Romuald Szymkiewicz

METODY NUMERYCZNE W INŻYNIERII WODNEJ

Wydawnictwo Politechniki Gdańskiej

Romuald Szymkiewicz

METODY NUMERYCZNE W INŻYNIERII WODNEJ

Gdańsk 2012

PRZEWODNICZĄCY KOMITETU REDAKCYJNEGO
WYDAWNICTWA POLITECHNIKI GDAŃSKIEJ
Janusz T. Cieśliński

RECENZENT
Wanda Kowalska

PROJEKT OKŁADKI
Katarzyna Olszonowicz

Wydanie II – 2007

Wydano za zgodą
Rektora Politechniki Gdańskiej

Publikacja dostępna tylko w wersji elektronicznej

ISBN 978-83-7348-457-3

WYDAWNICTWO POLITECHNIKI GDAŃSKIEJ

Wydanie II. Ark. wyd. 12,2, ark. druku 17,0, 1009/747

Spis treści

Przedmowa

Inżynieria wodna należy do tych dyscyplin, w których metody numeryczne są szczególnie szeroko stosowane jako ważne narzędzie analizy i rozwiązywania zagadnień inżynierskich. Praktycznie wszystkie istotne przypadki przepływu wód powierzchniowych i wód podziemnych, a także przepływy w rurociągach i ich sieciach oraz przepływy przez budowle hydrotechniczne (przelewy, zasuwy, upusty denne itd.) mogą być skutecznie rozwiązywane metodami numerycznymi. Są to zagadnienia, które spotykamy w takich dyscyplinach szczegółowych, jak hydromechanika, hydraulika, hydrologia czy hydrogeologia. Procesy przepływu wody należą do bardziej złożonych zjawisk fizycznych, nie zawsze w pełni rozpoznanych. Nawet wprowadzając daleko posunięte uproszczenia, zwykle otrzymujemy skomplikowane modele matematyczne. Zależnie od założeń co do zmienności parametrów charakteryzujących ruch wody w przestrzeni i czasie, mają one postać równań algebraicznych lub ich układów, równań różniczkowych zwyczajnych, bądź równań różniczkowych cząstkowych.

Nawet najprostsze równania algebraiczne opisujące ruch ustalony jednostajny w kanale otwartym lub w rurociągu, przepływ przez przelew itd., są równaniami nieliniowymi, które można rozwiązać tylko w sposób przybliżony. Identyczna sytuacja występuje w przypadku modeli w postaci równań różniczkowych lub ich układów. Zmienne w czasie i przestrzeni parametry modelowanych obiektów, ich nieregularny kształt, zmienne w czasie oddziaływanie zewnętrzne i nieliniowość równań uniemożliwiają znalezienie rozwiązań analitycznych nawet dla najprostszych warunków. Możliwe jest jednak znalezienie rozwiązań przybliżonych za pomocą metod numerycznych. Zarówno wymienione wyżej powody, jak i powszechna dostępność komputerów oraz gwałtownie rosnące ich możliwości obliczeniowe, a także potrzeby praktyczne spowodowały w ostatnich 35 latach duże zainteresowanie inżynierów-hydrotechników metodami numerycznego rozwiązywania zagadnień inżynierii wodnej.

Mając na uwadze rodzaje zagadnień występujących w inżynierii wodnej i typy równań opisujących je, w książce omówiono metody rozwiązywania algebraicznych równań nieliniowych i ich układów, równań różniczkowych zwyczajnych i ich układów oraz wszystkich typów równań różniczkowych o pochodnych cząstkowych, tj. hiperbolicznych, parabolicznych oraz eliptycznych. Ponadto omówiono takie zagadnienia, jak rozwiązywanie układów algebraicznych równań liniowych, powszechnie występujących w trakcie rozwiązywania wymienionych wcześniej równań, a także interpolację i aproksymację, które często stosowane są przy rozwiązywaniu wielu problemów w inżynierii wodnej.

Niniejsza książka przeznaczona jest jako pomoc dydaktyczna przede wszystkim dla studentów kierunku budownictwo (specjalność budownictwo wodne) oraz inżynieria środowiska, mających za sobą wykład z podstaw informatyki. Poszczególne zagadnienia omówiono w sposób przystępny, umożliwiający samodzielne rozwiązywanie typowych zadań inżynierii wodnej. Z tego powodu podstawy teoretyczne metod numerycznych ograniczono do niezbędnego minimum, podając jednocześnie źródła bibliograficzne, w których można znaleźć szczegółowe i wyczerpujące informacje na dany temat. Jednak

od czytelnika wymagana jest znajomość podstawowych działów matematyki w zakresie kursu matematyki dla wymienionych wcześniej kierunków studiów. W celu ułatwienia korzystania z książki w dodatku zamieszczonym na jej końcu przypomniano syntetyczne informacje o błędach w obliczeniach, reprezentacji maszynowej liczb itd. Zakłada się również, że czytelnik posiada znajomość fizycznych podstaw dotyczących przepływu wody w przewodach pod ciśnieniem, w kanałach otwartych i w ośrodkach porowatych, a także transportu domieszek rozpuszczonych w wodzie. Dla pełniejszej ilustracji omówionych metod i przedstawionych algorytmów, większość zagadnień zilustrowano przykładami rozwiązań.

Życzliwe zainteresowanie P.T. Czytelników dwoma wcześniejszymi wydaniami książki w wersji papierowej stało się zachętą do jej ponownego wydania. W stosunku do poprzednich, niniejsza wersja „Metod…" została w wielu miejscach poprawiona oraz uzupełniona o dwa dodatkowe rozdziały i kilka przykładów. Jednak w przeciwieństwie do wspomnianych wydań, tym razem książka jest udostępniana Czytelnikom w wersji elektronicznej w ramach Pomorskiej Biblioteki Cyfrowej.

Bardzo dziękuję Pani Katarzynie Olszonowicz za pomoc w przygotowaniu tekstu i rysunków.

Autor

1 Rozwiązywanie układów algebraicznych równań liniowych

1.1. Wprowadzenie

Typowym problemem pojawiającym się w trakcie rozwiązywania wielu zagadnień jest jedno z podstawowych zadań algebry liniowej, a mianowicie konieczność rozwiązywania układów algebraicznych równań liniowych. Ocenia się, że rozwiązywanie układów równań liniowych występuje w ok. 75% wszystkich zagadnień naukowych. Częstym źródłem takich układów jest interpolacja i aproksymacja za pomocą funkcji liniowych. Ponadto do tego typu układów prowadzą numeryczne metody rozwiązywania równań różniczkowych. Na przykład problem ten występuje w trakcie rozwiązywania równań różniczkowych o pochodnych cząstkowych. Zastosowanie do ich rozwiązania metody różnic skończonych lub metody elementów skończonych, omówionych w dalszych rozdziałach, wymaga ostatecznie rozwiązywania układów równań algebraicznych, i to zwykle o dużej liczbie niewiadomych. Nawet rozwiązywanie układów równań nieliniowych ostatecznie sprowadza się do cyklicznego rozwiązywania układów liniowych. Częsta konieczność rozwiązywania układów równań liniowych zaowocowała opracowaniem licznych metod, a także wprowadzeniem wielu ich realizacji do bibliotek maszyn cyfrowych.

Rozpatrzmy układ n równań liniowych z n niewiadomymi

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = 1, 2, \ldots, n. \tag{1.1}$$

Układ ten wygodnie jest zapisać w postaci macierzowej

$$\mathbf{A}\,\mathbf{X} = \mathbf{B}, \tag{1.2}$$

gdzie: $\mathbf{A} = (a_{ij})$ $(i = 1, 2, ..., n; j = 1, 2, ..., n)$ jest macierzą kwadratową n-tego stopnia, złożoną ze współczynników układu,

$\mathbf{X} = (x_i)$ $(i = 1, 2, ..., n)$ jest wektorem kolumnowym reprezentującym niewiadome,

$\mathbf{B} = (b_i)$ $(i = 1, 2, ..., n)$ jest wektorem prawych stron równań lub kolumną wyrazów wolnych

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \quad \mathbf{X} = \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix}, \quad \mathbf{B} = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{Bmatrix}.$$

Zakładamy tu, że macierz **A** i wektor **B** są rzeczywiste, a układ równań (1.1) ma rozwiązanie (Ralston, 1971).

Macierz układu w pewnych sytuacjach przyjmie szczególną postać. Jej struktura decyduje zwykle o nazwie. I tak:

— *macierz przekątniowa*, lub inaczej *diagonalna*, to taka, która ma niezerowe elementy jedynie na głównej przekątnej

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & 0 & d_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} = \operatorname{diag}(d_1, d_2, \cdots, d_n);$$

— *macierz jednostkowa* **I** stopnia *n* jest szczególnym przypadkiem macierzy diagonalnej

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \operatorname{diag}(1, 1, \cdots, 1),$$

tzn.

$$\mathbf{I} = (\delta_{ij}),$$

gdzie δ_{ij} jest deltą Kroneckera zdefiniowaną następująco:

$$\delta_{ij} = \begin{cases} 1 & \text{dla} \quad i = j, \\ 0 & \text{dla} \quad i \neq j, \end{cases} \qquad (i = 1, 2, \dots, n; \quad j = 1, 2, \dots, n);$$

— *macierzą trójkątną* nazywamy macierz mającą elementy niezerowe tylko po jednej stronie głównej przekątnej. Wyróżniamy zatem:

- macierz trójkątną dolną

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ l_{31} & l_{32} & l_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{bmatrix},$$

- macierz trójkątną górną

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ & u_{22} & u_{23} & \dots & u_{2n} \\ & & u_{33} & \dots & u_{3n} \\ & & & \ddots & \vdots \\ & & & & u_{nn} \end{bmatrix};$$

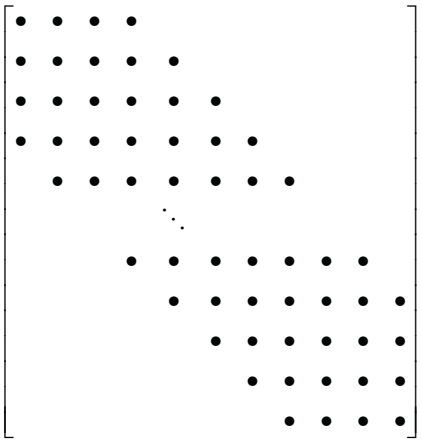
— *macierzą rzadką* nazywa się macierz mającą dużą liczbę elementów zerowych;

— *macierz pasmowa* (lub *wstęgowa*) to macierz mająca elementy niezerowe, zgrupowane w paśmie wokół głównej przekątnej (rys. 1.1).

Uwzględnienie cech macierzy współczynników układów równań prowadzi do poprawy efektywności algorytmu rozwiązania, ponieważ umożliwia redukcję czasu pracy komputera i wymaganej pamięci.

Znalezienie analitycznej postaci rozwiązania układu (1.1) nie sprawia kłopotu – otrzymuje się je ze wzorów Cramera. Istotnym problemem jest natomiast wykonanie obliczeń. Nawet dla układów niskiego stopnia wzory Cramera są nieprzydatne, gdyż wymagają wykonania wielkiej liczby działań. Z tego powodu opracowano bardziej skuteczne algorytmy rozwiązania układu (1.1).

Metody rozwiązania układów równań algebraicznych dzieli się na dwie grupy:
1) metody dokładne (lub bezpośrednie),
2) metody iteracyjne.



Rys. 1.1. Macierz pasmowa

Metody dokładne są to metody, które przy braku błędów zaokrągleń dają dokładne rozwiązanie po wykonaniu skończonej liczby operacji.

Metody iteracyjne dają zbieżny ciąg rozwiązań przybliżonych, którego granicą jest rozwiązanie dokładne. Umożliwiają one znalezienie jedynie rozwiązania przybliżonego.

Do oceny skuteczności algorytmów służą dwa podstawowe kryteria:
— liczba operacji, jaką trzeba wykonać,
— dokładność obliczonego rozwiązania.

Wybór metody zależy od własności układu równań. Jeśli macierz **A** jest pełna i niezbyt wielka, zwykle metody dokładne są skuteczniejsze. Na ogół metody dokładne stosuje się, gdy rozmiar układu nie przekracza kilkuset. Ze wzrostem liczby równań liczba operacji arytmetycznych gwałtownie rośnie, a błędy zaokrągleń stają się istotne. Jeśli macierz **A** jest rzadka i bardzo duża, skuteczniejsze są metody iteracyjne.

1.2. Dokładne metody rozwiązywania układów równań liniowych

1.2.1. Układy z macierzami trójkątnymi

Układ równań o macierzy trójkątnej rozwiązuje się szczególnie łatwo. Układ

$$\mathbf{U}\,\mathbf{X} = \mathbf{B} \tag{1.3}$$

z macierzą trójkątną górną **U** ma postać

$$\begin{aligned} u_{11}x_1 + \cdots + u_{1,n-1}x_{n-1} + u_{1,n}x_n &= b_1, \\ &\cdots \\ u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n &= b_{n-1}, \\ u_{n,n}x_n &= b_n. \end{aligned} \tag{1.4}$$

Przy założeniu, że $u_{ii} \neq 0$ (i = 1, 2, ..., n), niewiadome można obliczyć w kolejności: x_n, x_{n-1}, ..., x_1 ze wzorów:

$$x_n = \frac{b_n}{u_{n,n}},$$

$$x_{n-1} = \frac{b_{n-1} - u_{n-1,n} x_n}{u_{n-1,n-1}}, \tag{1.5}$$

...

$$x_1 = \frac{b_1 - u_{1,2} x_2 - u_{1,3} x_3 - \cdots - u_{1,n-1} x_{n-1} - u_{1,n} x_n}{u_{1,1}},$$

które można zapisać krócej

$$x_n = \frac{b_n}{u_{n,n}} \tag{1.6a}$$

$$x_i = \frac{b_i - \sum_{k=i+1}^{n} u_{i,k} x_k}{u_{i,i}} \quad (i = n-1, \ldots, 1). \tag{1.6b}$$

Ponieważ niewiadome wyznacza się w kolejności od ostatniej do pierwszej, ten algorytm nazywa się *podstawianiem wstecz* lub *postępowaniem odwrotnym*.

Układ równań liniowych

$$\mathbf{L}\,\mathbf{X} = \mathbf{B} \tag{1.7}$$

z macierzą trójkątną dolną **L** można rozwiązać podobnie. Przyjmując, że $l_{ii} \neq 0$ (i = 1, 2, ..., n), można wyznaczyć niewiadome metodą *podstawiania w przód*. Proces ten zapisuje się następująco:

$$x_1 = \frac{b_1}{l_{1,1}} \tag{1.8a}$$

$$x_i = \frac{b_i - \sum_{k=1}^{i-1} l_{i,k} x_k}{l_{i,i}} \quad (i = 2, \ldots, n). \tag{1.8b}$$

Z powyższych wzorów wynika, że rozwiązanie układu z macierzą trójkątną wymaga n dzieleń oraz

$$\sum_{i=1}^{n} (i-1) = \frac{1}{2} n(n-1) \approx \frac{1}{2} n^2$$

dodawań i mnożeń. Jest to liczba działań porównywalna z liczbą operacji arytmetycznych wykonanych przy mnożeniu wektora przez macierz trójkątną (Dahlquist i Bjorck, 1983).

1.2.2. Metoda eliminacji Gaussa

Jest to najważniejsza metoda dokładna rozwiązywania dowolnych układów liniowych. Jej idea polega na zastosowaniu takiego sposobu eliminacji niewiadomych, który doprowadzi do układu z macierzą trójkątną górną. Rozwiązanie zaś takiego układu, jak wynika z poprzedniego punktu, jest szczególnie proste.

Zapiszmy układ (1.1) w postaci:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n = b_3, \\ \qquad\qquad \ldots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n. \end{cases} \tag{1.9}$$

Ponieważ proces przekształcenia układu będzie polegał na kolejnych eliminacjach niewiadomych, wprowadźmy indeks eliminacji i nadajmy układowi wyjściowemu indeks jeden. Przyjmując $a_{ij}^{(1)} = a_{ij}$ oraz $b_i^{(1)} = b_i$, układ (1.9) zapiszemy:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1n}^{(1)}x_n = b_1^{(1)}, \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{31}^{(1)}x_1 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)}, \\ \qquad\qquad \ldots \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)}. \end{cases} \tag{1.10}$$

Załóżmy, że macierz układu jest nieosobliwa, zaś $a_{11}^{(1)} \neq 0$. Pomnóżmy kolejno pierwsze równanie przez $a_{i1}^{(1)} / a_{11}^{(1)}$ i odejmijmy od *i*-tego równania (i = 2, 3, ..., n). Otrzymamy wówczas pierwszy układ zredukowany

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)}, \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)}, \\ a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n &= b_3^{(2)}, \\ &\ldots \\ a_{n2}^{(2)}x_2 + a_{n3}^{(2)}x_3 + \ldots + a_{nn}^{(2)}x_n &= b_n^{(2)}, \end{aligned} \tag{1.11}$$

w którym współczynniki $a_{ij}^{(2)}$ są równe

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)}, \quad b_i^{(2)} = b_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} b_1^{(1)},$$

gdzie: i = 2, 3, ..., n; j = 2, 3, ..., n.

Wymiar układu zredukowanego wynosi $(n-1) \times (n-1)$, a jego współczynnikami są $a_{ij}^{(2)}$. Załóżmy teraz, że $a_{22}^{(2)} \neq 0$. Możemy zatem drugie równanie układu (1.11) pomnożyć przez

$a_{i2}^{(2)} \big/ a_{22}^{(2)}$ i odjąć od *i*-tego równania tego układu. Otrzymamy wówczas drugi układ zredukowany

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)}, \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)}, \\ a_{33}^{(3)}x_3 + \ldots + a_{3n}^{(3)}x_n &= b_3^{(3)}, \\ &\ldots \\ a_{n3}^{(3)}x_3 + \ldots + a_{nn}^{(3)}x_n &= b_n^{(3)}, \end{aligned} \tag{1.12}$$

w którym współczynniki $a_{ij}^{(3)}$ oraz $b_i^{(3)}$ są równe

$$a_{ij}^{(3)} = a_{ij}^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} a_{2j}^{(2)}, \quad b_i^{(3)} = b_i^{(2)} - \frac{a_{i1}^{(2)}}{a_{22}^{(2)}} b_2^{(2)},$$

gdzie: $i = 3, 4, ..., n$; $j = 3, 4, ..., n$.

Przedłużając ten sposób postępowania, po $n - 1$ krokach otrzymamy ostateczny układ

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)}, \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)}, \\ a_{33}^{(3)}x_3 + \ldots + a_{3n}^{(3)}x_n &= b_3^{(3)}, \\ &\ldots \\ a_{nn}^{(n)}x_n &= b_n^{(n)}. \end{aligned} \tag{1.13}$$

Jest to układ z macierzą trójkątną górną, który rozwiązuje się metodą podstawiania wstecz (1.6). Współczynniki przy niewiadomych wyznacza się przy użyciu wzorów:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}, \tag{1.14}$$

gdzie: $k = 1, 2, ..., n - 1$,
$j = k + 1, k + 2, ..., n$,
$i = k + 1, k + 2, ..., n$.

Elementy $a_{11}^{(1)}$, $a_{22}^{(2)}$, $a_{33}^{(3)}$, ..., $a_{nn}^{(n)}$ decydujące o procesie eliminacji nazywa się elementami głównymi. Układ (1.13) otrzymany został przy założeniu, że ich wartości są różne od zera. W wypadku, gdy w *k*-tym kroku eliminacji element główny $a_{kk}^{(k)}$ jest równy zeru, należy przestawić wiersze w taki sposób, aby na miejscu (k, k) znalazł się element niezerowy. Zatem w miejsce wiersza *k* wstawimy wiersz *l*, w którym $a_{lk}^{(k)} \neq 0$. Najlepiej jest wybrać wiersz, dla którego

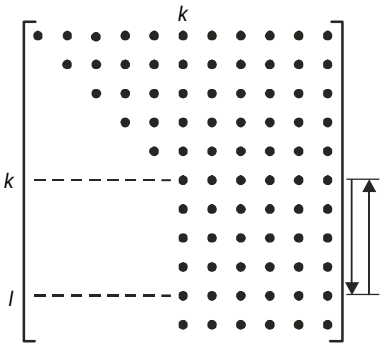
$$\left| a_{lk}^{(k)} \right| = \max_{k \le i \le n} \left| a_{ik}^{(k)} \right|, \tag{1.15}$$

gdzie *i* jest indeksem wiersza,
i zamienić go z wierszem *k*, przestawiając je wzajemnie $l \to k$, zaś $k \to l$ (rys. 1.2).

Przestawianie wierszy należy, jak wiadomo, do przekształceń elementarnych, których stosowanie daje macierz równoważną macierzy wyjściowej. Jeśli macierz układu (1.9) jest nieosobliwa, to przestawienia wierszy lub kolumn doprowadzą do macierzy, w której element $a_{kk}^{(k)}$ będzie różny od zera.

W dwóch wypadkach w metodzie eliminacji Gaussa elementy główne zawsze będą różne od zera i nie jest konieczne przestawianie wierszy. Są to takie układy, w których macierz **A** jest:

— macierzą z dominującą główną przekątną (diagonalą), tzn. gdy



Rys. 1.2. Macierz współczynników układu po *k* eliminacjach

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \qquad (i = 1, 2, \ldots, n), \tag{1.16}$$

— macierz jest symetryczna i dodatnio określona, tzn. gdy:

$$\mathbf{A} = \mathbf{A}^{\mathrm{T}} \text{ (symetryczna)} \tag{1.17}$$

i

$$\mathbf{X}^{T}\,\mathbf{A}\,\mathbf{X} > 0 \text{ dla każdego } \mathbf{X} \neq 0 \text{ (dodatnio określona).} \tag{1.18}$$

Dla danej macierzy **A** warunek (1.16) sprawdza się bardzo łatwo. Natomiast sprawdzenie warunku (1.18) jest bardziej złożone. Jednym z możliwych sposobów sprawdzenia, czy macierz symetryczna jest dodatnio określona, jest wykorzystanie kryterium Sylwestra (Dahlquist i Bjorck, 1983). Brzmi ono następująco: macierz **A** stopnia *n* jest dodatnio określona wtedy i tylko wtedy, gdy

$$\det(\mathbf{A}_k) > 0 \quad (k = 1, 2, ..., n),$$

gdzie $\mathbf{A}_k$ jest macierzą o wymiarze $k \times k$, utworzoną z *k* pierwszych wierszy i kolumn macierzy **A**. Kryterium to jest równoważne warunkowi $a_{kk}^{(k)} > 0$ ($k = 1, 2, ..., n$). Zatem gdy wykonujemy proces eliminacji Gaussa, nie przestawiając wierszy ani kolumn, i jeśli wszystkie elementy główne są dodatnie, to macierz **A** jest dodatnio określona.

1.2.3. Metoda rozkładu macierzy układu na macierze trójkątne

Załóżmy, że możemy rozłożyć macierz **A** układu (1.2) i przedstawić w postaci iloczynu macierzy trójkątnych – dolnej i górnej

$$\mathbf{A} = \mathbf{L}\,\mathbf{U}. \tag{1.19}$$

Jest to tzw. rozkład trójkątny lub rozkład **LU**. W takiej sytuacji układ wyjściowy (1.2) jest równoważny układowi

$$\mathbf{L}\,\mathbf{U}\,\mathbf{X} = \mathbf{B}, \tag{1.20}$$

który możemy z kolei rozłożyć na dwa układy z macierzami trójkątnymi:

$$\mathbf{L}\,\mathbf{Y} = \mathbf{B}, \tag{1.21}$$

$$\mathbf{U}\,\mathbf{X} = \mathbf{Y}. \tag{1.22}$$

W ten sposób uzyskujemy algorytm obliczeń bardziej ekonomiczny niż metoda eliminacji Gaussa. Liczba operacji tą metodą wyniesie, jak wiadomo, $2 \cdot n^2/2 = n^2$, podczas gdy metoda eliminacji Gaussa wymaga $n^3/3$ operacji (Dahlquist i Bjorck, 1983).

Istnieje twierdzenie o rozkładzie trójkątnym, które mówi, że jeśli det $(\mathbf{A}_k) \neq 0$ (k = 1, 2, ..., n) (gdzie $\mathbf{A}_k$ – macierz utworzona z k pierwszych wierszy i kolumn), to istnieje jedyny rozkład $\mathbf{A} = \mathbf{L}\,\mathbf{U}$ na czynniki takie, że macierz $\mathbf{L}$ jest macierzą trójkątną dolną i ma elementy $l_{ii} = 1$ (i = 1, 2, ..., n), a macierz $\mathbf{U}$ jest macierzą trójkątną górną.

W rzeczywistości, w każdej macierzy nieosobliwej $\mathbf{A}$ można tak przestawić wiersze, aby istniał rozkład trójkątny. Wynika to z równoważności eliminacji Gaussa i rozkładu trójkątnego. Mianowicie, wynik eliminacji można wykorzystać do przedstawienia macierzy układu (1.9) w postaci iloczynu $\mathbf{A} = \mathbf{L}\,\mathbf{U}$. Macierz górnotrójkątna $\mathbf{U}$ ma postać macierzy układu będącego wynikiem eliminacji (1.13). Z kolei macierz dolnotrójkątną $\mathbf{L}$ tworzą elementy $l_{ik} = a_{ik}^{(k)} \big/ a_{kk}^{(k)}$, występujące we wzorze (1.14), dla i = 2, 3, ..., n; k = 1, 2, ..., $i - 1$ oraz $l_{ii} = 1$ dla i = 1, 2, ..., n. Można więc zapisać:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2,n-1} & a_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & a_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ \vdots & \vdots & & & \\ l_{n1} & l_{n2} & \dots & l_{n,n-1} & 1 \end{bmatrix} \times \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1,n-1}^{(1)} & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \dots & a_{2,n-1}^{(2)} & a_{2n}^{(2)} \\ & & & & \vdots \\ & & & & a_{nn}^{(n)} \end{bmatrix}. \tag{1.23}$$

Jeśli macierz $\mathbf{A}$ jest symetryczna i dodatnio określona, to istnieje jedyna macierz górnotrójkątna $\mathbf{R}$ o dodatnich elementach na głównej przekątnej taka, że

$$\mathbf{A} = \mathbf{R}^T\,\mathbf{R}. \tag{1.24}$$

Ten wariant twierdzenia o rozkładzie trójkątnym można powiązać również z metodą eliminacji Gaussa w następujący sposób:

$$\mathbf{A} = \mathbf{R}^T\mathbf{R} = \mathbf{L}\,\mathbf{U} = \mathbf{U}^T\,\mathbf{D}^{-1}\,\mathbf{U}, \tag{1.25}$$

gdzie: $\mathbf{U}$ – wynikowa macierz eliminacji Gaussa,
$\mathbf{D} = \text{diag}\,(u_{11}, u_{22}, ..., u_{nn})$ – macierz przekątniowa utworzona z elementów głównych.

Rozwiązanie układu (1.2) $\mathbf{A}\,\mathbf{X} = \mathbf{B}$ sprowadza się w tym wypadku do rozwiązania dwóch układów z macierzami trójkątnymi:

$$\mathbf{U}^T\,\mathbf{Y} = \mathbf{B}, \tag{1.26}$$

$$\mathbf{U}\,\mathbf{X} = \mathbf{D}\,\mathbf{Y}. \tag{1.27}$$

Nie trzeba tutaj pamiętać macierzy dolnotrójkątnej $\mathbf{L}$ utworzonej przez mnożniki $l_{ik} = a_{ik}^{(k)} \big/ a_{kk}^{(k)}$.

Rozwiązując wiele różnych zagadnień (na przykład równania różniczkowe cząstkowe w przypadku jednowymiarowym), otrzymuje się układy równań (1.2), w których macierz **A** jest trójdiagonalna:

$$\mathbf{A}=\begin{bmatrix} \beta_1 & \gamma_1 & & & & \\ \alpha_2 & \beta_2 & \gamma_2 & & & \\ & \alpha_3 & \beta_3 & \gamma_3 & & \\ & & & \ddots & & \\ & & & \alpha_{n-1} & \beta_{n-1} & \gamma_{n-1} \\ & & & & \alpha_n & \beta_n \end{bmatrix}. \tag{1.28}$$

Taki układ równań można rozwiązać nadzwyczaj szybko, wykonując małą liczbę działań arytmetycznych. W algorytmie rozwiązania układu wykorzystuje się opisany wcześniej rozkład macierzy **A** o postaci (1.19), w którym:

$$\mathbf{L}=\begin{bmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & l_3 & 1 & & \\ & & \ddots & & \\ & & l_{n-1} & 1 & \\ & & & l_n & 1 \end{bmatrix}, \quad \mathbf{U}=\begin{bmatrix} u_1 & \gamma_1 & & & & \\ & u_2 & \gamma_2 & & & \\ & & u_3 & \gamma_3 & & \\ & & & \ddots & \ddots & \\ & & & & u_{n-1} & \gamma_{n-1} \\ & & & & & u_n \end{bmatrix}. \tag{1.29a,b}$$

Elementy $l_2, l_3, ..., l_n$ oraz $u_1, u_2, u_3, ..., u_n$ oblicza się rekurencyjnie ze wzorów:

$$\left.\begin{aligned} u_1 &= \beta_1 \\ l_i &= \frac{\alpha_i}{u_{i-1}} \\ u_i &= \beta_i - l_i \gamma_{i-1} \end{aligned}\right\} \quad \text{dla} \quad i = 2, 3, ..., n. \tag{1.30}$$

Dla zaoszczędzenia pamięci komputera, elementy l_i oraz u_i można zapamiętać w miejscach zajętych poprzednio przez elementy α_i i β_i macierzy **A**.

Aby znaleźć rozwiązanie **X** układu (1.20), należy kolejno rozwiązać dwa układy równań (1.21) i (1.22). Proces ich rozwiązywania przebiega następująco:

$$\begin{aligned} &y_1 = b_1, \\ &y_i = b_i - l_i y_{i-1} \quad \text{dla} \quad i = 2, 3, ..., n \end{aligned} \tag{1.31}$$

oraz

$$\begin{aligned} &x_n = y_n/u_n, \\ &x_i = (y_i - \gamma_i x_{i+1})/u_i \quad \text{dla} \quad i = n-1, n-2, ..., 1. \end{aligned} \tag{1.32}$$

Powyższy algorytm warto stosować, gdy układ rozwiązuje się co najmniej dla dwóch różnych wektorów prawych stron **B**. Jeśli układ rozwiązuje się tylko raz, to nie ma potrzeby pamiętać elementy l_i. Tok obliczeń będzie więc jeszcze prostszy:

$$u_1 = \beta_1; \quad y_1 = b_1,$$

$$\left.\begin{aligned} l &= \alpha_i / u_{i-1}, \\ u_i &= \beta_i - l\ \gamma_{i-1}, \\ y_i &= b_i - l\ y_{i-1}, \end{aligned}\right\} \quad \text{dla} \quad i = 2, 3, \cdots, n \tag{1.33}$$

oraz

$$\begin{aligned} x_n &= y_n / u_n, \\ x_i &= (y_i - \gamma_i x_{i+1}) / u_i \quad \text{dla} \quad i = n-1, n-2, ..., 1. \end{aligned} \tag{1.34}$$

Powyższy algorytm rozwiązania układu równań z macierzą trójdiagonalną nazywany jest algorytmem Thomasa lub „double sweep" (Fletcher, 1991). Łączna liczba działań arytmetycznych wynosi tutaj tylko $3(n-1)$ dodawań i mnożeń oraz $2n-1$ dzieleń. Co do liczby działań i obciążenia pamięci jest to algorytm optymalny (Dryja i Jankowscy, 1981).

Układy równań z macierzami pasmowymi są szczególnie efektywnie rozwiązywane metodą eliminacji Gaussa, gdyż nie zmienia ona struktury macierzy. Otrzymane macierze **L** oraz **U** są dalej macierzami pasmowymi. Jeśli w trakcie eliminacji nie trzeba dokonywać częściowego wyboru elementów głównych, to szerokość pasma macierzy **L** równa jest powiększonej o 1 liczbie niezerowych przekątnych macierzy **A** leżących pod diagonalą, zaś szerokość pasma macierzy **U** równa jest powiększonej o 1 liczbie niezerowych przekątnych leżących nad diagonalą macierzy **A**. Jeśli zaś dokonuje się częściowego wyboru elementów głównych, to szerokość pasma macierzy **L** jest taka jak poprzednio, zaś pasma macierzy **U** jest równa szerokości pasma macierzy **A** (Dahlquist i Bjorck, 1983).

Jeśli wewnątrz pasma macierzy **A** występują elementy zerowe, to w trakcie eliminacji pasmo zapełnia się, gdyż stają się one niezerowe. Jest to istotna niedogodność.

1.2.4. Metoda odwracania macierzy

Jeśli macierz układu (1.2) $\mathbf{A}\ \mathbf{X} = \mathbf{B}$ można odwrócić, to rozwiązanie układu można otrzymać wprost jako:

$$\mathbf{X} = \mathbf{A}^{-1}\ \mathbf{B}, \tag{1.35}$$

ponieważ

$$\mathbf{A}^{-1}\ \mathbf{A}\ \mathbf{X} = \mathbf{A}^{-1}\ \mathbf{B}, \tag{1.36}$$

zaś

$$\mathbf{A}^{-1}\ \mathbf{A} = \mathbf{I}. \tag{1.37}$$

Ta metoda jest kosztowna pod każdym względem i raczej nie stosuje się jej w praktyce. Szczególnie należy unikać jej w wypadku układów z macierzami pasmowymi. Wiadomo bowiem, że macierz odwrotna do macierzy pasmowej jest zwykle macierzą pełną. Zatem w wypadku układu z macierzą trójdiagonalną o wymiarze $n \times n$ zamiast pamiętać $3n$ elementów niezerowych, należałby pamiętać n^2 elementów macierzy $\mathbf{A}^{-1}$.

1.2.5. Metoda gradientów sprzężonych

Metodę tę można stosować do rozwiązania układu równań (1.2) $\mathbf{A}\ \mathbf{X} = \mathbf{B}$, gdy macierz **A** jest dodatnio określona. Tok obliczeń jest następujący (Dahlquist i Bjorck, 1983):

— przyjmuje się początkową wartość wektora niewiadomych $\mathbf{X}_0$;

— oblicza się residuum dla przyjętego $\mathbf{X}_0$;

$$\mathbf{r}_0 = \mathbf{B} - \mathbf{A}\ \mathbf{X}_0 \tag{1.38}$$

podstawiając

$$\mathbf{P}_0 = \mathbf{r}_0, \tag{1.39}$$

— dla $i = 0, 1, 2, ..., n - 1$ oblicza się kolejno

$$\alpha_i = \frac{\|\mathbf{r}_i\|^2}{\mathbf{P}_i^T \mathbf{A}\mathbf{P}_i}, \tag{1.40}$$

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \alpha_i \mathbf{P}_i, \tag{1.41}$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i \mathbf{A}\mathbf{P}_i, \tag{1.42}$$

$$\beta_i = \frac{\|\mathbf{r}_{i+1}\|^2}{\|\mathbf{r}_i\|^2}, \tag{1.43}$$

$$\mathbf{P}_{i+1} = \mathbf{r}_{i+1} + \beta_i \mathbf{P}_i. \tag{1.44}$$

W powyższych zależnościach $\|\mathbf{r}\|$ oznacza tzw. normę euklidesową wektora $\mathbf{r}$, zdefiniowaną następująco:

$$\|\mathbf{r}\| = \left(r_1^2 + r_2^2 + r_3^2 + \ldots + r_n^2\right)^{1/2}. \tag{1.45}$$

Jak wynika z przedstawionego algorytmu, w trakcie jego realizacji nie następuje przekształcanie macierzy współczynników $\mathbf{A}$ rozwiązywanego układu. Jest to bardzo atrakcyjna własność metody, szczególnie gdy rozwiązuje się wielki układ z macierzą rzadką. Wystarczy bowiem pamiętać wyłącznie niezerowe elementy macierzy $\mathbf{A}$ i tzw. macierz adresów zawierającą informacje o ich położeniu w $\mathbf{A}$. Umożliwia to wykonywanie operacji tylko na niezerowych elementach.

Jeśli nie ma błędów zaokrągleń, to ostateczny wektor $\mathbf{X}_n$ jest rozwiązaniem układu (1.2). W wielu wypadkach obliczeń nie trzeba prowadzić do końca, gdyż zadowalającym przybliżeniem rozwiązania jest wektor $\mathbf{X}_i$ dla pewnego $i < n$.

1.3. Metody iteracyjne rozwiązywania układów równań liniowych

Dokładne metody rozwiązywania układu równań liniowych są efektywne wtedy, gdy układ nie jest zbyt wielki, a macierz współczynników zawiera dużą liczbę elementów niezerowych. Tymczasem rozwiązując różne zagadnienia, szczególnie równania różniczkowe cząstkowe, natrafiamy na układy równań algebraicznych o macierzach rzadkich, mających wielkie wymiary. Rozwiązanie układu równań z macierzą współczynników, w której występuje duża liczba elementów zerowych, metodą eliminacji Gaussa jest niecelowe. Bardziej efektywne są *metody iteracyjne*, które – operując macierzą rzadką – nie zmieniają jej struktury. Iteracja (w języku łacińskim: powtarzanie) oznacza powtarzanie pewnej czynności lub procesu, w tym wypadku numerycznego, czyli kolejne przybliżenie. Metody iteracyjne startują z początkowego przybliżenia rozwiązania, które następnie poprawia się, aż do otrzymania rozwiązania dostatecznie dokładnego.

Rozpatrzmy układ równań liniowych (1.2) **A X** = **B**, który po rozpisaniu ma postać:

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1,\\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2,\\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n = b_3,\\
&\quad\vdots \quad\ \vdots \qquad\ \vdots \quad\ \vdots \quad\ \vdots \quad\ \vdots\\
&a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n.
\end{aligned} \tag{1.46}$$

Załóżmy, że układ został tak uporządkowany, że $a_{ii} \neq 0$ (i = 1, 2, ..., n). Przekształćmy układ (1.46), wyznaczając z kolejnych jego równań kolejne niewiadome:

$$\begin{aligned}
x_1 &= (-(a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n) + b_1)/a_{11},\\
x_2 &= (-(a_{21}x_1 + a_{23}x_3 + \ldots + a_{2n}x_n) + b_2)/a_{22},\\
x_3 &= (-(a_{31}x_2 + a_{32}x_3 + \ldots + a_{3n}x_n) + b_3)/a_{33},\\
&\quad\vdots \quad\ \vdots \qquad\ \vdots \quad\ \vdots \quad\ \vdots \quad\ \vdots\\
x_n &= (-(a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,n-1}x_{n-1}) + b_n)/a_{nn}.
\end{aligned} \tag{1.47}$$

Krócej można napisać go w postaci

$$x_i = \frac{-\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j + b_i}{a_{ii}} \quad \text{dla} \quad i = 1, 2, \ldots, n. \tag{1.48}$$

Najprostsza metoda iteracyjna polega na utworzeniu ciągu przybliżeń według wzoru

$$x_i^{(k+1)} = \frac{-\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i}{a_{ii}} \quad \text{dla} \quad i = 1, 2, \ldots, n. \tag{1.49}$$

gdzie k jest indeksem iteracji.
Metoda ta nazywa się *metodą Jacobiego* lub metodą *iteracji prostej*.
Jeśli

$$\lim_{k\to\infty} \mathbf{X}^{(k)} = \mathbf{X}, \tag{1.50}$$

to **X** jest rozwiązaniem układu wyjściowego **A X** = **B**.

W metodzie Jacobiego nowe wartości elementów wektora **X** w iteracji $k + 1$ oblicza się na podstawie ich wartości w iteracji k. Łatwo można zauważyć, że obliczając nowe przybliżenie elementu i-tego, czyli $x_i^{(k+1)}$, znamy już nowe przybliżenie elementów od 1 do $i - 1$ włącznie. Fakt ten można uwzględnić w formule iteracyjnej (1.49), otrzymując:

$$x_i^{(k+1)} = \frac{-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i}{a_{ii}} \quad \text{dla} \quad i = 1, 2, \ldots, n. \tag{1.51}$$

Powyższa metoda nosi nazwę *metody Gaussa-Seidela*. Jak wykazują doświadczenia, w większości wypadków metoda ta jest około dwukrotnie szybciej zbieżna niż metoda Jacobiego.

Istnieje możliwość znacznego przyspieszenia zbieżności metody Gaussa-Seidela. W tym celu należy dokonać jej modyfikacji. Wzór iteracyjny (1.51) można napisać w innej postaci

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + \mathbf{r}^{(k)}, \tag{1.52}$$

gdzie:
$$\mathbf{r}^{(k)} = \mathbf{B} - \mathbf{A}\,\mathbf{X}^{(k)} \tag{1.53}$$

jest wektorem residualnym w *k*-tej iteracji. Residuum *i*-tego równania określa wzór

$$r_i^{(k)} = \frac{-\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} + b_i}{a_{ii}} \quad \text{dla} \quad i = 1, 2, \ldots, n\,. \tag{1.54}$$

Metodę iteracyjną o postaci

$$x_i^{(k+1)} = x_i^{(k)} + \omega r_i^{(k)} \quad \text{dla} \quad i = 1, 2, ..., n \tag{1.55}$$

nazywa się *metodą nadrelaksacji* lub w skrócie *metodą SOR* (skrót od angielskich słów: *Successive Overrelaxation*). Parametr ω jest tzw. współczynnikiem relaksacji. Jego wartość należy tak dobrać, aby zbieżność procesu iteracyjnego była jak najszybsza. Parametr ten przyjmuje wartości z przedziału (0, 2). Dla $\omega = 1$ metoda nadrelaksacji staje się metodą Gaussa-Seidela. Natomiast dla $\omega = 2$ proces iteracyjny staje się niezbieżny.

Podstawowym problemem związanym ze stosowaniem iteracyjnych metod rozwiązywania liniowych układów równań jest zbieżność procesu iteracyjnego. Istnieje udowodnione twierdzenie (Stoer i Burlisch, 1980), które mówi, że proces iteracyjny jest zbieżny, jeśli macierz układu równań **A** jest macierzą z dominującą diagonalą, tzn. gdy

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad \text{dla} \quad i = 1, 2, \ldots, n\,. \tag{1.56}$$

Twierdzenie to dotyczy wszystkich trzech metod iteracyjnych, przy czym w metodzie nadrelaksacji o zbieżności dodatkowo decyduje parametr ω. W wypadku dowolnego układu równań algebraicznych jest trudno oszacować taką wartość ω, przy której zapewniona jest najszybsza zbieżność procesu iteracyjnego. W szczególnych wypadkach, gdy układ równań jest wynikiem rozwiązania równania różniczkowego cząstkowego (np. Laplace'a, Poissona), istnieją propozycje oszacowania optymalnej wartości ω. Do problemu tego powrócimy w podrozdziale 8.5, przy okazji numerycznego rozwiązywania tego typu równań.

Z rozwiązywaniem układów równań metodami iteracyjnymi wiążą się dodatkowe następujące problemy:

1) określenie początkowego przybliżenia poszukiwanego rozwiązania, czyli $\mathbf{X}^{(k=0)}$,
2) przerwanie procesu iteracyjnego.

Jeśli jest spełniony warunek (1.56), to wszystkie omówione metody są zbieżne dla dowolnego przybliżenia początkowego $\mathbf{X}^{(k=0)}$. Chociaż początkowe przybliżenie może być dowolne, to najczęściej wybiera się wektor zerowy, czyli

$$\mathbf{X}^{(k=0)} = \mathbf{0}. \tag{1.57}$$

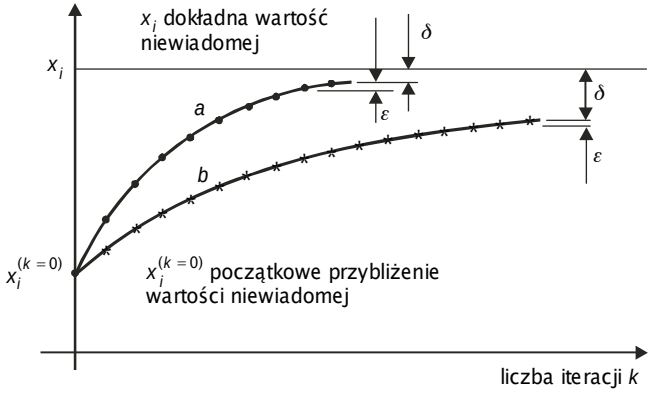
W niektórych wypadkach, na przykład rozwiązując równanie Laplace'a, mamy ogólną informację o rozwiązaniu. Jej wykorzystanie pozwala dokładniej oszacować początkowe przybliżenie i w konsekwencji skrócić czas obliczeń.

W celu przerwania obliczeń wykorzystuje się tzw. test końca procesu iteracyjnego. Test ten polega na oszacowaniu błędu rozwiązania w iteracji $k + 1$. W szczególnych wypadkach o błędzie można wnioskować, porównując dwa kolejne przybliżenia rozwiązania. Obliczenia kończy się, gdy

$$\left| x_i^{(k+1)} - x_i^{(k)} \right| \le \varepsilon \quad \text{dla } i = 1, 2, ..., n, \tag{1.58}$$

gdzie: k – indeks iteracji,
ε – dodatnia liczba określająca dokładność.

Powyższy test daje dobre wyniki, gdy proces iteracyjny jest względnie szybkozbieżny. Jeśli jest on wolnozbieżny, to pomimo niewielkiej różnicy rozwiązań w dwóch kolejnych iteracjach błąd rozwiązania δ powstały wskutek przerwania procesu iteracyjnego może być duży. Sytuację taką ilustruje rys. 1.3.



Rys. 1.3. Wpływ testu (1.58) na błąd rozwiązania przy: a) szybkozbieżnym procesie iteracyjnym, b) wolnozbieżnym procesie iteracyjnym

Inny test zakończenia procesu iteracyjnego polega na badaniu normy euklidesowej wektora residualnego w iteracji k-tej:

$$\left\| \mathbf{r}^{(k)} \right\| \le \varepsilon', \tag{1.59}$$

gdzie $\varepsilon' > 0$ określa dokładność rozwiązania.
Wektor $\mathbf{r}$ definiuje zależność (1.53), zaś jego normę – zależność (1.45).

1.4. Uwagi o układach równań liniowych

Opisane w poprzednich podrozdziałach metody rozwiązywania układów równań algebraicznych w większości sytuacji są skuteczne. Jednak zdarza się, że uzyskane rozwiązanie budzi wątpliwości, gdyż zdecydowanie odbiega od oczekiwanego. Jest to wynikiem tzw. złego uwarunkowania, na które niekiedy można natrafić przy rozwiązywaniu układów równań liniowych. Rozważmy układ

$$\begin{aligned} 0{,}1x + y &= 0{,}9 \\ 0{,}1x + 1{,}0001y &= 0{,}9001 \end{aligned} \tag{1.60}$$

którego rozwiązaniem jest para liczb $x = -1$, $y = 1$. Weźmy teraz pod uwagę układ

$$\begin{aligned} 0{,}1x + y &= 0{,}9 \\ 0{,}1x + 0{,}9999y &= 0{,}9002 \end{aligned} \tag{1.61}$$

mający rozwiązanie: $x = +29$, $y = -2$. Widać, że zmiana współczynnika a_{22} o $-0{,}0002$ i zmiana b_2 o $0{,}0001$ spowodowała drastyczną zmianę rozwiązania.

Załóżmy, że macierz współczynników układu (1.60) jest nieosobliwa. Rozwiązanie tego układu można zatem zapisać w postaci:

$$\mathbf{X} = \mathbf{A}^{-1}\,\mathbf{B}, \tag{1.62}$$

gdzie $\mathbf{A}^{-1}$ oznacza macierz odwrotną. Załóżmy ponadto, że elementy macierzy $\mathbf{A}$ zostały w ten sposób unormowane, że największy co do modułu jest rzędu 1. Załóżmy również, że niektóre elementy macierzy $\mathbf{A}^{-1}$ są bardzo duże i że jednym z nich jest

$$a_{ji}^{-1} = \frac{A_{ij}}{D}, \tag{1.63}$$

gdzie A_{ij} jest dopełnieniem algebraicznym elementu (a_{ij}), a więc nie zmienia się przy zmianie elementu a_{ij}, zaś D oznacza wyznacznik macierzy $\mathbf{A}$.

Założenie, że element a_{ji}^{-1} jest duży, oznacza, że A_{ij} musi być duże w stosunku do D. Ponieważ jeden ze składników rozwinięcia wyznacznika A względem elementów i-tego wiersza lub j-tej kolumny ma postać $a_{ij}\,A_{ij}$, więc mały błąd elementu a_{ij} w stosunku do 1, tzn. do normalizacji macierzy $\mathbf{A}$, może spowodować duży błąd względny wyznacznika D, a tym samym duży błąd względny elementu a_{ij} i ostatecznie duży błąd względny rozwiązania $\mathbf{X}$. Podobnie, mała zmiana składowej wektora $\mathbf{B}$ może znacznie zmienić $\mathbf{X}$. Można więc stwierdzić, że macierz $\mathbf{A}$, której elementami są liczby rzędu 1, i układ równań są źle uwarunkowane, jeżeli $\mathbf{A}^{-1}$ zawiera bardzo duże elementy. W wypadku, gdy największa wartość bezwzględna elementu macierzy $\mathbf{A}^{-1}$ jest rzędu 1, to macierz $\mathbf{A}$ można nazwać dobrze uwarunkowaną. Wracając do układu (1.61), należy zauważyć, że elementami macierzy współczynników dla tego układu są liczby rzędu 1. Natomiast macierz odwrotna ma postać następującą:

$$\mathbf{A}^{-1} = \begin{bmatrix} \dfrac{1{,}0001}{0{,}0001} & \dfrac{-1}{0{,}0001} \\ \dfrac{-1}{0{,}0001} & \dfrac{1}{0{,}0001} \end{bmatrix} = \begin{bmatrix} 10001 & -10000 \\ -10000 & 10000 \end{bmatrix}.$$

Widać więc, że elementy macierzy odwrotnej są rzędu 10^4, co wskazuje na złe uwarunkowanie macierzy $\mathbf{A}$.

Współczynniki układu (1.61) mogą być np. wynikiem pomiarów. Jeżeli nie można otrzymać tych wartości z odpowiednią dokładnością, to rozwiązanie może być obarczone dużym błędem, bez względu na to, jak dokładnie będą przeprowadzone obliczenia. Problem rozwiązania układu źle uwarunkowanego z dokładnością usprawiedliwioną przez dokładność danych jest jednym z najtrudniejszych problemów związanych z układami równań liniowych. Ten sam efekt mogą też wywołać błędy zaokrągleń powstające w czasie obliczeń.

W trakcie rozwiązywania układów równań liniowych występuje również problem błędu rozwiązania. Istnieją trzy źródła błędu. Pierwsze jest związane z błędami współczyn-

ników macierzy **A** i elementów wektora **B**. Gdy są to wielkości empiryczne, błędu tego nie można wyeliminować. Można jedynie, na podstawie oszacowania błędu pomiaru, ocenić dokładność rozwiązania. Drugim źródłem błędu są zaokrąglenia wykonywane w czasie obliczeń. Trzecim źródłem jest błąd metody. W metodach dokładnych (wzory Cramera, eliminacja Gaussa), które przy założeniu, że wszystkie działania wykonujemy dokładnie, prowadzą do dokładnego rozwiązania, błędu metody nie ma. Natomiast jeśli stosować metody iteracyjne, rozwiązanie układu ma postać zbieżnego nieskończonego procesu iteracyjnego. Proces ten przerywa się w momencie, gdy zostanie osiągnięta wymagana dokładność. Zatem z założenia otrzymujemy rozwiązanie obarczone błędem.

2 Rozwiązywanie przybliżone równań nieliniowych

2.1. Wprowadzenie

Rozwiązując wiele praktycznych zagadnień hydrauliki, bardzo często spotyka się algebraiczne równania nieliniowe. Na przykład, rozwiązanie tego typu równania jest konieczne przy wyznaczeniu głębokości normalnej w kanale otwartym. Głębokość normalna, przy której panuje natężenie przepływu Q, jest to głębokość w warunkach przepływu ustalonego jednostajnego w kanale o danym spadku dna i współczynniku szorstkości. Nawet dla najprostszego, bo prostokątnego, przekroju poprzecznego kanału, otrzymujemy algebraiczne równanie nieliniowe względem głębokości.

Nieliniową postać ma również formuła Colebrooka-White'a, opisująca współczynnik oporów liniowych λ w przewodzie, w którym płynie woda pod ciśnieniem. Do nieliniowego równania algebraicznego dochodzi się również, obliczając wydatek lub obciążenie przelewu z uwzględnieniem prędkości dopływającej wody. Typowym zagadnieniem hydrauliki kanałów otwartych jest obliczenie układu zwierciadła wody wywołanego podpiętrzeniem. Głębokości w kolejnych przekrojach kanału oblicza się, rozwiązując w każdym z nich równanie nieliniowe, którym jest dyskretne równanie energii mechanicznej płynącego strumienia.

Dodajmy, że problem rozwiązania algebraicznych równań nieliniowych występuje powszechnie w trakcie rozwiązywania nieliniowych równań różniczkowych zwyczajnych.

Z kolei układy algebraicznych równań nieliniowych w inżynierii wodnej mogą występować albo jako oddzielne problemy inżynierskie, albo jako fragment rozwiązania innego zagadnienia. Pierwsza sytuacja występuje na przykład w trakcie rozwiązywania zagadnienia ustalonego przepływu w sieci wodociągowej. Natomiast sytuacja druga ma miejsce w trakcie rozwiązywania równań różniczkowych cząstkowych, opisujących nieustalony przepływ w kanałach otwartych. Rozwiązywanie równań metodą różnic lub elementów skończonych ostatecznie prowadzi do układów algebraicznych równań nieliniowych, które rozwiązuje się dla kolejnych wartości rosnącego czasu. Niektóre z wymienionych problemów zostaną omówione szczegółowo w dalszych rozdziałach.

Jak wiadomo, tylko w wyjątkowych wypadkach można rozwiązać równanie nieliniowe w skończonej liczbie operacji. Zwykle stosuje się metody przybliżane. Są to tzw. metody iteracyjne. Istnieje szereg metod przybliżonego rozwiązywania algebraicznych równań nieliniowych. W rozdziale tym przedstawione zostaną tylko wybrane metody znajdowania pierwiastków rzeczywistych równania

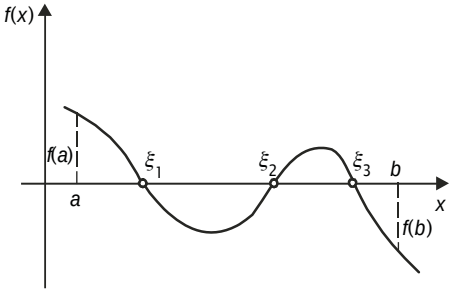
$$f(x) = 0, \tag{2.1}$$

gdzie x jest zmienną rzeczywistą, a $f(x)$ jest dowolną, wystarczająco regularną funkcją. Określenie funkcja wystarczająco regularna oznacza, że $f(x)$ jest ciągła oraz ma ciągłe pochodne potrzebnych rzędów.

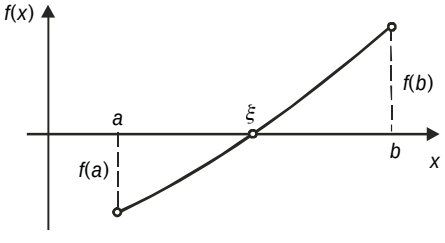
Każdą wartość ξ, która spełnia równanie (2.1), tzn. taką, że $f(\xi) = 0$, nazywamy pierwiastkiem równania. Zakładamy tu, że równanie (2.1) ma tylko pierwiastki odosobnione, tj. dla każdego pierwiastka równania (2.1) istnieje otoczenie, które nie zawiera innych pierwiastków tego równania. Niektóre metody wymagają początkowego przybliżenia, dostatecznie bliskiego szukanego pierwiastka. W innych wypadkach wystarcza znajomość przedziału, w którym jest pierwiastek, bez wymagania ciągłości pochodnej.

Proces obliczania przybliżonych wartości pierwiastków rzeczywistych równania (2.1) dzieli się na dwa etapy:

1) lokalizacja pierwiastków, czyli ustalenie możliwie wąskich przedziałów $\langle a, b\rangle$, które zawierają jeden i tylko jeden pierwiastek równania (2.1),
2) uściślenie wartości pierwiastków, czyli określenie ich wartości z żądaną dokładnością.



Rys. 2.1. Pierwiastki równania $f(x) = 0$



Rys. 2.2. Pojedynczy pierwiastek funkcji $f(x)$ w przedziale $\langle a, b\rangle$

Podstawą algorytmów procesu lokalizacji pierwiastków jest twierdzenie Bolzano-Cauchy'ego (Demidowicz, Maron i Szuwałowa, 1965). Brzmi ono następująco: jeżeli funkcja ciągła $f(x)$ ma na końcach przedziału domkniętego $\langle a, b\rangle$ różne znaki, a więc $f(a) \cdot f(b) < 0$, to wewnątrz tego przedziału istnieje co najmniej jeden pierwiastek równania $f(x) = 0$ (rys. 2.1).

Twierdzenie to nie rozstrzyga o liczbie pierwiastków wewnątrz przedziału $\langle a, b\rangle$. Jednakże w praktyce bardzo często mamy wystarczająco dużo wiadomości o szukanym pierwiastku, aby znalezienie przedziału $\langle a, b\rangle$ nie stanowiło problemu. Na przykład w hydraulice rozwiązaniem równania często jest głębokość wody lub średnica rurociągu. Z założenia muszą one mieć wartość dodatnią, co eliminuje z rozwiązań pierwiastki ujemne. W wypadku, gdy o przebiegu funkcji $f(x)$ mamy skąpe informacje, można – po znalezieniu przedziału $\langle a, b\rangle$ – zbadać znak pochodnej $f'(x)$ w tym przedziale. Jeśli w przedziale otwartym (a, b) istnieje pochodna funkcji $f(x)$ i nie zmienia w nim znaku, czyli jeśli $f'(x) > 0$ (lub $f'(x) < 0$) dla $a < x < b$, to ξ jest jedynym pierwiastkiem (rys. 2.2).

Omawiając w następnych podrozdziałach metody rozwiązania równań nieliniowych, założono, że znany jest przedział, w którym znajduje się jeden pierwiastek.

2.2. Metoda połowienia (bisekcji)

Przyjmujemy, że równanie nieliniowe dane jest w postaci (2.1) $f(x) = 0$, przy czym funkcja $f(x)$ jest ciągła w przedziale domkniętym $\langle a, b\rangle$, oraz że zachodzi nierówność $f(a) \cdot f(b) < 0$. Idea metody połowienia polega na podziale przedziału $\langle a, b\rangle$ na połowy

i sprawdzeniu, w której z nich znajduje się pierwiastek. W ten sposób określa się nowy węższy przedział zawierający pierwiastek ξ. Proces ten powtarzany jest tak długo, aż długość kolejnego przedziału będzie wystarczająco mała.

Algorytm prowadzący do znalezienia pierwiastka równania nieliniowego $f(x) = 0$ w przedziale $\langle a, b\rangle$ metodą połowienia jest następujący:

1) obliczyć punkt $c = (a + b)/2$,
2) obliczyć wartość funkcji w tym punkcie $f(c)$,
3) sprawdzić, czy $f(c) = 0$, jeżeli tak, to c jest szukanym pierwiastkiem; w przeciwnym razie
4) wybrać z przedziałów $\langle a, c\rangle$ i $\langle c, b\rangle$ ten, w którym znajduje się pierwiastek; praktycznie oznacza to sprawdzenie znaku iloczynu $f(a) \cdot f(c)$ (lub $f(c) \cdot f(b)$) i określenie nowego przedziału $\langle a_1, b_1\rangle$; jeśli $f(a) \cdot f(c) < 0$, to $b_1 = c$, w przeciwnym wypadku – $a_1 = c$;
5) sprawdzić, czy długość nowego przedziału $\langle a_1, b_1\rangle$ jest dostatecznie mała; jeżeli $b_1 - a_1 < \varepsilon$, to można przyjąć jako przybliżoną wartość pierwiastka $\xi_p = (a_1 + b_1)/2$; w przeciwnym razie proces obliczeń należy powtórzyć od punktu 1 dla przedziału $\langle a_1, b_1\rangle$.

W wyniku zastosowania opisanego algorytmu albo w pewnym kroku otrzymujemy pierwiastek dokładny, albo ciąg zstępujący przedziałów $\langle a_1, b_1\rangle$, $\langle a_2, b_2\rangle$, $\langle a_3, b_3\rangle$, ..., takich, że

$$f(a_i) \cdot f(b_i) < 0; \qquad i = 1, 2, \ldots \tag{2.2}$$

Po wykonaniu i kroków długość przedziału, w którym jest pierwiastek ξ, wynosi

$$b_i - a_i = \frac{1}{2^i}(b - a); \qquad (i = 1, 2, \ldots). \tag{2.3}$$

Ponieważ końce przedziałów $a_1, a_2, \ldots, a_i$ tworzą ciąg monotonicznie niemalejący, ograniczony z góry, a końce $b_1, b_2, \ldots, b_i$ ciąg nierosnący ograniczony z dołu, to na podstawie równości (2.3) istnieje ich wspólna granica

$$\xi = \lim_{i\to\infty} a_i = \lim_{i\to\infty} b_i\,, \tag{2.4}$$

która jest pierwiastkiem równania (2.1).

Metoda połowienia jest wolno zbieżna i dlatego stosowana jest często do wstępnego określania pierwiastków równania (2.1). Ma ona jednak bardzo istotne zalety:

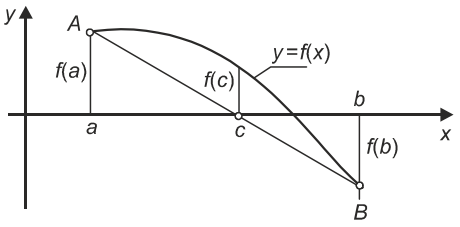
— jest zawsze zbieżna, co oznacza, że istnieje pewność, że w każdej iteracji szukany pierwiastek leży pomiędzy dwiema wartościami, w których $f(x)$ ma różne znaki, a zatem systematyczne powtarzanie obliczeń musi doprowadzić do uzyskania wyniku z żądaną dokładnością;
— jej algorytm można łatwo zaprogramować.

W metodzie połowienia nie zakłada się żadnych ograniczeń na funkcję $f(x)$, ani na pochodne tej funkcji. Wymagana jest jedynie ciągłość funkcji w przedziale $\langle a, b\rangle$.

2.3. Metoda interpolacji liniowej (siecznych)

Metoda ta należy do najstarszych metod rozwiązywania równania (2.1) $f(x) = 0$, będąc przy tym znacznie skuteczniejsza od metody połowienia.

Załóżmy, że znany jest przedział $\langle a, b\rangle$, w którym znajduje się pierwiastek (rys. 2.3).



Rys. 2.3. Schemat metody interpolacji liniowej

Przez punkty $A(a, f(a))$ i $B(b, f(b))$ można poprowadzić sieczną, która przetnie oś odciętych w punkcie c. Wartość odciętej c określona jest zależnością:

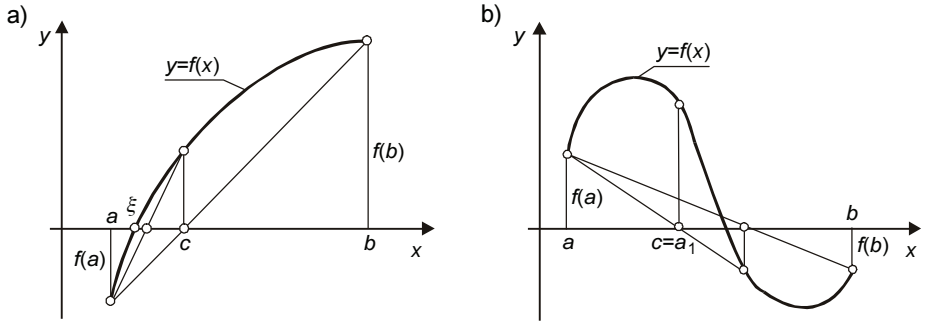
$$c = \frac{af(b) - bf(a)}{f(b) - f(a)}. \tag{2.5}$$

Stosowanie wzoru (2.5) w pobliżu pierwiastka jest jednak niewskazane, gdyż zarówno w liczniku, jak i w mianowniku pojawiają się różnice liczb tego samego rzędu, co powoduje znaczne błędy zaokrągleń. Wzór ten można jednak przekształcić, dodając i odejmując a. W konsekwencji można go zapisać w postaci sumy

$$c = a + f(a)\frac{a - b}{f(b) - f(a)}, \tag{2.6}$$

w której drugi składnik można uważać za poprawkę wartości a.

W równaniu (2.6) jeden lub oba końce mogą być „ruchome". Zależy to od właściwości funkcji wewnątrz przedziału $\langle a, b\rangle$ (rys. 2.4).



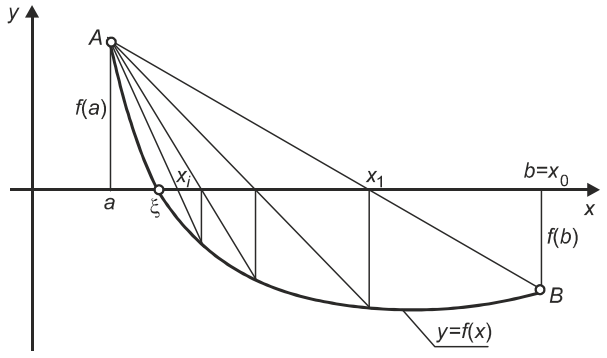
Rys. 2.4. Różne przebiegi funkcji w przedziale $\langle a, b\rangle$

Jeżeli założymy, że w przedziale $\langle a, b\rangle$ istnieje tylko jeden pierwiastek, a pochodna $f''(x)$ nie zmienia znaku w tym przedziale, to jeden z końców przedziału będzie nieruchomy. W wypadku gdy $f(a) > 0$ i $f''(x) > 0$ dla $a \le x \le b$, koniec a jest nieruchomy (rys. 2.5) i wzór (2.6) można zapisać następująco:

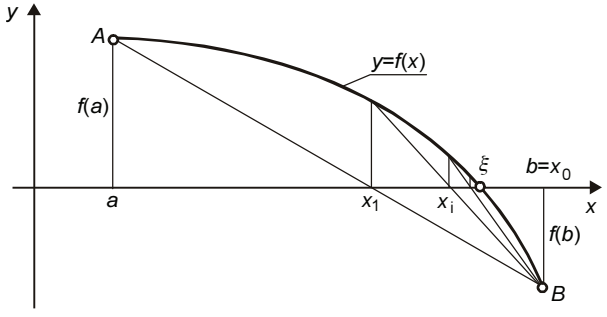
$$\begin{aligned} x_0 &= b, \\ x_{i+1} &= x_i - \frac{f(x_i)}{f(x_i) - f(a)}(x_i - a), \end{aligned} \tag{2.7}$$

gdzie: $i = 0, 1, 2, ...$

Ten szczególny przypadek metody siecznych nazywany jest *metodą regula falsi* (Fortuna, Macukow i Wąsowski, 1982). Analogiczną zależność można napisać w sytuacji, gdy koniec $x = b$ jest nieruchomy, jak ma to miejsce na rys. 2.6.



