzucane [53], a badania prowadzi się na zbiorze zawierającym pozostałe 100 danych z dwu klas. Tak też postąpiono w tym przykładzie.

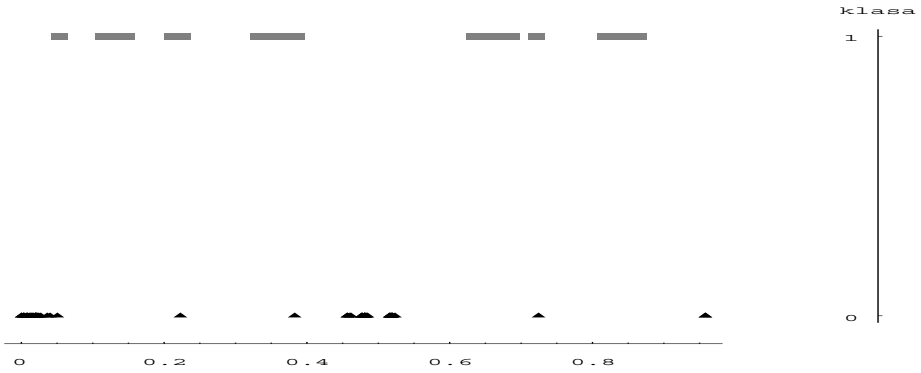
4–wymiarowe dane pochodzące z pozostałych dwu klas zostały najpierw przeskalowane do kostki $[0,1]^4$, a następnie transformowane za pomocą krzywej Hilberta z dokładnością gwarantującą podział kostki I_4 na 2^{40} równych podkostek.

Przetransformowany ciąg uczący przedstawiono na rysunku 7.3, w którym dane z różnych klas są zobrazowane w postaci funkcyjnej: dane należące do jednej klasy jako punkty odpowiadające wartości 0, natomiast dane z drugiej klasy jako punkty odpowiadające wartości 1.

Pierwsze 24 punkty z ciągu uczącego zamieszczone są w tabeli 7.6. Porównanie zawartości tej tabeli i rysunku 7.3 jasno wskazuje, że istnieją przedziały w I_1, w których obserwacje z obu klas są przemieszane i nie sposób uniknąć błędnej klasyfikacji.

Zastosowanie algorytmu Haara i krzywej Hilberta w odniesieniu do zbioru testowego *Iris* prowadzi do błędów klasyfikacji rzędu 7% (4,33% w przypadku uwzględnienia trzech klas). Zastąpienie krzywej Hilberta przez krzywą Peano pozwala na zmniejszenie poziomu błędnych klasyfikacji do 6%. Dla porównania, metoda k–NN w odniesieniu do danych 4–wymiarowych prowadzi do błędów rzędu 8% [53].

Wyniki obliczeń zebrano w tabeli 7.1, w której uwzględniono również zastosowane uporządkowanie cech.



Rys. 7.3. Zbiór danych *Irys* zawierający tylko dwie klasy po transformacji przy użyciu quasi-odwrotności krzywej Hilberta (szczegóły opisane w przykładzie 7.3)
Fig. 7.3. *Iris* data from two classes after transformation by the quasi-inverse of the Hilbert curve (see Example 7.3 for details)

Tabela 7.1. Błędy klasyfikacji uzyskane przy użyciu algorytmu Haara oraz krzywych Hilberta i Peano w odniesieniu do danych *Iris* (patrz opis przykładu 7.3). Dla porównania podano błędy uzyskiwane metodą k–NN

Krzywa i najlepsza kolejność cech	% Błędnych klasyfikacji	
	algorytm Haara	k–NN
Hilbert 1,2,3,4	7%	8% ⋆
Peano 1,2,3,4	6%	
⋆ Źródło [53]		

Przykład 7.4 *Zbiór danych Muscular dystrophy*

Dane *Muscular dystrophy* [3] składają się z 193 5–wymiarowych wektorów danych, które odpowiadają wiekowi badanych 193 osób oraz pomiarom poziomu czterech markerów mierzonych w surowicy krwi. Wśród badanych były osoby zdrowe oraz osoby będące nosicielami *Duchenne Muscular dystrophy* – choroby przekazywanej genetycznie.

Eksperymenty obliczeniowe przeprowadzono na pełnym zbiorze danych w przeciwieństwie do badań prowadzonych przez Friedmana [53], w których użyto zbiór edytowany, dotyczący 190 osób, prowadzący na ogół do mniejszych nieco wartości błędów. W przypadku klasycznej metody k–NN były to, odpowiednio, błędy

Tabela 7.2. Błędy klasyfikacji uzyskane za pomocą algorytmu Haara w odniesieniu do danych *Muscular dystrophy* (patrz opis przykładu 7.4)

Krzywa i najlepsza kolejność cech	% Błędnych klasyfikacji	
	algorytm Haara	k–NN
Hilbert 1,4,3,2,5	9,8%	11,4% *
Sierpiński mod. 1,2,3,4,5	10,3%	
* Otrzymane dla całego ciągu uczącego zawierającego 193 wzorce (bez edycji ciągu).		

klasyfikacji rzędu $11,4\%$ (dla pełnego ciągu uczącego) oraz $10,5\%$ (w przypadku ciągu edytowanego). Metoda testowania była tu taka sama jak w przypadku zbioru *Iris*. Otrzymano błędy klasyfikacji rzędu $9,8\%$ oraz $10,3\%$ w przypadku zastosowania odpowiednio krzywej Hilberta i krzywej Sierpińskiego. Podsumowanie wyników zamieszczono w tabeli 7.2.

Podsumowując, algorytm rozpoznawania oparty na rozwinięciu jednowymiarowych (przetransformowanych za pomocą krzywej wypełniającej) danych w szereg Haara, pozwala na otrzymanie szybkiego i łatwego w stosowaniu klasyfikatora, którego błąd klasyfikacji jest porównywalny z wymagającą dużych nakładów obliczeniowych metodą k najbliższych sąsiadów z metryką w przestrzeni d–wymiarowej. Jednocześnie należy pamiętać, że przedstawiany tu algorytm, w końcowym rezultacie, nie wymaga pamiętania całego ciągu uczącego, ani nawet obliczonych współczynników rozwinięcia w szereg Haara. Pamiętać należy jedynie punkty odcinka I_1, w których następuje zmiana przynależności do klas. Punktów takich, w praktyce, jest znacznie mniej niż elementów ciągu uczącego.

7.4 Szybki algorytm k najbliższych sąsiadów

Metoda k najbliższych sąsiadów (k–NN)[29], [41], [57] [36] jest jedną z najbardziej popularnych metod klasyfikacji. Polega ona na wyznaczeniu k najbliższych sąsiadów klasyfikowanej obserwacji x wśród elementów z ciągu uczącego, a następnie zaklasyfikowaniu x do tej klasy, do której należy większość znalezionych sąsiadów. Najczęściej stosowaną miarą odległości jest metryka euklidesowa. Mimo rozwoju wielu innych, bardzo wyrafinowanych technik nieparametrycznych, metoda k–NN jest jedną z najbardziej efektywnych i popularnych technik w rozpoznawniu [53]. Reguła k–NN zawdzięcza swą popularność nie tylko intuicyjnej prostocie sfor-

mułowania, lecz również dobrze zbadanym własnościom asymptotycznym [36]. Odnotujmy dla porządku, że reguła k–NN z ustaloną z góry liczbą sąsiadów nie zapewnia zbieżności $E[J_n]$, przy $n \to \infty$, do ryzyka Bayesa $J^\star$.

Proste implementacje metody k–NN prowadzą w przypadku wielowymiarowym do czasochłonnych obliczeń, a czas rozpoznawnia pojedynczego wzorca rośnie wraz ze wzrostem liczby sąsiadów k, długością ciągu uczącego n oraz wymiarem przestrzeni cech d. Także wybór odpowiedniej metryki [53] ma wpływ na efektywność metody klasyfikacji, zarówno ze względu na złożoność obliczeniową procesu rozpoznawania, jak i z uwagi na uzyskiwane błędy rozpoznawania (dopasowanie metryki do struktury danych). Szereg prac poświęcono efektywnym metodom wyznaczania k najbliższych sąsiadów w przestrzeni d–wymiarowej [20], [79], [49], [193], [34], [207], [39].

W niniejszym rozdziale zbadamy efektywność metody k–NN zastosowanej w odniesieniu do danych przetransformowanych na odcinek I_1 za pomocą krzywej wypełniającej. Metodę tę możemy interpretować jako wprowadzenie nowej metryki, która definiuje odległość między punktami przestrzeni wielowymiarowej poprzez odległość między odpowiadającymi im punktami na odcinku I_1.

Metryka ta ma postać:

$$d(x, x^{'}) = |\boldsymbol{\Psi}(x) - \boldsymbol{\Psi}(x^{'})|, \quad x, x^{'} \in I_d. \tag{7.37}$$

Łatwo sprawdzić, że $d(\cdot)$ spełnia wszystkie wymagane własności, to znaczy:
$d(x, x^{'}) \geq 0$,
$d(x, x^{'}) = 0$ wtedy i tylko wtedy, gdy $x = x^{'}$,
$d(x, x^{'}) = d(x^{'}, x)$,
$d(x, x^{'}) + d(x^{'}, x^{''}) \leq d(x, x^{''})$.

7.4.1 Algorytm k–CNN

Algorytm rozpoznawania (oznaczany dalej w skrócie k–CNN) polega na wyborze tej klasy, do której należy większość spośród k najbliższych (ze względu na położenie na krzywej wypełniającej) sąsiadów rozpoznawanego obiektu x.

Konstrukcja klasyfikatora, podobnie jak w przypadku przedstawionego wcześniej algorytmu Haara, wymaga przetransformowania całego ciągu uczącego $L_n = \{(X_i, Y_i),\ i = 1, 2, \ldots, n\}$ za pomocą wybranego odwzorowania quasi-odwrotnego $\boldsymbol{\Psi}$. Następnie nowy ciąg uczący $L_n^{\boldsymbol{\Psi}} = \{(t_i, Y_i),\ t_i = \boldsymbol{\Psi}(X_i),\ i = 1, \ldots, n\}$ należy uporządkować zgodnie z niemalejącymi wartościami t_i. Dla uproszczenia dalej będziemy zakładać, że ciąg uczący jest ponumerowany zgodnie z kolejnością położenia punktów na krzywej wypełniającej, czyli że $t_{(i-1)} \leq t_{(i)},\ i = 2, 3, \ldots, n$.

ALGORYTM

Niech $L_n^{\Psi} = ((t_{(1)}, Y_{(1)}), (t_{(2)}, Y_{(2)}), \ldots, (t_{(n)}, Y_{(n)}))$.

Krok 1. Aby zaklasyfikować obserwację x, oblicz wartość $t = \Psi(x)$.

Krok 2. Wyznacz zbiór

$$S = \{(t_{(r)}, Y_{(r)}), \ldots, (t_{(r+k-1)}, Y_{(r+k-1)})\},$$

zawierający k elementów ciągu uczącego o najmniejszej odległości od t, wyznaczonej jako $|t_i - t|$ (lub na podstawie innej metryki określonej na odcinku I_1).

Krok 3. Wyznacz k_i, $i = 0, 1, \ldots, M$, gdzie k_i oznacza liczbę sąsiadów ze zbioru S należących do i-tej klasy. W ogólnym przypadku $k = k_0 + k_1 + \cdots + k_{M-1}$, gdzie M jest liczbą klas.

Krok 4. Zaklasyfikuj x do klasy, która ma największą liczbę reprezentantów w zbiorze S, czyli jeśli

$$k_i = \max\{k_0, k_1, \ldots, k_{M-1}\},$$

to x jest klasyfikowane do klasy i-tej.

Jeżeli nastąpi poszerzenie ciągu uczącego o dodatkowy wzorzec, to aktualizacja zbioru L_n^{Ψ} polega na dołożeniu nowego elementu w taki sposób, by zbiór rozszerzony był dalej uporządkowany zgodnie z położeniami jego elementów na odcinku I_1. Wymaga to, poza transformacją nowego wzorca na odcinek I_1, jedynie $O(\log_2 n)$ operacji arytmetycznych.

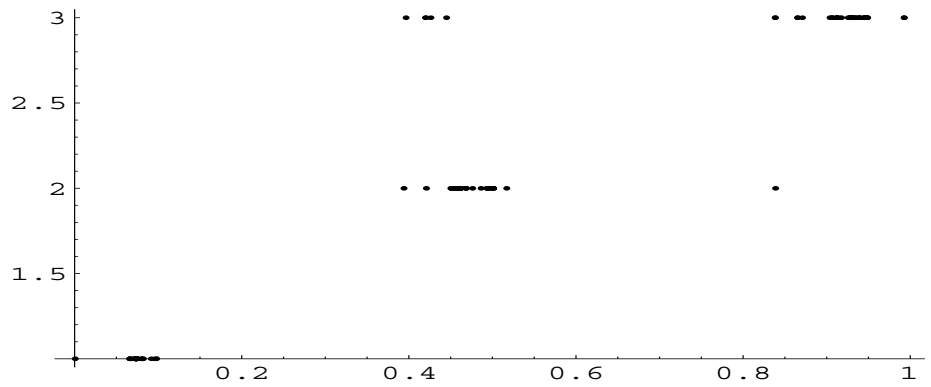
7.4.2 Kompresja danych w metodzie k–CNN

Reguła k najbliższych sąsiadów w przypadku danych jednowymiarowych, postaci: $(t_{(i)} = \Psi(X_{(i)}), Y_{(i)}) = (t_{(i)}, Y_{(i)})$, $i = 1, 2, \ldots, n$, prowadzi do podziału odcinka jednostkowego na $n + 1$ pododcinków:

$$[0, (t_{(1)} + t_{(k+1)})/2], [(t_{(1)} + t_{(k+1)})/2, (t_{(2)} + t_{(k+2)})/2], \ldots, [(t_{(n-k)} + t_{(n)})/2, 1],$$

w których decyzję o przynależności do klasy określa wybór najczęstszej klasy, odpowiednio w zbiorach

$$\{Y_{(1)}, \ldots, Y_{(k)}\}, \{Y_{(2)}, \ldots, Y_{(k+1)}\}, \ldots, \{Y_{(n-k+1)}, \ldots, Y_{(n)}\}.$$



Rys. 7.4. Zbiór danych *Iris* zawierający trzy klasy po transformacji przy użyciu quasi-odwrotności krzywej Peano (szczegóły opisane w przykładzie 7.3)
Fig. 7.4. *Iris* data from three classes after transformation by the quasi-inverse of the Peano curve (see Example 7.3 for details)

Zwykle wiele z sąsiadujących podprzedziałów ma przyporządkowany ten sam numer klasy, w związku z tym można je połączyć (skompresować).

Gdy zakładamy, że ciąg uczący pozostanie bez zmian (klasyfikator nie będzie modyfikowany), wtedy zamiast ciągu uczącego L_n^{Ψ} wystarczy zapamiętać położenie punktów na odcinku, w których następuje zmiana przypisywanej przynależności do klasy.

W konsekwencji, na czas potrzebny do rozpoznania nowego obiektu $x \in I_d$ składa się czas transformacji $\Psi(x)$ oraz czas potrzebny na znalezienie podprzedziału, do którego należy $\Psi(x)$, czyli sumarycznie jest to czas rzędu

$$O(d) + O(\log_2(m)),$$

gdzie m jest efektywną liczbą podprzedziałów odcinka, które są związane z różnymi decyzjami o przynależności do klasy.

Przykład 7.5 *Zbiór Iris*

Zbiór *Iris* zawiera 150 wzorców scharakteryzowanych przez cztery pomiary (szerokości i długości dwu typów płatków) trzech różnych gatunków irysów: *setosa*, *versicolor* i *virginica*, po 50 przykładów dla każdego z gatunków. W przeciwieństwie do przykładu (7.3) obliczenia przeprowadzono dla całego ciągu uczącego zawierającego elementy należące do wszystkich trzech klas.

Dane (po przeskalowaniu do kostki I_4) przetransformowano na odcinek I_1 za pomocą krzywej Peano (patrz rysunek 7.4). Badana metoda klasyfikacji prowadzi

Tabela 7.3. Błędy klasyfikacji dla danych *Iris* uzyskane metodą 1–CNN i 5–CNN

1–CNN	5–CNN	k–NN	krzywa
4,0	2,66	5,33	Peano
6,0	6,66	5,33	Sierpiński

do reguły decyzyjnej składającej się z 3–5 przedziałów (zależnie od liczby sąsiadów $k = 2, 3$). Liczba k została określona zgodnie z procedurą cross-validation CV1, w której 150 danych zostało w sposób losowy podzielonych na zbiór uczący zawierający 75 elementów i zbiór testujący również 75–elementowy.

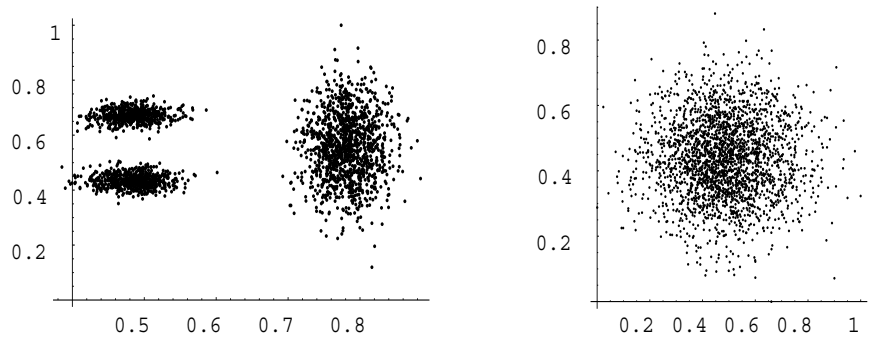
Błąd klasyfikacji (uśredniony na podstawie 10 różnych podziałów zbioru danych) wynosił 2,27 %, przy czym wartości błędu przed uśrednieniem zawierały się w przedziale $[0\% - 4,0\%]$. Łatwo zauważyć, że po transformacji danych *Iris*, klasa *setosa* jest nadal ściśle oddzielona w sensie geometrycznym od pozostałych klas. Na rysunku 7.4 jest to klasa oznaczona etykietą 1. W tabeli 7.3 przedstawiono dalsze wyniki badań, tym razem weryfikowane za pomocą procedury CV1, w której każdy element ze zbioru danych jest elementem testującym względem klasyfikatora zbudowanego na podstawie pozostałych $n - 1$ wzorców, traktowanych jako ciąg uczący. Względna liczba błędnych klasyfikacji służy jako estymata błędu metody klasyfikacji. W tabeli 7.3 podano także odpowiednie wyniki otrzymane przy zastosowaniu krzywej Sierpińskiego.

Przykład 7.6 *Clouds – dane symulowane*

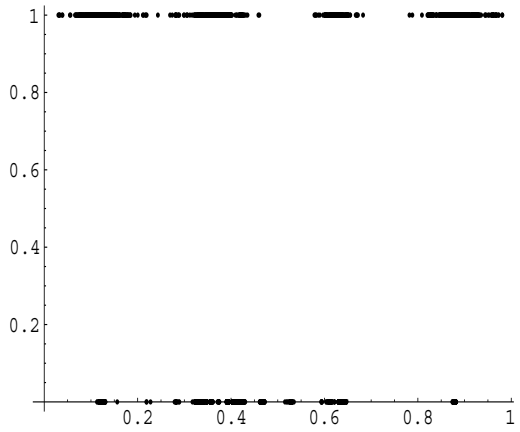
Zbiór danych *Clouds* składa się z $n = 5000$ dwuwymiarowych obserwacji należących do dwu klas z takimi samymi prawdopodobieństwami a priori. Dane te pochodzą z bazy danych Elena (patrz ftp://ftp.ucl.ac.be/pub/neural-nets...A/-databases/ARTIFIcial/clouds/clouds.txt). Dane z pierwszej klasy mają rozkład będący mieszaniną trzech różnych rozkładów normalnych:

$N[(0,0),\ \mathrm{diag}(0,2,\ 0,2)]$, $N[(0,2),\ \mathrm{diag}(0,2,\ 0,2)]$, $N[(2,1),\ \mathrm{diag}(0,2,\ 1)]$ z prawdopodobieństwami (wagami) odpowiednio równymi 1/4, 1/4, 1/2. Dane z drugiej klasy mają rozkład normalny $N[0,0),\ \mathrm{diag}(1,1)]$. Zbiór danych *Clouds*, po przeskalowaniu do I_2, przedstawia rysunek 7.5. Do transformacji danych (po wstępnym przeskalowaniu do kostki I_2) użyto krzywej Sierpińskiego. Efekty tej transformacji przedstawiono na rysunku 7.6.

Najmniejsze wartości błędu klasyfikacji otrzymano (procedura CV1) przy użyciu całego zbioru danych $n = 5000$. Błąd ten wynosił (metoda k–CNN, $k = 11$) około 11,32%. Dla porównania, przy zastosowaniu klasycznej metody k–NN na danych dwuwymiarowych otrzymano 10,94% błędnych klasyfikacji.