Rys. 2.5. Metoda interpolacji liniowej z nieruchomym końcem $x = a$



Rys. 2.6. Metoda interpolacji liniowej z nieruchomym końcem $x = b$

W celu oszacowania błędu przybliżonego pierwiastka można wykorzystać następujące twierdzenie (Demidowicz, Maron i Szuwałowa, 1965).
Niech ξ będzie dokładną, a $\bar{x}$ przybliżoną wartością pierwiastka równania (2.1), przy czym obie te liczby znajdują się w przedziale domkniętym $\langle a, b\rangle$. Jeśli dla $x \in \langle a, b\rangle$ zachodzi nierówność $|f'(x)| \geq m > 0$, to prawdziwe jest następujące oszacowanie:

$$\left|\bar{x} - \xi\right| \leq \frac{f(\bar{x})}{m}. \tag{2.8}$$

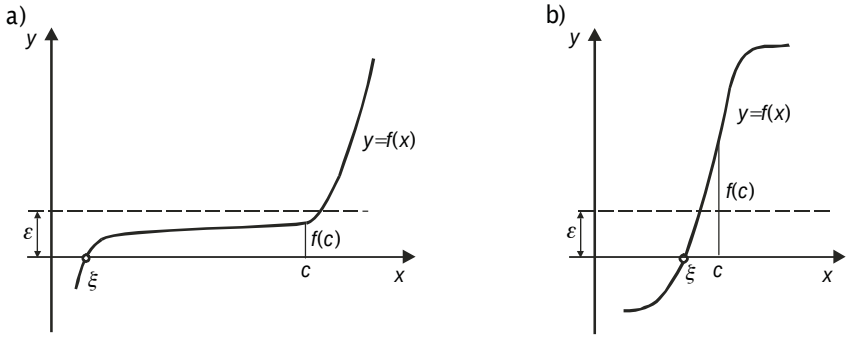
Błąd bezwzględny przybliżenia x_i, gdy znane są wartości x_i oraz przybliżenia poprzedniego x_{i-1}, można określić według wzoru (Demidowicz, Maron i Szuwałowa, 1965)

$$\left|\xi - \bar{x}\right| \leq \frac{M - m}{m}\left|x_i - x_{i-1}\right|, \tag{2.9}$$

gdzie za M i m można wziąć odpowiednio: największą i najmniejszą wartość $|f'(x)|$ w przedziale domkniętym $\langle a, b\rangle$. Jeśli natomiast przedział $\langle a, b\rangle$ jest na tyle wąski, że zachodzi nierówność $M \leq 2\,m$, to na mocy (2.9) otrzymujemy $|\xi - x_i| \leq |x_i - x_{i-1}|$. Zatem, gdy zachodzi nierówność $|x_i - x_{i-1}| < \varepsilon$, gdzie ε jest zadanym kresem górnym błędu bezwzględnego, to jest pewne, że nierówność $|\xi - x_i| < \varepsilon$ jest również prawdziwa.

Oszacowanie błędu bezwzględnego przybliżenia x_i na podstawie wzorów (2.8) lub (2.9) może być jednak kłopotliwe ze względu na konieczność badania pochodnej $f'(x)$

w przedziale $\langle a, b\rangle$. Dlatego często w praktyce warunkiem na zakończenie obliczeń jest sprawdzenie, czy $|f(x_i)| < \varepsilon$. Określenie ε tak, aby x_i można było uznać na dobre przybliżenie pierwiastka równania $f(x) = 0$, jest problemem, który należy rozwiązać indywidualnie w każdym zadaniu. Podany warunek oparty jest na dość rozpowszechnionym poglądzie, że x_i jest dobrym przybliżeniem pierwiastka ξ, jeżeli $|f(x_i)|$ jest małą liczbą, i na odwrót, jeżeli $|f(x_i)|$ ma dużą wartość, to liczba x_i zostaje uznana za złe przybliżenie. Jak wynika z rysunku 2.7, takie podejście nie musi być prawidłowe.



Rys. 2.7. Różne przebiegi $f(x)$ w pobliżu pierwiastka

Algorytm prowadzący do znalezienia pierwiastka równania $f(x) = 0$ metodą interpolacji liniowej jest bardzo zbliżony do algorytmu dla metody połowienia. Różnica występuje tylko w sposobie określania kolejnych przybliżeń c rozwiązania dokładnego.

■ *Przykład 2.1*

Obliczanie głębokości normalnej w kanale o przekroju trapezowym metodą siecznych

Jak wiadomo natężenie przepływu wody w kanale w warunkach ruchu ustalonego jednostajnego określa równanie Manninga (Czetwertyński i Utrysko, 1969):

$$Q = \frac{1}{n} R^{2/3} s^{1/2} A\,, \tag{2.1.1}$$

gdzie: Q – natężenie przepływu [m^3/s],
R – promień hydrauliczny [m],
s – spadek linii energii (w tym wypadku równy spadkowi dna) [/],
n – współczynnik szorstkości wg Manninga [$s/m^{1/3}$],
A – powierzchnia przekroju czynnego [m^2].

Głębokość H, która spełnia równanie (2.1.1), nosi nazwę głębokości normalnej. Równanie to można zapisać w postaci równoważnej

$$f(H) = Q - \frac{s^{1/2}}{n} R^{2/3} A = 0\,. \tag{2.1.2}$$

Jest ono ważne dla dowolnego kształtu przekroju poprzecznego. Dla dodatnich głębokości $f(H)$ jest funkcją malejącą (rys. 2.1.1). W prezentowanym przykładzie przyjęto, że przekrój kanału jest trapezem o nachyleniu skarp 1 : M i szerokości dna B (rys. 2.1.2). Powierzchnia przekroju czynnego w tym wypadku będzie równa

$$A = B\,H + MH^2, \qquad (2.1.3)$$

a promień hydrauliczny

$$R = \frac{BH + MH^2}{B + 2H\sqrt{1+M^2}}\ . \qquad (2.1.4)$$

Jeżeli zatem znamy: Q, n, s, B oraz M, to głębokość normalną H można obliczyć jako miejsce zerowe funkcji:

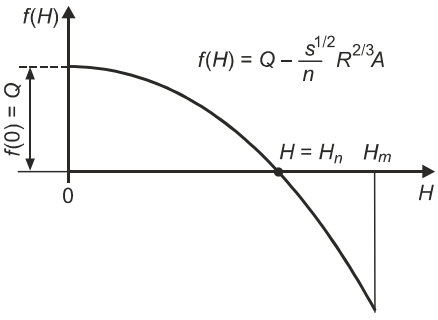
$$f(H) = Q - \frac{s^{1/2}}{n}\frac{(BH + MH^2)^{5/3}}{\left(B + 2H\sqrt{1+M^2}\right)^{2/3}}\ . \qquad (2.1.5)$$

Równanie $f(H) = 0$ jest nieliniowe. Do jego rozwiązania którego zastosujemy metodę siecznych.

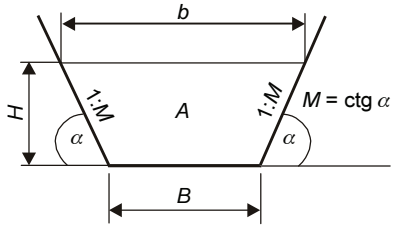
W algorytmie obliczeń pierwiastka wyróżniamy dwa etapy:

— lokalizację pierwiastka przez tablicowanie $f(H)$ z interwałem ΔH,

— obliczenie pierwiastka z żądaną dokładnością.



Rys. 2.1.1. Przebieg funkcji $f(H)$ opisanej wzorem (2.1)



Rys. 2.1.2. Przekrój poprzeczny kanału trapezowego

W celu lokalizacji pierwiastka dokonujemy tablicowania funkcji (2.1.5) z interwałem ΔH, badając znak iloczynu $f(H)$ na końcach przedziału. Następnie poszukujemy kolejnych przybliżeń pierwiastka według procedury opisanej w podrozdziale 2.3. Obliczenie kończymy, gdy spełniony jest warunek dokładności rozwiązania o postaci:

$$|f(H)| \le \varepsilon. \qquad (2.1.6)$$

Z równania (2.1.5) wynika, że $f(H)$ ma wymiar natężenia przepływu, co bardzo ułatwia dobór wartości parametru ε.

Tok obliczeń jest następujący:

1) podstaw $H_1 = 0$,
2) podstaw $H_2 = H_1 + \Delta H$,
3) sprawdź warunek: $f(H_1) \cdot f(H_2) > 0$,
   jeśli jest spełniony – $H_1 = H_2$, i powróć do 2, jeśli nie – przejdź do 4,
4) oblicz $H = H_1 + f(H_1)\ \dfrac{H_1 - H_2}{f(H_2) - f(H_1)}$,
5) sprawdź warunek: $|f(H)| \le \varepsilon$;
   jeśli jest spełniony – zakończ obliczenia, gdyż H jest poszukiwanym przybliżeniem pierwiastka;

jeśli nie – dokonaj przesunięcia jednego z końców przedziału, podstawiając
$H_2 = H$, gdy $f(H_1) \cdot f(H) < 0$ lub
$H_1 = H$, gdy $f(H_1) \cdot f(H) > 0$,
i przejdź do 4.

W celu oceny efektywności metody należy obliczać liczbę iteracji, koniecznych do znalezienia rozwiązania z żądaną dokładnością, tzn. liczbę wywołań funkcji $f(H)$.

Tabela 2.1.1

Wyniki obliczeń głębokości normalnej w kanale trapezowym metodą siecznych

Q [m^3/s]	H [m]	A [m^2]	Liczba iteracji	$f(H)$ [m^3/s]
4,0	0,828	4,342	4	0,0002
6,0	1,038	5,771	4	0,0004
8,0	1,216	7,079	4	0,0003
10,0	1,371	8,306	3	0,0005
12,5	1,544	9,755	4	0,0002
15,0	1,700	11,134	4	0,0002

Obliczona głębokość normalna determinuje inne, parametry kanału, będące istotną informacją z punktu widzenia hydrotechnika projektującego kanał. Są to: szerokość zwierciadła wody b, powierzchnia przekroju czynnego A, średnia prędkość przepływu v.

Opisany algorytm zastosowano do obliczeń głębokości normalnych w kanale o następujących parametrach: $B = 4{,}0$ m, $M = 1{,}50$, $s = 0{,}001$, $n = 0{,}025$ $m^{1/2} \cdot s^{-1}$. Obliczenia wykonano dla różnych wartości natężenia przepływu Q, a uzyskane wyniki przy $\Delta H = 0{,}50$ m zestawiono w tabeli 2.1.1. Zauważmy, że rozwiązanie z dokładnością $\varepsilon \le 0{,}001$ uzyskano w każdym wypadku po 3–4 iteracjach. ■

2.4. Metoda iteracji prostej

Jest to jedna z najważniejszych metod numerycznego rozwiązywania równań nieliniowych. Idea tej metody jest następująca. Równanie (2.1), $f(x) = 0$, gdzie $f(x)$ jest funkcją ciągłą, dla którego należy wyznaczyć pierwiastek rzeczywisty, przekształcamy do równania równoważnego

$$x = \varphi(x). \tag{2.10}$$

Załóżmy, że znane jest dostatecznie bliskie przybliżenie pierwiastka x_0. Wartość x_0 można podstawić do prawej strony równania (2.10). Otrzymamy wówczas liczbę

$$x_1 = \varphi(x_0). \tag{2.11}$$

Wartość x_1 podstawiamy ponownie do równania (2.10) i obliczamy $x_2 = \varphi(x_1)$. Powtarzając ten proces, otrzymujemy ciąg liczb:

$$\begin{aligned} &x_0, \\ &x_1 = \varphi(x_0), \\ &x_2 = \varphi(x_1), \\ &\vdots \\ &x_i = \varphi(x_{i-1}). \end{aligned} \tag{2.12}$$

Z powyższym procesem wiążą się następujące problemy:
— jak znaleźć funkcję φ?
— przy jakich warunkach ciąg x_i jest zbieżny?

Ciąg (2.12) jest zbieżny, jeżeli istnieje granica $\xi = \lim_{i\to\infty} x_i$. Przechodząc w równaniu (2.12) do granicy przy założeniu ciągłości funkcji $\varphi(x)$, mamy

$$\lim_{i\to\infty} x_i = \varphi(\lim_{i\to\infty} x_{i-1})$$

lub

$$\xi = \varphi(\xi). \tag{2.13}$$

Zatem granica ξ jest pierwiastkiem równania (2.10). Można ją obliczyć ze wzoru (2.12) z dowolną dokładnością. Jednak możliwość stosowania metody iteracji prostej uwarunkowana jest zbieżnością ciągu (2.12). Problem zbieżności procesu iteracyjnego rozstrzyga następujące twierdzenie (Demidowicz, Maron i Szuwałowa, 1965).
Niech funkcja $\varphi(x)$ będzie określona i różniczkowalna w przedziale domkniętym $\langle a, b\rangle$ i jej wartości należą do tego przedziału, tj. $\varphi(x) \in \langle a, b\rangle$. Wtedy, jeśli istnieje ułamek właściwy q taki, że

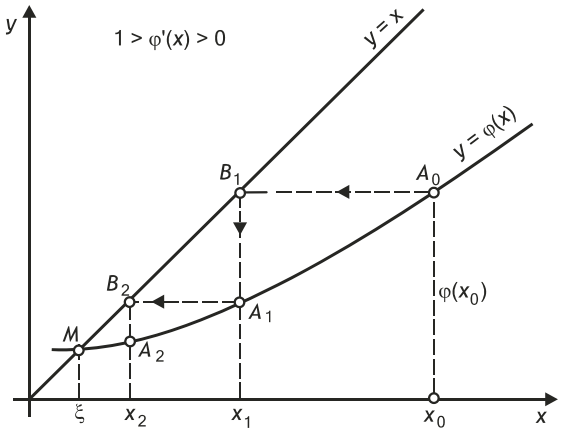
$$|\varphi'(x)| \le q < 1 \tag{2.14}$$

dla $a < x < b$, to:
— proces iteracyjny (2.12) jest zbieżny niezależnie od przybliżenia początkowego $x_0 \in \langle a, b\rangle$,
— wartość $\xi = \lim_{i\to\infty} x_i$ jest jedynym pierwiastkiem równania $x = \varphi(x)$ w przedziale domkniętym $\langle a, b\rangle$.

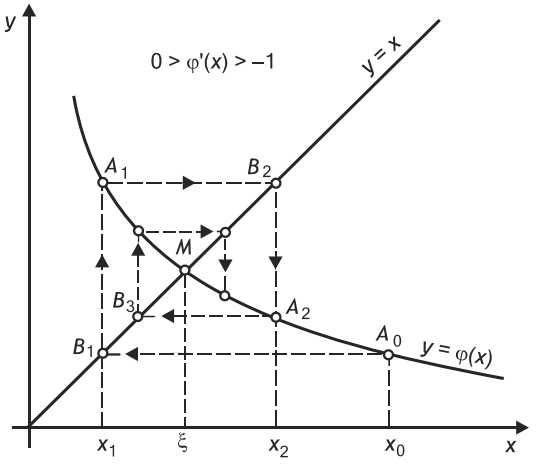
Powyższe twierdzenie pozostaje w mocy, gdy funkcja $\varphi(x)$ jest określona i różniczkowalna w przedziale nieskończonym $-\infty < x < +\infty$ i spełniona jest nierówność (2.14). Ponadto, przy założeniach twierdzenia metoda iteracji jest zbieżna dla dowolnie wybranego w przedziale $\langle a, b\rangle$ przybliżenia początkowego x_0. Zatem, w wypadku gdy błąd w obliczeniach nie wyprowadza kolejnego przybliżenia poza granice przedziału $\langle a, b\rangle$, metoda sama się koryguje, a błędna wartość jest traktowana jako nowe przybliżenie. Taki błąd nie wpływa na wynik końcowy. Własność autokorekcji stawia metodę iteracji prostej wśród najpewniejszych metod numerycznych.

Metodę iteracji prostej można zinterpretować geometrycznie. Na płaszczyźnie xy kreślimy wykresy funkcji $y = x$ i $y = \varphi(x)$. Pierwiastek rzeczywisty równania (2.10) jest odciętą punktu M, czyli punktu przecięcia krzywej $y = \varphi(x)$ z prostą $y = x$ (rys. 2.8). Proces iteracyjny zapisany wzorem (2.12) ilustruje linia łamana prowadzona przez punkty A_0, B_1, A_1, B_2, ... Jak widać z rysunku 2.8, odcięte punktów A_1, A_2, ... są kolejnymi przybliżeniami x_1, x_2, ... pierwiastka ξ.

Należy dodać, że proces iteracyjny zapisany wzorem (2.12) zależnie od wartości pochodnej $\varphi'(x)$ może przebiegać w taki sposób, że kolejne przybliżenia pierwiastka ξ będą znajdować się na przemian po jego lewej i prawej stronie (rys. 2.9), a także może być rozbieżny (rys. 2.10 i rys. 2.11).



Rys. 2.8. Schemat działania metody iteracji prostej dla $1 > \varphi'(x) > 0$



Rys. 2.9. Schemat działania metody iteracji prostej dla $0 > \varphi'(x) > -1$

Zaprogramowanie tej metody jest wyjątkowo proste. Wyjściowe równanie $f(x) = 0$ należy przekształcić do postaci $x = \varphi(x)$ oraz określić wartość przybliżenia początkowego x_0. Kolejne przybliżenia obliczane są według wzorów (2.12).

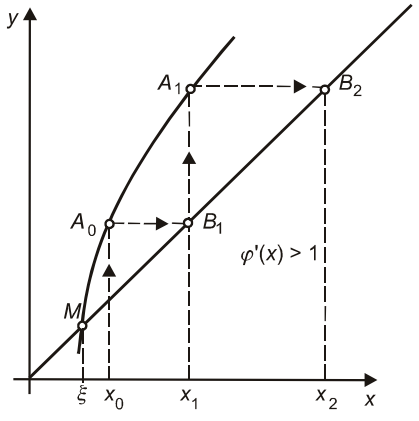
Proces iteracyjny należy kontynuować tak długo, aż dwa kolejne przybliżenia x_i oraz x_{i-1} spełnią nierówność

$$\left| x_i - x_{i-1} \right| \le \frac{1-q}{q} \varepsilon \,, \tag{2.15}$$

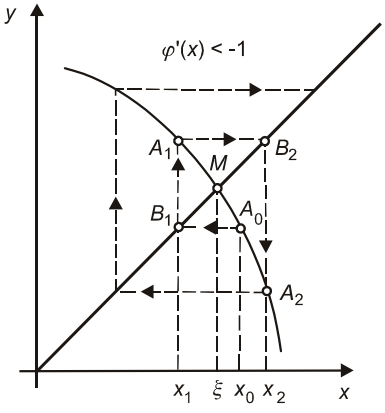
gdzie ε jest kresem górnym błędu bezwzględnego pierwiastka, zaś q powinno spełniać warunek $|\varphi'(x)| \le q < 1$.

Na ogół nie znamy dokładnie wartości q. Można jednak oszacować wartość przybliżoną w następujący sposób (Demidowicz, Maron i Szuwałowa, 1965)

$$q \approx \left| \frac{\varphi(x_{i-1}) - \varphi(x_{j-2})}{x_{i-1} - x_{i-2}} \right| = \left| \frac{x_i - x_{i-1}}{x_{i-1} - x_{i-2}} \right|. \tag{2.16}$$



Rys. 2.10. Rozbieżny proces iteracyjny



Rys. 2.11. Rozbieżny proces iteracyjny

Powodzenie wykorzystania tej metody zależy od zbieżności ciągu (2.12). W niektórych wypadkach możliwe jest określenie dostatecznych warunków zbieżności. Jednak w praktyce często trzeba się zadowolić eksperymentalnym sprawdzaniem zbieżności. Jeśli daje ono wynik negatywny, należy dokonać innego przekształcenia równania (2.1) w (2.10). Przekształcenie to może być wykonane na dowolnie wiele sposobów. Trafność wyboru takiego przekształcenia, oprócz zapewnienia uzyskania samego rozwiązania, decyduje również o szybkości zbieżności procesu iteracyjnego (2.12). Informacją o szybkości zbieżności metody jest tzw. wykładnik zbieżności. Mówi się, że metoda jest rzędu p, jeśli istnieje stała K taka, że dla dwóch kolejnych przybliżeń x_i oraz x_{i+1} pierwiastka ξ zachodzi relacja (Fortuna, Macukow i Wąsowski, 1982):

$$\left| x_{i+1} - \xi \right| \le K \left| x_i - \xi \right|^p. \tag{2.17}$$

Dla $p = 1$ metodę określa się jako liniową, dla $p = 2$ – jako kwadratową, a dla $p = 3$ – jako sześcienną.

■ *Przykład 2.2*

Wpływ sposobu przekształcenia $f(x) = 0 \rightarrow x = \varphi(x)$ na własności procesu iteracyjnego (Dahlquist i Bjorck, 1983)

Rozważmy równanie postaci

$$x + \ln x = 0, \tag{2.2.1}$$

którego pierwiastek wynosi $\xi \approx 0{,}5$ (dokładniej 0,5671). Przekształćmy je do postaci (2.10), tj. $x = \varphi(x)$, na trzy sposoby:

1) $$x = -\ln x\,, \tag{2.2.2}$$

z ogólnym wzorem rekurencyjnym

$$x_i = -\ln x_{i-1}, \quad (i = 1, 2, ...); \tag{2.2.3}$$

2) $$x = e^{-x}, \tag{2.2.4}$$

ze wzorem rekurencyjnym

$$x_i = e^{-x_{i-1}} \quad (i = 1, 2, ...), \tag{2.2.5}$$

oraz

3) $$x = \frac{1}{2}(x + e^{-x})\,, \tag{2.2.6}$$

dla którego wzór rekurencyjny jest następujący:

$$x_i = \frac{1}{2}(x_{i-1} + e^{-x_{i-1}}) \qquad (i = 1, 2, ...). \tag{2.2.7}$$

W sposób oczywisty nasuwają się tutaj następujące pytania:
— którego z tych wzorów można użyć?
— którego wzoru należy użyć?

Odpowiedź na pytanie pierwsze wynika bezpośrednio z warunku (2.14) $|\varphi'(\xi)| < 1$. Dla przypadku (2.2.2) będzie

$$\left|\varphi'(\xi)\right| = \left|\frac{1}{\xi}\right| \approx \frac{1}{0{,}5} = 2\,.$$

Ciąg kolejnych przybliżeń określony przez (2.2.3) jest więc rozbieżny, dla nawet bardzo bliskiego pierwiastkowi ξ przybliżenia początkowego x_0 (zbieżność jest możliwa wyłącznie dla $x_0 = \xi$).

Dla przypadku (2.2.4) otrzymamy

$$\left|\varphi'(\xi)\right| = e^{-\xi} = e^{-0{,}5} \approx 0{,}6\,.$$

Tutaj ciąg kolejnych przybliżeń (2.2.5) jest zbieżny, dla dowolnego przybliżenia początkowego.

Dla przypadku (2.2.6) będziemy mieli

$$\left|\varphi'(\xi)\right| = \frac{1}{2}\left|1 - e^{-\xi}\right| = \frac{1}{2}\left|1 - e^{-0{,}5}\right| \approx 0{,}2\,.$$

Zatem ciąg kolejnych przybliżeń (2.2.7) jest również zbieżny.

Odpowiedź na pytanie drugie wynika pośrednio z obliczeń kolejnych przybliżeń dla omawianego zadania. Przyjmijmy, że założona dokładność rozwiązania wynosi $\varepsilon = 0{,}00001$, a przybliżenie początkowe $x_0 = 1{,}0$. Dla przekształcenia (2.2.4) rozwiązanie otrzymuje się po 21 iteracjach, uzyskując wartość pierwiastka równą 0,5671. W tabeli 2.2.1 pokazano, ile iteracji trzeba wykonać dla przekształceń (2.2.4) oraz (2.2.6), gdy zmienia się przybliżenie początkowe.

Lepsze wyniki, ze względu na liczbę iteracji, uzyskano dla wzoru (2.2.6), dla którego pochodna φ' jest mniejsza. Zatem można wnioskować, że tempo zbieżności procesu iteracyjnego określa wartość pochodnej φ' – im wartość tej pochodnej jest mniejsza, tym więk-

sze jest tempo zbieżności procesu iteracyjnego (2.12). Istotne jest więc pytanie: czy można polepszyć tempo zbieżności? Zapiszmy wzór (2.2.6) do nieco zmienionej postaci

$$x_i = \frac{\alpha x_{i-1} + e^{-x_{i-1}}}{\alpha + 1}, \tag{2.2.8}$$

któremu odpowiada pochodna

$$\varphi'(x) = \frac{\alpha - e^{-x}}{\alpha + 1}. \tag{2.2.9}$$

Tabela 2.2.1

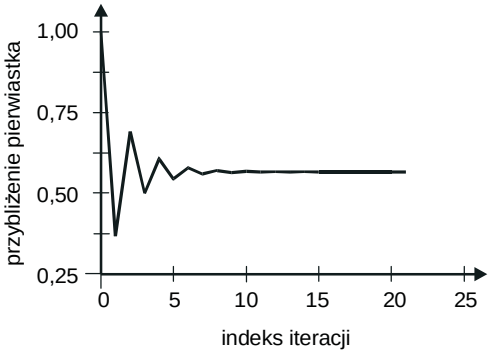
Zależność liczby iteracji od przybliżenia początkowego dla równań (2.2.4) i (2.2.6)

Równanie / Przybl. począt.	(2.2.4)	(2.2.6)
1,0	21	9
10,0	23	12
1000,0	23	19

Po uwzględnieniu (2.2.4) zapiszemy ją następująco:

$$\varphi'(\xi) = \frac{\alpha - \xi}{\alpha + 1}. \tag{2.2.10}$$

Dla wartości α bliskiej wartości ξ wartość pochodnej będzie mała, a zatem zbieżność powinna być szybsza aniżeli dla wzoru (2.2.6).



Rys. 2.2.1. Przybliżenie rozwiązania w kolejnych iteracjach dla równania (2.2.8) przy $x_0 = 1$ i $\alpha = 0{,}6$

Obliczenia wykonane dla wyjściowych danych jak w poprzednich przykładach oraz dla $\alpha = 0{,}6$ potwierdzają to przypuszczenie. Rozwiązanie z dokładnością $\varepsilon = 0{,}00001$ otrzymuje się dla:

$x_0 =$ 1,0 po 5 iteracjach,
$x_0 =$ 10,0 po 7 iteracjach,
$x_0 =$ 1000,0 po 11 iteracjach.

Porównując powyższe wyniki z wynikami zawartymi w tabeli 2.2.1, zauważamy zdecydowaną poprawę tempa zbieżności. ■

■ *Przykład 2.3*

Obliczenie współczynnika oporu λ wg Colebrooka-White'a metodą iteracji prostej

Jedną z częściej stosowanych formuł empirycznych dla określenia współczynnika oporu λ jest wzór Colebrooka-White'a (Czetwertyński i Utrysko, 1969)

$$\frac{1}{\sqrt{\lambda}} = -2\log\left(\frac{2{,}51}{R_e\sqrt{\lambda}} + \frac{e}{3{,}72d}\right), \tag{2.3.1}$$

gdzie: λ – współczynnik oporu,
R_e – liczba Reynoldsa [/],
e – chropowatość przewodu [m],
d – średnica przewodu [m].

Równanie to ze względu na λ jest równaniem nieliniowym, które można rozwiązać metodą iteracji prostej (Burzyński, Granatowicz, Piwecki i Szymkiewicz, 1991). Stosując podstawienie $Z = 1/\sqrt{\lambda}$, zapisujemy je w postaci

$$Z = -2\log\left(\frac{2{,}51Z}{R_e} + \frac{e}{3{,}72d}\right), \tag{2.3.2}$$

tzn.

$$Z = \varphi(Z). \tag{2.3.3}$$

Jak wiadomo, metoda iteracji prostej jest zbieżna, gdy jest spełniony warunek $|\varphi'(Z)| < 1$. Uwzględniając, że lg N = ln N/ln 10 oraz wprowadzając oznaczenia:

$$A = \frac{2{,}51}{R_e}, \qquad B = \frac{e}{3{,}72d}, \qquad C = -\frac{2}{\ln 10} = -0{,}8685\,, \tag{2.3.4a,b,c}$$

otrzymujemy:

$$\varphi(Z) = C \ln (A\,Z + B) \tag{2.3.5}$$

oraz

$$\frac{d\varphi}{dZ} = C\frac{A}{AZ + B}\,. \tag{2.3.6}$$

W celu oszacowania maksymalnej wartości pochodnej $d\varphi/dZ$ równanie (2.3.6) sprowadzamy do postaci

$$\frac{d\varphi}{dZ} = \frac{C}{Z + B/A} \approx \frac{-8{,}685}{10Z + R_e\,e/d}\,. \tag{2.3.7}$$

Z badań eksperymentalnych wiadomo (Czetwertyński i Utrysko, 1969), że $\lambda \in \langle 0{,}005 - 0{,}08\rangle$; zatem

$$Z_{\min} = \frac{1}{\sqrt{\lambda_{\max}}} \approx 3{,}5\,. \tag{2.3.8}$$

Minimalna wartość liczby Reynoldsa dla przepływu burzliwego w rurociągu wynosi 2300, a minimalna wartość chropowatości dla rur stalowych e = 0,02 mm. Jako maksymalną

średnicę rurociągu przyjęto d = 1000 mm. Uwzględniając powyższe dane, otrzymujemy oszacowanie

$$\left.\frac{d\varphi}{dZ}\right|_{\max} \approx \frac{-8{,}685}{10 \cdot 3{,}5 + 2300 \cdot 0{,}02/1000} \approx 0{,}25 < 1 . \tag{2.3.9}$$

Jak widać, warunek $|\varphi'(Z)| < 1$ jest spełniony dla wszystkich wartości R_e, e, d. Nie jest więc konieczne poszukiwanie przedziału, w którym znajduje się pierwiastek. Za pierwsze przybliżenia poszukiwanej wartości współczynnika λ można przyjąć na przykład wartość 0,04. Wynika stąd, że

$$Z_0 = \frac{1}{\sqrt{0{,}04}} = 5{,}0 . \tag{2.3.10}$$

Omawiając metodę iteracji prostej, podano że obliczanie kolejnych przybliżeń powinno być wykonywane tak długo, aż spełniona będzie nierówność

$$|Z_i - Z_{i-1}| \le \frac{1-q}{q}\varepsilon . \tag{2.3.11}$$

W tym wypadku za q można przyjąć

$$q = \left.\frac{d\varphi}{dz}\right|_{\max} = 0{,}25 .$$

Dla potrzeb projektowania wystarcza określenie λ z dokładnością do 0,00001. Jeśli wziąć pod uwagę, że $Z = 1/\sqrt{\lambda}$,

$$dZ = -\frac{1}{2\lambda^{3/2}} d\lambda .$$

Zatem:

$$\Delta Z = -\frac{0{,}000005}{1{,}5}, \quad \text{zaś} \quad \varepsilon = \frac{0{,}000005}{0{,}08^{1.5}} \approx 0{,}00082 .$$

Wyniki obliczeń λ dla danych R_e, d oraz e zestawiono w tabeli 2.3.1. W ostatniej kolumnie podano liczbę wykonanych iteracji.

Tabela 2.3.1

Wyniki obliczeń współczynnika oporu

R_e	e [mm]	d [mm]	λ	Liczba iteracji
2500	1,00	100,0	0,05388	6
10000	1,00	100,0	0,04307	5
40000	1,00	100,0	0,03930	3
100000	1,00	100,0	0,03844	3
200000	1,00	100,0	0,03814	3

Potwierdzają one znany fakt zaniku wpływu liczby Reynoldsa przy jej dużych wartościach na wartość współczynnika λ. ■

2.5. Metoda Newtona (stycznych)

Rozwiązujemy równanie nieliniowe o postaci (2.1) $f(x) = 0$. Załóżmy, że znamy x_i, przybliżoną wartość pierwiastka tego równania. Niech kolejne przybliżenie pierwiastka leży w odległości h od przybliżenia poprzedniego, czyli

$$x_{i+1} = x_i + h, \tag{2.18}$$

gdzie h jest wielkością małą.
Zakładając, że funkcja $f(x)$ jest ciągła i różniczkowalna, jej wartość w punkcie x_{i+1} można określić z rozwinięcia w szereg Taylora

$$f(x_i + h) = f(x_i) + hf'(x_i) + \frac{h^2}{2} f''(x_i) + \ldots . \tag{2.19}$$

Metoda Newtona polega na pominięciu w równaniu (2.19) wyrazów z pochodnymi drugiego rzędu i wyższymi. Ponieważ można przyjąć, że $f(x_i + h) = 0$, to

$$f(x_i) + h f'(x_i) = 0, \tag{2.20}$$

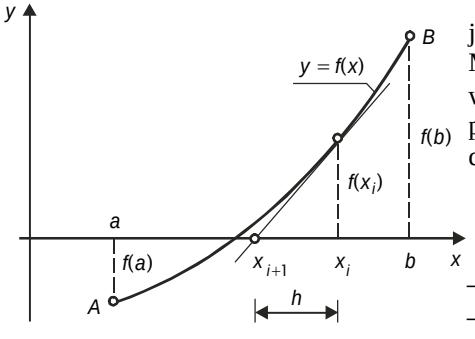
skąd

$$h = -\frac{f(x_i)}{f'(x_i)}. \tag{2.21}$$

Podstawiając otrzymaną poprawkę do wzoru (2.18), określamy następne przybliżenie pierwiastka

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}. \tag{2.22}$$

Wzór (2.22) można także otrzymać, wychodząc z równania stycznej do krzywej $y = f(x)$ w punkcie x_i.



Rys. 2.12. Graficzna interpretacja metody Newtona

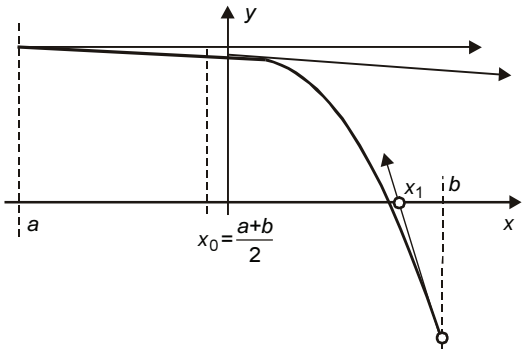
Istotnym problemem w metodzie Newtona jest przyjęcie przybliżenia początkowego. Można udowodnić (Demidowicz, Maron i Szuwałowa, 1965), że w przedziale $\langle a, b\rangle$ dobrym przybliżeniem początkowym będzie taki punkt, dla którego spełniona jest nierówność

$$f(x_0)\, f''(x_0) > 0. \tag{2.23}$$

Jeżeli zachodzą następujące warunki:

- $f(a) \cdot f(b) < 0$,
- $f''(x)$ jest ciągła w przedziale $\langle a, b\rangle$ i nie zmienia w nim znaku,
- styczne do krzywej $y = f(x)$ poprowadzone w punktach o odciętych a i b przecinają oś x wewnątrz przedziału $\langle a, b\rangle$,

to jako przybliżenie początkowe x_0 można przyjąć dowolną wartość z przedziału $\langle a, b\rangle$ (Fortuna, Macukow i Wąsowski 1982).
Skutki niespełnienia ostatniego warunku ilustruje rys. 2.13.

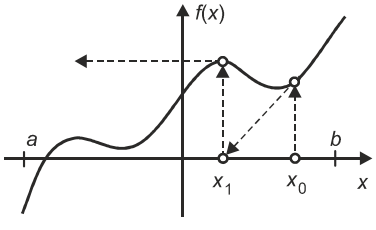


Rys. 2.13. Metoda Newtona – przykład drastycznie złej zbieżności w wypadku złego przyjęcia przybliżenia początkowego

Przyjęcie przybliżenia początkowego zarówno w punkcie a (pochodna funkcji bardzo bliska zeru), jak i w punkcie leżącym w środku przedziału $\langle a, b\rangle$ (argument następnego przybliżenia jest bardzo dużą liczbą) prowadzi do załamania lub wydłużenia obliczeń w trakcie realizacji komputerowej algorytmu. Natomiast przyjmując punkt startowy na prawym krańcu przedziału, uzyskujemy gwarancję uzyskania rozwiązania. Nasuwa się tutaj następujący wniosek: w realizacji komputerowej, przed użyciem metody Newtona, należy sprawdzić, na którym krańcu przedziału poszukiwania pierwiastka pochodna funkcji ma większą wartość. Przybliżeniem początkowym zapewniającym najlepsze rozwiązanie będzie ten kraniec przedziału, gdzie pochodna ma wartość większą.

Testy numeryczne potwierdzają skuteczność stosowania metody Newtona nawet dla wartości początkowych dalekich od pierwiastka, chociaż formuła (2.22) była wyprowadzona intuicyjnie tylko dla wartości x bliskich pierwiastka ξ. Wyjaśnienie tego problemu podaje Legras (1974).

Ze wzoru (2.22) wynika, że im większa jest wartość pochodnej $f'(x)$ w otoczeniu pierwiastka ξ, tym wartość poprawki jest mniejsza. Oznacza to, że metoda Newtona jest bardzo efektywna, gdy krzywa $f(x)$ jest stroma w otoczeniu pierwiastka. W wypadku, gdy funkcja $f(x)$ ma lokalne ekstrema ($f'(x) = 0$), obliczenie pierwiastka za pomocą tej metody może być niemożliwe. Sytuację taką pokazano na rys. 2.14.



Rys. 2.14. Metoda Newtona – funkcja z lokalnymi ekstremami

Chociaż dla metody Newtona istnieje teoretyczne oszacowanie błędu przybliżenia, to ze względu na praktyczne kłopoty z jego stosowaniem, w wypadku gdy funkcja $f(x)$ jest dostatecznie gładka, dokładność obliczanego przybliżenia x_i określa się na podstawie nierówności

$$\left|f(x_i)\right| < \varepsilon \,. \tag{2.24}$$

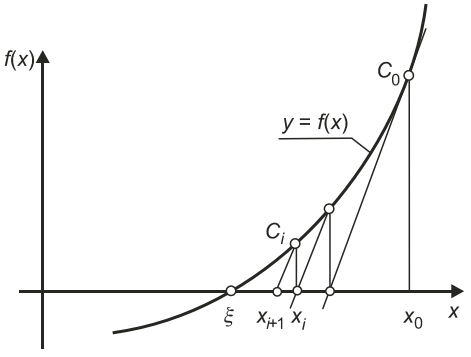
Określenie wartości ε tak, aby x_i można było uważać za dobre przybliżenie pierwiastka ξ, jest problemem, który należy rozwiązać indywidualnie w każdym zadaniu.

Iteracyjną formułę metody Newtona (2.22) można zmodyfikować. Jeśli pochodna $f'(x)$ zmienia się w przedziale domkniętym $\langle a, b\rangle$ nieznacznie, to we wzorze (2.22) można

przyjąć $f'(x_i) = f'(x_0)$. Kolejne przybliżenia pierwiastka ξ równania $f(x) = 0$ można więc obliczać według wzoru

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_0)}, \quad i = 0, 1, 2, \ldots . \tag{2.25}$$

W interpretacji geometrycznej wzór (2.25) oznacza zamianę stycznych w punktach $C_i(x_i, f(x_i))$ na proste równoległe do stycznej, poprowadzonej przez punkt $C_0(x_0, f(x_0))$, leżący na krzywej $y = f(x)$ (rys. 2.15). Taka modyfikacja metody Newtona jest szczególnie użyteczna w wypadku, gdy pochodna $f'(x)$ jest złożona. Można udowodnić, że jeśli pochodne $f'(x)$ i $f''(x)$ nie zmieniają znaków, to proces iteracyjny opisany wzorem (2.25) jest zbieżny.



Rys. 2.15. Graficzna interpretacja metody Newtona ze stałą styczną

W wielu problemach występujących w hydrotechnice określenie analitycznej postaci $f'(x)$ może być kłopotliwe. W takich wypadkach można zastąpić dokładną wartość pochodnej jej przybliżeniem przez aproksymację ilorazem różnicowym wynikającym z rozwinięcia w szereg Taylora

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}. \tag{2.26}$$

Wówczas wzór (2.22) przyjmie postać

$$x_{i+1} = x_i - \frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} \cdot f(x_i). \tag{2.27}$$

Postać tego wzoru jest podobna do (2.6). Różnica polega na tym, że we wzorze (2.6) jeden z końców przedziału jest stały.

■ *Przykład 2.4*

Obliczanie głębokości krytycznej w kanale metodą Newtona

Ustalony przepływ w kanale otwartym, zależnie od istniejących warunków, może mieć charakter przepływu nadkrytycznego, krytycznego lub podkrytycznego. Kryterium klasyfikacji stanowi głębokość, z jaką odbywa się przepływ. Głębokość krytyczna spełnia równanie (Czetwertyński i Utrysko, 1969):

$$\frac{\alpha Q^2}{g} = \frac{A^3}{b}, \tag{2.4.1}$$

gdzie: Q – natężenie przepływu,
α – współczynnik de Saint-Venanta,
g – przyspieszenie ziemskie,
A – powierzchnia przekroju czynnego,
b – szerokość zwierciadła wody.

Równanie to jest ważne dla dowolnego kształtu przekroju poprzecznego. Przyjęto tu, że przekrój jest trapezem o nachyleniu skarp 1: *M* i szerokości dna *B*. Powierzchnię przekroju czynnego można w tym wypadku wyrazić wzorem (2.1.3), zaś szerokość zwierciadła wody opisuje wzór

$$b = B + 2\,H\,M. \tag{2.4.2}$$

Jeżeli zatem znamy: α, Q, g, B, M, to głębokość krytyczną można wyznaczyć jako pierwiastek równania

$$f(H) = (B + 2HM) - g\,\frac{(BH + H^2M)^3}{\alpha Q^2}. \tag{2.4.3}$$

Do rozwiązania powyższego równania zastosujemy metodę Newtona. Różniczkując równanie (2.4.3) otrzymujemy wyrażenie definiujące pochodną funkcji $f(H)$.

$$f'(H) = 2M - \frac{g}{\alpha Q^2}\cdot 3(BH + H^2M)^2(B + 2HM). \tag{2.4.4}$$

Podobnie jak w przykładzie (2.1), w algorytmie rozwiązania można wyróżnić dwa etapy:
— lokalizację pierwiastka w przedziale o zadanej szerokości ΔH,
— obliczenie pierwiastka w przedziale z żądaną dokładnością.
Po znalezieniu przedziału zawierającego pierwiastek pierwszym przybliżeniem jest odcięta punktu leżącego w jego środku, czyli $H = (H_1 + H_2)/2{,}0$.

Dla kanału o szerokości B = 6,0 m i nachyleniu skarp M = 1,5 oraz dla wartości współczynnika α = 1,1 wykonano obliczenia dla natężeń przepływu Q zmieniających się od 20 do 50 m^3/s. Wyniki obliczeń zestawiono w tabeli 2.4.1.

Tabela 2.4.1

Wyniki obliczeń głębokości krytycznej w kanale trapezowym metodą Newtona

Q [m^3/s]	H [m]	Liczba iteracji
20	0,987	4
25	1,131	3
30	1,262	2
40	1,497	4
50	1,708	2

Uzyskano je dla ΔH = 0,50 m oraz ε = 0,001 m.

2.6. Metody hybrydowe

Omówione wcześniej metody są podstawowymi sposobami numerycznego rozwiązywania równań nieliniowych.

Jak to już pokazano wcześniej, nawet gdy pierwiastek występuje w wyizolowanym podprzedziale, funkcja może przebiegać różnie – często tak, że bezpośrednie zastosowanie metody jest kłopotliwe lub wręcz niemożliwe. Z tego względu opracowano szereg metod będących ulepszeniami metod podstawowych, najczęściej poprzez połączenie dwóch podstawowych w jedną wykorzystującą ich zalety. Dalej krótko zostaną omówione te, które nie

wymagają znajomości analitycznej postaci pochodnej funkcji. Należą do nich: metoda wynikająca z połączenia metod „regula falsi" oraz bisekcji, zwana metodą Riddersa, metoda Wegsteina, która wynika z połączenia metod iteracji prostej oraz siecznych, a także metoda Steffensena, która jest modyfikacją metody siecznych.

2.6.1. Metoda Riddersa

Jest to jeden z wariantów metody „regula falsi", połączonej z metodą bisekcji (Press, Teukolsky, Vetterling i Flannery; 1992). Załóżmy, że pierwiastek równania nieliniowego postaci (2.1) $f(x) = 0$ został zlokalizowany w przedziale $\langle x_1, x_2 \rangle$. Kolejność postępowania w metodzie Riddersa jest następująca. W każdym kroku iteracyjnym:

1. Określa się wartość funkcji $f(x)$ w punkcie środkowym przedziału $x_3 = (x_1 + x_2)/2$;
2. Zakłada się, że dobrym przybliżeniem $f(x)$ dającym małą odchyłkę funkcji w jej miejscu zerowym jest funkcja wykładnicza zbudowana na wartościach $f(x)$ w znanych punktach x_1, x_2, x_3, o postaci

$$f(x_1) - 2f(x_3)e^{\alpha} + f(x_2)e^{2\alpha} = 0\,. \tag{2.27}$$

Jest to równanie kwadratowe względem zmiennej, którą jest czynnik e^{α}. Jego rozwiązanie ma postać

$$e^{\alpha} = \frac{f(x_3) + \operatorname{sign}(f(x_2))\sqrt{f(x_3)^2 - f(x_1)f(x_2)}}{f(x_2)}, \tag{2.28}$$

3. Następnym krokiem jest zastosowanie metody „regula falsi", lecz nie do wartości $f(x_1)$, $f(x_3)$, $f(x_2)$, ale do wartości $f(x_1)$, $f(x_3)\,e^{\alpha}$, $f(x_2)\,e^{2\alpha}$, która umożliwia wyznaczenie nowego przybliżenia x_4 pierwiastka równania. Uwzględniając rozwiązanie (2.28), otrzymujemy ostatecznie

$$x_4 = x_3 + (x_2 - x_1)\frac{\operatorname{sign}(f(x_1) - f(x_2))f(x_3)}{\sqrt{f(x_3)^2 - f(x_1)f(x_2)}}\,. \tag{2.29}$$

4. Kolejnym działaniem jest sprawdzenie znaków funkcji w punktach x_1 do x_4, co ostatecznie umożliwia określenie sposobu zawężenia przedziału występowania pierwiastka. I tak:
   - jeżeli $f(x_3) \cdot f(x_4) < 0$, to $x_1 = x_3$, $x_2 = x_4$;
   - jeżeli natomiast $f(x_3) \cdot f(x_4) > 0$, trzeba jeszcze dodatkowo sprawdzić, czy $f(x_2) \cdot f(x_4) > 0$. Gdy ostatni warunek jest spełniony, to $x_2 = x_4$, w przeciwnym razie $x_1 = x_4$.

Równanie (2.29) ma ciekawe własności. Po pierwsze gwarantuje ono, że przybliżenie x_4 zawsze znajduje się wewnątrz przedziału $\langle x_1, x_2 \rangle$, po drugie – tempo zbieżności metody jest wyższe niż metody bisekcji. Przyspieszenie zbieżności uzyskuje się przez wykorzystanie współczynników wykładniczych w (2.27). Ponadto w algorytmie uwzględniamy tylko wartości funkcji. Nie jest potrzebna znajomość jej pochodnej.

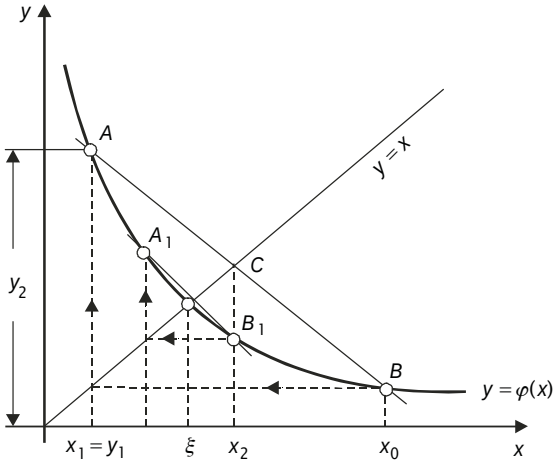
Testem zakończenia procesu iteracyjnego może być sprawdzenie warunku $|x_2 - x_1| \le \varepsilon$, gdzie ε jest przyjętą dokładnością rozwiązania.

2.6.2. Metoda Wegsteina

Metoda ta dobrze nadaje się do równań (2.1) przekształconych do postaci:

$$x = \varphi(x). \tag{2.30}$$

Sama metoda jest połączeniem opisanych wcześniej metod iteracji prostej oraz siecznych. Rozpatrzmy graficzną interpretację metody Wegsteina zastosowanej do równania (2.30) dla przypadku, gdy $-1 < \varphi'(x) < 0$ (rys. 2.16).



Rys. 2.16. Graficzna ilustracja metody Wegsteina, gdy $-1 < \varphi'(x) < 0$ (x_0 – początkowe przybliżenie, ξ – pierwiastek równania)

Startując z przybliżenia początkowego x_0, któremu odpowiada punkt B na krzywej $y = \varphi(x)$ o współrzędnych $(x_0, \varphi(x_0))$, wyznaczamy punkt A o współrzędnych $(x_1, \varphi(x_1))$. Kolejnym krokiem jest poprowadzenie siecznej przez punkty A oraz B i wyznaczenie współrzędnej punktu jej przecięcia z prostą $y = x$. Otrzymujemy ją z zależności:

$$\frac{y_2 - y_1}{x_0 - x_1} = \frac{y_2 - x_2}{x_2 - x_1}. \tag{2.31}$$

gdzie: $y_1 = \varphi(x_0)$, $x_1 = y_1$, $y_2 = \varphi(x_1)$.

Współrzędną tę wyliczamy z przekształcenia (2.31):

$$x_2 = \frac{y_2 x_0 - y_1 x_1}{y_2 - y_1 - x_1 + x_0}. \tag{2.32}$$

i traktujmy jako następne przybliżenie poszukiwanego pierwiastka.
Dla $i + 1$ przybliżenia pierwiastka, ogólny wzór rekurencyjny ma postać:

$$x_{i+1} = \frac{\varphi(x_i)x_{i-1} - \varphi(x_{i-1})x_i}{\varphi(x_i) - \varphi(x_{i-1}) - x_i + x_{i-1}}, \tag{2.33}$$

gdzie: $i = 1, 2, \ldots$, przy czym $x_1 = \varphi(x_0)$.

Wzór (2.33) można przekształcić, dodając i odejmując od licznika wyrażenie $\varphi^2(x_i) - \varphi(x_i)\,\varphi(x_{i-1}) - \varphi(x_i)\,x_i$. Otrzymujemy wtedy korzystniejszą w obliczeniach komputerowych postać:

$$x_{i+1} = \varphi(x_i) - \frac{(\varphi(x_i) - \varphi(x_{i-1}))(\varphi(x_i) - x_i)}{\varphi(x_i) - \varphi(x_{i-1}) - x_i + x_{i-1}} \tag{2.34}$$

Zalety tej wersji metody, to własność samokorekcji charakteryzująca metodę iteracji prostej oraz bardzo prosty algorytm.

2.6.3. Metoda Steffensena

Jest to zmodyfikowana metoda siecznych, w której używa się dodatkowo tylko jednej nowej wartości funkcji w każdym kroku iteracyjnym (Dahlquist i Bjorck, 1983). Efektem modyfikacji jest wzrost tempa zbieżności metody. Dla równań postaci (2.10) metoda określona jest wzorem rekurencyjnym

$$x_{i+1} = x_i - \frac{f(x_i)}{g(x_i)} \quad (i = 0, 1, 2, \ldots), \tag{2.35}$$

gdzie:

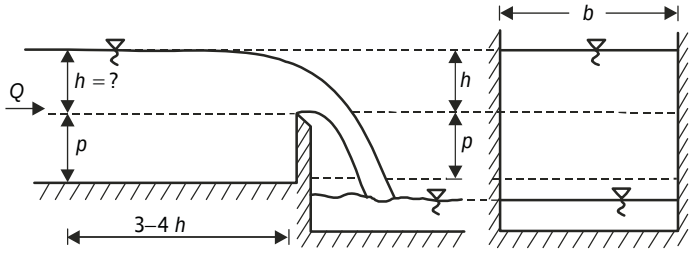
$$g(x_i) = \frac{f(x_i + f(x_i)) - f(x_i)}{f(x_i)}.$$

Metoda ta, ściśle związana z metodą siecznych, jest szczególnie użyteczna dla układów równań nieliniowych.

■ *Przykład 2.5*

Obliczenie obciążenia przelewu o ostrej krawędzi

Kanałem prostokątnym o szerokości b, wyposażonym na całej szerokości w przelew o ostrej krawędzi i wysokości p jak na rys. 2.5.1, płynie woda z natężeniem Q. Należy obliczyć obciążenie hydrauliczne h przelewu, uwzględniając prędkość wody dopływającej.



Rys. 2.5.1. Przelew o ostrej krawędzi

Jeśli oznaczyć przez v_0 średnią prędkość dopływu, wzór określający natężenie przepływu przez przelew prostokątny ma postać

$$Q = \frac{2}{3}\mu b\sqrt{2g}\left[\left(h + \frac{v_0^2}{2g}\right)^{3/2} - \left(\frac{v_0^2}{2g}\right)^{3/2}\right], \tag{2.5.1}$$

gdzie: h – obciążenie przelewu,
b – szerokość przelewu,
g – przyspieszenie ziemskie,
μ – współczynnik wydatku.

Prędkość wody dopływającej określa wzór

$$v_0 = \frac{Q}{b(p+h)} \tag{2.5.2}$$

Współczynnik wydatku dla przelewu prostokątnego o ostrej krawędzi, bez bocznego zwężenia, dany jest wzorem

$$\mu = 0{,}615\left(1 + \frac{1}{1000h + 1{,}6}\right)\left[1 + 0{,}5\left(\frac{h}{h+p}\right)^2\right], \tag{2.5.3}$$

ważnym gdy $p \geq 0{,}3$ m; $0{,}025 \leq h \leq 0{,}8$ m oraz $h/p \leq 1$.
Równanie (2.5.1), z zależnościami (2.5.2) i (2.5.3), jest równaniem nieliniowym względem h o postaci $f(h) = 0$, przy czym:

$$f(h) = Q - \frac{2}{3}\mu b\sqrt{2g}\left[\left(h + \frac{v_0^2}{2g}\right)^{3/2} - \left(\frac{v_0^2}{2g}\right)^{3/2}\right]. \tag{2.5.4}$$

Obliczenia wykonano metodami niewymagającymi znajomości pochodnej funkcji $f(h)$, tj. bisekcji, siecznych, Riddersa, Wegsteina, iteracji prostej oraz Steffensena, dla następujących danych:

— natężenie przepływu $Q = 7{,}20$ m^3/s,
— szerokość kanału $b = 5{,}0$ m,
— wysokość przelewu $p = 1{,}0$ m.

W przypadku metody iteracji prostej oraz Wegsteina, równanie (2.5.4) przekształcono do postaci $h = \varphi(h)$ poprzez dodanie i odjęcie h od prawej strony (2.5.4). Nowa funkcja będzie wtedy dana relacją:

$$\varphi(h) = h + Q - \frac{2}{3}\mu b\sqrt{2g}\left[\left(h + \frac{v_0^2}{2g}\right)^{3/2} - \left(\frac{v_0^2}{2g}\right)^{3/2}\right]. \tag{2.5.5}$$

Tak jak w poprzednich przykładach, pierwszym etapem rozwiązania była lokalizacja pierwiastka w przedziale o zadanej długości $\Delta h = 2$ m. Wyniki obliczeń w kolejnych iteracjach, dla dokładności $\varepsilon = 0{,}00001$, przedstawiono w tabeli 2.5.1.

Jak należało oczekiwać, zbieżność metody bisekcji jest najwolniejsza. Najefektywniejsza okazała się metoda Riddersa. Zauważmy, że pomimo zastosowania najprostszego przekształcenia wyjściowej funkcji (2.5.4) do postaci (2.5.5), metoda Wegsteina zakończyła się sukcesem w odróżnieniu od metody iteracji prostej, która – jak pokazano w tabeli 2.5.1, daje rozbieżny ciąg kolejnych przybliżeń poszukiwanego pierwiastka.

Tabela 2.5.1

Zestawienie wyników iteracyjnego rozwiązania równania (2.5.4) różnymi metodami

h w kolejnych iteracjach obliczone metodą:					
bisekcji	siecznych	Riddersa	Wegsteina	iteracji prostej	Steffensena
1,000010	0,425214	0,771840	0,000010	1,00001	0,464892
0,500010	1,047816	0,789766	0,042470	$-1{,}02749 \cdot 10^{1}$	0,677061
0,750010	0,744417	0,789799	0,083670	$2{,}33255 \cdot 10^{3}$	0,764307
0,875010	0,784867	0,789799	1,409764	$-1{,}24394 \cdot 10^{8}$	0,789733
0,812510	0,789904		0,582819	$3{,}57022 \cdot 10^{17}$	0,789799
0,781260	0,789798		0,739922		
0,796885	0,789799		0,795161		
0,789072			0,789673		
0,792979			0,789798		
0,791026			0,789799		
0,790049					
0,789561					
0,789805					
0,789683					
0,789744					
0,789774					
0,789790					
0,789797					

■

2.7. Rozwiązywanie układów równań nieliniowych

Układy równań nieliniowych mogą nie mieć rozwiązań, mogą mieć jedno, skończoną ich liczbę lub nieskończenie wiele rozwiązań. Warunki istnienia i jednoznaczności rozwiązania, ściśle związane z konkretnymi przypadkami, powinny być badane indywidualnie. Wskazówki ogólne są przeważnie dość trudne do wykorzystania praktycznego.

Wszystkie metody numeryczne rozwiązywania układów równań nieliniowych są metodami iteracyjnymi. Zależnie od przyjętego sposobu opisu procesu iteracyjnego nadaje się im różne nazwy.

Układ równań nieliniowych przedstawia się w ogólnej postaci wektorowej

$$\mathbf{F}(\mathbf{X}) = \mathbf{0}, \tag{2.36}$$

gdzie: $\mathbf{X} = (x_1, x_2, x_3, ..., x_n)^T$,
$\mathbf{F}(\mathbf{X}) = (f_1(\mathbf{X}), f_2(\mathbf{X}), ..., f_n(\mathbf{X}))^T$,
T – symbol transpozycji,
n – wymiar układu.

Po rozpisaniu ma on postać:

$$\begin{aligned} f_1(x_1, x_2, ..., x_n) &= 0, \\ f_2(x_1, x_2, ..., x_n) &= 0, \\ &\vdots \\ f_n(x_1, x_2, ..., x_n) &= 0. \end{aligned} \tag{2.37}$$

Rozwiązanie układu równań nieliniowych polega na znalezieniu wektora **X** spełniającego ten układ.

2.7.1. Metoda iteracji prostej (Picarda)

Wyznaczając kolejne niewiadome z równań układu (2.37), przekształcamy go do postaci:

$$\begin{aligned} x_1 &= g_1 (x_1, x_2, ..., x_n), \\ x_2 &= g_2 (x_1, x_2, ..., x_n), \\ &\vdots \\ x_n &= g_n(x_1, x_2, ..., x_n), \end{aligned} \tag{2.38}$$

co można zapisać krócej w postaci macierzowej

$$\mathbf{X} = \mathbf{G}(\mathbf{X}). \tag{2.39}$$

Mając początkową wartość wektora $\mathbf{X}^{(0)}$, tworzymy ciąg określony zależnością

$$\mathbf{X}^{(k+1)} = \mathbf{G}(\mathbf{X}^{(k)}), \tag{2.40}$$

gdzie: k – indeks iteracji.

Jeśli powyższy ciąg dąży do granicy **X**, gdy $k \to \infty$, to **X** jest rozwiązaniem układu (2.36). Ogólne warunki zbieżności tego ciągu przybliżeń są mało przydatne w konkretnych sytuacjach. Skuteczność metody należy określić na podstawie eksperymentów numerycznych. Z doświadczenia wynika, że metoda iteracji prostej jest metodą wolnozbieżną.

2.7.2. Metoda Newtona

Wprowadźmy następujące oznaczenia:

$\mathbf{X}^{(k)}$ – przybliżenie rozwiązania układu (2.36) w iteracji k-tej,

$\mathbf{\Delta}^{(k)}$ – wektor odchyłek-różnic w iteracji k-tej pomiędzy rozwiązaniem dokładnym i przybliżonym układu (2.36).

Pomiędzy tymi wektorami a rozwiązaniem dokładnym układu (2.36) zachodzi związek

$$\mathbf{X} = \mathbf{X}^{(k)} + \mathbf{\Delta}^{(k)}. \tag{2.41}$$

Zakładamy, że funkcje f_i tworzące układ są określone, ciągłe i różniczkowalne względem wektora **X**. Ponieważ z definicji wiadomo, że $\mathbf{F}(\mathbf{X}) = 0$, można napisać

$$\mathbf{F}(\mathbf{X}^k + \mathbf{\Delta}^{(k)}) \approx \mathbf{F}(\mathbf{X}^{\{k\}}) + \frac{\partial \mathbf{F}(\mathbf{X}^{(k)})}{\partial \mathbf{X}} \mathbf{\Delta}^{(k)} \approx 0\,. \tag{2.42}$$

Powyższa formuła zapisu wynika z rozwinięcia funkcji **F** w szereg Taylora wokół punktu $\mathbf{X}^{(k)}$, z zaniedbaniem członów szeregu z pochodnymi rzędu wyższego niż 1. Wektor odchyłek określamy następująco:

$$\boldsymbol{\Delta}^{(k)} = \mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}, \tag{2.43}$$

gdyż przybliżeniem $\mathbf{X}$ będzie $\mathbf{X}^{(k+1)}$. Zatem można napisać, że

$$\frac{\partial \mathbf{F}(\mathbf{X}^{(k)})}{\partial \mathbf{X}}(\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}) = -\mathbf{F}(\mathbf{X}^{(k)}), \tag{2.44}$$

gdzie:

$$\frac{\partial \mathbf{F}(\mathbf{X}^{(k)})}{\partial \mathbf{X}} = \mathbf{J}^{(k)} = \begin{bmatrix} \dfrac{\partial f_1\left(x_1^{(k)}, x_2^{(k)}, \ldots x_n^{(k)}\right)}{\partial x_1} & \cdots & \dfrac{\partial f_1\left(x_1^{(k)}, x_2^{(k)}, \ldots x_n^{(k)}\right)}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f_n\left(x_1^{(k)}, x_2^{(k)}, \ldots x_n^{(k)}\right)}{\partial x_1} & \cdots & \dfrac{\partial f_n\left(x_1^{(k)}, x_2^{(k)}, \ldots x_n^{(k)}\right)}{\partial x_n} \end{bmatrix}$$

jest jakobianem układu (2.36).
Metodę Newtona można ostatecznie zapisać w postaci

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - (\mathbf{J}^{(k)})^{-1}\, \mathbf{F}(\mathbf{X}^{(k)}). \tag{2.45}$$

Jak widać, metoda ta jest zarówno pracochłonna, jak i wymaga znacznej pamięci komputera. W każdej iteracji należy obliczyć jakobian układu o wymiarach $n \times n$, a następnie dokonać jego odwrócenia. W obliczeniach praktycznych zwykle metodę Newtona zapisuje się w poprzedniej postaci, tzn. (2.44)

$$\mathbf{J}^{(k)}\, \boldsymbol{\Delta}^{(k)} = -\,\mathbf{F}^{(k)}. \tag{2.46}$$

Zamiast odwracania macierzy $\mathbf{J}$, rozwiązujemy układ równań liniowych względem wektora odchyłek, skąd następnie obliczamy nowe przybliżenie wektora niewiadomych:

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + \boldsymbol{\Delta}^{(k)}. \tag{2.47}$$

Ten sposób jest bardziej ekonomiczny, szczególnie wtedy, gdy $\mathbf{J}$ jest macierzą pasmową.

Innym zabiegiem stosowanym w celu poprawienia ekonomiczności metody Newtona jest stosowanie stałego jakobianu. Zamiast obliczać jakobian w każdej iteracji, wykorzystuje się jego wartość z iteracji $k = 0$. Będzie więc

$$\mathbf{J}^{(0)}\, \boldsymbol{\Delta}^{(k)} = -\,\mathbf{F}^{(k)}. \tag{2.48}$$

Tak określony proces jest wolniej zbieżny, lecz nie wymaga ciągłego obliczania $\mathbf{J}$, co jest kłopotliwe szczególnie wtedy, gdy pochodne trzeba liczyć numerycznie.

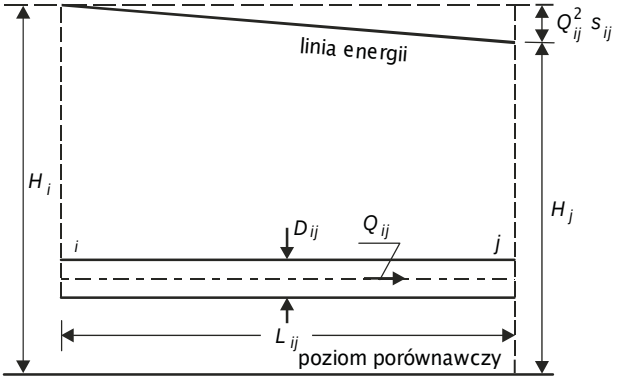
■ *Przykład 2.6*

Obliczenie przepływów w sieci wodociągowej

Jednym z typowych zagadnień inżynierii sanitarnej jest zagadnienie przepływu ustalonego wody w sieci wodociągowej. Pod pojęciem sieci należy rozumieć układ pojedynczych przewodów połączonych w węzłach. W węźle zbiegają się co najmniej trzy przewody lub dwa z lokalnym dopływem/odpływem wody. Sieć może być zamknięta, tworząc szereg pierścieni. Rozwiązanie sieci polega na obliczeniu natężenia przepływu w jej przewodach oraz ciśnień w jej węzłach (Mitosek, 2001). W tym celu wykorzystuje się dwa rodzaje równań. Mianowicie dla każdej gałęzi sieci można napisać uproszczone równanie zachowania energii mechanicznej (Bernoulliego) (rys. 2.6.1):

$$H_i - H_j = Q_{ij}^2 s_{ij}\,, \tag{2.6.1}$$

gdzie: i – indeks węzła początkowego,
j – indeks węzła końcowego,
H_i – rzędna linii ciśnienia piezometrycznego w węźle i,
H_j – rzędna linii ciśnienia piezometrycznego w węźle j,
Q_{ij} – natężenie przepływu w gałęzi,
s_{ij} – opór hydrauliczny przewodu rozumiany jako strata energii w przewodzie przy przepływie jednostkowym.