Rys. 7.5. Dane symulowane: klasa 0 – lewa strona rysunku, klasa 1 – prawa strona rysunku
Fig. 7.5. Simulated data: class 0 – left panel, class 1 – right panel



Rys. 7.6. Dane symulowane *Clouds* po transformacji do 1-D za pomocą krzywej Sierpińskiego
Fig. 7.6. Data *Clouds* after transformation to 1-D via the Sierpiński curve

Tabela 7.4. % błędnych klasyfikacji, dyspersja błędu σ, liczba przedziałów jednowymiarowej funkcji decyzyjnej m i stopień kompresji n/m uzyskany przy różnych długościach ciągu uczącego n

n	5000	4000	3000	2500	2000	1500	1000	500
błąd %	11,32	11,58	11,29	11,69	11,62	12,05	12,81	13,43
σ		1,04	0,62	0,44	0,66	0,43	0,64	0,60
m	63	63	54	39	33	42	33	20
n/m	79	63	55	63	61	45	30	25

Oszacowania prawdopodobieństw błędnej klasyfikacji przy różnych długościach ciągu uczącego (n =4000, 3000, 2500, 2000, 1500, 1000, 500) zamieszczono w tabeli 7.4. W każdym przypadku (dla ustalonego n) wygenerowano 10 niezależnych ciągów uczących, dla których wyniki testowano na pozostałych $5000 - n$ elementach zbioru *Clouds*. W każdym z przypadków podano także średnią wartość liczby przedziałów funkcji decyzyjnej o stałych wartościach na odcinku. Część prezentowanych powyżej wyników umieszczono w pracy autorki [172]. Szczegółowe wyniki eksperymentów przeprowadzonych na testowych zbiorach danych, zaczerpniętych z [3] oraz [53], są także zamieszczone w pracach autorki [174], [160]. Z eksperymentów tych wynika, że metoda k–CNN pozwala na uzyskiwanie podobnych, a często nawet mniejszych błędów klasyfikacji, niż w przypadku metody k–NN realizowanej w przestrzeni wielowymiarowej, przy równoczesnym znacznym stopniu kompresji funkcji decyzyjnej i skróceniu czasu klasyfikacji.

Podobne eksperymenty przeprowadzane na danych symulowanych opisano w pracy autorki [173]. Na przykład, w odniesieniu do zbioru danych *Muscular dystrophy* (por. przykład 7.4, podrozdział 7.3.4) uzyskano, przy zastosowaniu krzywej Sierpińskiego, błędy klasyfikacji rzędu 11,4%, czyli takie same jak otrzymywane metodą k–NN w odniesieniu do oryginalnych, wielowymiarowych danych.

W przypadku danych symulowanych z przykładu 7.2 metoda k–CNN prowadzi natomiast do znacznej poprawy, czyli do błędów klasyfikacji rzędu 28% (dla krzywej Peano) w porównaniu z 34% otrzymywanymi innymi metodami (por. też wyniki przedstawione w podrozdziale 7.3.4).

7.5 Wektorowa kwantyzacja w rozpoznawaniu

Algorytmy rozpoznawania, które były proponowane i badane w poprzednich podrozdziałach miały klarowne uzasadnienia teoretyczne. Celem niniejszego podrozdziału jest wskazanie, że metodyka stosowania krzywych wypełniających może być użyteczna także w udoskonalaniu takich metod rozpoznawania, które nie doczekały się pełnego opracowania teoretycznego, lecz ich praktyczna użyteczność jest w literaturze doceniana [83], [84], [78].

Metoda LVQ (*learning vector quantization*) – procedura skojarzonego uczenia typu samoorganizujących odwzorowań Kohonena – stosowana w odniesieniu do obiektów pochodzących z różnych klas, zaproponowana przez Kohonena [80], [82], [76], [176], [77], [78], jest metodą tworzenia zbiorów obiektów najlepiej reprezentujących rozkłady danych w klasach. Zbiory te nazywane są prototypami, zbiorami wektorów odniesienia bądź książkami kodowymi (kwantyzatorami).

W przestrzeni danych wejściowych, którą tu będziemy utożsamiać z przestrzenią cech, należy wybierać skończony zbiór reprezentantów, z których każdy skoja-

rzony jest jednoznacznie z numerem pewnej klasy. Na podstawie ciągu uczącego L_n, w trakcie procedury uczenia, tworzony jest nowy ciąg wzorców, który dalej może być traktowany jako zbiór odniesienia używany następnie do klasyfikacji metodą najbliższego sąsiada 1–NN.

Proces uczenia LVQ pełni tu dwie funkcje: służy kondensacji zbioru uczącego oraz edycji tego zbioru, przy czym operacje te należy rozumieć nieco szerzej niż klasyczne pojęcia edycji i kondensacji stosowane w odniesieniu do reguły najbliższego sąsiada [57], [36].

W trakcie uczenia sieci LVQ, na podstawie ciągu uczącego tworzy się nowy ciąg prototypów, na ogół znacznie mniej liczny niż ciąg uczący. Sieć taka różni się od sieci samoorganizujących SOM (omawianych w rozdziale 6) brakiem struktury topologicznej. Poza tym neurony zaopatrzone są w etykiety określające przynależność danego neuronu do konkretnej klasy.

Po procesie uczenia sieć LVQ działa podobnie jak sieć SOM. To znaczy, sieć LVQ wybiera taki prototyp (neuron), który jest najbliższym sąsiadem prezentowanego wzorca, wektora x względem przyjętej metryki. Jako wynik klasyfikacji podawany jest numer klasy, który jest etykietą wybranego prototypu.

Jak widać, sieć LVQ działa zgodnie z regułą najbliższego sąsiada, stosowaną jednakże nie w odniesieniu do całego ciągu uczącego, lecz w stosunku do ustalonego zbioru prototypów. W procesie uczenia sieci LVQ położenie wybranych prototypów jest korygowane w taki sposób, by minimalizować empiryczny błąd klasyfikacji, obliczany względem ciągu uczącego.

Znane są różne warianty algorytmów uczenia LVQ, które mają mniej lub bardziej heurystyczny charakter. Algorytmy te, znane jako algorytmy LVQ1, LVQ2, LVQ2.1, LVQ2.2, LVQ3, OLVQ1 [83] [76] oraz ich modyfikacje [38], [33], dokonują równocześnie zarówno kwantyzacji danych wejściowych (przestrzeni cech), jak i ich klasyfikacji.

Wprawdzie żaden ze wspomnianych algorytmów (w przypadku wielowymiarowym) nie gwarantuje zbieżności do ryzyka bayesowskiego, są one jednak bardzo użyteczne ze względu na prostotę samego algorytmu klasyfikacji i znaczną kompresję zbioru danych (ciągu uczącego).

7.5.1 Algorytmy LVQ z użyciem krzywej wypełniającej

W podejściu tutaj rozważanym każdy d–wymiarowy wektor x jest transformowany do punktu $\Psi(x)$ na odcinku jednostkowym I_1. W związku z tym zbiór prototypów V jest zbiorem liczb z I_1 oznaczonych numerem klasy, do której są przypisane:

$$\begin{gathered} V = ((v_1, c_1), (v_2, c_2), \ldots, (v_N, c_N)) \\ v_i \in I_1, \;\; c_i \in \{0, 1, \ldots, M-1\}, \;\; i = 1, 2, \ldots, N, \end{gathered} \tag{7.38}$$

gdzie N oznacza liczbę prototypów, a M liczbę klas. W przypadku, gdy mamy do czynienia z dychotomią, $M = 2$, etykiety przynależności do klasy przyjmują wartości ze zbioru $\{0, 1\}$. Jeżeli dane wejściowe są jednowymiarowe, skalarne, to algorytm LVQ, niezależnie od jego konkretnej postaci, pozwala na uzyskanie bardzo prostej sieci klasyfikującej, o relatywnie małej liczbie prototypów N.

Algorytm uczenia LVQ w wersji ze zmiennymi (optymalizowanymi) współczynnikami uczenia – OLVQ1 [83] – połączony z transformacją wielowymiarowych danych, przytaczamy za pracą autorki [167]. Ma on następującą postać:

ALGORYTM OLSQ1 – wersja skalarna OLVQ1

W poniższym algorytmie l jest numerem iteracji.

Krok 1. Przekształć obserwacje (X_j, Y_j), do postaci jednowymiarowej (t_j, Y_j), $t_j = \boldsymbol{\Psi}(X_j)$, $j = 1, 2, \ldots, n$. Ustal początkowe wartości

$$V = ((v_1(0), c_1), (v_2(0), c_2), \ldots, (v_N(0), c_N))$$

oraz indywidualne współczynniki uczenia α_i, $i = 1, \ldots, N$.

Krok 2. Wybierz losowo element z ciągu uczącego $t = t_i = \boldsymbol{\Psi}(X_i)$ $c = Y_i$. Znajdź jednowymiarowy prototyp v_k, najbliższy rozpatrywanej obserwacji t, czyli $|v_k(l) - t| \leq |v_i(l) - t|$, $i = 1, \ldots, N$.

Krok 3. Zmień pozycję wybranego v_k oraz jego współczynnik uczenia zgodnie z regułą:

$$v_k(l+1) = v_k(l) + s(l)\,\alpha_k(l)\,(t - v_k(l)),$$

$$\alpha_k(l+1) = \alpha_k(l)/(1 + s(l)\,\alpha_k(l)),$$

gdzie $s(l) = 1$, jeśli klasa wylosowanego obiektu c jest zgodna z klasą wybranego prototypu $c = c_k$ oraz $s(l) = -1$ w przeciwnym przypadku $(c \neq c_k)$.

ALGORYTM LSQ2 – wersja skalarna LVQ2

W algorytmie LVQ2.1 modyfikowane są pozycje dwu najbliższych sąsiadów, oznaczmy je przez v_k i v_r, jeśli jeden z nich należy do tej samej klasy co aktualny wzorzec, a drugi należy do innej klasy, przy czym nie jest istotne, który z nich jest najbliższym sąsiadem. Ponadto t powinno znajdować się w przedziale $O = [v_{\min} + \frac{s}{s+1}\,|v_k - v_r|, v_{\min} + \frac{|v_k - v_r|}{s+1}]$, gdzie $v_{\min} = \min\{v_k, v_r\}$, a s jest parametrem określającym szerokość okna O, przy czym $s = (1 - w)/(1 + w)$, gdzie $w = 0,2 - 0,3$. W konsekwencji proces uczenia odbywa się tylko wtedy, gdy

$$\min\{|t - v_k|/|t - v_r|, |t - v_r|/|t - v_k|\} > s.$$

Proces uczenia ma postać

$$v_r(l+1) = v_r(l) + \alpha\,(t - v_r(l)),$$

jeśli klasa wylosowanego obiektu c jest zgodna z klasą wybranego prototypu $c = c_r$ oraz

$$v_k(l+1) = v_k(l) - \alpha\,(t - v_k(l)),$$

gdy wylosowany obiekt należy do innej klasy niż prototyp v_k, $c \neq c_k$.

W kolejnej modyfikacji algorytmu LVQ, zwanej LVQ3, proces uczenia ma miejsce także wtedy, gdy oba prototypy v_r i v_k należą do tej samej klasy co aktualnie klasyfikowany wzorzec. Współczynnik uczenia α jest wtedy dodatkowo zmniejszany do wartości $\epsilon\,\alpha$, $\epsilon < 1$. Algorytm ten może być stosowany także w wersji z optymalizowanymi, indywidualnymi współczynnikami uczenia.

W celu poprawienia pozycji prototypów V zastosowano prostą modyfikację algorytmu LVQ2, która dalej nazywana będzie OLSQm [167], [170]. Algorytm ten ma szanse działać skutecznie tylko lokalnie, w pobliżu rozwiązania bliskiego optymalnemu.

ALGORYTM OLSQm

W tej wersji algorytmu modyfikowane są położenia jedynie tych prototypów, które aktualnie powodują błędną klasyfikację, to znaczy:

$$v_k(l+1) = v_k(l) - \alpha_k(l)\,(t - v_k(l)), \quad \text{gdy} \quad c_k \neq c,$$

w przeciwnym przypadku V pozostaje bez zmian.

7.5.2 Wybór początkowej zawartości zbioru prototypów

Wstępny wybór zbioru prototypów V może być dokonywany w rozmaity sposób. Najprostszym rozwiązaniem jest losowy wybór prototypów spośród elementów zbioru uczącego. Lepszym rozwiązaniem jest jednak wybór V za pomocą różnych metod kwantyzacji, na przykład algorytmu SOM lub algorytmu K–średnich [41], [105], [83], [100].

Zadanie kwantyzacji może dotyczyć każdej klasy z osobna lub całego zbioru uczącego [38]. W tym drugim przypadku określenie etykiety prototypu może się odbywać poprzez wybór klasy, która lokalnie minimalizuje empiryczne ryzyko [109], [115]. Inną możliwością jest zastosowanie jednego ze znanych algorytmów rozpoznawania, np. algorytmu k najbliższych sąsiadów [6].

7.5.3 Kondensacja zbioru prototypów

Bez straty ogólności możemy założyć, że V jest zbiorem uporządkowanym na odcinku I_1 w ten sposób, że $v_i \leq v_{i+1}$, $i = 1, \ldots, N-1$. W związku z tym granice między obszarami decyzji o różnej przynależności do klas wyznaczane są przez punkty $(v_i + v_{i+1})/2$, $c_i \neq c_{i+1}$, $i = 1, 2, \ldots, N-1$. Zbiór N prototypów prowadzi do funkcji decyzyjnej składającej się co najwyżej z N podprzedziałów o stałej wartości.

Zauważmy, że każdy prototyp $v_i, i = 2, \ldots, N-1$ może być usunięty ze zbioru prototypów V bez zmiany wartości funkcji decyzyjnej, gdy obaj jego sąsiedzi, v_{i+1} oraz v_{i-1}, należą do tej samej co on klasy, czyli $c_i = c_{i-1} = c_{i+1}$.

W tym miejscu potrzebny jest pewien komentarz. W przypadku rozpoznawania opartego na regule szukania „najbliższego prototypu" pojawia się szereg prototypów, które leżą we wnętrzu obszaru związanego z daną klasą. Prototypy te można usunąć, bez zmiany wartości funkcji decyzyjnej, jednakże w przypadku wielowymiarowym jest to bardzo skomplikowany problem obliczeniowy, ściśle związany z wyznaczaniem wielowymiarowych diagramów Voronoia. Jeśli operuje się danymi jednowymiarowymi, to wyznaczenie diagramu Voronoia staje się bardzo proste.

7.5.4 Przykłady zastosowania algorytmów LSQ do rozpoznawania

Efektywność użycia procedur OLSQ i OLSQm zbadano na danych rzeczywistych *Iris* oraz danych symulowanych (patrz przykład 7.2). W celu estymacji błędu rozpoznawania w odniesieniu do zbioru danych rzeczywistych zastosowano metodę CV1, procedurę typu cross-validation, w której każdy z wzorców pochodzących ze zbioru danych jest rozpoznawany za pomocą klasyfikatora skonstruowanego na podstawie pozostałych danych używanych jako ciąg uczący oraz metodę resubstytucji, w której jako zbiór testujący wybierany jest cały zbiór uczący [57].

Wyniki uśredniano dla 10 różnych prototypów uzyskanych za pomocą algorytmu uczenia OLSQ i OLSQm. W przypadku danych symulowanych generowano za każdym razem – dla ustalonej liczby prototypów – pięć niezależnych ciągów uczących. Do testowania użyto niezależny zbiór testujący składający się z 2000 obserwacji.

Przykład 7.7 *Zbiór danych Iris w wersji z trzema klasami*

Zbiór danych *Iris* opisano już w przykładzie 7.3. Średnie błędy klasyfikacji otrzymane dla zbiorów prototypów uzyskanych metodą OLSQ1 i OLSQm (użytymi sekwencyjnie) przedstawiono w tabeli 7.5. Dane przetransformowano na odcinek za pomocą krzywej Peano.

Tabela 7.5. Średnie błędy klasyfikacji otrzymane dla zbiorów prototypów uzyskanych metodą OLSQ1 i OLSQm (użytymi sekwencyjnie) w zastosowaniu do zbioru danych *Iris* przetransformowanych za pomocą krzywej Peano

Liczba prototypów	% Błędnych klasyfikacji		
	Resubstytucja	CV1	Najlepszy – najgorszy wynik w metodzie CV1
12	4,0%	4,0%	4,0%–4,0%
30	3,1%	4,5%	2,66%–5,3%
60	2,4%	3,9%	2,0%–5,3%

Liczba prototypów podana w tabeli 7.5 dotyczy wartości bez kompresji. Po usunięciu niepotrzebnych prototypów ich liczba zmniejszyła się do co najwyżej 16 (przy $N = 60$). Dla porównania, w przypadku klasycznej procedury uczenia LVQ (dla czterowymiarowych danych) najlepszy uzyskany wynik był rzędu 5% w przypadku 10–elementowego zbioru prototypów (wektorów 4-D) [16]. Błąd klasyfikacji uzyskany metodą k–NN jest rzędu 5,33 % w przypadku oryginalnego zbioru *Iris* oraz 2,66% w przypadku danych przetransformowanych za pomocą krzywej Peano [174].

Przykład 7.8 *Dane symulowane opisane w przykładzie 7.2*

Jak już wspomniano, w przykładzie tym występują dwie klasy reprezentowane przez 10-wymiarowy rozkład normalny. Klasy są rozdzielone, a granicę między nimi stanowi powierzchnia sfery o promieniu $\sqrt{9,8}$. Teoretyczny błąd (Bayesa) jest równy $0,0\%$, ale przykład mimo to uchodzi za trudne zadanie rozpoznawania [53]. Zbiór prototypów wyznaczano na podstawie ciągu uczącego o długości $n = 500$. Podobnie jak poprzednio, do transformacji danych zastosowano quasi-odwrotność krzywej Peano. Metoda k najbliższych sąsiadów prowadzi w przypadku badanego przykładu do błędów rozpoznawania rzędu 34% [53]. Przypomnijmy, że podobne wyniki daje klasyfikator w postaci szeregu Haara (patrz podrozdział 7.3.4).

Badany tu algorytm OLSQ pozwala zmniejszyć błąd klasyfikacji do 26% (z odchyleniem standardowym $0,01$) w przypadku 40 prototypów (20 prototypów w każdej z dwu klas). Kondensacja redukuje liczbę prototypów z 40 do 25. Zmiana początkowej liczby prototypów (będących punktem startowym procedury uczenia) nie prowadzi do zmniejszenia liczby błędnych klasyfikacji uzyskiwanych w procesie testowania. Otrzymane metodą LSQ wyniki są porównywalne jedynie z rezultatem uzyskanym przez Friedmana [53]. Podsumowując, eksperymenty obliczeniowe wskazują, że badane tu algorytmy prowadzą do uzyskiwania nie tylko szybkich i łatwych w realizacji, ale i efektywnych metod rozpoznawania.

7.6 Uwagi dotyczące realizacji algorytmów rozpoznawania

Pierwszy krok realizacji algorytmu jest oczywisty – przed procesem konstruowania klasyfikatora na podstawie ciągu uczącego (procesem uczenia) należy przetransformować wielowymiarowe wzorce (wektory) do punktów zawartych w kostce jednostkowej I_d. Można wybrać dowolną, liniową bądź nieliniową, ale ściśle monotoniczną transformację poszczególnych współrzędnych, która pozwala uzyskać wzajemnie jednoznaczne odwzorowanie skończonej liczby punktów z przestrzeni R^d w punkty z I_d. W kolejnych eksperymentach zastosowano jedynie liniowe transformacje.

Ważnym problemem jest także wybór konkretnej krzywej wypełniającej F. Można przy tym skorzystać z procedury typu cross-validation.

Eksperymenty obliczeniowe pokazały, że najlepszym wyborem jest najczęściej krzywa Peano (szczególnie, jeśli dane są skupione w centrum kostki I_d) lub krzywa Sierpińskiego (w przypadku danych rozrzuconych po obrzeżach kostki I_d). Krzywa Hilberta (tak często stosowana w przypadku przetwarzania obrazów) była sporadycznie wybierana przez procedurę cross-validation.

Poza wyborem typu krzywej, następnym etapem projektowania algorytmu, w którym można dokonywać optymalizacji, jest procedura ustalania kolejności uporządkowania poszczególnych współrzędnych wektora cech x_i, $i = 1, 2, \ldots, n$. Naturalnie, optymalny błąd klasyfikacji (Bayesa) jest niezależny od sposobu numeracji współrzędnych. Zmianę kolejności numeracji współrzędnych należy tu interpretować jako użycie nieco innej krzywej wypełniającej.

Eksperymenty, zarówno na danych rzeczywistych jak i symulowanych, wskazują jednak, że numeracja współrzędnych, podobnie jak użycie innego rodzaju krzywej, może w istotny sposób wpływać na efektywność algorytmu klasyfikacji – uzyskiwane eksperymentalnie wartości błędów. Minimalizacja empirycznego błędu ze względu na kolejność (numerację) cech użytą przy transformacji danych za pomocą wybranej krzywej wypełniającej jest skomplikowanym problemem kombinatorycznym, który zależy nie tylko od konkretnej postaci ciągu uczącego, ale także od użytej metody klasyfikacji. Dokładne rozwiązanie tego problemu w przypadku dużej liczby cech może być trudne. Na szczęście, nie wydaje się potrzebne badanie wszystkich możliwych permutacji $x_1, x_2, \ldots, x_d$.

Można zaobserwować, że te zmienne, które dają lepsze rezultaty w rozdzielaniu danych pochodzących z różnych klas, powinny się znajdować na końcu wektora cech. Jest to uwaga słuszna w przypadku używania krzywych wypełniających opisanych w niniejszej monografii. Dobre rezultaty dają następujące reguły heurystyczne [174]:

Niech $M_i = (m_{i0}, m_{i1})$ oznacza wektor wartości oczekiwanych rozkładu i–tej zmiennej. W praktyce zamiast wartości oczekiwanej używamy średniej z danych pochodzących z ciągu uczącego. Wektor wartości oczekiwanych może także zostać zastąpiony przez wektor median.

$$\hat{M}_i = \{n_j^{-1} \sum_{k=1}^{n} x_{ik} Y_k\}_{j=0}^{1},$$

gdzie

$$Y_k = \begin{cases} 1, \text{ gdy } X_k \text{ należy do klasy } 1, \\ 0, \text{ w przeciwnym przypadku,} \end{cases}$$

a n_j oznacza liczbę wzorców (elementów ciągu uczącego) należących do j–tej klasy ($j = 0, 1$).