Rys. 2.6.1. Schemat odcinka sieci wodociągowej

Opór hydrauliczny określa zależność

$$s_{ij} = \frac{L_{ij}}{K_{ij}^2}\,, \tag{2.6.2}$$

gdzie: L_{ij} – długość przewodu,
K_{ij} – charakterystyka przewodu definiowana jako

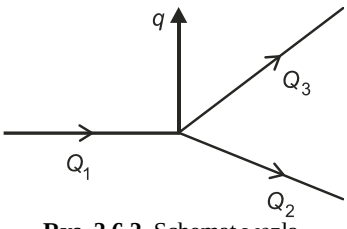
$$K_{ij} = \frac{1}{n} R_{ij}^{2/3} A_{ij}\,, \tag{2.6.3}$$

gdzie: n – współczynnik szorstkości według Manninga,
A – powierzchnia przekroju poprzecznego przewodu,
R – promień hydrauliczny przewodu.

Dla dowolnej sieci można napisać tyle równań o postaci (2.6.1), ile gałęzi zawiera sieć.

Do drugiego typu równań, które stosuje się do obliczeń przepływu w sieci, należy równanie ciągłości. Dla każdego węzła, w którym łączą się przewody lub odprowadzana jest woda (rys. 2.6.2) można napisać równanie zachowania masy, które przy stałej gęstości ma postać bilansu objętości:

$$\sum_{i=1}^{N} Q_i = q \tag{2.6.4}$$



Rys. 2.6.2. Schemat węzła, w którym łączą się 3 gałęzie

gdzie: Q_i – dopływ/odpływ do/z węzła gałęzią,
q – dopływ/odpływ zewnętrzny (pobór wody),
N – liczba gałęzi łączących się w węźle.

Zapisując powyższe równanie, należy brać pod uwagę kierunek przepływu wody w przewodach. Zwykle przyjmuje się następującą konwencję: dopływy do węzła traktuje się jako dodatnie, zaś odpływy mają znak ujemny.

Załóżmy, że określamy przepływ w sieci o strukturze przedstawionej na rysunku (2.6.3). Jest to sieć wielopierścieniowa, zasilana z jednego zbiornika, o znanej wartości ciśnienia piezometrycznego. Sieć tę tworzy 26 gałęzi o danych średnicach, długościach i charakterystykach hydraulicznych. Gałęzie łączą się w 17 węzłach, w których dodatkowo istnieją odpływy q. Można więc dla niej napisać 26 równań typu (2.6.1) oraz 16 równań typu (2.6.4), co daje łącznie 42 równania. W sieci należy określić:

— 17 wartości ciśnień w węzłach,

— 26 wartości przepływów w gałęziach.

W węźle 1 reprezentującym źródło zasilania znane jest ciśnienie, a zatem do układu można dodać równanie, które domknie go. Będziemy mieli zamknięty układ o wymiarze 43×43.

Ponieważ równania (2.6.1) są nieliniowe względem natężeń przepływów, układ ten jest układem nieliniowych równań algebraicznych. Można zapisać go następująco:

$$\mathbf{A}\mathbf{X} = \mathbf{B} \tag{2.6.5}$$

gdzie: $\mathbf{A}$ – macierz współczynników,
$\mathbf{B}$ – wektor wyrazów wolnych,
$\mathbf{X}$ – wektor niewiadomych.

Wektor niewiadomych $\mathbf{X}$ utworzą wszystkie wartości węzłowe ciśnień oraz przepływy w gałęziach. Ma on więc następującą strukturę:

$$\mathbf{X} = (H_1, H_2, ..., H_{17}, Q_{1,2}, Q_{2,3}, ..., Q_{13,\,17})^T$$

Macierz układu $\mathbf{A}$ jest macierzą pasmową bardzo rzadką. Tylko niewielka część jej elementów ma wartości różne od zera. Niektóre z nich zależą od niewiadomych, co powoduje nieliniowość układu. Wektor wyrazów wolnych $\mathbf{B} = (b_1, b_2, ..., b_{43})^T$ ma następujące elementy:

$b_1 = H_z$ (H_z – zadane ciśnienie w węźle 1),
$b_2, b_3, ..., b_{27} = 0$,
$b_{28} = q_2$,
$b_{29} = q_3$,
...
$b_{43} = q_{17}$.

Układ (2.6.5) w postaci $\mathbf{A}\mathbf{X} - \mathbf{B} = \mathbf{0}$ rozwiązujemy metodą Newtona. Zgodnie z jej opisem, algorytm iteracyjny tej metody ma postać

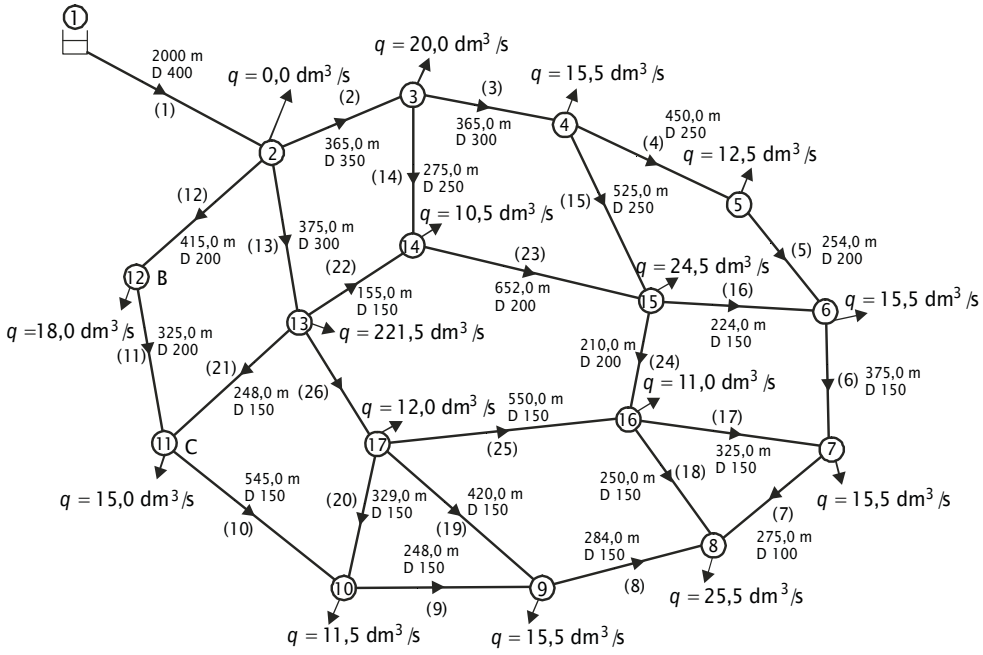
$$\mathbf{J}^{(k)}\,\Delta\mathbf{X}^{(k+1)} = -\,\mathbf{F}^{(k)} \tag{2.6.6}$$

gdzie: $\mathbf{J}$ – jakobian układu (2.6.5),
$\Delta\mathbf{X}^{(k+1)} = \mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}$ – wektor poprawek,
$\mathbf{F}^{(k)} = \mathbf{A}^{(k)}\,\mathbf{X}^{(k)} - \mathbf{B}$ – wektor residualny układu (2.6.5),
k – indeks iteracji.

Elementy macierzy Jacobiego w tym wypadku oblicza się bardzo łatwo. Kolejne równania typu (2.6.1) i (2.6.4) różniczkujemy względem kolejnych niewiadomych. Z oczywistych powodów macierz $\mathbf{J}$, podobnie jak $\mathbf{A}$, jest również macierzą pasmową, bardzo rzadką. Proces iteracyjny (2.6.6) prowadzimy do chwili uzyskania rozwiązania z postulowaną dokładnością.

Rozwiązując zadanie przepływów ustalonych w sieci rurociągów, należy wprowadzić stosowne dane, wymagane przez algorytm obliczeń. Mianowicie należy określić:

— liczbę gałęzi w sieci,
— charakterystykę każdej gałęzi, w tym:
  – indeks początkowego węzła gałęzi,
  – indeks końcowego węzła gałęzi,
  – średnicę wewnętrzną rurociągu,
  – długość rurociągu,
  – współczynnik szorstkości wg Manninga (jeśli ta formuła jest stosowana do obliczenia oporów ruchu);
— liczbę węzłów w sieci,
— charakterystykę każdego węzła, w tym:
  – indeksy przewodów tworzących węzeł (z uwzględnieniem kierunku przepływu w przewodzie),
  – pobór wody z węzła;
— liczbę źródeł zasilania oraz wartości ciśnień panujących w źródłach,
— dokładności rozwiązania ε_H oraz ε_Q stosownie dla ciśnień i przepływów.



Rys. 2.6.3. Schemat sieci pierścieniowej zasilanej z jednego źródła (Wichowski, 2002)

Powyższe dane dla sieci przedstawionej na rys. 2.6.3 zestawiono w tabelach 2.6.1 oraz 2.6.2.

Tabela 2.6.1

Charakterystyka sieci i przewodów

Nr przewodu	Nr węzła początkowego	Nr węzła końcowego	Długość [m]	Średnica [m]	Współczynnik Manninga
1	1	2	2000,0	0,400	0,012
2	2	3	365,0	0,350	0,012
3	3	4	365,0	0,300	0,012
4	4	5	450,0	0,250	0,012
5	5	6	254,0	0,200	0,012
6	6	7	375,0	0,150	0,012
7	7	8	275,0	0,100	0,012
8	9	8	284,0	0,150	0,012
9	10	9	246,0	0,150	0,012
10	11	10	545,0	0,150	0,012
11	12	11	325,0	0,200	0,012
12	2	12	415,0	0,200	0,012
13	2	13	375,0	0,300	0,012
14	3	14	275,0	0,200	0,012
15	4	15	525,0	0,250	0,012
16	15	6	224,0	0,150	0,012
17	16	7	325,0	0,150	0,012
18	16	8	250,0	0,150	0,012
19	17	9	420,0	0,150	0,012
20	17	10	320,0	0,150	0,012
21	13	11	248,0	0,150	0,012
22	13	14	156,0	0,150	0,012
23	14	15	652,0	0,200	0,012
24	15	16	210,0	0,200	0,012
25	17	16	550,0	0,150	0,012
26	13	17	245,0	0,250	0,012

Tabela 2.6.2

Rozbiory wody w węzłach

Nr węzła	Pobór wody q [l/s]	Nr węzła	Pobór wody q [l/s]
2	0,00	10	11,50
3	20,00	11	15,50
4	15,50	12	18,00
5	12,50	13	21,50
6	16,50	14	10,50
7	15,50	15	24,50
8	25,50	16	11,00
9	15,50	17	12,00

Dla przyjętych dokładności ε_H = 0,001 m oraz ε_Q = 0,01 dm^3/s rozwiązanie uzyskano po 6 iteracjach. Obliczone wartości ciśnień w węzłach oraz przepływów w gałęziach zestawiono w tabelach 2.6.3 i 2.6.4.

Tabela 2.6.3

Rzędne ciśnień piezometrycznych w węzłach

Nr węzła	Ciśnienie [m H_2O]	Nr węzła	Ciśnienie [m H_2O]
1	100,000	9	69,148
2	76,315	10	70,266
3	74,128	11	72,335
4	70,016	12	72,868
5	70,946	13	73,496
6	70,225	14	73,011
7	68,934	15	70,660
8	68,222	16	69,830
		17	72,089

Tabela 2.6.4

Przepływy w gałęziach sieci

Nr gałęzi	Przepływ [l/s]	Nr gałęzi	Przepływ [l/s]
1	245,50	14	22,64
2	122,31	15	32,75
3	79,67	16	7,26
4	31,42	17	8,66
5	18,92	18	13,23
6	9,68	19	13,80
7	2,85	20	12,45
8	9,42	21	11,29
9	11,12	22	9,20
10	10,17	23	21,34
11	14,38	24	22,33
12	32,38	25	10,57
13	90,81	26	48,83

■

3 Interpolacja i aproksymacja funkcji

3.1. Uwagi wstępne

W wielu wypadkach zachodzi konieczność zastąpienia funkcji zadanej skomplikowanym wzorem analitycznym, innym wyrażeniem, bardziej prostym. Konieczność takiego przybliżenia wynikać może np. z powodu trudności całkowania funkcji. Często też funkcja zadana jest w sposób dyskretny. Ma to np. miejsce przy wszelkich eksperymentach pomiarowych. Uzyskaną zależność dobrze jest przedstawić w postaci równania ze względu na łatwość i wygodę dalszego stosowania wyników badań. Zadaną funkcję, którą należy zastąpić inną, nazywa się funkcją przybliżaną lub aproksymowaną, natomiast funkcję, za pomocą której dokonuje się przybliżenia, nazywa się funkcją przybliżającą lub aproksymującą.

W praktyce najczęściej postać funkcji aproksymujących ogranicza się do wyrażeń będących kombinacją liniową pewnej liczby funkcji bazowych wybranych a priori. Jeśli przez

$$u_0(x), u_1(x), \ldots, u_n(x) \tag{3.1}$$

oznaczymy funkcje bazowe, to dana funkcja będzie przybliżana kombinacją typu:

$$a_0 u_0(x) + a_1 u_1(x) + \ldots + a_n u_n(x), \tag{3.2}$$

zależną od $n + 1$ współczynników a_i, które należy wyliczyć.

Podstawowe zasady aproksymacji funkcji ciągłych w przedziale domkniętym podał Weierstrass (Ralston, 1971). Są to dwa następujące twierdzenia:

1) dla dowolnej funkcji ciągłej w pewnym przedziale domkniętym istnieje zbieżny do niej w tym przedziale ciąg wielomianów;
2) dla dowolnej funkcji ciągłej o okresie 2π istnieje jednostajnie zbieżny do niej na całej osi rzeczywistej ciąg wielomianów trygonometrycznych.

Istnieje szereg możliwych do wyboru baz. Wybór ten zależy od typu zagadnienia, które jest przedmiotem rozwiązywania. Dla wielu zastosowań jako bazę wygodnie jest przyjąć np. jednomiany

$$1, x, x^2, x^3, \ldots, x^n. \tag{3.3}$$

Przyjęcie takie sprowadza zagadnienie aproksymacji do przybliżenia badanej funkcji wielomianami n-tego stopnia. Przypadek tego typu aproksymacji jest wystarczający dla omawianych w niniejszym podręczniku zagadnień, i z tego względu zostanie omówiony dokładnie. Natomiast szczegółowe wiadomości związane z doborem i stosowaniem innego rodzaju baz podaje Legras (1974).

3.2. Wielomiany interpolacyjne

Niech w przedziale $\langle a, b\rangle$ określona będzie funkcja rzeczywista $f(x)$ i niech będzie ustalonych $m + 1$ wartości argumentu x_i ($i = 0, 1, 2, 3, ..., m$), takich że

$$x_0 < x_1 < x_2 < x_3 \ldots < x_m \tag{3.4}$$

Punkty o tych odciętych nazywa się węzłami lub punktami interpolacji. Zadanie interpolacyjne polega na znalezieniu wielomianu $P(x)$ stopnia nie większego niż pewne n, który w węzłach interpolacji przyjmie takie same wartości, jak funkcja przybliżana, tzn. dla którego zachodzą równości

$$f(x_i) = P(x_i) \quad (i = 0, 1 ,2, ..., m). \tag{3.5}$$

Wielomian taki nazywany jest wielomianem interpolacyjnym przybliżającym funkcję $f(x)$. Ma on ogólną postać:

$$P(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots a_n \cdot x^n. \tag{3.6}$$

Dowód istnienia wielomianu $P(x)$ spełniającego warunek (3.5) jest nieskomplikowany. Wystarczy w tym celu rozpatrzyć wielomian (3.6) o współczynnikach a_i ($i = 0, 1, 2, ..., n$), na razie nieznanych, i podstawić do równania (3.5). Otrzymuje się w ten sposób układ $m + 1$ równań liniowych z $n + 1$ niewiadomymi $a_0, a_1, a_2, ..., a_n$ o postaci:

$$\begin{aligned} &a_0 + a_1 x_0 + a_2 x_0^2 + \ldots a_n x_0^n = f(x_1), \\ &a_0 + a_1 x_1 + a_2 x_1^2 + \ldots a_n x_1^n = f(x_2), \\ &\ldots \\ &\ldots \\ &a_0 + a_1 x_m + a_2 x_m^2 + \ldots a_n x_m^n = f(x_m). \end{aligned} \tag{3.7}$$

Wyznacznik utworzony z pierwszych k wierszy i kolumn macierzy współczynników tego układu równań, jest różny od zera, ponieważ wszystkie punkty interpolacji są różne (Ralston, 1971). Dlatego też rząd macierzy współczynników układu równań (3.7) równa się mniejszej z dwu liczb $m + 1$ i $n + 1$. Jeśli $m = n$, układ ma dokładnie jedno rozwiązanie. Zatem liczba naturalna n, dla której zadanie interpolacyjne ma rozwiązanie, powinna być równa zmniejszonej o 1 liczbie węzłów interpolacji. W celu znalezienia wielomianu wystarczy rozwiązać układ (3.7). Wielomian ten (przy $n = m$) nazywany jest wielomianem interpolacyjnym Lagrange'a.

3.3. Wielomian interpolacyjny Lagrange'a

Poszukuje się wielomianu $P(x)$ stopnia n, który przyjmuje identyczne wartości, jak pewna funkcja $f(x)$ w punktach $x_0, x_1, x_2, ..., x_n$. Poszukiwany wielomian zapisać można w postaci

$$P(x) = L_0(x)\, f(x_0) + L_1(x)\, f(x_1) + \ldots + L_i(x)\, f(x_i) + \ldots + L_n(x)\, f(x_n), \tag{3.8}$$

gdzie: $L_i(x)$ – wielomiany stopnia niższego lub równego n, które spełniają warunek

$$L_i(x_j) = \delta_{ij} = \begin{cases} 0 & \text{dla} \quad i \neq j, \\ 1 & \text{dla} \quad i = j. \end{cases} \tag{3.9}$$

Jeśli $L_i(x)$ jest wielomianem zerującym się dla $x = x_0, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n$, można go zapisać w postaci

$$L_i(x) = C_i\,[(x - x_0)\,(x - x_1)\,...\,(x - x_{i-1})\,(x - x_{i+1})\,...\,(x - x_n)]. \tag{3.10}$$

Z drugiej strony, z (3.9) wynika, że $L_i(x_i) = 1$, zatem

$$C_i = \frac{1}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}. \tag{3.11}$$

Po podstawieniu do (3.10), otrzymujemy

$$\begin{aligned} L_i(x) &= \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} = \\ &= \frac{\prod(x-x_j)}{\prod(x_i-x_j)}, \\ &\text{dla } i = 0, 1, ..., n;\quad j = 0, 1, ..., n, \text{ z wyjątkiem } i = j, \end{aligned} \tag{3.12}$$

gdzie: $\prod$ jest symbolem iloczynu.

Ostatecznie formułę interpolacyjną (3.8) otrzymujemy w postaci

$$\begin{aligned} P(x) &= \sum_{i=0}^{n} f(x_i)\,\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i+1})(x_i-x_{i+1})\ldots(x_i-x_n)} = \\ &= \frac{\displaystyle\sum_{i=0}^{n} f(x_i) \prod_{j=0,\,j\neq i}^{n}(x-x_j)}{\displaystyle\prod_{j=0,\ j\neq i}^{n}(x_i-x_j)}. \end{aligned} \tag{3.13}$$

Jest to tzw. formuła interpolacyjna Lagrange'a. Zależność tę można zapisać, stosując notację macierzową. W tym celu wzór (3.12) zapisujemy w formie:

$$L_i(x) = B_0^i + B_1^i x + B_2^i x^2 + \ldots + B_n^i x^n, \tag{3.14}$$

Wprowadzając następujące oznaczenia:

$$\mathbf{L} = \begin{bmatrix} B_0^0 & B_0^1 & \ldots & B_0^n \\ B_1^0 & B_1^1 & \ldots & B_1^n \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ B_n^0 & B_n^1 & \ldots & B_n^n \end{bmatrix}, \quad \mathbf{f} = \begin{Bmatrix} f(x_0) \\ f(x_1) \\ . \\ . \\ . \\ f(x_n) \end{Bmatrix}, \quad \mathbf{X} = \left[1, x, x^2, x^3, ...x^n\right],$$

wielomian $P(x)$ można przedstawić w postaci iloczynu:

$$P(x) = \mathbf{X}\,\mathbf{L}\,\mathbf{f}. \tag{3.15}$$

Macierz **L** nazywana jest macierzą Lagrange'a. Oznaczając dalej

$$\mathbf{L}\,\mathbf{f} = \mathbf{A} = \begin{Bmatrix} A_0 \\ A_1 \\ \cdot \\ \cdot \\ \cdot \\ A_n \end{Bmatrix},$$

można zapisać, że

$$P(x) = A_0 + A_1\,x + A_2\,x^2 + \ldots + A_n\,x^n = \mathbf{X}\,\mathbf{A}. \tag{3.16}$$

Jak widać, wielomian ten jest zdefiniowany poprzez macierz kolumnową **A**. W wypadku różniczkowania lub całkowania tego wielomianu w przedziale $\langle 0, x\rangle$ wystarczy więc zróżniczkować lub scałkować macierz wierszową **X**.

■ *Przykład 3.1*

Interpolacja wielomianami Lagrange'a

Przez 3 węzły interpolacji ($x_0 = 0$, $f(x_0) = 1$), ($x_1 = 1$, $f(x_1) = 3$), ($x_2 = 2$, $f(x_2) = 7$) poprowadzić wielomian Lagrange'a (rys. 3.1.1).

Korzystając z równania (3.12), można kolejno wyliczać:

$$L_0(x) = \frac{(x-1)(x-2)}{(0-1)(0-2)} = 1 - \frac{3}{2}x + \frac{1}{2}x^2, \tag{3.1.1}$$

$$L_1(x) = \frac{(x-0)(x-2)}{(1-0)(1-2)} = 2x - x^2, \tag{3.1.2}$$

$$L_2(x) = \frac{(x-0)(x-1)}{(2-0)(2-1)} = -\frac{1}{2}x + \frac{1}{2}x^2. \tag{3.1.3}$$

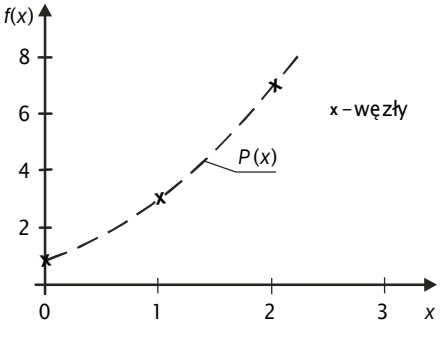
Macierz Lagrange'a przyjmie postać:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ -1{,}5 & 2 & -0{,}5 \\ 0{,}5 & -1 & 0{,}5 \end{bmatrix}, \tag{3.1.4}$$

zaś jej iloczyn przez wektor wartości węzłowych jest równy

$$\mathbf{A} = \mathbf{L}\,\mathbf{f} = \begin{bmatrix} 1 & 0 & 0 \\ -1{,}5 & 2 & -0{,}5 \\ 0{,}5 & -1 & 0{,}5 \end{bmatrix} \cdot \begin{Bmatrix} 1 \\ 3 \\ 7 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix}. \tag{3.1.5}$$

Wielomian interpolujący będzie więc następujący:



Rys. 3.1.1. Przykład obliczonego wielomianu interpolacyjnego Lagrange'a

$$P(x) = 1 + x + x^2. \tag{3.1.6}$$

Jego wykres przedstawiono na rys. 3.1.1. ■

3.4. Interpolacja funkcjami sklejanymi

Interpolację tego typu stosuje się głównie w celu połączenia punktów funkcji zadanej dyskretnie – krzywą „możliwie gładką". Przyjmuje się przy tym, że miarą „gładkości" jest wartość „krzywizny całkowitej" rozumianej jako

$$E = \int_a^b \left| f''(x) \right|^2 dx \,. \tag{3.17}$$

Krzywą „najgładszą" będzie ta krzywa, która da minimalną wartość „krzywizny całkowitej".

Funkcja sklejana jest funkcją rzeczywistą rzędu k, zdefiniowaną dla każdego $x \in (-\infty, +\infty)$, jeśli w każdym przedziale

$$(-\infty, x_1\rangle, \langle x_1, x_2\rangle, \ldots, \langle x_{n-1}, x_n\rangle, \langle x_n, +\infty),$$

gdzie $n \geq k$:

— jest ona wielomianem stopnia k,
— ma ciągłe pochodne, aż do rzędu $k - 1$ włącznie.

Funkcja sklejana jest zatem kawałkami złożona z n wielomianów stopnia k tak, że sama funkcja i jej pochodne nie mają w punktach x_i nieciągłości. W praktyce najczęściej stosuje się funkcje trzeciego stopnia, i tutaj omówiony zostanie tylko ten przypadek. Omówienie funkcji wyższych stopni można znaleźć w literaturze (Stoer, 1979).

Dany jest przedział $\langle a, b\rangle$. Przedział ten podzielony jest na podprzedziały za pomocą szeregu punktów leżących wewnątrz i spełniających warunek

$$a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b.$$

W każdym z punktów tak podzielonego przedziału znane są wartości funkcji $f(x)$, czyli:

$$f(x_0) = y_0,\ f(x_1) = y_1, \ldots,\ f(x_{n-1}) = y_{n-1},\ f(x_n) = y_n.$$

W tym wypadku funkcja sklejana, oznaczana dalej symbolem $S(x)$, będzie kawałkami złożona z n wielomianów 3. stopnia tak, że sama funkcja i jej pochodne nie będą miały w punktach x_i nieciągłości.

Istnieją dwa twierdzenia traktujące o jednoznaczności określenia i własności minimalizacji krzywizny takiej funkcji (Stoer, 1979). Mówią one, że spośród wszystkich funkcji mających ciągłe pochodne do drugiego rzędu włącznie właśnie funkcja sklejana $S(x)$, mająca dodatkowo jedną z poniższych własności:

— zerowanie się drugich pochodnych na końcach przedziału, czyli $S''(a) = S''(b) = 0$,
— okresowość,
— określone wartości pierwszej pochodnej na końcach przedziału $S'(a) = \alpha$, $S'(b) = \beta$, minimalizuje wyrażenie (3.17).

Ponieważ $|f''(x)|^2$ jest miarą krzywizny funkcji w punkcie x, wynik ten wskazuje, że pomiędzy wszystkimi funkcjami mającymi ciągłą pochodną drugiego rzędu w $\langle a, b\rangle$ funkcja sklejana $S(x)$ jest „najgładsza" w tym sensie, że daje najmniejszą z możliwych wartości krzywizny całkowitej (3.17).

Funkcję sklejaną można wyznaczyć na podstawie podanej wcześniej jej definicji. Oznaczając:

$$h_{j+1} = x_{j+1} - x_j \quad (j = 0, 1 , 2, ..., n-1),$$
$$S''(x_j) = M_j \qquad (j = 0, 1, 2, ..., n)$$

i mając na uwadze fakt, że jest to funkcja 3. stopnia, czyli że jej druga pochodna jest funkcją liniową, można napisać następującą formułę interpolacyjną:

$$S''(x) = M_j \frac{x_{j+1} - x}{h_{j+1}} + M_{j+1} \frac{x - x_j}{h_{j+1}}, \quad x \in \left\langle x_j, x_{j+1} \right\rangle . \tag{3.18}$$

Całkując, otrzymujemy dla $x \in \langle x_j, x_{j+1}\rangle$ ($j = 0, 1, 2, ..., n - 1$) wyrażenie na pierwszą pochodną

$$S'(x) = -M_j \frac{(x_{j+1} - x)^2}{2h_{j+1}} + M_{j+1} \frac{(x - x_j)^2}{2h_{j+1}} + A_j , \tag{3.19}$$

a następnie na samą funkcję

$$S(x) = M_j \frac{(x_{j+1} - x)^3}{6h_{j+1}} + M_{j+1} \frac{(x - x_j)^3}{6h_{j+1}} + A_j (x - x_j) + B_j \tag{3.20}$$

ze stałymi A_j i B_j.
Ponieważ $S(x_j) = y_j$ i $S(x_{j+1}) = y_{j+1}$, otrzymujemy dwa równania

$$\begin{aligned} M_j \frac{h_{j+1}^2}{6} + B_j &= y_j , \\ M_{j+1} \frac{h_{j+1}^2}{6} + A_j h_{j+1} + B_j &= y_{j+1}, \end{aligned} \tag{3.21}$$

z których wyznaczamy obie stałe A_j i B_j:

$$\begin{aligned} B_j &= y_j - M_j \frac{h_{j+1}^2}{6}, \\ A_j &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6}(M_{j+1} - M_j). \end{aligned} \tag{3.22}$$

Z drugiej strony, funkcję $S(x)$ można rozwinąć w otoczeniu punktu x_j w szereg Taylora. Uwzględniając w rozwinięciu wyrazy zawierające pochodne do III rzędu włącznie, otrzymujemy

$$S(x_j + \Delta x) \approx S(x_j) + \frac{\Delta x}{1!} S'(x_j) + \frac{(\Delta x)^2}{2!} S''(x_j) + \frac{(\Delta x)^3}{3!} S'''(x_j) . \tag{3.23}$$

Ponieważ w przedziale $\langle x_j, x_{j+1}\rangle$ $x = x_j + \Delta x$, stąd $\Delta x = x - x_j$, a więc

$$S(x) \approx S(x_j) + \frac{x - x_j}{1!} S'(x_j) + \frac{(x - x_j)^2}{2!} S''(x_j) + \frac{(x - x_j)^3}{3!} S'''(x_j) \tag{3.24}$$

dla $x \in \langle x_j, x_{j+1}\rangle$.

Wyrażając dalej pochodne za pomocą M_j, uzyskuje się końcową postać funkcji sklejanej

$$S(x) = y_j + \alpha_j (x - x_j) + \beta_j (x - x_j)^2 + \gamma_j (x - x_j)^3, \tag{3.25}$$

gdzie:

$$\alpha_j = S'(x_j) = -\frac{M_j h_{j+1}}{2} + A_j = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{2M_j + M_{j+1}}{6} h_{j+1},$$

$$\beta_j = \frac{S''(x_j)}{2} = \frac{M_j}{2},$$

$$\gamma_j = \frac{S'''(x_j)}{6} = \frac{M_{j+1} - M_j}{6h_{j+1}}.$$

W ten sposób funkcja $S(x)$ wyrażona jest poprzez jej pochodne II rzędu w punktach x_j. Ich wartość można wyznaczyć, wykorzystując warunek ciągłości pierwszej pochodnej $S'(x)$ (równanie (3.19)) w punktach x_j (tzn. $S'(x_{j-}) = S'(x_{j+})$). Otrzymuje się zależność

$$\frac{y_j - y_{j-1}}{h_j} + \frac{h_j}{3} M_j + \frac{h_j}{6} M_{j-1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{3} M_j - \frac{h_{j+1}}{6} M_{j+1} \tag{3.26}$$

$$(j = 1, 2, ..., n - 1),$$

którą po uporządkowaniu zapisujemy następująco:

$$\frac{h_j}{h_j + h_{j+1}} M_{j-1} + 2M_j + \frac{h_{j+1}}{h_j + h_{j+1}} M_{j+1} = \left[\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \right] \frac{6}{h_j + h_{j+1}}. \tag{3.27}$$

W ten sposób otrzymujemy $n - 1$ równań zawierających $n + 1$ niewiadomych M_0, M_1, M_2, ..., M_n. Dwa brakujące równania uzyskuje się z warunków, jakie funkcja $S(x)$ spełnia na brzegach przedziału, np.:

$$S''(a) = S''(b) = 0, \quad \text{czyli } M_0 = M_n = 0.$$

Układ ten można zapisać przejrzyściej w postaci równań

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j \quad (j = 1, 2, ..., n-1), \tag{3.28}$$

w których:

$$\lambda_j = \frac{h_{j+1}}{h_j + h_{j+1}}, \quad \mu_j = \frac{h_j}{h_j + h_{j+1}} = 1 - \lambda_j,$$

$$d_j = \frac{6}{h_j + h_{j+1}} \left[\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \right].$$

Jeśli dodatkowo wprowadzimy zależności wynikające z przyjętych warunków na brzegach: $\lambda_0 = 0$, $\mu_n = 0$, $d_0 = 0$, $d_n = 0$, to przyjmie on postać

$$\begin{bmatrix} 2 & \lambda_0 & & & & & & & \\ \mu_1 & 2 & \lambda_1 & & & & & & \\ & \mu_2 & 2 & \lambda_2 & & & & & \\ & & \mu_3 & 2 & \lambda_3 & & & & \\ & & & . & . & . & & & \\ & & & & . & . & . & & \\ & & & & & . & . & . & \\ & & & & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & & & & & \mu_n & 2 \end{bmatrix} \begin{Bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \\ . \\ . \\ . \\ M_{n-1} \\ M_n \end{Bmatrix} = \begin{Bmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \\ . \\ . \\ . \\ d_{n-1} \\ d_n \end{Bmatrix}. \qquad (3.29)$$

Przedstawiony układ równań jest układem dobrze uwarunkowanym. Widać, że $\lambda_j \geq 0$, $\mu_j \geq 0$ oraz $\lambda_j + \mu_j = 1$. Ponadto współczynniki λ_j i μ_j zależą tylko od sposobu podziału przedziału $\langle a, b \rangle$, a nie od wartości funkcji y_i w punktach podziału. Ta własność może być wykorzystana do udowodnienia twierdzenia (Stoer, 1979), że macierz współczynników układu (3.29) jest nieosobliwa dla dowolnego podziału przedziału $\langle a, b \rangle$. Układ ten można rozwiązać metodą rozkładu macierzy współczynników na dwie macierze trójkątne, co, jak wiadomo, prowadzi do bardzo efektywnego algorytmu opisanego w punkcie 1.2.3.

■ *Przykład 3.2*

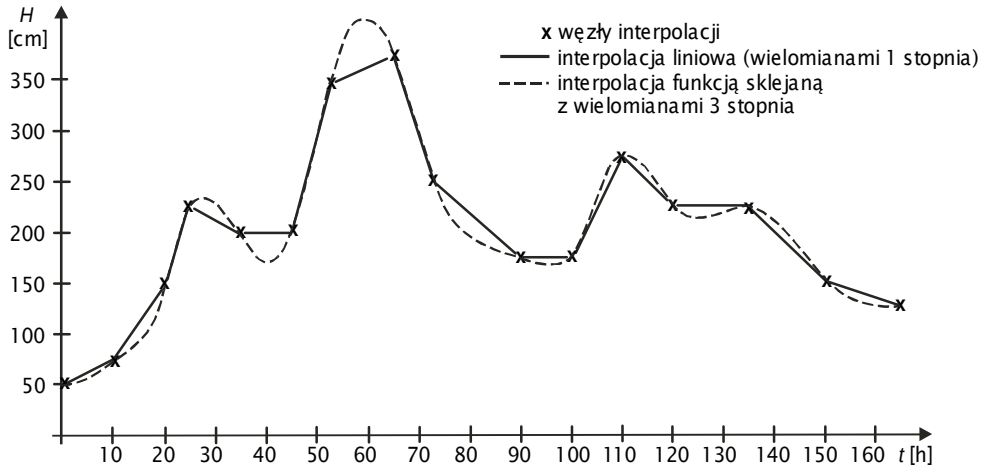
Interpolacja limnigramu za pomocą funkcji sklejanej

Limnigramem nazywa się krzywą reprezentującą przebieg zmian stanów wody w czasie, w danym przekroju rzeki. Zatem zmienną niezależną jest tutaj czas *t*, zaś zmienną zależną stan wody *H*, czyli poziom zwierciadła wody liczony od przyjętego poziomu porównawczego – zera wodowskazu. Ponieważ stany wody mierzy się w określonych terminach, limnigram jest funkcją zadaną w postaci tabelarycznej. W rozwiązywanym tutaj przykładzie jest on opisany za pomocą 16 punktów i przedstawiony w tabeli 3.2.1.

Tabela 3.2.1

Obserwowany limnigram stanów wody

Nr punktu	*t* [h]	*H* [cm]	Nr punktu	*t* [h]	*H* [cm]
1	0	50	9	73	250
2	10	75	10	90	175
3	20	150	11	100	175
4	25	225	12	110	275
5	35	200	13	120	225
6	45	200	14	135	225
7	53	350	15	150	150
8	65	375	16	165	125



Rys. 3.2.1. Wynik interpolacji limnigramu funkcją sklejaną

Tabela 3.2.2

Obliczone współczynniki funkcji sklejanej interpolującej limnigram

j	α_j	β_j	γ_j
1	1,2500	0,0000	0,0125
2	0,1956	0,3750	0,0355
3	13,9647	1,4413	–0,2469
4	9,8640	–2,2614	0,1025
5	–11,3409	0,8137	0,0320
6	13,0588	1,7749	–0,1329
7	17,8582	–1,4156	0,0084
8	–12,4793	–1,1125	0,0899
9	–14,2917	1,0454	–0,0273
10	–2,7372	–0,3473	0,0621
11	5,6966	1,5157	–0,1085
12	3,4497	–1,7404	0,0895
13	–6,3646	0,9459	–0,0348
14	–0,4183	–0,6189	0,0209
15	–4,8805	0,3214	–0,0071

Zgodnie z podanym opisem funkcji sklejanej, zadanie interpolacji polega na wyznaczeniu współczynników równania (3.25). Współczynniki te wyznacza się przez rozwiązanie układu (3.29) metodą rozkładu macierzy współczynników (algorytm Thomasa). Proces obliczeń przebiega według poniższego schematu:

1) wczytaj dane, tzn. ilość węzłów interpolacji i ich współrzędne (x_j, y_j);
2) zbuduj trójdiagonalną macierz współczynników układu (3.29), obliczając współczynniki λ_j, μ_j równań (3.28), oraz oblicz składowe wektora prawych stron d_j;

3) uzupełnij układ, wprowadzając równania wynikające z warunków zerowania się drugiej pochodnej na końcach przedziału ($M_0 = 0$, $M_n = 0$);
4) rozwiąż układ równań, stosując metodę rozkładu macierzy;
5) oblicz współczynniki funkcji sklejanej (3.25): α_j, β_j, γ_j ($j = 0, 1, 2, ..., n - 1$).

Otrzymane w wyniku obliczeń wartości współczynników równania (3.25) przedstawiono w tabeli 3.2.2). Na rysunku 3.2.1 pokazano uzyskany wynik interpolacji funkcją sklejaną. Dla porównania zaznaczono również wynik interpolacji liniowej.

■

3.5. Aproksymacja funkcji za pomocą wielomianów

Załóżmy, że $y = f(x)$ jest funkcją ciągłą w przedziale $\langle a, b\rangle$. Znalezienie przybliżenia (aproksymacji) tej funkcji polega na znalezieniu współczynników pewnego wielomianu $P(x)$, który będzie „dobrze" przybliżał w tym przedziale funkcję $f(x)$. Pojęcie dobrej aproksymacji ma sens tylko w wypadku, kiedy zdefiniowane jest kryterium charakteryzujące błąd między funkcją $f(x)$ a wielomianem $P(x)$. Zwykle stosuje się jedno z dwóch następujących kryteriów (Legras, 1974).

Kryterium Czebyszewa

Współczynniki wielomianu aproksymującego należy tak dobrać, aby maksymalna różnica między wartością funkcji $f(x)$ a wielomianem $P(x)$ osiągnęła minimum. Zatem wymaga się tutaj, aby

$$E = \max_{x\in\langle a,b\rangle} \left|\varepsilon(x)\right| \to \min, \quad \text{gdzie} \quad \varepsilon(x) = f(x) - P(x) .$$

Ten rodzaj aproksymacji nazywa się aproksymacją jednostajną lub aproksymacją Czebyszewa.

Kryterium najmniejszego błędu kwadratowego

Współczynniki wielomianu aproksymującego należy tak dobrać, aby suma kwadratów różnic między wartością funkcji $f(x)$ a wielomianem $P(x)$ w $\langle a, b\rangle$ osiągnęła minimum, czyli

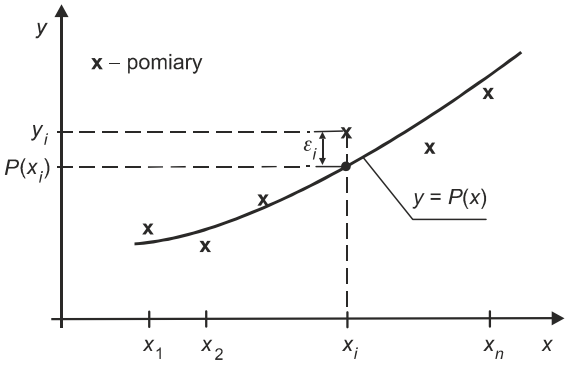
$$E = \int_a^b \varepsilon^2(x)\,dx \to \min, \quad \text{gdzie } \varepsilon(x) = f(x) - P(x).$$

Aproksymację przy zastosowaniu tego rodzaju kryterium nazywa się aproksymacją w sensie najmniejszych kwadratów lub metodą najmniejszych kwadratów.

Dalsze rozważania ograniczymy do aproksymacji w sensie najmniejszych kwadratów, w wypadku, kiedy funkcja aproksymowana $y = f(x)$ zadana jest w sposób dyskretny. Wynika to z faktu, że w inżynierii wodnej ten rodzaj aproksymacji jest stosowany najczęściej.

Pierwsze sformułowanie metody najmniejszych kwadratów podał Legendre w roku 1806. Rozwinął ją później Gauss, podając jej podstawy matematyczne, które są stosowane do dzisiaj bez większych zmian. Zakłada się, że wielkości x i y związane są ze sobą nieznaną zależnością $y = f(x)$. Jest jednak ona określona na płaszczyźnie $x - y$ w sposób dyskretny

za pomocą n punktów o współrzędnych (x_i, y_i) (i = 1, 2, 3, ..., n). Zadanie polega na znalezieniu funkcji $P(x)$, jak najlepiej przybliżającej funkcję tak określoną (rys. 3.1).



Rys. 3.1. Ilustracja problemu aproksymacji zależności $y = f(x)$ wielomianem $P(x)$

W tym celu należy wybrać wielomian $y = P(x)$ zależny od k parametrów a_1, a_2, ..., a_k i określić wartości tych parametrów tak, aby zminimalizować funkcję kryterialną $E(a_1, a_2, ..., a_k)$. Ma ona postać

$$E(a_1, a_2, \ldots, a_k) = \sum_{i=1}^{n} [y_i - P(x_i)]^2 \tag{3.31}$$

i jak widać, jest funkcją wielu zmiennych, która osiąga ekstremum w punkcie, gdzie:

$$\frac{\partial E}{\partial a_1} = \frac{\partial E}{\partial a_2} = \ldots = \frac{\partial E}{\partial a_k} = 0 .$$

Różniczkując funkcję E, uzyskuje się następujący układ równań:

$$\begin{aligned}
&\frac{\partial E}{\partial a_1} = \sum_{i=1}^{n} 2[y_i - P(x_i)]\left(-\frac{\partial P}{\partial a_1}\right) = 0, \\
&\frac{\partial E}{\partial a_2} = \sum_{i=1}^{n} 2[y_i - P(x_i)]\left(-\frac{\partial P}{\partial a_2}\right) = 0, \\
&\ldots \\
&\frac{\partial E}{\partial a_k} = \sum_{i=1}^{n} 2[y_i - P(x_i)]\left(-\frac{\partial P}{\partial a_k}\right) = 0.
\end{aligned} \tag{3.32}$$

Mając na uwadze postać funkcji kryterialnej (3.31), można wykazać, że powyższy warunek definiuje jej ekstremum minimum. Jest to układ k równań algebraicznych o k niewiadomych, który ma jednoznaczne rozwiązanie tylko w wypadku, gdy funkcja aproksymująca $P(x)$ zależy liniowo od parametrów a_1, a_2, ..., a_k. Warunek ten jest spełniony, gdyż $P(x)$ jest wielomianem. Cechą charakterystyczną układu (3.32) jest symetria macierzy współczynników. W wypadku nieliniowej zależności funkcji f od parametrów a_i układ ten jest układem równań nieliniowych, a rozwiązanie takiego układu jest, jak wiadomo, zagadnieniem złożonym. Z tego względu w zasadzie nie stosuje się do aproksymacji funkcji nieliniowych

względem ich parametrów, z wyjątkiem takich, które przez proste przekształcenie można sprowadzić do równania liniowego. W tabeli 3.1 przedstawia się przykłady tego typu przekształceń.

Tabela 3.1

Przekształcenia linearyzujące niektóre funkcje (Kacprzyński, 1974)

$y = f(x)$	Nowe zmienne sprowadzające funkcję $y = f(x)$ do postaci liniowej $z = a_1 + a_2 u$			
$y = a\,b^x$	$z = \ln y$	$a_1 = \ln a$	$a_2 = \ln b$	$u = x$
$y = \dfrac{1}{a + b\,x}$	$z = \dfrac{1}{y}$	$a_1 = a$	$a_2 = b$	$u = x$
$y = q^{a + b\,x}$	$z = \lg y$	$a_1 = a$	$a_2 = b$	$u = x$
$y = a\,x^b$	$z = \ln y$	$a_1 = \ln a$	$a_2 = b$	$u = \ln x$
$y = a + \dfrac{b}{x}$	$z = y$	$a_1 = a$	$a_2 = b$	$u = \dfrac{1}{x}$
$y = \dfrac{x}{a + b\,x}$	$z = \dfrac{1}{y}$	$a_1 = b$	$a_2 = a$	$u = \dfrac{1}{x}$

Zastosowanie metody najmniejszych kwadratów zilustrowano poniżej przykładem aproksymacji krzywej przepływu dla danego przekroju cieku.

■ *Przykład 3.3*

Aproksymacja krzywej przepływów metodą najmniejszych kwadratów

W hydraulice koryt otwartych bardzo często wykorzystuje się tzw. krzywą przepływów (Lambor, 1971). Jest to zależność pomiędzy stanem wody a natężeniem przepływu w danym przekroju rzeki. Zależność tę uzyskuje się w wyniku pomiarów, a zatem zadana jest ona w sposób tabelaryczny:

$$(H_1, Q_1),\ (H_2, Q_2),\ \ldots,\ (H_i, Q_i),\ \ldots,\ (H_n, Q_n),$$

gdzie: H_i – stan wody,
Q_i – natężenie przepływu,
n – liczba pomiarów.

Dla łatwiejszego korzystania z tej zależności poszukuje się funkcji, która w miarę dobrze opisze zależność $Q(H)$ w zakresie zmienności H. Z praktyki wiadomo, że zależność tę można skutecznie aproksymować między innymi krzywą potęgową o ogólnym równaniu (Lambor, 1971):

$$Q = a(H - h_0)^b, \tag{3.3.1}$$

przy czym bardzo często przyjmuje się, że $h_0 = 0$. W takiej sytuacji równanie upraszcza się do postaci

$$Q = a\,H^b. \tag{3.3.2}$$

Jak widać, jest to równanie nieliniowe względem współczynnika b. W celu zlinearyzowania

równania można dokonać przekształcenia polegającego na obustronnym logarytmowaniu (patrz tabela 3.1). Otrzymuje się w ten sposób równanie:

$$Z = a_1 + a_2\, u, \tag{3.3.3}$$

gdzie: $Z = \ln Q$, $a_1 = \ln a$, $a_2 = b$, $u = \ln H$.

Zgodnie z (3.31) parametry a_1 i a_2 należy tak dobrać, aby wyrażenie

$$E(a_1, a_2) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Z_i - (a_1 + a_2 u_i))^2 \tag{3.3.4}$$

osiągnęło minimum. Jak wiadomo, ma to miejsce, gdy

$$\frac{\partial E}{\partial a_1} = \sum_{i=1}^{n} [2(Z_i - a_1 - a_2 u_i)(-1)] = 0\,,$$

$$\frac{\partial E}{\partial a_2} = \sum_{i=1}^{n} [2(Z_i - a_1 - a_2 u_i)(-u_i)] = 0\,. \tag{3.3.5}$$

Zatem w celu wyznaczenia optymalnych wartości parametrów a_1 i a_2 należy rozwiązać powyższy układ równań. Układ ten można zapisać w postaci macierzowej

$$\begin{bmatrix} n & \sum_{i=1}^{n} u_i \\ \sum_{i=1}^{n} u_i & \sum_{i=1}^{n} u_i^2 \end{bmatrix} \cdot \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum_{i=1}^{n} Z_i \\ \sum_{i=1}^{n} Z_i u_i \end{Bmatrix}. \tag{3.3.6}$$

Stosując do rozwiązania tego układu wzory Cramera, otrzymuje się:

$$a_1 = \frac{W_1}{W}, \quad a_2 = \frac{W_2}{W}\,,$$

gdzie:

$$W = n \sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i \right)^2,$$

$$W_1 = \sum_{i=1}^{n} Z_i \sum_{i=1}^{n} u_i^2 - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} Z_i u_i\,,$$

$$W_2 = n \sum_{i=1}^{n} Z_i u_i - \sum_{i=1}^{n} Z_i \sum_{i=1}^{n} u_i\,.$$

Po uwzględnieniu przekształcenia linearyzującego wartości parametrów a oraz b w równaniu wyjściowym wyniosą:

$$a = \exp(W_1/W),\ b = W_2/W.$$

Zgodnie z opisanym wyżej tokiem postępowania, wykonano przykładowe obliczenie współczynników krzywej przepływów dla danych zamieszczonych w tabeli 3.3.1. Wynikiem są następujące wartości współczynników:

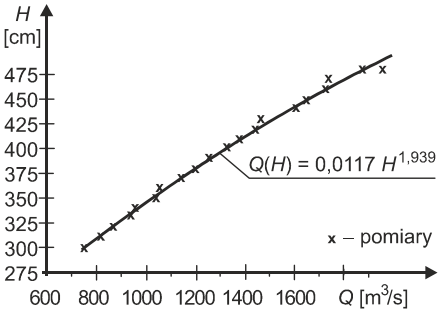
$$a = 0{,}01179, \quad b = 1{,}939.$$

Zatem funkcja (IX.2) w tym wypadku ma postać:

$$Q(H) = 0{,}01179\, H^{1{,}939}, \qquad (3.3.7)$$

przy czym H należy zadawać w cm, otrzymując wynik Q w m^3/s.

Na rysunku 3.3.1 porównano pomiary z tabeli 3.3.1 z wartościami obliczonymi wg równania (3.3.7).



Rys. 3.3.1. Wynik aproksymacji krzywej przepływów

Tabela 3.3.1

Obserwowana zależność $Q(H)$

i	H_i [cm]	Q_i [m^3/s]	i	H_i [cm]	Q_i [m^3/s]	i	H_i [cm]	Q_i [m^3/s]
1	300	740	7	360	1130	13	430	1460
2	311	810	8	380	1190	14	440	1600
3	320	840	9	390	1250	15	450	1640
4	332	930	10	402	1322	16	460	1715
5	340	950	11	411	1370	17	470	1730
6	349	1050	12	420	1430	18	480	1950

■

3.6. Porównanie interpolacji i aproksymacji

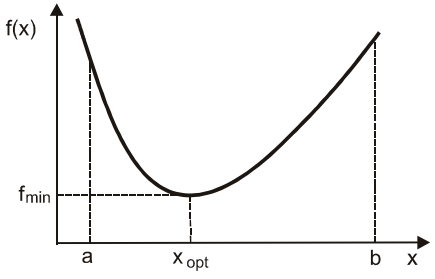
Proces interpolacji jest wrażliwy na wybór węzłów interpolacji. Ewentualne, często nieznaczne, zakłócenia wartości funkcji w punktach interpolacji mogą bardzo znacznie zmieniać funkcję interpolującą. Spośród metod interpolacji, cennymi zaletami – w stosunku do interpolacji wielomianowej – charakteryzuje się metoda interpolacji funkcjami sklejanymi. Daje ona krzywą o minimalnej krzywiźnie całkowitej w przedziale $\langle a, b\rangle$, a ponadto przy zagęszczaniu podziału $\langle a, b\rangle$ funkcje sklejane są zbieżne do funkcji, którą interpolują. Własności tej nie mają wielomiany interpolacyjne. Metody interpolacji pozwalają na przybliżenie funkcji określonych złożonymi równaniami za pomocą prostych wyrażeń. Stanowią one podstawę metod całkowania numerycznego, które będą omówione w rozdziale 4.

Z definicji zadania interpolacji wiadomo, że funkcja interpolująca musi przechodzić przez punkty (x_i, y_i), gdzie $i = 1, 2, ..., n$. Ponieważ pomiary obarczone są błędami, zwykle dla odciętej x_i otrzymuje się różne wartości rzędnej y_i. W konsekwencji interpolacja staje się niemożliwa. W wypadku, kiedy trzeba przybliżać funkcje określone eksperymentalnie, co w zasadzie jest regułą we wszelkich zagadnieniach technicznych, należy stosować metody aproksymacyjne. Aproksymacja jest mało wrażliwa na wybór węzłów aproksymacji, jeśli tylko liczba węzłów jest wystarczająco duża i są one właściwie rozmieszczone w przedziale $\langle a, b\rangle$. W metodach tych następuje „wygładzanie", co redukuje wpływ błędów wartości funkcji aproksymowanej w węzłach aproksymacji.

4 Metody poszukiwania ekstremum funkcji jednej zmiennej

4.1. Przedstawienie problemu

Bardzo często rozwiązanie problemu z zakresu inżynierii wodnej polega na znalezieniu najlepszego rozwiązanie z punktu widzenia przyjętego kryterium jakości, co wymaga stosowania technik optymalizacyjnych. W rozdziale tym rozpatrzymy tylko najprostszy przypadek, gdy kryterium jakości lub inaczej funkcja celu, jest funkcją jednej zmiennej. Innymi słowy, będziemy poszukiwali ekstremum funkcji $f(x)$ w przedziale, przy czym zwykle będzie to ekstremum minimum (rys. 4.1). Zakładamy przy tym, że funkcja ta jest ciągła.



Rys. 4.1. Funcja $f(x)$ posiadająca ekstremum minimum w punkcie $x = x_{opt}$

Jeśli rozpatrywana funkcja jest różniczkowalna, to poszukiwany punkt ekstremalny można określić wykorzystując jej pierwszą pochodną. Warunek istnienia ekstremum:

$$\frac{df}{dx} = 0 \tag{4.1}$$

umożliwia obliczenie wartość zmiennej x dla które funkcja $f(x)$ przyjmuje wartość ekstremalną. Jeśli w tym punkcie jej pochodna II rzędu jest dodatnia

$$\frac{d^2 f}{dx^2} > 0 \tag{4.2}$$

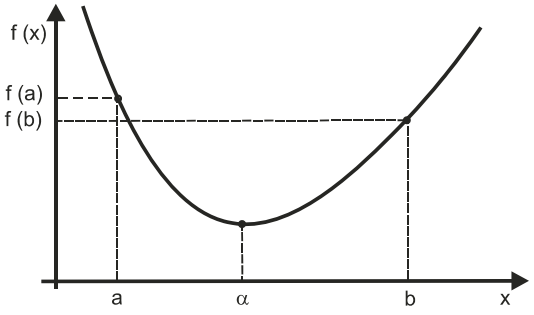
to jest to ekstremum minimum.

Jeśli funkcja $f(x)$ nie jest zadana w jawnej postaci i jej właściwości nie są znane, to najlepiej jest użyć numerycznych metod poszukiwania jej ekstremów. Załóżmy, że:

— w przedziale $\langle a, b\rangle$ funkcja $f(x)$ posiada minimum w punkcie $\alpha \in \langle a, b\rangle$;

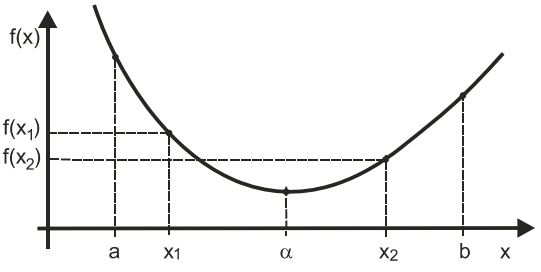
— funkcja $f(x)$ jest malejąca w przedziale $\langle a, \alpha\rangle$ i rosnąca w przedziale $\langle \alpha, b\rangle$.

Funkcję o powyższych właściwościach (rys. 4.2) nazywa się funkcją unimodalną (Kusiak, Danielewska-Tułecka, Oprochta, 2009).



Rys. 4.2. Funkcja unimodalna $f(x)$ w przedziale $\langle a, b\rangle$

W celu znalezienia mniejszego przedziału zawierającego punkt ekstremalny α, wystarczy obliczyć wartości funkcji $f(x)$ tylko w 2 punktach leżących wewnątrz przedziału $\langle a, b\rangle$.



Rys. 4.3. Wybór dwóch punktów w przedziale $\langle a, b\rangle$ umożliwiający określenie krótszego przedziału zawierającego ekstremum funkcji $f(x)$**.**

W tym celu należy wybrać dwa punkty x_1 oraz x_2 (jak na rys. 4.3) tak, aby $a < x_1 < x_2 < b$. Jeżeli

$$f(x_1) \le f(x_2) \quad \text{to} \quad \alpha \in \langle a, x_2\rangle . \tag{4.3}$$

W przypadku przeciwnym, jeżeli

$$f(x_1) > f(x_2) \quad \text{to} \quad \alpha \in \langle x_1, b\rangle . \tag{4.4}$$

W ten sposób otrzymuje się szereg podprzedziałów

$$\langle a^{(i)}, b^{(i)}\rangle \quad \text{dla} \quad i = 1, 2, 3, \ldots \tag{4.5}$$

takich, że każdy następny podprzedział zawarty jest w poprzednim

$$\langle a^{(i+1)}, b^{(i+1)}\rangle \in \langle a^{(i)}, b^{(i)}\rangle. \tag{4.6}$$

W powyższej zależności indeks i oznacza numer iteracji przy czym $a^{(0)} = a$ zaś $b^{(0)} = b$.

Jeżeli długość kolejnych podprzedziałów zmierza do zera

$$b^{(i)} - a^{(i)} \to 0 \tag{4.7}$$

to szeregi utworzone odpowiednio przez początki przedziałów $a^{(i)}$ i ich końce $b^{(i)}$ są zbieżne do punktu α. W podejściu tym położenie ekstremum minimum określane jest z pewną

dokładnością. Będziemy poszukiwali takiego punktu x, dla którego spełnione jest następujące kryterium:

$$\alpha \in \langle x - \varepsilon,\ x + \varepsilon \rangle \tag{4.8}$$

gdzie ε jest dodatnią liczbą reprezentującą założoną dokładność.

Wszystkie znane metody poszukiwanie ekstremum minimum funkcji jednej zmiennej różnią się między sobą sposobem podziału przedziału $\langle a, b \rangle$, w którym występuje punkt ekstremalny. Zanim jednak przystąpimy do ich prezentacji omówmy pewną nietypową metodę poszukiwania ekstremum funkcji $f(x)$ polegającą na rozwiązaniu algebraicznego równania nieliniowego.

4.2. Wyznaczenie ekstremum funkcji przez rozwiązanie równania nieliniowego

Rozpatrzmy funkcję unimodalną, $f(x)$, która w przedziale $\langle a, b \rangle$ posiada ekstremum. Jego położenie można wyznaczyć rozwiązując następujące równanie wynikające z dobrze znanego warunku istnienia ekstremum:

$$g(x) = \frac{df}{dx} = 0\,. \tag{4.9}$$

Innymi słowy, punkt ekstremalny wyznacza rozwiązanie równania algebraicznego. Ponieważ zwykle jest to równanie nieliniowe, do jego rozwiązania należy zastosować jedną z metod omówionych w rozdziale 2, jak metoda bisekcji, metoda siecznych, metoda iteracji prostej, lub metoda Newtona.. Znak pochodnej II rzędu określa rodzaj ekstremum.

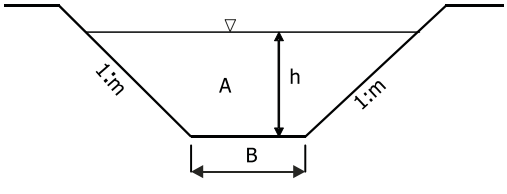
Dokładniejsze wyjaśnienie toku postępowania zawiera poniższy przykład.

■ *Przykład 4.1*

Rozpatrzmy jednostajny przepływ w kanale otwartym o trapezowym przekroju poprzeczny. Jego parametry są następujące (rys. 4.1.1):

— pole powierzchni przekroju czynnego A,
— podłużny spadek kanału s,
— współczynnik szorstkości według Manninga n,
— nachylenie skarp kanału $1 : m$.

Należy obliczyć szerokość dna kanału B (rys. 4.1.1), który zapewni maksymalne natężenie przepływu Q. Do rozwiązania zadania zastosujmy metodę siecznych.



Rys. 4.1.1. Przekrój poprzeczny rozpatrywanego kanału

Rozpocznijmy rozwiązanie zadania od sformułowania funkcji celu. Przepływ jednostajny opisuje równanie Manninga

$$Q = \frac{1}{n} R^{2/3} s^{1/2} A \,. \tag{4.1.1}$$

W równaniu (4.1.1) R jest promieniem hydraulicznym definiowanym następująco

$$R = \frac{A}{p} \tag{4.1.2}$$

gdzie p jest obwodem zwilżonym przekroju poprzecznego kanału.. Przy zadanych wartościach parametrów A, s, n oraz m z równania (4.1.2) wynika, że natężenie przepływu w kanale Q rośnie wraz ze wzrostem promienia hydraulicznego R. Zatem

$$Q \to \max \quad \text{gdy} \quad R \to \max \,. \tag{4.1.3}$$

Ponieważ równanie (4.1.2) wskazuje, że przy danej wartości pola powierzchni przekroju czynnego A promień hydrauliczny R będzie rósł dla malejących wartości obwodu zwilżonego p

$$R \to \max \quad \text{gdy} \quad p \to \min \,. \tag{4.1.4}$$

to oznacza, że należy znaleźć taką wartość szerokości dna kanału B, która zapewni minimalną wartość obwodu zwilżonego p. Dla przekroju trapezowego promień hydrauliczny jest równy

$$p = B + 2h\sqrt{1+m^2} \tag{4.1.5}$$

więc funkcję celu możemy sformułować następująco:

$$p(B) = B + 2h\sqrt{1+m^2} \quad \to \quad \min \,. \tag{4.1.6}$$

Głębokość h występującą w powyższej funkcji należy wyrazić poprzez znane parametry kanału. W tym celu wykorzystuje się wzór na pole powierzchni przekroju czynnego kanału w kształcie trapezu:

$$A = B \cdot h + m \cdot h^2 \tag{4.1.7}$$

który następnie zapisujemy w postaci:

$$m \cdot h^2 + B \cdot h - A = 0 \,. \tag{4.1.8}$$

Jest to równanie kwadratowe względem h. Jego rozwiązanie przebiega w dobrze znany sposób. Kolejno obliczamy:

$$\Delta = b^2 - 4a \cdot c = B^2 + 4m \cdot A \tag{4.1.9}$$

$$h_{1,2} = \frac{-b \pm \sqrt{\Delta}}{2a} \tag{4.1.10}$$

Z oczywistych powodów poszukiwanym rozwiązaniem jest tylko pierwiastek dodatni, czyli:

$$h = \frac{-b + \sqrt{\Delta}}{2a} = \frac{-B + (B^2 + 4m \cdot A)^{1/2}}{2m} \,. \tag{4.1.11}$$

Po wstawieniu równania (4.1.11) ido równania (4.1.5) funkcja celu przyjmuje ostateczną postać:

$$p(B) = B + 2\left(\frac{-B + (B^2 + 4m \cdot A)^{1/2}}{2m}\right)\sqrt{1+m^2} \quad \rightarrow \quad \min . \tag{4.1.12}$$

Warunek (4.9), który w tym przypadku przyjmuje postać:

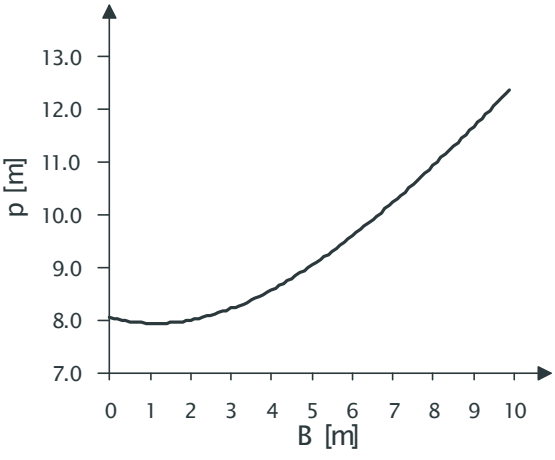
$$g(B) = \frac{dp}{dB} = 0 \tag{4.1.13}$$

prowadzi do następującego równania:

$$g(B) = 1 + 2\left(\frac{-1 + B\,(B^2 + 4m \cdot A)^{-1/2}}{2m}\right)\sqrt{1+m^2} = 0 . \tag{4.1.14}$$

Jego pierwiastek określa poszukiwane ekstremum funkcji $p(B)$.

Dla następującego zestawu danych: $A = 7{,}5$ m^2, $s = 0{,}0005$, $n = 0{,}030$ oraz $m = 1{,}5$ pierwiastek równania (4.1.14) jest położony w przedziale $\langle 0,\ 10\ \text{m}\rangle$. Wykres funkcji celu (4.1.12) w tym przedziale przedstawiono na rysunku 4.1.2.



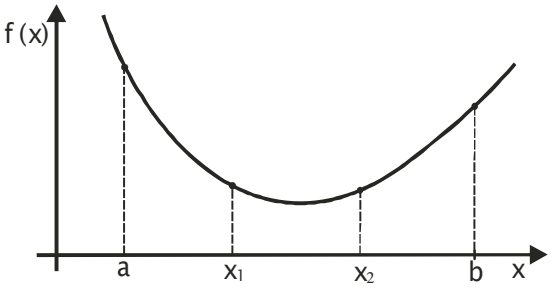
Rys. 4.1.2. Wykres funkcji celu $p(B)$ dla przyjętych danych

Wartość pierwiastka $B = 1{,}14$ m równania (4.1.14) obliczono metodą siecznych opisaną w rozdziale 2 z dokładnością $\varepsilon = 0{,}001$. Proces iteracyjny wymagał wykonania 3 iteracji. Przy obliczonej szerokości dna kanału w kanale panuje maksymalne możliwe natężenie przepływu $Q = 5{,}38$ m^3/s zaś głębokość wynosi $h = 1{,}887$ m.

■

4.3. Metoda podziału przedziału na trzy równe części

Rozpatrzmy przedział $\langle \alpha, b \rangle$ zawierajcy ekstremum funkcji $f(x)$ jak na rys. 4.4.