Heurystyczna reguła porządkowania współrzędnych wektora cech polega na ich uporządkowaniu zgodnie z niemalejącymi wartościami odległości między klasami (patrz również [57]):

$$\delta_i = |\hat{m}_{i0} - \hat{m}_{i1}|. \tag{7.39}$$

Poszczególne współrzędne mogą zostać również uporządkowane zgodnie z ich ustandaryzowanymi średnimi odległościami, czyli zgodnie z δ_i / S_i, gdzie S_i oznacza odchylenie standardowe i–tej zmiennej, liczone łącznie względem średniej po wszystkich klasach. Efektywność proponowanych reguł można sprawdzić graficznie, analizując rysunki obrazujące dane po transformacji za pomocą krzywej, takie, jak na przykład rysunek 7.3. Podprzedziały, w których występują przemieszane dane z różnych klas (dla ustalonej numeracji współrzędnych) są na rysunku 7.3 wyraźnie widoczne. Powyższe reguły wyboru kolejności zmiennych łatwo uogólnić na przypadek większej liczby klas [174].

Dokładność obliczeń związanych z transformacją wielowymiarowych danych określona jest przez liczbę cyfr w dwójkowym (lub trójkowym) rozwinięciu współrzędnych punktów z I_d, uwzględnianych w transformacji $\boldsymbol{\Psi}$. Należy zauważyć, że wymagana dokładność powinna zależeć od długości ciągu uczącego n. W praktyce jednak stosunkowo łatwo zweryfikować, czy ustalony stopień aproksymacji jest wystarczający ze względu na możliwość rozróżniania danych z ciągu uczącego. W tym celu wystarczy porównać liczbę różnych elementów ciągu uczącego przed i po transformacji na odcinek I_1. Jeżeli długości obu ciągów są takie same, to przyjęty stopień dokładności jest wystarczający. W przeciwnym razie dokładność należy zwiększyć, gdyż przypadek ten oznacza, że przeciwobrazy pewnych elementów ciągu uczącego są nierozróżnialne w I_1 (wielu danym odpowiada ten sam punkt na odcinku I_1). Problem wyboru odpowiedniej dokładności ilustrują dane z tabeli 7.6. Tabela ta zawiera część danych pochodzących z często używanego

przykładu testowego, tak zwanego zbioru *Iris* [3], już po przetransformowaniu za pomocą quasi-odwrotności krzywej Hilberta. Cały zbiór danych *Iris* po transformacji na odcinek I_1 przedstawiono na rysunku 7.3. Jak łatwo zaobserwować, wielowymiarowe dane są często reprezentowane przez różniące się między sobą bardzo nieznacznie liczby rzeczywiste. W związku z tym muszą być one zapamiętywane z dużo większą dokładnością niż wartości poszczególnych współrzędnych danych przed transformacją.

Tabela 7.6. Część ciągu danych *Iris* po transformacji dokonanej przy użyciu quasi-odwrotności krzywej Hilberta, z dokładnością do 12 pozycji w rozwinięciu dwójkowym współrzędnych punktów z I_d

1-D Wzorzec	Y	1-D Wzorzec	Y
0,0270507868281129	0	0,0362309979554993	0
0,0402179964821698	0	0,0504533983194050	0
0,0528795113805244	1	0,1150935027335435	1
0,1152063914796599	1	0,1188845443584796	1
0,1330598780696163	1	0,1346415936459379	1
0,1346415936459379	1	0,1364780050298577	1
0,1467180817053304	1	0,2110204322889330	1
0,2227368386147645	0	0,2246379094249277	1
0,2251184958413432	1	0,3328100090002408	1
0,3557664855361508	1	0,3592026557489589	1
0,3644336041807036	1	0,3652626701223198	1
0,3828793777238388	0	0,3847363761842643	1

7.7 Wizualizacja wielowymiarowych danych za pomocą krzywych wypełniających

Analiza wielowymiarowych danych jest ściśle związana z możliwością ich graficznego przedstawienia. Wizualizację często wiąże się z problemem redukcji wymiaru danych w taki sposób, by jak najmniej zniekształcić ich wewnętrzną strukturę.

7.7.1 Uwagi na temat znanych metod wizualizacji

Jedną z możliwości jest zastosowanie klasycznych metod redukcji wymiaru, takich jak metoda analizy komponentów głównych (*Principal Component Analysis* – PCA) [191] czy odwzorowanie Sammona [139], [40], [52] i to zarówno w odniesieniu do oryginalnych danych, jak i danych wstępnie przekształconych, na przykład za pomocą odwzorowania realizowanego przez sieć SOM.

W metodzie PCA ustala się liniową kombinację kolumn danych z maksymalną (bądź minimalną) wariancją. Komponenty główne wyznaczane są jako wektory własne macierzy kowariancji danych wejściowych bądź też macierzy korelacji danych. Należy zaznaczyć, że komponenty główne zależą od sposobu skalowania zmiennych. Także zastosowanie macierzy korelacji zamiast macierzy kowariancji daje te same wyniki tylko wówczas, gdy zmienne wejściowe mają zerowe wartości oczekiwane. W statystyce najczęściej używa się PCA opartej na macierzy kowariancji, natomiast w przypadku sieci neuronowych częściej wybierana jest macierz korelacji [191], [116], [13].

PCA jest ściśle związana z transformacją Karhuena–Loeve [57], stosowaną w przetwarzaniu sygnałów traktowanych jako procesy stochastyczne. Użycie tylko części komponentów głównych prowadzi do redukcji wymiaru przestrzeni danych. Podprzestrzeń rozpięta na zredukowanej liczbie największych komponentów głównych stanowi najlepszą reprezentację (wizualizację, gdy ograniczymy się do jednego, dwu lub trzech komponentów) wielowymiarowych danych w tym sensie, że reprezentacja ta ma zarówno maksymalną macierz kowariancji (ze względu na ślad i wyznacznik macierzy), jak i w najlepszy sposób aproksymuje oryginalne dane (minimalizuje sumę kwadratów odległości punktów od ich projekcji) [191].

Algorytm SOM równocześnie realizuje dwa zadania – wektorowej kwantyzacji (kompresja danych) oraz zadanie odtwarzania przestrzennej organizacji danych wejściowych w ograniczonej, na ogół, strukturze topologicznej sieci SOM [192], [2], [14], [62], [96]. Są to odwzorowania topograficzne prowadzące do redukcji wymiaru, wizualizacji danych przestrzennych itd. Z punktu widzenia każdego z tych zadań rozpatrywanych odrębnie, algorytm SOM nie jest rozwiązaniem idealnym, stąd wiele pomysłów łączenia algorytmów uczenia konkurencyjnego z różnymi transformacjami danych wejściowych [32], [103], [163], [164], [75]. Obrazowanie wielowymiarowych danych w postaci punktów na płaszczyźnie w zastosowaniu do zadań rozpoznawania było szeroko badane w pracy [144]. Każde ze wspomnianych wyżej rozwiązań wymaga znacznych nakładów obliczeniowych.

7.7.2 Wizualizacja za pomocą powierzchni i par krzywych wypełniających

W niniejszym podrozdziale przedstawimy inne podejście do wizualizacji wielowymiarowych danych polegające na redukcji ich wymiaru za pomocą krzywych wypełniających, a dokładniej odwzorowań quasi-odwrotnych do krzywych wypełniających.

Jak wiadomo, krzywe wypełniające, których użyciem zajmujemy się w niniejszej monografii, mają tę własność, że punkty bliskie sobie w przestrzeni wielowymiarowej są obrazami (w odwzorowaniu względem krzywej) punktów bliskich

sobie na odcinku jednostkowym. Naturalnie odwzorowanie pseudoodwrotne może prowadzić do rozseparowania na odcinku I_1 punktów będących obrazami bliskich sobie danych z przestrzeni wielowymiarowej, powodując pewne zniekształcenia ich geometrycznej struktury, ale zachowując równocześnie ich własności statystyczne. Przedstawienie przetransformowanych danych na odcinku jednostkowym jest najprostszą formą wizualizacji za pomocą krzywych wypełniających [119], [182].

W niniejszym rozdziale skoncentrujemy się na odwzorowaniach danych wielowymiarowych na płaszczyznę. Odwzorowania danych na płaszczyźnie mogą być łatwo analizowane przez człowieka, pozwalając mu szybko podejmować decyzje motywowane różnymi, trudno poddającymi się formalizacji, czynnikami.

Dalej zdefiniujemy odwzorowanie analogiczne do krzywej wypełniającej, które przekształca kwadrat jednostkowy $I_2 = [0,1]\times[0,1]$ w wielowymiarową kostkę I_d. Będziemy postulować, by odwzorowanie to, podobnie jak krzywa wypełniająca, było odwzorowaniem ciągłym i zachowywało miarę Lebesgue'a.

D–wymiarowa kostka I_d może być analizowana jako produkt kartezjański dwu (lub więcej) kostek o niższym wymiarze, czyli $I_d = I_r \times I_s, \quad s + r = d$ w tym sensie, że $(x_1, \ldots, x_{r+s}) \in I_d$ wtedy i tylko wtedy, gdy $(x_1, \ldots, x_r) \in I_r$ oraz $(x_{r+1}, \ldots, x_{r+s}) \in I_s$.

Niech F_d oznacza krzywą wypełniającą kostkę d–wymiarową I_d, spełniającą warunki **C1**–**C3** (patrz rozdział 4.5), i odpowiednio niech $\boldsymbol{\Psi}_d$ będzie quasi-odwrotnością względem F_d.

Zdefiniujmy odwzorowanie $W_{r,s} : I_2 \to I_d$ jako

$$W_{r,s}(x,y) = F_r(x)(\times)F_s(y) = (x_{1r}(x), \ldots, x_{rr}(x), x_{1s}(y), \ldots, x_{ss}(y)),$$

gdzie $F_r(x) = (x_{1r}(x), \ldots, x_{rr}(x))$, $x \in I_1$ oraz $F_s(y) = (x_{1s}(y), \ldots, x_{ss}(y))$, $y \in I_1$ są odpowiednio r i s wymiarowymi krzywymi wypełniającymi, które spełniają warunek Höldera (warunek **C1** z rozdziału 4.5) dla dowolnego $r, s > 1$, czyli warunek

$$\| F_r(t_1) - F_r(t_2) \| \leq c_r (| t_1 - t_2 |)^{1/r}, \quad t_1, t_2 \in I_1,$$

gdzie c_r jest pewną stałą zależną od wymiaru r.

Twierdzenie 7.7.1 *Odwzorowanie $W_{r,s} = F_r(\times)F_s$ jest ciągłym odwzorowaniem, spełniającym warunek Höldera z wykładnikiem $1/\max(r,s)$, czyli*

$$\| W_{r,s}(x_1,y_1) - W_{r,s}(x_2,y_2) \| \leq \sqrt{2}\, c_v (\| (x_1,y_1) - (x_2,y_2) \|)^{1/v}, \qquad (7.40)$$

$$(x_1,y_1), (x_2,y_2) \in I_2,$$

gdzie $v = \max(r,s)$, natomiast $c_v = \max(c_r, c_s)$.

Dowód. $W_{r,s}$ jest odwzorowaniem ciągłym jako produkt kartezjański dwu odwzorowań ciągłych. Każda z jego współrzędnych jest odwzorowaniem ciągłym. Przejdziemy teraz do dowodu warunku Höldera (7.40).

Zgodnie z definicją $W_{r,s}$ mamy:

$$\begin{aligned}
&\| W_{r,s}(x_1,y_1) - W_{r,s}(x_2,y_2) \|^2 \\
&= ||F_r(x_1) - F_r(x_2)||^2 + ||F_s(y_1) - F_s(y_2)||^2 \\
&\leq c_r^2 |x_1 - x_2|^{2/r} + c_s^2 |y_1 - y_2|^{2/s} \\
&\leq c_v^2 (|x_1 - x_2|^2)^{2/v} + c_v^2 (|y_1 - y_2|^2)^{2/v} \\
&\leq c_v^2 (|x_1 - x_2|^2 + |y_1 - y_2|^2)^{2/v} + c_v^2 (|x_1 - x_2|^2 + |y_1 - y_2|^2)^{2/v} \\
&\leq 2c_v^2 ||(x_1,y_1) - (x_2,y_2)||^{2/v},
\end{aligned}$$

co kończy dowód twierdzenia □.

Zdefiniujmy odwzorowanie quasi-odwrotne $\boldsymbol{\Psi}_{r,s}(z) \in W_{r,s}^{-1}(z)$, $z \in I_{r+s}$. Niech

$$\boldsymbol{\Psi}_{r,s}(z) = (\boldsymbol{\Psi}_r(x_{1r}, \ldots, x_{rr}), \boldsymbol{\Psi}_s(x_{1s}, \ldots, x_{ss})),$$

przy czym

$$z = (x_{1r}, \ldots, x_{rr}, x_{1s}, \ldots, x_{ss}) \in I_d = I_r \times I_s,$$

$$(x_{1r}, \ldots, x_{rr}) \in I_r \text{ oraz } (x_{1s}, \ldots, x_{ss}) \in I_s,$$

a $\boldsymbol{\Psi}_r$ i $\boldsymbol{\Psi}_s$ są, odpowiednio, quasi-odwrotnościami krzywych F_r i F_s. Łatwo sprawdzić, że $W_{s,r}^{-1}(x_{1r}, \ldots, x_{rr}, x_{1s}, \ldots, x_{ss})$ jest równe

$$\{(x,y) : F_r(x) = (x_{1r}, \ldots, x_{rr}),\ F_s(y) = ((x_{1s}, \ldots, x_{ss})\}.$$

Ponieważ zachodzi

$$F_r(\boldsymbol{\Psi}_r(x_{1r}, \ldots, x_{rr})) = (x_{1r}, \ldots, x_{rr})$$

oraz

$$F_s(\boldsymbol{\Psi}_s(x_{1s}, \ldots, x_{ss})) = (x_{1s}, \ldots, x_{ss}),$$

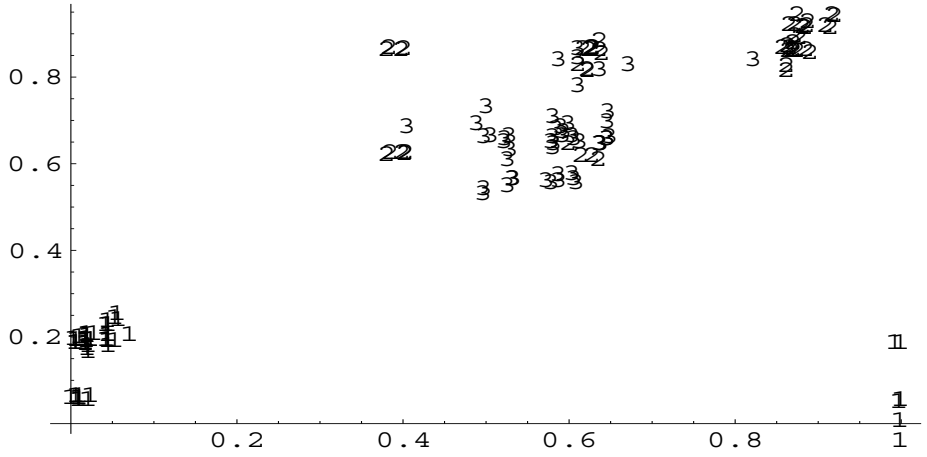
zatem

$$\begin{aligned}
&W_{r,s}(\boldsymbol{\Psi}_{r,s}(x_{1r}, \ldots, x_{rr}, x_{1s}, \ldots, x_{ss})) \\
&= W_{r,s}(\boldsymbol{\Psi}_r(x_{1r}, \ldots, x_{rr}), \boldsymbol{\Psi}_s(x_{1s}, \ldots, x_{ss})) \\
&= (F_r(\boldsymbol{\Psi}_r(x_{1r}, \ldots, x_{rr})), F_s(\boldsymbol{\Psi}_s(x_{1s}, \ldots, x_{ss}))) \\
&= (x_{1r}, \ldots, x_{rr}, x_{1s}, \ldots, x_{ss}).
\end{aligned}$$

Odwzorowanie $\boldsymbol{\Psi}_{r,s}$ można stosować bezpośrednio do wizualizacji wielowymiarowych danych na płaszczyźnie. Odwzorowanie to gwarantuje, podobnie jak w przypadku poprzednio rozpatrywanych przekształceń wielowymiarowego zbioru danych na odcinek I_1, że punkty leżące blisko siebie na odcinku są obrazami

punktów leżących blisko siebie w przestrzeni wielowymiarowej I_d. Z twierdzenia 7.7.1 wynika bezpośrednio, że jeśli $||(x_1, y_1) - (x_2, y_2)|| = \Delta$, to odległość odpowiadających (x_1, y_1) oraz (x_2, y_2) punktów w przestrzeni d–wymiarowej, czyli $||W_{r,s}(x_1, y_1) - W_{r,s}(x_2, y_2||$ jest nie większa niż $c_v \Delta^{1/v}$, gdzie $v = \max(r, s)$, $s + r = d$.

Inną metodą zastosowania krzywych wypełniających do wizualizacji wielowymiarowych danych jest połączenie transformacji tych samych danych za pomocą różnego typu krzywych, na przykład transformacji z równoczesnym użyciem krzywej Peano oraz krzywej Hilberta. Ponieważ różne krzywe wypełniające w różny sposób przenoszą geometryczne własności danych, mogą być one traktowane jako różne ,,projekcje" wielowymiarowych danych, które połączone razem dają ich (czyli danych) pełniejszy obraz. Tego typu wizualizacje zachowują wszelkie własności odwzorowań bazujących na pojedynczych krzywych wypełniających.



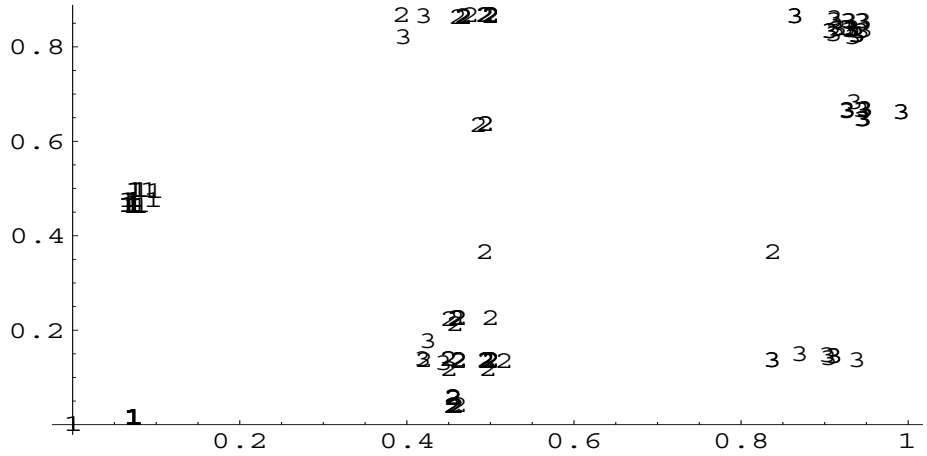
Rys. 7.7. Dane *Iris* po transformacji za pomocą krzywej Sierpińskiego w odniesieniu do par zmiennych o numerach odpowiednio $(1, 3)$ na jednej osi oraz $(2, 4)$ na drugiej osi współrzędnych
Fig. 7.7. *Iris* data after Sierpiński's curve transformation on $(1, 3)$ and $(2, 4)$ variables

7.7.3 Przykłady wizualizacji wielowymiarowych danych na płaszczyźnie

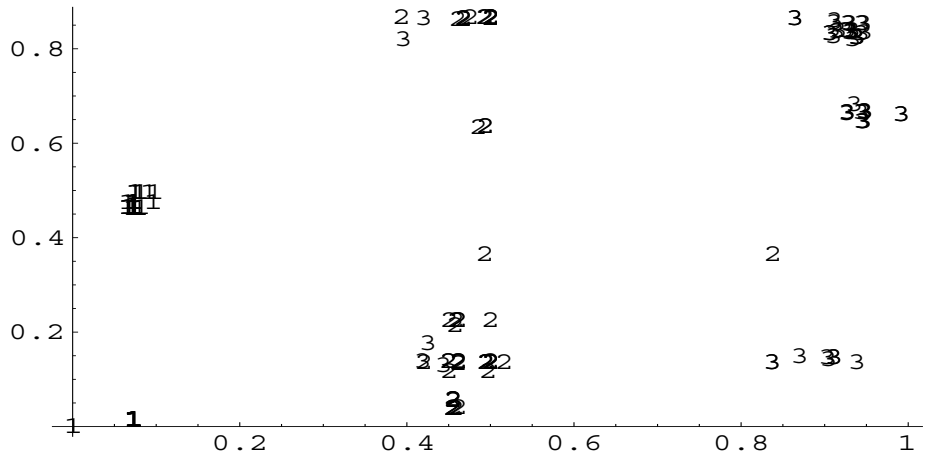
Pierwszy przykład dotyczy wizualizacji znanego zbioru danych *Iris* [51] opisanego w przykładzie 7.3, który zawiera pomiary płatków trzech różnych gatunków irysa: *iris setosa, versicolor* i *virginica*.

Odwzorowanie zbioru *Iris* na płaszczyźnie przedstawiono na rysunku 7.7. Długości dwu typów płatków oraz osobno ich szerokości zakodowano za pomocą dwu-

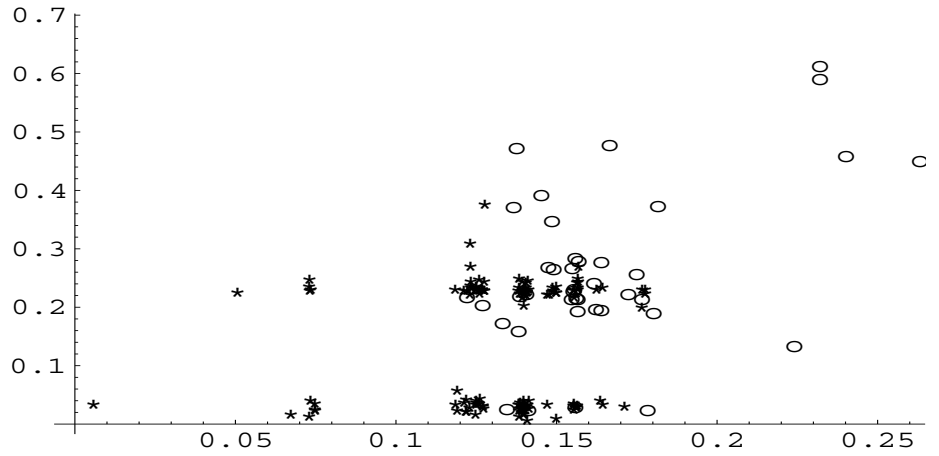
wymiarowej krzywej Sierpińskiego – jako zmienne o numerach (1,3) oraz odpowiednio (2,4). Z rysunku tego widać wyraźnie, że gatunek *iris setosa* (oznakowany cyfrą 1) jest dokładnie oddzielony od pozostałych dwu gatunków oznaczonych cyframi 2 i 3.



Rys. 7.8. Dane *Iris* po transformacji za pomocą krzywej Peano (oś pozioma) oraz krzywej Hilberta (oś pionowa)
Fig. 7.8. *Iris* data after parallel Peano's and Hilbert's curve transformation on all variables



Rys. 7.9. Dane *Iris* po transformacji za pomocą krzywej Peano (oś pozioma) oraz krzywej Sierpińskiego (oś pionowa)
Fig. 7.9. *Iris* data after parallel Peano's and Sierpiński's curve transformation on all variables



Rys. 7.10. Dane *Muscular dystrophy* po transformacji za pomocą krzywej Hilberta w odniesieniu do zmiennych o numerach odpowiednio $(1,2,3)$ na jednej osi oraz $(4,5)$ na drugiej osi współrzędnych
Fig. 7.10.*Muscular dystrophy* data after Hilberts's curve transformation on $(1,2,3)$ and $(4,5)$ variables

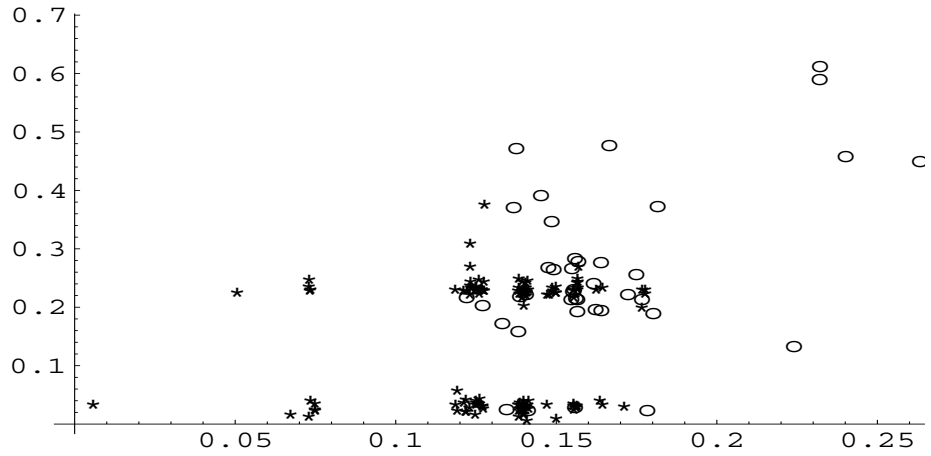
Lepszą jeszcze wizualizację danych uzyskano za pomocą równoległej transformacji danych przy użyciu dwu różnych typów krzywej wypełniającej.

Na rysunku 7.8 przedstawiono transformację danych za pomocą krzywych Peano i Hilberta – natomiast na rysunku 7.9 wizualizację zbioru *Iris* otrzymaną przy zastosowaniu krzywej Peano i Sierpińskiego. W obu przypadkach podział płaszczyzny na trzy obszary odpowiadające poszczególnym klasom dokonany przez obserwatora (wybrano granice podziału dyskryminacyjnego w postaci kombinacji prostych równoległych do obu osi układu współrzędnych) prowadzi do błędów rozpoznawania rzędu $0,04$, czyli mniejszych niż uzyskane metodą k najbliższych sąsiadów [53].

W drugim przypadku rozpatrzymy dane ze zbioru *Duchenne Muscular dystrophy* opisanego w przykładzie 7.4. Dane te zawierają 193 pomiary pięciu parametrów $d = 5$. Transformacja za pomocą krzywej Hilberta w odniesieniu do dwu grup zmiennych zależy od sposobu ich podziału.

W pierwszym wariancie podział zmiennych był typu $(1,2,3)$ oraz $(4,5)$ – patrz rysunek 7.10; w drugim wariancie zmienne podzielono na grupy $(4,2,3)$ oraz $(1,5)$ – patrz rysunek 7.11. Pierwsze rozwiązanie nie jest szczególnie korzystne z punktu widzenia rozpoznawania. Osiągnięto błąd klasyfikacji rzędu 20%.

Drugi sposób wizualizacji pozwolił natomiast na uzyskanie błędów tego samego rzędu, jak uzyskiwane metodą CART [19], [53], czyli błędów rzędu $12,4\%$.



Rys. 7.11. Dane *Muscular dystrophy* po transformacji za pomocą krzywej Hilberta w odniesieniu do zmiennych o numerach odpowiednio $(4,2,3)$ na jednej osi oraz $(1,5)$ na drugiej osi współrzędnych
Fig. 7.11. *Muscular dystrophy* data after Hilberts's curve transformation on $(4,2,3)$ and $(1,5)$ variables

7.8 Podsumowanie

W powyższym rozdziale omówiono i zbadano własności różnych algorytmów rozpoznawania, których wspólną cechą było zastosowanie krzywych wypełniających, a dokładnie ich quasi-odwrotności, jako narzędzia do transformacji wielowymiarowych danych do postaci jednowymiarowej.

Nakład obliczeń potrzebny do przetransformowania za pomocą krzywej wypełniającej d–wymiarowego punktu na odcinek I_1 jest rzędu $O(d)$, gdyż do transformacji nie musimy generować całej krzywej, lecz każdy wzorzec można transformować niezależnie od pozostałych.

Podstawowym rezultatem zawartym w powyższym rozdziale jest twierdzenie 7.2.3, z którego wynika, iż krzywa wypełniająca, która zachowuje miarę Lebesgue'a i jest odwracalna prawie wszędzie, z dokładnością do zbioru miary zero, może być użyta (a dokładnie jej quasi-odwrotność) jako odwzorowanie, które zachowuje ryzyko Bayesa dla dowolnego rozkładu danych o nośniku zawartym w wielowymiarowej kostce I_d.

Należy zaznaczyć, że transformacja bazująca na krzywej wypełniającej nie stanowi sposobu na uniknięcie konieczności posiadania bardzo dużej liczby obserwacji w przypadku, gdy wymiar przestrzeni jest duży (bellmanowskie „przekleństwo wielowymiarowości"). Przetransformowane dane wymagają bowiem zachowania

dużej dokładności, by można je było rozróżnić między sobą. Ceną za redukcję wymiaru jest znaczne zwiększenie wymaganej dokładności reprezentowania jednowymiarowych danych.

Ze względu na zachowywanie ryzyka Bayesa przez transformacje oparte na krzywych wypełniających, możemy w odniesieniu do przetransformowanych danych użyć dowolnego, dostatecznie wrażliwego klasyfikatora, nie tracąc nic z informacji statystycznych zawartych w oryginalnych danych. Jeżeli reguła klasyfikacji w jednym wymiarze jest zbieżna, w jakimś sensie, do ryzyka Bayesa, to klasyfikator taki, zastosowany w odniesieniu do przetransformowanych danych, jest w takim samym sensie asymptotycznie optymalny.

Podstawową cechą rozważanych algorytmów jest szybkość samego procesu podejmowania decyzji. W końcowym etapie konstrukcji klasyfikatora otrzymujemy dyskryminacyjny podział odcinka I_1 na m pododcinków odpowiadających poszczególnym klasom, przy czym m jest nie większe niż długość ciągu uczącego. W związku z tym nakład pracy potrzebny do dokonania samego procesu klasyfikacji nowego obiektu x jest rzędu $O(d) + \log_2 m$.

Szybkie algorytmy klasyfikacji używające krzywych wypełniających wykazują pewne podobieństwo z algorytmami bazującymi na podziałach danych [36], [19]. Podobieństwo to można wiązać z tym, iż struktura każdej krzywej wypełniającej jest w istocie rzeczy związana z pewnym rekurencyjnym podziałem przestrzeni danych.

Rozdział 8

Krzywe wypełniające w problemach statystycznego sterowania produkcją

Zadanie szybkiego wykrywania zmian zachodzących w procesie stochastycznym [199] na podstawie ciągu obserwacji ma bardzo wiele istotnych zastosowań, poczynając od sterowania i kontroli jakości przemysłowych procesów produkcyjnych [114], [9], [104], [86], poprzez automatyczne wykrywanie uszkodzeń w systemie dynamicznym [195], [85], a skończywszy na uaktualnianiu współczynników w adaptacyjnych algorytmach sterowania [140]. Przy projektowaniu algorytmów służących do detekcji zmian zachodzących w procesie, jako punkt odniesienia dla jakości funkcjonowania algorytmu przyjmuje się zazwyczaj stosunek liczby fałszywych alarmów – czyli liczby błędnych decyzji podjętych przez algorytm – w odniesieniu do liczby wszystkich podejmowanych decyzji. Podstawowym zadaniem umożliwiającym efektywne sterowanie procesem jest natomiast jak najszybsze wykrycie momentu zmiany. W konsekwencji głównym kryterium oceny jakości algorytmu wykrywającego zmiany zachodzące w procesie na podstawie ciągu obserwacji jego przebiegu powinna być minimalizacja średniego czasu, jaki upływa między momentem pojawienia się zmiany w procesie, a momentem wykrycia tej zmiany.

Wartość oczekiwana czasu do „fałszywego alarmu" w sytuacji, gdy sterowany proces nie uległ zmianom (*in control average run length* ARL) powinna być jak największa, musi być jednak skończona, gdyż w przeciwnym przypadku system decyzyjny nie reagowałby na jakiekolwiek zmiany. Dlatego zadanie formułuje się zwykle w następujący sposób. Należy skonstruować procedurę decyzyjną, która minimalizuje średni czas do wykrycia zmiany procesu, przy ograniczeniu, że średni czas do fałszywego alarmu będzie nie krótszy niż pewna, z góry zadana,

wartość. Główną ideą proponowanego tu podejścia jest zastosowanie transformacji quasi-odwrotnej do krzywej wypełniającej w celu przekształcenia wielowymiarowych danych do postaci skalarnej oraz podejmowanie decyzji w przestrzeni jednowymiarowej.

Należy zwrócić uwagę, że tego typu transformacje mają całkiem odmienny charakter niż liniowe przekształcenia danych z jednej strony, bądź transformacje bazujące na odległości Mahalanobisa [113], [57], z drugiej strony.

Dane po przetransformowaniu na odcinek są używane do wyznaczenia obszaru akceptacji, czyli obszaru, który na zadanym poziomie ufności ,,gwarantuje", że dana obserwacja pochodzi z tego samego rozkładu określającego stan normalny procesu.

W niniejszym rozdziale badany jest tylko problem wykrywania zmian w wielowymiarowym procesie jedynie na podstawie aktualnej obserwacji procesu. Proponowane rozwiązanie problemu decyzyjnego w postaci multi-karty pozwala na szybkie podejmowanie decyzji. Złożoność obliczeniowa procedury (poza nakładem związanym z transformacją wielowymiarowego wektora obserwacji do postaci jednowymiarowej) jest proporcjonalna do logarytmu z liczby przedziałów multi-karty.

8.1 Sformułowanie problemu wykrywania zmian w procesie

Niech $X_1, X_2, \ldots$ będą niezależnymi, wielowymiarowymi, zmiennymi losowymi (wektorami losowymi), obserwowanymi sekwencyjnie w kolejnych momentach czasowych $t = 1, 2, \ldots$ Ponadto zakładamy, że wszystkie zmienne losowe od X_1 do X_{q-1} mają ten sam rozkład, opisany funkcją gęstości prawdopodobieństwa f, natomiast począwszy od momentu q zmienne losowe $X_q, X_{q+1}, \ldots$ mają inny rozkład prawdopodobieństwa z funkcją gęstości $f_1 \neq f$. Moment czasowy q jest nieznany, należy natomiast podjąć pewne działania w momencie pojawienia się zmian w obserwowanym procesie, które przejawiają się jedynie w postaci obserwacji $X_q, X_{q+1}, \ldots$ Szybkość wykrycia zmian może być bardzo istotna ze względu na pojawianie się dodatkowych kosztów związanych z nieprawidłowo przebiegającym procesem. Podstawowym zadaniem umożliwiającym efektywne sterowanie procesem jest jak najszybsze wykrycie momentu zmiany. W konsekwencji głównym kryterium oceny jakości algorytmu wykrywającego zmiany zachodzące w procesie jest szybkość wykrywania zmiany.

W odróżnieniu od klasycznych prac, w których o f i f_1 przyjmowano, że mają rozkład normalny, w rozważanym tutaj problemie nie zakłada się znajomości postaci rozkładu opisującego monitorowany proces. Przyjmujemy, że istnieją

rozkłady prawdopodobieństwa opisujące proces w stanie prawidłowym (*in control*) oraz proces zakłócony (*out of control*) i dla uproszczenia przyjmujemy, że rozkłady te są absolutnie ciągłe względem miary Lebesgue'a, czyli istnieją odpowiednie gęstości tych rozkładów f oraz f_1, natomiast nie są one znane. Bez straty ogólności będziemy zakładać, że f oraz f_1 mają ograniczone nośniki zawarte w I_d. Założenia dotyczące charakteru zmian w procesie nie są dokładnie sprecyzowane. Generalnie rzecz traktując, można dopuszczać znacznie szerszą klasę zmian niż tylko zmiany wartości średnich i wariancji.

Naszym zadaniem jest, na podstawie ciągu niezależnych obserwacji $X_1, \ldots$ $X_t = (x_{t1}, \ldots, x_{td})$, podjąć decyzję, czy X_t jest zmienną losową o rozkładzie f – czyli monitorowany proces jest w stanie prawidłowym lub X_t jest zmienną losową o innym rozkładzie – proces jest w stanie rozregulowania, innymi słowy – w procesie pojawiły się istotne zmiany.

Istnieje obszerna literatura na temat algorytmów detekcji zmian w procesie (patrz [9] i cytowana tam bibliografia). Znanych jest wiele różnych wielowymiarowych kart kontrolnych, czyli procedur stosowanych do wykrywania zmian w monitorowanym procesie. Statystyczna kontrola procesu często przybiera postać ciągłego (powtarzanego w czasie) testowania hipotez [113]. Jeśli decyzje opierają się na pojedynczych obserwacjach (bądź wielu równoczesnych obserwacjach pobranych w tym samym momencie czasu), wtedy typową, polecaną w tym przypadku procedurą, jest karta kontrolna stosująca statystykę Hotelling'a T^2 lub inne tego typu algorytmy, oparte faktycznie na odległości Mahalanobisa [113], [104], [9].

Jeżeli rozkłady o gęstościach f oraz f_1 są wielowymiarowymi rozkładami normalnymi, różniącymi się tylko wektorem wartości średnich (macierz kowariancji nie ulega zmianie), i jeśli pominiemy problem estymacji macierzy kowariancji (co w praktyce nie jest łatwe), to karta kontrolna bazująca na odległości Mahalanobisa jest łatwa do zastosowania. Wystarczy sprawdzić warunek

$$T^2 = (X_i - M)^T \Sigma^{-1} (X_i - M) > h, \tag{8.1}$$

gdzie M jest wektorem wartości oczekiwanych, Σ macierzą kowariancji procesu w stanie prawidłowym, natomiast h jest określoną wartością graniczną. Bez spełnienia wymienionych powyżej założeń problem wykrywania zmian w wielowymiarowym procesie staje się ekstremalnie trudny.

Karta T^2 jest uogólnieniem jednowymiarowej karty Shewharta [114], która używa przedziału ufności wartości oczekiwanej rozkładu wyznaczonego na podstawie rozkładu normalnego (przy znanej wariancji rozkładu) lub na podstawie rozkładu t–Studenta (gdy wariancja jest estymowana z próby).

W przypadku wielowymiarowym równanie $(X_i - M)^T \Sigma^{-1} (X_i - M) = h$ opisuje powierzchnię elipsoidy o środku w punkcie M. Jeśli macierz kowariancji Σ

jest znana, to $(X_i - M)^T \Sigma^{-1} (X_i - M)$ ma rozkład χ^2 z d stopniami swobody. Nierówność $T^2 \leq h$ dla wartości h wybranej jako wartość krytyczna rozkładu χ^2 przy ustalonym poziomie ufności $1 - p$, opisuje elipsoidę ufności (na tym samym poziomie ufności $1 - p$) o minimalnej objętości [113].

Kryterium jakości kart kontrolnych [114], [9] jest najczęściej średni czas wykrycia zmiany w rozkładzie obserwowanego procesu ARL. Innymi słowy, ARL to średnia liczba obserwacji od momentu, w którym nastąpiła zmiana, do momentu jej wykrycia, czyli podjęcia decyzji o alarmie.

Decyzja o zmianie charakteru procesu może zostać podjęta także w przypadku braku faktycznej zmiany. Wartość ARL_0, średni czas do fałszywego alarmu, jest istotnym parametrem opisującym kartę kontrolną. Wartość ta jest związana z błędem I rodzaju stosowanego testu statystycznego, podczas gdy ARL jest związany z błędem II rodzaju.

W przypadku niezależności obserwowanych zmiennych losowych ARL_0 odpowiada wartości oczekiwanej liczby prób do osiągnięcia pierwszego „sukcesu" (zgodnie ze schematem Bernoulliego), przy czym „sukces" jest utożsamiany z podjęciem decyzji o alarmie. Liczba takich prób ma rozkład geometryczny o wartości oczekiwanej $1/p$ i wariancji $(1-p)/p^2$, gdzie $1-p$ jest poziomem ufności, prawdopodobieństwem uznania obserwacji za zgodną z aktualnym rozkładem. Wartość $1-p$ jest określona przez granice kontrolne (obszar akceptacji) karty. Naturalnie, różne granice kontrolne mogą prowadzić do tej samej wartości poziomu ufności.

Jeżeli nie ma żadnych informacji o typie zmian, które mogą zachodzić w procesie, to standardowym i naturalnym sposobem postępowania jest wyznaczanie obszaru ufności o minimalnej objętości (minimalnej długości w przypadku jednowymiarowym).

Typowym przykładem może być pierwsza karta kontrolna Shewharta [114], [9], w której obszar akceptacji jest wyznaczany przez przedział $+/-3\sigma$ względem wartości średniej obserwowanej wielkości. Tylko w najprostszych przypadkach, przy pełnej znajomości rozkładu i ściśle ustalonych granicach kontrolnych, wartość ARL_0 można określić analitycznie.

Jeszcze trudniejsza sytuacja występuje w przypadku określania wartości ARL, gdyż zależy ona od rodzaju odstępstwa od normy. W przypadku niezależnych obserwacji wartość ARL jest równa $1/(1-\beta)$, gdzie β jest aktualną, związaną z charakterem i wielkością zmiany w rozkładzie procesu, wartością błędu II rodzaju.

Najczęściej wartości ARL dla różnych rodzajów zmian rozkładu są wyznaczane eksperymentalnie poprzez badania symulacyjne. Ze względu na dużą wariancję zmiennej losowej, która wyznacza liczbę obserwacji procesu od momentu zmiany do podjęcia decyzji o wszczęciu alarmu, eksperymenty te wymagają ogromnych nakładów obliczeniowych.

8.2 Wykrywanie zmian w wielowymiarowym procesie przy użyciu multi–karty

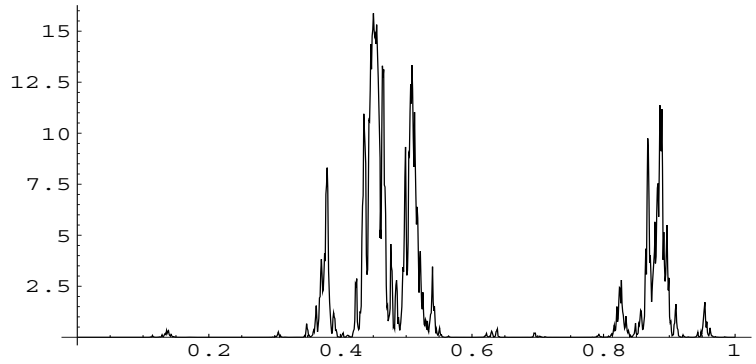
W przypadku obserwacji wielowymiarowego procesu zwykle konstruowanych jest tyle niezależnych kart kontrolnych, ile jest obserwowanych zmiennych (jaki jest wymiar wektora X). W praktyce zmienne te są często statystycznie zależne i konieczna jest ich łączna analiza wielowymiarowa. Proponowane tu podejście bierze pod uwagę możliwe zależności statystyczne między składowymi wektora obserwacji. Jego istotą jest przetransformowanie poszczególnych wielowymiarowych obserwacji na odcinek I_1 za pomocą quasi–odwrotności krzywej wypełniającej.