Rys. 4.4. Podział przedziału $\langle a, b \rangle$ na 3 równe części

Wewnątrz przedziału $\langle a, b \rangle$ przyjmijmy dwa punkty o współrzędnych

$$x_1^{(i)} = \frac{2}{3}a + \frac{1}{3}b, \qquad x_2^{(i)} = \frac{1}{3}a + \frac{2}{3}b\,. \tag{4.10, 4.11}$$

Rozpatrywany przedział podzielony został na 3 równe części. Ponieważ w każdej iteracji część niezawierająca ekstremum jest odrzucana, to jego początkowa długość jest redukowana 3/2 raza:

$$b^{(1)} - a^{(1)} = \frac{b^{(0)} - a^{(0)}}{3/2} = \frac{2}{3}\left(b^{(0)} - a^{(0)}\right). \tag{4.12}$$

Po wykonaniu K iteracji ekstremum funkcji $f(x)$ znajdzie się w następującym przedziale:

$$b^{(K)} - a^{(K)} = \left(\frac{2}{3}\right)^K \left(b^{(0)} - a^{(0)}\right). \tag{4.13}$$

Należy zauważyć, że w tej metodzie wartość funkcji celu $f(x)$ jest obliczana $2K$ razy.

■ *Przykład 4.2*

W wyniku pomiarów polowych w wybranym przekroju poprzecznym cieku dysponujemy zestawem pomierzonych wartości stanów wody H oraz odpowiadających im natężeń przepływów Q. Zatem znany jest zestaw danych (H_i, Q_i) dla $i = 1, 2, \ldots, n$, przy czym n oznacza liczbę wykonanych pomiarów. Dokonujemy aproksymacji funkcji $Q(H)$ zadanej w sposób dyskretny krzywą o równaniu:

$$Q = \alpha(H - \gamma)^{\beta}. \tag{4.2.1}$$

gdzie α, β oraz γ są parametrami, które należy określić.

Zauważmy, że zadanie to jest bardzo podobne do zadania rozwiązywanego w przykładzie 3.3. Jednak w poprzednim przypadku stosowana formuła aproksymująca krzywą przepływu była prostsza. Była nią, bowiem formuła (4.2.1), w której a priori przyjęto $\gamma = 0$. To pozwo-

liło dokonać linearyzacji równania aproksymującego i wyznaczyć jego nieznane współczynniki α oraz β metodą najmniejszych kwadratów. W przypadku równania (4.2.1) taki sposób postępowania nie jest możliwy. Jego struktura nie pozwala na linearyzację względem wszystkich trzech parametrów, co uniemożliwia określenie ich wartości wspomnianą metodą najmniejszych kwadratów. Jednak sformułowane wyżej zadanie można rozwiązać kojarząc omówioną w rozdziale 3 metodę najmniejszych kwadratów z zaprezentowaną wcześniej metodą optymalizacji.

W równaniu (4.2.1) H jest tzw. stanem wody i reprezentuje położenie zwierciadła wody w cieku w stosunku do arbitralnie przyjętego poziomu porównawczego. Zatem jest to wartość względna. Ponieważ stan wody wyrażany jest w centymetrach to jak wynika z równania (4.2.1), parametr γ powinien być wyrażony w tych samych jednostkach. Zauważmy jednocześnie, że jego wartość musi być większa od najniższego pomierzonego stanu wody.

Rozwiązanie zacznijmy od sformułowania funkcji celu, której ekstremum minimum będziemy poszukiwali. Przyjmijmy, że ma ona postać analogiczną do stosowanej w metodzie najmniejszych kwadratów, czyli:

$$F(\alpha, \beta, \gamma) = \sum_{i=1}^{n} [Q_i - \alpha(H_i - \gamma)^{\beta}]^2 \rightarrow \quad \min . \tag{4.2.2}$$

Załóżmy również, że wartość parametru γ jest znana. Zatem w równaniu należy określić wartości dwóch pozostałych parametrów a mianowicie α oraz β. Ponieważ przy znanej wartości γ równanie krzywej aproksymującej (4.2.1) jest nieliniowe względem parametru β, dokonajmy jego obustronnego logarytmowania:

$$\ln Q = \ln \alpha + \beta \ln(H - \gamma) . \tag{4.2.3}$$

Po wprowadzeniu nowych zmiennych:

$$z = \ln Q, \quad a_0 = \ln \alpha, \quad a_1 = \beta, \quad u = \ln (H - \gamma)$$

otrzymuje się następujący wielomian I stopnia:

$$z(u) = a_0 + a_1 u . \tag{4.2.4}$$

Kryterium najmniejszego błędu kwadratowego (3.31) w tym przypadku przyjmuje postać:

$$E(a_0, a_1) = \sum_{i=1}^{n} [z_i - (a_0 + a_1 u_i)]^2 \tag{4.2.5}$$

zaś układ równań, który spełniają współczynniki funkcji aproksymującej minimalizujące wyrażenie (4.2.5) jest następujący:

$$\begin{bmatrix} n & \sum_{i=1}^{n} u_i \\ \sum_{i=1}^{n} u_i & \sum_{i=1}^{n} u_i^2 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} z_i \\ \sum_{i=1}^{n} z_i u_i \end{bmatrix} \tag{4.2.6}$$

Zastosowanie wzorów Cramera daje następujące rozwiązanie układu:

$$a_0 = \frac{W_1}{W}, \quad a_1 = \frac{W_2}{W} \tag{4.2.7}$$

gdzie:

$$W = n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2, \quad W_1 = \sum_{i=1}^{n} z_i \sum_{i=1}^{n} u_i^2 - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} z_i u_i, \quad W_2 = n\sum_{i=1}^{n} z_i u_i - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} u_i .$$

W konsekwencji dla zadanej wartości parametru γ metoda najmniejszych kwadratów umożliwiła określenie wartości dwóch pozostałych parametrów:

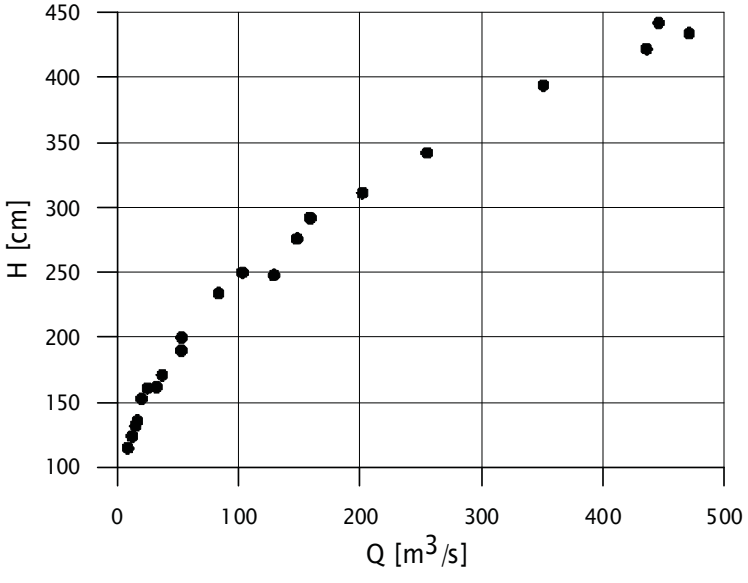
$$\alpha = e^{a_0} \quad \text{i} \quad \beta = a_1 .$$

Zatem znając ostatecznie wartości wszystkich trzech parametrów można obliczyć odpowiadającą przyjętej wartości γ wartość funkcji kryterialnej (4.2.2):

$$F(\gamma) = \sum_{i=1}^{n} [Q_i - \alpha(H_i - \gamma)^{\beta}]^2 \tag{4.2.8}$$

Postępując w przedstawiony wyżej sposób, dla kolejno zakładanych wartości parametru $\gamma = \gamma_1, \gamma_2, \gamma_3, \ldots$ można obliczyć odpowiadające wartości funkcji celu $F = F_1, F_2, F_3, \ldots$. Proces ten można prowadzić w sposób zorganizowany wykorzystując poleconą w treści zadania metodę podziału przedziału, w którym znajduje się minimum funkcji (4.2.8) na 3 równe części.

Zastosujmy przedstawiony sposób postępowania do aproksymacji zależności $Q(H)$ przedstawionej na rysunku 4.2.1.



Rys. 4.2.1. Zależność $Q(H)$ uzyskana w wyniku pomiarów polowych

Współrzędne poszczególnych punktów aproksymowanej zależności $Q(H)$ zestawiono w tab. 4.2.1.

Tabela 4.2.1

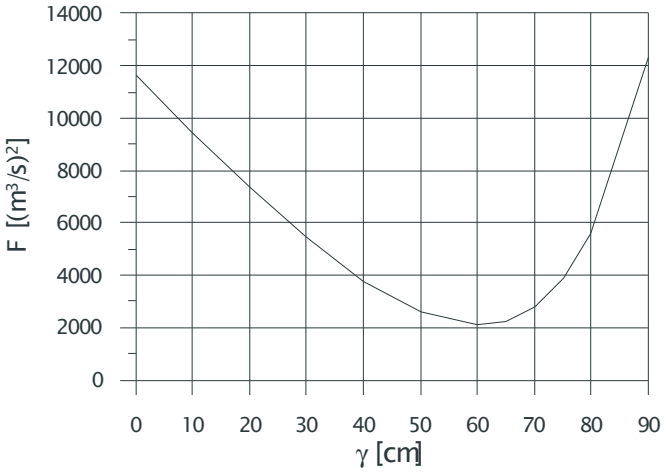
Pomierzone stany wody H i odpowiadające im natężenia przepływów Q

Numer punktu	H [cm]	Q [m^3/s]
1	161	32,1
2	115	12,2
3	124	12,0
4	132	15,0
5	136	16,6
6	248	129,0
7	153	19,9
8	161	24,7
9	171	36,9
10	190	52,8
11	200	53,0
12	234	83,5
13	250	103,0
14	276	148,0
15	292	159,0
16	311	202,0
17	342	255,0
18	394	351,0
19	442	446,0
20	434	446,0
21	422	436,0

Ponieważ omawiana metoda poszukiwania ekstremum funkcji jest ważna tylko dla funkcji unimodalnych, warto sprawdzić wykres przyjętej funkcji kryterialnej (4.2.8). Jej wykres dla przyjętego zestawu danych przedstawiono na rysunku 4.2.2. Jak widać, funkcja celu $F(\gamma)$ ma w przyjętym przedziale jedno ekstremum minimum czyli spełnia warunek unimodalności. Zatem położenie ekstremum można określić bardziej efektywnie stosując metodę optymalizacji polegającą na podziale na 3 równe części.

Obliczenia wykonane zgodnie z opisanym algorytmem dla przyjętych danych dały następujące wyniki:

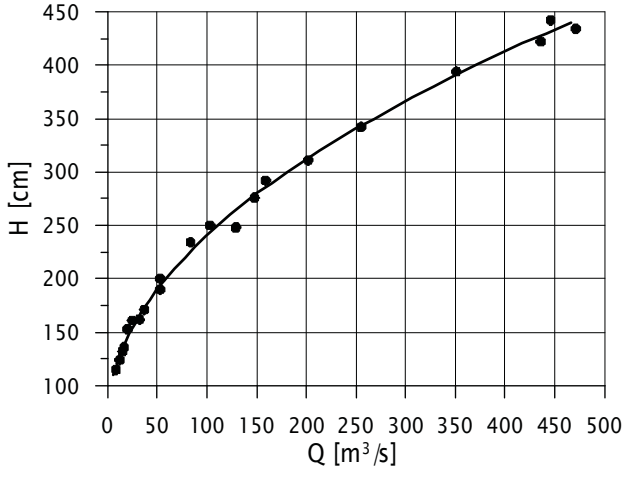
— minimalną wartość funkcja kryterialnej $F(\alpha, \beta, \gamma)$ przyjmuje dla $\gamma = 60$ cm;
— optymalnej wartości γ odpowiadają następujące wartości pozostałych parametrów $\alpha = 0{,}00188$ oraz $\beta = 2{,}09113$.



Rys. 4.2.2. Wykres funkcji celu $F(\gamma)$ dla krzywej przepływu zadanej w tab. 4.2.1 i aproksymowanej równaniem (4.2.1)

Zatem formuła aproksymująca zależność przedstawioną na rysunku 4.2.1 ma postać:

$$Q(H) = 0{,}00188(H - 60)^{2{,}09113} \tag{4.2.9}$$



Rys. 4.2.3. Zależność $Q(H)$ otrzymana w wyniku pomiarów (czarne punkty) i jej aproksymacja równaniem (4.2.9) (linia ciągła)

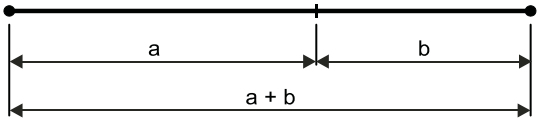
Wykorzystując formułę (4.2.9) należy pamiętać o zastosowanym w trakcie aproksymacji układzie jednostek. Stany wody należy podstawiać w cm otrzymując natężenie przepływu w m^3/s. Oczywiście formuła (4.2.9) obowiązuje w przedziale w którym wykonano aproksymację funkcji $Q(H)$. Jej wykres, na tle zadanych danych pomiarowych, przedstawiono na rysunku 4.2.2. Warto zwrócić uwagę na fakt wynikający z rysunku 4.2.2. Otóż zastosowana funkcja aproksymująca (4.2.1) znacznie lepiej przybliża zależność $Q(H)$ niż zastosowana

w rozdziale 3 funkcja (3.5) lub inaczej funkcja (4.2.1) z parametrem $\gamma = 0$. Wartość funkcji celu dla $\gamma = 60$ cm wynosi ok. 2000, podczas gdy jej wartość dla $\gamma = 0$ jest równa 11 500 (rys. 4.2.2).

■

4.4. Metoda złotego podziału

Rozpatrzmy odcinek o długości $a + b$ jak pokazano na rys. 4.5.



Rys. 4.5. Podział odcinka o długości $a + b$ na dwie części z wykorzystaniem złotej liczby

Przyjmuje się, że odcinek ten dzielony jest na dwie części zgodnie z zasadą złotego podziału, jeśli spełniona jest następująca relacja:

$$\frac{a}{a+b} = \frac{b}{a} \tag{4.14}$$

Wprowadzenie nowej zmiennej

$$\tau = \frac{b}{a} \tag{4.15}$$

umożliwia zapisanie równania (4.14) w następującej postaci

$$1 + \tau = \frac{1}{\tau} \tag{4.16}$$

lub równoważnie

$$\tau^2 + \tau - 1 = 0\,. \tag{4.17}$$

Wyróżnik tego równania jest równy

$$\Delta = 1 + 4 = 5$$

co oznacza, że ma ono dwa pierwiastki

$$\tau_{1,2} = \frac{-1 \pm \sqrt{5}}{2}\,. \tag{4.18}$$

Z definicji (4.15) wynika, że τ musi być liczbą dodatnią a zatem należy wybrać tylko dodatni pierwiastek

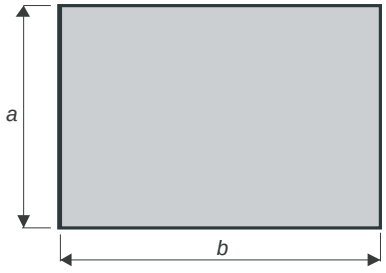
$$\tau = \frac{-1 + \sqrt{5}}{2} = 0{,}618\,. \tag{4.19}$$

Odwrotność otrzymanej liczby

$$\frac{1}{\tau} = \frac{b}{a} = \frac{1}{0{,}618} = 1{,}618 = \Phi \tag{4.20}$$

nazywa się złotą liczbą.

Historia złotej liczby jest bardzo długa, jako, że pochodzi ze Starożytnej Grecji. Jej definicja wiąże się z tzw. złotym prostokątem, w którym stosunek krótszego boku do dłuższego jest równy $a/b = 0{,}618$ (rys. 4.6)



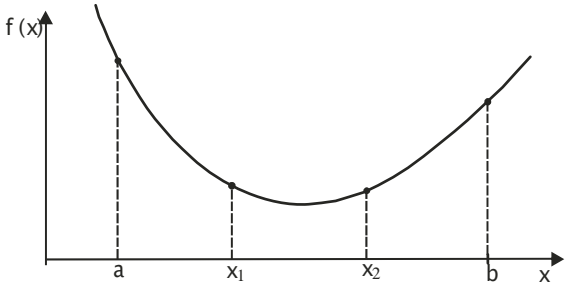
Rys. 4.6. Złoty prostokąt, w którym $a/b = 0{,}618$

Zależność ta jest powszechnie stosowana w architekturze jako wyrażająca pożądane proporcje pomiędzy wymiarami bryły budynku.

Zasadę złotego podziału można zastosować w procesie poszukiwania ekstremum funkcji $f(x)$. Załóżmy, że w przedziale $\langle a^{(i)}, b^{(i)} \rangle$ jej wartość jest znana w dwóch wybranych punktach $x_1^{(i)}$ oraz $x_2^{(i)}$. Punkty te należy wybrać tak, aby były spełnione następujące warunki: $x_1^{(i)} < x_2^{(i)}$ oraz

$$\frac{x_2^{(i)} - a^{(i)}}{b^{(i)} - a^{(i)}} = \frac{b^{(i)} - x_1^{(i)}}{b^{(i)} - a^{(i)}} = \tau \quad \text{i} \quad \frac{b^{(i)} - x_2^{(i)}}{b^{(i)} - x_1^{(i)}} = \tau \tag{4.21a, b}$$

gdzie $\tau = 0{,}618$.



Rys. 4.7. Podział przedziału $\langle a, b \rangle$ z zastosowaniem złotego podziału

Następnie przedział jest redukowany zgodnie z zasadami podanymi w poprzedniej metodzie podziału na 3 równe części.

Proces obliczeniowy według metody złotego podziału przebiega następująco:

1. **Inicjacja metody**

   Z zależności (4.21a)

$$\frac{b^{(0)} - x_1^{(0)}}{b^{(0)} - a^{(0)}} = \tau \quad \text{i} \quad \frac{x_2^{(0)} - a^{(0)}}{b^{(0)} - a^{(0)}} = \tau \tag{4.22, 4.23}$$

otrzymuje się odpowiednio

$$x_1^{(0)} = b^{(0)} - \tau(b^{(0)} - a^{(0)}) \tag{4.24}$$

i

$$x_2^{(0)} = a^{(0)} + \tau(b^{(0)} - a^{(0)}) \,. \tag{4.25}$$

2. Redukcja długości przedziału

Jeśli $f(x_2^{(i)}) > f(x_1^{(i)})$ to

$$\begin{aligned} b^{(i+1)} &= x_2^{(i)}; \quad a^{(i+1)} = a^{(i)} \\ x_2^{(i+1)} &= x_1^{(i)} \\ x_1^{(i+1)} &= b^{(i)} - \tau(b^{(i+1)} - a^{(i)}) \,. \end{aligned} \tag{4.26}$$

W przypadku przeciwnym

$$\begin{aligned} a^{(i+1)} &= x_1^{(i)}; \quad b^{(i+1)} = b^{(i)} \\ x_1^{(i+1)} &= x_2^{(i)} \\ x_2^{(i+1)} &= a^{(i)} - \tau(b^{(i)} - a^{(i+1)}). \end{aligned} \tag{4.27}$$

3. Warunek zakończenia procesu iteracyjnego

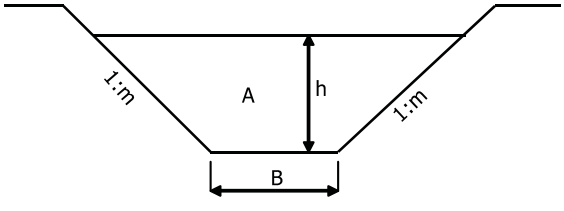
Jeśli

$$b^{(i+1)} - a^{(i+1)} \le \varepsilon \tag{4.28}$$

to proces obliczeń jest kończony, jeśli nie – należy powrócić do punktu 2.

■ *Przykład 4.3*

Rozpatrzmy problem budowy kanału o długości L i przekroju trapezowym, jak pokazany na rys. 4.3.1, w którym natężenie przepływu powinno wynosić Q. Warunki gruntowe w miejscu wykonania kanału pozwalają przyjąć nachylenie jego skarp równe $1 : m$.



Rys. 4.3.1. Przekrój poprzeczny projektowanego kanału

Całkowity koszt wykonania kanału jest sumą kosztu wykonania wykopu oraz umocnienia i uszczelnienia jego dna i skarp. Pomijając szczegóły związane z projektowaniem kanałów przyjmijmy, że łączny koszt jego budowy wyniesie:

$$K = C_1 \cdot A \cdot L + C_2 \cdot p \cdot L \tag{4.3.1}$$

gdzie: L – długość kanału,
A – pole przekroju czynnego kanału,
p – obwód zwilżony przekroju,

C_1 – koszt wykopania 1 m bieżącego kanału,
C_2 – koszt wykonania umocnienia kanału na długości 1 m.

Dzieląc obustronnie równanie (4.3.1) przez długość kanału L otrzymamy łączny koszt wykonania 1 mb kanału k:

$$k = \frac{K}{L} = C_1 \cdot A + C_2 \cdot p \,. \tag{4.3.2}$$

Zadanie polega na tym, aby znając następujące dane:
— natężenie przepływu Q,
— spadek dna kanału s,
— współczynnik szorstkości wg Manninga n,
— pochylenie skarp kanału m,
— jednostkowa cena wykonania wykopu C_1,
— jednostkowa cena umocnienia kanału C_2

dobrać taką szerokość dna kanału B, przy której całkowity koszt wykonania kanału będzie minimalny. Zatem rozwiązanie sprowadza się do znalezienia ekstremum minimum zależnej od szerokości dna kanału B funkcji celu zdefiniowanej wzorem (4.3.2):

$$k(B) = C_1 \cdot A + C_2 \cdot p \quad \rightarrow \quad \min \tag{4.3.3}$$

Do rozwiązania należy zastosować opisaną wcześniej metodę złotego podziału.

Kluczową kwestią jest tutaj obliczenie wartości funkcji $k(B)$. W tym celu należy wykorzystać dobrze znane równanie przepływu ustalonego jednostajnego. Przepływ ustalony jednostajny w kanale otwartym opisuje równanie Manninga:

$$Q = \frac{1}{n} R^{2/3} \cdot s^{1/2} \cdot A \tag{4.3.4}$$

gdzie R jest promieniem hydraulicznym zdefiniowanym wzorem

$$R = \frac{A}{p} \,. \tag{4.3.5}$$

Pole powierzchni przekroju czynnego kanału trapezowego A i jego obwód zwilżony p oblicza się stosowanymi już formułami:

$$p = B + 2h\sqrt{1+m^2} \,, \qquad A = B \cdot h + m \cdot h^2 \tag{4.3.6, 4.3.7}$$

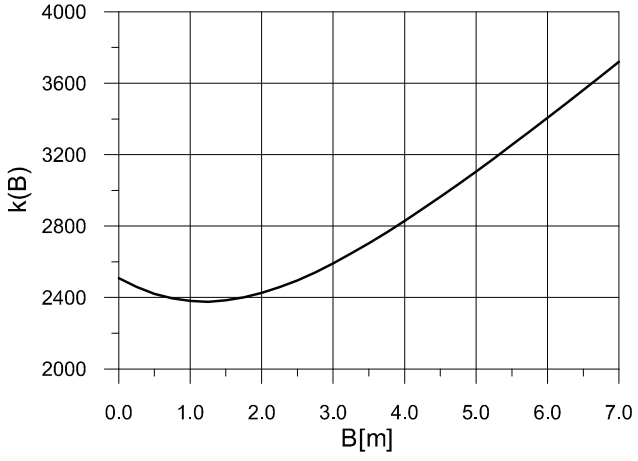
Wartości funkcji celu $k(B)$ liczymy następująco:
— Przyjmujemy wartość szerokości dna kanału B.
— Obliczamy głębokość normalną w kanale rozwiązując równanie Manninga (4.3.4) względem h czyli poszukujemy pierwiastka równania:

$$f(h) = Q - \frac{1}{n} R^{2/3} \cdot s^{1/2} \cdot A = 0 \,. \tag{4.3.8}$$

Ponieważ równanie to jest nieliniowe, do jego rozwiązania stosujemy jedną ze znanych metod przybliżonych jak metodę bisekcji, siecznych, Newtona itp.
— Obliczamy parametry przekroju trapezowego, czyli pole powierzchni przekroju i obwód zwilżony według wzorów (4.3.6) i (4.3.7),
— Obliczamy wartość funkcji celu (4.3.3)

Proces poszukiwania ekstremum minimum poleconą metodą złotego podziału przebiega zgodnie z podanym wcześniej jej algorytmem, to znaczy najpierw określa się przedział, w którym funkcja $k(B)$ ma ekstremum minimum a następnie z przyjętą dokładnością ε oblicza się jego przybliżone położenie.

Dla następującego, arbitralnie przyjętego zestawu danych: $Q = 2{,}5$ m^3/s, $s = 0{,}0006$, $n = 0{,}025$, $m = 1{,}0$, $C_1 = 200$ zł/m^3/m przekroju, $C_2 = 300$ zł/m/m funkcja celu $k(B)$ ma kształt jak przedstawiony na rys. 4.3.2. Jednocześnie stwierdzono, że ekstremum funkcji znajduje się w przedziale $\langle 0{,}0,\ 7{,}0\ \text{m}\rangle$.



Rys. 4.3.2. Wykres funkcji celu $k(B)$

Dla założonej dokładności rozwiązania $\varepsilon = 0{,}01$ m, po 14 iteracjach otrzymano optymalną szerokość dna kanału $B = 1{,}23$ m, której odpowiada głębokość normalna $h = 1{,}451$ m. Przy tej szerokości koszt wykonania 1 metra bieżącego kanału wynosi $k_{\min} = 2377$ zł.

■

4.5. Metoda wykorzystująca ciąg liczb Fibonacciego

Spośród wszystkich metod podziału, metoda wykorzystująca ciąg liczb Fibonacciego wymaga najmniejszej liczby iteracji. Ciąg Fibonacciego jest definiowany następująco:

$$F_i = F_{i-1} + F_{i-2} \tag{4.29}$$

dla $i = 2, 3, 4, \ldots$ przy czym $F_0 = F_1 = 1$.

Kolejnymi wyrazami ciągu są następujące liczby: 1, 1, 2, 3, 5, 8, 13, … .

Szereg (4.29) zaproponował matematyk włoski Fibonacci żyjący w Pizzie w latach 1180–1250. Jego nazwisko jest dobrze znane w historii nauki, ponieważ to on zainicjował wprowadzenie cyfr arabskich do Europy. Wzór (4.29) jest rezultatem badań jakie prowadził Fibonacci nad populacją królików.

Fibonacci analizował następujący problem: para królików żyjących na łące każdego miesiąca wydaje na świat dwójkę potomstwa – samca i samicę. Urodzone króliki nie giną i również każdego miesiąca rodzą parę królików (samca i samicę). Fibonacci poszukiwał

odpowiedzi na następujące pytanie: ile par królików będzie żyło na łące po upływie jednego roku?

Ciąg Fibonacciego ma bardzo interesującą właściwość. W miarę wzrostu liczby jego wyrazów stosunek dwóch kolejnych wyrazów zmierza do wartości odpowiadającej poprzednio zdefiniowanej złotej liczby Φ, tzn.:

$$\frac{F_i}{F_{i-1}} \to \Phi \quad \text{dla} \quad i \to \infty . \tag{4.30}$$

Na przykład, dla $i = 16$ otrzymuje się

$$\frac{F_{16}}{F_{15}} = 1{,}61803 .$$

Proces obliczeń przebiega następująco:

1. Inicjacja metody

Proces rozpoczyna się od obliczenia następującej wartości

$$c = \frac{b^{(0)} - a^{(0)}}{\varepsilon} \tag{4.31}$$

gdzie ε oznacza założoną dokładność rozwiązania. Następnie poszukuje się takiej wartości N dla której zachodzi relacja

$$F_{N-1} < c \le F_N . \tag{4.32}$$

2. Redukcja długości bieżącego przedziału

W kolejnych iteracjach $i = 1, 2, 3, \ldots, N–2$ w aktualnie rozpatrywanym przedziale wybierane są dwa punkty o następujących współrzędnych:

$$x_1^{(i)} = \frac{F_{N-i-1}}{F_{N-i+1}}\left(b^{(i-1)} - a^{(i-1)}\right) + a^{(i-1)} \tag{4.33}$$

$$x_2^{(i)} = \frac{F_{N-i}}{F_{N-i+1}}\left(b^{(i-1)} - a^{(i-1)}\right) + a^{(i-1)} . \tag{4.34}$$

Nowe granice przedziału zawierającego ekstremum są obliczane alternatywnie:

- Jeśli $f(x_1^{(i)}) \le f(x_2^{(i)})$ to lewy brzeg przedziału nie ulega zmianie, czyli $a^{(i)} = a^{(i-1)}$, natomiast prawy tak – $b^{(i)} = x_2^{(i)}$.
- Jeśli $f(x_1^{(i)}) > f(x_2^{(i)})$ to lewy brzeg jest przesuwany $a^{(i)} = x_1^{(i)}$, natomiast prawy nie – $b^{(i)} = b^{(i-1)}$.

3. Zbieżność procesu iteracyjnego

Po wykonaniu $N–2$ iteracji początkowa długość przedziału zostaje zredukowana do wartości:

$$b^{(N-2)} - a^{(N-2)} = \frac{F_{N-1}}{F_N} \cdot \frac{F_{N-2}}{F_{N-1}} \cdots \frac{F_2}{F_3}\left(b^{(0)} - a^{(0)}\right) = 2\frac{b^{(0)} - a^{(0)}}{F_N} \le 2\varepsilon \tag{4.35}$$

Warto dodać, że w trakcie procesu iteracyjnego wartość funkcji kryterialnej $f(x)$ jest obliczana $N–1$ razy.

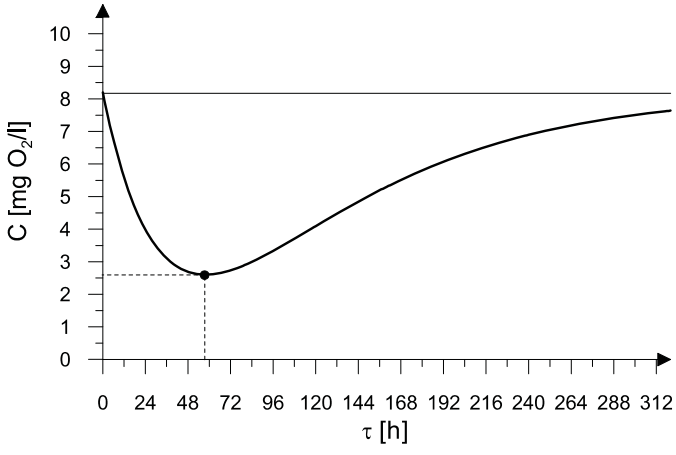
■ *Przykład 4.4*

Załóżmy, że przed punktem zrzutu ścieków woda płynąca w rzece jest nasycona tlenem zaś poniżej tego punktu nastąpiło dobre wymieszanie ścieków z wodą. W takiej sytuacji koncentrację tlenu rozpuszczonego w rzece poniżej zrzutu ścieków określa równanie Streetera – Phelpsa zapisane w następującej postaci (Adamski, 2002):

$$C(\tau) = C_s - \frac{k_b \cdot L_0}{k_a - k_b}\left(e^{-k_a \cdot \tau} - e^{-k_b \cdot \tau}\right) \tag{4.4.1}$$

gdzie: C – koncentracja tlenu rozpuszczonego w wodzie [mg/l],
C_s – koncentracja tlenu rozpuszczonego w stanie nasycenia [mg/l],
L_0 – biochemiczne zapotrzebowanie tlenu w punkcie wymieszania ścieków [mg/l],
k_b – stała szybkości usuwania BZT [h^{-1}],
k_a – stała reaeracji (napowietrzania) [h^{-1}],
τ – czas przemieszczania się zanieczyszczeń w cieku [h].

Rozpatrywana funkcja posiada ekstremum minimum. Jego położnie określa przekrój rzeki, w którym mogą wystąpić krytyczne warunki tlenowe dla organizmów zwierzęcych. Metodą Fibonacci'ego należy określić krytyczny czas wystąpienia minimum koncentracji tlenu (ekstremum minimum funkcji $C(\tau)$), jego wartość oraz odpowiadające mu położenie minimum wzdłuż osi rzeki. Temperatura wody w rzece wynosi +25°C zaś średnia prędkość jej przepływu jest równa V_0 = 0,52 m/s = 1,872 km/h. Obliczenia należy wykonać dla następujących danych: C_s = 8,20 mg O_2/l, k_b = 0,26 1/d, k_a = 0,62 1/d, L_0 = 25,0 mg O_2/l.



Rys. 4.4.1. Wykres koncentracji tlenu $C(\tau)$

Optymalizacja z wykorzystaniem ciągu liczb Fibonacci'ego wymagała wykonania 11 iteracji dla przyjętej dokładności rozwiązania ε = 0,1. Obliczona krytyczna koncentracja tlenu rozpuszczonego w wodzie o wartości C_{kr} = 2,60 mg/l, wystąpi po czasie τ_{kr} = 58,33 h w przekroju rzeki położonym w odległości

$$x_{kr} = V_0 \quad \tau_{kr} = 1{,}87 \cdot 58{,}33 \approx 105 \text{ km}$$

od punktu zrzutu ścieków.

■

5 Całkowanie numeryczne

5.1. Całki oznaczone w inżynierii wodnej

W trakcie rozwiązywania wielu zagadnień z zakresu inżynierii wodnej pojawia się potrzeba numerycznego obliczenia wartości całki oznaczonej. Poniżej przedstawiono kilka przykładów występowania całek oznaczonych.

Analizując różne przypadki przepływu wody hydrolodzy operują tzw. krzywymi sumowymi. Krzywa sumowa to funkcja czasu zdefiniowana jako całka oznaczona z wielkości podlegającej sumowaniu. Wartość całki określa łączną ilość danej wielkości, która wystąpiła od umownie przyjętego czasu początkowego. Do najczęściej stosowanych w projektowaniu hydrologicznym krzywych tego typu należą:

— krzywa sumowa przepływu

$$V(t) = \int_{t_0}^{t} Q(\tau) \cdot d\tau \tag{5.1}$$

gdzie: V – objętość wody jaka przepłynęła przez wybrany przekrój rzeki w przedziale czasowym $\langle t_0, t \rangle$,
t_0 – umownie przyjęty czas początkowy,
t – czas,
Q – natężenie przepływu,
τ – zmienna całkowania;

— krzywa sumowa infiltracji

$$F(t) = \int_{t_0}^{t} f(\tau) \cdot d\tau \tag{5.2}$$

gdzie: F – objętość wody jaka wnika przez powierzchnię terenu do gruntu w przedziale czasowym $\langle t_0, t \rangle$,
t_0 – umownie przyjęty czas początkowy,
t – czas,
f – natężenie infiltracji,
τ – zmienna całkowania;

— krzywa sumowa opadów atmosferycznych

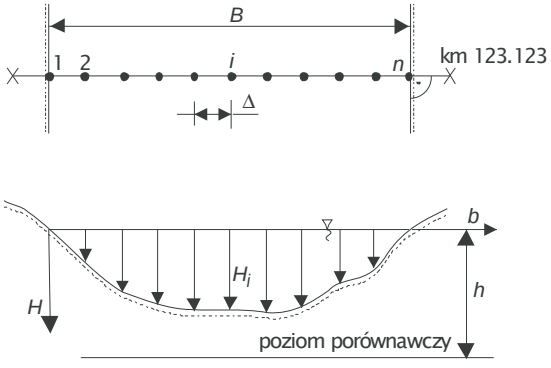
$$S(t) = \int_{t_0}^{t} P(\tau) \cdot d\tau \tag{5.3}$$

gdzie: S – objętość wody jaka spadła na powierzchnię terenu w przedziale czasowym $\langle t_0, t \rangle$,
t_0 – umownie przyjęty czas początkowy,
t – czas,

P – natężenie opadów,
τ – zmienna całkowania.

Jak można zauważyć, wyznaczenie każdej z wymienionych krzywych sumowych wymaga numerycznego obliczania wartości całki oznaczonej.

Jednym z ważniejszych parametrów charakteryzujących ciek wodny w wybranym miejscu jest pole przekroju czynnego. Dla cieków naturalnych informacji o przekrojach dostarczają pomiary polowe. W konsekwencji przekrój poprzeczny cieku jest zdefiniowany w sposób dyskretny, za pomocą zbioru punktów, których współrzędne zostały pomierzone. (rys. 5.1).



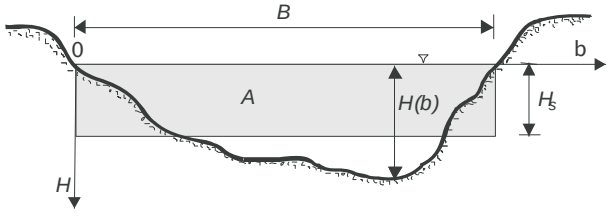
Rys. 5.1. Przekrój poprzeczny kanału naturalnego

Jeśli dla dowolnego napełnienia kanału – stanu wody jest wymagana znajomość odpowiadającego mu pola powierzchni przekroju, to jego wartość otrzymuje się po obliczeniu następującej całki:

$$A = \int_0^B H(b) \cdot db \,. \tag{5.4}$$

Taką samą całkę należy policzyć, jeśli chcemy wyznaczyć średnią głębokość wody w analizowanym przekroju cieku

$$H_s = \frac{A}{B} = \frac{1}{B} \int_0^B H(b) \cdot db \tag{5.5}$$



Rys. 5.2. Średnia głębokość w przekroju cieku naturalnego

W wielu zagadnieniach hydrauliki kanałów występuje potrzeba uśrednienia danej funkcji w przekroju poprzecznym kanału lub w wybranym pionie pomiarowym od dna kanału do zwierciadła wody. Problem ten dotyczy uśredniania takich wielkości jak prędkość przepływu, temperatura wody lub koncentracja domieszki rozpuszczonej w wodzie. Uśrednienie w pionie wymienionych wielkości wykonuje się za pomocą następujących formuł:

$$U = \int_Z^h u(z) \cdot dz, \quad T = \int_Z^h \tau(z) \cdot dz, \quad C = \int_Z^h c(z) \cdot dz, \tag{5.6a,b,c}$$

gdzie: Z – wzniesienie dna kanału ponad przyjęty poziom porównawczy,
h – wzniesienie zwierciadła wody w kanale ponad przyjęty poziom porównawczy,
U – uśredniona w pionie prędkość przepływu $u(z)$,
T – uśredniona w pionie temperatura wody $\tau(z)$,
C – uśredniona w pionie koncentracja $c(z)$,
z – zmienna całkowania.

Ponieważ funkcje podcałkowe $u(z)$, $\tau(z)$ oraz $c(z)$ otrzymywane są poprzez pomiary, zatem są one przedstawione w postaci dyskretnej, co wymaga numerycznego obliczania całek.

Kolejny przykład dotyczy gospodarki wodnej. Ważną informacją dla użytkowników zbiorników retencyjnych, wykorzystywanych dla celów ochrony przed powodzią, jest objętość prognozowanej fali wezbraniowej. Obliczamy ją na podstawie obserwowanego lub prognozowanego natężenia przepływu w rozpatrywanym przekroju rzeki:

$$W = \int_0^T Q(t) \cdot dt \tag{5.7}$$

gdzie: W – objętość fali wezbraniowej,
t – czas,
Q – natężenie przepływu,
T – czas trwania wezbrania.

Identyczne całki obliczane są w początkowym i końcowym przekroju rozpatrywanego odcinka rzeki w trakcie bilansowania masy lub objętości wody.

Warto dodać, że wiele znanych formuł numerycznego rozwiązywania równań różniczkowych zwyczajnych szeroko stosowanych w hydraulice można wyprowadzić stosując metody numerycznego całkowania. Aby krótko wyjaśnić tę kwestię rozpatrzmy problem rozwiązania zagadnienia początkowego następującego równania różniczkowego zwyczajnego:

$$\frac{dy}{dx} = f(x, y). \tag{5.8}$$

Załóżmy, że w punkcie x_i wartość funkcji y jest znana: $y(x_i) = y_i$. Naszym celem jest znalezienia przybliżonej wartości funkcji y w następnym punkcie, to znaczy $y(x_{i+1} = x_i + \Delta x)$ oznaczonej symbolem y_{i+1}. W tym celu całkujemy równanie (5.8):

$$\int_{y_i}^{y_{i+1}} dy = \int_{x_i}^{x_{i+1}} f(x, y) dx \tag{5.9}$$

otrzymując

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f(x, y)dx \,. \tag{5.10}$$

Zależnie od zastosowanej formuły przybliżającej wartość całki oznaczonej w równaniu (5.10), otrzymuje się odpowiednią metodę numerycznego rozwiązywania równania różniczkowego zwyczajnego (5.8). Problem ten zostanie omówiony bardziej szczegółowo w rozdziale 6.

5.2. Numeryczne obliczanie wartości pojedynczych całek oznaczonych

Przedstawione w poprzednim podrozdziale całki oznaczone łączy wspólna cecha. Ich wartości dokładnych nie można obliczyć. W związku z tym będziemy poszukiwali ich przybliżonych wartości. Zatem będziemy mieli funkcję jednej zmiennej $f(x)$ zadaną w przedziale $\langle a, b\rangle$. Celem postępowania jest obliczenie przybliżonej wartości całki funkcji $f(x)$ w tym przedziale

$$I = \int_a^b f(x) \cdot dx \,. \tag{5.11}$$

W tym celu rozpatrywany przedział dzielony jest na podprzedziały o długości h za pomocą zbioru punktów leżących wewnątrz $\langle a, b\rangle$:

$$a = x_0 < x_1 < \ldots < x_i < \ldots < x_{n-1} < x_n = b$$

gdzie

$$n = \frac{b-a}{h} \tag{5.12}$$

jest liczbą podprzedziałów. Całkę (5.11) można przepisać w równoważnej postaci:

$$I = \int_a^b f(x) \cdot dx = \sum_{i=1}^{n} I_i = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x) \cdot dx \tag{5.13}$$

Przedmiotem przybliżonych obliczeń jest całka I_i:

$$I_i = \int_{x_{i-1}}^{x_i} f(x) \cdot dx \tag{5.14}$$

Istnieje szereg dobrze znanych metod numerycznego obliczania całek oznaczonych (tzw. kwadratur) jak różne wersje metody prostokątów, metoda trapezowa, metoda Simpsona i inne. Wszystkie wymienione metody można wyprowadzić stosując jednolite podejście w postaci formuł Newtona-Cotesa.

Ogólna idea tego podejścia polega na aproksymacji funkcji $f(x)$ w przedziale $\langle a, b\rangle$ wielomianem interpolacyjnym

$$I = \int_a^b f(x) \cdot dx \cong \int_a^b F_m(x) \cdot dx \tag{5.15}$$

gdzie $F_m(x)$ jest wielomianem stopnia m

$$F_m(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \cdots + a_{m-1} \cdot x^{m-1} + a_m \cdot x^m \tag{5.16}$$

Zależnie od przyjętego stopnia wielomianu m otrzymuje się różne typy ostatecznych formuł. Rozpatrzmy niektóre z nich.

Wielomian stopnia $m = 0$: $F_0(x) = a_0$

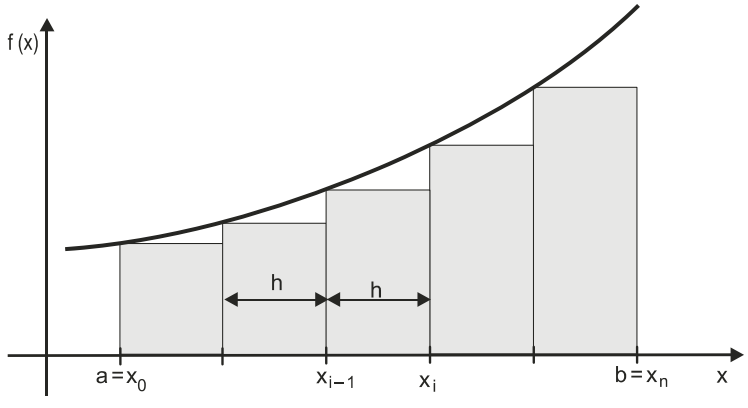
Wielomian ten przyjmuje stałą wartość w całym rozpatrywanym przedziale. Zależnie od założonej jego wartości w podprzedziale $\langle x_{i-1}, x_i \rangle$ otrzymuje się rodzinę formuł zaliczanych do metody prostokątów. Są one następujące:

— **wersja a**: przyjmując $F_0 = f_{i-1}$ otrzymuje się

$$I_i = \int_{x_{i-1}}^{x_i} F_0(x) \cdot dx = \int_{x_{i-1}}^{x_i} f_{i-1} \cdot dx \approx h \cdot f_{i-1} \tag{5.17}$$

i w konsekwencji (rys. 5.3)

$$I = \int_a^b f(x) \cdot dx = \sum_{i=1}^{n} I_i \approx h \cdot \sum_{i=1}^{n} f_{i-1} \tag{5.18}$$



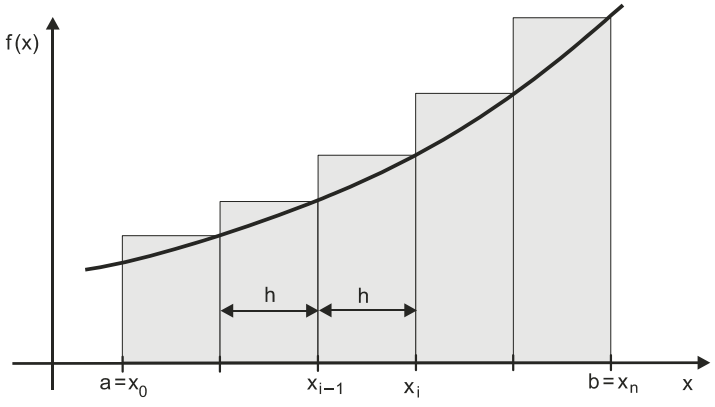
Rys. 5.3. Metoda prostokątów – wersja a

— **wersja b**: przyjmując $F_0 = f_i$ otrzymuje się

$$I_i = \int_{x_{i-1}}^{x_i} F_0(x) \cdot dx = \int_{x_{i-1}}^{x_i} f_i \cdot dx \approx h \cdot f_i \tag{5.19}$$

i w konsekwencji (rys. 5.4)

$$I = \int_a^b f(x) \cdot dx = \sum_{i=1}^{n} I_i \approx h \cdot \sum_{i=1}^{n} f_i \tag{5.20}$$



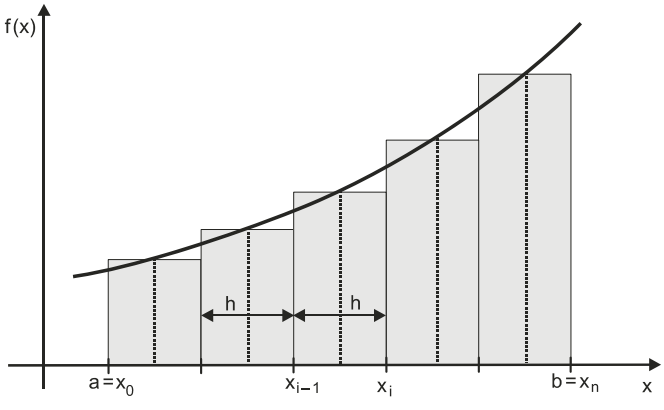
Rys. 5.4. Metoda prostokątów – wersja b

— **wersja c**: przyjmując $F_0 = f_{i-1/2}$ otrzymuje się

$$I_i = \int_{x_{i-1}}^{x_i} F_0(x) \cdot dx = \int_{x_{i-1}}^{x_i} f_{i-1/2} \cdot dx \approx h \cdot f_{i-1/2} \tag{5.21}$$

i w konsekwencji (rysunek 5.5)

$$I = \int_a^b f(x) \cdot dx = \sum_{i=1}^{n} I_i \approx h \cdot \sum_{i=1}^{n} f_{i-1/2} \tag{5.22}$$



Rys. 5.5. Metoda prostokątów – wersja c

Występująca we wzorach (5.21) i (5.22) zmienna $f_{i-1/2}$ oznacza wartość funkcji $f(x)$ w środku przedziału $\langle x_{i-1}, x_i \rangle$.

Wielomian stopnia $m = 1$: $F_1(x) = a_0 + a_1 \cdot x$

Wielomian ten zmienia się liniowo. W podprzedziale $\langle x_{i-1}, x_i \rangle$ zdefiniowany jest następująco:

$$F_1(x) = f_{i-1} + \frac{f_i - f_{i-1}}{h}(x - x_{i-1}) = \left(f_{i-1} - \frac{f_i - f_{i-1}}{h} x_{i-1} \right) + \left(\frac{f_i - f_{i-1}}{h} \right) x \tag{5.23}$$

Zatem

$$a_0 = \left(f_{i-1} - \frac{f_i - f_{i-1}}{h} x_{i-1} \right), \qquad a_1 = \left(\frac{f_i - f_{i-1}}{h} \right). \tag{5.24a, b}$$

Przybliżoną wartość całki

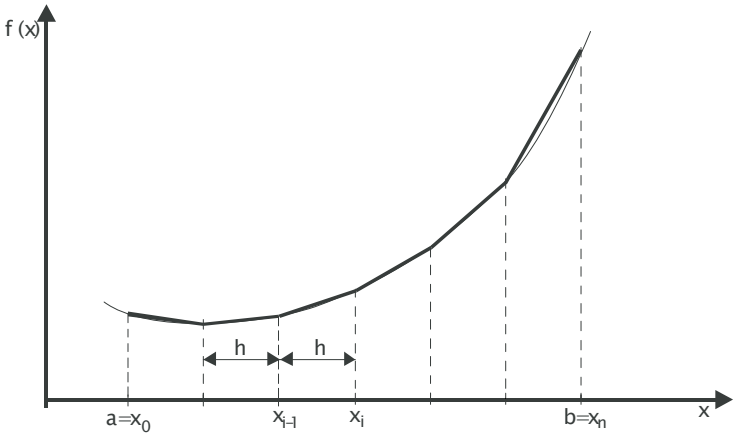
$$I_i = \int_{x_{i-1}}^{x_i} F_1(x) \cdot dx = \int_{x_{i-1}}^{x_i} (a_0 + a_1 \cdot x) \cdot dx \tag{5.25}$$

oblicza się następująco:

$$\begin{aligned} &\int_{x_{i-1}}^{x_i} (a_0 + a_1 \cdot x) \cdot dx = \left(a_0 \cdot x + a_1 \cdot \frac{x^2}{2} \right) \Bigg|_{x_{i-1}}^{x_i} = \\ &= a_0 \cdot (x_i - x_{i-1}) + a_1 \cdot \left(\frac{x_i^2}{2} - \frac{x_{i-1}^2}{2} \right) = a_0 \cdot h + \frac{a_1}{2} h \cdot (x_i + x_{i-1}). \end{aligned} \tag{5.26}$$

Podstawienie współczynników a_0 oraz a_1 do równania (5.26) i proste przekształcenia prowadzą do następującej wartości całki w podprzedziale $\langle x_{i-1}, x_i \rangle$

$$I_i = \int_{x_{i-1}}^{x_i} f(x) \cdot dx \cong h \cdot \frac{f_{i-1} + f_i}{2} \tag{5.27}$$



Rys. 5.6. Ilustracja przybliżenia całki metodą trapezów

Przy jednostajnym podziale przedziału $\langle a, b\rangle$ (h = const), całkę I oblicza się następująco:

$$I = \int_a^b f(x)\cdot dx \cong h\cdot \sum_{i=1}^{n} \frac{f_{i-1}+f_i}{2} = h\left(\frac{f_0}{2} + f_1 + f_2 + \cdots + f_{n-1} + \frac{f_n}{2}\right) \tag{5.28}$$

natomiast przy podziale niejednostajnym, gdy odległości h pomiędzy punktami wewnątrz przedziału $\langle a, b\rangle$ są różne, całka I jest równa:

$$I = \int_a^b f(x)\cdot dx \cong \sum_{i=1}^{n} h_i \cdot \frac{f_{i-1}+f_i}{2}\,. \tag{5.29}$$

Jeśli chodzi o dokładność obliczenia całki, najmniej dokładne są dwie pierwsze wersje metody prostokątów (a oraz b). Interesujące jest porównanie dokładności metody trapezów oraz metody prostokątów wersji c. Oszacowania ich błędów są następujące (Chapra i Canale, 2006):

— dla metody trapezowej

$$E = \frac{h^2}{12}(b-a)f''(\xi); \tag{5.30}$$

— dla metody prostokątów (wersja c)

$$E = -\frac{h^2}{24}(b-a)f''(\xi) \tag{5.31}$$

gdzie $\xi \in \langle a, b\rangle$.

Zauważmy, że ten wariant metody prostokątów zapewnia większą dokładność niż metoda trapezowa. Jak można zauważyć, błąd metody jest zdeterminowany przez krok całkowania h. Jeśli funkcję podcałkową opisuje znana formuła to błąd metody można zmniejszyć zmniejszając krok całkowania h. Jednakże bardzo często ten sposób zwiększenia dokładności nie może być stosowany, ponieważ całka musi być liczona z zadanym wynikającym z pomiarów terenowych lub laboratoryjnych, krokiem całkowania h_i.

Wielomian stopnia $m = 2$: $F_2(x) = a_0 + a_1 x + a_2 x^2$

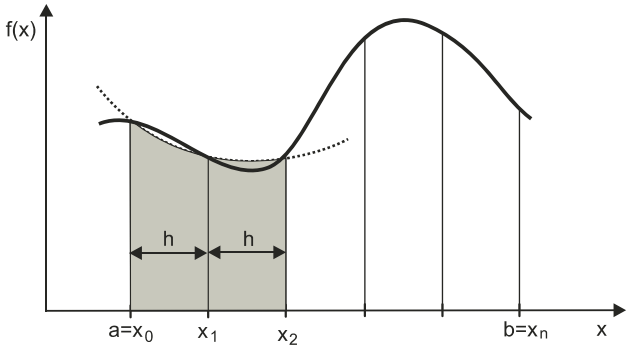
Wielomian ten prowadzi do dobrze znanej i powszechnie stosowanej metody Simpsona. Jak poprzednio, przedział $\langle a, b\rangle$ jest jednostajnie dzielony (h = const) na n podprzedziałów za pomocą $n + 1$ węzłów włączając skrajne węzły. Obliczenie całki (5.11) można wykonać wykorzystując wielomian 2. stopnia, którym odcinek po odcinku zastępujemy funkcję $f(x)$ (rys. 5.7). W tym celu uwzględniamy podprzedziały o długości $2h$ czyli dwukrotnie dłuższe niż stosowane w metodzie prostokątów i trapezów. Zatem mamy 3 węzły interpolacji, w których znane są wartości funkcji $f(x)$:

$$(x_0,\ f(x_0)),\quad (x_1,\ f(x_1)),\quad (x_2,\ f(x_2)).$$

Wykorzystując powyższe węzły można skonstruować, wielomian Lagrange'a 2. stopnia ($m = 2$).

Ogólna formuła wielomianu Lagrange'a (3.13) w tym przypadku przyjmie następującą postać:

$$F_2(x) = f(x_0)\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f(x_1)\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f(x_2)\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \tag{5.30}$$



Rys. 5.7. Ilustracja metody Simpsona

W podprzedziale $\langle x_0, x_2\rangle$ całkę z powyższego wielomianu obliczamy następująco:

$$\int_{x_0}^{x_2} F_2(x)\cdot dx =$$

$$= \int_{x_0}^{x_2}\left(f(x_0)\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f(x_1)\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f(x_2)\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\right)\cdot dx =$$

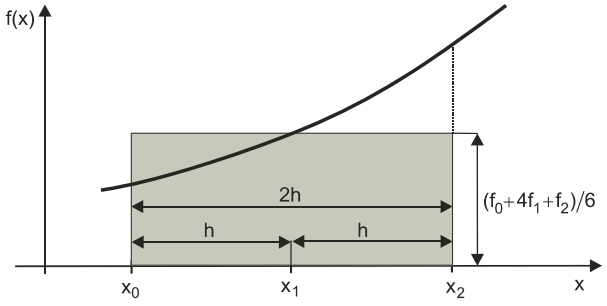
$$= \frac{h}{3}\big(f(x_0) + 4f(x_1) + f(x_2)\big). \tag{5.31}$$

Przedstawmy powyższą formułę w sposób podobny do formuł poprzednio omawianych:

$$\int_{x_0}^{x_2} F_2(x)\cdot dx = \frac{h}{3}\big(f(x_0) + 4f(x_1) + f(x_2)\big) = (2h)\cdot\frac{f(x_0) + 4f(x_1) + f(x_2)}{6} \tag{5.32}$$

Jak można zauważyć, w rzeczywistości jest to formuła zbliżona do formuły metody prostokątów. Całka jest równa iloczynowi podstawy prostokąta o szerokości $2h$ oraz jego wysokości, równej pewnej średniej ważonej (rys. 5.8):

$$\frac{f(x_0) + 4f(x_1) + f(x_2)}{6}.$$



Rys. 5.8. Interpretacja metody Simpsona

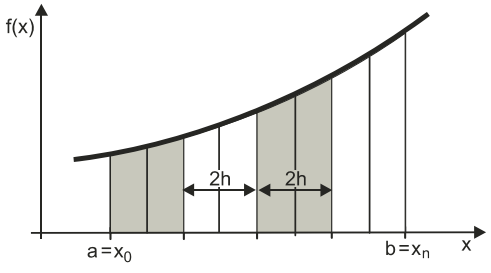
Dla dowolnego podprzedziału $\langle x_{i-1}, x_{i+1}\rangle$ formuła Simpsona przyjmuje postać:

$$\int_{x_{i-1}}^{x_{i+1}} F_2(x)\cdot dx = \frac{h}{3}\left(f(x_{i-1}) + 4f(x_i) + f(x_{i+1})\right) \tag{5.33}$$

natomiast dla całego przedziału całkowania $\langle a, b\rangle$ otrzymuje się

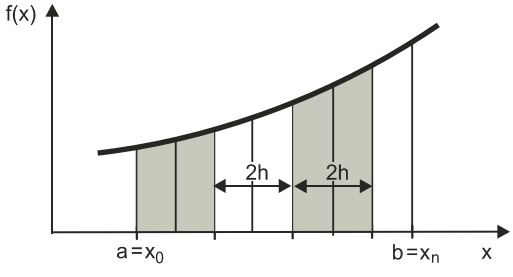
$$I = \int_a^b f(x)\cdot dx = \frac{h}{3}\sum_{i=0}^{n/2-1}\left(f_{2i} + 4f_{2i+1} + f_{2i+2}\right). \tag{5.34}$$

Wzór (5.34) obowiązuje dla parzystej liczby podprzedziałów n (rys. 5.9).



Rys. 5.9. Całkowanie metodą Simpsona w wypadku parzystej liczby podprzedziałów

Jeżeli liczba podprzedziałów n jest nieparzysta, to w ostatnim podprzedziale $\langle b-h, b\rangle$ należy obliczyć dodatkową całkę (rys. 5.10):



Rys. 5.10. Numeryczne całkowanie metodą Simpsona przy nieparzystej liczbie podprzedziałów

$$\int_{b-h}^{b} F_2(x)\cdot dx = \frac{h}{12}\left(-f(x_{n-2}) + 8f(x_{n-1}) + 5f(x_n)\right). \tag{5.35}$$

Ostatecznie w tym wypadku przybliżona wartość całki oznaczonej I jest równa:

$$\begin{aligned} I &= \int_a^b f(x)\cdot dx = \int_a^{b-h} f(x)\cdot dx + \int_{b-h}^{b} f(x)\cdot dx = \\ &= \frac{h}{3}\sum_{i=0}^{(n-1)/2-1}(f_{2i} + 4f_{2i+1} + f_{2i+2}) + \frac{h}{12}\left(-f_{n-2} + 8f_{n-1} + 5f_n\right). \end{aligned} \tag{5.36}$$

Inny wariant metody Simpsona można otrzymać, jeśli do interpolacji zastosujemy wielomian Lagrange'a 3. stopnia ($m = 3 - F_3(x) = a_0 + a_1\cdot x + a_2\cdot x^2 + a_3\cdot x^3$).

■ *Przykład 5.1*

Wykorzystując pomiary hydrometryczne wybranego przekroju poprzecznego cieku przedstawione na rysunku 5.1.1 oraz zestawione w tabeli 5.1.1, obliczyć pole przekroju czynnego kanału metodą trapezową.

Jak widać, przekrój kanału opisany jest za pomocą 9 punktów. Ich położenie definiują współrzędne (X_i, H_i), przy czym X_i jest odległością od lewego brzegu zaś H_i jest głębokością wody (rys. 5.1.1).

Tablica 5.1.1

Pomierzone głębokości wody w wybranych punktach przekroju poprzecznego cieku

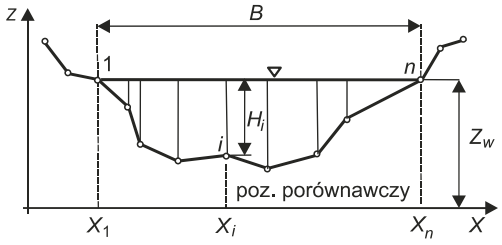
i	X_i [m]	H_i [m]
1	0,00	0,00
2	1,25	0,48
3	2,00	0,95
4	3,50	1,27
5	5,00	1,09
6	6,25	1,44
7	7,75	1,03
8	8,75	0,78
9	11,25	0,00

W tym wypadku całka (5.4) przybliżona formułą metody trapezowej przyjmie postać:

$$A = \int_0^B H(X)\cdot dX \cong \sum_{i=2}^{n}(X_i - X_{i-1})\cdot\frac{H_{i-1} + H_i}{2} \tag{5.1.1}$$

gdzie: n – liczba pomierzonych punktów,
X_i – odcięta punktu i,
H_i – głębokość w punkcie i,

B – szerokość cieku na poziomie zwierciadła wody.



Rys. 5.1.1. Schemat pomierzonego przekroju poprzecznego cieku

Po podstawieniu danych do wzoru (5.1.1) otrzymujemy:

$$A \cong (1{,}25-0{,}00)\frac{0{,}0+0{,}45}{2}+(2{,}00-1{,}25)\frac{0{,}45+0{,}95}{2}+(3{,}50-2{,}00)\frac{0{,}95+1{,}27}{2}+$$
$$+(5{,}00-3{,}50)\frac{1{,}27+1{,}09}{2}+(6{,}25-5{,}00)\frac{1{,}09+1{,}44}{2}+(7{,}75-6{,}25)\frac{1{,}44+1{,}03}{2}+$$
$$+(8{,}75-7{,}75)\frac{1{,}03+0{,}78}{2}+(11{,}25-8{,}75)\frac{0{,}78+0{,}00}{2}\cong 9{,}555\ \mathrm{m}^2$$

■

5.3. Numeryczne obliczanie wartości całek podwójnych

Całki podwójne występują w hydraulice i w hydrologii równie często jak omówione w poprzednim podrozdziale całki pojedyncze. Jako typowy przykład rozpatrzmy problem określenia natężenia przepływu kanale otwartym. Z wyjątkiem kanałów laboratoryjnych, natężenie przepływu w kanałach otwartych określamy metodą pośrednią, na podstawie pomierzonych prędkości przepływu w wybranych punktach jego przekroju poprzecznego. W tym celu wykorzystujemy jego definicję. Przez objętościowe natężenie przepływu w kanale rozumie się objętość wody, która przepływa w jednostce czasu przez jego przekrój poprzeczny zorientowany prostopadle do wektora prędkości. Można, więc napisać:

$$Q(x)=\iint_A u(x,y,z)\cdot \mathrm{d}A \tag{5.37}$$

gdzie: Q – natężenie przepływu,
u – lokalna prędkość przepływu,
A – pole powierzchni przekroju czynnego kanału.

Inną typową sytuacją w której pojawiają się całki podwójne jest proces eliminacji jednego z wymiarów, gdy zagadnienie dwuwymiarowe jest redukowane do zagadnienia jednowymiarowego. W takim przypadku zamiast niejednostajnego rozkładu danej wielkości w przekroju poprzecznym wprowadzamy rozkład jednostajny. Oznacza to stałą wartość danej wielkości w przekroju poprzecznym. W trakcie uśredniania należy przestrzegać następującej ogólnej zasady: całkowity strumień wielkości przez przekrój przy jej rzeczywistym rozkładzie musi być równy strumieniowi wynikającemu z wielkości uśrednionej. Zastosowanie tej zasady pozwala napisać wyrażenie na średnią prędkość przepływu w przekroju cieku: (rys. 5.11):

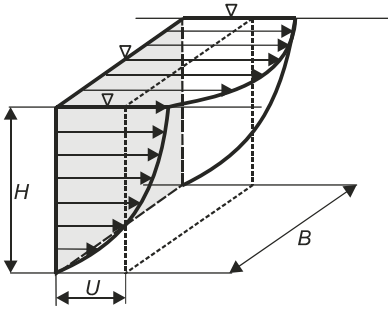
$$U(x) = \frac{Q}{A} = \frac{1}{A}\iint_A u(x, y, z)\cdot \mathrm{d}A \tag{5.38}$$

gdzie $U(x)$ jest średnią prędkością przepływu w przekroju.

Jeśli tak uśrednioną prędkość wprowadzimy do równania energii lub pędu płynącego strumienia wody, to odpowiednie bilanse wymienionych wielkości będą zakłócone. Aby skorygować wynikające z uśrednienia nieścisłości wprowadza się współczynniki korekcyjne. Są one definiowane następująco:

$$\alpha = \frac{\iint_A u^3 \cdot \mathrm{d}A}{U^3 \cdot A}, \qquad \beta = \frac{\iint_A u^2 \cdot \mathrm{d}A}{U^2 \cdot A} \tag{5.39a, b}$$

gdzie α jest współczynnikiem korygującym energię kinetyczną (współczynnik Coriolisa lub de Saint Venanta) zaś β jest współczynnikiem korygującym pęd. Jak można zauważyć obliczenie każdego z wymienionych współczynników wymaga obliczenia całki podwójnej.