W ogólnym przypadku należy się spodziewać, że nawet gdy gęstość wielowymiarowego rozkładu jest funkcją bardzo gładką (wielokrotnie różniczkowalną), to po przetransformowaniu do postaci jednowymiarowej za pomocą odwzorowania quasi–odwrotnego do krzywej wypełniającej otrzymujemy funkcję gęstości, która nie jest gładka (nie jest różniczkowalna).

Dla ilustracji, na rysunku 8.1 przedstawiono gęstość dwuwymiarowego rozkładu normalnego po przetransformowaniu na odcinek I_1 za pomocą quasi-odwrotności krzywej Peano. Należy zwrócić uwagę na fraktalną strukturę przetransformowanej, jednowymiarowej funkcji gęstości prawdopodobieństwa, która zgodnie z lematami (4.16) i (4.14) jest funkcją hölderowską z wykładnikiem 1/2.

Karty kontrolne, w których wnioskowanie odbywa się na podstawie pojedynczej obserwacji mają charakter testu istotności i wymagają konstrukcji obszaru krytycznego o zadanym poziomie istotności p. Jeżeli obserwacja należy do obszaru krytycznego, oznacza to, że zaszło zdarzenie o bardzo małym (nie większym niż p) prawdopodobieństwie i możemy podjąć decyzję o odrzuceniu hipotezy, że obserwacja pochodzi z ustalonego rozkładu, dla którego wyznaczono obszar krytyczny. Jest to podstawą do podjęcia decyzji, że nastąpiła zmiana w procesie, a prawdopodobieństwo fałszywego alarmu jest równe właśnie poziomowi istotności p. Obszar akceptacji (ufności) jest dopełnieniem zbioru krytycznego. Wybór obszaru krytycznego nie jest problemem posiadającym jednoznaczne rozwiązanie. Przy braku hipotezy alternatywnej rozsądnym rozwiązaniem jest wybór zbioru krytycznego o maksymalnej mierze Lebesgue'a, w konsekwencji miara zbioru akceptacji jest w ten sposób minimalizowana.

Nawet wtedy, gdy postać gęstości rozkładu na odcinku $g(t) = f(F(t)),\ t \in I_1$ jest dokładnie znana, to i tak problem wyznaczenia zbioru ufności o minimalnej mierze Lebesgue'a na odcinku nie jest problemem łatwym do rozwiązania. W ogólnym przypadku zbiór ten może składać się z wielu rozłącznych pododcinków zawartych w I_1. Każdy z tych pododcinków może być używany jako przedział kontrolny typowej karty kontrolnej. Pojawienie się obserwacji leżącej poza



Rys. 8.1. Gęstość dwuwymiarowego rozkładu normalnego o parametrach $\Sigma = \mathrm{diag}(0,1\ 0,1)$, $M = (0,6\ 0,6)$ przetransformowana za pomocą quasi-odwrotności krzywej Peano
Fig. 8.1. 2-dimensional normal with $\Sigma = \mathrm{diag}(0.1\ 0.1)$, $M = (0.6\ 0.6)$, after transformation via quasi-inverse of the Peano space-filling curve

przedziałem kontrolnym jest sygnałem do ,,wszczęcia alarmu". Złożenie większej liczby rozłącznych przedziałów kontrolnych będziemy nazywać multi–kartą. Podstawą idei konstrukcji multi–karty jest niezmienniczość odpowiednich prawdopodobieństw w stanach: prawidłowym i rozregulowania, po przetransformowaniu wektorów mierzonych wielkości na odcinek I_1.

Zauważmy ponadto, że jesteśmy w stanie aproksymować wspomniane prawdopodobieństwa z zadaną dokładnością, naturalnie przy założeniu, że posiadamy dostatecznie dużą liczbę obserwacji zachowania procesu w stanie prawidłowym. Dokładność ta może być określana za pomocą liczby rozłącznych podprzedziałów, które aproksymują odpowiednie obszary decyzji w odniesieniu do procesów reprezentowanych na odcinku I_1. Wyznaczone podprzedziały w I_1 definiują multi–kartę.

Niech A_d będzie obszarem akceptacji oryginalnego problemu obserwowanego w przestrzeni wielowymiarowej I_d, wyznaczonym w taki sposób, że wartość ARL_0 w stanie prawidłowym jest równa $1/p$. Pomijamy tu problem konstrukcji obszaru $A_d \subset I_d$, gdyż trudności w jego konstruowaniu są właśnie jednym z powodów konstrukcji multi–karty na podstawie przetransformowanych, jednowymiarowych danych. Dla celów rozważań teoretycznych wystarczy założenie, że taki obszar istnieje.

Oznaczmy przez $A = \boldsymbol{\Psi}(A_d)$ obszar akceptacji po transformacji na odcinek I_1. Z własności transformacji za pomocą krzywej wypełniającej F mamy

$$\int_{A_d} f(x)dx = \int_A g(t)dt = 1 - p, \tag{8.2}$$

gdzie $g(t) = f(F(t))$. Ponieważ transformacja $\boldsymbol{\Psi}$ zachowuje miarę Lebesgue'a, zatem jeśli zbiór A_d jest zbiorem o minimalnej mierze Lebesgue'a w I_d, takim że $\int_{A_d} f(x)dx = 1 - p$, to także A jest zbiorem o minimalnej długości w I_1, wśród wszystkich zbiorów, dla których zachodzi $\int_A g(t)dt = 1 - p$.

Niech N oznacza liczbę rozłącznych podprzedziałów zawartych w I_1 i mających postać: $[a_1, b_1], [a_2, b_2], \ldots, [a_N, b_N]$ takich, że $0 \leq a_1 < \ldots a_i < b_i < a_{i+1} \ldots b_N \leq 1$. Ponadto niech

$$\mathcal{A}_N = \cup_{i=1}^{N} [a_i, b_i].$$

Korzystając z powyższych oznaczeń, możemy sformułować następującą własność:

Lemat 8.1 *Dla każdego $\epsilon > 0$ istnieje liczba naturalna N oraz zbiór $\mathcal{A}_N$ złożony z N rozłącznych pododcinków zawartych w I_1 taki, że*

$$\sum_{i=1}^{N} \int_{a_i}^{b_i} f(F(t))dt = 1 - p \quad \textit{oraz} \quad \mu_1(\mathcal{A}_N \ominus A) < \epsilon,$$

gdzie $A \ominus B \stackrel{\triangle}{=} A \cup B - A \cap B$, $\mu_1(\cdot)$ oznacza miarę Lebesgue'a na odcinku, natomiast $0 < p < 1$ jest ustalonym poziomem istotności. □

Zauważmy również, że jeżeli funkcja gęstości $f(x)$ jest ograniczona od góry przez wartość $f_{\max}$, to z warunku $\mu_1(\mathcal{A}_N \ominus \boldsymbol{\Psi}(A)) < \epsilon$ wynika także spełnienie warunku $P[\mathcal{A}_N \ominus \boldsymbol{\Psi}(A)] < \epsilon \cdot f_{\max}$.

ALGORYTM MULTI–KARTY TYPU SHEWHARTA

Krok 1. Przetransformuj obserwację X do postaci jednowymiarowej $t = \boldsymbol{\Psi}(X)$.

Krok 2. Wyznacz

$$c = \arg \min_{0 \leq i \leq N} |V_i - t|, \tag{8.3}$$

gdzie $V_i = (b_i + a_i)/2, \;\; i = 1, \ldots, N$.

Krok 3. Jeżeli $t \in [a_c, b_c]$, podejmij decyzję: „proces w stanie prawidłowym", w przeciwnym przypadku podejmij decyzję o alarmie – „proces w stanie rozregulowania".

Pozostaje pokazać, jak w powyższym algorytmie należy dobierać granice przedziałów a_k, b_k, co zostanie omówione w następnych podrozdziałach.

8.3 Konstrukcja multi–karty na bazie histogramu

Niech $Z_1, \ldots, Z_n$ będą niezależnymi obserwacjami wielowymiarowego procesu w stanie prawidłowym po przetransformowaniu na odcinek I_1. Na ich podstawie konstruowana będzie multi–karta. Wyznaczmy histogram $g_n(t), t \in I_1$, który jest estymatą gęstości rozkładu $g(t) = f(F(t)),\ t \in I_1$ na odcinku I_1. Oznaczmy przez h_n szerokość przedziałów histogramu. Dalej, niech p, jak do tej pory, określa poziom istotności. Z własności histogramu wynika, że funkcja $g_n(t)$ określona na odcinku I_1 jest funkcją prostą złożoną z co najwyżej $\lceil h_n^{-1} \rceil < n$ przedziałów o stałej wartości. W związku z tym zbiór ufności o minimalnej mierze Lebesgue'a na odcinku, oznaczany A_n, można wyznaczyć za pomocą prostego algorytmu zachłannego o złożoności $O(n \log n)$. Dla uproszczenia notacji wygodniej będzie, zamiast zbioru A_n, konstruować wprost jego dopełnienie $C_n = I_1 - A_n$. C_n jest zbiorem określającym obszar krytyczny.

Zbiór A_n składa się ze skończonej, nie większej niż $\lceil h_n^{-1} \rceil$, liczby rozłącznych przedziałów będących podzbiorami I_1. Podobnie, zbiór C_n jest sumą skończonej liczby rozłącznych przedziałów otwartych zawartych w I_1. W konsekwencji $\mathcal{A}_N = A_n$, przy czym liczba podprzedziałów $N \leq \lceil h_n^{-1} \rceil$.

Dalej, niech $C^\star$ oznacza optymalny zbiór krytyczny na odcinku I_1 wyznaczony na poziomie istotności p. W konsekwencji zachodzi

$$\int_{C^\star} g(t)dt = p, \quad \int_{C_n} g_n(t)dt = p,$$

gdzie $C^\star$ oraz C_n są zbiorami o maksymalnej długości.

Stąd możemy wnioskować:

$$\int_C g(t)dt = p \implies \mu_1(C) \leq \mu_1(C^\star)$$

oraz

$$\int_C g_n(t)dt = p \implies \mu_1(C) \leq \mu_1(C_n).$$

Niech

$$\varepsilon_n = \int_0^1 |g_n(t) - g(t)|dt \tag{8.4}$$

oznacza błąd estymacji $g(t)$ według normy w L_1.

Wiadomo, że w przypadku, gdy gęstość $g(t)$ estymowana jest za pomocą histogramu o szerokości przedziałów h_n, zależnej od liczby obserwacji n, można udowodnić (por. [35]), że zachodzi co następuje:

Lemat 8.2 *Jeśli*

$$h_n \to 0, \quad \textit{oraz} \quad n\,h_n \to \infty, \tag{8.5}$$

gdy $n \to \infty$, *to*

$$\text{i)} \quad \varepsilon_n \to 0 \quad \textit{według prawdopodobieństwa.}$$

$$\text{ii)} \quad E[\varepsilon_n] \to 0.$$

Dla każdego $\epsilon > 0$ *istnieje takie* n_0, *że dla* $n > n_0$

$$\text{iii)} \quad P[\varepsilon_n - E[\varepsilon_n]| > \epsilon] \leq 2\exp(-n\epsilon^2/2),$$

a w rezultacie

$$\text{iv)} \quad \varepsilon_n \to 0 \quad \textit{z prawdopodobieństwem 1.}$$

□

W tym miejscu możemy wskazać na twierdzenie Abou–Jaoude (patrz [35] twierdzenie 5 z rozdziału 2) jako na jeden z możliwych konstruktywnych dowodów lematu 8.1.

Z własności histogramu określonego na zbiorze domkniętym I_1 wynika, że $\int_0^1 g_n(t)dt = 1$. W związku z tym możemy skorzystać z równości Scheffego [35], [37] w postaci:

$$\sup_{B\in\mathcal{B}} |\int_B g_n(t)dt - \int_B g(t)dt| = 1/2 \int_0^1 |g_n(t) - g(t)|dt, \tag{8.6}$$

gdzie $\mathcal{B}$ jest klasą wszystkich zbiorów borelowskich w I_1.

Z twierdzenia Scheffego wynika, że

$$|\int_{C^\star} g_n(t)dt - p| = |\int_{C^\star} g_n(t)dt - \int_{C^\star} g(t)dt| \leq 1/2 \int_0^1 |g_n(t) - g(t)|dt = \varepsilon_n/2$$

oraz analogicznie

$$|p - \int_{C_n} g(t)dt| = |\int_{C_n} g_n(t)dt - \int_{C_n} g(t)dt| \leq \varepsilon_n/2.$$

Otrzymujemy zatem:

$$|\int_{C^\star} g_n(t)dt - \int_{C_n} g_n(t)dt| = |\int_{C^\star} g_n(t)dt - p| \leq \varepsilon_n/2 \tag{8.7}$$

oraz

$$|\int_{C^\star} g(t)dt - \int_{C_n} g(t)dt| = |p - \int_{C_n} g(t)dt| \leq \varepsilon_n/2. \tag{8.8}$$

Z nierówności (8.8) wynika następujące twierdzenie.

Twierdzenie 8.3.1 *Jeśli spełnione są założenia lematu 8.2, czyli warunki (8.5), to przy* $n \to \infty$ *wartość* $\int_{C_n} g(t)dt$ *dąży do* p *z prawdopodobieństwem* 1. □

Z powyższego twierdzenia nie możemy wyciągać żadnych wniosków na temat asymptotycznych własności miary Lebesgue'a zbioru C_n oraz zbioru $C^\star \ominus C_n$, możemy natomiast udowodnić co następuje.

Twierdzenie 8.3.2 *Jeśli spełnione są założenia lematu 8.2, to przy* $n \to \infty$ *miara Lebesgue'a zbioru* C_n, *czyli* $\mu_1(C_n)$, *dąży do* $\mu_1(C^\star)$ *z prawdopodobieństwem* 1. □

Zanim przejdziemy do dowodu powyższego twierdzenia, należy zauważyć, że twierdzenie to, w ogólnym przypadku, nie może gwarantować zbieżności zbioru C_n do zbioru $C^\star$ w tym sensie, że $\mu_1(C^\star \ominus C_n)$ dąży do zera przy $n \to \infty$.

Dowód. Zakładamy, że h_n spełnia warunki (8.5). Z nierówności (8.7) wynika, że

$$|\int_{C^\star} g_n(t)dt - p| \leq \varepsilon_n/2.$$

Jeśli $\int_{C^*} g_n(t)dt \leq p$, to istnieje taki zbiór Δ_n, rozłączny z $C^\star$ i taki, że

$$\int_{C^\star \cup \Delta_n} g_n(t)dt = p.$$

Wtedy, zgodnie z definicją C_n, mamy

$$\mu_1(C^\star \cup \Delta_n) = \mu_1(C^\star) + \mu_1(\Delta_n) \leq \mu_1(C_n).$$

Ponieważ $\mu_1(\Delta_n) \geq 0$, zatem $\mu_1(C^\star) \leq \mu_1(C_n)$. W przypadku, gdy zachodzi $\int_{C^*} g_n(t)dt \geq p$, istnieje taki zbiór Δ_n zawarty w $C^\star$, że

$$\int_{C^\star - \Delta_n} g_n(t)dt = p.$$

W konsekwencji

$$\mu_1(C^\star - \Delta_n) = \mu_1(C^\star) - \mu_1(\Delta_n) \leq \mu_1(C_n).$$

Jeżeli ponadto

$$\varepsilon_n/2 \leq p(1 - \mu_1(C^\star)),$$

to istnieje taki zbiór Δ_n zawarty w $C^\star$, że dodatkowo $g_n(t) \geq p$ dla każdego $t \in \Delta_n$ i w konsekwencji $\mu_1(\Delta_n) \leq \varepsilon_n/(2p)$.

Analogicznie, z nierówności (8.8) otrzymujemy

$$|\int_{C_n} g(t)dt - p| \leq \varepsilon_n/2.$$

Jeśli $\int_{C_n} g(t)dt \leq p$, to istnieje taki zbiór $\hat{\Delta}_n$, rozłączny z C_n, że

$$\int_{C_n \cup \hat{\Delta}_n} g(t)dt = p.$$

Wtedy zgodnie z definicją $C^\star$, mamy

$$\mu_1(C_n \cup \hat{\Delta}_n) = \mu_1(C_n) + \mu_1(\hat{\Delta}_n) \leq \mu_1(C^\star).$$

Ponieważ $\mu_1(\hat{\Delta}_n) \geq 0$, więc $\mu_1(C^\star) \geq \mu_1(C_n)$.

Gdy natomiast $\int_{C_n} g(t)dt \geq p$, wtedy istnieje taki zbiór $\hat{\Delta}_n$ zawarty w C_n, że

$$\int_{C_n - \hat{\Delta}_n} g(t)dt = p.$$

W konsekwencji

$$\mu_1(C_n - \hat{\Delta}_n) = \mu_1(C_n) - \mu_1(\hat{\Delta}_n) \leq \mu_1(C^\star).$$

Jeżeli ponadto, analogicznie jak poprzednio, spełniony jest warunek

$$\varepsilon_n/2 \leq p(1 - \mu_1(C^\star)),$$

to istnieje taki zbiór $\hat{\Delta}_n$ zawarty w C_n, że dodatkowo dla każdego $t \in \hat{\Delta}_n$ zachodzi $g(t) \geq p$ i otrzymujemy $\mu_1(\hat{\Delta}_n) \leq \varepsilon_n/(2p)$.

W dowodzie będziemy dalej korzystać z lematu Borela–Cantelli [142], co wiąże się z koniecznością oszacowania dla dowolnego $\epsilon > 0$ prawdopodobieństwa zdarzenia $|\mu_1(C_n) - \mu_1(C^\star)| > \epsilon$, a mianowicie

$$\begin{aligned} &P[|\mu_1(C_n) - \mu_1(C^\star)| > \epsilon] \\ &\leq P[\mu_1(C_n - \mu_1(C^\star) > \epsilon] + P[-\mu_1(C_n + \mu_1(C^\star) > \epsilon] \\ &\leq P[\mu_1(\hat{\Delta}_n) > \epsilon] + P[\mu_1(\Delta_n) > \epsilon]. \end{aligned} \tag{8.9}$$

Dalej mamy

$$P[\mu_1(\Delta_n) > \epsilon] = 1 - P[\mu_1(\Delta_n) \leq \epsilon].$$

Z kolei $P[\mu_1(\Delta_n) \leq \epsilon]$ jest nie mniejsze niż

$$\begin{aligned} &P[\varepsilon_n \leq 2p(1 - \mu_1(C^\star))]\, P[\varepsilon_n/p \leq \epsilon] \\ &= (1 - P[\varepsilon_n > 2p(1 - \mu_1(C^\star))])\,(1 - P[\varepsilon_n/p > \epsilon]). \end{aligned}$$

Z iii) lematu 8.2 wynika, że

$$\begin{aligned} &P[\varepsilon_n > 2p(1 - \mu_1(C^\star))] \\ &\leq 2\exp[-n(2p(1 - \mu_1(C^\star)) - E[\varepsilon_n])^2/2] = 2\exp(-nc_{1n}) \end{aligned}$$

oraz

$$P[\varepsilon_n > p\epsilon] \leq 2\exp[-n(p\epsilon - E[\varepsilon_n])^2/2] = 2\exp(-nc_{2n}),$$

stąd

$$\begin{aligned} P[\mu_1(\Delta_n) > \epsilon] &\leq 1 - (1 - 2\exp(-nc_{1n}))(1 - 2\exp(-nc_{2n})) \\ &< 4(\exp(-nc_{1n}) + \exp(-nc_{2n})). \end{aligned}$$

Ponieważ $E[\varepsilon_n]$ dąży do zera, dla każdego $\epsilon_1 > 0$ istnieje więc takie naturalne n_0, że dla wszystkich $n > n_0$ wartość $E[\varepsilon_n] < \epsilon_1$. Dalej otrzymujemy $\exp(-nc_{in}) < \exp(-nc_{n_0})$, $i = 1, 2$, gdzie $c_{n_0} > 0$. W konsekwencji

$$\sum_{n=n_0}^{\infty} P[\mu_1(\Delta_n) > \epsilon] \leq 8\exp(-nc_{n_0}) < \infty.$$

Analogiczne rozumowanie możemy przeprowadzić w odniesieniu do $\mu_1(\hat{\Delta}_n)$. Rozumowanie to pozwala pokazać zbieżność zupełną (por. [142]), odpowiednio, ciągu zmiennych losowych $\mu_1(\Delta_n)$ oraz $\mu_1(\hat{\Delta}_n)$. Na podstawie (8.9) możemy dalej wnioskować o zbieżności zupełnej do zera ciągu zmiennych losowych $\mu_1(C_n) - \mu_1(C^*)$. Zbieżność zupełna pociąga za sobą zbieżność z prawdopodobieństwem 1, co kończy dowód twierdzenia. □

8.4 Neuronowe algorytmy konstrukcji multi-karty

Do konstrukcji multi-karty może zostać zastosowany algorytm uczenia typu wektorowej kwantyzacji Kohonena [83] (patrz rozdział 6). Mówiąc w znacznym uproszczeniu, algorytm Kohonena spełnia tu rolę estymatora gęstości rozkładu.