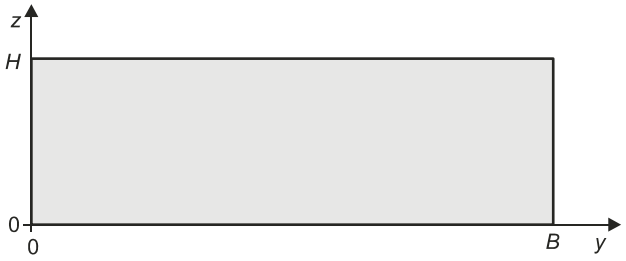
Rys. 5.11. Rozkład prędkości w przekroju poprzecznym kanału otwartego

Podobnie, całki podwójne muszą być obliczane w trakcie uśredniania w poprzecznym przekroju kanału takich wielkości skalarnych jak koncentracja rozpuszczonej w wodzie domieszki lub temperatura wody

$$C(x) = \frac{1}{A}\iint_A c(x, y, z)\mathrm{d}A, \qquad T(x) = \frac{1}{A}\iint_A \tau(x, y, z)\mathrm{d}A \tag{5.40}$$

gdzie: $C(x)$ – średnia koncentracja rozpuszczonej domieszki w przekroju cieku,
$c(x,y,z)$ – koncentracje pomierzona lokalnie,
$T(x)$ – średnia temperatura wody w przekroju cieku,
$\tau(x,y,z)$ – temperatura wody pomierzona lokalnie.

Rozpatrzmy przepływ w kanale otwartym. Jego przekrój ma kształt prostokąta o wymiarach $B \times H$ (rys. 5.12). Załóżmy, że w przekroju poprzecznym znane są prędkości przepływu $u(y, z)$. Zadanie polega na obliczeniu natężenia przepływu. W tym celu wykorzystujemy jego definicję wyrażoną wzorem (5.37). Prędkość przepływu należy scałkować po powierzchni prostokątnego przekroju czynnego przedstawionego na rysunku 5.12.



Rys. 5.12. Obszar całkowania w trakcie obliczania natężenia przepływu w kanale prostokątnym

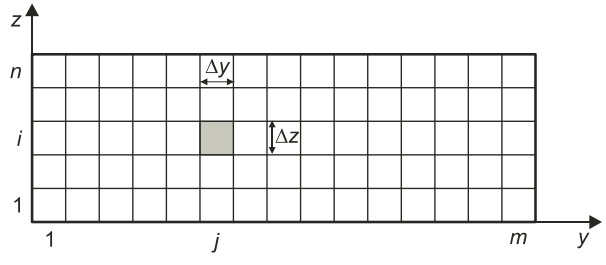
Dla prostokątnego obszaru całkowania o wymiarach $B \times H$, można napisać:

$$Q = \iint_A u(y,z)\cdot \mathrm{d}A = \int_0^B \left(\int_0^H u(y,z)\cdot dz \right) \cdot dy = \int_0^H \left(\int_0^B u(y,z)\cdot dy \right) \cdot dz \tag{5.41}$$

Kolejność obliczania całek pojedynczych nie jest istotna. Obydwie całki mogą być obliczone metodami przybliżonymi omówionymi wcześniej. W tym celu rozpatrywany obszar całkowania jest dzielony na mniejsze podobszary ze stałym krokiem przestrzennym Δy w kierunku y oraz w kierunku z ze stałym krokiem równym Δz. W rezultacie obszar ciągły pokrywa $N = n \times m$ komórek-podobszarów o wymiarach $\Delta z \times \Delta y$ jak na rysunku 5.13. Natężenie przepływu w kanale obliczane jest następująco:

$$Q = \iint_A u(y,z)\cdot \mathrm{d}A = \int_0^H \left(\int_0^B u(y,z)\cdot dy \right) \cdot dz \approx \sum_{i=1}^{n} \left(\sum_{j=1}^{m} u_{ij}^{*} \cdot \Delta y \right) \cdot \Delta z = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} u_{ij}^{*} \cdot \Delta A \right) \tag{5.42}$$

gdzie: m – indeks komórki odniesiony do kierunku y,
n – indeks komórki odniesiony do kierunku z,
ΔA – pole powierzchni komórki,
u_{ij}^{*} – prędkość przepływu reprezentatywna dla komórki (i, j).



Rys. 5.13. Prostokątny obszar całkowania pokryty przez $N = n \times m$ komórek o wymiarach $\Delta z \times \Delta y$

Formuła (5.42) jest ważna przy jednostajnym podziale przekroju kanału na komórki, gdy ΔA = const. Jednak w praktyce typową sytuacją jest pomiar prędkości w punktach rozłożonych w przekroju kanału w sposób nierównomierny. Zatem każdemu punktowi, w którym pomierzono prędkość przepływu przypisuje się komórkę o innej powierzchni. W takiej sytuacji natężenie przepływu obliczamy według nieco zmodyfikowanego wzoru:

$$Q \approx \sum_{k=1}^{N} u_k \cdot \Delta A_k \tag{5.43}$$

gdzie: k – indeks punktu, w którym pomierzono prędkość,
N – całkowita liczba punktów pomiarowych w rozpatrywanym przekroju kanału,
u_k – pomierzona prędkość przepływu w punkcie k,
ΔA_k – pole powierzchni komórki przypisanej punktowi k.

Analogiczne formuły stosujemy przy uśrednianiu koncentracji domieszki i temperatury wody na podstawie wykonanych pomiarów punktowych. Odpowiednie wzory mają następującą postać

$$C \approx \frac{1}{A} \sum_{k=1}^{N} c_k \cdot \Delta A_k, \qquad T \approx \frac{1}{A} \sum_{k=1}^{N} \tau_k \cdot \Delta A_k \tag{5.44, 5.45}$$

w których całkowite pole powierzchni przekroju czynnego kanału jest równe

$$A = \sum_{k=1}^{N} \Delta A_k \ . \tag{5.46}$$

W wymienionych wyżej przypadkach uśredniania dokładność obliczenia całki jest zdeterminowana przez wykonane punktowe pomiary uśrednianej wielkości. Jej zwiększenie można uzyskać poprzez zwiększenie liczby punktów pomiarowych w przekroju kanału.

6 Rozwiązywanie równań różniczkowych zwyczajnych

6.1. Wprowadzenie

W wielu dziedzinach techniki bardzo częsta jest sytuacja, gdy funkcja jednej zmiennej opisująca dany proces występuje w równaniu pod znakiem pochodnej. Chcąc określić tę funkcję, należy równanie scałkować. Innymi słowy, dane jest równanie różniczkowe zwyczajne, a celem postępowania jest znalezienie funkcji spełniającej je.

W praktyce inżynierskiej bardzo często równanie różniczkowe opisujące badany proces nie daje się scałkować żadnym ze znanych sposobów całkowania. W takiej sytuacji jedynym wyjściem jest skorzystanie z metod przybliżonego rozwiązywania równań różniczkowych. Stosując je, otrzymujemy poszukiwaną funkcję w postaci dyskretnej, tzn. w postaci zbioru liczb opisujących jej wartość w wybranych punktach przedziału całkowania.

Przyjmijmy, że dane jest równanie

$$y' = \frac{dy}{dx} = f(x, y) . \tag{6.1}$$

Jego rozwiązanie polega na znalezieniu funkcji $y(x)$, która w przedziale $\langle a, b \rangle$ spełni je. Jednak od funkcji $y(x)$ wymaga się ponadto, aby spełniała ona pewne dodatkowe warunki. W zagadnieniach technicznych przedmiotem poszukiwania jest zawsze jedna funkcja spełniająca równanie różniczkowe. Tymczasem wiadomo, że ogólne rozwiązanie równania (6.1) ma postać

$$y(x) = \int f dx + C . \tag{6.2}$$

Istnieje więc nieskończenie wiele rozwiązań różniących się stałą C. Chcąc otrzymać jedno konkretne rozwiązanie, na funkcję y nakłada się, oprócz warunku spełnienia równania (6.1), dodatkowe warunki. Istnieją dwie możliwości formułowania problemu rozwiązania różniące się sposobem formułowania warunków dodatkowych. Rozpatrzmy je kolejno.

Dane jest równanie różniczkowe zwyczajne (6.1)

$$y' = f(x, y). \tag{6.3}$$

Należy znaleźć funkcję $y(x)$, która w przedziale $\langle a, b \rangle$ spełni powyższe równanie, a na początku przedziału przyjmie zadaną wartość

$$y(a) = y_{a.} \tag{6.4}$$

Zatem spośród nieskończonej liczby rozwiązań równania, którymi są krzywe różniące się stałą całkowania C, wybieramy tylko tę, która przechodzi przez punkt początkowy o współrzędnych (a, y_a). Tak sformułowane zadanie rozwiązania równania (6.3) nazywa się *zagad-*

nieniem początkowym równania różniczkowego zwyczajnego. Zagadnienie to ma prawie zawsze rozwiązanie. Wystarczy, aby funkcja $f(x, y)$ była ciągła na płaszczyźnie $x - y$ oraz ograniczona.

Zagadnienie początkowe formułuje się także dla równań różniczkowych zwyczajnych wyższych rzędów. Na przykład dla równania drugiego rzędu

$$y'' = f(x, y, y') \tag{6.5}$$

brzmi ono następująco: należy znaleźć funkcję $y(x)$, która w przedziale $\langle a, b\rangle$ spełni powyższe równanie, a na początku przedziału zarówno funkcja, jak i jej pierwsza pochodna, przyjmie zadane wartości

$$y(a) = y_a, \tag{6.6}$$

$$y'(a) = y'_a. \tag{6.7}$$

Ogólnie można powiedzieć, że w zagadnieniu początkowym równania różniczkowego zwyczajnego n-tego rzędu należy zadać n warunków, tzn. na funkcję i jej pochodne do $n - 1$ rzędu włącznie.

Drugi sposób formułowania problemu rozwiązania równań różniczkowych zwyczajnych dotyczy równań wyższych rzędów lub układów równań. Na przykład dla równania II rzędu (4.5) brzmi on następująco: należy znaleźć funkcję $y(x)$, która w przedziale $\langle a, b\rangle$ spełni wymienione równanie różniczkowe, a na końcach przedziału przyjmie zadane wartości:

$$y(a) = y_a \quad \text{oraz} \quad y(b) = y_b. \tag{6.8}$$

Zatem rozwiązanie $y(x)$ musi przejść przez punkty o współrzędnych (a, y_a) oraz (b, y_b). Tak sformułowane zagadnienie nosi nazwę *zagadnienia brzegowego równania różniczkowego zwyczajnego*. W przeciwieństwie do zagadnienia początkowego, którego rozwiązanie istnieje prawie zawsze, znalezienie funkcji spełniającej równanie oraz narzucone warunki na obu brzegach przedziału $\langle a, b\rangle$ może być niemożliwe.

W inżynierii wodnej równania różniczkowe zwyczajne występują w opisie wielu wypadków przepływu wody, przy czym formułowane są dla nich zarówno zagadnienia początkowe, jak i brzegowe. Szczególnie często występuje problem początkowy układu równań różniczkowych zwyczajnych, jako etap rozwiązywania różniczkowych równań cząstkowych typu hiperbolicznego i parabolicznego metodą elementów skończonych.

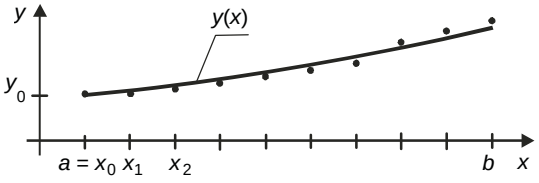
6.2. Numeryczne rozwiązywanie zagadnień początkowych równań różniczkowych zwyczajnych

W zagadnieniu początkowym równania różniczkowego typu (6.3) chodzi o obliczenie dla x w pewnym przedziale $x_0 \le x \le b$ takiego rozwiązania $y(x)$, które przyjmie w punkcie $x = x_0$ zadaną wartość początkową y_0. Zakłada się przy tym istnienie i jednoznaczność tego rozwiązania. Stosowane w tym celu metody numeryczne umożliwiają obliczenie przybliżenia $y_1, y_2, ..., y_k, ...$ wartości rozwiązania dokładnego $y(x_1), y(x_2), ..., y(x_k)$ w wybranych punktach $x_1, x_2, ..., x_k$ przedziału $\langle x_0, b\rangle$. O wyborze metody obliczeń decyduje w głównej mierze liczba kroków oraz żądana dokładność rozwiązania.

Niech będzie dane równanie różniczkowe zwyczajne rzędu pierwszego (6.1)

$$y' = \frac{dy}{dx} = f(x, y). \tag{6.9}$$

Przyjmuje się, że funkcja $f(x, y)$ jest ciągła w pewnym obszarze D zawartym w płaszczyźnie zmiennych rzeczywistych x, y. Ponadto zakłada się, że funkcja $f(x,y)$ jest ograniczona w tym obszarze oraz że spełnia warunek Lipschitza (Krupowicz, 1986; Legras, 1974; Ralston, 1971). W obszarze tym szukamy takiego rozwiązania $y(x)$ równania (6.9), które przejdzie przez dany punkt początkowy o współrzędnych $x = x_0$, $y = y_0$. W tym wypadku będziemy oczywiście szukać rozwiązania przybliżonego, tzn. przybliżeń y_j wartości $y(x_j)$ rozwiązania ścisłego w punktach $x_j (j = 1, 2, 3, ...)$ (rys. 6.1). Różnica $\Delta x_j = x_{j+1} - x_j$ nosi nazwę *długości kroku całkowania* i oznacza się ją zwykle symbolem h.



Rys. 6.1. Funkcja $y(x)$ i jej numeryczne przybliżenie

Większość stosowanych metod opiera się na równaniu powstającym z (6.9) przez jego scałkowanie

$$\int_{y(x_j)}^{y(x_{j+1})} dy = \int_{x_j}^{x_{j+1}} f(x, y)dx . \tag{6.10}$$

Prowadzi to do formuły

$$y(x_{j+1}) = y(x_j) + \int_{x_j}^{x_{j+1}} f(x, y(x))dx , \tag{6.11}$$

którą można zapisać następująco:

$$y_{j+1} = y_j + \int_{x_j}^{x_{j+1}} f(x, y(x))dx . \tag{6.12}$$

Całkę po prawej stronie (6.12) zastępuje się wyrażeniem przybliżającym. W zależności od sposobu jej obliczania, otrzymuje się odpowiedniego typu metodę przybliżonego całkowania równań różniczkowych zwyczajnych.

Stosowane metody rozwiązywania równań różniczkowych dzieli się na dwie zasadnicze grupy:

1) metody jednokrokowe,
2) metody wielokrokowe.

Metodą jednokrokową nazywa się taką metodę, w której przybliżoną wartość funkcji y_{j+1} oblicza się tylko na podstawie wartości y z poprzedniego kroku, czyli y_j. Metodą wielokrokową jest metoda, w której do obliczenia y_{j+1} konieczna jest znajomość co najmniej dwóch wcześniejszych wartości, czyli $y_j, y_{j-1}, y_{j-2}, ... , y_{j-l}$.

Metody rozwiązywania równań różniczkowych zwyczajnych mogą być jawne – jeśli wyrażenie na y_{j+1} jest w postaci $y_{j+1} = G(x, y_j, y_{j-1}, y_{j-2} ..., y_{j-l})$, lub niejawne – jeśli dają zależność typu $G(x, y_{j+1}, y_j, y_{j-1}, y_{j-2} ..., y_{j-l}) = 0$, czyli w postaci równania nieliniowego względem y_{j+1}.

6.2.1. Metody jawne jednokrokowe

Istnieje wiele jawnych metod jednokrokowych rozwiązywania równań różniczkowych zwyczajnych. Ich przegląd podają Ralston (1971), Krupowicz (1986) i inni. Dalej omawia się tylko niektóre metody.

Rozwinięcie w szereg Taylora

Jeśli znany jest punkt należący do krzywej $y(x)$ o współrzędnych (x_j, y_j), można obliczyć przybliżoną wartość y w punkcie $x_{j+1} = x_j + h$, wykorzystując rozwinięcie tej funkcji w szereg Taylora:

$$y(x_{j+1}) = y(x_j) + hy'(x_j) + \frac{h^2}{2} y''(x_j) + \ldots + \frac{h^p}{p!} y^{(p)}(x_j) + \ldots , \tag{6.13}$$

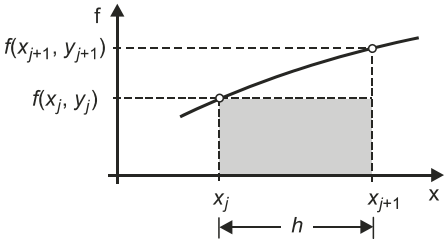
gdzie: $y(x_j) = y_j, y'(x_j) = f(x_j, y_j), y''(x_j) = f_x'(x_j, y_j) + f(x_j, y_j) f_y'(x_j, y_j)$ itd.

Metoda ta jest rzadko stosowana ze względu na konieczność obliczania pochodnych wyższych rzędów.

Metoda Eulera (łamanych)

Metoda ta polega na zastąpieniu krzywej w przedziale h jej styczną w punkcie (x_j, y_j). Otrzymuje się więc

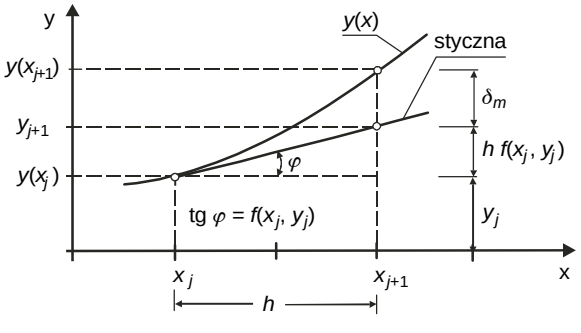
$$\begin{aligned} x_{j+1} &= x_j + h, \\ y_{j+1} &= y_j + hf(x_j, y_j). \end{aligned} \tag{6.14}$$



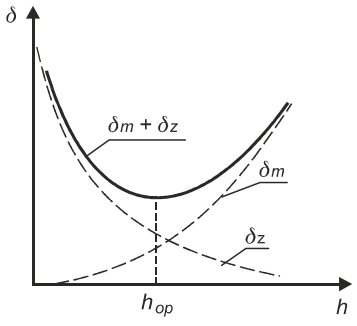
Rys. 6.2. Zastąpienie pola pod krzywą $f(x, y)$ polem prostokąta – metoda Eulera

Widać, że jest to metoda, którą można wyprowadzić z poprzedniej, pomijając w rozwinięciu Taylora człony z wyższymi pochodnymi niż pierwsza. Identyczną zależność otrzymuje się, gdy w (6.12) zastąpimy całkę polem prostokąta (rys. 6.2), czyli

$$y_{j+1} = y_j + hf(x_j, y_j) . \tag{6.15}$$



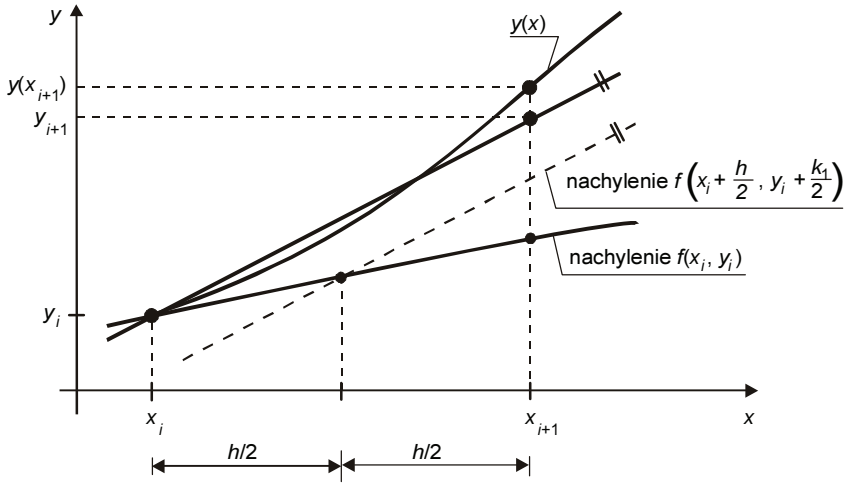
Rys. 6.3. Geometryczna interpretacja metody Eulera



Rys. 6.4. Zależność błędów od wielkości kroku całkowania

Interpretację geometryczną tej metody przedstawiono na rys. 6.3. Krzywą $y(x)$ zastępujemy odcinkiem prostej, która jest styczną do krzywej w punkcie x_j. Przybliżenie y_{j+1} wartości funkcji $y(x_{j+1})$ oblicza się, dodając do wartości poprzedniej przyrost $h \cdot f(x_j, y_j)$. Przybliżenie to jest tym lepsze, im mniejszy jest krok h, co ilustruje rys. 6.3. Zależność błędu metody od wielkości h jest funkcją rosnącą, co schematycznie przedstawiono na rys. 6.4, oznaczając błąd przez δ_m. Jednakże zmniejszanie kroku całkowania h powoduje wzrost ilości obliczeń, a zatem i wzrost błędu wynikającego z zaokrągleń, oznaczonego przez δ_z. Gdy h jest dostatecznie małe, błąd zaokrągleń może przeważać i dalsze zmniejszanie h może powodować nie poprawę, a pogorszenie wyniku obliczeń, gdyż całkowity błąd jest sumą błędu metody i błędu zaokrągleń. Wynika z tego, że istnieje pewna optymalna wartość h_{op}, dla której całkowity błąd jest minimalny.

Powyższe uwagi o błędach odnoszą się do wszystkich metod całkowania.



Rys. 6.5. Geometryczna interpretacja ulepszonej metody Eulera

Metoda Eulera ulepszona

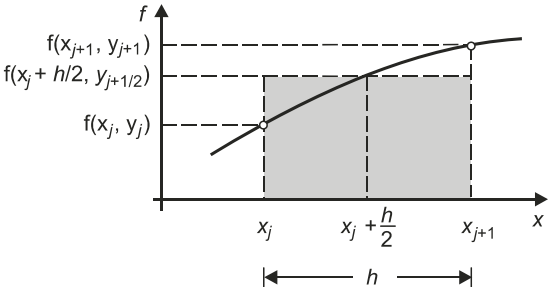
W metodzie tej zastępuje się krzywą $y(x)$ odcinkiem prostej, który jest do niej styczny w połowie przedziału $\langle x_j, x_{j+1}\rangle$, zamiast – jak w poprzedniej metodzie – w punkcie (x_j, y_j). Do obliczenia przybliżonej wartości y w punkcie $x_j + h/2$ wykorzystuje się metodę Eulera, czyli

$$\begin{aligned} y_{j+1/2} &= y_j + \frac{h}{2} f(x_j, y_j), \\ y_{j+1} &= y_j + hf\left(x_j + \frac{h}{2}, y_{j+1/2}\right), \end{aligned} \tag{6.16}$$

lub oznaczając inaczej

$$k_1 = hf(x_j, y_j),$$
$$k_2 = hf\left(x_j + \frac{h}{2}, y_j + \frac{k_1}{2}\right), \tag{6.17}$$
$$y_{j+1} = y_j + k_2.$$

Z kolei biorąc pod uwagę wzór (6.12), łatwo można zauważyć, że analogiczną formułę do (6.17) daje on, gdy całkę zastąpimy polem prostokąta o wysokości $f(x_j + h/2, y_j + k_1/2)$, co przedstawiono na rys. 6.6.



Rys. 6.6. Obliczanie całki w (4.12) w ulepszonej metodzie Eulera

Metoda Rungego-Kutty

Metoda ta jest algebraicznym uogólnieniem ulepszonej metody Eulera. Polega ona na znalezieniu na y_{j+1} wyrażenia zgodnego z rozwinięciem w szereg Taylora, aż do pewnej potęgi h, bez obliczania pochodnych. Przyjmując zgodność wyrazów zawierających h, aż do potęgi p, mamy:

$$y_{j+1} = y_j + \alpha_1 k_1 + \alpha_2 k_2 + \ldots + \alpha_p k_p, \tag{6.18}$$

gdzie: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf(x_j + \mu_1 h, \quad y_j + \lambda_{11} k_1),$$
$$k_3 = hf(x_j + \mu_2 h, \quad y_j + \lambda_{21} k_1 + \lambda_{22} k_2),$$
$$\ldots$$
$$k_p = hf(x_j + \mu_{p-1} h, \quad y_j + \lambda_{p-1,1} k_1 + \lambda_{p-1,2} k_2 + \ldots + \lambda_{p-1,p-1} k_{p-1}),$$

gdzie α_i, μ_i, λ_{ij} są współczynnikami, które należy określić. Sposób określenia tych współczynników można prześledzić na przykładzie, w którym zostaną uwzględnione wyrazy rozwinięcia do h^2 włącznie, czyli dla $p = 2$. W tym wypadku formuły (6.18) przyjmą postać:

$$y_{j+1} = y_j + \alpha_1 k_1 + \alpha_2 k_2, \tag{6.19}$$

przy czym: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf(x_j + \mu h, y_j + \lambda k_1).$$

Wstawienie do (6.19) wyrażenia na k_1 oraz zastąpienie wyrażenia na k_2 rozwinięciem w szereg Taylora wokół punktu (x_j, y_j) z zachowaniem pochodnych I rzędu prowadzi do formuły

$$y_{j+1} = y_j + \alpha_1 hf(x_j, y_j) + \alpha_2 hf(x_j, y_j) + \alpha_2 \mu h^2 f'_x(x_j, y_j) + \\ + \alpha_2 \lambda h k_1 f'_y(x_j, y_j).$$

Po jej uporządkowaniu otrzymujemy

$$y_{j+1} = y_j + (\alpha_1 + \alpha_2) hf(x_j, y_j) + \alpha_2 h^2 \left[\mu f'_x(x_j, y_j) + \lambda \frac{k_1}{h} f'_y(x_j, y_j) \right] = \\ = y_j + (\alpha_1 + \alpha_2) hf(x_j, y_j) + h^2 [\alpha_2 \mu f'_x(x_j, y_j) + \alpha_2 \lambda f(x_j, y_j) f'_y(x_j, y_j)]. \tag{6.20}$$

Z drugiej strony, z rozwinięcia Taylora wynika wyrażenie

$$y_{j+1} = y_j + hy'_j + \frac{h^2}{2} y''_j = \\ = y_j + hf(x_j, y_j) + \frac{h^2}{2} [f'_x(x_j, y_j) + f(x_j, y_j) f'_y(x_j, y_j)]. \tag{6.21}$$

Porównując ze sobą (4.20) i (4.21), otrzymuje się zależności

$$\alpha_1 + \alpha_2 = 1, \quad \alpha_2 \mu = \frac{1}{2}, \quad \alpha_2 \lambda = \frac{1}{2}, \tag{6.22a,b,c}$$

z których wynika, że:

$$\mu = \lambda, \quad \alpha_1 = \frac{2\mu - 1}{2\mu}, \quad \alpha_2 = \frac{1}{2\mu}. \tag{6.23a,b,c}$$

Przyjmując $\mu = 1/2$, otrzymujemy $\alpha_1 = 0$ i $\alpha_2 = 1$. Zależność (6.19) przyjmie więc postać

$$y_{j+1} = y_j + k_2, \tag{6.24}$$

gdzie: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}k_1\right).$$

Są to formuły identyczne z (6.17).

Z kolei przyjmując w (6.23) $\mu = 1$, otrzymujemy $\alpha_1 = 1/2$ i $\alpha_2 = 1/2$, czyli

$$y_{j+1} = y_j + \frac{1}{2}(k_1 + k_2), \tag{6.25}$$

gdzie: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf(x_j + h, y_j + k_1).$$

Ten wariant metody Rungego-Kutty nazywa się również metodą Eulera-Cauchy.

W podobny sposób, uwzględniając kolejne wyrazy rozwinięcia Taylora, można uzyskać szereg wariantów tej metody (Legras, 1974). I tak mamy:

— uwzględniając h^3

$$y_{j+1} = y_j + \frac{1}{6}(k_1 + 4k_2 + k_3), \tag{6.26}$$

gdzie: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}k_1\right),$$

$$k_3 = hf(x_j + h, \quad y_j + k_1 - 2k_2);$$

— uwzględniając h^4

$$y_{j+1} = y_j + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \tag{6.27}$$

gdzie: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}k_1\right),$$

$$k_3 = hf\left(x_j + \frac{1}{2}h, \quad y_j + \frac{1}{2}k_2\right),$$

$$k_4 = hf(x_j + h, \quad y_j + k_3);$$

— uwzględniając h^5

$$y_{j+1} = y_j + \frac{1}{13770}(1445k_1 + 6561k_3 + 3264k_4 + 2500k_5), \tag{6.28}$$

gdzie: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf\left(x_j + \frac{2}{9}h, \quad y_j + \frac{2}{9}k_1\right),$$

$$k_3 = hf\left(x_j + \frac{1}{3}h, \quad y_j + \frac{1}{12}(k_1 + 3k_2)\right),$$

$$k_4 = hf\left(x_j + \frac{3}{4}h, \quad y_j + \frac{3}{128}(23k_1 - 81k_2 + 90k_3)\right),$$

$$k_5 = hf\left(x_j + \frac{9}{10}h, \quad y_j + \frac{9}{10000}(-345k_1 + 2025k_2 - 1224k_3 + 544k_4)\right);$$

— uwzględniając h^6

$$y_{j+1} = y_j + \frac{1}{192}(23k_1 + 125k_2 - 81k_5 + 125k_6), \tag{6.29}$$

gdzie: $k_1 = hf(x_j, y_n)$,

$$k_2 = hf\left(x_j + \frac{1}{3}h, \quad y_j + \frac{1}{3}k_1\right),$$

$$k_3 = hf\left(x_j + \frac{2}{5}h, \quad y_j + \frac{1}{25}(6k_2 + 4k_1)\right),$$

$$k_4 = hf\left(x_j + h, \quad y_j + \frac{1}{4}(k_1 - 12k_2 + 15k_3)\right),$$

$$k_5 = hf\left(x_j + \frac{2}{3}h, \quad y_j + \frac{1}{81}(6k_1 + 90k_2 - 50k_3 + 8k_4)\right),$$

$$k_6 = hf\left(x_j + \frac{4}{5}h, \quad y_j + \frac{1}{75}(6k_1 + 36k_2 + 10k_3 + 8k_4) \right);$$

— uwzględniając h^8

$$y_{j+1} = y_j + \frac{1}{840}(41k_1 + 216k_3 + 27k_4 + 272k_5 + 27k_6 + 216k_7 + 41k_8), \qquad (6.30)$$

gdzie: $k_1 = hf(x_j, y_j)$,

$$k_2 = hf\left(x_j + \frac{1}{9}h, \quad y_j + \frac{1}{9}k_1 \right),$$

$$k_3 = hf\left(x_j + \frac{1}{6}h, \quad y_j + \frac{1}{24}(k_1 + 3k_2) \right),$$

$$k_4 = hf\left(x_j + \frac{1}{3}h, \quad y_j + \frac{1}{6}(k_1 - 3k_2 + 4k_3) \right),$$

$$k_5 = hf\left(x_j + \frac{1}{2}h, \quad y_j + \frac{1}{8}(-5k_1 + 27k_2 - 24k_3 + 6k_4) \right),$$

$$k_6 = hf\left(x_j + \frac{2}{3}h, \quad y_j + \frac{1}{9}(221k_1 - 981k_2 + 867k_3 - 102k_4 + k_5) \right),$$

$$k_7 = hf\left(x_j + \frac{5}{6}h, \quad y_j + \frac{1}{48}(-183k_1 + 678k_2 - 472k_3 - 66k_4 + 80k_5 + 3k_6) \right),$$

$$k_8 = hf\left(x_j + h, y_j + \frac{1}{82}(716k_1 - 2072k_2 + 1002k_3 + 834k_4 - 454k_5 - 9k_6 + 72k_7) \right).$$

6.2.2. Metody jawne wielokrokowe

Całkując równanie $y = f(x, y)$, otrzymuje się znaną już zależność (6.12), czyli

$$y_{j+1} = y_j + \int_{x_j}^{x_{j+1}} f(x, y(x))dx. \qquad (6.31)$$

Funkcja zmiennej niezależnej x – $f(x, y(x))$ nieznana jest w przedziale $\langle x_j, x_{j+1} \rangle$, ale znana jest w punktach $x_0, x_1, x_2, ..., x_j$. Oznacza to, że znane są wartości $f(x_0, y(x_0))$, $f(x_1, y(x_1))$, $f(x_2, y(x_2))$, ..., $f(x_j, y(x_j))$. Można więc zastosować wielomian interpolacyjny Lagrange'a w celu przybliżenia funkcji $f(x, y(x))$ między węzłami x_j i x_{j+1}. Jeśli dokona się ekstrapolacji funkcji f na ten przedział, to otrzymuje się proste wyrażenie na y_{j+1} typu $y_{j+1} = F(x, y_0, y_1, ..., y_j)$, czyli tak zwany schemat jawny.

Ogólna postać formuły jawnej jest następująca (Ralston, 1971):

$$y_{j+1} = y_{j-p} + h\sum_{i=0}^{q} \beta_i f_{j-i}. \qquad (6.32)$$

Stosując wielomiany interpolacyjne Lagrange'a, można otrzymać szereg wariantów w zależności od wartości p i q. W szczególnym przypadku, przyjmując $p = 0$, otrzymujemy grupę formuł typu Adamsa-Bashfortha, które dla różnych wartości q mają postać:

$$q=1 \quad y_{j+1}=y_j+\frac{h}{2}(3f_j-f_{j-1}), \tag{6.33a}$$

$$q=2 \quad y_{j+1}=y_j+\frac{h}{12}(23f_j-16f_{j-1}+5f_{j-2}), \tag{6.33b}$$

$$q=3 \quad y_{j+1}=y_j+\frac{h}{24}(55f_j-59f_{j-1}+37f_{j-2}-9f_{j-3}). \tag{6.33c}$$

Sposób obliczenia współczynników formuły ekstrapolacyjnej przedstawia się dla $q = 2$, czyli dla wypadku, kiedy dokonuje się interpolacji za pomocą wielomianu Lagrange'a 2. stopnia. Wielomian ten, jak wiadomo z podrozdziału 3.3, można zapisać w postaci

$$P(x) = \mathbf{X}\,\mathbf{L}\,\mathbf{F}. \tag{6.34}$$

Przyjmując, że węzłem początkowym jest punkt o współrzędnych (x_{j-2}, y_{j-2}), otrzymujemy

$$P(x)=\left[1,\ \frac{x}{h},\ \frac{x^2}{h^2}\right]\begin{bmatrix}1 & 0 & 0\\ -\frac{3}{2} & 2 & -\frac{1}{2}\\ \frac{1}{2} & -1 & \frac{1}{2}\end{bmatrix}\begin{Bmatrix}f_{j-2}\\ f_{j-1}\\ f_j\end{Bmatrix}. \tag{6.35}$$

Całkując ten wielomian w przedziale między x_j i x_{j+1}, tzn. w granicach od $2h$ do $3h$, ponieważ początek został sprowadzony do punktu $j - 2$, otrzymuje się

$$\int_{2h}^{3h} dx = h, \quad \int_{2h}^{3h} \frac{x}{h}dx = \frac{5}{2}h, \quad \int_{2h}^{3h} \frac{x^2}{h^2}dx = \frac{19}{3}h,$$

czyli

$$\int_{2h}^{3h} P(x)dx = h\left[\frac{5}{12},\ -\frac{4}{3},\ \frac{23}{12}\right]\begin{Bmatrix}f_{j-2}\\ f_{j-1}\\ f_j\end{Bmatrix}.$$

Ostatecznie równanie (4.31) przyjmie postać

$$y_{j+1}=y_j+\frac{h}{12}(5f_{j-2}-16f_{j-1}+23f_j). \tag{6.36}$$

W podobny sposób można wyprowadzić inne formuły interpolacyjne.

■ *Przykład 6.1*

Transformacja fali wezbraniowej przez pojedynczy zbiornik

Pojedynczy zbiornik o liniowej charakterystyce transformuje falę wezbraniową zgodnie z równaniem (Szymkiewicz, 2000)

$$\frac{dQ}{dt}=\frac{1}{K}(P(t)-Q(t)), \tag{6.1.1}$$

gdzie: $P(t)$ – znana funkcja zależna od czasu, reprezentująca natężenie dopływu do zbiornika,
$Q(t)$ – poszukiwana funkcja czasu reprezentująca natężenie odpływu ze zbiornika,
K – stała retencji.

Należy wyznaczyć funkcję $Q(t)$ dla $t \in \langle 0, N \cdot h \rangle$, rozwiązując powyższe równanie metodą ekstrapolacyjną Adamsa (6.33b) i wiedząc, że:

— warunek początkowy ma postać

$$Q(t = 0) = P_0; \tag{6.1.2}$$

— dopływ do zbiornika zmienia się w czasie zgodnie z równaniem

$$P(t) = P_0 + P_m \left(\frac{t}{T_m} \right)^2 \exp\left(1 - \left(\frac{t}{T_m} \right)^2 \right),$$

gdzie: P_0, P_m, T_m – zadane parametry;

— krok całkowania wynosi h, a liczba kroków – N.

Zalecona do rozwiązania formuła Adamsa ma postać

$$Q_{j+1} = Q_j + \frac{h}{12}(5Q'_{j-2} - 16Q'_{j-1} + 23Q'_j)\,. \tag{6.1.3}$$

Jak widać, do rozpoczęcia obliczeń wymagana jest znajomość Q'_j w punktach $j = 1, 2, 3$. Do wyznaczenia ich zastosujemy metodę jednokrokową Eulera (ulepszoną), którą zgodnie z (6.17) zapiszemy

$$Q_{j+1} = Q_j + hQ'(t_j + 0{,}5h, Q_j + 0{,}5k_1)\,, \tag{6.1.4}$$

gdzie: $k_1 = h\,Q'(t_j, Q_j)$.

Metodę tę należy stosować dla dwóch pierwszych kroków, czyli dla $j = 1, 2$, gdyż Q_1 znane jest jako warunek początkowy.

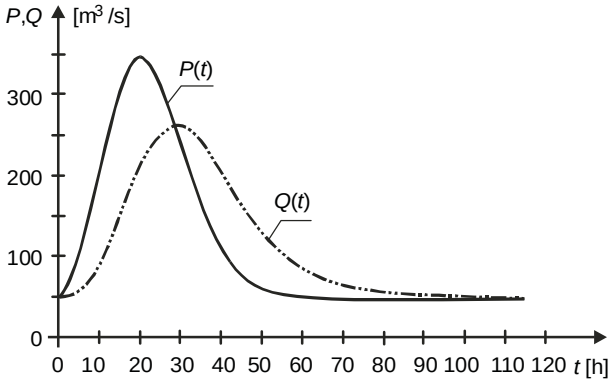
Algorytm rozwiązania opisanego wcześniej zagadnienia można przedstawić w następującej postaci:

1) podstaw: $t_1 = 0$, $Q_1 = P_0$;
2) dla $j = 1, 2$ powtórz operacje:
   – podstaw $t_{j+1} = t_j + h$,
   – oblicz Q_{j+1} według wzoru (6.1.4);
3) dla $j = 3, 4, ..., N - 1$ powtórz operacje:
   – podstaw $t_{j+1} = t_j + h$,
   – oblicz Q_{j+1} według wzoru (6.1.3).

Działanie algorytmu zilustrowano przykładem obliczeń. Przyjęto, że:

— krok całkowania wynosi $h = 1$ h,
— stała retencji wynosi $K = 12$ h,
— parametry funkcji $P(t)$ są równe: $P_0 = 50$ m^3/s, $P_m = 300$ m^3/s, $t_m = 20$ h,
— liczba kroków całkowania wynosi $N = 120$.

Obliczoną funkcję $Q(t)$ przedstawiono na rys. 6.1.1, na którym dla porównania zaznaczono również funkcję $P(t)$.



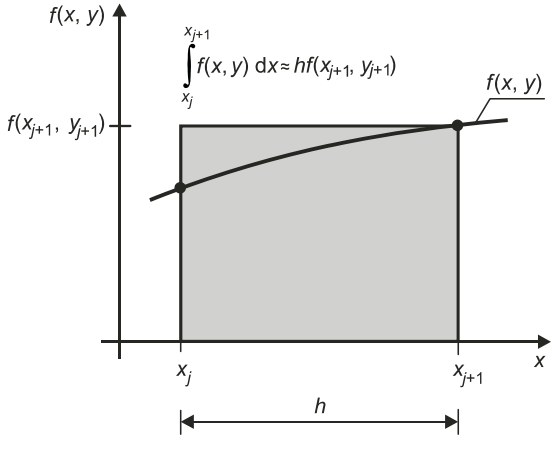
Rys. 6.1.1. Funkcja $Q(t)$ otrzymana jako wynik transformacji funkcji $P(t)$ przez zbiornik ■

6.2.3. Metody niejawne jednokrokowe

Do tej grupy metod rozwiązania równań różniczkowych zwyczajnych zalicza się niejawną metodę Eulera oraz niejawną metodę trapezową. Obydwie metody można łatwo wyprowadzić, wykorzystując ogólny wzór (6.12).

Metodę niejawną Eulera otrzymujemy jako wynik przybliżenia pola trapezu krzywoliniowego polem prostokąta o wymiarach $h \cdot f(x_{j+1}, y_{j+1})$ (rys. 4.7), czyli

$$\int_{x_j}^{x_{j+1}} f(x, y(x))\,dx \approx h \cdot f(x_{j+1}, y_{j+1})\,. \tag{6.37}$$



Rys. 6.7. Zastąpienie pola pod krzywą $f(x, y)$ polem prostokąta – niejawna metoda Eulera

Równanie (6.12) przyjmie więc postać

$$y_{j+1} = y_j + h \cdot f(x_{j+1}, y_{j+1}). \tag{6.38}$$

Jeśli w równaniu (6.12) całkę zastąpimy polem trapezu prostoliniowego (rys. 6.8)

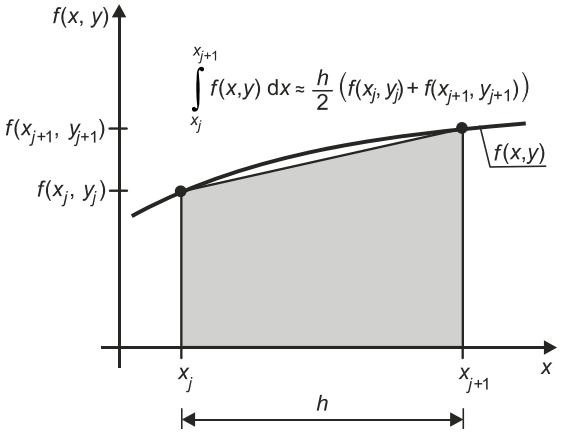
$$\int_{x_j}^{x_{j+1}} f(x, y(x))dx \approx \frac{h}{2}(f(x_j, y_j) + f(x_{j+1}, y_{j+1})), \tag{6.39}$$

to otrzymamy niejawną metodę trapezową

$$y_{j+1} = y_j + \frac{h}{2}(f(x_j, y_j) + f(x_{j+1}, y_{j+1})). \tag{6.40}$$

Jak widzimy, formuły metody niejawnej Eulera oraz niejawnej metody trapezów prowadzą do algebraicznych równań nieliniowych. Zatem w każdym kroku całkowania poszukiwaną wartość y_{j+1} otrzymamy, rozwiązując nieliniowe równanie algebraiczne dowolną metodą opisaną w rozdziale 2. Najczęściej stosowany wariant metody iteracji prostej, nazywany algorytmem „predyktor-korektor", przedstawiono w punkcie 6.2.4.

Obydwie wymienione metody należą do najczęściej stosowanych metod rozwiązywania układów równań różniczkowych zwyczajnych.



Rys. 6.8. Zastąpienie pola pod krzywą $f(x, y)$ polem trapezu – niejawna metoda trapezowa

■ *Przykład 6.2*

Czas opróżniania zbiornika retencyjnego

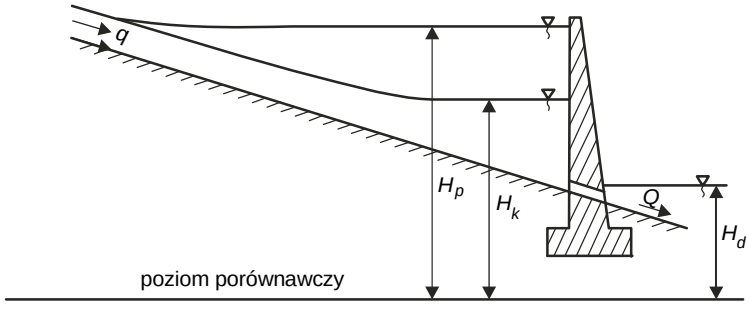
Obliczyć czas konieczny do przygotowania zbiornika retencyjnego na przyjęcie fali wezbraniowej, to znaczy czas, w którym zwierciadło wody w zbiorniku $H(t)$ z początkowego poziomu H_p (rys. 6.2.1) obniży się do poziomu pożądanego H_k.

Obliczenie należy wykonać, przyjmując następujące założenia i dane:

— początkowy poziom wody w zbiorniku wynosi H_p = 15 m npp, zaś poziom, do którego należy obniżyć zwierciadło wody, wynosi H_k = 10 m npp;

— dopływ wody do zbiornika nie zmienia się w czasie i wynosi q = 3 m^3/s;

— odpływ wody odbywa się przez upust denny przy jego całkowitym otwarciu. Jego powierzchnia wynosi $A = 2{,}50$ m^2, zaś współczynnik wydatku $\varphi = 0{,}65$;
— poziom wody na stanowisku dolnym jest stały i wynosi $H_d = 5$ m npp;
— powierzchnię zbiornika na poziomie zwierciadła wody, będącą funkcją napełnienia zbiornika $F(H)$, definiuje tabela 6.2.1 i rys. 6.2.2.



Rys. 6.2.1. Schemat zbiornika retencyjnego

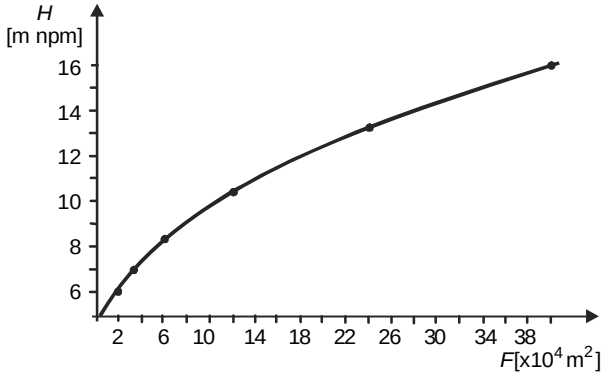
Tabela 6.2.1

Zależność powierzchni zbiornika od jego napełnienia

H_i [m npp]	5,00	6,00	7,40	8,30	10,40	13,25	16,00
F_i [m^2]	1000	20000	40000	60000	120000	240000	400000

Dla pośrednich wartości H wartość $F(H)$ oblicza się, interpolując liniowo pomiędzy punktami. Zatem $F(H)$ zdefiniowana jest następująco:

$$F(H)=\begin{cases} F_1 & \text{dla} \quad H \le H_1 \\ F_i + \dfrac{F_{i+1}-F_i}{H_{i+1}-H_i}(H-H_i) & \text{dla} \quad H_i < H < H_{i+1} (i=1,2,3,4,5,6,7) \\ F_6 & \text{dla} \quad H \ge H_6 \ . \end{cases} \tag{6.2.1}$$



Rys. 6.2.2. Zależność powierzchni zbiornika od jego napełnienia

Dla zbiornika przedstawionego na rys. 6.2.1 można napisać następujące różniczkowe równanie retencji:

$$\frac{dH}{dt} = \frac{1}{F(H)}(q(t) - Q(t)), \tag{6.2.2}$$

w którym chwilowy odpływ $Q(t)$ jest funkcją różnicy poziomów wody w zbiorniku i poniżej zapory:

$$Q(t) = \varphi \cdot A \cdot \sqrt{2g}\ (H(t) - H_d)^{1/2}. \tag{6.2.3}$$

Do rozwiązania zastosujmy niejawny schemat Eulera (6.38), który ma postać:

$$H_{j+1} = H_j + \Delta t\, H'_{j+1}, \tag{6.2.4}$$

gdzie: j – indeks punktu obliczeniowego,
Δt – krok całkowania w czasie.

Podstawiając przyjęte dane i zależności, otrzymujemy:

$$H_{j+1} = H_j + \frac{\Delta t}{F(H_{j+1})}\left(q - \varphi A\sqrt{2g}\,(H_{j+1} - H_d)^{1/2}\right). \tag{6.2.5}$$

Ponieważ zastosowany schemat całkowania jest niejawny, otrzymane równanie jest nieliniowe. Zastosujmy więc metodę iteracji prostej. Załóżmy jednocześnie, że pierwsze przybliżenie obliczymy jawną metodą Eulera (6.15):

$$H_{j+1}^{(k=0)} = H_j + \frac{\Delta t}{F(H_j)}\left(q - \varphi A\sqrt{2g}\,(H_j - H_d)^{1/2}\right). \tag{6.2.6}$$

Tak obliczony wynik korygujemy, stosując (6.2.5) w postaci:

$$H_{j+1}^{(k+1)} = H_j + \frac{\Delta t}{F(H_{i+1}^{(k)})}\left(q - \varphi A\sqrt{2g}\,(H_{j+1}^{(k)} - H_d)^{1/2}\right), \tag{6.2.7}$$

gdzie: $k = 0, 1, 2, \ldots$ – indeks iteracji.

Proces kontynuujemy do chwili, gdy różnica wyników w dwóch kolejnych iteracjach spełni warunek:

$$\left|H_{j+1}^{(k+1)} - H_{j+1}^{(k)}\right| \le \varepsilon, \tag{6.2.8}$$

gdzie: ε – przyjęta dokładność rozwiązania.

Zatem odpowiedź na pytanie o czas opróżnienia zbiornika otrzymujemy, rozwiązując równanie różniczkowe zwyczajne (6.2.2) z warunkiem początkowym:

$$H(t = 0) = H_p. \tag{6.2.9}$$

Nie interesuje nas postać funkcji $H(t)$, lecz czas t, po którym zwierciadło wody w zbiorniku osiągnie wartość H_k. Równanie różniczkowe należy więc całkować z krokiem Δt do chwili spełnienia warunku

$$H(t) \le H_k \tag{6.2.10}$$

Algorytm obliczeń można zapisać w następującej ogólnej postaci:

1) podstaw $j = 1$, $H_j = H_p$;

2) oblicz pierwsze przybliżenie $H_{j+1}^{(0)}$ według (6.2.6);
3) oblicz następne przybliżenie $H_{j+1}^{(1)}$ według (6.2.7);
4) sprawdź warunek (6.2.8):
   — jeśli jest spełniony:
   podstaw: $H_{j+1} = H_{j+1}^{(1)}$ na j podstaw $j + 1$
   i przejdź do punktu 5,
   — jeśli nie jest spełniony:
   podstaw: $H_{j+1}^{(0)} = H_{j+1}^{(1)}$
   i przejdź do punktu 3;
5) sprawdź warunek (6.2.10);
   — jeśli nie jest spełniony: przejdź do punktu 2;
   — jeśli jest spełniony, zakończ obliczenia, ponieważ $t = (j - 1) \cdot \Delta t$ jest poszukiwanym czasem, po którym zwierciadło wody osiągnie poziom H_k.

Obliczenia wykonane według powyższego algorytmu dla przyjętych danych wykazały, że przygotowanie zbiornika na przyjęcie fali wezbraniowej wymaga ok. 17 godzin i 36 minut. Wynik ten uzyskano, przyjmując dokładność rozwiązania równania nieliniowego $\varepsilon = 0{,}0005$ m. Krok całkowania Δt zmieniano w kolejnych wariantach zadania w granicach od $\Delta t = 180$ s do $\Delta t = 720$ s. Zmiany te nie miały istotnego wpływu na wynik ostateczny. ■

■ *Przykład 6.3*

Obliczenie krzywej spiętrzenia

Obliczenie krzywej spiętrzenia polega na wyznaczeniu profilu zwierciadła wody w kanale, który wywołuje budowla piętrząca przy danym natężeniu przepływu (rys. 6.3.1). W tym celu należy scałkować równanie różniczkowe zwyczajne opisujące ruch ustalony wolnozmienny w kanale z warunkiem początkowym odpowiadającym spiętrzeniu w przekroju posadowienia budowli. W układzie współrzędnych, jak na rys. 6.3.1, równanie to można zapisać w następującej postaci (Szymkiewicz, 2000):

$$\frac{dE}{dx} = S \,, \tag{6.3.1}$$

gdzie: $E(x)$ – całkowita energia mechaniczna liczona względem przyjętego poziomu odniesienia,
S – spadek linii energii.

Energię mechaniczną E definiuje wyrażenie

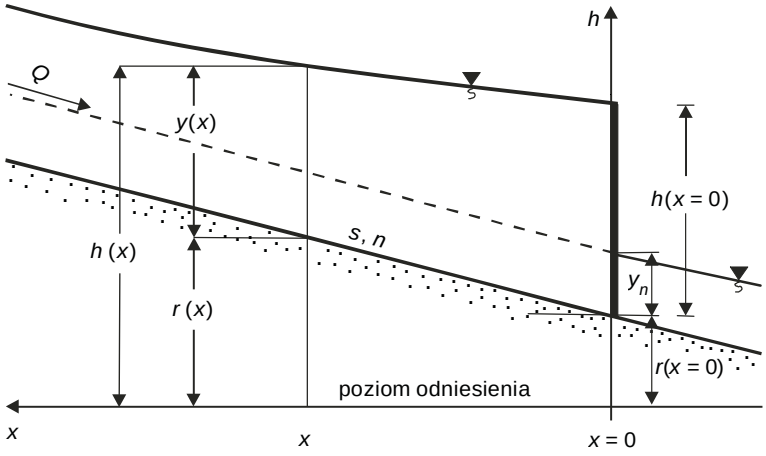
$$E = h + \frac{\alpha Q^2}{2gA^2} \,, \tag{6.3.2}$$

zaś spadek linii energii przy wyrażeniu oporów ruchu przez równanie Manninga określa wzór

$$S = \frac{n^2 Q^2}{R^{4/3} A^2} \,, \tag{6.3.3}$$

gdzie: $h(x)$ – rzędna zwierciadła wody w punkcie x,
α – współczynnik de Saint-Venanta,

Q – natężenie przepływu,
g – przyspieszenie grawitacyjne,
A – powierzchnia przekroju czynnego kanału,
R – promień hydrauliczny przekroju czynnego,
n – współczynnik szorstkości kanału wg Manninga.



Rys. 6.3.1. Profil zwierciadła wody wywołany budowlą piętrzącą

Równanie (6.3.1) rozwiązujemy niejawną metodą trapezową (6.40)

$$E_{j+1} = E_j + \frac{\Delta x}{2}(S_j + S_{j+1}) , \tag{6.3.4}$$

gdzie: j – indeks przekroju obliczeniowego,
Δx – krok całkowania.

Podstawiając do powyższego wzoru wyrażenia (6.3.2) i (6.3.3), otrzymujemy

$$h_{j+1} + \frac{\alpha Q^2}{2gA_{j+1}^2} = h_j + \frac{\alpha Q^2}{2gA_j^2} + \frac{\Delta x}{2}\left(\frac{n^2Q^2}{R_j^{4/3}A_j^2} + \frac{n^2Q^2}{R_{j+1}^{4/3}A_{j+1}^2}\right). \tag{6.3.5}$$

Ponieważ w przekroju budowli piętrzącej o indeksie $j = 1$ znana jest rzędna zwierciadła wody h_p, wynikająca z przyjętego piętrzenia

$$h(x = 0) = h_1 = h_p, \tag{6.3.6}$$

to tym samym w równaniu (6.3.5) wszystkie zmienne z indeksem j są znane. Równanie to jest więc równaniem z jedną niewiadomą h_{j+1}, reprezentującą rzędną zwierciadła wody w przekroju 2 odległym od zapory o Δx. Równanie jest jednak nieliniowe, gdyż $A_{j+1} = A(h_{j+1})$ oraz $R_{j+1} = R(h_{j+1})$. Rozwiązując je jedną z metod rozwiązywania równań algebraicznych nieliniowych, otrzymujemy przybliżoną wartość h_2. Następnie identyczny tok postępowania powtarzamy dla następnego odcinka kanału. Obliczenia prowadzimy do miejsca, w którym spełniony jest następujący warunek:

$$(h_j - r_j) - Y_n \le 0{,}02 \text{ m} , \tag{6.3.7}$$

gdzie: Y_n – głębokość normalna w kanale odpowiadająca natężeniu przepływu Q.

Powyższy warunek określa koniec krzywej spiętrzenia.

Tabela 6.3.1

Obliczony profil zwierciadła wody w kanale

Indeks przekroju	km	Rzędna dna r [m npp]	Rzędna zw. wody h [m npp]	Głębokość y [m]
1	0,000	5,000	8,500	3,500
2	0,250	5,063	8,501	3,438
3	0,500	5,125	8,501	3,376
4	0,750	5,188	8,502	3,314
5	1,000	5,250	8,503	3,253
6	1,250	5,313	8,504	3,191
7	1,500	5,375	8,504	3,129
8	1,750	5,438	8,505	3,068
9	2,000	5,500	8,506	3,006
10	2,250	5,563	8,508	2,945
11	2,500	5,625	8,509	2,884
12	2,750	5,688	8,510	2,823
13	3,000	5,750	8,512	2,762
14	3,250	5,813	8,513	2,701
15	3,500	5,875	8,515	2,640
16	3,750	5,938	8,517	2,579
17	4,000	6,000	8,519	2,519
18	4,250	6,063	8,521	2,459
19	4,500	6,125	8,524	2,399
20	4,750	6,188	8,527	2,339
21	5,000	6,250	8,530	2,280
22	5,250	6,313	8,533	2,221
23	5,500	6,375	8,537	2,162
24	5,750	6,438	8,541	2,104
25	6,000	6,500	8,546	2,046
26	6,250	6,563	8,551	1,988
27	6,500	6,625	8,557	1,932
28	6,750	6,688	8,564	1,876
29	7,000	6,750	8,571	1,821
30	7,250	6,813	8,579	1,767
31	7,500	6,875	8,588	1,713
32	7,750	6,938	8,598	1,661
33	8,000	7,000	8,610	1,610
34	8,250	7,063	8,623	1,560
35	8,500	7,125	8,637	1,512

Opisany algorytm rozwiązania zastosowano do wyznaczenia krzywej spiętrzenia w kanale trapezowym o następujących parametrach:
— szerokość kanału $B = 5$ m,
— nachylenie skarp kanału $M = 1{,}5$,
— spadek dna kanału $s = 0{,}00025$,
— współczynnik szorstkości wg Manninga $n = 0{,}030$,
— głębokość normalna $Y_n = 1{,}5$ m,
— natężenie przepływu $Q = 3$ m^3/s,
— rzędna zwierciadła wody po podpiętrzeniu w przekroju budowli: $h_p = 8{,}5$ m.

Wyniki obliczeń, uzyskane przy kroku całkowania $\Delta x = 250$ m oraz dokładności rozwiązania równania nieliniowego (6.3.5) metodą siecznych $\varepsilon = 0{,}0005$, przedstawiono w tabeli 6.3.1. ■

6.2.4. Metody niejawne wielokrokowe

Jeśli do interpolacji funkcji f w równaniu (6.31) wykorzystany zostanie również punkt nieznany $(x_{j+1}, f(x_{j+1}, y(x_{j+1})))$, otrzymujemy zależność na y_{j+1} w postaci równania nieliniowego: $G(x, y_1, ..., y_j, y_{j+1}) = 0$, czyli tzw. schemat niejawny. Dla grupy otrzymanych w ten sposób schematów ogólna formuła ma następującą postać (Ralston, 1971):

$$y_{j+1} = y_{j-p} + h\sum_{i=-1}^{q} \beta_{i+1} f_{j-i}\,. \tag{6.41}$$

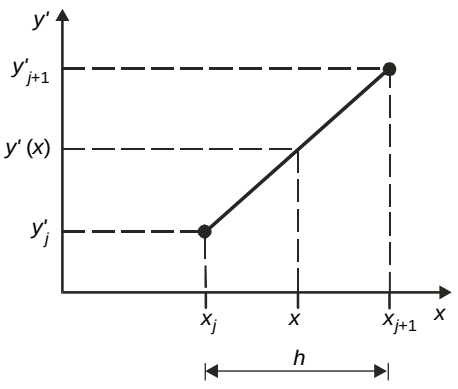
Współczynniki β można obliczyć, jak w wypadku schematów jawnych, za pomocą wielomianów interpolacyjnych Lagrange'a, przyjmując odpowiednie wartości parametrów p i q. Zakładając $p = 0$, otrzymuje się formuły typu Adamsa-Moultona:

$$q = 0 \quad y_{j+1} = y_j + \frac{h}{2}(f_j + f_{j+1}), \tag{6.42a}$$

$$q = 1 \quad y_{j+1} = y_j + \frac{h}{12}(-f_{j-1} + 8f_j + 5f_{j+1}), \tag{6.42b}$$

$$q = 2 \quad y_{j+1} = y_j + \frac{h}{24}(f_{j-2} - 5f_{j-1} + 19f_j + 9f_{j+1}). \tag{6.42c}$$

Sposób wyprowadzenia powyższych wzorów zilustrujemy najprostszym wariantem metody. W celu wyprowadzenia formuły (6.42a), przez węzły x_j oraz x_{j+1} prowadzimy liniowy wielomian interpolacyjny Lagrange'a (rys. 6.9).



Rys. 6.9. Schemat do wyprowadzenia niejawnej formuły Adamsa-Moultona dla $q = 0$

Pochodną y'_j w węźle x_j znamy z poprzedniego kroku obliczeń lub z zadanego warunku początkowego. Natomiast nie znamy jej w węźle x_{j+1}. Liniowy wielomian ma postać

$$y'(x) = y'_j + \frac{y'_{j+1} - y'_j}{x_{j+1} - x_j}(x - x_j) \quad \text{dla} \quad x_j \le x \le x_{j+1}. \tag{6.43}$$

Wprowadźmy nowy układ współrzędnych, taki że $X = x - x_j$. Wzór powyższy zapiszemy następująco:

$$y'(X) = y'_j + \frac{y'_{j+1} - y'_j}{h} X \quad \text{dla} \quad 0 \le X \le h. \tag{6.44}$$

Obliczmy całkę z pochodnej $y'(X)$ w przedziale $\langle 0, h\rangle$. Otrzymamy

$$\begin{aligned} \int_0^h y'(X)dX &= \int_0^h \left(y'_j + \frac{y'_{j+1} - y'_j}{h} X \right) dX = \\ &= y'_j h + \frac{y'_{j+1} - y'_j}{h} \frac{h^2}{2} = \frac{h}{2}(y'_j - y'_{j+1}). \end{aligned} \tag{6.45}$$

Wstawiając otrzymany wynik całkowania do ogólnego wzoru (6.12), otrzymuje się równanie

$$y_{j+1} = y_j + \frac{h}{2}(y'_j + y'_{j+1}), \tag{6.46}$$

czyli poszukiwaną formułę (6.42a). W podobny sposób można wyprowadzić pozostałe formuły typu Adamsa-Moultona.

Jak wynika z równań (6.42), w wypadku schematów niejawnych otrzymuje się wyrażenie na y_{j+1} w postaci równania nieliniowego $G(y_{j+1}, y_j, y_{j-1}, ...) = 0$, którego rozwiązanie jest bardziej skomplikowane. Do rozwiązywania tego typu równań stosuje się zwykle metodę „predyktor-korektor". Idea tej metody jest następująca:

— znając wartości y w punktach aż do j-tego włącznie, oblicza się pierwsze przybliżenie y w punkcie $j + 1$, stosując schemat jawny, czyli dokonuje się predykcji

$$y_{j+1}^{(k=0)} = y_j + h\sum_{i=0}^{q} \beta_i f_{j-i}; \tag{6.47}$$

— znając to przybliżenie, dokonuje się korekty wyniku metodą iteracji prostej, czyli

$$y_{j+1}^{(k+1)} = y_j + h\sum_{i=0}^{q} \beta_{i+1} f_{j-i} + h\beta_0 f_{j+1}^{(k)}, \tag{6.48}$$

gdzie $f_{j+1}^{(k)} = f(x_{j+1}, y_{j+1}^{(k)})$, zaś k jest indeksem iteracji.

Proces ten prowadzi się do momentu uzyskania wystarczająco dokładnego rozwiązania. Można powiedzieć, że metoda „predyktor-korektor" jest praktyczną realizacją niejawnej metody wielokrokowej.

Metody wielokrokowe wymagają wcześniejszej znajomości pewnej liczby punktów. Punkty te można obliczyć, stosując metody jednokrokowe. Zauważmy jednak, że równanie (6.42a) jest identyczne z formułą niejawnej metody trapezowej (6.40), czyli w zasadzie jest metodą jednokrokową.

Z doświadczeń wynika, że schematy jawne są mniej dokładne od niejawnych, jednakże te z kolei wymagają dość dokładnego określenia początkowej wartości niewiadomej y_{j+1}.

6.2.5. Rozwiązywanie układów równań różniczkowych zwyczajnych

Poznane metody rozwiązywania zagadnienia początkowego równania różniczkowego można uogólnić na przypadek *n* równań. Załóżmy, że układ równań rzędu pierwszego sprowadzić można do postaci

$$\begin{aligned} \frac{dy_1}{dx} &= f_1(x, y_1, y_2, \ldots, y_n), \\ \frac{dy_2}{dx} &= f_2(x, y_1, y_2, \ldots, y_n), \\ &\vdots \\ \frac{dy_n}{dx} &= f_n(x, y_1, y_2, \ldots, y_n), \end{aligned} \tag{6.49}$$

gdzie: *n* oznacza rozmiar układu.

Powyższy układ w zapisie wektorowym przyjmie postać

$$\frac{d\mathbf{Y}}{dx} = \mathbf{F}(x, \mathbf{Y}), \tag{6.50}$$

gdzie: $\mathbf{Y}$ – wektor o składowych $y_1, y_2, \ldots, y_n$,
$\mathbf{F}$ – funkcja wektorowa o składowych $f_1(x, y_1, y_2, \ldots, y_n)$, $f_2(x, y_1, y_2, \ldots, y_n)$, ..., $f_n(x, y_1, y_2, \ldots, y_n)$.