W rozważanym przypadku proces adaptacyjnego ustalania pozycji kwantyzatorów powinien uwzględniać możliwość odrzucania obserwacji, jeśli wskazuje ona na prawdopodobieństwo rozregulowania procesu. Wymaga to wprowadzenia dodatkowych parametrów sieci, które definiują obszar oddziaływania danego kwantyzatora V_i w postaci przedziału $[V_i - s_i, V_i + s_i]$.

Zaproponowany poniżej algorytm uczenia służy optymalizacji pozycji kwantyzatorów (liczb z odcinka I_1), przy równoczesnym zachowaniu zadanego poziomu istotności p. W konsekwencji zaprojektowanie multi-karty wymaga nie tylko ustalenia pozycji kwantyzatorów $V_1, V_2, \ldots, V_N$, ale także szerokości indywidualnych stref akceptacji s_i, $i = 1, \ldots, N$. Na ich podstawie można podać wprost parametry multi-karty: $a_i = V_i - s_i$, $b_i = V_i + s_i$.

Niech ciąg $t_1 = \boldsymbol{\Psi}(X_1), \ldots, t_n = \boldsymbol{\Psi}(X_n)$ będzie ciągiem niezależnych obserwacji wielowymiarowego procesu w stanie normalnym, przekształconych do postaci jednowymiarowej za pomocą transformacji quasi-odwrotnej $\boldsymbol{\Psi}$.

Na podstawie tego ciągu będziemy konstruować multi-kartę o zadanym poziomie istotności p.

Zdefiniujmy funkcję

$$K_i(t) = \begin{cases} 1, & \text{gdy } t \in [V_i - s_i, V_i + s_i], \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Estymatorem (typu Monte–Carlo) rzeczywistego prawdopodobieństwa akceptacji obserwacji przez multi-kartę, zdefiniowaną przez przedziały $[V_i - s_i, V_i + s_i]$, $i = 1, \ldots, N$, jest wyrażenie

$$1/n \sum_{j=1}^{n} \sum_{i=1}^{N} K_i(t_j). \tag{8.10}$$

Pod koniec procesu uczenia wartość wyrażenia (8.10) powinna być bliska zadanemu poziomowi akceptacji $1 - p$. Podobnie,

$$1/n \sum_{j=1}^{n} \sum_{i=1}^{N} (1 - K_i(t_j)) = N - 1/n \sum_{j=1}^{n} \sum_{i=1}^{N} K_i(t_j) \tag{8.11}$$

jest estymatorem aktualnego poziomu istotności multi-karty.

Na poziom ten składają się indywidualne wartości poziomu odrzucania, związane z każdym z kwantyzatorów, które określają częstość odrzucania realizowaną indywidualnie przez każdy z przedziałów multi-karty. Wartość tę szacuje następujące wyrażenie:

$$1/n \sum_{j=1}^{n} (1 - K_i(t_j)) = 1 - 1/n \sum_{j=1}^{n} K_i(t_j). \tag{8.12}$$

Poniżej przedstawimy algorytm uczenia multi-karty, którego celem jest ustalenie takich wartości V_i, s_i, by wszystkie wartości (8.12) były równe p/N, a tym samym, by projektowana multi-karta osiągnęła poziom odrzucania równy zadanej wartości p.

ALGORYTM I

Algorytm uczenia składa się z dowolnej liczby epok. Jedna epoka uczenia obejmuje przetworzenie całego ciągu uczącego $t_1, \ldots, t_n$. Początkowe parametry multi-karty V_1, s_i, dla pierwszej epoki uczenia, są ustalane arbitralnie.

Przyjmij początkowe wartości $p_i = 0$, $i = 1, 2, \ldots, N$. Dla każdej obserwacji $t_j \in [0, 1]$ związanej z momentem czasowym $j = 1, 2, \ldots, n$ wykonaj kroki 1–4.

Krok 1. Wyznacz numer kwantyzatora $V_{i^\star}$, który jest najbliższym sąsiadem względem obserwacji t_j, czyli $i^\star = \arg\min_i |V_i - t_j|$.

Krok 2. Jeśli $|V_{i^\star} - t_j| > s_i$, to odrzuć obserwację t_j i zaktualizuj indywidualną wartość poziomu odrzucania $p(i^\star) = p(i^\star) + 1$.

Krok 3. Jeżeli $|V_{i^\star} - t| \leq s_i$, to zaakceptuj obserwację i zaktualizuj pozycje kwantyzatora, czyli podstaw $V_{i^\star} := V_{i^\star} + \eta_1(t - V_{i^\star})$, gdzie η_1 jest współczynnikiem uczenia.

Krok 4. Zaktualizuj indywidualne szerokości stref akceptacji $s_i := s_i + \eta_2(p_i/j - p/N)$, gdzie $\eta_2 \in (0,1)$ jest wybranym parametrem uczenia.

Ostatni krok algorytmu można wykonywać co jakiś czas, w skrajnym przypadku tylko raz pod koniec całej epoki uczenia, gdyż wiąże się on z procesem całkowania, którego efekty ujawniają się dopiero po upływie pewnego czasu (określonej liczby iteracji). Proces uczenia można zakończyć, gdy indywidualne poziomy odrzucania, wyznaczone empirycznie, czyli wartości p_i/n są wystarczająco bliskie wymaganym wartościom p/N.

W uproszczonej wersji algorytmu wszystkie szerokości stref akceptacji s_i mogą być takie same. Przy większej liczbie kwantyzatorów daje to potencjalnie te same rezultaty jak w przypadku indywidualnych szerokości stref akceptacji.

Odpowiednią modyfikację algorytmu przedstawiono poniżej.

ALGORYTM II

Rozpocznij z wartością $\hat{p} = 0$. Dla każdej obserwacji $t_j \in [0,1]$ związanej z momentem czasowym $j = 1, 2, \ldots, n$ wykonaj kroki 1–4.

Krok 1. Wyznacz kwantyzator $V_{i^\star}$, który jest najbliższym sąsiadem w stosunku do obserwacji t_j, czyli wyznacz
$i^\star = \arg\min_i |V_i - t_j|$.

Krok 2. Jeśli $|V_{i^\star} - t_j| > s$, to odrzuć obserwację t_j i zaktualizuj licznik odrzuconych obserwacji $\hat{p} := \hat{p} + 1$.

Krok 3. Jeżeli $|V_{i^\star} - t_j| \leq s$, to zaakceptuj obserwację t_j oraz zaktualizuj pozycje wybranego kwantyzatora, czyli dokonaj podstawienia

$$V_{i^\star} := V_{i^\star} + \eta_1(t_j - V_{i^\star}).$$

Krok 4. Zaktualizuj szerokość strefy akceptacji na podstawie aktualnej liczby odrzuconych obserwacji $s := s + \eta_2(\hat{p}/n - p)$.

Podobnie jak w przypadku algorytmu I, krok 4. może być wykonywany incydentalnie, na przykład co 1000 iteracji lub tylko raz na końcu całej epoki uczenia. Proces uczenia kończy się, gdy empiryczna ocena prawdopodobieństwa odrzucenia $\hat{p}/n$ jest wystarczająco bliska ustalonemu poziomowi istotności p.

Oba algorytmy mogą prowadzić do rozwiązań, w których pewne pary przedziałów akceptacji zachodzą na siebie. W konsekwencji efektywna liczba przedziałów multi-karty może być znacznie mniejsza od ustalonej wstępnie liczby N.

8.5 Przykład działania multi-karty

Dane do badania multi-karty zostały wygenerowane z dwuwymiarowego rozkładu normalnego o wektorze średnich $M = (0,5,\ 0,5)$ i diagonalnej macierzy kowariancji $\Sigma = \mathrm{diag}(0,1,\ 0,1)$. Elipsoida ufności ma w tym przypadku kształt okręgu o promieniu zależnym od wartości poziomu istotności p.

Prawdopodobieństwo pojawienia się obserwacji pochodzących spoza kwadratu I_2 jest w badanym przypadku nie tylko bardzo małe, ale także obserwacje te znajdują się poza elipsoidą ufności dla ustalonej tu, typowej wartości $p = 0,005$. Obie zmienne opisujące stacjonarny proces ,,w stanie prawidłowym" są realizacjami niezależnych zmiennych losowych o rozkładzie normalnym i przyjmują wartości z przedziału $[0, 1]$ na poziomie ufności $0,0000003$ (w odniesieniu do każdej zmiennej niezależnie).

Badania przeprowadzono dla multi-karty, w której parametry wyznaczono przy użyciu algorytmu II, na podstawie danych przetransformowanych na odcinek za pomocą quasi-odwrotności krzywej Peano.

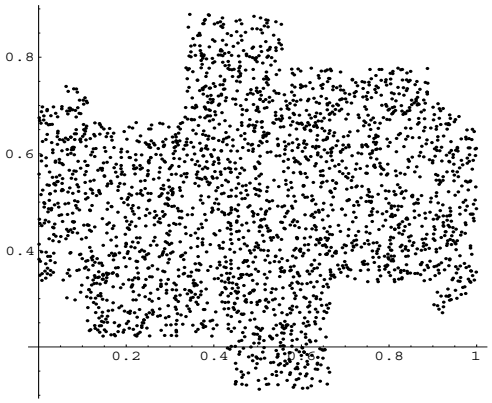
Na rysunkach 8.2, 8.3 przedstawiono wielowymiarowe obszary akceptacji powstałe poprzez przetransformowanie, za pomocą krzywej Peano, multi-karty złożonej z trzech (rysunek 8.2) bądź czterech (rysunek 8.3) rozłącznych przedziałów z odcinka I_1.

Jako punkt odniesienia przy badaniu multi-karty przyjęto zmiany procesu polegające na addytywnej zmianie wektora wartości oczekiwanych o wektor ΔM_q, począwszy od pewnego momentu czasowego q. Stąd obserwowany proces ma postać:

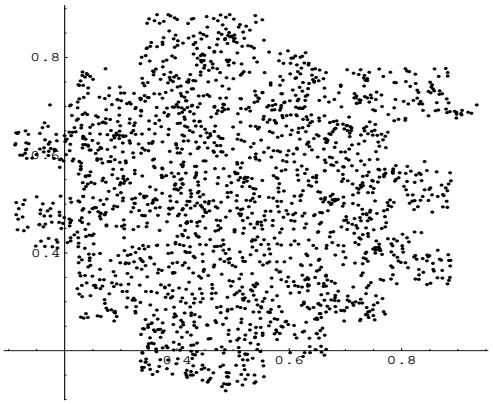
$$X_k = Y_k + \Delta M_q,\ k \geq q,$$

gdzie Y_k ma rozkład taki jak rozkład obserwacji procesu w stanie prawidłowym (bez zakłóceń). Dla uproszczenia przyjęto dalej, że $q = 1$.

Typową kartą kontrolną, którą stosuje się, gdy możemy założyć, że Y_k mają rozkład normalny, jest karta oparta na statystyce T^2 (por. [104], [86]) zgodnie z regułą (8.1). Wprowadźmy współczynnik $\lambda = ||\Delta M_q||/0,1$.



Rys. 8.2. Obszar akceptacji (trzy przedziały na odcinku) dwuwymiarowego rozkładu normalnego o parametrach $\Sigma = \mathrm{diag}(0,1,\ 0,1)$, $M = (0,5,\ 0,5)$ przetransformowanego do I_2 za pomocą krzywej Peano, $p = 0,005$. Obszar akceptacji multi-karty został wyznaczony za pomocą algorytmu II i przetransformowany z powrotem do kwadratu I_2
Fig. 8.2. Acceptance region (three subintervals) of the 2-dimensional normal distribution, after transformation via Peano space-filling curve. $\Sigma = \mathrm{diag}(0.1,\ 0.1)$, $M = (0.5,\ 0.5)$, $p = 0.005$. The acceptance region was calculated using Algorithm II and transformed back to I_2



Rys. 8.3. Obszar akceptacji (cztery przedziały na odcinku) dwuwymiarowego rozkładu normalnego o parametrach $\Sigma = \mathrm{diag}(0,1,\ 0,1)$, $M = (0,5,\ 0,5)$ z $p = 0,005$, przetransformowany do I_2 za pomocą krzywej Peano. Przedziały multi-karty zostały wyznaczone za pomocą algorytmu II
Fig. 8.3. Acceptance region (four subintervals) of the 2-dimensional normal distribution, transformed by Peano space-filling curve. $\Sigma = \mathrm{diag}(0.1,\ 0.1)$, $M = (0.5,\ 0.5)$, $p = 0.005$. The multi-cart acceptance region was calculated using Algorithm II and transformed back to I_2

Współczynnik λ ocenia wielkość zmiany wartości oczekiwanej względem dyspersji równej $0,1$. Wartość $\lambda = 0$ oznacza, że w procesie nie zaszły żadne zmiany.

ARL dla procesu „normalnego" (czyli ARL_0) został oszacowany na podstawie średniej z 10^5 symulacji, natomiast ARL wykrycia zmiany uśredniano na podstawie $2 \cdot 10^4$ przebiegów. Wrażliwość multi-karty na zmianę wartości średniej procesu zależy nie tylko od względnej wielkości zmiany λ, ale także od kierunku, w jakim nastąpiła zmiana (patrz rysunki 8.2, 8.3).

By móc porównać wyniki z kartą T^2, wartości ARL odpowiadające ustalonemu $\lambda > 0$ zostały wyznaczone poprzez uśrednienie odpowiednich wyników otrzymanych dla ośmiu różnych kierunków wektora o stałej długości ΔM_q.

W tabeli 8.1 przedstawiono porównanie wartości ARL obu kart: klasycznej karty T^2 i multi-karty z rysunku 8.2. W kolumnie czwartej tabeli podano odchylenie standardowe eksperymentalnych wartości ARL badanej multi-karty. Osiągnięcie podanej dokładności wymagało od 10^4 do 10^5 powtórzeń procedury użycia multi-karty polegających na wygenerowaniu ciągu obserwacji „w stanie zakłóconym", ze zmienionym wektorem wartości oczekiwanych, aż do momentu, gdy kolejna obserwacja zostanie odrzucona przez multi-kartę. Wynikiem pojedynczego eksperymentu jest numer obserwacji, która została odrzucona.

Tabela 8.1. Porównanie ARL multi-karty typu Shewharta dla danych przetransformowanych za pomocą krzywej Peano z wynikami uzyskanymi dla statystyki Hotellinga T^2. Druga i trzecia kolumna zawierają średni czas do fałszywego alarmu w sytuacji „normalnej" ($\lambda = 0$) oraz średni czas wykrycia zmiany $\lambda > 0$

λ	ARL dla T^2 ($h = 10,6$)	ARL dla multi-karty	Odchylenie standardowe ARL multi-karty
0,0	200	203	0,6
0,5	116	107	0,1
3,0	2,2	2,77	0,02

Dane dla wielowymiarowej karty T^2 ($d = 2$) pochodzą z pracy [104], choć naturalnie można je również wyznaczyć korzystając z tablic rozkładu χ^2. Jak widać, już w przypadku prostej multi-karty złożonej z trzech przedziałów można, tylko na podstawie obserwacji procesu w stanie prawidłowym, skonstruować algorytm wykrywania zmian w procesie, porównywalny (w sensie średnim) z algorytmem T^2. Należy zauważyć, że algorytm T^2, w przypadku rozkładu normalnego, jest optymalnym algorytmem wykrywania zmiany procesu na podstawie pojedynczych obserwacji.

8.6 Podsumowanie

W powyższym rozdziale zaproponowano nową metodykę wykrywania zmian w monitorowanym procesie. Wstępne idee proponowanego podejścia były prezentowane jako zaproszony referat na International Workshop on New Developments in Quality Control (July 2,3, 1999, Frankfurt) [166].

Krzywe wypełniające stanowią wygodne narzędzie do transformacji wielowymiarowych danych do postaci jednowymiarowej. Transformacja ta jest niezmiennicza względem miary probabilistycznej w tym sensie, że zachowuje prawdopodobieństwa odpowiadających sobie zdarzeń w przestrzeni wielowymiarowej i po transformacji na odcinek I_1.

Wielowymiarowe obszary decyzji, to znaczy obszary związane z decyzjami o stanie procesu: ,,proces w stanie prawidłowym" lub ,,proces zakłócony", transformowane są do postaci kilku rozłącznych przedziałów zawartych w I_1. Przedziały te tworzą multi-kartę kontrolną pozwalającą oceniać stan wielowymiarowego procesu na podstawie jednowymiarowej obserwacji.

Proponowane podejście można stosować nie tylko w odniesieniu do oryginalnych obserwacji wielowymiarowego procesu, ale także do danych przetworzonych (na przykład wielowymiarowych residuów itd.).

Ze względu na brak założeń dotyczących postaci rozkładów, analizowane algorytmy mają charakter nieparametryczny. Założenia czynione w odniesieniu do monitorowanego procesu dotyczą zatem tylko istnienia odpowiednich rozkładów prawdopodobieństwa oraz ewentualnie niezależności kolejnych obserwacji procesu.

Zbadano własności teoretyczne multi-karty wyznaczonej na podstawie histogramu. Jeśli gęstość rozkładu na odcinku jest estymowana za pomocą histogramu o szerokości przedziałów h_n, zależnej od liczby obserwacji n, w taki sposób, że $h_n \to 0$ w odpowiednim tempie, to przedziały ufności multi-karty empirycznej mają asymptotycznie te same prawdopodobieństwa i łączną długość co przedziały wyznaczone przy znajomości teoretycznego rozkładu.

Oprócz powyższego algorytmu, posiadającego pełną podbudowę teoretyczną, podano również dwa algorytmy heurystyczne konstrukcji multi-karty, w których używa się algorytmów uczenia stosowanych w samoorganizujących sieciach neuronowych [83], [165], [168].

Algorytm uczenia typu Kohonena pozwala na efektywne wyznaczenie granic decyzji w multi-karcie oraz na adaptacyjne, prowadzone na bieżąco na podstawie aktualnych obserwacji, modyfikacje granic decyzji wyrażonych w postaci przedziałów multi-karty.

Zbadany w niniejszym rozdziale algorytm (typu Kohonena) nie ma wprawdzie w pełni określonych własności asymptotycznych, jednakże działa na bieżąco

i w konsekwencji nie wymaga przechowywania i łącznego przetwarzania dużej liczby danych.

Stosowane do tej pory metody statystyczne wymagają znajomości rozkładu monitorowanego procesu i na ogół zakładają, że rozkład ten jest rozkładem normalnym. Zaproponowane tu podejście stanowi atrakcyjne narzędzie w przypadku gdy wiadomo, że sterowany proces nie spełnia tego założenia.

Transformacja wielowymiarowych danych do postaci jednowymiarowej otwiera nowe możliwości w konstruowaniu algorytmów jakościowego diagnozowania stanu procesu.

Rozdział 9

Uwagi końcowe

W monografii opracowano jednolitą, stosującą układy równań funkcyjnych, metodykę definiowania krzywych wypełniających, prowadzącą wprost do efektywnych algorytmów obliczania wartości krzywej i jej quasi-odwrotności. Podano rekurencyjne metody wyznaczania odwzorowań quasi-odwrotnych do wielowymiarowych krzywych typu Peano, Hilberta i Sierpińskiego. Złożoność obliczeniowa uzyskiwanych algorytmów jest liniowo zależna od wymiaru wypełnianej przez krzywą przestrzeni.

Rozdziały 2–4 stanowią oryginalne opracowanie na temat definiowania, obliczania i badania własności krzywych wypełniających zachowujących miarę Lebesgue'a. W szczególności podano w niektórych przypadkach niepoprawialne wartości stałych w odpowiednich warunkach Höldera.

Na uwagę zasługuje także, podana w twierdzeniu 3.4.1 własność, dotycząca maksymalnej wartości wykładnika Höldera uzyskiwanej dla wielowymiarowych krzywych wypełniających, powstałych poprzez złożenie odwzorowań dwuwymiarowych. Z twierdzenia tego wynika słaba przydatność praktyczna tego typu metod definiowania wielowymiarowych krzywych wypełniających.

Zidentyfikowano własności krzywych wypełniających, które są istotne z punktu widzenia zastosowań w problemach decyzyjnych. Są to własności zachowywania miary Lebesgue'a przez krzywą wypełniającą oraz jej quasi-odwrotność, a także optymalna dla danego wymiaru przestrzeni wartość wykładnika w warunku Höldera. Uzyskanie tych własności gwarantują, wprost lub pośrednio, sformułowane w rozdziale 4 warunki **C1**–**C3**.

Wyprowadzono teoretyczną zależność między wymiarem pudełkowym wielowymiarowych danych a wymiarem pudełkowym danych przetransformowanych przy użyciu krzywej wypełniającej. Wyniki tych badań stały się podstawą do zaproponowania nowej, łatwiejszej obliczeniowo, metody oceny wymiaru fraktalnego obiektów wielowymiarowych. Monitorowanie wymiaru fraktalnego bieżących

obserwacji może stanowić podstawę do podejmowania decyzji na temat aktualnego stanu systemu.

Zdefiniowano nową klasę krzywych wypełniających, które zachowują zadaną miarę probabilistyczną. Krzywe te mogą być teoretycznym narzędziem, mającym zastosowanie w kwantyzacji wielowymiarowych danych.

Pokazano, że asymptotyczna wartość dystorsji (6.11) kwantyzatorów przetransformowanych z odcinka I_1 do kostki I_d maleje wraz z N w sposób typowy dla wielowymiarowych kwantyzatorów.

Udowodniono, że wyspecyfikowana w monografii klasa krzywych wypełniających zachowuje ryzyko Bayesa dla dowolnego rozkładu danych o nośniku zawartym w wielowymiarowej kostce I_d. Zbadano własności separujące krzywych wypełniających.

Wprowadzono pojęcie powierzchni wypełniającej, które jest ciągłym odwzorowaniem przekształcającym kwadrat jednostkowy $I_2 = [0,1] \times [0,1]$ w wielowymiarową kostkę I_d i odpowiednie odwzorowanie quasi-odwrotne oraz zbadano ich podstawowe własności.

Na podstawie wymienionych wyżej rezultatów opracowano metodologię rozwiązywania wielowymiarowych problemów decyzyjnych, polegającą na przekształceniu zbioru wielowymiarowych danych, które stanowią punkt wyjścia w procesie decyzyjnym, w ciąg danych jednowymiarowych, zawartych w odcinku jednostkowym. Ważnym elementem tej metodologii jest dobre zdefiniowanie odwzorowania quasi-odwrotnego do krzywej wypełniającej. Złożoność obliczeniowa takiej transformacji danych jest liniowa ze względu na wymiar danych. Transformacja każdego elementu zbioru danych za pomocą quasi-odwrotności krzywej wypełniającej może odbywać się niezależnie, w dowolnym momencie czasowym i nie wymaga konstrukcji całej krzywej wypełniającej. Transformacja quasi-odwrotna pozwala uporządkować liniowo dane z przestrzeni wielowymiarowej.

Metodykę tę zastosowano w odniesieniu do wielowymiarowych problemów rozpoznawania, problemów statystycznego wykrywania zmian w procesie oraz problemów kwantyzacji. Zaowocowało to powstaniem wielu szczegółowych metod o określonych własnościach teoretycznych. W wielu przypadkach zaproponowano także na ich bazie proste obliczeniowo metody heurystyczne, w większości oparte na własnościach jednowymiarowych sieci Kohonena.
W szczególności opracowano:

- Metody kwantyzacji, które łączą transformację wielowymiarowych danych za pomocą quasi-odwrotności wybranej krzywej wypełniającej ze skalarną kwantyzacją w odniesieniu do danych przetransformowanych na odcinek I_1. Zastosowanie różnych wariantów algorytmów uczenia na bazie danych skalarnych umożliwia kształtowanie gęstości rozkładu kwantyzatorów.

- Podano algorytm aproksymacji krzywej zachowującej zadaną, lecz nieznaną, miarę probabilistyczną na podstawie niezależnych obserwacji wielowymiarowej zmiennej losowej o tymże nieznanym rozkładzie.

- Zaproponowano i zbadano rodzinę klasyfikatorów w postaci rozwinięć w szeregi ortogonalne, w których współczynniki rozwinięcia optymalnej funkcji decyzyjnej estymowane są na podstawie przetransformowanych, jednowymiarowych danych. Wykazano zgodność tego typu klasyfikatorów (przy stosunkowo słabych założeniach dodatkowych). Szczegółowo zbadano algorytm rozpoznawania w przypadku zastosowania układu funkcji Haara na odcinku I_1 jako szeregu ortonormalnego.

- Opracowano szybki algorytm rozpoznawania oparty na metodzie k najbliższych sąsiadów stosowanej do danych przetransformowanych na odcinek.

- Połączenie metody LVQ Kohonena z transformacją danych za pomocą krzywej doprowadziło do uzyskania jeszcze jednej klasy efektywnych algorytmów rozpoznawania.

- Zaproponowano metodę wizualizacji wielowymiarowych danych za pomocą quasi-odwrotności ciągłego odwzorowania, które przekształca kwadrat jednostkowy $I_2 = [0,1] \times [0,1]$ w wielowymiarową kostkę I_d. Metoda ta może być użyta jako podstawa do tworzenia heurystycznych, prostych reguł rozpoznawania wygenerowanych przez obserwatora. Odwzorowanie to można użyć bezpośrednio do reprezentacji wielowymiarowych danych na płaszczyźnie. Pokazano zastosowanie tego typu wizualizacji w rozwiązywaniu problemów rozpoznawania.

- Zdefiniowano pojęcie multi-karty, która jest uogólnieniem tradycyjnej karty kontrolnej i pozwala oceniać stan wielowymiarowego procesu na podstawie przekształconych, skalarnych obserwacji.

- Zbadano własności teoretyczne multi-karty wyznaczonej na podstawie histogramu.

- Zaproponowano również dwa algorytmy heurystyczne konstrukcji multi-karty, stosujące algorytmy uczenia używane w samoorganizujących sieciach neuronowych.

- Uzyskane wyniki teoretyczne poparto licznymi badaniami symulacyjnymi, przy czym jako punkt odniesienia przyjęto stosowane w każdej z dziedzin klasyczne problemy testowe.

Podstawową cechą opracowanych algorytmów jest szybkość samego procesu podejmowania decyzji. W każdym z przypadków należy najpierw rozwiązać pewien, często skomplikowany, problem optymalizacyjny. Jednakże, w końcowym efekcie konstrukcji rozwiązania, otrzymujemy podział odcinka I_1 na zbiór pododcinków, odpowiadających obszarom poszczególnych decyzji.

Proces podejmowania decyzji odbywa się na odcinku, a jego podstawą jest aktualna obserwacja przetransformowana również do postaci jednowymiarowej. Nakład obliczeń potrzebny do przetransformowania za pomocą krzywej wypełniającej d–wymiarowego punktu na odcinek I_1 jest rzędu $O(d)$. W związku z tym nakład pracy potrzebny do dokonania samego procesu podejmowania decyzji ma złożoność obliczeniową rzędu $O(d) + \log_2 m$, gdzie m jest liczbą przedziałów decyzyjnych na odcinku jednostkowym.

Prezentowane wyniki mogą znaleźć zastosowanie w konstrukcji algorytmów sterowania i diagnozowania, w upraszczaniu algorytmów identyfikacji systemów o parametrach rozłożonych itd. Jest naturalne, że już w przypadku omawianej w monografii klasy krzywych wypełniających istnieje wiele szczegółowych problemów w dziedzinie projektowania systemów sterowania i diagnozowania, które wymagają dalszego rozwoju badań teoretycznych i praktycznych.

Wydaje się celowa kontynuacja badań dotyczących zarówno własności wyspecyfikowanej klasy krzywych wypełniających, jak i szukania innych typów krzywych, które poszerzą obszar rozwiązywanych przy ich pomocy problemów decyzyjno-optymalizacyjnych. Rozwijanie tematyki krzywych wypełniających wiąże się także z nierozpatrywanym w monografii problemem szukania odwzorowań zwiększających wymiar problemu, lecz upraszczających potencjalny opis przetwarzanych obserwacji (sygnałów). Pytanie o użyteczność krzywych wypełniających w tym zakresie jest otwartym problemem badawczym.

Literatura

[1] ABEND K., HARLEY T.J., KANAL L.N., *Classification of binary random patterns*, IEEE Trans. on Information Theory, 1965, 11, 538–544.

[2] ALHONIEMI E., SIMULA O., VESANTO J., *Process monitoring and modeling using the self-organizing map*, Integrated Computer Aided Engineering, 1999, 6, 3–14.

[3] ANDREWS D.F., HERZBERG A.M., *Data. A Collection of Problems for Many Fields for Student and Research Workers*, Springer-Verlag, Berlin, 1985.

[4] ANHALT C.S., KRISHNAMURTHY K.S., CHEN P., MELTON D.E., *Competitive learning algorithms for vector quantization,* Neural Networks, 1990, 3, 277–290.

[5] ANSARI A., FINEBERG A., *Image data ordering and compression using Peano scan and LOT* , IEEE Trans. on Consumer Electronics 1992, 38, 436–444.

[6] BARAS J.S., DEY S., *Combined compression and classification with learning vector quantization*, IEEE Trans. on Information Theory, 1999, 45(6), 1911– 1920.

[7] BARNSLEY M., *Fractals Everywhere*, Academic Press, New York, 1988.

[8] BARTHOLDI J.J., PLATZMAN L.K., *Heuristics based on space-filling curves for combinatorial problems in Euclidean space*, Management Science, 1988, 34, 291–305.

[9] BASSEVILLE M., NIKIFOROV I.V., *Detection of Abrupt Changes. Theory and Applications*, Prentice-Hall, Englewood Cliffs, 1993.

[10] BAUER H.U., DER R., HERMAN M., *Controlling the magnification factor of self-organizing feature maps*, Neural Computation, 1996, 8, 757–771.

[11] BENAIM M, FORT J.C., PAGES G., *The AS convergence of the one-dimensional Kohonen algorithm*, Advances in Applied Probability, 1998, 30, 850–869.

[12] BENVENISTE A., METIVIER M., PRIOURET P., *Adaptive Algorithms and Stochastic Approximations*, Springer Verlag, Berlin, New York, 1990.

[13] BISHOP C.M., *Neural Network for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[14] BISHOP C.M., SVENSEN M., WILIAMS C.K., *GMT: the generative topographic mapping*, Neural Computation, 1998, 10(1), 215–234.

[15] BILINGSLEY P., *Probability and Measure*, John Wiley and Sons, New York, 1979.

[16] BLAYO F., CHENEVAL Y. et al., *Enhanced Learning for Evolutive Neural Architecture* – ESPRIT Basic Research Report 6891, 1995.

[17] BOTTOU C., BENGIO Y., *Convergence properties of the K-means algorithms*, Advances in Neural Information Systems,1995, 7, 585–592.

[18] BOUTON C., PAGES G., *Self-Organization and a.s. convergence of the one-dimensional Kohonen algorithm with non-uniformly distributed stimuli*, Stochastic Processes and Their Applications, 1993, 47, 249–274.

[19] BREIMAN L., FRIEDMAN J.H., OLSEN R.A., STONE C.J., *Classification and Regression Trees*, Wadsworth & Brooks, Pacific Grove, 1984.

[20] BRODER A.J., *Strategies for efficient incremental nearest neighbor search*, Pattern Recognition, 1990, 23, 171–178.

[21] BUDINICH M., *Sorting with self-organizing maps*, Neural Computation, 1995, 7, 1188–1190.

[22] BUDINICH M., *A self-organizing neural network for the traveling salesman problem that is competitive with simulated annealing*, Neural Computation, 1994, 8, 416–424.

[23] BUTZ A.R., *Space-filling curves and mathematical programming*, Information and Control, 1968, 12, 314–330.

[24] BUTZ A.R., *Convergence with Hilbert's space-filling curve*, Journal of Computer and System Science, 1969, 3, 128–146.

[25] BUTZ A.R., *Alternative algorithm for Hilbert's space-filling curve*, IEEE Trans. on Computing, 1971, 20, 424–426.

[26] CHERBIT G. (Eds), *Fractals. Non-integral Dimensions and Applications*, John Wiley and Sons, New York, 1990.

[27] CHINRUNGRUENG Ch., SEQUIN C., *Optimal adaptive K-means algorithm with dynamic adjustment of learning rate*, IEEE Trans. on Neural Networks, 1995, 6, 157–169.

[28] COLE A.J., *Halftoning without dither or edge enhancement*, Visual Computer, 1991, 7, 232–246.

[29] COVER T.M., HART P.E., *Nearest neighbor pattern classification*, IEEE Trans. on Information Theory, 1967, 13, 21–27.

[30] CYPKIN J.Z.,*Podstawy teorii układów uczących*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1973.

[31] DEGROOT M.H., *Optymalne decyzje statystyczne*, PWN, Warszawa, 1981.

[32] DEMARTINES P., HERAULT J., *Curvilinear component analysis: a self-Organizing neural network for nonlinear mapping of data sets*, IEEE Trans. on Neural Networks, 1997, 8, 148–154.

[33] DE SA V.R., BALLARD D., *A note on learning vector quantization*, Advances in Neural Information Processing Systems, 1993, 5, 220–227.

[34] DEVROYE L., *Lecture Notes on Bucket Algorithms*, Birhauser, Boston, 1986.

[35] DEVROYE L., GYÖRFI L.,*Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

[36] DEVROYE L., GYÖRFI L., LUGOSI G., *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[37] DEVROYE L., LUGOSI G., *Combinatorial Methods in Density Estimation*, Springer, New York, 2001.

[38] DIAMANTINI C., SPALVIERI S., *Quantizing for minimum average misclassification risk*, IEEE Trans. on Neural Networks, 1998, 9, 174–182.

[39] DJOUADI A., BOULTACHE E., *A Fast algorithm for the nearest-neighbor classifier*, IEEE Trans. PAMI, 1997, 19(3), 277–282.

[40] DUCH W., *Quantitative measures for the self-organizing topographic maps*, Open Systems and Infomation Dynamics, 1995, 2, 295-302.

[41] DUDA R.O., HART P.E., *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.

[42] DUDA R., *Wprowadzenie do topologii*, PWN, Warszawa, 1986.

[43] DUGUNDJI J., *Topology*, Allyn and Bacon Inc., Boston, 1966.

[44] DURBIN R., WILLSHOW D., *An analogue approach to the travelling salesman problem using an elastic net method*, Nature (London), 1987, 336, 689–691.

[45] ENGEL J., *Density estimation with Haar series*, Statistics and Probability Letters, 1990, 9, 111-117.

[46] ENGEL J., *The multiresolution histogram*, Metrica, 1997, 46, 41-57.

[47] ENGELKING R., *Zarys topologii ogólnej*, PWN, Warszawa, 1968.

[48] FALCONER K., *Fractal Geometry. Mathematical Foundations and Applications*, John Wiley and Sons, New York, 1990.

[49] FARAGO A., LINDER T., LUGOSI G., *Fast nearest neighbor search in dissimilarity spaces*, IEEE Trans. PAMI, 1993, 15, 957–962.

[50] FAVATA F., WALKER R., *A study of the applications of Kohonen-type neural networks to travelling salesman problem*, Biological Cybernetics, 1991, 64, 463–468.

[51] FISCHER R.A., *The use of multiple measurements in taxonomic problems*, Annals of Eugenica, 1936, VII, 179–188.

[52] FLEXER A., *Limitations of self-organizing maps for vector quantization and multidimensional scaling*, Advances in Neural Information Processing Systems , 1997, 9, 445–451.

[53] FRIEDMAN J. H., *Flexible Metric Nearest Neighbor Classification*, Technical Report, 1994, Stanford University.

[54] FRITZKE B., *A growing neural gas network learns topologies*, Advances in Neural Information Processing Systems, 1995, 7, 625–632.

[55] FRITZKE B., *The LBG-U method for vector quantization over LGB inspired from neural networks*, Neural Processing Letters, 1997, 5, 35–45.

[56] FORT J.C., PAGES G., *On the a.s. convergence of the Kohonen algorithm with general neighborhood function*, Annals of Applied Probability, 1995, 5(4), 1177–1216.

[57] FUKUNAGA K., *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.

[58] GERSHO A., GRAY R.M., *Vector Quantization and Signal Compression*, Kluwer Academic Publisher, Boston, 1993.

[59] GILBERT E.N., *Gray codes and paths on the n-cube*, Bell System Techn. J., 1957, 37, 815–826.

[60] GILBERT W.J., *A cube-filling Hilbert curve*, Mathematical Intelligencer, 1984, 6(3), 78.

[61] GOTSMAN C., LINDENBAUM M., *On the metric properties of discrete space-filling curves*, IEEE Trans. on Image Processing, 1996, 5(5), 794–797.

[62] GRAEPEL T., OBERMAYER K., *A stochastic self-organizing map for proximity data*, Neural Computation, 1999, 11, 139–155.

[63] GRAY R.M., *Vector quantization*, IEEE ASSP Magazine, 1984, 1, 4–29.

[64] GRAY R.M., NEUHOFF D.L., *Quantization*, IEEE Trans. on Information Theory 1998, 44, 2325–2383.

[65] GRABOWSKI J., SKUBALSKA E., *Maximal flow problem in a network with variable structure*, Zastosowania Matematyki, 1982, XVII, 2, 293–307.

[66] GRABOWSKI J., SKUBALSKA E., SMUTNICKI C., *On flow shop scheduling with release and due dates to minimize maximum lateness*, J. Operational Res. Society, 1983, 34, 7, 615–620.

[67] GRABOWSKI J., SKUBALSKA E., *Optymalizacja struktury sieci transportowej przy kryterium minimalizacji kosztów przepływu*, Archiwum Automatyki i Telemechaniki, 1985, XXX, 1, 3–21.

[68] GREBLICKI W., *Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probabilities densities*, IEEE Trans. on Information Theory, 1981, 27, 364–366.

[69] GRIFFITHS J.G ., YANG C.G., *Algorithm for generating improved images of curved surfaces by distributing errors along Hilbert curve*, Computer-Aided Design, 1987, 19, 299.

[70] HASTIE T., TIBSHIRANI R., *Discriminant Adaptive Nearest Neighbor Classification*, Technical report, Stanford University, 1994.

[71] HESKES T., KAPPEN B., *Self-organization and parametric regression*, Proc. ICANN95, Paris, France, 1995, 81–86.

[72] HILBERT D., *Ueber die stetige Abbildung einer Linie auf ein Flaechenschtueck*, Mathematische Annalen, 1891, 38, 459–469.

[73] HUTCHINSON J.E., *Fractals and self-similarity*, Indiana Univ. Math. J., 1981, 30, 713–747.

[74] KAMATA S., NIIMI M., KAWAGUCHI E., *A gray image compression using a Hilbert scan*, w: Proceedings of the 13th ICPR, 2, 905–909, Vienna, Austria, August 25–29, 1996.

[75] KAMBHATLA N., LEEN T.K., *Dimension reduction by local principal component analysis*, Neural Computation, 1997, 9, 1493–1516.

[76] KANGAS J., KOHONEN T., LAAKSONEN J., *Variants of self-organizing maps*, IEEE Trans. on Neural Networks, 1990, 1, 93–99.

[77] KARAYIANNIS N.B., BEZDEK J.C., PAL N.R., HATHAWAY R.J., PIN-I PAI, *Repairs to GLVQ: a new family of competitive learning schemes*, IEEE Trans. on Neural Networks, 1996, 7, 1062–1071.

[78] KARAYIANNIS N.B., *A methodology for constructing fuzzy algorithms for learning vector quantization*, IEEE Trans. on Neural Networks, 1997, 8, 505–518.

[79] KIM B.S., PARK S.B., *A fast kNN finding algorithm based upon the ordered partition*, IEEE PAMI, 1986, 8(6), 761–766.

[80] KOHONEN T.,*Learning Vector Quantization for Pattern Recognition*, Technical Report TKK-F-A, 601, Finland: Helsinki University of Technology, 1986.

[81] KOHONEN T., *An introduction to neural computing*, Neural Networks, 1988, 1, 3–16.

[82] KOHONEN T., *Self-organizing maps*, Proc. of the IEEE, 1990, 78, 1464–1480.

[83] KOHONEN T., *Self-organizing Maps*, Springer Verlag, Berlin, Heidelberg, 1995.

[84] KORBICZ J., *Metody rozpoznawania obrazów w diagnostyce procesów przemysłowych*, materiały III Krajowej Konferencji Naukowo-Technicznej, 7–10 września 1998, Jurata, 93–100.

[85] KORBICZ J., OBUCHOWSKI A., UCIŃSKI D., *Sztuczne sieci neuronowe. Podstawy i zastosowania*, Akademicka Oficyna Wydawnicza PLJ, Warszawa 1994.

[86] KORONACKI J., THOMSON J.R., *Statystyczne sterowanie procesem. Metoda Deminga etapowej optymalizacji jakości*, Akademicka Oficyna Wydawnicza PLJ, Warszawa, 1994.

[87] KRZYŻAK A., *On exponential bounds on the Bayes risk of the kernel classification rule*, IEEE Trans. on Information Theory, 1991, 37, 490-499.

[88] KRZYŻAK A., RAFAJŁOWICZ E., *Aproksymacja funkcji przy pomocy jednokierunkowych sieci neuronowych*, w: Biocybernetyka i Inżynieria Biomedyczna 2000 pod red. M. Nałęcza, t.6, Sieci neuronowe, 371–388, Akademicka Oficyna Wydawnicza Exit, Warszawa, 2000.

[89] KRZYŻAK A., RAFAJŁOWICZ E., SKUBALSKA-RAFAJŁOWICZ E., *Clipped median and space-filling curves in image filtering*, Nonlinear Analysis, Theory, Methods and Applications, 2001, 47, 1, 303–314.

[90] KUDREWICZ J., *Fraktale i chaos*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1993.

[91] KULLBACK S., *Information Theory and Statistics*, wyd. II, Dover Publications Inc., Mineola, 1997.

[92] KULKARNI S.R., LUGOSI G., VENATESH S., *Learning pattern classification – a survey*, IEEE Trans. on Information Theory, 1998, 44, 2178-2206.

[93] KURATOWSKI K., *Wstęp do teorii mnogości i topologii*, PWN, Warszawa, 1972.

[94] KURZYŃSKI M., *Rozpoznawanie obiektów. Metody statystyczne*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 1997.

[95] LAMARQUE C.H ., ROBERT F., *Image analysis using space-filling curves and 1D wavelet basis*, Pattern Recognition, 1996, 29, 1309–1322.

[96] LAMPINEN J., OJA E., *Clustering properties of hierarchical self-organizing maps*, Journal of Mathematical Imaging and Vision, 1992, 2, 261–272.

[97] LEMPEL A., ZIV J., *Compression of two-dimensional data*, IEEE Trans. on Information Theory, 1986, 32, 2–8.

[98] LIKHOVIDOV V., *Variational approach to unsupervised learning algorithm of neural networks*, Neural Networks, 1997, 10, 273–289.

[99] LIN S., SI J., *Weight-value convergence of the SOM algorithm for discrete inputs*, Neural Computation, 1998, 10, 807–814.