Warunki początkowe określa wektor $\mathbf{Y}_0$ o składowych:

$$y_1(x_0), y_2(x_0), \ldots, y_n(x_0)$$

Wszystkie poprzednio omówione metody dają się bez trudu uogólnić na układ równań, czyli na przypadek wektorów i funkcji wektorowych. Na przykład, wzory Rungego-Kutty rzędu drugiego dla układu 3 równań można zapisać w postaci wektorowej następująco:

$$\begin{aligned} \mathbf{K}_1 &= h\mathbf{F}(x_j, \mathbf{Y}_j), \\ \mathbf{K}_2 &= h\mathbf{F}\left(x_j + \frac{1}{2}h, \mathbf{Y}_j + \frac{1}{2}\mathbf{K}_1\right), \\ \mathbf{Y}_{j+1} &= \mathbf{Y}_j + \mathbf{K}_2, \end{aligned} \tag{6.51}$$

gdzie: $\mathbf{K}_1 = \begin{Bmatrix} k_1 \\ l_1 \\ m_1 \end{Bmatrix}$, $\mathbf{K}_2 = \begin{Bmatrix} k_2 \\ l_2 \\ m_2 \end{Bmatrix}$, $\mathbf{F}(x) = \begin{Bmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{Bmatrix}$, $\mathbf{Y}(x) = \begin{Bmatrix} u(x) \\ v(x) \\ w(x) \end{Bmatrix}$,

$k_1 = h\, f_1(x_j, u_j, v_j, w_j)$,
$l_1 = h\, f_2(x_j, u_j, v_j, w_j)$,
$m_1 = h\, f_3(x_j, u_j, v_j, w_j)$,

$$k_2 = hf_1\left(x_j + \frac{1}{2}h, u_j + \frac{1}{2}k_1, v_j + \frac{1}{2}l_1, w_j + \frac{1}{2}m_1\right),$$

$$l_2 = hf_2\left(x_j + \frac{1}{2}h,\, u_j + \frac{1}{2}k_1,\, v_j + \frac{1}{2}l_1,\, w_j + \frac{1}{2}m_1\right),$$

$$m_2 = hf_3\left(x_j + \frac{1}{2}h,\, u_j + \frac{1}{2}k_1,\, v_j + \frac{1}{2}l_1,\, w_j + \frac{1}{2}m_1\right),$$

$u_{j+1} = u_j + k_2,$
$v_{j+1} = v_j + l_2,$
$w_{j+1} = w_j + m_2.$

Analogicznie można przedstawić w zapisie wektorowym każdą inną metodę rozwiązywania równań różniczkowych zwyczajnych.

W wielu zagadnieniach otrzymuje się układy równań różniczkowych zwyczajnych, których nie warto sprowadzać do rozpatrywanej wcześniej postaci (6.50). Na przykład, rozwiązując równania różniczkowe cząstkowe metodą elementów skończonych (o czym będzie mowa w rozdziale 7), napotykamy problem rozwiązania układu równań różniczkowych zwyczajnych o postaci:

$$\mathbf{S}\frac{d\mathbf{Y}}{dt} + \mathbf{AY} = \mathbf{0}, \tag{6.52}$$

gdzie **S** jest zwykle macierzą stałą, **A** jest macierzą zmienną zależną od *t*, a w zagadnieniach nieliniowych również od **Y**.

Diagonalizacja tego układu w celu sprowadzenia go do postaci (6.50) jest nieopłacalna. Macierze **S** i **A** są zwykle pasmowe. Ich elementy niezerowe zlokalizowane są w ograniczonej odległości od głównej przekątnej, zaś poza pasmem wszystkie elementy są zerowe. Do rozwiązywania tego typu równań różniczkowych zwyczajnych najlepiej jest stosować metody A-stabilne, czyli takie metody, których obszarem absolutnej stabilności jest cała lewa półpłaszczyzna zespolona ($Re\ z < 0$) (Dahlquist i Bjorck, 1983; Palczewski, 1999). Okazuje się, że spośród wszystkich metod rozwiązywania układów równań różniczkowych zwyczajnych metodami A-stabilnymi są tylko te metody, które spełniają następujące warunki (Jankowscy, 1981):

— są metodami niejawnymi,
— są rzędu nie wyższego niż 2.

Warunki te spełniają tylko 2 metody spośród metod klasycznych, a mianowicie:

— metoda niejawna Eulera,
— metoda niejawna trapezów (niejawny schemat Adamsa-Moultona II rzędu).

Metody te opisać można jedną formułą wspólną

$$\mathbf{Y}_{t+\Delta t} = \mathbf{Y}_t + \Delta t((1-\theta)\,\mathbf{Y}_t' + \theta\mathbf{Y}'_{t+\Delta t}) \tag{6.53}$$

Dla $\theta = 1$ jest to schemat Eulera niejawny, dla $\theta = 0{,}5$ jest to niejawny schemat trapezowy. W niektórych przypadkach dobre rezultaty uzyskuje się przy $\theta = 2/3$, co odpowiada tzw. schematowi Galerkina.

Wyznaczając z równania (6.52) wektor pochodnych, otrzymamy

$$\mathbf{Y}' = \mathbf{S}^{-1}(-\mathbf{AY}). \tag{6.54}$$

Po wstawieniu tego wyrażenia do równania metody (6.53), otrzymuje się zależność

$$\mathbf{Y}_{t+\Delta t} = \mathbf{Y}_t + \Delta t((1-\theta)\,\mathbf{S}^{-1}(-\mathbf{A}_t\mathbf{Y}_t) + \theta\mathbf{S}^{-1}(-\mathbf{A}_{t+\Delta t}\,\mathbf{Y}_{t+\Delta t})),$$

którą po przekształceniu i uporządkowaniu można zapisać następująco:

$$(\mathbf{S} + \Delta t\theta\mathbf{A}_{t+\Delta t})\mathbf{Y}_{t+\Delta t} = (\mathbf{S} - \Delta t(1-\theta)\,\mathbf{A}_t)\mathbf{Y}_t), \tag{6.55}$$

lub krócej

$$\mathbf{R}_{t+\Delta t}\mathbf{Y}_{t+\Delta t} = \mathbf{F}, \tag{6.56}$$

gdzie: $\mathbf{R}_{t+\Delta t} = \mathbf{S} + \Delta t\ \theta\mathbf{A}_{t+\Delta t}$,
$\mathbf{F} = (\mathbf{S} - \Delta t(1 - \theta)\,\mathbf{A}_t)\mathbf{Y}_t$.

Ostatecznie więc problem rozwiązania układu równań różniczkowych zwyczajnych w postaci (6.52) sprowadzony został do problemu rozwiązania układu równań algebraicznych. Poszukiwany wektor $\mathbf{Y}_{t+\Delta t}$ otrzymujemy, rozwiązując w każdym kroku czasowym układ (6.56). Układ ten jest liniowy, jeśli elementy macierzy **A** nie zależą od **Y**. Jeśli **A** zależy od **Y**, układ jest nieliniowy. W tym drugim przypadku należy zastosować iteracyjne rozwiązanie układu równań, stosując np. metodę Newtona albo metodę iteracji prostej. Jeśli macierze **S** oraz **A** są pasmowe, to macierz **R** jest również pasmowa.

6.2.6. Rozwiązywanie równań różniczkowych zwyczajnych rzędu wyższego niż jeden

Dane jest równanie różniczkowe rzędu III

$$\frac{d^3y}{dx^3} = f\left(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}\right). \tag{6.57}$$

Należy znaleźć funkcję $y(x)$ spełniającą je oraz zadane warunki początkowe:

$$\text{dla } x = x_0\text{: } y = y_0,\ \frac{dy}{dx} = \left.\frac{dy}{dx}\right|_{x_0},\quad \frac{d^2y}{dx^2} = \left.\frac{d^2y}{dx^2}\right|_{x_0}.$$

Równanie to można zastąpić układem równań w sposób następujący:

$$\frac{dy}{dx} = u,\quad \frac{d^2y}{dx^2} = \frac{du}{dx} = v,\quad \frac{d^3y}{dx^3} = \frac{d^2u}{dx^2} = \frac{dv}{dx} = f(x, y, u, v), \tag{6.58a,b,c}$$

czyli:

$$\frac{dy}{dx} = u, \tag{6.59a}$$

$$\frac{du}{dx} = v, \tag{6.59b}$$

$$\frac{dv}{dx} = f(x,\ y,\ u,\ v) \tag{6.59c}$$

z warunkami początkowymi: dla $x = x_0$: $u_0 = y_0'$, $v_0 = y_0''$, $y = y_0$.
Powyższy tok postępowania ma charakter ogólny, gdyż każde równanie n-tego rzędu można zastąpić układem n równań rzędu pierwszego.

■ *Przykład 6.4*

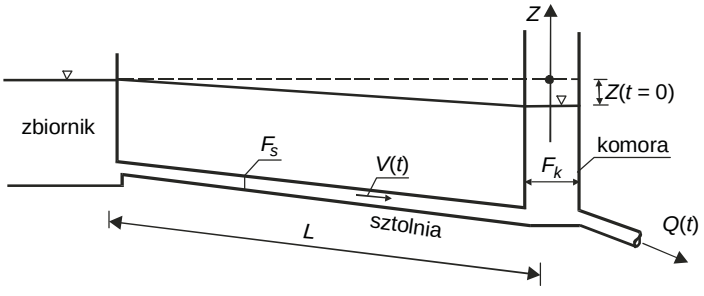
Rozwiązanie układu równań różniczkowych zwyczajnych opisujących przepływ wody w sztolni i komorze wyrównawczej

Przepływ nieustalony w sztolni i komorze wyrównawczej (rys. 6.4.1) opisuje układ równań różniczkowych zwyczajnych o postaci (Czetwertyński i Utrysko, 1969):

$$\frac{L}{g}\frac{dV}{dt} = -(Z + KV|V|)\,, \tag{6.4.1}$$

$$F_k \frac{dZ}{dt} = (F_s V - Q(t))\,, \tag{6.4.2}$$

gdzie: $V(t)$ – prędkość wody w sztolni [m/s],
$Z(t)$ – wzniesienie zwierciadła wody w komorze ponad przyjęty poziom porównawczy [m],
L – długość sztolni [m],
F_s – pole przekroju poprzecznego sztolni [m^2],
F_k – pole przekroju poprzecznego komory [m^2],
g – przyspieszenie grawitacyjne [m/s^2],
$Q(t)$ – znana funkcja czasu reprezentująca odpływ ze sztolni [m^3/s], $K = Ln^2/R^{4/3}$,
n – współczynnik szorstkości sztolni wg Manninga,
R – promień hydrauliczny sztolni [m].



Rys. 6.4.1. Schemat systemu zbiornik-sztolnia-komora wyrównawcza

Warunki początkowe dla tego układu mają postać:
dla $t = 0$

$$V = \frac{Q(t)}{F_s}, \quad Z = -KV|V|\,. \tag{6.4.3a,b}$$

Układ ten należy rozwiązać metodą Rungego-Kutty, stosując formuły IV rzędu, czyli (6.27). Całkowanie należy prowadzić do chwili zaniku oscylacji w komorze wyrównawczej.

Układ równań (6.4.1, 6.4.2) można przekształcić do postaci:

$$\frac{dV}{dt} = (-Z - K|V|V)\frac{g}{L} = F(V, Z), \tag{6.4.4}$$

$$\frac{dZ}{dt} = (F_s V - Q(t))\frac{1}{F_k} = G(t,V). \tag{6.4.5}$$

Zastosowanie przyjętego wariantu metody Rungego-Kutty dla tego układu równań prowadzi do formuł:

$$k_1 = hF(V_j, Z_j)\,, \tag{6.4.6a}$$

$$l_1 = hG(t_j, V_j)\,, \tag{6.4.6b}$$

$$k_2 = hF\left(V_j + \frac{k_1}{2},\quad Z_j + \frac{l_1}{2}\right), \tag{6.4.6c}$$

$$l_2 = hG\left(t_j + \frac{h}{2},\quad V_j + \frac{k_1}{2}\right), \tag{6.4.6d}$$

$$k_3 = hF\left(V_j + \frac{k_2}{2},\quad Z_j + \frac{l_2}{2}\right), \tag{6.4.6e}$$

$$l_3 = hG\left(t_j + \frac{h}{2},\quad V_j + \frac{k_2}{2}\right), \tag{6.4.6f}$$

$$k_4 = hF(V_j + k_3, Z_j + l_3)\,, \tag{6.4.6g}$$

$$l_4 = hG(t_j + h, V_j + k_3)\,. \tag{6.4.6h}$$

$$V_{j+1} = V_j + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)\,, \tag{6.4.7a}$$

$$Z_{j+1} = Z_j + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)\,. \tag{6.4.7b}$$

Algorytm obliczeń przedstawić można w następującej postaci:

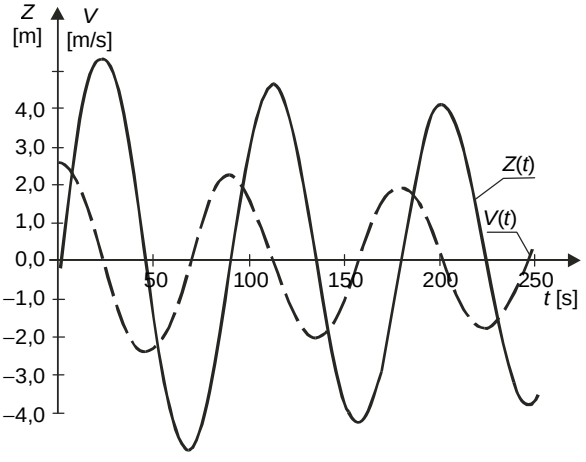
1) podstaw $j = 1$, $t_j = 0$ i oblicz początkowe wartości V_j oraz Z_j według wzorów (6.4.3a, b);
2) oblicz współczynniki: k_1, l_1, k_2, l_2, k_3, l_3, k_4, l_4, realizując formuły (6.4.6a, b, c, d, e, f, g, h);
3) oblicz $t_{j+1} = t_j + h$ oraz V_{j+1} i Z_{j+1} według wzorów (6.4.7a, b);
4) sprawdź warunek ustalenia się przepływu w sztolni: $|V_{j+1} - V_j| \le \varepsilon$
   - jeśli nie jest spełniony, na j podstaw $j + 1$ i przejdź do punktu 2;
   - jeśli jest spełniony, zakończ obliczenia, pamiętając, że liczba punktów całkowania wynosi $N = j + 1$.

Dla zilustrowania działania algorytmu wykonano obliczenia, w których przyjęto, że:

— krok całkowania wynosi $h = 1$ s,
— dopuszczalna zmiana prędkości wody w sztolni w ruchu ustalonym $\varepsilon = 0{,}001$,
— długość sztolni wynosi $L = 300$ m,
— średnica sztolni wynosi $D_s = 3{,}5$ m,
— średnica komory wynosi $D_k = 9$ m,
— współczynnik szorstkości sztolni wg Manninga wynosi $n = 0{,}012$,
— następuje gwałtowne zamknięcie dopływu wody do turbin, tzn. funkcja $Q(t)$ ma postać:

$$Q(t) = \begin{cases} Q_0 & \text{dla} \quad t \le t_0, \\ 0 & \text{dla} \quad t > t_0; \end{cases}$$

— przepływ początkowy w sztolni wynosi $Q_0 = 25$ m^3/s, zaś $t_0 = 0$.
Otrzymane wyniki obliczeń przedstawiono na rys. 6.4.2 oraz w tabeli 6.4.1.



Rys. 6.4.2. Obliczone funkcje $Z(t)$ i $V(t)$

Tabela 6.4.1

Wyniki obliczeń komory wyrównawczej

t [s]	Z [m]	V [m/s]	Q [m^3/s]	t [s]	Z [m]	V [m/s]	Q [m^3/s]
0,0	–0,35	2,60	25,00	150,0	–3,78	–0,97	0,00
10,0	3,27	2,00	0,00	160,0	–4,23	0,40	0,00
20,0	5,24	0,51	0,00	170,0	–2,69	1,56	0,00
30,0	4,71	–1,18	0,00	180,0	0,08	1,94	0,00
40,0	2,00	–2,26	0,00	190,0	2,72	1,41	0,00
50,0	–1,55	–2,24	0,00	200,0	4,02	0,25	0,00
60,0	–4,25	–1,19	0,00	210,0	3,42	–1,01	0,00
70,0	–4,89	0,37	0,00	220,0	1,22	–1,77	0,00
80,0	–3,23	1,73	0,00	230,0	–1,48	–1,67	0,00
90,0	–0,11	2,23	0,00	240,0	–3,42	–0,80	0,00
100,0	2,96	1,67	0,00	250,0	–3,72	0,42	0,00
110,0	4,55	0,36	0,00	260,0	–2,26	1,42	0,00
120,0	3,98	–1,08	0,00	270,0	0,22	1,72	0,00
130,0	1,56	–1,98	0,00	280,0	2,53	1,21	0,00
140,0	–1,51	–1,92	0,00	290,0	3,60	0,16	0,00

■

6.3. Numeryczne rozwiązywanie zagadnień brzegowych układów równań różniczkowych zwyczajnych

Jeśli wymagamy, aby rozwiązywany układ równań różniczkowych zwyczajnych spełniał zadane warunki nie tylko w punkcie początkowym przedziału całkowania $\langle a, b\rangle$, ale także na jego końcu, to tak sformułowany problem jest zagadnieniem brzegowym układu równań różniczkowych zwyczajnych.

W poprzednim przypadku warunki zadane na początku przedziału całkowania zapewniały jednoznaczne rozwiązanie. Otrzymywano je, przemierzając przedział $\langle a, b\rangle$ z krokiem całkowania h i obliczając funkcje w kolejnych węzłach. Natomiast w tym przypadku warunki zadane na początku przedziału nie zapewniają jednoznacznego rozwiązania problemu, gdyż przyjęte przypadkowo pozostałe warunki na brzegu $x = a$ zwykle nie zapewniają spełnienia warunków wymaganych na końcu przedziału całkowania $x = b$. Jest to zasadnicza różnica pomiędzy zagadnieniem początkowym opisanym w punkcie 6.2.5 a zagadnieniem brzegowym. W takiej sytuacji należy oczekiwać, że metody stosowane do rozwiązania zagadnienia brzegowego będą miały charakter iteracyjny. W konsekwencji nakład pracy w przypadku zagadnienia brzegowego jest nieporównywalnie większy niż w przypadku zagadnienia początkowego.

Druga, zasadnicza różnica pomiędzy zagadnieniem początkowym i brzegowym dotyczy istnienia rozwiązania. O ile to pierwsze praktycznie zawsze ma rozwiązanie, jeśli tylko spełniony jest warunek Lipschitza, to rozwiązanie zagadnienia brzegowego może nie istnieć (Stoer i Bulirsch, 1980).

Standardowe sformułowanie zagadnienia brzegowego układu n równań różniczkowych zwyczajnych o postaci

$$\frac{dy_i(x)}{dx} = g_i(x, y_1, y_2, \cdots, y_n) \tag{6.60}$$

gdzie: $i = 1, 2, ..., n$

brzmi następująco: poszukujemy rozwiązania układu (6.60) w przedziale $\langle a, b\rangle$ takiego, aby spełniało ono n_1 warunków zadanych w punkcie $x = a$ oraz $n_2 = n - n_1$ warunków zadanych w punkcie $x = b$. Zauważmy, że tak zdefiniowany problem może być formułowany również dla równań różniczkowych zwyczajnych rzędu wyższego niż 1. Jak bowiem wiadomo z punktu 6.2.6, każde równanie rzędu n można zastąpić układem n równań rzędu pierwszego.

Do rozwiązania zagadnienia brzegowego układu równań różniczkowych zwyczajnych stosuje się dwie klasy metod. Są to metody strzału oraz metody różnicowe nazywane niekiedy relaksacyjnymi (Press i inni, 1992). Metody strzału polegają na przyjęciu wartości wszystkich zmiennych zależnych, nie tylko zadanych jako warunek brzegowy na początku przedziału całkowania i następnie na cyklicznym rozwiązywaniu zagadnienia początkowego. W trakcie rozwiązywania nieznane wartości w punkcie początkowym są tak modyfikowane, aby otrzymać spełnienie warunku brzegowego zadanego na końcu przedziału całkowania. Natomiast metody różnicowe wymagają zastąpienia równań różniczkowych przybliżającymi je równaniami algebraicznymi. Otrzymuje się w ten sposób układ równań algebraicznych, który należy domknąć, wykorzystując zadane warunki brzegowe, a następnie rozwiązać go. W rezultacie otrzymuje się wartości funkcji w węzłach obliczeniowych przedziału całkowania, które jednocześnie spełniają zadane warunki na jego brzegach.

W hydraulice kanałów otwartych formułować można zarówno zagadnienia początkowe, jak i zagadnienia brzegowe równań różniczkowych zwyczajnych. Rozwiązanie zagadnienia początkowego równania ruchu ustalonego wolnozmiennego, w wyniku którego otrzymano tzw. krzywą spiętrzenia w kanale wywołaną budowlą piętrzącą, przedstawiono

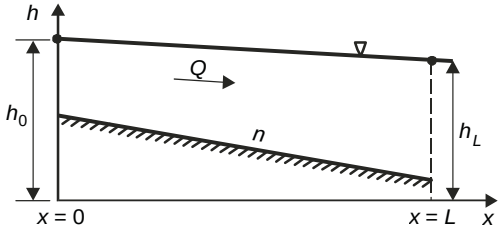
w przykładzie 6.3. Dla tego samego równania można sformułować taki problem rozwiązania, który da się sprowadzić do standardowego zagadnienia brzegowego.

Rozpatrzmy równanie przepływu ustalonego wolnozmiennego w kanale otwartym, zapisane w następującej postaci (Szymkiewicz, 2000).

$$\frac{d}{dx}\left(h+\frac{\alpha Q^2}{2gA^2}\right)=-\frac{n^2Q^2}{R^{4/3}A^2}, \tag{6.61}$$

gdzie: h – rzędna zwierciadła wody liczona od przyjętego poziomu porównawczego,
Q – natężenie przepływu,
A – powierzchnia przekroju czynnego kanału,
R – promień hydrauliczny przekroju czynnego,
α – współczynnik de Saint-Venanta,
g – przyspieszenie grawitacyjne,
n – współczynnik szorstkości wg Manninga.

Załóżmy, że na odcinku kanału o długości L bez dopływu bocznego panuje przepływ ustalony niejednostajny (rys. 6.10).



Rys. 6.10. Odcinek kanału

Załóżmy również, że w wyniku pomiarów znane są:
— rzędna zwierciadła wody na początku odcinka $h(x = 0) = h_0$,
— rzędna zwierciadła wody na końcu odcinka $h(x = L) = h_L$,
— natężenie przepływu Q = const.
Przyjmijmy ponadto, że współczynnik szorstkości wg Manninga, który nie zmienia się na długości kanału (n = const), jest nieznany.

Sformułujmy następujący, bardzo praktyczny, problem: wyznaczyć położenie zwierciadła wody $h(x)$ $(0 \le x \le L)$, które na końcach odcinka kanału przyjmie zadane wartości $h(x = 0) = h_0$ oraz $h(x = L) = h_L$, a także odpowiadającą tej sytuacji wartość współczynnika szorstkości n. Zauważmy, że tak sformułowany problem jest niczym innym jak zagadnieniem brzegowym następującego układu równań różniczkowych zwyczajnych:

$$\frac{d}{dx}\left(h+\frac{\alpha Q^2}{2gA^2}\right)=-\frac{n^2Q^2}{R^{4/3}A^2}, \tag{6.62}$$

$$\frac{dn}{dx}=0. \tag{6.63}$$

Podobne zagadnienie można sformułować w sytuacji, gdy znany jest współczynnik szorstkości n, natomiast nieznane jest natężenie przepływu Q. Rozwiązując problem brzegowy układu równań

$$\frac{d}{dx}\left(h+\frac{\alpha Q^2}{2gA^2}\right)=-\frac{n^2Q^2}{R^{4/3}A^2}, \tag{6.64}$$

$$\frac{dQ}{dx}=0, \tag{6.65}$$

otrzymujemy profil zwierciadła wody spełniający narzucone warunki na końcach odcinka kanału oraz natężenie przepływu Q.

Rozwiązanie układu (6.62) i (6.63) metodą różnicową oraz metodą strzału przedstawiono odpowiednio w przykładzie 6.5 oraz w przykładzie 6.6.

■ *Przykład 6.5*

Wyznaczenie profilu zwierciadła wody oraz natężenia przepływu w kanale metodą różnicową

Przedział całkowania równań (6.64) i (6.65) $\langle 0, L\rangle$ dzielimy za pomocą N węzłów na odcinki o długości $\Delta x_i (i = 1, 2, ..., N-1)$. Równanie (6.64) aproksymujemy niejawną metodą trapezową (6.40)

$$\frac{E_{i+1}-E_i}{\Delta x_i}+\frac{1}{2}(S_i+S_{i+1})=0, \tag{6.5.1}$$

Po zastąpieniu energii E oraz spadku linii energii S wyrażeniami (6.3.2) i (6.3.3) otrzymuje się równanie algebraiczne

$$\left(h_{i+1}+\frac{\alpha Q^2}{2gA_{i+1}^2}\right)-\left(h_i+\frac{\alpha Q^2}{2gA_i^2}\right)+\frac{\Delta x_i}{2}\left(\frac{n^2Q^2}{R_i^{4/3}A_i^2}+\frac{n^2Q^2}{R_{i+1}^{4/3}A_{i+1}^2}\right)=0. \tag{6.5.2}$$

Podobne równania można zapisać dla każdego przedziału Δx_i ($i = 1, 2, ..., N-1$). Otrzymujemy w ten sposób układ $N-1$ równań. W równaniach tych występuje $N+1$ niewiadomych. Jest to N rzędnych zwierciadła wody w węzłach h_i ($i = 1, 2, 3, ..., N$) oraz natężenie przepływu Q. Układ zamykamy, wprowadzając dwa zadane warunki na brzegach kanału

$$h_1=h_0 \quad \text{i} \quad h_N=h_L.$$

Jest to układ nieliniowych równań algebraicznych, który można zapisać w następującej postaci:

$$\mathbf{AX}=\mathbf{B}, \tag{6.5.3}$$

gdzie: $\mathbf{B}=(h_0, 0, ..., 0, h_L, 0)^T$ – wektor prawych stron,
$\mathbf{X}=(h_1, h_2, ..., h_{N-1}, h_N, Q)^T$ – wektor niewiadomych,
T – symbol transpozycji.

Macierz **A** jest macierzą rzadką o wymiarze $(N + 1) \times (N + 1)$ i strukturze przedstawionej na rys. 6.5.1. Jej elementy są zdefiniowane następująco:

$$a_{1,1} = 1, a_{N,N} = 1, \tag{6.5.4a,b)}$$

$$a_{i,i} = 1, \quad a_{i,i-1} = -1 \quad \text{dla } i = 2, 3, ..., N - 1, \tag{6.5.4c,d}$$

$$a_{i,N+1} = -\frac{\alpha Q}{2gA_{i-1}^2} + \frac{\alpha Q}{2gA_i^2} + \frac{\Delta x_{i-1}}{2}\left(\frac{n^2|Q|}{R_{i-1}^{4/3}A_{i-1}^2} + \frac{n^2|Q|}{R_i^{4/3}A_i^2}\right)$$

$$\text{dla } i = 2, 3, ..., N - 1, \tag{6.5.4e}$$

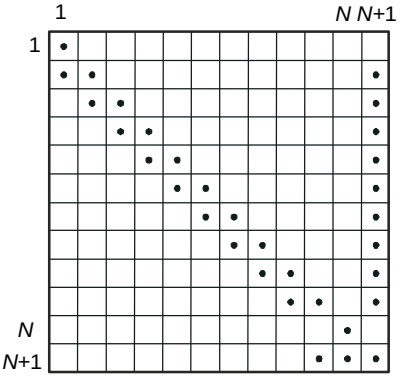
$$a_{N+1,N+1} = -\frac{\alpha Q}{2gA_{N-1}^2} + \frac{\alpha Q}{2gA_N^2} + \frac{\Delta x_{N-1}}{2}\left(\frac{n^2|Q|}{R_{N-1}^{4/3}A_{N-1}^2} + \frac{n^2|Q|}{R_N^{4/3}A_N^2}\right). \tag{6.5.4.f}$$

Celem uwzględnienia właściwego znaku członu tarcia w równaniu (6.5.2) wyrażenie Q^2 zastąpiono wyrażeniem $Q|Q|$. Rozwiązanie układu (6.5.3) metodą Newtona lub metodą iteracji prostej jest kłopotliwe, gdyż w wielu przypadkach proces iteracyjny jest niezbieżny. Skuteczną metodą rozwiązania tego układu jest następujący algorytm będący modyfikacją metody iteracji prostej

$$\mathbf{A}^* \mathbf{X}^{(k+1)} = \mathbf{B}, \tag{6.5.5}$$

w którym

$$\mathbf{A}^* = \mathbf{A}\left(\frac{\mathbf{X}^{(k)} + \mathbf{X}^{(k-1)}}{2}\right), \tag{6.5.6}$$



Rys. 6.5.1. Struktura macierzy **A**

gdzie: k jest indeksem iteracji.

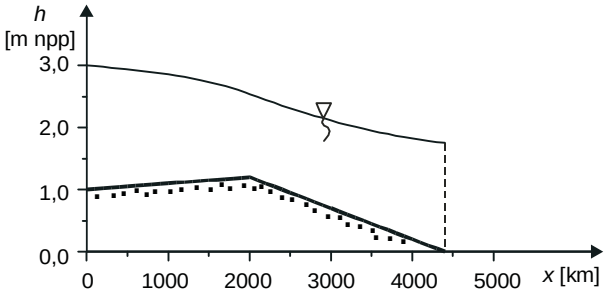
Dla $k = 0$ należy przyjąć $\mathbf{A}^* = \mathbf{A}(\mathbf{X}^{(0)})$. Po przyjęciu pierwszego przybliżenia, proces iteracyjny kontynuuje się do chwili, gdy kolejne przybliżenie spełnia następujące warunki dokładności rozwiązania

$$\left|X_i^{(k+1)} - X_i^{(k)}\right| \le \varepsilon_H \text{ dla } i = 1,N \quad \text{i} \quad \left|X_{N+1}^{(k+1)} - X_{N+1}^{(k)}\right| \le \varepsilon_Q, \tag{6.5.7a,b}$$

gdzie ε_H i ε_Q oznaczają dokładność obliczenia odpowiednio rzędnej zwierciadła wody h_i oraz natężenia przepływu Q.

Opisaną wyżej metodę rozwiązania zagadnienia brzegowego zastosujmy do wyznaczenia układu zwierciadła wody i natężenia przepływu w hipotetycznym kanale. Kanał ma długość L = 4000 m, przekrój trapezowy o szerokości dna 5 m i nachyleniu skarp 1 : 1,5, spadek dna jest zmienny i wynosi s = –0,0001 w pierwszej połowie oraz s = 0,0005 w drugiej. Współczynnik szorstkości wg Manninga wynosi n = 0,030. Kanał podzielono na N = 40 odcinków o stałej długości Δx = 100 m. Rzędne dna rosną liniowo od 1,0 m na początku kanału do 1,20 m w środku jego długości, a następnie zmniejszają się do 0,0 m na końcu kanału.

W przekroju górnym przyjęto poziom wody $h_0 = 3{,}0$ m, co odpowiada głębokości $H_0 = 2{,}0$ m. W przekroju dolnym przyjęto $h_N = h_L = 2$ m, czyli głębokość $H_L = 2$ m. Wyniki obliczeń przedstawiono na rys. 6.5.2. Otrzymanemu profilowi zwierciadła wody odpowiada natężenie przepływu $Q = 6{,}618$ m^3/s.



Rys. 6.5.2. Układ zwierciadła wody w kanale o zmiennym spadku dna

Pierwsze przybliżenie rzędnych zwierciadła wody oraz natężenia przepływu przyjęto arbitralnie. Mianowicie, założono hydrostatyczne napełnienie odcinka kanału do rzędnej $h = 3$ m oraz $Q = 50$ m^3/s. Rozwiązanie z dokładnością $\varepsilon_H = 0{,}002$ oraz $\varepsilon_Q = 0{,}1$ uzyskano po 15 iteracjach. Układ równań liniowych w każdej iteracji rozwiązywano metodą eliminacji Gaussa, w wersji uwzględniającej wyłącznie niezerowe elementy macierzy **A***. ■

■ *Przykład 6.6*

Wyznaczenie profilu zwierciadła wody oraz wartości współczynnika wg Manninga metodą strzału

Rozwiązanie zagadnienia brzegowego układu równań (6.62) i (6.63) można otrzymać, rozwiązując ciąg zagadnień początkowych o postaci

$$\frac{dE}{dx} = -S\ , \tag{6.6.1}$$

$$E(x = 0) = E_0\ . \tag{6.6.2}$$

Problem ten ma jedno rozwiązanie. Ponieważ $E = E(x, n)$, rozwiązanie to będzie zależało od wyboru początkowej wartości współczynnika szorstkości n. Rozwiązaniem zagadnienia brzegowego będzie rozwiązanie zagadnienia początkowego przy tak dobranej wartości n, aby spełniony był drugi warunek na brzegu $x = L$ (rys. 6.6.1)

$$E(L, n) = E_L. \tag{6.6.3}$$

Wprowadźmy funkcję

$$F(n) = E(L, n) - E_L, \tag{6.6.4}$$

definiującą różnicę pomiędzy obliczoną wartością E w punkcie $x = L$ dla danego n i zadanym warunkiem brzegowym w tym punkcie E_L. Rozwiązanie zagadnienia brzegowego sprowadza się w tej sytuacji do wyznaczenia miejsca zerowego powyższej funkcji. War-

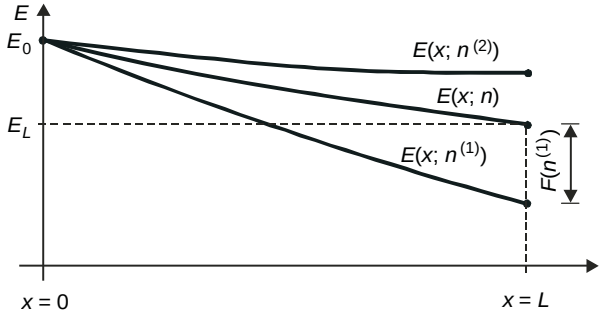
tość $F(n)$ można obliczyć dla dowolnej wartości n. W tym celu należy wyznaczyć wartość rozwiązania zagadnienia początkowego (6.6.4) w punkcie $x = L$ $E(L, n)$. Jeśli znana jest para wartości n takich, że

$$F(n^{(1)}) \cdot F(n^{(2)}) < 0, \tag{6.6.5}$$

to pierwiastek równania $F(n) = 0$ można obliczyć jedną spośród znanych metod poszukiwania miejsca zerowego funkcji jak metoda połowienia, siecznych, Newtona itd. Proces iteracyjny kończy się, gdy spełniony jest warunek:

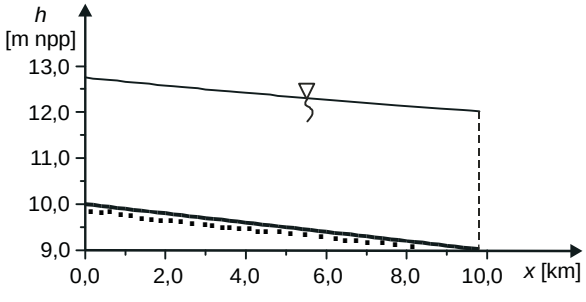
$$\left|F(n)\right| \le \varepsilon \ , \tag{6.6.6}$$

gdzie: ε – dokładność obliczenia rzędnej zwierciadła wody.



Rys. 6.6.1. Rozwiązanie zagadnienia brzegowego jako rozwiązanie ciągu zagadnień początkowych

Postępując w opisany wyżej sposób, rozwiążmy następujące zagadnienie: w kanale prostym o długości L = 9 800 m, mającym przekrój trapezowy o szerokości w dnie B = 5 m i nachyleniu skarp 1 : 1,5 oraz stały spadek dna s = 0,0001, należy wyznaczyć profil zwierciadła wody oraz współczynnik szorstkości według Manninga, jeśli wiadomo, że natężenie przepływu wynosi Q = 30 m^3/s, zaś rzędne zwierciadła wody na końcach kanału są równe h_0 = 12,750 m npp oraz h_L = 12,020 m npp. Rzędne dna kanału zmieniają się liniowo od wartości 10,000 m pp do 9,020 m npp. Przyjęto stały przestrzenny krok całkowania Δx = 200 m. Przedział, w którym jest pierwiastek równania (6.6.4), określono tablicując funkcję $F(n)$ z interwałem Δn = 0,015. Rozwiązanie zadania z dokładnością ε = 0,0001 otrzymano już po 3 iteracjach. Wyliczona wartość współczynnika szorstkości wynosi n = 0,011, zaś obliczony profil zwierciadła wody przedstawiono na rys. 6.6.2.



Rys. 6.6.2. Układ zwierciadła wody w kanale obliczany metodą strzału ■

7 Równania różniczkowe o pochodnych cząstkowych

7.1. Przykłady równań w inżynierii wodnej

Wiele przypadków przepływu wody, ważnych z inżynierskiego punktu widzenia, opisywanych jest za pomocą równań różniczkowych cząstkowych. Równania takie otrzymujemy, gdy w analizowanym zjawisku występuje więcej niż jedna zmienna niezależna. Jest to zatem typowa sytuacja w inżynierii wodnej, gdyż zjawiska przepływu są zwykle zmienne w przestrzeni i w czasie. Równania różniczkowe o pochodnych cząstkowych są wynikiem lokalnego stosowania podstawowych zasad zachowania (masy, energii, pędu). Z tego powodu liczba typów równań lub ich układów jest ograniczona. Podobne równania opisują bardzo różne procesy i zjawiska. Ogólnie nazywa się je równaniami fizyki matematycznej.

Jednym z istotnych działów inżynierii wodnej jest przepływ wody w ośrodku porowatym, nazywany zwykle filtracją w gruncie. Równania opisujące ruch wody w gruncie wyprowadza się z zasady zachowania masy oraz z zachowania pędu, która w tym przypadku sprowadza się do znanego równania Darcy'ego.

Jak wiadomo, równanie dwuwymiarowej nieustalonej filtracji może być przedstawione w następującej ogólnej formie:

$$A\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(B\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(B\frac{\partial h}{\partial y}\right) + w\,. \tag{7.1}$$

gdzie: t – czas,
x, y – współrzędne przestrzenne,
$h(x, y)$ – funkcja reprezentująca tzw. ciśnienie piezometryczne,
w – człon źródłowy,
A, B – współczynniki równania.

Zależnie od znaczenia, jakie przypiszemy parametrom A i B w powyższym równaniu, otrzymać możemy następujące podstawowe przypadki.

— *Filtracja obszarowa ze swobodnym zwierciadłem wody* (*rys. 7.1*)

Przyjmując $A = \mu$, $B = k(h - z)$, otrzymujemy uogólnioną postać równania Boussinesqa

$$\mu\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(k(h-z)\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(k(h-z)\frac{\partial h}{\partial y}\right) + w\,, \tag{7.2}$$

w którym: $h = \mathrm{h}(x,y,t)$ – rzędna swobodnego zwierciadła wody gruntowej,
$z = z(x,y)$ – rzędna spągu warstwy wodonośnej,
$k = k(x,y)$ – współczynnik filtracji,

$\mu = \mu(x,y)$ – odsączalność (porowatość efektywna),
$w = w(x,y,t)$ – zasilanie zewnętrzne niezależne od h (np. infiltracja).

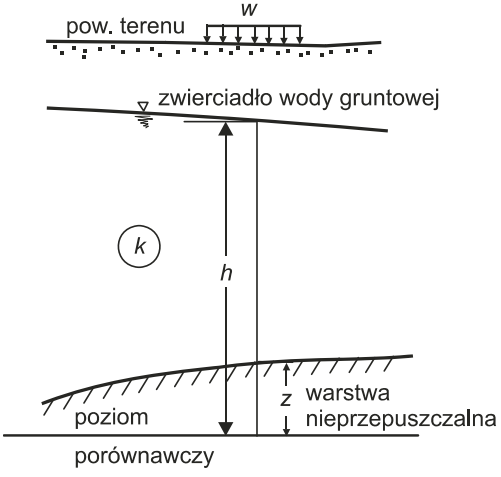
W równaniu (7.2) współczynnik B przy pochodnych II rzędu zależy od poszukiwanej funkcji h, co komplikuje problem jego rozwiązania. Do celów praktycznych często dokonuje się linearyzacji, przyjmując np. $k\,(h - z) = a = \text{const}$, co prowadzi do równania liniowego, którego rozwiązanie jest zdecydowanie łatwiejsze w porównaniu z (7.2). Właściwe przyjęcie stałej a wymaga dobrej znajomości modelowanego zjawiska.

— *Nieustalona filtracja obszarowa pod ciśnieniem* (*rys. 7.2*)

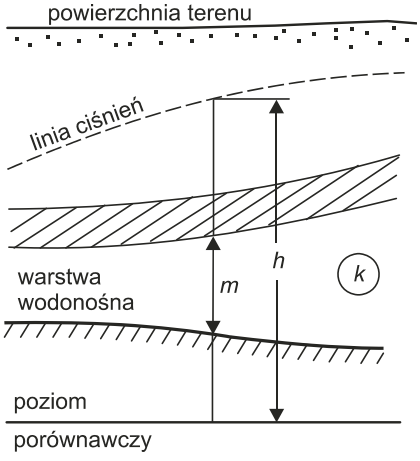
Przyjmując $A = s$, $B = T = k\,m$, otrzymujemy równanie opisujące nieustaloną filtrację wody w warstwie wodonośnej pod ciśnieniem

$$s\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(T\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(T\frac{\partial h}{\partial y}\right) + w\,. \tag{7.3}$$

przy czym: $h = h(x, y, t)$ – rzędna linii ciśnień,
$T = T(x, y) = k\,m$ – przewodność warstwy równa iloczynowi współczynnika filtracji $k(x, y)$ oraz miąższości warstwy wodonośnej $m(x, y)$,
$s = s(x, y)$ – współczynnik zasobności sprężystej,
$w = w(x, y, t)$ – zasilanie zewnętrzne warstwy wodonośnej.



Rys. 7.1. Filtracja obszarowa ze swobodnym zwierciadłem



Rys. 7.2. Filtracja nieustalona pod ciśnieniem

— *Filtracja ustalona ze swobodną powierzchnią*

Przyjmując $\partial h/\partial t = 0$ oraz $k(h - z) = a = \text{const}$, otrzymujemy równanie

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} = F\,, \tag{7.4}$$

w którym $h = h(x, y)$ jest ciśnieniem piezometrycznym, zaś $F = F(x, y) = -\,w/a$ reprezentuje zasilanie wód podziemnych przez czynniki zewnętrzne. Jak widzimy, ten przypadek filtracji opisuje dobrze znane równanie Poissona.

— *Filtracja ustalona pod ciśnieniem* (*rys. 7.3*)

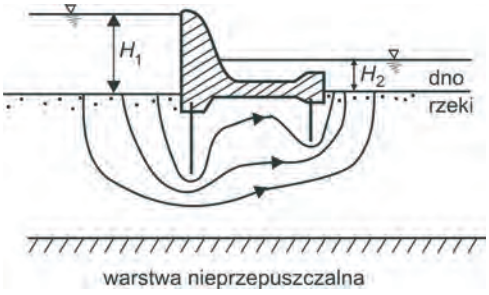
Przyjmując

$$\frac{\partial h}{\partial t} = 0\,,\ B = k \cdot m = \text{const},\ w = 0,$$

otrzymujemy dobrze znane równanie Laplace'a

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} = 0\,, \tag{7.5}$$

w którym $h(x, y)$ jest tzw. ciśnieniem piezometrycznym.
Powyższe równanie opisuje klasyczne zagadnienie budownictwa wodnego, jakim jest filtracja pod budowlą piętrzącą w jednorodnej warstwie gruntu (rys. 7.3).



Rys. 7.3. Ustalona filtracja pod ciśnieniem

W niektórych sytuacjach można przyjąć, że parametry ruchu wody są identyczne we wszystkich płaszczyznach pionowych równoległych do x. Oznacza to, że analizowany przepływ wody jest zagadnieniem jednowymiarowym. Jego równanie ogólne otrzymuje się zaniedbując w (7.1) człon reprezentujący zmienność funkcji w kierunku y. Podstawowe przypadki filtracji jednowymiarowej otrzymuje się z równania ogólnego w sposób analogiczny do przedstawionego dla filtracji dwuwymiarowej.

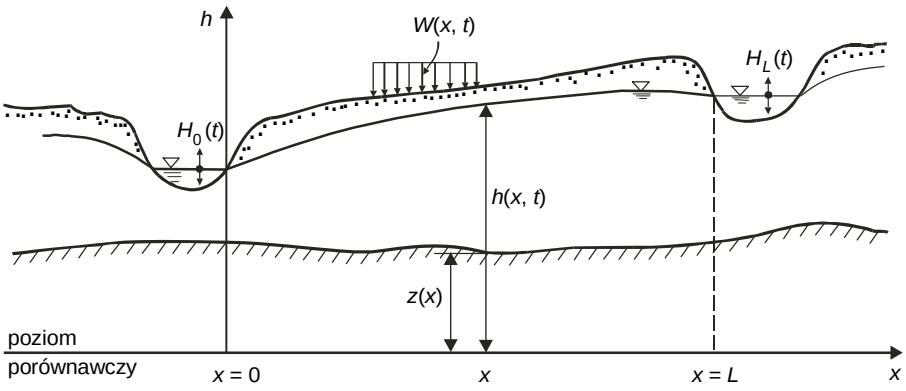
— *Nieustalona filtracja w przypadku jednowymiarowym*

Przyjmijmy w równaniu (7.2), że parametry ruchu nie zależą od y. Wówczas równanie to możemy zapisać w następującej postaci:

$$\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial h}{\partial x}\right) + \frac{w}{\mu}\,, \tag{7.6}$$

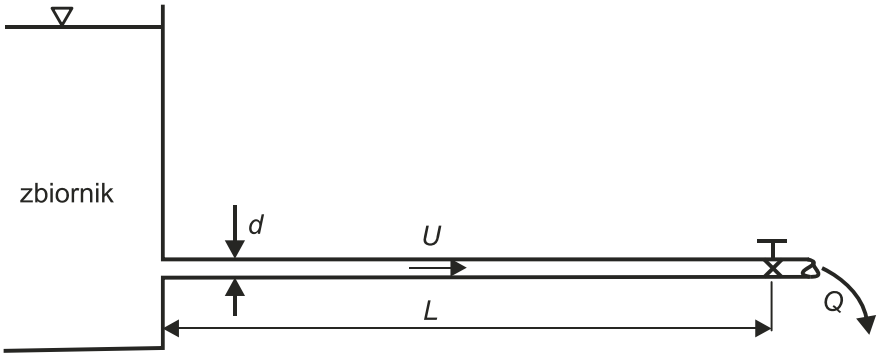
gdzie: $D = k(h - z)/\mu$ jest współczynnikiem dyfuzji.

Równanie to jest dobrze znanym jednowymiarowym równaniem dyfuzji. Może ono opisywać np. przepływ wody w obszarze między dwoma biegnącymi równolegle (w kierunku osi y) kanałami, w których poziomy wody zmieniają się w czasie niezależnie od siebie (rys. 7.4).



Rys. 7.4. Przekrój warstwy wodonośnej

— *Nieustalony przepływ w przewodach pod ciśnieniem*



Rys. 7.5. Schemat rurociągu

Jeśli w rurociągu, w którym płynie ciecz ze stałym natężeniem Q, zmienimy warunki przepływu przez przymknięcie zaworu na jego końcu, wywołamy przepływ nieustalony. Skrajną sytuację wywołamy przez nagłe i całkowite zamknięcie przepływu. Przepływ nieustalony wywoływany jest również przez nagłe otwarcie zaworu.

Równania opisujące ten przypadek przepływu cieczy wyprowadzamy z zasady zachowania pędu oraz zasady zachowania masy. Jeśli założymy:

— sprężystość ścian rurociągu,
— ściśliwość cieczy,
— jednostajny rozkład prędkości i ciśnienia w przekroju poprzecznym rurociągu,
— stałe napełnienie rurociągu cieczą,
— opory hydrauliczne wywołane tarciem jak w przepływie ustalonym

otrzymamy układ równań różniczkowych cząstkowych o postaci (Mitosek, 2001):

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + g\frac{\partial h}{\partial x} + \frac{f}{2d}U|U| = 0\,, \tag{7.7}$$

$$\frac{\partial h}{\partial t} + U\frac{\partial h}{\partial x} + \frac{c^2}{g}\frac{\partial U}{\partial x} = 0\,, \tag{7.8}$$

gdzie: x – współrzędna przestrzenna,
t – czas,
h – ciśnienie piezometryczne,
U – prędkość przepływu cieczy,
f – współczynnik oporów liniowych,
d – średnica wewnętrzna rury,
g – przyspieszenie ziemskie,
c – prędkość fali ciśnienia.

Prędkość fali ciśnienia wyrażona jest następująco:

$$c = \frac{1}{\sqrt{\rho\left(\frac{1}{K} + \frac{d}{Ee}\right)}}\,, \tag{7.9}$$

gdzie: ρ – gęstość cieczy,
K – współczynnik ściśliwości cieczy,
E – moduł Younga materiału rurociągu,
e – grubość ścianki rurociągu.

Równania (7.7) i (7.8) często nazywa się równaniami uderzenia hydraulicznego.

— *Nieustalony przepływ w kanałach otwartych*

Przy założeniu, że:
— przepływ wody jest wolnozmienny
— przepływ jest jednowymiarowy,
— obowiązuje hydrostatyczny rozkład ciśnienia,
— jedyną siłą masową jest siła ciężkości,
— pochylenie dna kanału jest niewielkie,
— rozkład prędkości w pionie jest jednostajny,
— dopływ boczny nie występuje,
otrzymuje się następujący układ równań (Cunge, Holly i Verwey, 1980; Szymkiewicz, 2000):

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + g\frac{\partial H}{\partial x} = (s - S)\,, \tag{7.10}$$

$$\frac{\partial H}{\partial t} + \frac{\partial}{\partial x}(UH) = 0\,, \tag{7.11}$$

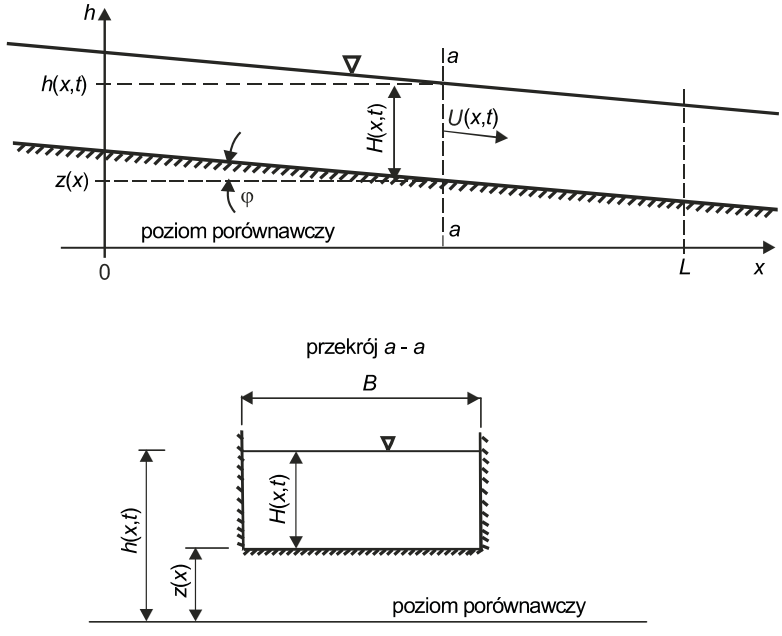
w którym poszczególne symbole mają następujące znaczenie (rys. 7.6):
x – współrzędna przestrzenna,
t – czas,
$U(x, t)$ – średnia prędkość przepływu,
$H(x, t)$ – głębokość,
s – spadek dna kanału,
S – spadek hydrauliczny (linii energii),
g – przyspieszenie grawitacyjne.

Spadek hydrauliczny, jak wspomniano w założeniach, oblicza się jak w ruchu ustalonym. Po przyjęciu formuły Manninga ma on następującą postać:

$$S = \frac{n^2}{H^{4/3}} U|U|, \tag{7.12}$$

gdzie: n – współczynnik szorstkości wg Manninga.

Równanie (7.10), nazywane dynamicznym, wyprowadza się z zasady zachowania pędu, zaś równanie (7.11) jest równaniem ciągłości ruchu, wyprowadzonym z zasady zachowania masy.



Rys. 7.6. Schemat kanału otwartego

— *Przenoszenie masy w strumieniu płynącej wody*

Jeśli do płynącego strumienia wody wprowadzona zostanie domieszka łatwo rozpuszczająca się w wodzie i niezmieniająca dynamiki strumienia (tzw. domieszka pasywna), rozpocznie się proces jej przenoszenia. Przenoszenie masy jest wynikiem adwekcji wywołanej ruchem strumienia wody oraz dyfuzji wywołanej gradientem koncentracji domieszki. Ogólnie mówi się o procesie przenoszenia adwekcyjno-dyfuzyjnego, zaś równanie opisujące ten proces nazywa się równaniem adwekcji-dyfuzji. W przypadku dwuwymiarowym, gdy można założyć jednostajny rozkład koncentracji w pionie, równanie to ma następującą postać (Szymkiewicz, 2000):

$$\frac{\partial}{\partial t}(HC) + \frac{\partial}{\partial x}(UHC) + \frac{\partial}{\partial y}(VHC) - \frac{\partial}{\partial x}\left(HD_x \frac{\partial C}{\partial x}\right) - \frac{\partial}{\partial y}\left(HD_y \frac{\partial C}{\partial y}\right) + H\delta = 0, \tag{7.13}$$

gdzie: x,y – współrzędne przestrzenne,

t – czas,
$C(x,y,t)$ – koncentracja domieszki,
$U(x,y,t)$ – składowa wektora prędkości w kierunku x,
$V(x,y,t)$ – składowa wektora prędkości w kierunku y,
$H(x,y,t)$ – głębokość,
D_x, D_y – współczynniki przenoszenia dyfuzyjnego w kierunku x i y,
δ – człon źródłowy.

Wykorzystując równanie ciągłości przepływu ze swobodną powierzchnią oraz zakładając stałą głębokość H i stałe wartości współczynników $D_x = D_y = D$, a także pomijając człon źródłowy, otrzymujemy prostszą, dobrze znaną, postać równania (7.13):

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} + V\frac{\partial C}{\partial y} - D\frac{\partial^2 C}{\partial x^2} - D\frac{\partial^2 C}{\partial y^2} = 0\,. \tag{7.14}$$

Jest to dwuwymiarowy przypadek równania adwekcji-dyfuzji o stałych współczynnikach.

Analizując przepływ w rzekach i kanałach, często można założyć równomierny rozkład koncentracji w przekroju poprzecznym. W takiej sytuacji równanie (7.13) można przekształcić do postaci jednowymiarowej:

$$\frac{\partial}{\partial t}(AC) + \frac{\partial}{\partial x}(QC) - \frac{\partial}{\partial x}\left(AD\frac{\partial C}{\partial x}\right) + A\sigma = 0\,, \tag{7.15}$$

gdzie: $A(x,t)$ – pole powierzchni przekroju czynnego,
$Q(x,t)$ – natężenie przepływu,
$C(x,t)$ – koncentracja (uśredniona w przekroju),
D – współczynnik dyfuzji.

W szczególnym przypadku przepływu ustalonego jednostajnego i domieszki nierozkładalnej, równanie (7.15) upraszcza się do jednowymiarowego równania adwekcji-dyfuzji o stałych współczynnikach

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} - D\frac{\partial^2 C}{\partial x^2} = 0\,, \tag{7.16}$$

w którym: $U = Q/A$ jest średnią prędkością przepływu.

Dwa szczególne przypadki powyższego równania, to:
— równanie adwekcji

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} = 0\,, \tag{7.17}$$

— równanie dyfuzji

$$\frac{\partial C}{\partial t} - D\frac{\partial^2 C}{\partial x^2} = 0\,. \tag{7.18}$$

Równaniem tego typu jest również równanie (7.6) opisujące jednowymiarową filtrację nieustaloną.

— *Uproszczone modele propagacji fal wezbraniowych*

W inżynierii wodnej równania adwekcji-dyfuzji, adwekcji oraz dyfuzji opisują nie tylko przenoszenie rozpuszczonych w wodzie domieszek, ale i inne procesy. I tak, równaniem adwekcji-dyfuzji jest tzw. równanie fali dyfuzyjnej, będące uproszczonym modelem propagacji fali wezbraniowej (Eagleson, 1978)

$$\frac{\partial Q}{\partial t} + c\frac{\partial Q}{\partial x} - \nu\frac{\partial^2 Q}{\partial x^2} = 0\,, \tag{7.19}$$

gdzie: $Q(x,t)$ – natężenie przepływu,
c – prędkość kinematyczna fali,
ν – współczynnik dyfuzji hydraulicznej.

Prędkość kinematyczną definiuje się następująco:

$$c = \frac{1}{\alpha m Q^{m-1}}\,, \tag{7.20}$$

przy czym $$\alpha = \frac{1}{\left(\dfrac{s^{1/2}}{np^{2/3}}\right)}, \quad m = \frac{3}{5}\,,$$

gdzie: s – spadek dna kanału,
n – współczynnik szorstkości wg Manninga,
p – obwód zwilżony przekroju czynnego.

Natomiast współczynnik dyfuzji hydraulicznej określa następujące wyrażenie:

$$\nu = \frac{Q}{2Bs}\,, \tag{7.21}$$

gdzie: B – szerokość kanału.

Z kolei tzw. model fali kinematycznej ma postać wymienionego wcześniej równania adwekcji:

$$\frac{\partial Q}{\partial t} + c\frac{\partial Q}{\partial x} = 0\,, \tag{7.22}$$

w którym prędkość kinematyczną c definiuje zależność (7.20).

7.2. Klasyfikacja równań różniczkowych cząstkowych i poprawne formułowanie problemu ich rozwiązywania

Równanie różniczkowe cząstkowe drugiego rzędu o dwóch zmiennych niezależnych można napisać w następującej ogólnej postaci:

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} + fu = F(x,y)\,, \tag{7.23}$$

gdzie: $u = u(x, y)$ – szukana funkcja,
$a, b, c, ..., f$ – współczynniki równania,

$F(x, y)$ – funkcja źródłowa,
x, y – zmienne niezależne.

Z powyższym równaniem związany jest tzw. wyróżnik

$$\Delta = b^2 - 4ac \,. \tag{7.24}$$

Zależnie od znaku wyróżnika Δ mamy do czynienia z równaniem różniczkowym cząstkowym określonego typu. I tak, równanie jest:
— typu eliptycznego, jeżeli $\Delta < 0$,
— typu parabolicznego, jeżeli $\Delta = 0$,
— typu hiperbolicznego, jeżeli $\Delta > 0$.

Powyższy sposób klasyfikacji jest bardzo prosty i – jeśli tylko można – powinien być stosowany. Niestety, ma on istotną wadę, którą jest ograniczenie do równań II rzędu z dwiema zmiennymi niezależnymi. Tymczasem, wśród równań wymienionych w poprzednim podrozdziale, występują zarówno równania II rzędu z 3 zmiennymi niezależnymi, jak i równania I rzędu oraz ich układy. Bardziej ogólny sposób klasyfikacji równań różniczkowych cząstkowych polega na badaniu ich charakterystyk.

Zgodnie z definicją (Fletcher, 1991) charakterystykami są linie w czasoprzestrzeni (lub powierzchnie w przypadku dwu- i trójwymiarowym), na których wyjściowe równanie lub układ równań nie ma jednoznacznego rozwiązania. Inaczej mówiąc, na charakterystykach występuje nieciągłość rozwiązania. Warunek ten pozwala wyznaczyć równania charakterystyk, a na podstawie ich natury wnioskować o typie równania lub układu równań. I tak, równanie jest typu:
— hiperbolicznego, jeśli wszystkie charakterystyki są rzeczywiste,
— parabolicznego, jeśli tylko jedna charakterystyka jest rzeczywista,
— eliptycznego, jeśli wszystkie charakterystyki są urojone.

Rozpatrzmy układ równań różniczkowych cząstkowych I rzędu z dwiema zmiennymi niezależnymi reprezentującymi położenie x oraz czas t:

$$a\frac{\partial u}{\partial t} + b\frac{\partial v}{\partial t} + e\frac{\partial u}{\partial x} + f\frac{\partial v}{\partial x} = F_1 \,, \tag{7.25a}$$

$$c\frac{\partial u}{\partial t} + d\frac{\partial v}{\partial t} + g\frac{\partial u}{\partial x} + h\frac{\partial v}{\partial x} = F_2 \,, \tag{7.25b}$$

w którym: $u(x, t)$ i $v(x, t)$ – są poszukiwanymi funkcjami,
$a, b, ..., g, h$ – współczynniki,
F_1, F_2 – człony źródłowe.

W notacji macierzowej układ ten można zapisać następująco:

$$\mathbf{A}\frac{\partial \boldsymbol{\phi}}{\partial t} + \mathbf{B}\frac{\partial \boldsymbol{\phi}}{\partial x} = \mathbf{F} \,, \tag{7.26}$$

gdzie: $\boldsymbol{\phi} = \begin{Bmatrix} u \\ v \end{Bmatrix}$, $\mathbf{F} = \begin{Bmatrix} F_1 \\ F_2 \end{Bmatrix}$, $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$.

Warunek, jaki musi być spełniony na charakterystykach układu, ma postać (Fletcher, 1991)

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{E}dt & \mathbf{E}dx \end{bmatrix} \begin{Bmatrix} \dfrac{\partial \boldsymbol{\phi}}{\partial t} \\ \dfrac{\partial \boldsymbol{\phi}}{\partial x} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F} \\ \mathbf{0} \end{Bmatrix}, \tag{7.27}$$

gdzie **E** jest macierzą jednostkową o wymiarze równym wymiarowi układu.

Jeśli wymaga się, aby układ (7.25) nie miał jednoznacznego rozwiązania na charakterystykach, to wyznacznik główny układu (7.27) musi być równy zero:

$$\det(\mathbf{A}dx - \mathbf{B}dt) = 0. \tag{7.28}$$

Dla układu (7.26) warunek ten przyjmie postać:

$$\begin{vmatrix} a & b & e & f \\ c & d & g & h \\ dt & 0 & dx & 0 \\ 0 & dt & 0 & dx \end{vmatrix} = 0. \tag{7.29}$$

Po obliczeniu wyznacznika otrzymujemy równanie

$$(ad - bc)\left(\frac{dx}{dt}\right)^2 - (ah - cf + ed - hb)\frac{dx}{dt} + (eh - fg) = 0, \tag{7.30}$$

które należy rozwiązać względem dx/dt. Jego wyróżnik ma wartość

$$\Delta = (ah - cf + ed - hb)^2 - 4(ad - bc)(eh - fg). \tag{7.31}$$

Wartość wyróżnika określa naturę pierwiastków równania (7.30), a tym samym naturę charakterystyk układu (7.25). I tak układ będzie typu:

— hiperbolicznego, gdy $\Delta > 0$,
— parabolicznego, gdy $\Delta = 0$,
— eliptycznego, gdy $\Delta < 0$.

Równania charakterystyk układu (7.25) otrzymujemy, rozwiązując równanie kwadratowe (7.30).

Równanie różniczkowe o pochodnych cząstkowych nie zmienia swego typu w obszarze rozwiązania, jeśli jest ono liniowe. Równanie (7.23) i układ (7.25) są liniowe, jeśli współczynniki są stałe lub zależą tylko od zmiennych niezależnych (w równaniu (7.23) są nimi x i y, zaś w układzie (7.25) są to x i t). Jeśli współczynniki równania zależą nie tylko od zmiennych niezależnych, ale i od zmiennej zależnej: $a = a(x, y, u)$, $b = b(x, y, u)$, ..., to równanie jest quasi-liniowe. Jeśli natomiast w równaniu występują pochodne w potędze innej niż 1, co oznacza, że jego współczynniki są funkcją również pochodnej nieznanej funkcji, to równanie nazywamy równaniem nieliniowym. Zatem w przypadku równań różniczkowych cząstkowych o zmiennych współczynnikach, quasi-liniowych lub nieliniowych, typ równania może mieć charakter lokalny.

Korzystając z zaprezentowanych sposobów klasyfikacji równań różniczkowych cząstkowych, określmy typ niektórych równań przedstawionych w podrozdziale 7.1.

Dla równań Poissona (7.4) oraz Laplace'a (7.5) współczynniki przy pochodnych drugiego rzędu są równe: $a = 1$, $b = 0$, $c = 1$. Wyróżnik (7.24) będzie więc równy:

$$\Delta = b^2 - 4ac = 0 - 4 \cdot 1 \cdot 1 = -4 < 0 .$$

Zatem równania Poissona i Laplace'a są typu eliptycznego.

Z kolei w jednowymiarowym równaniu dyfuzji (7.6) współczynniki są równe $a = D$, $b = 0$, $c = 0$. Wyróżnik (7.24) przyjmie następującą wartość:

$$\Delta = b^2 - 4ac = 0 - 4 \cdot D \cdot 0 = 0 .$$

Równanie dyfuzji jest więc równaniem typu parabolicznego. Równaniami typu parabolicznego są również równania adwekcji-dyfuzji (7.16) i (7.19).

Z kolei w celu sklasyfikowania układu równań (7.7) i (7.8), opisujących uderzenie hydrauliczne, musimy wyznaczyć równania charakterystyk. Dla tego układu warunek (7.29) przyjmie postać:

$$\begin{vmatrix} 1 & 0 & U & g \\ 0 & 1 & \dfrac{c^2}{g} & U \\ dt & 0 & dx & 0 \\ 0 & dt & 0 & dx \end{vmatrix} = 0 . \tag{7.32}$$

W wyniku obliczenia wyznacznika otrzymujemy kwadratowe równanie algebraiczne

$$\left(\frac{dx}{dt}\right)^2 - 2U\frac{dx}{dt} - (c^2 - U^2) = 0 . \tag{7.33}$$

Wyróżnik tego równania jest równy:

$$\Delta = 4U^2 + 4(c^2 - U^2) = 4c^2 , \tag{7.34}$$

a jego pierwiastek ma wartość

$$\sqrt{\Delta} = 2c . \tag{7.35}$$

Zatem równanie (7.33) ma dwa pierwiastki:

$$\left.\frac{dx}{dt}\right|_+ = U + c, \quad \left.\frac{dx}{dt}\right|_- = U - c . \tag{7.36a,b}$$

Równanie pierwsze określa tzw. charakterystykę dodatnią, zaś drugie – ujemną.
Równania te definiują charakterystyki układu (7.7) i (7.8). Są one rzeczywiste, a zatem równania te tworzą układ równań różniczkowych cząstkowych typu hiperbolicznego.

Przeprowadzona w analogiczny sposób analiza charakterystyk układu równań de Saint-Venanta (7.10) i (7.11), opisujących przepływ nieustalony w kanałach otwartych, pozwala wykazać, że układ ten ma również dwie charakterystyki rzeczywiste:

$$\left.\frac{dx}{dt}\right|_+ = U + \sqrt{gH}, \quad \left.\frac{dx}{dt}\right|_- = U - \sqrt{gH} , \tag{7.37a,b}$$

a zatem jest on także układem typu hiperbolicznego.

Na zakończenie problemu klasyfikacji określmy jeszcze typ równania adwekcji (7.17). W tym przypadku warunek (7.28) będzie miał postać:

$$\begin{vmatrix} 1 & U \\ dt & dx \end{vmatrix} = 0 . \tag{7.38}$$

Wynika z niego równanie charakterystyki

$$\frac{dx}{dt} = U . \tag{7.39}$$

Zatem równanie adwekcji ma jedną charakterystykę. Jest ona rzeczywista, a więc zgodnie z definicją równanie to jest typu hiperbolicznego.