[100] LINDE Y., BUZO A., GRAY R.M., *An algorithm for vector quantizer design*, IEEE Trans. on Communication, 1980, 28, 84-95.

[101] LINDGREN B.W, *Elementy teorii decyzji*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1977.

[102] LJUNG L., *Analysis of recursive stochastic algorithms*, IEEE Trans. on Automatic Control, 1977, 22, 551–575.

[103] LOWE D., TIPPING M., *Feed-forward neural networks and topographic mappings for exploratory data analysis*, Neural Computing and Applications, 1996, 4, 83–95.

[104] LOWRY C.A., WOODALL W.H. , *A multivariate exponentially weighted moving average control chart*, Technometrics, 1962, 34(1), 46–53.

[105] MACQUEEN J., *Some methods for classification and analysis of multivariate observation*, Proc. of the Fifth Berkeley Symposium on Math. Stat. and Prob., 1, 281–296, 1967.

[106] MANDELBROT B.B., *The Fractal Geometry of Nature*, W.H. Freeman, New York, 1983.

[107] MARTINETZ T.M., BERCOVICH S.G., SCHULTEN K.J., *Neural-gas network for vector quantization and its application to time-series prediction*, IEEE Trans. on Neural Networks, 1993, 4, 558–559.

[108] MARTYN T., *Fraktale i obiektowe algorytmy ich wizualizacji*, Nakom, Poznań, 1996.

[109] MCLEAN G.F., *Vector quantization for texture classification*, IEEE Trans. on Systems, Man and Cybernetics, 1993, 23(3), 637–649.

[110] MILNE S.C., *Peano curves and smoothness of functions*, Advances in Mathematics, 1980, 35, 129–157.

[111] MOORE E.H., *On certain crinkly curves*, Trans. Amer. Math. Soc., 1900, 1, 72–90.

[112] MORA G ., CHERRUAULT Y., *Characterization and generation of alpha-dense curves*, Computers and Mathematics with Applications, 1997, 33(9), 83–91.

[113] MORRISON D.F., *Wielowymiarowa analiza statystyczna*, PWN, Warszawa, 1990.

[114] MONTGOMERY D.C., *Introduction to Statistical Quality Control*, Wiley, New York, 1996.

[115] OEHLER K.L., GRAY R.M.,*Combining image compression and classification using vector quantization*, IEEE Trans. PAMI, 1995, 17(5), 461–473.

[116] OJA E.,*Neural networks, principal components, and subspaces*, Journal of Neural Systems, 1989, 1, 61–68.

[117] OSOWSKI S.,*Sieci neuronowe w ujęciu algorytmicznym*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1996.

[118] PAL N.R., BEZDEK J.C., HATHAWEY R.J., *Sequential competitive learning and the fuzzy c–means clustering algorithm*, Neural Networks, 1996, 9, 787–796.

[119] PATRICK E.D., ANDERSON D.R., BECHTEL F.K.,*Mapping multidimensional space to one dimension for computer output display*, IEEE Trans. on Comput., 1968, 17, 949–953.

[120] PATRICK E.A., *Fundamentals of Pattern Recognition*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1972.

[121] PEANO G., *Sur une courbe qui remplit toute une aire plane*, Math. Ann., 1890, 36, 157–160.

[122] PEITGEN H.-O., JÜRGENS H., SAUPE D., *Granice chaosu. Fraktale*, PWN, Warszawa, 1995.

[123] PHUVAN S., OH T.K., CARVIS N., LI Y., SZU H.H., *Texture analysis by space-filling curves and one-dimensional Haar wavelets*, Optical Engineering, 1992, 31, 1899–1906.

[124] PLATZMAN L.K ., BARTHOLDI J.J., *Spacefilling curves and the planar traveling salesman problem*, Journal of the ACM, 1989, 36(4), 719–737.

[125] HANAN J., PRUSINKIEWICZ P., *Lindenmayer systems, fractals, and plants*, Lecture Notes in Biomathematics, Springer, New York, 1989.

[126] QUWEIDER M.K., SALARI E., *Peano scanning partial distance search for vector quantization*, IEEE Signal Processing Letters, 1995, 2, 170–171.

[127] RAFAJŁOWICZ E., SKUBALSKA-RAFAJŁOWICZ E., *FFT in calculating nonparametric regression estimate based on trigonometric series*, Applied Mathematics and Computer Science, 1993, 3(4), 713–720.

[128] RAFAJŁOWICZ E., SKUBALSKA-RAFAJŁOWICZ E., *Nonparametric regression estimation by Bernstein-Durmayer polynomials*, Tatra Mountains Mathematical Publications, 1999, 17, 227–239.

[129] REGAZZONI C.S., TESCHIONI A., *A new approach to vector median filtering based on space filling curves*, IEEE Trans. on Image Processing, 1997, 6, 1025–1037.

[130] RIESZ F., SZÖKEFALVI-NAGY B., *Functional Analysis*, Dover Publications, Inc., New York, 1994.

[131] RIPLEY B., *Pattern Recognition and Neural Networks*, Cambridge Univ. Press, 1996.

[132] RITTER H., *Asymptotic level density for a class of vector quantization processes*, IEEE Trans. on Neural Networks, 1991, 2, 173–175.

[133] RITTER H., SCHULTEN K., *On the stationary state of Kohonen's self-organizing sensory mapping*, Biological Cybernetics, 1986, 54, 99–106.

[134] RITTER H., SCHULTEN K.L., *Convergence properties of Kohonen's topology conserving map: Fluctuations, stability, and dimension detection*, Biological Cybernetics, 1988, 60, 59–71.

[135] RITTER H., SCHULTEN K.L., *Kohonen's self-organizing maps: Exploring their computational capabilities*, w: Proc. of the 1988 IEEE Conference on Neural Networks, San Diego, 109–116, 1988.

[136] ROBBINS H., MONRO S., *A stochastic approximation method*, Ann. Math. Stat., 1951, 22, 400–407.

[137] SAGAN H., *Space-filling Curves*, Springer-Verlag, New York, 1994.

[138] SALEM R., ZYGMUND A., *Lacunary power series and Peano curves*, Duke Journal, 1945, 21, 383–390.

[139] SAMMON J.W., *A nonlinear mapping for data structure analysis*, IEEE Trans. on Computers, 1969, 18, 401–409.

[140] SAVARAGI Y., SUNAHARA Y., NAKAMIZO T., *Statistical Decision Theory in Adaptive Control Systems*, Mathematics in Science and Engineering 39, Academic Press, New York, 1967.

[141] SCHUSTER H.G., *Deterministic Chaos*, VGH Verlagsgesellschaft, Weinheim, 1988.

[142] SERFLING R.J., *Twierdzenia graniczne statystyki matematycznej*, PWN, Warszawa, 1991.

[143] SERGEYEV Y.D., *An information global optimization algorithm with local tunning*, SIAM Journal on Optimization, 1995, 5(4), 858–870.

[144] SIEDLECKI W., SIEDLECKA K., SKLANSKY J., *An overview of mapping techniques for exploratory pattern analysis*, Pattern Recognition, 1988, 21, 411-429.

[145] SIERPIŃSKI W., *O pewnej krzywej wypełniającej kwadrat. Sur une nouvelle courbe continue qui remplit toute une aire plane*, Bulletin de l'Acad. des Sciences de Cracovie, 1912, 463–478.

[146] SIERPIŃSKI W., *Remarque sur la courbe peannienne*, Wiadomości Matematyczne, 1936, 42, 1.

[147] SKARBEK W., *Metody reprezentacji obrazów cyfrowych*, Akademicka Oficyna Wydawnicza PLJ, Warszawa, 1993.

[148] SKUBALSKA E., *Przegląd zagadnień ustalania tras pojazdów ze szczególnym uwzględnieniem problemów harmonogramowania*, Zeszyty Naukowe AGH, 1981, 866, 146, 159–176.

[149] SKUBALSKA E., *Algorytm optymalnego ustalania tras pojazdów*, Zeszyty Naukowe AGH, 1982, 928, 32, 443–453.

[150] SKUBALSKA E., *Zastosowanie metody podziału i ograniczeń do optymalnego ustalania tras pojazdów*, Przegląd Statystyczny, 1984, XXXI, 1/2, 65–81.

[151] SKUBALSKA E., *Zagadnienia planowania tras pojazdów. Modele matematyczne*, Archiwum Automatyki i Telemechaniki, 1984, XXIX, 4, 483–499.

[152] SKUBALSKA-RAFAJŁOWICZ E., *Zadanie harmonogramowania produkcji w sytuacjach awaryjnych*, Zeszyty Naukowe PL. Sl. Ser. Automatyka, 1986, 895, 85, 233–241.

[153] SKUBALSKA-RAFAJŁOWICZ E., *Zagadnienie wrażliwości rozwiązań optymalnych zadań szeregowania*, Zeszyty Naukowe PL. Sl. Ser. Automatyka, 1988, 970, 94, 277–289.

[154] SKUBALSKA-RAFAJŁOWICZ E., *Zastosowanie matematyki interwałowej w rozwiązywaniu zadań szeregowania*, Zeszyty Naukowe PL. Sl. Ser. Automatyka, 1990, 1082, 100, 275–285.

[155] SKUBALSKA-RAFAJŁOWICZ E., *Problemy selektywnego wyboru i szeregowania zadań z przezbrojeniami*, Zeszyty Naukowe PL. Sl. Ser. Automatyka, 1992, 1175, 109, 224–232.

[156] SKUBALSKA-RAFAJŁOWICZ E., *Algorytm rozwiązywania zadania komiwojażera w przestrzeni wielowymiarowej z metryką euklidesową*, Zeszyty Naukowe PL. Sl. Ser. Automatyka, 1994, 1250, 114, 241–250.

[157] SKUBALSKA-RAFAJŁOWICZ E., *Szybki algorytm selekcji zadań*, Zeszyty Naukowe PL. Sl. Ser. Automatyka, 1994, 1250, 114, 251–258.

[158] SKUBALSKA-RAFAJŁOWICZ E., *The Closed Curve Filling Multidimensional Cube*, Technical Report, Inst. of Eng. Cybern., Technical University of Wrocław, (46/94), 1994.

[159] SKUBALSKA-RAFAJŁOWICZ E., *Visualization of Multidimensional Data Using Spacefilling Curves*, Technical Report, ICT Technical University of Wrocław, 40/95, 1995.

[160] SKUBALSKA-RAFAJŁOWICZ E. *Nearest Neighbor Classification with Metric on Space-filling Curve – Comparative Studies on Real Data*, Technical Report, Inst. of Eng. Cybern., Technical University of Wrocław, 26/29, 1995.

[161] SKUBALSKA-RAFAJŁOWICZ E., *Transformations of Multidimensional Data Using Spacefilling Curves – Implementations in Mathematica*, Technical Report, Inst. of Eng. Cybern., Technical University of Wrocław, 1996, 52/96.

[162] SKUBALSKA-RAFAJŁOWICZ E., *Szybki algorytm rozwiązywania dużych zadań komiwojażera*, Zeszyty Naukowe PL. Sl. Ser. Automatyka, 117, 253–263, 1996.

[163] SKUBALSKA-RAFAJŁOWICZ E., *Applications of the space-filling curves with data driven measure–preserving property*, Nonlinear Analysis, Theory, Methods and Applications, 30(3), 1305–1310, 1997.

[164] SKUBALSKA-RAFAJŁOWICZ E., *Space-filling curves and Kohonen's 1-D SOM as a method of a vector quantization with known asymptotic level density*, Proc. of the 3rd Conf. Neural Networks and Their Applications, Kule 97, 161–166, 1997.

[165] SKUBALSKA-RAFAJŁOWICZ E., *Krzywe wypełniające w przetwarzaniu wielowymiarowych danych a samoorganizujące sieci Kohonena*, V Krajowa Konferencja ,,Komputerowe wspomaganie badań naukowych" KOWBAN, Polanica Zdrój 1998, 429–434.

[166] SKUBALSKA-RAFAJŁOWICZ E., *On using space-filling curves for constructing multidimensional control charts. Detection of changes in time series*, zaproszony referat wygłoszony na International Workshop on New Developments in Quality Control, Frankfurt, 2-3 July 1999, Technical Report, ICT Technical University of Wrocław, SPR 17/99.

[167] SKUBALSKA-RAFAJŁOWICZ E., *Learning vector quantization with multidimensional data transformation based on space-filling curve*, Proc. of the 4th Conf. Neural Networks and Their Applications, Zakopane 1999, 226–231.

[168] SKUBALSKA-RAFAJŁOWICZ E., *On using space-filling curves and vector quantization for constructing multidimensional control charts*, Proc. of the 5th Conf. Neural Networks and Their Applications, Zakopane 2000, 162–167.

[169] SKUBALSKA-RAFAJŁOWICZ E., *Samoorganizujące sieci neuronowe*, w: Biocybernetyka i Inżynieria Biomedyczna 2000 pod red. M. Nałęcza, t.6, Sieci neuronowe, 179–226, Akademicka Oficyna Wydawnicza Exit, Warszawa, 2000.

[170] SKUBALSKA-RAFAJŁOWICZ E., *One-dimensional Kohonen LVQ nets for multidimensional pattern recognition*, Applied Mathematics and Computer Science, 10(4), 767–778, 2000.

[171] SKUBALSKA-RAFAJŁOWICZ E., *Pattern recognition algorithm based on space-filling curves and orthogonal expansion*, IEEE Trans. on Information Theory, 2001, 47, 5, 1915–1927.

[172] SKUBALSKA-RAFAJŁOWICZ E., *Data compression for pattern recognition based on space-filling curve pseudo-inverse mapping*, Nonlinear Analysis, Theory, Methods and Applications, 2001, 47/1, 315–326.

[173] SKUBALSKA-RAFAJŁOWICZ E., KRZYŻAK, A., *Data sorting along a space-filling curve for fast pattern recognition*, Proc. of the Second Int. Symp. on Methods and Models in Automation and Robotics, Międzyzdroje, Poland, 1, 339–344, 1995.

[174] SKUBALSKA-RAFAJŁOWICZ E., KRZYŻAK, A., *Fast k-NN classification rule using metric on space-filling curves*, Proc. of the 13th ICPR, Vienna, Austria, August 25–29, 1996, 1996, II, 221–225, IEEE Computer Society Press, Los Alamitos CA.

[175] SKUBALSKA-RAFAJŁOWICZ E., RAFAJŁOWICZ E., *Serching for optimal experimental designs using space-filling curves*, Applied Mathematics and Computer Science, 1998, 8, 3, 647–656.

[176] SONG H., LEE S.W., *LVQ combined with simulated annealing for optimal design of large-set reference models*, Neural Networks, 1996, 9(2), 329–336.

[177] SPECHT D., *Series estimation of probability density function*, Technometrics, 1971, 13, 409–424.

[178] STEELE J.M., *Efficacy of spacefilling heuristics in euclidean combinatorial optimization*, Operations Reserch Letters, 1989, 8, 237–239.

[179] STEINHAUS H., *La courbe de peano et les fonctions independantes*, w: Hugo Steinhaus Selected papers, 487–488, PWN, Warszawa, 1985.

[180] STEINHAUS H., *Sur la courbe peanienne de W. Sierpiński*, w: Hugo Steinhaus Selected papers, 489–492, PWN, Warszawa, 1985.

[181] STEINHAUS H., *La courbe de peano et les fonctions independantes*, Comptes Rendus Acad. Sci., Paris, 1936, 202, 1961–1966.

[182] STEVENS R.J., LEHAR A.F., PRESTON F.H., *Manipulation and presentation of multidimensional image data using the Peano scan*, IEEE Trans. PAMI, 1983, 5, 520–525.

[183] STONE C.J., *Optimal global rate of convergence for nonparametric regression*, Ann. Statist., 1982, 10, 1040 – 1053.

[184] STROGIN R.G., SERGEYEV Y.D., *Global multidimensional optimization on parallel computer*, Parallel Computing, 1992, 18(11), 1259–1273.

[185] TADEUSIEWICZ R., *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa 1993.

[186] TADEUSIEWICZ R., FLASIŃSKI M., *Rozpoznawanie obrazów*, PWN, Warszawa 1991.

[187] THIRAN P., HASLER M., *Self-organization of a one-dimensional Kohonen network with quantized weights and inputs*, Neural Networks, 1994, 7, 1427–1439.

[188] TOLAT V.V., *An analysis of Kohonen's self-organizing maps using a system of energy functions*, Biological Cybernetics, 1990, 64, 155-164.

[189] TRICOT C., *Curves and Fractal Dimension*, Springer-Verlag, Warszawa, New York, Berlin, 1993.

[190] ULTSCH A., SIEMION H., *Kohonen's self-organizing feature maps for exploratory data analysis*, Proc. INNC'90, Dordrecht, Netherlands, 305–308, 1990.

[191] VENABLES W.N., RIPLEY B.D., *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York, 1994.

[192] VESANTO J., *SOM-based data visualization methods*, Intelligence Data Analysis, 1999, 3(2), 111–126.

[193] VIDAL, R.E., *An algorithm for finding nearest neighbors in (approximately) constant time*, Pattern Recognition Letters, 1986, 4, 145–157.

[194] WANG X.Z., *Data Mining and Knowledge Discovery for Process Monitoring and Control. (Advances in industrial control)*, Springer-Verlag, London, Berlin, 1999.

[195] WĘGRZYN S., *Podstawy automatyki*, wyd. V, PWN, Warszawa, 1980.

[196] WHITEHEAD B.A., CHAOTE T.D., *Evolving space-filling curves to distribute radial basis functions over an input space*, IEEE Trans. on Neural Networks, 1994, 5, 15–23.

[197] WITTEN I.H., NEAL R.M., *Using Peano curves for bilevel display of continuous-tone images*, IEEE Computer Graphics and Applications, 1982, 2, 47–52.

[198] WOLVERTON C.T., WAGNER T.J., *Asymptotically optimal discriminant functions for pattern classification*, IEEE Trans. on Information Theory, 1969, 15, 258-265.

[199] WONG E., *Procesy stochastyczne w teorii informacji i układach dynamicznych*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1976.

[200] WUNDERLICH W., *Irregular curves and functional equations*, Proc. Benares Math. Society, 1954, 5, 215–230.

[201] WUNDERLICH W., *Ueber Peano–Kurven*, Elem. Math., 1973, 28, 1–10.

[202] YAIR E., ZEGER K., GERSHO A., *Competitive learning and soft competition for vector quantizer design* , IEEE Trans. on Signal Processing, 1992, 40, 294–309.

[203] YIN H., ALLISON N.M., *On the distribution and convergence of feature space in self-organizing maps*, Neural Computation, 1995, 7, 1178—1187.

[204] YI ZHENG, GREENLEAF J.F., *The effect of concave and convex weight adjustments on self-organizing maps*, IEEE Trans. on Neural Networks, 1996, 7, 87–96.

[205] YONG-ZAI LU, *Industrial Intelligent Control*, John Wiley and Sons, New York, 1996.

[206] ZADOR P.L., *Asymptotic quantization error of continuous signals and the quantization dimension*, IEEE Trans. on Information Theory, 1982, 28, 139–149.

[207] ZAKARAUSKAS P., OZARD J.M., *Complexity analysis for partitioning nearest-neighbor searching algorithms*, IEEE Trans. PAMI, 1996, 18(6), 663–668.

[208] ŻURADA J., BARSKI M., JĘDRUCH W., *Sztuczne sieci neuronowe*, PWN, Warszawa 1996.

Space-filling curves in decision problems

In the monograph, a methodology of constructing and validating decision algorithms is proposed. This methodology forms a new approach to the problems of multidimensional decision making. The main idea is to transform each multivariate observation to univariate one, using a quasi-inverse of a carefully chosen space-filling curve as a transformation. Then, the decision problem is solved for univariate data. The advantages of using such transformations lie in the dimensionality reduction and in data compression without loosing the information associated with the spatial structure of multivariate observations. These advantages allow us to construct fast decision algorithms, which can be easily implemented as on-line methods.

The methodology proposed is contained in the stream of research which explores novel strategies for control of complex industrial processes. These strategies are directed towards the design of flexible control systems, which combine model--based classical techniques with a variety of learning based methods. The learning is based on past observations, which are used for a control system adaptation and enhancement.

Important prerequisites for developing the decision making methodology are criteria for selecting space-filling curves and algorithms for calculating their quasi--inverses as quickly as possible. In the monograph, a method of forming functional equations for space-filling curves and their quasi-inverses is developed for the Peano, the Hilbert and the Sierpiński curves. It is also proved that these functional equations can be solved by backward recursions, providing exact values of the curves on dense sets in a finite number of iterations.

It is also shown that the space-filling based transformations retain essential statistical information, which is important in decision-making. In particular, it is proved that the Bayes risk is invariant under these transformations for every distribution which has a bounded support. Also a fractal dimension, scaled by the dimension of the data, occurs to be invariant. A new class of curves which

retain a prescribed probability measure is defined and used for solving vector quantization problems. In particular, an asymptotic distortion error of quantizers is investigated using the dimension error techniques. The notion of the space-filling curves is generalized also in another direction, namely, space-filling surfaces and their quasi-inverses are defined and investigated.

The above theoretical results form foundations for deriving new classes of decision-making algorithms, which are applicable in a variety of control system tasks. In particular, the notion of a multi-chart is introduced, together with a new method of detecting system faults. In addition to new fast algorithms of vector quantization and pattern recognition, also their asymptotics is investigated, providing a theoretical foundations for their use in monitoring and diagnosis of a system state. In some cases simple heuristic algorithms, having solid support in the simulations, are also discussed.

The important feature of all the decision algorithms, resulting from the methods considered in the monograph, is their low computational complexity.

Spis treści

Contents