Wykonana wyżej analiza typu równań wykazuje, że w inżynierii wodnej występują wszystkie typy równań. Są więc równania typu hiperbolicznego, parabolicznego i eliptycznego. Równania hiperboliczne opisują zjawiska falowe. Równania typu parabolicznego opisują dysypacyjne procesy zmienne w czasie, czyli nieustalone. Natomiast równania typu eliptycznego opisują procesy ustalone i są charakterystyczne dla zagadnień równowagi.

Chcąc uzyskać rozwiązanie równania różniczkowego o pochodnych cząstkowych w zadanym obszarze ograniczonym, należy poprawnie sformułować problem jego rozwiązania. Przyjmuje się, że problem ten jest poprawnie sformułowany, gdy spełnione są następujące warunki (Fletcher, 1991):

— rozwiązanie równania istnieje,
— rozwiązanie równania jest jedyne,
— rozwiązanie zależy w sposób ciągły od dodatkowych warunków zadanych na granicy obszaru rozwiązania.

Omówmy krótko poszczególne warunki.

Problem istnienia rozwiązania, istotny z teoretycznego punktu widzenia, zwykle nie jest badany. Przyjmuje się, że rozwiązanie równań wyprowadzonych z podstawowych zasad zachowania istnieje. Taka sytuacja ma miejsce w zagadnieniach przepływu wody.

Problem jednoznaczności rozwiązania związany jest ściśle z dodatkowymi warunkami zadawanymi na granicy obszaru rozwiązania. Jeśli na granicy zadamy za mało informacji dodatkowych, otrzymamy nieskończenie wiele rozwiązań. Jeśli zaś zadamy ich za dużo, otrzymamy rozwiązanie o niefizycznym charakterze lub sprzeczne. Sposób zadawania dodatkowych warunków związany jest ściśle z typem równania. Zatem dokonanie klasyfikacji równania, któremu wcześniej poświęcono dużo miejsca, ma zasadnicze znaczenie praktyczne, gdyż umożliwia poprawne sformułowanie problemu rozwiązania równania.

Jak wiadomo, rozwiązaniem równania różniczkowego cząstkowego jest funkcja, która spełnia jednocześnie w obszarze rozwiązania C to równanie, jak również warunki na granicach obszaru C. Wymagany rodzaj i ilość warunków granicznych zależą od typu równania.

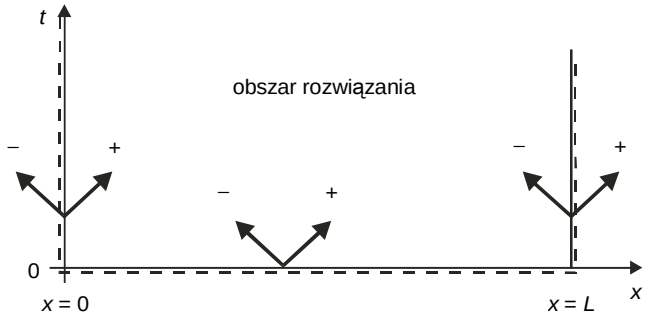
Dla równań typu hiperbolicznego obowiązuje zasada zadawania na każdej granicy tylu warunków, ile charakterystyk wchodzi z tej granicy do obszaru rozwiązania (Godunow, 1975). Sprawdźmy konsekwencje tego wymagania w odniesieniu do przytoczonych w podrozdziale 7.1 równań tego typu, a mianowicie równań uderzenia hydraulicznego (7.7) i (7.8), przepływu nieustalonego w kanałach otwartych (7.10) i (7.11) oraz równania adwekcji (7.17). Obszar rozwiązania wymienionych równań jest następujący:

$$0 \le x \le L, \quad t \ge 0 ,$$

gdzie: L oznacza długość rurociągu lub kanału.

Równania charakterystyk są następujące:

$$\left.\frac{dx}{dt}\right|_{+} = U + c, \quad \left.\frac{dx}{dt}\right|_{-} = U - c \, .$$



Rys. 7.7. Obszar rozwiązania i układ charakterystyk układu równań uderzenia hydraulicznego

Ponieważ $c >> U$, gdyż prędkość fali ciśnienia c jest rzędu setek, a nawet tysiąca m/s, podczas gdy prędkość przepływu U jest rzędu m/s, to układ charakterystyk będzie zawsze wyglądał jak na rys. 7.7. Jedna charakterystyka będzie miała zawsze nachylenie dodatnie, zaś druga – ujemne. Zatem przez granicę $t = 0$ do obszaru rozwiązania wchodzą dwie charakterystyki. Oznacza to, że na tym odcinku granicy należy zadać dwa warunki: $U(x, t = 0) = U_p(x)$ i $H(x, t = 0) = H_p(x)$ dla $O \le x \le L$. Informacje na granicy $t = 0$ nazywa się warunkami początkowymi. Z każdego z brzegów, którymi są fizyczne końce rurociągu, jedna charakterystyka wchodzi do obszaru rozwiązania. W konsekwencji na każdym z nich należy zadać jedną informację: $U(x = 0, t) = U_0(t)$ albo $H(x = 0, t) = H_0(t)$ i $U(x = L, t) = U_L(t)$ albo $H(x = L, t) = H_L(t)$ dla $t \ge 0$. Warunki zadane na brzegach fizycznych nazywa się warunkami brzegowymi. Zespół sformułowanych wyżej warunków początkowych i brzegowych nazywa się warunkami granicznymi (Kącki, 1992). Sformułowane w powyższy sposób zagadnienie nazywa się zagadnieniem początkowo-brzegowym układu równań różniczkowych typu hiperbolicznego.

Bardziej złożoną strukturę charakterystyk ma układ równań de Saint-Venanta (7.10) i (7.11). Są one następujące (7.37a,b):

$$\left.\frac{dx}{dt}\right|_{+} = U + \sqrt{gH}, \quad \left.\frac{dx}{dt}\right|_{-} = U - \sqrt{gH} \, .$$

W tym przypadku nachylenie stycznych do charakterystyk zależy od rodzaju przepływu wody w kanale. W warunkach ruchu nadkrytycznego (spokojnego), gdy liczba Froude'a jest mniejsza od jedności ($F_r < 1$), zachodzi relacja

$$U < \sqrt{gH} \, ,$$

z której wynika różne nachylenie charakterystyk. Ich układ przedstawiono na rys. 7.8a. W tym przypadku należy zadać następujące warunki graniczne:

— warunki początkowe

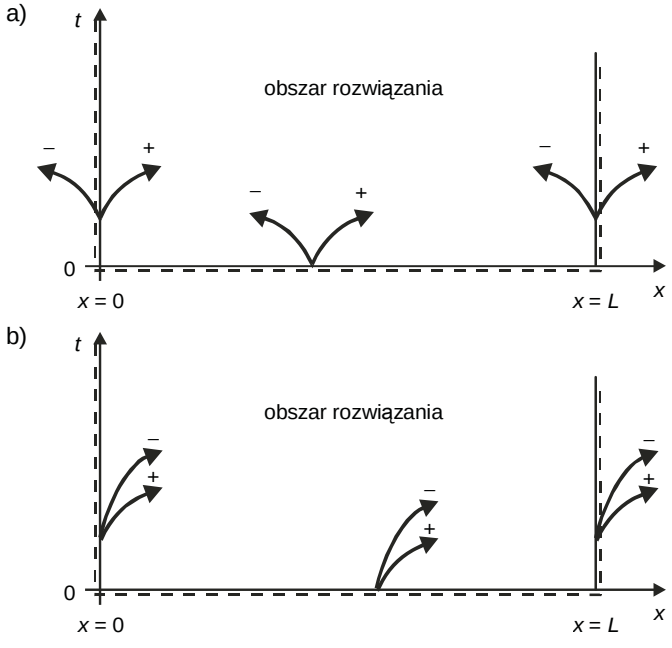
$U(x, t = 0) = U_p(x)$, $H(x, t = 0) = H_p(x)$ dla $0 \le x \le L$,

— warunki brzegowe
$U(x = 0, t) = U_0(t)$ lub $H(x = 0, t) = H_0(t)$ dla $t \geq 0$
i $U(x = L, t) = U_L(t)$ lub $H(x = L, t) = H_L(t)$ dla $t \geq 0$.

Natomiast w warunkach przepływu podkrytycznego (rwącego), gdy liczba Froude'a jest większa od jedności ($F_r > 1$), zachodzi relacja:

$$U > \sqrt{gH} \ .$$

Wynika z niej, że obydwie charakterystyki mają podobne nachylenie. Ich układ przedstawiono na rys. 7.8b. Poprawnie zadane warunki na granicach są następujące:
— warunki początkowe
$U(x, t = 0) = U_p(x)$ i $H(x, t = 0) = H_p(x)$ dla $0 \leq x \leq L$,
— warunki brzegowe
na brzegu $x = 0$ zadaje się dwa warunki:
$U(x = 0, t) = U_0(t)$ i $H(x = 0, t) = H_0(t)$ dla $t > 0$,
zaś na brzegu $x = L$ nie zadaje się żadnego warunku (żadna charakterystyka nie wchodzi do obszaru rozwiązania).



Rys. 7.8. Układ charakterystyk układu de Saint-Venanta przy przepływie spokojnym (a) i rwącym (b)

W przypadku równania adwekcji (7.17) charakterystyki (7.39) o postaci:

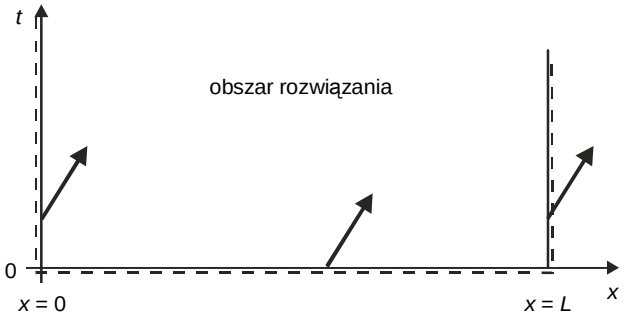
$$\frac{dx}{dt} = U$$

mają nachylenie zależne od znaku prędkości U. Przy $U > 0$ wyglądają one jak na rys. 7.9.

Warunki graniczne należy zadać następująco:

— warunek początkowy
$C(x, t = 0) = C_p(x)$ dla $0 \le x \le L$,

— warunki brzegowe
$x = 0$: $C(0, t) = C_0(t)$ dla $t \ge 0$,
$x = L$: nie zadaje się żadnej informacji.



Rys. 7.9. Układ charakterystyk równania adwekcji

Oczywiście przy $U < 0$ nachylenie charakterystyk będzie odwrotne i w konsekwencji warunek brzegowy należy zadać na brzegu $x = L$. Zasadą jest zadawanie informacji na brzegu, przez który następuje napływ wody.

W wypadku równania parabolicznego, jak na przykład (7.6), warunki graniczne tworzyć będą:

— warunek początkowy w postaci zadanej funkcji h na granicy $t = 0$: $h(x, t) = h_p(x)$ dla $0 \le x \le L$;

— warunki brzegowe w postaci zadanej funkcji h lub jej pochodnej w punktach brzegowych obszaru $x = 0$ oraz $x = L$, czyli

$$h(x = 0, t) = h_0(t) \text{ dla } t \ge 0 \text{ i } h(x = L, t) = h_L(t) \text{ lub } \left.\frac{\partial h}{\partial x}\right|_{x=L} = \varphi_L(t) \text{ dla } t \ge 0.$$

Zagadnienie rozwiązania równania różniczkowego cząstkowego typu parabolicznego jest również tzw. zagadnieniem początkowo-brzegowym.

Inaczej formułuje się warunki graniczne dla równań typu eliptycznego, takich jak równanie Poissona lub Laplace'a. Jednoznaczność rozwiązania tego typu równania wymaga, aby poszukiwana funkcja spełniała w obszarze rozwiązania to równanie oraz zadane warunki na brzegu obszaru. Nie mówi się tutaj nic o warunku początkowym, gdyż w wymienionych równaniach nie występuje czas. Zagadnienie rozwiązania równania różniczkowego cząstkowego typu eliptycznego jest tzw. zagadnieniem brzegowym. Wyróżnia się trzy zasadnicze rodzaje zagadnienia brzegowego:

a) zagadnienie brzegowe I rodzaju,
b) zagadnienie brzegowe II rodzaju,
c) zagadnienie brzegowe III rodzaju.

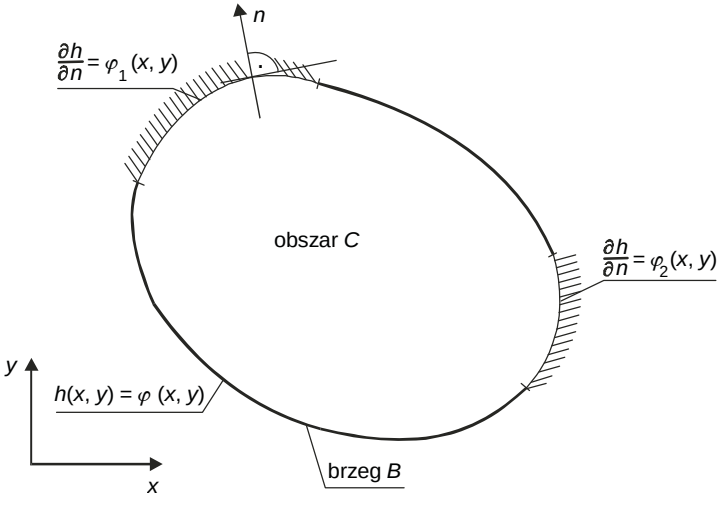
Ad a)

Na brzegu B obszaru C dana jest funkcja $\varphi(x, y)$. Znaleźć funkcję $h(x, y)$ spełniającą we wnętrzu obszaru C równanie na przykład (7.5) i przyjmującą na brzegu B dane wartości $\varphi(x, y)$, czyli $h(x, y) = \varphi(x, y)$ dla $x, y \in B$ (rys. 7.10).

Ad b)

Na brzegu B obszaru C dana jest funkcja $\varphi_1(x, y)$. Znaleźć funkcję $h(x, y)$ spełniającą we wnętrzu obszaru równanie (7.5), taką że jej pochodna w kierunku normalnym zewnętrznym do brzegu B w każdym jego punkcie przyjmuje wartość φ_1, czyli

$$\frac{\partial h}{\partial n} = \varphi_1(x, y) .$$



Rys. 7.10. Schemat obszaru C ograniczonego brzegiem B oraz rodzaje warunków brzegowych

Ad c)

Zagadnienie III rodzaju występuje wtedy, gdy na części brzegu B zadany jest warunek pierwszego rodzaju (tj. funkcja), a na pozostałej części brzegu zadany jest warunek drugiego rodzaju (tzn. pochodna). Ten typ zagadnienia jest najczęściej spotykanym w problemach filtracji.

Mając na uwadze znaczenie funkcji $h(x, y)$ oznaczającej ciśnienie piezometryczne, możemy łatwo zinterpretować fizyczny sens powyższych sformułowań. Widzimy, że w zagadnieniu I rodzaju na całym obwodzie obszaru C zadane jest ciśnienie piezometryczne $h = \varphi(x, y)$. Poszukiwany jest rozkład tego ciśnienia we wnętrzu obszaru C, zależny od wartości na brzegu oraz kształtu obszaru C. W zagadnieniu II rodzaju na całym brzegu zadane jest natężenie przepływu, co wynika z prawa Darcy'ego, zgodnie z którym

$$v = k\frac{\partial h}{\partial n}, \quad \text{skąd} \quad \frac{\partial h}{\partial n} = v/k = \varphi_1(x, y) ,$$

gdzie: k – współczynnik filtracji,
v – prędkość filtracji.

W zagadnieniu III rodzaju na części brzegu obszaru wymuszane jest ciśnienie piezometryczne, a na pozostałej części wymuszany jest przepływ.

Bardzo często spotykany jest przypadek, kiedy brzeg obszaru jest nieprzepuszczalny. Konsekwencją braku przepływu jest warunek:

$$\varphi_1 = 0, \text{ czyli } \frac{\partial h}{\partial n} = 0 .$$

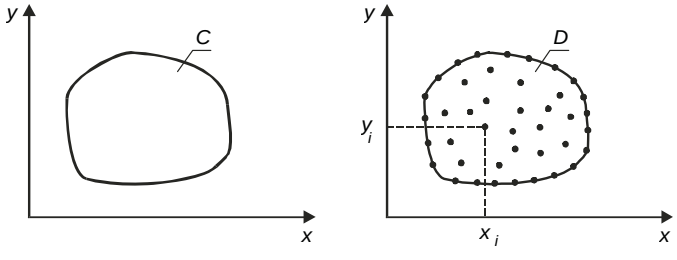
Zadane na brzegach warunki określa się jako:
— warunki Dirichleta, gdy zadana jest funkcja,
— warunki Neumanna, gdy zadana jest pochodna normalna do brzegu,
— warunki mieszane, gdy na części brzegu zadany jest warunek Dirichleta, a na części – warunek Neumanna.

Z tego powodu sformułowane wyżej typy zagadnień brzegowych nazywa się odpowiednio: zagadnieniem Dirichleta, Neumanna, mieszanym. Należy dodać, że zagadnienie Neumanna nie ma jednoznacznego rozwiązania z wyjątkiem przypadku, kiedy rozwiązanie powinno spełniać dodatkowy znany warunek, na przykład wynikający z zasady zachowania masy.

Trzeci warunek poprawnego formułowania problemu rozwiązania równania różniczkowego cząstkowego wymaga ciągłej zależności rozwiązania od zadanych warunków dodatkowych. Spełnienie tego warunku zapewnia, że niewielka zmiana zadanych warunków początkowych lub brzegowych wywoła również niewielką zmianę rozwiązania. W przypadku drastycznej zmiany rozwiązania warunek ciągłej zależności nie jest spełniony.

Rozwiązania analityczne równań różniczkowych cząstkowych są na ogół trudne do uzyskania, zwłaszcza dla bardziej skomplikowanych przypadków, z jakimi zwykle spotykamy się w praktyce inżynierskiej. Stąd do ich rozwiązywania stosuje się metody numeryczne. Na szeroką skalę metody te weszły w użycie wraz z rozpowszechnieniem się komputerów. Metody numeryczne są niezwykle skutecznym narzędziem, któremu poddają się zarówno równania cząstkowe liniowe, jak i nieliniowe, bardzo często występujące w zagadnieniach technicznych.

Jak wiadomo z poprzednich rozdziałów, podstawową cechą metod numerycznych jest operowanie liczbami. Liczby reprezentują zarówno dane, jak i wyniki. Ta cecha implikuje sposób postępowania w trakcie aplikacji metod numerycznych. Ponieważ równania różniczkowe o pochodnych cząstkowych wyprowadzane są przy założeniu ciągłości ośrodka, zaś metody numeryczne mogą być stosowane tylko w ośrodkach dyskretnych, konieczne jest zastąpienie obszaru ciągłego, w którym poszukuje się rozwiązania, obszarem dyskretnym. Proces ten nazywa się dyskretyzacją. W jego wyniku obszar ciągły C zastępuje się zbiorem wyizolowanych punktów, zwanych węzłami (rys. 7.11). Tworzą one obszar dyskretny D.



Rys. 7.11. Transformacja obszaru ciągłego

Położenie każdego węzła w przyjętym układzie definiują jego współrzędne. Metody numeryczne umożliwiają obliczenie wartości rozwiązania równania różniczkowego cząstkowego tylko w węzłach obszaru dyskretnego. Fakt ten wymaga wykonania drugiej, typowej dla metod numerycznych, operacji, a mianowicie zastąpienia równania różniczkowego cząstkowego lub układu takich równań układem równań algebraicznych, w którym niewiadomymi będą wartości rozwiązania w węzłach:

$$\frac{\partial u}{\partial t} + \cdots = 0 \rightarrow \mathbf{AX} = \mathbf{B},$$

gdzie: **A** – macierz współczynników układu równań,
X – wektor niewiadomych utworzony przez nieznane wartości rozwiązania w węzłach,
B – wektor wyrazów wolnych.

Proces ten nazywa się aproksymacją równania różniczkowego cząstkowego. Rozwiązując otrzymany układ równań algebraicznych, uzyskujemy rozwiązanie równania różniczkowego. Zwykle rozwiązanie to ma charakter przybliżony ze względu na błędy wprowadzone w procesie aproksymacji. Tylko w szczególnych przypadkach metody numeryczne dają rozwiązania identyczne z rozwiązaniem dokładnym. Opisany sposób postępowania jest wspólny dla wszystkich metod numerycznych. Różnice pomiędzy poszczególnymi metodami polegają na różnych sposobach aproksymacji równań.

Formułując problem numerycznego rozwiązania równania różniczkowego o pochodnych cząstkowych, należy zwrócić uwagę na jego poprawność. Przez analogię do warunków poprawnego formułowania problemu rozwiązania równania różniczkowego można sformułować warunki poprawności dla numerycznego rozwiązania. Są one następujące (Fletcher, 1991):

— rozwiązanie numeryczne równania istnieje,
— rozwiązanie numeryczne równania jest jednoznaczne,
— rozwiązanie numeryczne równania zależy w sposób ciągły od dodatkowych warunków zadanych na granicach obszaru rozwiązania.

W konsekwencji, poprawnie sformułowany problem numerycznego rozwiązania równania różniczkowego cząstkowego da wynik bliski wynikowi dokładnemu poprawnie sformułowanego problemu rozwiązania równania różniczkowego.

Istnieje szereg metod numerycznego rozwiązywania równań różniczkowych cząstkowych. Do najbardziej znanych należy metoda różnic skończonych oraz metoda elementów skończonych.

7.3. Metoda różnic skończonych

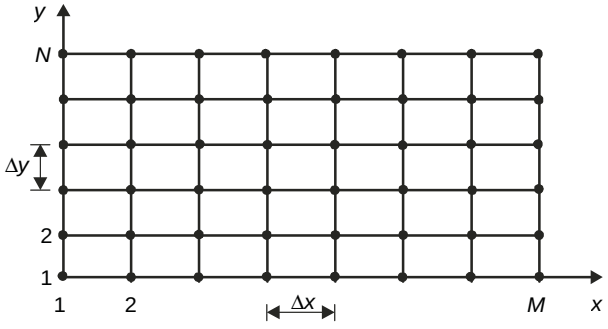
Jedną z metod numerycznych najczęściej stosowanych do rozwiązania równań różniczkowych cząstkowych jest metoda różnic skończonych. W tej metodzie obszar ciągły C zastępuje się obszarem dyskretnym D, złożonym z węzłów siatki, jaką pokrywa się obszar C.

Obszar dyskretny tworzy zbiór punktów otrzymanych w przecięciach rodzin prostych równoległych do osi układu współrzędnych (Michlin i Smolicki, 1972). W przypadku płaskim na płaszczyźnie (x, y) obszar ciągły zastępujemy zbiorem punktów – węzłów otrzymanych w miejscu przecięcia prostych równoległych do osi x oraz prostych równoległych do osi y (rys. 7.12). Proste te oddalone są od siebie odpowiednio o wartości Δx i Δy. Mogą one być stałe: $\Delta x = \text{const}$, $\Delta y = \text{const}$, lub zmienne. Ponadto mogą one być jednakowe. Zależnie od zastosowanego podejścia, otrzymuje się siatkę prostokątną lub kwadratową.

W otrzymanym w ten sposób obszarze dyskretnym nie można operować równaniami różniczkowymi. Do obliczenia poszukiwanych wartości funkcji w węzłach siatki służą tzw. równania różnicowe. Są to równania algebraiczne, które otrzymuje się w wyniku aproksymacji pochodnych występujących w równaniu różniczkowym. Podstawą aproksymacji jest rozwinięcie funkcji w szereg Taylora.

Rozpatrzmy funkcję $u(x)$ ciągłą i różniczkowalną w otoczeniu punktu o współrzędnej x_0. Funkcję tę, w otoczeniu punktu x_0 o promieniu Δx, można rozwinąć w szereg Taylora

$$u(x_0+\Delta x) = y(x_0) + \sum_{m=1}^{\infty} \frac{\Delta x^m}{m!} \frac{\partial^m u}{\partial x^m}\bigg|_{x_0} . \tag{7.40}$$



Rys. 7.12. Dyskretyzacja obszaru ciągłego w metodzie różnic skończonych

Dla dodatniej i ujemnej wartości Δx powyższy wzór rozpisuje się odpowiednio:

$$u(x_0+\Delta x) = u(x_0) + \frac{\Delta x}{1!}\frac{\partial u}{\partial x}\bigg|_{x_0} + \frac{\Delta x^2}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{x_0} + \frac{\Delta x^3}{3!}\frac{\partial^3 u}{\partial x^3}\bigg|_{x_0} + \dots . \tag{7.41}$$

i

$$u(x_0-\Delta x) = u(x_0) - \frac{\Delta x}{1!}\frac{\partial u}{\partial x}\bigg|_{x_0} + \frac{\Delta x^2}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{x_0} - \frac{\Delta x^3}{3!}\frac{\partial^3 u}{\partial x^3}\bigg|_{x_0} + \dots . \tag{7.42}$$

Równanie (7.41) można zapisać następująco:

$$u(x_0+\Delta x) = u(x_0) + \Delta x\frac{\partial u}{\partial x}\bigg|_{x_0} + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}\bigg|_{x_0} + O(\Delta x^3) . \tag{7.43}$$

Człon $O(\Delta x^3)$ oznacza resztę wzoru Taylora, wynikającą z obcięcia go po wyrazie z pochodną II rzędu. Informuje on o zmianie błędu wynikającego z faktu obcięcia szeregu wykorzystanego do obliczenia wartości funkcji $u(x_0 + \Delta x)$ w zależności od zmiany wartości Δx. Na przykład dwukrotne zwiększenie Δx spowoduje ośmiokrotny wzrost błędu, natomiast dwukrotne zmniejszenie Δx wywoła ośmiokrotną redukcję błędu obcięcia (patrz dodatek).

Wyznaczmy z szeregu (7.41) pochodną I rzędu w punkcie x_0. Jest ona równa

$$\frac{\partial u}{\partial x}\bigg|_{x_0} = \frac{u(x_0+\Delta x)-u(x_0)}{\Delta x} - \frac{\Delta x}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{x_0} - \frac{\Delta x^2}{3!}\frac{\partial^3 u}{\partial x^3}\bigg|_{x_0} - \cdots . \tag{7.44}$$

Jeśli więc zastąpimy pochodną wyrażeniem

$$\frac{\partial u}{\partial x}\bigg|_{x_0} \approx \frac{u(x_0+\Delta x)-u(x_0)}{\Delta x}, \tag{7.45}$$

to będzie ono aproksymowało ją z dokładnością $O(\Delta x)$. Oznacza to, że błąd aproksymacji będzie zmieniał się liniowo ze zmianą Δx. Formuła (7.45) nazywana jest aproksymacją pierwszej pochodnej ilorazem różnicowym przednim. W analogiczny sposób można wyznaczyć aproksymację pochodnej I rzędu z szeregu (7.42). Otrzymamy:

$$\left.\frac{\partial u}{\partial x}\right|_{x_0} \approx \frac{u(x_0)-u(x_0-\Delta x)}{\Delta x}. \tag{7.46}$$

Ten sposób aproksymacji nazywa się aproksymacją wstecznym ilorazem różnicowym i zapewnia dokładność $O(\Delta x)$.

Jeżeli od równania (7.41) odejmiemy równanie (7.42) i odrzucimy człony zawierające pochodne III rzędu i wyższe, otrzymamy

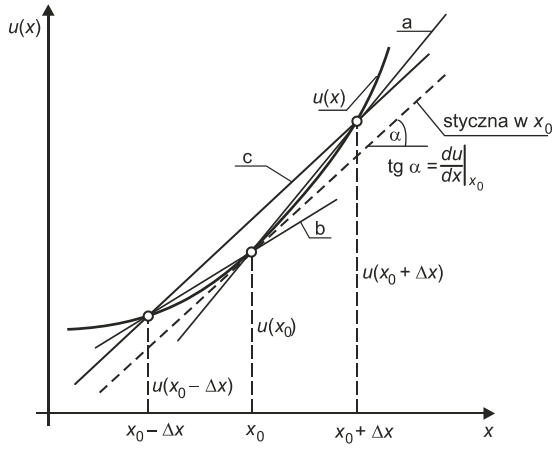
$$u(x_0+\Delta x)-u(x_0-\Delta x) \approx 2\Delta x\left.\frac{\partial u}{\partial x}\right|_{x_0}, \tag{7.47}$$

skąd

$$\left.\frac{\partial u}{\partial x}\right|_{x_0} \approx \frac{u(x_0+\Delta x)-u(x_0-\Delta x)}{2\Delta x}. \tag{7.48}$$

W tym wypadku pochodna w punkcie x_0 aproksymowana jest za pomocą ilorazu różnicowego centralnego z dokładnością $O(\Delta x^2)$.

Graficzną interpretację aproksymacji pierwszej pochodnej w punkcie x_0 przedstawiono na rys. 7.13.



Rys. 7.13. Aproksymacja pierwszej pochodnej funkcji $u(x)$ w punkcie x_0 ilorazem różnicowym: a – przednim, b – wstecznym, c – centralnym

Prosta przechodząca przez punkty x_0 i x_0-Δx stanowi przybliżenie stycznej w punkcie x_0, określone za pomocą ilorazu różnicowego wstecznego, prosta przechodząca przez x_0 i $x_0 + \Delta x$ – przybliżenie określone za pomocą ilorazu różnicowego przedniego, natomiast prosta przechodząca przez $x_0 - \Delta x$ i $x_0 + \Delta x$ – przybliżenie określone za pomocą ilorazu różnicowego centralnego. Jest oczywiste, że najdokładniejsza jest aproksymacja pochodnej za pomocą ilorazu centralnego (wzór 7.48). Wynika to z faktu, że w rozwinięciu w szereg

Taylora odrzucone zostają w tym wypadku jedynie wyrazy z pochodnymi III rzędu i wyższymi. Jest to również widoczne na rysunku.

Zauważmy, ze szeregi (7.41) i (7.42) można wykorzystać do aproksymacji pochodnej II rzędu. Mianowicie, dodając do siebie równania (7.41) i (7.42) i pomijając wyrazy zawierające pochodne IV rzędu i wyższe, otrzymujemy

$$u(x_0 + \Delta x) + u(x_0 - \Delta x) \approx 2u(x_0) + \Delta x^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{x_0}, \tag{7.49}$$

skąd

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{x_0} \approx \frac{u(x_0 - \Delta x) - 2u(x_0) + u(x_0 + \Delta x)}{\Delta x^2}. \tag{7.50}$$

Powyższe aproksymacje otrzymano wprost z rozwinięcia funkcji w szereg Taylora. Bardziej ogólną metodą określania ilorazów różnicowych, aproksymujących pochodne, jest zastosowanie ogólnego wyrażenia na aproksymację pochodnej. W przypadku pierwszej pochodnej może ono mieć postać:

$$\left.\frac{\partial u}{\partial x}\right|_j = a u_{j-1} + b u_j + c u_{j+1} + O(\Delta x^m), \tag{7.51}$$

gdzie: j – indeks węzła o współrzędnej x_j,
a, b, c – współczynniki, które należy określić,
$O(\Delta x^m)$ – wskazuje dokładność wynikającą z zastosowanej aproksymacji.

Wykorzystując szereg Taylora (7.40) do wyznaczenia wartości u_{j-1} oraz u_{j+1}, można napisać:

$$\begin{aligned} &a u_{j-1} + b u_j + c u_{j+1} = \\ &= (a+b+c)u_j + (-a+c)\Delta x \left.\frac{\partial u}{\partial x}\right|_j + (a+c)\frac{\Delta x^2}{2}\left.\frac{\partial^2 u}{\partial x^2}\right|_j + (-a+c)\frac{\Delta x^3}{6}\left.\frac{\partial^3 u}{\partial x^3}\right|_j + \cdots. \end{aligned} \tag{7.52}$$

Porównując lewą stronę (7.51) z prawą stroną (7.52), otrzymujemy związki na nieznane współczynniki:

$$a + b + c = 0,\ (-a + c)\,\Delta x = 1,$$

z których dla dowolnego c można wyznaczyć a oraz b:

$$a = c - \frac{1}{\Delta x}, \qquad b = -2c + \frac{1}{\Delta x}.$$

Przyjęcie c tak, aby w (7.52) trzeci człon zniknął, daje najdokładniejszą aproksymację możliwą do uzyskania dla przypadku trzech występujących w równaniu parametrów. Zatem ostatecznie otrzymujemy:

$$c = -a = \frac{1}{2\Delta x} \quad \text{i} \quad b = 0.$$

Podstawienie powyższych wartości do (7.52) daje

$$\left.\frac{\partial u}{\partial x}\right|_j = \frac{1}{2\Delta x}(-u_{j-1} + u_{j+1}) - \frac{\Delta x^2}{6}\left.\frac{\partial^3 u}{\partial x^3}\right|_j + \cdots. \tag{7.53}$$

W ten sposób otrzymano iloraz różnicowy z różnicą centralną, czyli równanie (7.50), który daje błąd rzędu $O(\Delta x^2)$.

Postępując w podobny sposób, można wyznaczyć aproksymację drugiej pochodnej. Z porównania równania (7.51) i (7.52) otrzymujemy:

$$a + b + c = 0, \qquad (-a + c)\,\Delta x = 0, \quad (a + c)\frac{\Delta x^2}{2} = 1\,.$$

Wynika stąd, że

$$a = \frac{1}{\Delta x^2}, \quad b = -\frac{2}{\Delta x^2}, \quad c = \frac{1}{\Delta x^2},$$

a zatem

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_j = \frac{u_{j-1} - 2u_j + u_{j+1}}{\Delta x^2} + O(\Delta x^2)\,. \tag{7.54}$$

Przedstawiona metoda może być wykorzystana do konstrukcji jednostronnych (niesymetrycznych) ilorazów różnicowych lub dla określania formuł różnicowych na siatce niejednorodnej. Na przykład dla trzypunktowego niesymetrycznego ilorazu różnicowego ogólną formułę zapisujemy w postaci

$$\left.\frac{\partial u}{\partial x}\right|_j = au_j + bu_{j+1} + cu_{j+2} + O(\Delta x^m)\,. \tag{7.55}$$

Określając u_{j+1} oraz u_{j+2} z rozwinięcia funkcji w szereg Taylora wokół punktu j, otrzymujemy

$$\begin{aligned}\left.\frac{\partial u}{\partial x}\right|_j &= (a+b+c)u_j + (b\Delta x + c2\Delta x)\left.\frac{\partial u}{\partial x}\right|_j + \\ &+ \left(\frac{b\Delta x^2}{2} + \frac{c(2\Delta x)^2}{2}\right)\left.\frac{\partial^2 u}{\partial x^2}\right|_j + \cdots.\end{aligned} \tag{7.56}$$

Z (7.56) wynika, że celem otrzymania jak najmniejszego błędu, na parametry a, b i c należy nałożyć następujące warunki:

$$a + b + c = 0, \quad b\Delta x + c2\Delta x = 1, \quad \frac{b\Delta x^2}{2} + \frac{c(2\Delta x)^2}{2} = 0,$$

które umożliwiają określenie wartości parametrów:

$$a = -\frac{1{,}5}{\Delta x}, \quad b = \frac{2}{\Delta x}, \quad c = -\frac{0{,}5}{\Delta x}.$$

Zatem poszukiwana formuła ma postać:

$$\left.\frac{\partial u}{\partial x}\right|_j = \frac{-1{,}5u_j + 2u_{j+1} - 0{,}5u_{j+2}}{\Delta x} - \frac{\Delta x^2}{3}\cdot\left.\frac{\partial^3 u}{\partial x^3}\right|_j + \cdots. \tag{7.57}$$

Wprowadzany przez nią błąd jest rzędu $O(\Delta x^2)$. W podobny sposób można konstruować dowolne formuły aproksymujące pochodne występujące w równaniach różniczkowych cząstkowych.

Ostatecznym efektem zastąpienia pochodnych odpowiednimi ilorazami różnicowymi jest równanie algebraiczne, przybliżające rozwiązywane równanie różniczkowe w punkcie, w którym dokonano aproksymacji. Powtórzenie tej operacji dla kolejnych węzłów prowadzi do układu równań algebraicznych, w którym występuje większa liczba wartości węzłowych poszukiwanej funkcji niż równań. Układ zamyka się, wprowadzając dodatkowe równania wynikające z zadanych warunków brzegowych.

7.4. Metoda elementów skończonych

Obok opisanej w poprzednim podrozdziale metody różnic skończonych, do rozwiązywania równań różniczkowych cząstkowych można wykorzystać metodę elementów skończonych. Celem postępowania w tej metodzie jest, podobnie jak w metodzie różnic skończonych, przybliżone rozwiązanie równania lub układu równań różniczkowych cząstkowych.

Rozpatrzmy zatem dowolne równanie różniczkowe ważne w obszarze ciągłym C (rys. 7.14), którym może być każde z równań wymienionych w rozdziale 7.1, zapisane symbolicznie

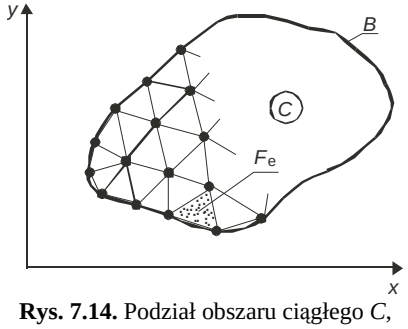
$$\Omega(h) = 0, \tag{7.58}$$

z warunkami brzegowymi

$$B(h) = 0, \tag{7.59}$$

gdzie h reprezentuje poszukiwaną funkcję.

Jeśli w zagadnieniu wystąpi układ równań, a więc i kilka funkcji niewiadomych, w (7.58) wystąpi wektor o odpowiednim rozmiarze. W celu uzyskania przybliżonego rozwiązania zagadnienia (7.58) obszar rozwiązania C dzieli się na mniejsze podobszary – „elementy skończone” (rys. 7.14). Zakłada się, że elementy łączą się ze sobą tylko w skończonej liczbie punktów znajdujących się na ich obwodzie. Punkty te nazywa się węzłami. Nieznaną funkcję h aproksymuje się wewnątrz obszaru C wyrażeniem



Rys. 7.14. Podział obszaru ciągłego C, ograniczonego brzegiem B, na elementy skończone

$$\hat{h} = \mathbf{N}\,\mathbf{h}\,, \tag{7.60}$$

gdzie: $\hat{h}$ – przybliżenie funkcji h,
$\mathbf{N} = [..., N_i, N_j, N_k, ...]$,
$\mathbf{h} = (..., h_i, h_j, h_k, ...)^T$.

W równaniu tym wektor $\mathbf{h}$ jest ciągiem wartości funkcji h w węzłach obszaru. Jeśli rozwiązujemy zagadnienie nieustalone, w którym jedną ze zmiennych niezależnych jest czas, wówczas wartości węzłowe będą funkcjami czasu. Natomiast $\mathbf{N}$ jest zbiorem funkcji zależnych tylko od współrzędnych przestrzennych. Funkcje te należy tak dobrać, aby równanie (7.60) było spełnione, gdy wstawi się do niego współrzędne odpowiednich węzłów obszaru. Są to tzw. funkcje kształtu lub funkcje bazowe.

Jeżeli funkcja $\hat{h}$ określona przez (7.60) jest przybliżeniem rozwiązania dokładnego h w obszarze rozwiązania C, to nie spełni ona dokładnie równania różniczkowego (7.58) i będzie

$$\Omega(\hat{h}) = R \neq 0\,. \tag{7.61}$$

Rozwiązaniem najdokładniejszym będzie takie, które zredukuje resztę R do wartości najmniejszych w całym obszarze rozwiązania. Zatem powinno ono spełniać warunek

$$\int_C WR dC = 0 , \tag{7.62}$$

gdzie: W jest pewną funkcją wagową zależną tylko od współrzędnych.

Jeżeli poszukiwana funkcja h aproksymowana jest według (7.60) na podstawie M nieznanych jej wartości w M węzłach, to można wybrać M – liniowo niezależnych funkcji wagowych W_i, uzyskując w ten sposób odpowiednią liczbę równań

$$\int_C W_i \Omega(\hat{h})\ dC = \int_C W_i \Omega\left(\mathbf{Nh}\right) dC = 0 , \quad i = 1, 2, ..., M, \tag{7.63}$$

z których należy wyznaczyć **h** jako rozwiązanie problemu.
Jeżeli przyjmiemy, że $W_i = N_i$, tzn. że funkcjami wagowymi będą funkcje kształtu, to proces taki, znany pod nazwą procedury Galerkina dla metody ważonej residualnej, prowadzi z reguły do najlepszego przybliżenia (Zienkiewicz, 1972).

Zatem procedura Galerkina sprowadza rozwiązanie problemu określonego równaniem różniczkowym (7.58) do następującego układu równań:

$$\int_C \Omega\left(\mathbf{Nh}\right) N_i\ dC = 0 , \tag{7.64}$$

gdzie: $i = 1, 2, 3, ..., M$,
M – liczba węzłów w obszarze C.

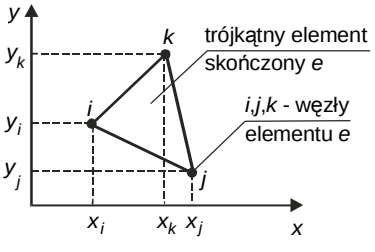
Dokładniejszy opis metody wraz z wyprowadzeniem procedury Galerkina podaje Zienkiewicz (1972).

Przyjmuje się dalej, że całka (7.64) liczona w całym obszarze C może być zastąpiona sumą całek liczonych w poszczególnych elementach obszaru. Zatem

$$\int_C \Omega\left(\mathbf{Nh}\right)\mathbf{N}^T dC = \sum_{e=1}^{m} \int_{F_e} \Omega(\mathbf{N}_e\mathbf{h}_e)\mathbf{N}_e^T dF_e = 0 , \tag{7.65}$$

gdzie: m – całkowita liczba elementów, na jaką podzielono obszar C,
F_e – reprezentuje powierzchnię „elementu skończonego" o numerze e, jak na rys. 7.15.

Oznacza to możliwość obliczania całki w każdym elemencie oddzielnie. Zsumowanie wyników całkowania prowadzi do układu finalnego. Nie wnikając w tej chwili w szczegóły wyznaczenia poszczególnych całek układu (7.65), należy stwierdzić, że ważną rolę w ich obliczaniu odgrywają funkcje kształtu, które definiują aproksymację rozwiązania poprzez podział obszaru na elementy skończone.

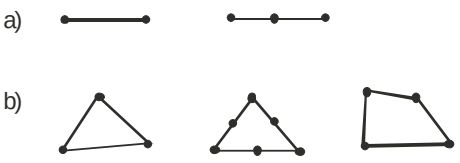


Rys. 7.15. Trójkątny element skończony

Wybór funkcji kształtu związany jest ściśle z warunkami, jakie muszą one spełniać, oraz z kształtem zastosowanych elementów skończonych. Od funkcji tych wymaga się, aby:

— były ciągłe w elemencie,

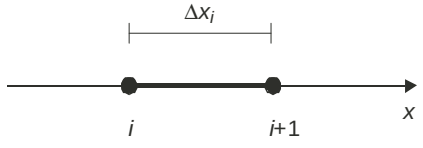
— zapewniały ciągłość funkcji *h* na styku elementów.



Rys. 7.16. Niektóre kształty elementów skończonych

Dla zagadnień jednowymiarowych elementem skończonym może być np. odcinek zbudowany na dwu lub więcej węzłach (rys. 7.16a), natomiast dla zagadnień dwuwymiarowych będzie nim dowolna figura płaska, jak na rys. 7.16b.

W dalszych rozważaniach ograniczymy się do omówienia niektórych funkcji kształtu dla elementów jednowymiarowych oraz elementów dwuwymiarowych trójkątnych. Nie wyczerpuje to oczywiście pełnej listy możliwych funkcji kształtu, sposobów ich tworzenia oraz możliwości metody – jest tylko pewną jej ilustracją. Szeroko problem ten przedstawiony jest przez Zienkiewicza (1972).



Rys. 7.17. Element liniowy

Jak wspomniano wcześniej, dla zagadnień jednowymiarowych naturalnym elementem skończonym jest odcinek. Rozpatrzmy zatem jednowymiarowy element *e* będący odcinkiem ograniczonym węzłami *i* oraz *i* + 1 o długości $\Delta x_i = x_{i+1} - x_i$, jak na rys. 7.17. W najprostszym wypadku funkcję *h* można aproksymować wewnątrz elementu za pomocą wyrażenia liniowego

$$\hat{h} = ax + b\,. \tag{7.66}$$

Współczynniki *a* i *b* można wyznaczyć, mając na uwadze fakt, że w węzłach na końcach elementów wartość wyrażenia (7.66) musi być równa wartości funkcji *h* w tych węzłach. Otrzymujemy następujące zależności:

$$h_i = ax_i + b,$$

$$h_{i+1} = ax_{i+1} + b,$$

Rozwiązanie powyższego układu pozwala napisać formułę (7.66) w postaci

$$\hat{h} = \frac{h_{i+1} - h_i}{x_{i+1} - x_i} x - \frac{h_{i+1} - h_i}{x_{i+1} - x_i} x_i + h_i\,,$$

Można ją zapisać w standardowej formie (7.60)

$$\hat{h} = \mathbf{N}_e \mathbf{h}_e\,,$$

gdzie:
$$\mathbf{N}_e = [N_i, N_{i+1}], \quad \mathbf{h}_e = \begin{Bmatrix} h_i \\ h_{i+1} \end{Bmatrix},$$

przy czym

$$N_i(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \quad N_{i+1}(x) = \frac{x - x_i}{x_{i+1} - x_i} \quad \text{dla} \quad x_i \le x \le x_{i+1}, \tag{7.67}$$

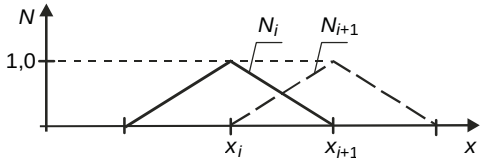
zaś indeks *e* oznacza element, w którym dokonujemy aproksymacji.

Jak widać, w tym wypadku funkcjami kształtu są liniowe wielomiany Lagrange'a znane z podrozdziału 3.3.

Analogiczną parę funkcji kształtu formułujemy w każdym elemencie. Łatwo można zauważyć, że dowolna funkcja $N_i(x)$ wystąpi w elementach współtworzonych przez węzeł i. Zatem liniowa funkcja kształtu zdefiniowana będzie następująco:

$$N_i(x) = \begin{cases} 0 & \text{dla} \quad x < x_{i-1}, \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{dla} \quad x_{i-1} \le x < x_i, \\ 1 & \text{dla} \quad x = x_i, \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{dla} \quad x_i < x \le x_{i+1}, \\ 0 & \text{dla} \quad x > x_{i+1}\ . \end{cases} \tag{7.68}$$

Warunki nałożone na tak utworzone funkcje kształtu są spełnione, bowiem:
$N_i = 1$ w węźle i oraz $N_i = 0$ w węźle $i + 1$, a także $N_{i+1} = 0$ w węźle i oraz $N_{i+1} = 1$ w węźle $i + 1$. Równanie (7.68) opisuje tzw. „funkcje daszkowe" określone na całej osi liczbowej x (rys. 7.18).



Rys. 7.18. Liniowe funkcje kształtu N_i oraz N_{i+1}

Aproksymacja funkcji h według formuły (7.60) w pojedynczym elemencie ma w tej sytuacji postać

$$\widehat{h} = N_i h_i + N_{i+1} h_{i+1}\,. \tag{7.69}$$

Różniczkowanie funkcji kształtu w takim elemencie prowadzi do wyrażeń

$$\frac{dN_i}{dx} = \frac{-1}{x_{i+1} - x_i} = -\frac{1}{\Delta x_i}; \quad \frac{dN_{i+1}}{dx} = \frac{1}{x_{i+1} - x_i} = \frac{1}{\Delta x_i}, \tag{7.70}$$

gdzie: Δx_i – długość elementu.

Natomiast całkowanie funkcji kształtu w elemencie daje wynik

$$\int_{x_i}^{x_{i+1}} N_i dx = \int_{x_i}^{x_{i+1}} \frac{x_{i+1} - x}{\Delta x_i} dx = \frac{1}{2}\Delta x_i\,. \tag{7.71}$$

Słuszny jest ponadto następujący wzór rekurencyjny, ułatwiający obliczanie całek z iloczynu funkcji kształtu

$$\int_{x_i}^{x_{i+1}} N_i^{\alpha} N_j^{\beta} dx = \int_{x_i}^{x_{i+1}} N_i^{\beta} N_j^{\alpha} dx = \frac{\alpha!\,\beta!}{(\alpha + \beta + 1)!}\Delta x_i = \frac{A}{B}\Delta x_i\,, \tag{7.72}$$

gdzie: α, β – całkowite wykładniki potęg.

Wzór ten jest bardzo użyteczny przy obliczaniu całki (7.65), w której zwykle występują iloczyny funkcji kształtu. Niektóre wartości całek ułatwiających obliczanie (7.65) przedstawiono w tabeli 7.1.

Tabela 7.1

Wartości całek z iloczynów funkcji kształtu dla elementów liniowych

$\alpha+\beta$	α	β	A	B
1	1	0	1	2
2	2 1	0 1	2 1	6
3	3 2	0 1	3 1	12
4	4 3 2	0 1 2	12 3 2	60
5	5 4 3	0 1 2	10 2 1	60
6	6 5 4 3	0 1 2 3	60 10 4 3	420

Do określenia funkcji kształtu 2. stopnia na elemencie jednowymiarowym, jak na rys. 7.19, potrzebne są 3 węzły, ponieważ wielomian stopnia 2. można poprowadzić przez 3 punkty. Można tutaj skorzystać wprost z wielomianu Lagrange'a i utworzyć bezpośrednio funkcje kształtu dla takiego elementu. Jeżeli element zbudowany jest na węzłach (*i*, *j*, *k*), to:

$$N_i(x)=\frac{(x-x_j)(x-x_k)}{(x_i-x_j)(x_i-x_k)}, \tag{7.73}$$

$$N_j(x)=\frac{(x-x_i)(x-x_k)}{(x_j-x_i)(x_j-x_k)}, \tag{7.74}$$

$$N_k(x)=\frac{(x-x_i)(x-x_j)}{(x_k-x_i)(x_k-x_j)}. \tag{7.75}$$

Przebieg powyższych funkcji wewnątrz elementu pokazano na rys. 7.19. Poszukiwana funkcja *h* przybliżona będzie wewnątrz elementu *e* w sposób następujący:

$$\hat{h}=N_i h_i+N_j h_j+N_k h_k\,. \tag{7.76}$$

W przypadku zagadnień dwuwymiarowych elementami skończonymi mogą być figury płaskie, jak np. na rys. 7.16b. Ograniczymy się tutaj do rodziny elementów trójkątnych ze względu na ich korzystne cechy. Niewątpliwą zaletą elementów o tych kształtach jest większa, niż w przypadku np. elementów kwadratowych lub prostokątnych, elastyczność w trakcie odwzorowania geometrii obszarów. Dla każdego elementu trójkątnego można, w zależności od liczby węzłów, dobrać wielomiany gwarantujące wymagane własności

funkcji kształtu. Zajmiemy się sposobem bezpośredniego tworzenia funkcji kształtu dla takich elementów, podobnie jak to uczyniliśmy dla elementów liniowych.

Najprostszą funkcją kształtu dla trójkątnego elementu skończonego e zbudowanego na węzłach (i, j, k), jak na rys. 7.15, będzie funkcja liniowa o postaci

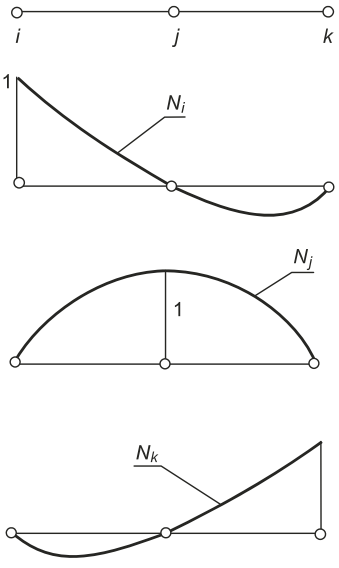
$$\widehat{h} = \alpha_0 + \alpha_1 x + \alpha_2 y \,. \tag{7.77}$$

Niech jej wartości w węzłach wynoszą odpowiednio h_i, h_j, h_k. Wartości współczynników α_0, α_1, α_2 w (7.77) należy tak dobrać, aby w węzłach i, j, k funkcja h przyjęła wartości h_i, h_j, h_k. W tym celu możemy dla kolejnych węzłów elementu zgodnie z (7.77) napisać następujące zależności:

$$h_i = \alpha_0 + \alpha_1 x_i + \alpha_2 y_i \,, \tag{7.78a}$$

$$h_j = \alpha_0 + \alpha_1 x_j + \alpha_2 y_j \,, \tag{7.78b}$$

$$h_k = \alpha_0 + \alpha_1 x_k + \alpha_2 y_k \,. \tag{7.78c}$$



Rys. 7.19. Kwadratowe funkcje kształtu dla elementu liniowego

Równania te tworzą układ, który po rozwiązaniu względem α_0, α_1, α_2 umożliwia zapisanie wyrażenia (7.77) w postaci

$$\begin{aligned}\widehat{h} = \frac{1}{2F_e}(a_i + b_i x + c_i y)h_i + \frac{1}{2F_e}(a_j + b_j x + c_j y)h_j \\ + \frac{1}{2F_e}(a_k + b_k x + c_k y)h_k,\end{aligned} \tag{7.79}$$

lub krócej w standardowej formie (7.60)

$$\widehat{h} = N_i h_i + N_j h_j + N_k h_k = \mathbf{N}_e \mathbf{h}_e \,, \tag{7.80}$$

gdzie:

$$N_i(x,y) = \frac{1}{2F_e}(a_i + b_i x + c_i y) \,, \tag{7.81a}$$

$$N_j(x,y) = \frac{1}{2F_e}(a_j + b_j x + c_j y) \,, \tag{7.81b}$$

$$N_k(x,y) = \frac{1}{2F_e}(a_k + b_k x + c_k y) \,, \tag{7.81c}$$

a F_e jest polem elementu e

$$F_e = \frac{1}{2}\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix} = \frac{1}{2}(c_k b_j - c_j b_k) \,, \tag{7.82}$$

natomiast

$$a_i = x_j y_k - x_k y_j, \quad a_j = x_k y_i - x_i y_k, \quad a_k = x_i y_j - x_j y_i, \tag{7.83a,b,c}$$

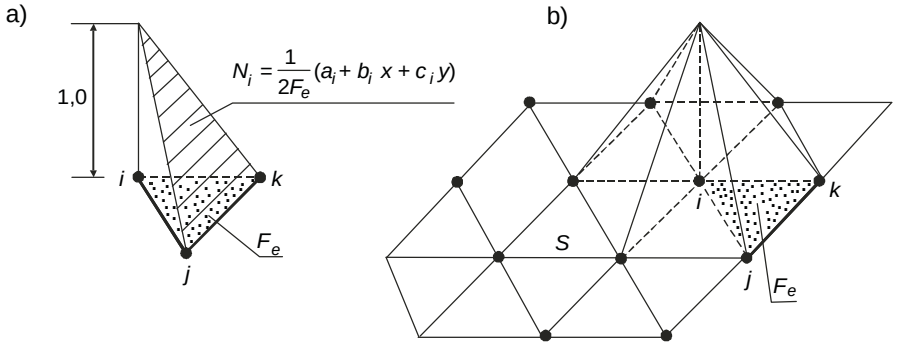
$$b_i = y_j - y_k, \quad b_j = y_k - y_i, \quad b_k = y_i - y_j, \tag{7.84a,b,c}$$

$$c_i = x_k - x_j, \quad c_j = x_i - x_k, \quad c_k = x_j - x_i. \tag{7.85a,b,c}$$

Funkcje N_i, N_j, N_k są zależne od geometrii elementu i spełniają warunki nałożone na funkcje kształtu, bowiem:

$N_i = 1$	w węźle i	oraz	$N_i = 0$	w węzłach j oraz k,
$N_j = 1$	w węźle j	oraz	$N_j = 0$	w węzłach i oraz k,
$N_k = 1$	w węźle k	oraz	$N_k = 0$	w węzłach i oraz j.

Wykres takiej liniowej funkcji kształtu dla węzła i w elemencie pokazano na rys. 7.20a. Natomiast kształt $N_i(x, y)$ w całym obszarze przedstawiono na rys. 7.20b.



Rys. 7.20. Liniowe funkcje kształtu dla elementów trójkątnych

Ma ona kształt ostrosłupa o wysokości 1 w węźle i. Jej wartość zmienia się liniowo do zera w kierunku węzłów elementów współtworzonych przez węzeł i. Poza tymi elementami w całym obszarze $N_i(x, y) = 0$.

Różniczkowanie funkcji kształtu względem zmiennych przestrzennych daje następujące wyniki:

$$\frac{\partial N_i}{\partial x} = \frac{b_i}{2F_e}, \quad \frac{\partial N_j}{\partial x} = \frac{b_j}{2F_e}, \quad \frac{\partial N_k}{\partial x} = \frac{b_k}{2F_e}, \tag{7.86a,b,c}$$

$$\frac{\partial N_i}{\partial y} = \frac{c_i}{2F_e}, \quad \frac{\partial N_j}{\partial y} = \frac{c_j}{2F_e}, \quad \frac{\partial N_k}{\partial y} = \frac{c_k}{2F_e}. \tag{7.87a,b,c}$$

W konsekwencji różniczkowanie funkcji $\widehat{h}$ względem x oraz y daje następujący wynik:

$$\frac{\partial \widehat{h}}{\partial x} = \frac{1}{2F_e}(b_i h_i + b_j h_j + b_k h_k), \tag{7.88}$$

$$\frac{\partial \widehat{h}}{\partial y} = \frac{1}{2F_e}(c_i h_i + c_j h_j + c_k h_k). \tag{7.89}$$

Z kolei różniczkowanie względem czasu prowadzi do zależności

$$\frac{\partial \hat{h}}{\partial t} = N_i \frac{dh_i}{dt} + N_j \frac{dh_j}{dt} + N_k \frac{dh_k}{dt}. \tag{7.90}$$

Całka z funkcji kształtu w elemencie *e* zgodnie z definicją jest równa objętości ostrosłupa o podstawie elementu i wysokości jednostkowej, czyli

$$\iint_{Fe} N_i(x,y)dxdy = \frac{F_e}{3}. \tag{7.91}$$

Do obliczenia całki po powierzchni elementu *e* z iloczynu funkcji kształtu podniesionych do potęgi całkowitej można wykorzystać następującą formułę rekurencyjną:

$$\iint_{F_e} N_i^{\alpha} N_j^{\beta} N_k^{\gamma} dxdy = \frac{\alpha!\beta!\gamma!}{(\alpha+\beta+\gamma+2)!} 2F_e = \frac{A}{B} F_e. \tag{7.92}$$

Niektóre wartości *A* i *B* ułatwiające obliczenie takich całek przedstawiono w tabeli 7.2.

Tabela 7.2

Wartości całek z iloczynów funkcji kształtu dla elementów trójkątnych

$\alpha+\beta+\gamma$	α	β	γ	*A*	*B*
1	1	0	0	1	3
2	2 1	0 1	0 0	2 1	12 12
3	3 2 1	0 1 1	0 0 1	6 2 1	60
4	4 3 2 2	0 1 2 1	0 0 0 1	12 3 2 1	180
5	5 4 3 3 2	0 1 2 1 2	0 0 0 1 1	60 12 6 3 2	1260
6	6 5 4 4 3 3	0 1 2 1 3 2	0 0 0 1 0 1	60 10 4 2 3 1	1680

7.5. Elementy teorii numerycznego rozwiązywania równań różniczkowych cząstkowych

Przedstawiony w podrozdziale 7..3 opis metody różnic skończonych objaśnia sposób zamiany równania różniczkowego na równoważny mu układ równań algebraicznych. Dla lepszej ilustracji toku postępowania rozpatrzmy przykład w postaci równania czystej adwekcji. Równanie tego typu, zaprezentowane w podrozdziale 7.1, opisuje na przykład adwekcyjne przenoszenie domieszki przez wodę płynącą z dużą prędkością, czyli przy dużych liczbach Reynoldsa, lub też przemieszczanie się fali wezbraniowej w kanale o znacznym spadku dna, gdy dopuszcza się stosowanie modelu fali kinematycznej. Wybór tego typu równania wydaje się odpowiedni ze względu na jego właściwości. Jest to, jak wiadomo, równanie typu hiperbolicznego. Zatem na jego przykładzie można dokonać dyskusji nie tylko ogólnych, ważnych dla równań każdego typu problemów całkowania numerycznego, ale również specyficznych, bo istotnych dla równań typu hiperbolicznego, zagadnień dyssypacji i dyspersji numerycznej.

Rozpatrzmy zatem równanie (7.17)

$$\frac{\partial C}{\partial t} + U \frac{\partial C}{\partial x} = 0 \,, \tag{7.93}$$

w którym: t jest czasem, x jest zmienną przestrzenną, $C(x, t)$ jest na przykład koncentracją rozpuszczonej domieszki w płynącej wodzie, U jest prędkością płynącej wody, czyli tzw. prędkością adwekcji.

Równanie to będziemy rozwiązywali w obszarze $0 \le x \le L$ (gdzie L jest długością odcinka cieku), $t > 0$. Załóżmy najprostszy wariant adwekcji. Woda płynie ze stałą prędkością zgodnie z kierunkiem osi x, czyli $U = \text{const} > 0$, zaś początkowy rozkład koncentracji ($t = t_0$) wzdłuż osi x ma postać trójkąta równoramiennego o wysokości jednostkowej (rys. 7.21).

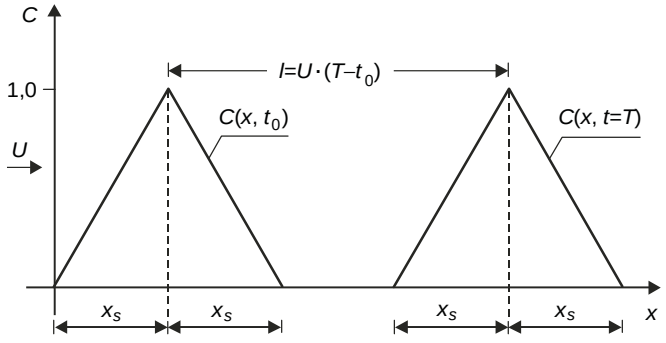
Przy znanej prędkości adwekcji U równanie (7.93) ma rozwiązanie analityczne o postaci (Abbott i Basco, 1989)

$$C(x,t) = C\left(x - \int_{t_0}^{t} U \cdot dt, t_0 \right). \tag{7.94}$$

Jeśli $U = \text{const}$, to powyższe równanie dla $t = T$ przyjmie postać

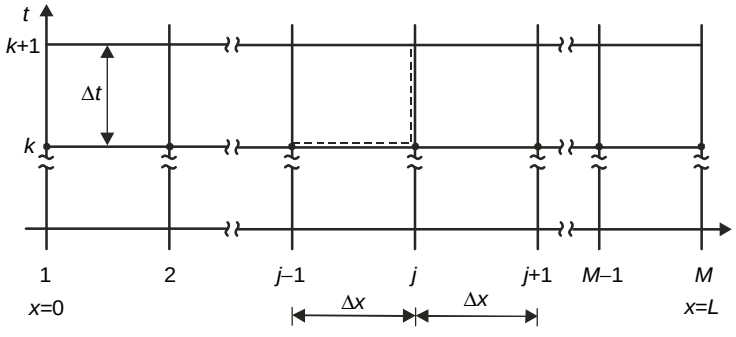
$$C(x,T) = C(x - U(T - t_0), t_0) \,. \tag{7.95}$$

Jest to formalny zapis procesu czystej translacji. Zależność tę interpretuje się następująco: koncentracja w chwili $t = T$ w dowolnym przekroju kanału o współrzędnej x równa jest koncentracji w chwili t_0, występującej w przekroju o współrzędnej $x - U(T - t_0)$. Inaczej mówiąc, cząstka, która w chwili $t = t_0$ znajdowała się w położeniu x_0, przemieści się wzdłuż osi kanału i w chwili $t = T$ znajdzie się w punkcie o współrzędnej $x_0 + U(T - t_0)$. Nawiązując do rys. 7.21, można stwierdzić, że jeśli przenoszenie domieszki rozpuszczonej w wodzie wynika jedynie z adwekcji, to początkowy rozkład koncentracji wzdłuż osi kanału będzie przemieszczał się bez deformacji jego kształtu, by w chwili $t = T$ znaleźć się w odległości wynikającej z prędkości adwekcji: $l = U(T - t_0)$. Chociaż, jak widać, równanie adwekcji (7.93) ma bardzo proste rozwiązanie analityczne, wiele bardzo ważnych informacji i wniosków wynika z jego numerycznego rozwiązania. Rozwiążmy więc równanie metodą różnic skończonych.



Rys. 7.21. Początkowy rozkład koncentracji $C(x, t_0)$ i analityczne rozwiązanie równania adwekcji dla $t = T$

Obszar rozwiązania równania ($0 \le x \le L$, $t \ge 0$) zastępujemy obszarem dyskretnym. Fragment przyjętej siatki węzłów przedstawiono na rys. 7.22. Jej wymiary wynoszą Δx = const w kierunku osi x oraz Δt = const w kierunku osi czasu. Załóżmy, że we wszystkich węzłach na poziomie czasu k znane są wartości koncentracji: C_j^k (j = 1, 2, ..., M). Celem postępowania jest określenie koncentracji na następnym poziomie czasowym k + 1: C_j^{k+1} (j = 2, 3, ..., M). W węźle pierwszym koncentracja jest znana. Definiuje ją zadany warunek brzegowy: $C_1^{k+1} = C_0(t_{k+1})$, gdzie $C_0(t)$ jest znaną funkcją.



Rys. 7.22. Siatka węzłów zastosowana do rozwiązania równania adwekcji

Dokonajmy aproksymacji pochodnych występujących w równaniu (7.93) w węźle (j, k) odpowiednio ilorazem różnicowym przednim i wstecznym:

$$\left.\frac{\partial C}{\partial t}\right|_j^k \approx \frac{C_j^{k+1} - C_j^k}{\Delta t}, \tag{7.96}$$

$$\left.\frac{\partial C}{\partial x}\right|_j^k \approx \frac{C_j^k - C_{j-1}^k}{\Delta x}. \tag{7.97}$$

Po wstawieniu powyższych formuł do równania (7.93), otrzymujemy wyrażenie

$$\frac{C_j^{k+1} - C_j^k}{\Delta t} + U \frac{C_j^k - C_{j-1}^k}{\Delta x} = 0 , \tag{7.98}$$

z którego można wyznaczyć jedyną niewiadomą C_j^{k+1}. Jest ona równa

$$C_j^{k+1} = C_j^k - \frac{U\Delta t}{\Delta x}(C_j^k - C_{j-1}^k) . \tag{7.99}$$

Wprowadźmy tzw. adwekcyjną liczbę Couranta

$$C_a = \frac{U\Delta t}{\Delta x} . \tag{7.100}$$

Jest to bezwymiarowy parametr interpretowany jako stosunek prędkości adwekcji U do tzw. prędkości na siatce: $\Delta x/\Delta t$, który jest zawsze większy od zera. Równanie (7.99) można przepisać następująco:

$$C_j^{k+1} = (1 - C_a)C_j^k + C_a \cdot C_{j-1}^k . \tag{7.101}$$

Jak już wspomniano, wartość koncentracji w pierwszym węźle na poziomie czasu $k + 1$ (C_1^{k+1}) określa zadany warunek brzegowy. Zatem przemierzając w dowolnej kolejności węzły siatki na poziomie $k + 1$, tzn. przyjmując $j = 2, 3, ..., M$, obliczamy wartości koncentracji w pozostałych węzłach.

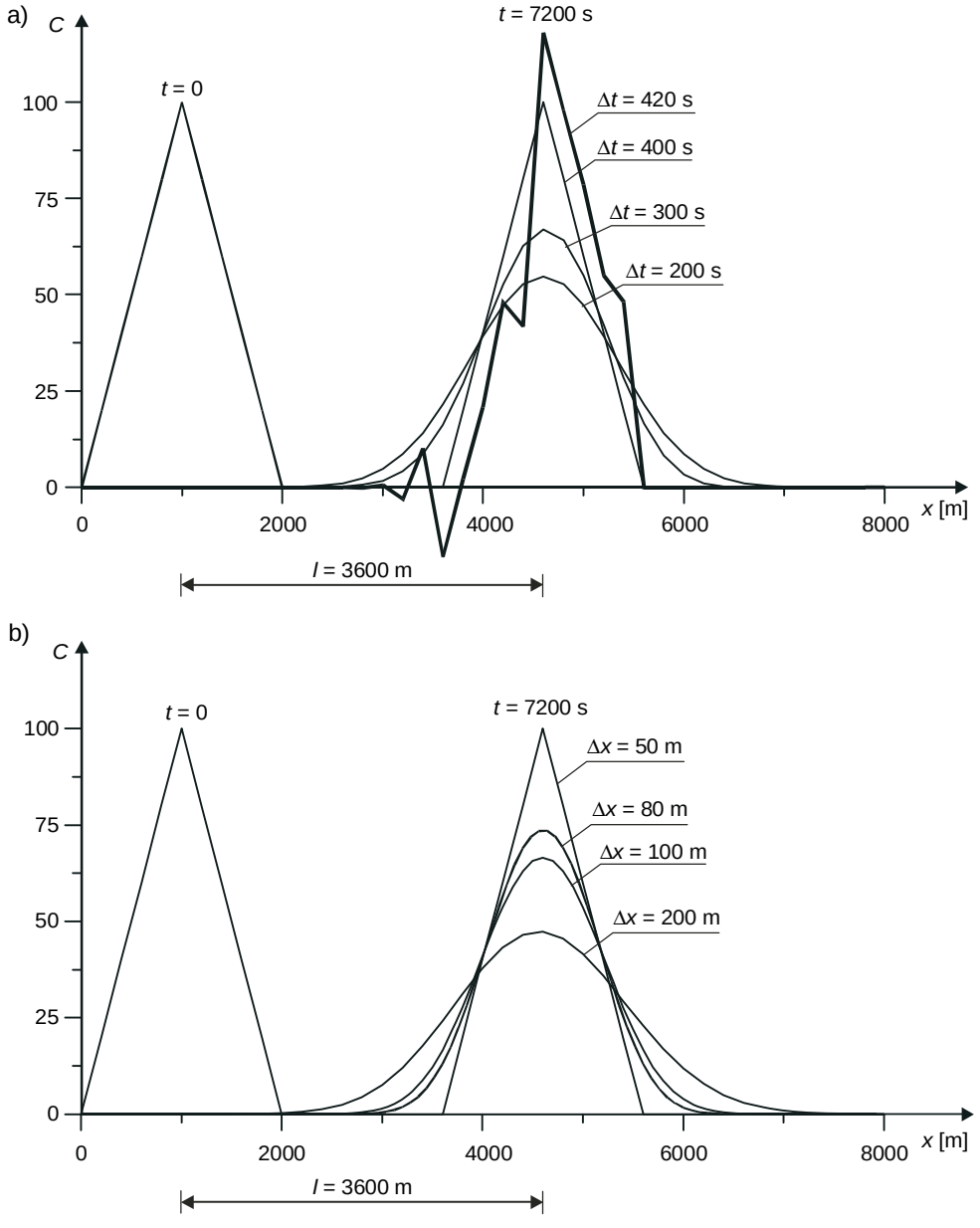
Opisany wyżej sposób rozwiązania równania adwekcji nazywa się schematem „pod prąd” (w języku angielskim: up-wind scheme). Nazwa pochodzi od sposobu aproksymacji pochodnej przestrzennej. Zawsze bierze się wartość koncentracji w węźle aproksymacji oraz w węźle sąsiednim w kierunku, skąd napływa woda (dokładniej: skąd wieje wiatr). Przy $U > 0$ jest to węzeł $(j - 1, k)$, zaś przy $U < 0$ będzie nim węzeł $(j + 1, k)$ (rys. 7.22). Schemat ten zalicza się do grupy tzw. schematów jawnych. Niewiadomą wyznacza się wprost z równania aproksymującego równanie różniczkowe, przy czym, jak już wspomniano, nie ma tutaj wymagań co do kolejności obliczeń. Do problemu jawności i niejawności powrócimy nieco dalej.

Niewątpliwą zaletą opisanego schematu jest jego prostota. Algorytm rozwiązania jest wyjątkowo nieskomplikowany i łatwy w realizacji komputerowej. Zastosujmy ten schemat do rozwiązania następującego hipotetycznego problemu:

— w kanale płynie woda ze stałą prędkością $U = 0{,}5$ m/s,
— w chwili $t = 0$ rozkład koncentracji domieszki ma kształt trójkąta równoramiennego, jak na rys. 7.21, przy czym połowa jego podstawy jest równa $x_s = 1000$ m.

Kolejne rozwiązania numeryczne, których efektem jest rozkład koncentracji po czasie $T = 7200$ s, otrzymane dla różnych, arbitralnie przyjętych wymiarów siatki węzłów Δx oraz Δt, przedstawiono na rys. 7.23. Analizując przedstawione wykresy obliczonej koncentracji, zauważamy istotne różnice zależnie od przyjętych wartości Δx oraz Δt. I tak, dla $\Delta x = 200$ m i $\Delta t = 400$ s rozwiązanie numeryczne pokrywa się z rozwiązaniem dokładnym. Jest to początkowy rozkład koncentracji przesunięty wzdłuż osi kanału o odległość $l = U \cdot T = 0{,}5 \cdot 7200 = 3600$ m. Zwiększenie wartości kroku czasowego i przyjęcie $\Delta t = 420$ s powoduje zmianę charakteru otrzymanego rozwiązania. Jest ono zupełnie niepodobne do spodziewanego rozkładu koncentracji. Występują liczne oscylacje, których obecność nie ma żadnego związku z fizycznym przebiegiem procesu. Z kolei zmniejszenie kroku całkowania w czasie Δt powoduje systematyczne rozmycie rozwiązania. Rozkład koncentracji ma

kształt zdeformowanego przez wygładzenie trójkąta. Z kolei, jak wynika z rys. 7.23b, przy stałej wartości Δt = 100 s zmniejszanie wartości Δx daje rozwiązanie coraz bardziej podobne do rozwiązania dokładnego.



Rys. 7.23. Rozwiązanie równania adwekcji po czasie T = 7200 przy: a) U = 0,5 m/s, Δx = 200 m i różnych wartości Δt, b) U = 0,5 m/s, Δt = 100 s i różnych wartościach Δx

Przedstawione wyniki obliczeń wykazują zasadniczy wpływ przyjmowanych wartości Δx oraz Δt na jakość otrzymanego rozwiązania numerycznego. Okazuje się jednak, że zależy ona nie tylko od samych wartości Δx i Δt, ale także od relacji pomiędzy nimi. Wpływ wymiarów siatki węzłów na jakość otrzymanego rozwiązania można zinterpretować. Wykonanie badania zastosowanego schematu numerycznego umożliwia nie tylko wyjaśnienie jego właściwości, ale pozwala wręcz przewidzieć jakościowy charakter spodziewanego wyniku bez konieczności wykonywania obliczeń testowych. Badanie takie wykonujemy, bazując na teorii metod numerycznego rozwiązywania równań różniczkowych cząstkowych.

Podstawową właściwością każdej metody numerycznej rozwiązania równania różniczkowego jest zbieżność. Rozwiązanie układu równań algebraicznych aproksymujących równanie różniczkowe nazywa się zbieżnym, jeśli przy wymiarach siatki zmierzających do zera zmierza ono do rozwiązania dokładnego równania w każdym punkcie obszaru rozwiązania. Wymagamy więc, aby

$$C_j^k \to C(x_j, t_k), \quad \text{gdy} \quad \Delta x, \Delta t \to 0,$$

gdzie: C_j^k – przybliżona wartość funkcji C w węźle (j, k),
$C(x_j, t_k)$ – dokładna wartość funkcji C w punkcie (x_j, t_k).

Różnicę pomiędzy rozwiązaniem dokładnym równania różniczkowego cząstkowego a dokładnym rozwiązaniem układu równań algebraicznych nazywa się błędem rozwiązania. Jest on równy

$$e_j^k = C(x_j, t_k) - C_j^k , \tag{7.102}$$

gdzie: e_j^k – błąd rozwiązania w węźle (j, k).

Wielkość tego błędu zależy od wymiarów siatki oraz od wartości członów szeregu Taylora pominiętych w procesie aproksymacji pochodnych występujących w równaniu różniczkowym.

Chociaż pojęcie zbieżności schematu jest intuicyjnie zrozumiałe i oczywiste, to wykazanie tej własności nie jest łatwe. Najprostszym sposobem wykazania zbieżności jest podejście empiryczne. Polega ono na cyklicznym rozwiązywaniu równania przy systematycznej redukcji wymiarów siatki i porównywaniu otrzymanego rozwiązania z rozwiązaniem dokładnym. Podejście to jest możliwe tylko w przypadku, gdy znamy rozwiązanie dokładne. Jest to rzadko spotykana sytuacja, na ogół bez większego znaczenia praktycznego. Ma ona jednak miejsce w przypadku rozwiązywanego wcześniej równania adwekcji. Jedna z serii eksperymentów wykonana dla różnych wartości Δx, której wyniki zamieszczono na rys. 7.23b, ilustruje ten problem. Dla malejących wartości Δx rozwiązanie numeryczne zmierza do rozwiązania dokładnego. Tendencja ta sugeruje, że zastosowany do rozwiązania równania schemat „pod prąd" jest schematem zbieżnym.

W zagadnieniach praktycznych dokładne rozwiązanie równania nie jest znane. W takiej sytuacji często można wykazać zbieżność schematu numerycznego metodą pośrednią. W tym celu wykorzystuje się twierdzenie Laxa (Fletcher, 1991), które brzmi: „*Dla danego poprawnie postawionego liniowego zagadnienia początkowego i jego różnicowej aproksymacji spełnienie warunku zgodności i stabilności jest koniecznym i wystarczającym warunkiem zbieżności*". Zatem jeśli dla danego schematu wykażemy jego zgodność i stabilność, to tym samym wykażemy jego zbieżność, gdyż zgodnie z twierdzeniem Laxa

zbieżność = zgodność + stabilność

Wadą tego podejścia jest jego ograniczenie do zagadnień liniowych. Pomimo to, jest ono bardzo użyteczne, gdyż zwykle można wnioskować o zbieżności zagadnień nieliniowych na podstawie wyników analizy ich zlinearyzowanych wariantów. Wykonanie zaś analizy zgodności oraz stabilności jest względnie łatwe. Omówmy je kolejno.

Układ równań różnicowych nazywa się zgodnym z równaniem różniczkowym, jeśli w granicy, przy wymiarach siatki zmierzających do zera, układ ten w każdym węźle siatki zmierza do równania różniczkowego, które aproksymuje. Badanie zgodności jest procesem odwrotnym do dyskretyzacji. Polega ono na zastąpieniu w równaniu aproksymującym wartości węzłowych funkcji ich rozwinięciami w szereg Taylora wokół badanego węzła. Zgodność wymaga, aby otrzymane tą drogą równanie zawierało rozwiązywane równanie różniczkowe cząstkowe oraz dodatkowe człony, które powinny znikać w trakcie redukcji wymiarów siatki. W celu zilustrowania problemu zgodności rozpatrzmy schemat „pod prąd".

Schemat „pod prąd" pozwolił zastąpić równanie adwekcji równaniem algebraicznym (7.98). Występują w nim następujące wartości węzłowe funkcji C: C_j^k, C_j^{k+1}, C_{j-1}^k. Wyznaczmy C_j^{k+1} oraz C_{j-1}^k, rozwijając funkcję C w szereg Taylora wokół węzła (j, k) (rys. 7.22).

$$C_j^{k+1} = C_j^k + \Delta t \left.\frac{\partial C}{\partial t}\right|_j^k + \frac{\Delta t^2}{2} \left.\frac{\partial^2 C}{\partial t^2}\right|_j^k + \ldots, \tag{7.103}$$

$$C_{j-1}^k = C_j^k - \Delta x \left.\frac{\partial C}{\partial x}\right|_j^k + \frac{\Delta x^2}{2} \left.\frac{\partial^2 C}{\partial x^2}\right|_j^k + \ldots. \tag{7.104}$$

Po podstawieniu powyższych zależności do równania (7.98), otrzymujemy

$$\frac{C_j^k + \Delta t \left.\frac{\partial C}{\partial t}\right|_j^k + \frac{\Delta t^2}{2} \left.\frac{\partial^2 C}{\partial t^2}\right|_j^k + \ldots - C_k^j}{\Delta t} + U \frac{C_j^k - \left(C_j^k - \Delta x \left.\frac{\partial C}{\partial x}\right|_j^k + \frac{\Delta x^2}{2} \left.\frac{\partial^2 C}{\partial x^2}\right|_j^k + \ldots \right)}{\Delta x} = 0. \tag{7.105}$$

Po uproszczeniu i uporządkowaniu otrzymujemy następujące równanie ważne w węźle (j, k):

$$\frac{\partial C}{\partial t} + U \frac{\partial C}{\partial x} = \frac{U \Delta x}{2} \frac{\partial^2 C}{\partial x^2} - \frac{\Delta t}{2} \frac{\partial^2 C}{\partial t^2} + \cdots. \tag{7.106}$$

Zgodnie z oczekiwaniem zawiera ono wyjściowe równanie adwekcji (7.93) oraz dodatkowe człony. Człony te znikają, gdy $\Delta x \to 0$ i $\Delta t \to 0$. Zatem w granicy równanie (7.106) zmierza w węźle (j, k) do rozwiązywanego równania adwekcji (7.93). Wnioskujemy stąd, że zastosowany schemat „pod prąd" jest schematem zgodnym.

Drugim warunkiem, którego spełnienia wymaga zbieżność, jest warunek stabilności numerycznej. Przez stabilność rozumie się zdolność schematu do wytłumienia dowolnych przypadkowych zakłóceń rozwiązania (np. wynikających z błędów zaokrągleń) w trakcie obliczeń. W przypadku rozwiązywanego wcześniej równania adwekcji wyniki przedstawione

na rys. 7.23a zawierają symptomy niestabilności. Są to oscylacje o niefizycznym charakterze. Z kolei przykładem stabilnych rozwiązań są wyniki przedstawione na rys. 7.23b.

Stosowane schematy numeryczne mogą być absolutnie stabilne, warunkowo stabilne lub absolutnie niestabilne. Schemat absolutnie stabilny to taki schemat, który nie powoduje nigdy nieograniczonego wzrostu błędów powstających w trakcie obliczeń. Ta cecha jest oczywiście pożądana i stanowi istotną zaletę schematu. Schemat warunkowo stabilny to taki schemat, który zapobiega lawinowemu narastaniu błędów tylko wtedy, gdy spełniony jest pewien warunek. Warunek ten określa dopuszczalną relację pomiędzy wymiarami siatki i nazywany jest warunkiem stabilności lub kryterium stabilności. Jeśli warunek stabilności nie jest spełniony, schemat jest niestabilny. Z kolei schemat absolutnie niestabilny to taki, który nigdy nie zabezpiecza przed nieograniczonym wzrostem błędu. Z oczywistych powodów schemat o takich cechach jest nieinteresujący.

Dokonajmy analizy stabilności numerycznej schematu „pod prąd" zastosowanego do rozwiązania równania adwekcji, w efekcie czego otrzymano układ równań algebraicznych (7.101)

$$C_j^{k+1} = (1 - C_a) C_j^k + C_a C_{j-1}^k \quad (j = 2, 3, \ldots, M) . \tag{7.107}$$

Przypomnijmy, że wartość C_1^{k+1} określa zadany warunek brzegowy. Rozwiązanie powyższego układu równań da w każdym węźle przybliżoną wartość funkcji C_j^{k+1}, zamiast dokładnej $C(x_j, t_{k+1})$. Błąd tego rozwiązania definiuje zależność (7.102). Podstawmy tę zależność do równań (7.107). Otrzymamy

$$C(x_j, t_{k+1}) - e_j^{k+1} = (1 - C_a)(C(x_j, t_k) - e_j^k) + C_a(C(x_{j-1} t_k) - e_{j-1}^k) \tag{7.108}$$
$$(j = 2, 3, ..., M).$$

Ponieważ rozwiązanie dokładne $C(x, t)$ musi spełniać równanie schematu (7.107), to powyższy układ uprości się do postaci

$$e_j^{k+1} = (1 - C_a) e_j^k + C_a e_{j-1}^k \quad (j = 2, 3, \ldots, M) . \tag{7.109}$$

Zauważmy, że błąd rozwiązania e spełnia jednorodny układ równań wynikający z (7.107). Ma on identyczną strukturę jak (7.107). Do układu tego dodaje się warunki początkowe i brzegowe. Ponieważ przyjmuje się, że są one dokładne, to błąd wynikający z ich przyjęcia jest równy zeru. Zatem $e_j^0 = 0$ dla $j = 2, 3, ..., M$ oraz $e_1^k = 0$ dla $k = 0, 1, 2, ...$.

Jedną z najczęściej stosowanych metod badania stabilności jest metoda Neumanna (Fletcher, 1991; Potter, 1977; Szymkiewicz, 2000). W metodzie tej błędy na danym poziomie czasowym w węzłach wzdłuż osi x rozwija się w skończony szereg zespolony Fouriera. Określenie stabilności bądź niestabilności polega na zbadaniu zachowania pojedynczej składowej szeregu. Jeśli przy przejściu z poziomu czasowego k na $k + 1$ amplituda fali Fouriera jest tłumiona, to schemat jest stabilny. Jeśli jednak następuje wzmocnienie amplitudy – schemat jest niestabilny.

W punkcie x_j błąd rozwiązania równania wyrażamy następująco:

$$e_j^k = \sum_{n=1}^{N} A_n^k e^{i \cdot n \cdot m \cdot j \cdot \Delta x} , \tag{7.110}$$

gdzie: $j = 2, 3, ..., M$,
$i = (-1)^{1/2}$ – jednostka urojona,
n – indeks składowej szeregu – fali Fouriera,

m – liczba falowa zdefiniowana wzorem

$$m = \frac{2\pi}{\lambda}, \tag{7.111}$$

λ – długość fali,
Δx – wymiar przestrzenny siatki,
A^k_n – współczynnik Fouriera – amplituda fali o indeksie n na poziomie czasu k,
N – indeks o skończonej wartości.

W przypadku równania liniowego, jak rozwiązywane tutaj równanie adwekcji, można ograniczyć się do analizy pojedynczej składowej szeregu (7.110). Zatem dla $n = 1$ będzie

$$e^k_j = A^k e^{i\varphi j}. \tag{7.112}$$

Symbol φ oznacza kąt zdefiniowany następująco:

$$\varphi = m\Delta x = \frac{2\pi}{K}, \tag{7.113}$$

gdzie $K = \lambda/\Delta x$ jest ilością interwałów Δx mieszczących się na fali o długości λ.
Wstawiając wyrażenie (7.112) do równania (7.109) opisującego błąd, otrzymujemy

$$A^{k+1}e^{i\cdot\varphi\cdot j} = (1 - C_a)A^k e^{i\cdot\varphi\cdot j} + C_a A^k e^{i\cdot\varphi(j-1)}. \tag{7.114}$$

Równanie powyższe dzielimy obustronnie przez $e^{i\varphi j}$, a następnie przez A^k. W efekcie uzyskuje się zależność

$$\frac{A^{k+1}}{A^k} = 1 - C_a + C_a \cdot e^{-i\varphi}. \tag{7.115}$$

Wprowadźmy oznaczenie

$$G = \frac{A^{k+1}}{A^k}. \tag{7.116}$$

Zmienną G interpretuje się jako tzw. współczynnik wzmocnienia amplitudy n-tej składowej Fouriera funkcji błędu przy przejściu z poziomu czasowego k na $k + 1$. Z równania tego wynika zależność

$$A^{k+1} = G \cdot A^k. \tag{7.117}$$

Błąd rozwiązania nie będzie miał tendencji do wzrostu, jeśli bezwzględna wartość współczynnika wzmocnienia nie będzie większa od 1 dla wszystkich składowych Fouriera. Warunek stabilności przyjmie więc postać

$$|G| \le 1. \tag{7.118}$$

Wykorzystując uogólniony wzór Eulera (Bronsztejn i Siemindiajew, 1990) o postaci

$$e^{-i\cdot\varphi} = \cos\varphi - i\sin\varphi, \tag{7.119}$$

równanie (7.115) zapiszemy następująco:

$$G = 1 - C_a + C_a(\cos\varphi - i\sin\varphi). \tag{7.120}$$

Wykorzystując znane zależności trygonometryczne, powyższe wyrażenie można przekształcić do postaci:

$$G = 1 - 2C_a \cdot \sin^2 \frac{\varphi}{2} - i \cdot 2 \cdot C_a \cdot \sin \frac{\varphi}{2} \cos \frac{\varphi}{2}. \tag{7.121}$$

Jak widać, współczynnik wzmocnienia jest liczbą zespoloną. Jego moduł jest więc równy

$$|G| = \left(1 - 4C_a \cdot \sin^2 \frac{\varphi}{2} + 4 \cdot C_a^2 \cdot \sin^2 \frac{\varphi}{2} \right)^{1/2}. \tag{7.122}$$

Warunek stabilności (7.118) przyjmie postać

$$\left(1 - 4C_a \cdot \sin^2 \frac{\varphi}{2} + 4 \cdot C_a^2 \cdot \sin^2 \frac{\varphi}{2} \right)^{1/2} \le 1. \tag{7.123}$$

Ponieważ funkcja $\sin^2 \varphi/2$ przyjmuje wartości z zakresu od 0 do 1, relację powyższą wystarczy zbadać dla jej skrajnych wartości.

Dla $\sin^2 \varphi/2 = 0$ nierówność (7.118) spełniona jest zawsze. Natomiast dla $\sin^2 \varphi/2 = 1$ otrzymujemy relację

$$1 - 4C_a + 4C_a^2 \le 1, \tag{7.124}$$

która będzie spełniona, gdy

$$C_a \le 1. \tag{7.125}$$

Powyższy warunek jest tzw. kryterium stabilności dla numerycznego rozwiązania równania adwekcji schematem „pod prąd". Schemat zapewni stabilne rozwiązanie tylko wtedy, gdy adwekcyjna liczba Couranta nie będzie większa od jedności. Wprowadzając jej definicję (7.100), otrzymamy relację

$$\frac{U \cdot \Delta t}{\Delta x} \le 1, \tag{7.126}$$

z której wynika warunek

$$\Delta t \le \frac{\Delta x}{U}, \tag{7.127}$$

narzucający ograniczenie na wartość maksymalną czasowego kroku całkowania przy przyjętym kroku przestrzennym Δx. Powyższe kryterium stabilności wyjaśnia oscylujące rozwiązanie równania adwekcji, otrzymane przy Δt = 420 s i przedstawione na rys. 7.23a.

Jak wcześniej wykazano, zastosowany do rozwiązania równania adwekcji schemat „pod prąd" jest schematem zgodnym z rozwiązywanym równaniem, a także jest schematem stabilnym, gdy spełniony jest warunek (7.125). Zatem, zgodnie z twierdzeniem Laxa, przy spełnieniu tego warunku jest on schematem zbieżnym. Spełnia on więc podstawowe wymagania stawiane metodom numerycznym rozwiązywania równań różniczkowych cząstkowych.

Zarówno analiza stabilności, jak i zgodności pozwala interpretować niektóre efekty obserwowane w rozwiązaniu numerycznym. Niestety, nie wyjaśniają one wszystkich właściwości metody numerycznej. Na przykład, nie dają one żadnych informacji o przyczynach niefizycznego wygładzania wyników oraz niefizycznych oscylacji, często obecnych w rozwiązaniu nawet wtedy, gdy kryterium stabilności numerycznej jest spełnione. Okazuje się, że wymienione efekty numeryczne mają genezę, którą można objaśnić i zinterpretować. W tym celu należy wykonać analizę dokładności rozwiązania. Jednak powinna być ona wykonana w sposób szczególny, mianowicie metodą tzw. równania zmodyfikowanego.

Jak wiadomo, metody numeryczne różnią się między sobą sposobem przekształcenia równania różniczkowego w układ równań algebraicznych. Podstawę procesu tego przekształcenia stanowi rozwinięcie funkcji w szereg Taylora. Stosowne formuły aproksymujące pochodne otrzymuje się pomijając wyrazy szeregu, co wprowadza tzw. błąd obcięcia szeregu. Determinuje on własności numeryczne schematu. Zwykle z błędem obcięcia wiąże się bezpośrednio pojęcie rzędu dokładności formuły aproksymującej odpowiednią pochodną. Rząd ten określa najwyższa pochodna w obciętym szeregu Taylora, zachowana w formule aproksymacyjnej. Niestety, rząd dokładności metody numerycznej jest tylko ogólną informacją, niemówiącą nic o wielkości błędu.

Jak już wspomniano, rozwiązania równań typu hiperbolicznego opisujących wiele zagadnień z zakresu hydrodynamiki mają postać fal propagujących bez zmiany amplitudy lub z jej niewielką zmianą. Jest sprawą zasadniczej wagi, aby metody numeryczne stosowane do rozwiązywania takich równań nie wprowadzały sztucznej dyssypacji. Równie ważną sprawą jest, aby metody numeryczne nie zakłócały prędkości propagacji fal, to znaczy, aby nie generowały sztucznej dyspersji. Obecność sztucznej dyssypacji obserwuje się w postaci nadmiernego wygładzenia rozwiązania. Natomiast obecność sztucznej dyspersji zauważa się w postaci jego oscylacji o niefizycznym charakterze.

Dla funkcji gładkich wartość błędu obcięcia szeregu Taylora zdeterminowana jest przez wartość pierwszego wyrazu obciętej części szeregu. Można wykazać, że rodzaj dominującego błędu w rozwiązaniu w takiej sytuacji zależy od parzystości bądź nieparzystości rzędu pochodnej w pierwszym wyrazie obciętej części szeregu Taylora. W tym celu rozpatrzymy problem propagacji fali płaskiej ulegającej jednocześnie dyssypacji i dyspersji. Opisuje ją równanie (Fletcher, 1991; Tan Weiyan, 1992)

$$C(x, t) = Ae^{-p(m)\cdot t}\, e^{-im(x-q(m)\cdot t)}, \tag{7.128}$$

gdzie: A – amplituda fali,
x – położenie,
t – czas,
$i = (-1)^{1/2}$ – jednostka urojona,
m – liczba falowa związana z długością fali λ zależnością (7.111),
$p(m)$ – parametr determinujący tempo tłumienia amplitudy w czasie,
$q(m)$ – prędkość propagacji fali.

Rozpatrzmy najprostszy przypadek równania hiperbolicznego, kiedy falę opisuje liniowe równanie adwekcji (7.93). Podstawiając (7.128) do równania (7.93), otrzymuje się:

$$p(m) = 0 \text{ oraz } q(m) = U, \tag{7.129a,b}$$

co oznacza propagację fali (7.128) ze stałą prędkością bez tłumienia jej amplitudy.

Rozpatrzmy teraz liniowe równanie adwekcji-dyfuzji (7.16)

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} - D\frac{\partial^2 C}{\partial x^2} = 0\,. \tag{7.130}$$

Podstawienie (7.128) do (7.130) daje zależności:

$$p(m) = Dm^2 \text{ oraz } q(m) = U\,. \tag{7.131a,b}$$

Otrzymany wynik wskazuje, że ruch fali opisanej równaniem adwekcji-dyfuzji przebiega ze stałą prędkością U i jednoczesnym tłumieniem jej amplitudy. Tempo tłumienia zależy od długości fali i jest intensywniejsze dla fal krótkich niż długich, ponieważ $m = 2\pi/\lambda$.

Z kolei dla fali, której ruch opisuje równanie Kortwega-de Vriesa

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} + \varepsilon\frac{\partial^3 C}{\partial x^3} = 0\,, \tag{7.132}$$

po wykonaniu podobnych operacji otrzymuje się:

$$p(m) = 0 \;\text{ oraz }\; q(m) = U - \varepsilon m^2\,. \tag{7.133a,b}$$

W przypadku tego równania amplituda fali nie zmienia się, lecz fala propaguje z prędkością zależną od jej długości. Jeśli występują fale o różnych długościach, każda z nich propaguje z różną prędkością, co oznacza ich dyspersję. W tym przypadku stopień zakłócenia prędkości zależy również od długości fali i jest większy dla fal krótkich (duża wartość m). Przy $\varepsilon > 0$ fale krótkie przemieszczają się wolniej, zaś przy $\varepsilon < 0$ szybciej niż wynika to z prędkości adwekcji.

Podsumowując, można przyjąć, że zmianę amplitudy propagującej fali wywołują człony zawierające pochodne parzystych rzędów, natomiast zmianę prędkości propagacji – człony zawierające pochodne nieparzystych rzędów.

Powyższy wniosek sugeruje, że dyssypatywne i dyspersyjne własności metod numerycznych można badać, analizując błąd obcięcia interpretowany w szczególny sposób zaproponowany przez Warminga i Hyetta (1974). Mianowicie, jeśli $\Omega(C) = 0$ reprezentuje równanie algebraiczne aproksymujące równanie różniczkowe, to można napisać, że

$$\Omega_a(C) = \Omega(C) + E(C) = 0, \tag{7.134}$$

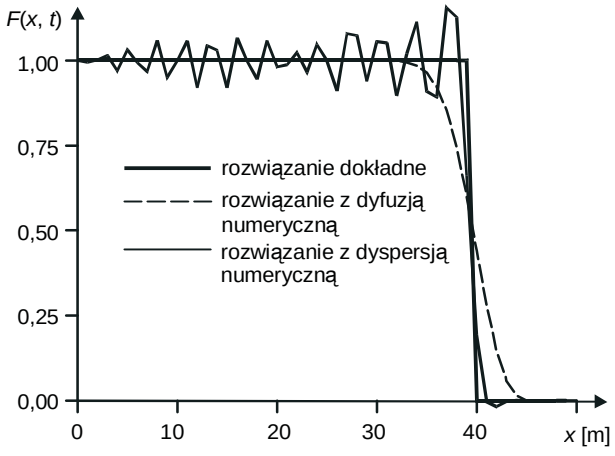
gdzie $E(C)$ reprezentuje błąd obcięcia. Równanie (7.134) otrzymujemy, wyrażając wartości węzłowe równania algebraicznego za pomocą rozwinięcia w szereg Taylora i tak przekształcając je, aby błąd $E(C)$ był funkcją jedynie pochodnych przestrzennych. $\Omega_a(C)$ jest równaniem różniczkowym z nieskończoną liczbą członów. Jest ono nazywane równaniem zmodyfikowanym. Równanie to można wykorzystać do badania zgodności metody numerycznej oraz rzędu jej dokładności. Dla danego równania różniczkowego $\Omega(C)$ postać równania zmodyfikowanego zależy od metody numerycznej zastosowanej do rozwiązania, a jej własności numeryczne można oszacować, analizując $E(C)$. Ogólnie można powiedzieć, że $E(C)$ zawiera wszystkie człony szeregu Taylora pominięte w procesie aproksymacji pochodnych. W równaniu zmodyfikowanym, podobnie jak w omówionych równaniach (7.130) i (7.132), człony z pochodnymi rzędu parzystego związane są z dyssypacją, zaś z pochodnymi rzędu nieparzystego – z dyspersją. Ponieważ dla funkcji gładkich o wartości błędu obcięcia decyduje pierwszy wyraz obciętej części szeregu Taylora, zatem pierwszy wyraz $E(C)$ będzie decydował o rodzaju błędu dominującego w rozwiązaniu. Jeśli będzie on zawierał pochodną nieparzystą, schemat będzie dyspersyjny. Natomiast jeśli będzie to wyraz z pochodną parzystą, schemat będzie dyssypatywny, przy czym z najniższym rzędem parzystym wiąże się proces dyfuzji numerycznej. Zatem zależnie od dominującego członu w błędzie obcięcia będziemy otrzymywali albo wygładzenie rozwiązania, albo jego oscylacje. Efekty te mają naturę numeryczną i nie wynikają z charakteru równań. Przykład obecności w rozwiązaniu równania adwekcji opisanych wyżej efektów numerycznych przedstawiono na rys. 7.24. Rozwiązanie dokładne zawiera nieciągłość. Natomiast rozwiązania numeryczne – zależnie od typu błędu dominującego – są albo sztucznie wygładzane, albo zawierają niefizyczne oscylacje. O ile stosunkowo łatwo można zidentyfikować przyczynę niefizycznych oscylacji rozwiązania, tzn. dyspersyjność schematu, o tyle interpretacja wy-

ników obliczeń w przypadku schematu dyssypatywnego może nastręczać pewnych trudności. Ich przyczyną jest identyczny efekt dyssypacji fizycznej i numerycznej.

W przypadku występowania nieciągłości funkcji $C(x, t)$, błąd obcięcia nie musi być zdominowany przez pierwszy wyraz obciętej części szeregu Taylora, a przez następny, zawierający pochodną wyższego rzędu. W takim wypadku niefizyczne oscylacje rozwiązania wystąpią nawet wtedy, gdy zastosowany schemat numeryczny jest umiarkowanie dyssypatywny.

Ze względu na swoją genezę dyfuzja numeryczna jest zwykle postrzegana jako uciążliwy problem sztucznego, niewynikającego z równania wygładzania rozwiązania. Taka sytuacja występuje w przypadku najprostszego rozwiązania równania typu hiperbolicznego, jakim jest jednowymiarowe równanie adwekcji (7.93). Równanie to, aproksymowane schematem „pod prąd", staje się równaniem algebraicznym (7.98). Badając zgodność tego równania, wykonano proces odwrotny do dyskretyzacji. W efekcie otrzymano równanie (7.106). Zauważmy, że równanie to ma lewą stronę identyczną z równaniem adwekcji (7.93). Natomiast prawa strona, w przeciwieństwie do (7.93), nie jest równa zero. Przekształćmy ją w taki sposób, aby wyeliminowane zostały pochodne po czasie wyższych rzędów i aby pozostały jedynie pochodne względem zmiennej x. W tym celu równanie adwekcji różniczkuje się kolejno, najpierw względem czasu t, a następnie względem x. Otrzymujemy

$$\frac{\partial^2 C}{\partial t^2} + U\frac{\partial^2 C}{\partial x \partial t} = 0, \quad \frac{\partial^2 C}{\partial t \partial x} + U\frac{\partial^2 C}{\partial x^2} = 0\,. \tag{7.135a,b}$$



Rys. 7.24. Przykład rozwiązania równania adwekcji w przypadku występowania nieciągłości (Szymkiewicz, 2001)

Eliminacja pochodnej mieszanej prowadzi do zależności

$$\frac{\partial^2 C}{\partial t^2} = U^2\frac{\partial^2 C}{\partial x^2}\,. \tag{7.136a,b}$$

Po jej wstawieniu do równania (7.106), otrzymujemy

$$\frac{\partial C}{\partial t}+U\frac{\partial C}{\partial x}=\frac{U\Delta x-\Delta t U^2}{2}\frac{\partial^2 C}{\partial x^2}+ \\ +\frac{U\Delta x^2}{6}\left(-1+3\frac{U\Delta t}{\Delta x}-2\frac{U^2\Delta t^2}{\Delta x^2}\right)\frac{\partial^3 C}{\partial x^3}+\cdots. \tag{7.137}$$

Wprowadźmy następujące oznaczenie wyrażeń przy pochodnej II i III rzędu:

$$D_n=\frac{U\Delta x}{2}\left(1-\frac{U\Delta t}{\Delta x}\right), \tag{7.138}$$

$$E_n=\frac{U\Delta x^2}{6}\left(-1+3\frac{U\Delta t}{\Delta x}-2\frac{U^2\Delta t^2}{\Delta x^2}\right), \tag{7.139}$$

które po uwzględnieniu definicji adwekcyjnej liczby Couranta (7.100) można zapisać w postaci:

$$D_n=\frac{U\Delta x}{2}\left(1-C_a\right), \tag{7.140}$$

$$E_n=\frac{U\Delta x^2}{6}\left(-1+3C_a-2C_a^2\right). \tag{7.141}$$

Równanie (7.137) można ostatecznie zapisać następująco:

$$\frac{\partial C}{\partial t}+U\frac{\partial C}{\partial x}=D_n\frac{\partial^2 C}{\partial x^2}+E_n\frac{\partial^3 C}{\partial x^3}+\cdots. \tag{7.142}$$

Otrzymane w ten sposób równanie nazywamy równaniem zmodyfikowanym. W tym przypadku jest to zmodyfikowane przy użyciu schematu „pod prąd" równanie adwekcji (7.93). Występujące w nim współczynniki przy pochodnych wyższych rzędów nazywa się odpowiednio: D_n – współczynnikiem dyfuzji numerycznej; E_n – współczynnikiem dyspersji numerycznej.

Dyssypatywna metoda numeryczna zastosowana do rozwiązania równania adwekcji (7.93) modyfikuje je do postaci:

$$\frac{\partial C}{\partial t}+U\frac{\partial C}{\partial x}=D_n\frac{\partial^2 C}{\partial x^2}. \tag{7.143}$$

W efekcie, nawet przy U = const, zamiast czystej translacji wzdłuż osi x otrzymuje się rozwiązanie typowe dla równania adwekcji-dyfuzji, w którym zadany rozkład na przykład koncentracji nie tylko przemieszcza się z prędkością adwekcji ale równocześnie podlegając dyfuzji ulega spłaszczeniu. Problem ten jest niezależny od wymiarowości zagadnienia.

Dla zagadnień dwu- i trójwymiarowych zamiast współczynnika dyfuzji pojawia się tensor dyfuzji numerycznej. Problem ten dobrze ilustruje rozwiązanie dwuwymiarowego równania adwekcji przy użyciu schematu „pod prąd". Równaniem tego typu jest równanie (7.14) przy D = 0 w przypadku dwuwymiarowym. Ma ono następującą postać:

$$\frac{\partial C}{\partial t}+u\frac{\partial C}{\partial x}+v\frac{\partial C}{\partial y}=0\,. \tag{7.144}$$

Rozwiązujemy je na siatce prostokątnej o wymiarach $\Delta x \cdot \Delta y$ (rys. 7.25), przy założeniu, że u, $v > 0$. Aproksymację pochodnych przestrzennych w węźle (i, j) z zastosowaniem iloraz ów różnicowych wstecznych wykonujemy na poziomie czasu t, co prowadzi do równania algebraicznego o postaci:

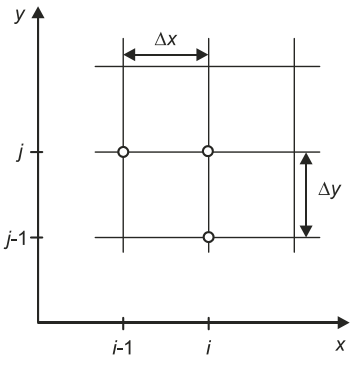
$$\frac{C_{i,j}^{k+1}-C_{i,j}^{k}}{\Delta t}+u\frac{C_{i,j}^{k}-C_{i-1,j}^{k}}{\Delta x}+v\frac{C_{i,j}^{k}-C_{i,j-1}^{k}}{\Delta y}=0\,, \tag{7.145}$$

gdzie: k – indeks poziomu czasowego,
Δt – krok czasowy.

Poszukiwaną wartość funkcji C na poziomie czasowym $k + 1$ oblicza się następująco:

$$C_{i,j}^{k+1}=C_{i,j}^{k}-C_x\left(C_{i,j}^{k}-C_{i-1,j}^{k}\right)+C_y\left(C_{i,j}^{k}-C_{i,j-1}^{k}\right), \tag{7.146}$$

gdzie:



Rys. 7.25. Siatka węzłów stosowana w schemacie „pod prąd” dla dwuwymiarowego równania adwekcji

$$C_x=u\Delta t/\Delta x,\quad C_y=v\Delta t/\Delta y \tag{7.147a,b}$$

są adwekcyjnymi liczbami Couranta odpowiednio w kierunku x oraz y.
Celem wyznaczenia równania zmodyfikowanego w równaniu (7.145) wartości węzłowe wyraża się za pomocą rozwinięcia funkcji C w szereg Taylora w węźle (i, j) na poziomie czasu t. W szeregu uwzględnia się wyrazy z pochodnymi do drugiej włącznie. Wykorzystując dodatkowo następującą zależność wynikającą z różniczkowania równania (7.144):

$$\frac{\partial^2 C}{\partial t^2}=u^2\frac{\partial^2 C}{\partial x^2}+v^2\frac{\partial^2 C}{\partial y^2}+2uv\frac{\partial^2 C}{\partial x\partial y}, \tag{7.148}$$

otrzymuje się w punkcie aproksymacji równanie:

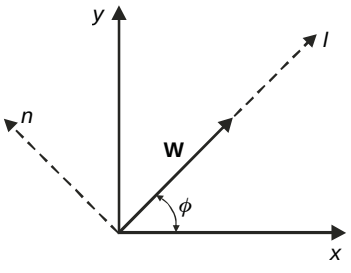
$$\begin{aligned}&\frac{\partial C}{\partial t}+u\frac{\partial C}{\partial x}+v\frac{\partial C}{\partial y}=\\&=\left[\frac{1}{2}\Delta t u^2+\frac{u\Delta x}{2}\right]\frac{\partial^2 C}{\partial x^2}+\left[\frac{1}{2}\Delta t\, v^2+\frac{v\Delta y}{2}\right]\frac{\partial^2 C}{\partial y^2}-\Delta t\, u\, v\frac{\partial^2 C}{\partial x\partial y}+\cdots.\end{aligned} \tag{7.149}$$

Powyższe równanie można zapisać w postaci

$$\frac{\partial C}{\partial t}+u\frac{\partial C}{\partial x}+v\frac{\partial C}{\partial y}=\operatorname{div}(\mathbf{D}^n\nabla C)+\cdots, \tag{7.150}$$

gdzie:

$$\mathbf{D}^n=\begin{bmatrix}\frac{1}{2}\Delta t u^2+\frac{u\Delta x}{2} & \frac{1}{2}\Delta t uv\\ \frac{1}{2}\Delta t vu & \frac{1}{2}\Delta t v^2+\frac{v\Delta y}{2}\end{bmatrix}. \tag{7.151}$$



Rys. 7.26. Obrót osi układu współrzędnych $x - y$ o kąt ϕ

Jest to zmodyfikowane wskutek aproksymacji równanie adwekcji (7.144). Po jego prawej stronie pojawiły się dodatkowe człony. Ich postać zależy od dokładności zastosowanej aproksymacji. Gdyby zastosowana aproksymacja względem położenia i czasu była II rzędu, w (7.150) nie wystąpi pierwszy człon po prawej stronie. Dla aproksymacji III rzędu nie wystąpi również drugi człon itd. Analiza dodatkowych członów występujących w równaniu zmodyfikowanym dostarcza informacji o własnościach metody.

Zmodyfikowane równanie adwekcji (7.150) jest podobne do równania adwekcji-dyfuzji (7.14). W obu równaniach występuje człon dyfuzyjny. W równaniu (7.14) człon ten opisuje fizyczny proces transportu w przypadku ogólnym zdefiniowany tensorem dyfuzji fizycznej **D**. Natomiast w równaniu (7.150) człon dyfuzji reprezentuje efekt błędu numerycznego wynikający z aproksymacji równania różniczkowego. Przez analogię do (7.14) tensor $\mathbf{D}^n$ występujący w tym członie nazywa się tensorem dyfuzji numerycznej.

Dokonajmy transformacji układu współrzędnych $x - y$ przez obrót o kąt ϕ, tak aby jego oś l była równoległa, a oś n prostopadła do wektora prędkości $\mathbf{w} = u\mathbf{i} + v\mathbf{j}$ (rys. 7.26). Przy obrocie układu dowolny tensor **R** transformuje się do nowego układu wg reguły (Sawicki, 1994):

$$\overline{\mathbf{R}} = \mathbf{Q}\,\mathbf{R}\,\mathbf{Q}^T, \tag{7.152}$$

gdzie: $\overline{\mathbf{R}}$ – tensor w układzie $l - n$,
$\mathbf{R}$ – tensor w układzie $x - y$,
T – symbol transpozycji,

$$\mathbf{Q} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \text{ – macierz transformacji.} \tag{7.153}$$

Po wykonaniu mnożenia (7.152) otrzymuje się tensor $\overline{\mathbf{R}}$ o elementach równych:

$$\overline{R}_{11} = R_{11}\cos^2\phi + R_{22}\sin^2\phi + R_{12}\sin 2\phi, \tag{7.154a}$$

$$\overline{R}_{22} = R_{11}\sin^2\phi + R_{22}\cos^2\phi - R_{12}\sin 2\phi, \tag{7.154b}$$

$$\overline{R}_{12} = \overline{R}_{21} = -\frac{1}{2}\left(R_{11} - R_{22}\right)\sin 2\phi + R_{12}\cos 2\phi. \tag{7.154c}$$

Jeśli przyjmiemy $\Delta x = \Delta y = \Delta$, tensor dyfuzji numerycznej $\mathbf{D}^n$ (7.151) w układzie współrzędnych $l - n$ zgodnie z powyższymi wzorami będzie miał następujące elementy:

$$\overline{D}_{11}^n = -\frac{1}{2}\Delta t w^2 \sin^2 2\phi + \frac{\Delta}{4} w \sin 2\phi\left(\sin\phi + \cos\phi\right), \tag{7.155a}$$

$$\overline{D}_{22}^n = \frac{\Delta}{4} w \sin 2\phi\left(\sin\phi + \cos\phi\right), \tag{7.155b}$$

$$\overline{D}_{12}^n = \overline{D}_{21}^n = \frac{\Delta}{4} w \sin 2\phi\left(\sin\phi - \cos\phi\right). \tag{7.155c}$$

Z powyższych zależności wynika, że schemat z różnicami wstecznymi generuje maksymalną dyfuzję numeryczną, gdy przepływ skierowany jest pod kątem 45° do osi układu $x - y$. Dla $\phi = \pi/4$ tensor ten ma następującą postać:

$$\overline{\mathbf{D}}^n = \begin{bmatrix} -\frac{1}{2}\Delta t w^2 + \frac{\Delta}{2\sqrt{2}} w & 0 \\ 0 & \frac{\Delta}{2\sqrt{2}} w \end{bmatrix}. \tag{7.156}$$

Jak można zauważyć, schemat zawsze generuje dyfuzję numeryczną w kierunku normalnym do wektora prędkości, gdyż $\overline{\mathrm{D}}^n_{22} \neq 0$. Natomiast dyfuzja numeryczna w kierunku przepływu zależy od relacji pomiędzy wymiarami siatki Δ oraz krokiem czasowym. Zauważmy, że dyfuzja w kierunku stycznym będzie zawsze mniejsza od dyfuzji w kierunku normalnym, gdyż $\overline{\mathrm{D}}^n_{11} < \overline{\mathrm{D}}^n_{22}$. Jednak warunek poprawnego sformułowania problemu początkowo-brzegowego dla równania adwekcji-dyfuzji, jakim jest równanie zmodyfikowane (7.150), wymaga, aby elementy tensora dyfuzji numerycznej były nieujemne. Zatem musi być spełniona relacja

$$-\frac{1}{2}\Delta t w^2 + \frac{\Delta}{2\sqrt{2}} w \geq 0 . \tag{7.157}$$

Wynika z niej warunek

$$\Delta t \leq \frac{\sqrt{2}}{2}\frac{\Delta}{w}, \tag{7.158}$$

który zapewnia istnienie rozwiązania numerycznego. W skrajnym przypadku, gdy

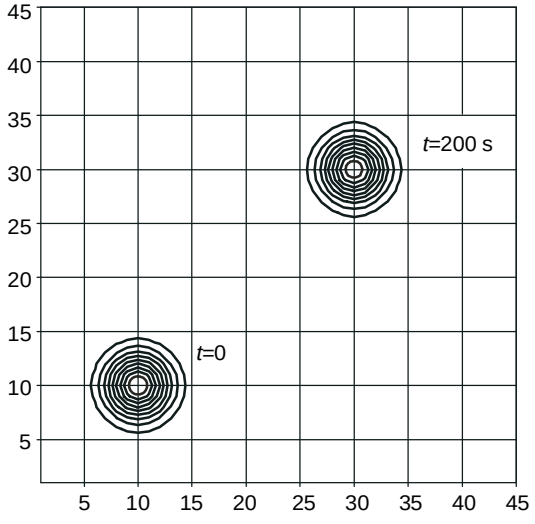
$$\Delta t = \frac{\sqrt{2}}{2}\frac{\Delta}{w}, \tag{7.159}$$

dyfuzja w kierunku stycznym nie wystąpi. Pozostanie tylko dyfuzja w kierunku normalnym. W przypadku przepływu jednowymiarowego przy liczbie Couranta równej jedności otrzymuje się rozwiązanie dokładne. Tensor dyfuzji numerycznej (7.151) będzie miał zerowe elementy. Rozwiązanie numeryczne będzie zgodne z rozwiązaniem analitycznym, ponieważ schemat jest dokładną aproksymacją równania adwekcji.

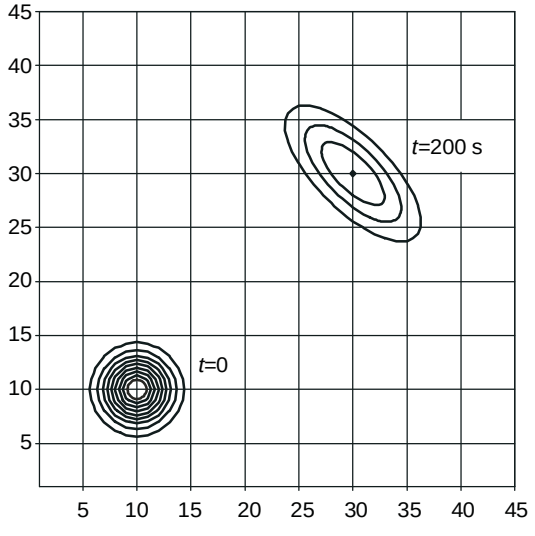
Przedstawione własności schematu potwierdzają wyniki testów numerycznych przedstawione na rysunkach 7.27 i 7.28. W zbiorniku o wymiarach 45×45 m, podzielonym na oczka o wymiarach $\Delta x = \Delta y = 1$ m, panuje przepływ ustalony z jednorodnym polem prędkości. Warunek początkowy $C(t = 0, x, y)$ określa rozkład Gaussa o parametrach $\sigma = 2$ m, $x_s = y_s = 9$ m i wartości maksymalnej $C_{\max} = 1$. Dla przepływu wzdłuż przekątnej zbiornika, gdy $u = 0{,}1$ m/s i $v = 0{,}1$ m/s, rozwiązaniem dokładnym jest przemieszczenie początkowego rozkładu bez deformacji (rys. 7.27). Po czasie $T = 200$ s maksimum rozkładu początkowego zostało przesunięte do punktu o współrzędnych $x = 29$ m, $y = 29$ m. Tymczasem w rozwiązaniu numerycznym zawsze występuje silna dyfuzja numeryczna. Dla $C_x = C_y = 1/2$, co odpowiada $\Delta t = 5$ s, otrzymuje się rozkład niesymetryczny. Ponieważ $\overline{D}^n_{11} < \overline{D}^n_{22}$, rozkład jest wydłużony w kierunku normalnym do wektora prędkości (rys. 7.28) i jednocześnie silnie spłaszczony. Maksymalna koncentracja została zredukowana do wartości $C = 0{,}406$. Jest to typowy wynik dla jawnego schematu „pod prąd".

Omawianie problemu dyfuzji numerycznej zakończmy podsumowaniem jej roli podczas rozwiązywania różnych zagadnień i możliwych konsekwencji ewentualnego zaniedbania wykonania analizy dokładności. Rozwiązanie równania adwekcji-dyfuzji (7.130) metodą dyssypatywną modyfikuje je następująco:

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} - (D + D_n)\frac{\partial^2 C}{\partial x^2} = 0\,. \tag{7.160}$$



Rys. 7.27. Czysta adwekcja wzdłuż przekątnej zbiornika



Rys. 7.28. Zniekształcenie początkowego rozkładu koncentracji przez proces dyfuzji numerycznej w przypadku rozwiązania dwuwymiarowego równania adwekcji schematem „pod prąd”

Oznacza to silniejsze wygładzenie rozwiązania niż wynikające z dyfuzji fizycznej, którą charakteryzuje współczynnik D. Zauważmy, że brak oszacowania dyfuzji numerycznej w tej sytuacji doprowadzi do błędnego oszacowania wartości współczynnika D, ponieważ w rozwiązaniu występuje łączny efekt dyfuzji fizycznej i numerycznej. Proces dyfuzji fizycznej zostanie częściowo zastąpiony procesem dyfuzji numerycznej. Typowe schematy numeryczne dla praktycznie stosowanych wymiarów siatki Δx oraz Δt generują dyfuzję numeryczną o intensywności porównywalnej z dyfuzją fizyczną ($D_n \approx D$) (Szymkiewicz, 2000).

Są jednak przypadki, kiedy dyfuzję numeryczną wprowadza się do rozwiązania celowo z pełną świadomością. Celem takiego podejścia jest wprowadzenie do rozwiązania efektów niemożliwych do uzyskania, gdyż procesy je wywołujące nie są reprezentowane w równaniu problemu.

W niektórych przypadkach dyfuzja numeryczna wykorzystywana jest w celu zapewnienia gładkiego rozwiązania. W tym sensie można mówić o jej pozytywnej roli. Mianowicie, rozwiązując równania hiperboliczne, w których nie są reprezentowane procesy dyssypacyjne lub równania paraboliczne – w których procesy dyssypacyjne odgrywają niewielką rolę – niekiedy celowo wprowadza się dyfuzję numeryczną. Celem tego zabiegu jest zapewnienie gładkiego rozwiązania. Stopień wygładzenia rozwiązania jest na tyle niewielki, że wyniki mogą być uznane za poprawne. Przykładem tego typu podejścia jest metoda „sztucznej lepkości" (w j. angielskim: artificial viscosity) stosowana do rozwiązania równań nieustalonego przepływu ze swobodną powierzchnią (Cunge, Holly i Verwey, 1980) oraz równań przepływu płynów ściśliwych (Potter, 1973). Wprowadzanie dyfuzji numerycznej jest często stosowanym zabiegiem w wielu specjalnych algorytmach rozwiązywania nieliniowych równań hiperbolicznych z nieciągłościami, w celu eliminacji niefizycznych oscylacji (Chang, 1995).

Niestety, znane są również przypadki, kiedy dyfuzja numeryczna generowana przez zastosowaną metodę numeryczną wprowadzana jest w sposób nieświadomy. Efekt jej działania w postaci wygładzania rozwiązania i tłumienia amplitudy fal jest błędnie interpretowany jako efekt procesów fizycznych, reprezentowanych w równaniach opisujących problem.

8 Algorytmy rozwiązania równań różniczkowych o pochodnych cząstkowych metodą różnic skończonych

8.1. Rozwiązanie jednowymiarowego równania filtracji nieustalonej

Zasady numerycznego całkowania równania filtracji nieustalonej metodą różnic skończonych omówimy na przykładzie jednowymiarowego równania dyfuzji (7.6), rozpatrując trzy podstawowe schematy numeryczne:

a) schemat jawny (explicit scheme),
b) schemat całkowicie niejawny (implicit scheme),
c) schemat Cranka-Nicolsona.

Zakładając, że parametry ruchu są identyczne we wszystkich płaszczyznach pionowych równoległych do osi x, czyli ruch nie zależy od y oraz zaniedbując człon źródłowy ($w = 0$), równanie filtracji (7.6) możemy zapisać w postaci równania dyfuzji (7.18).

$$\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(\frac{k(h-z)}{\mu} \frac{\partial h}{\partial x} \right). \tag{8.1}$$

Równanie tej postaci opisuje szereg różnych procesów. Między innymi opisuje ono przepływ wody w obszarze między dwoma biegnącymi równolegle (w kierunku osi y) kanałami, w których poziomy wody zmieniają się w czasie niezależnie od siebie (rys. 7.4), a miąższość warstwy wodonośnej jest na tyle mała, że uzasadnia założenie jednostajnego rozkładu prędkości filtracji w pionie, a tym samym zredukowanie wymiarowości zagadnienia. W tym przypadku współczynnik dyfuzji definiowany jest następująco:

$$D = \frac{k \cdot (h-z)}{\mu} \ , \tag{8.2}$$

gdzie: k – współczynnik filtracji,
h – wysokość piezometryczna,
z – rzędna spągu warstwy wodonośnej,
μ – porowatość efektywna.

Z powyższego wzoru wynika, że współczynnik dyfuzji hydraulicznej jest funkcją rozwiązania równania (8.1): $D = D(h)$. Konsekwencją tego faktu będzie nieliniowość układu równań algebraicznych aproksymujących równanie (8.1). Celem uproszczenia algorytmu rozwiązania można dokonać linearyzacji. Polega ona na przyjęciu stałego współczynnika dyfuzji, obliczonego na podstawie średnich wartości parametrów występujących w równaniu (8.2) $D = \bar{k} \cdot (\bar{h} - \bar{z})/\mu$ gdzie kreska nad symbolem oznacza parametr wymieniony w (8.2),

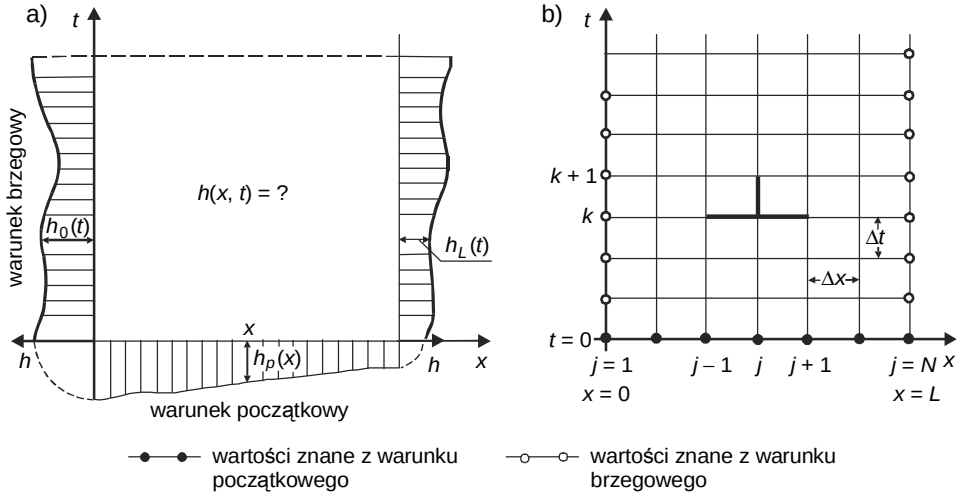
o wartości uśrednionej w obszarze całkowania. Podejście takie jest dopuszczalne, jeśli zakres zmienności funkcji $h(x, t)$ oraz $z(x)$ jest odpowiednio mały. Powyższe uproszczenie umożliwia zapisanie równania (8.1) w postaci liniowego równania dyfuzji

$$\frac{\partial h}{\partial t} = D\frac{\partial^2 h}{\partial x^2}, \tag{8.3}$$

Jak wspomniano w podrozdziale 7.2, do rozwiązania równania różniczkowego cząstkowego konieczne jest podanie warunków dodatkowych na granicy obszaru rozwiązania. W rozważanym wypadku warunki te będą następujące:

— warunek początkowy $h(x, t = 0) = h_p(x)$ dla $0 \le x \le L$,
— warunki brzegowe $h(0, t) = h_0(t)$ dla $t \ge 0$,
$h(L, t) = h_L(t)$ dla $t \ge 0$,

przy czym funkcje h_p, h_0 i h_L są zadane, zwykle w postaci tabelarycznej. Rozwiązanie równania (8.3) polega na znalezieniu funkcji $h(x, t)$ spełniającej to równanie w obszarze $0 \le x \le L$ i $t \ge 0$ oraz zadane warunki początkowe i brzegowe (rys. 8.1a).



Rys. 8.1. Warunki początkowe i brzegowe (a) oraz siatka węzłów (b) dla równania (8.3)

Zastąpmy ciągły obszar, w którym poszukujemy rozwiązania w płaszczyźnie (x, t), obszarem dyskretnym, złożonym z punktów przecięcia się linii poziomych, poprowadzonych równolegle do osi x z odstępem Δt oraz linii pionowych, poprowadzonych równolegle do osi t z odstępem Δx. Rozwiązania poszukiwać będziemy w węzłach powstałej siatki prostokątnej (rys. 8.1b). Jak wynika z przyjętych warunków początkowo-brzegowych, wartości funkcji $h(x, t)$ są znane (zadane) w węzłach na linii poziomej $t = 0$ (warunek początkowy) oraz na pionowych $x = 0$ i $x = L$ (warunki brzegowe, rys. 8.1a). Metoda różnic skończonych umożliwia znalezienie wartości h w pozostałych węzłach siatki.

Ad a) ***Jawny schemat różnicowy***

Zastąpmy pochodne w równaniu (8.3) odpowiednimi ilorazami różnicowymi. Stosując dla pochodnej względem czasu iloraz różnicowy przedni, a dla pochodnej II rzędu względem x iloraz różnicowy centralny na poziomie czasowym k, otrzymujemy dla węzła (j, k) wyrażenie

$$\frac{h_j^{k+1} - h_j^k}{\Delta t} = D\frac{h_{j-1}^k - 2h_j^k + h_{j+1}^k}{\Delta x^2}. \tag{8.4}$$

W powyższym równaniu wartości funkcji h we wszystkich węzłach na poziomie k są znane – najpierw z warunku początkowego, potem z bieżących obliczeń. Równanie (8.4) zawiera więc tylko jedną wartość niewiadomą, mianowicie h_j^{k+1} równą

$$h_j^{k+1} = h_j^k + C_d(h_{j-1}^k - 2h_j^k + h_{j+1}^k), \tag{8.5}$$

gdzie:
$$C_d = \frac{D\Delta t}{\Delta x^2}, \tag{8.6}$$

jest tzw. dyfuzyjną liczbą Couranta.

Znając wartości $h_j^{k=0}$ w węzłach $j = 1, 2, ..., N$, możemy za pomocą powyższego wzoru obliczyć wartości $h_j^{k=1}$ dla wszystkich węzłów wewnętrznych $2 \le j \le N-1$ (rys. 8.1b). Ponieważ wartości $h_1^{k=1}$ i $h_N^{k=1}$ są znane z zadanych warunków brzegowych, więc wartości poszukiwanej funkcji określane są w całym wierszu $k = 1$, czyli na poziomie $t + \Delta t$. Powtarzając ten cykl dla kolejnych wartości k, możemy uzyskać rozwiązanie dla dowolnego czasu $t > 0$.

Omówiony wyżej schemat obliczeń, jak widać bardzo prosty, ma pewną wadę. Mianowicie schemat ten jest stabilny, jeśli $C_d \le 0{,}5$. Wynika stąd ograniczenie na dopuszczalną wielkość kroku czasowego Δt. Mianowicie, ponieważ musi zachodzić relacja

$$\frac{D\Delta t}{\Delta x^2} \le \frac{1}{2},$$

to numeryczna stabilność wymaga spełnienia warunku

$$\Delta t \le \frac{\Delta x^2}{2D}. \tag{8.7}$$

Wartość C_d ma także wpływ na dokładność aproksymacji równania (8.1). Zastosowana powyżej aproksymacja jest najdokładniejsza dla $C_d = 1/6$ (Demidowicz, Maron i Szuwałowa, 1965). Warto także zauważyć, że dla $C_d = 0{,}5$, co jest górną dopuszczalną granicą, wzór (8.5) sprowadza się do postaci

$$h_j^{k+1} = \frac{1}{2}(h_{j-1}^k + h_{j+1}^k). \tag{8.8}$$

Obliczenia za pomocą powyższego wzoru prowadzić można nawet ręcznie.

Ad b) ***Całkowicie niejawny schemat różnicowy***

Rozważmy powtórnie zlinearyzowane równanie (8.3)

$$\frac{\partial h}{\partial t} = D\frac{\partial^2 h}{\partial x^2}, \tag{8.9}$$

z warunkiem początkowym $h(x, t = 0) = h_p(x)$ dla $0 \le x \le L$ oraz warunkami brzegowymi:

$$h(x = 0, t) = h_0(t), \quad h(x = L, t) = h_L(t).$$

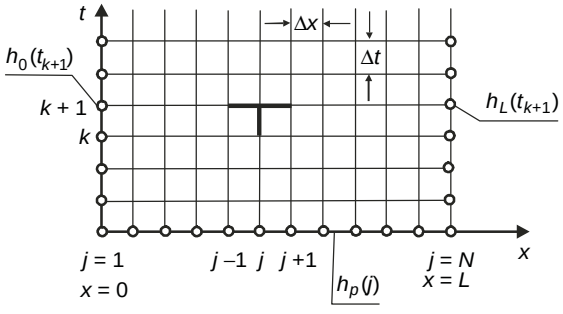
Zdyskretyzujmy obszar $0 \le x \le L$, $t \ge 0$ jak poprzednio, oraz zastąpmy pochodne w równaniu (8.9) ilorazami różnicowymi w sposób następujący (rys. 8.2):

$$\frac{h_j^{k+1} - h_j^k}{\Delta t} = D \frac{h_{j-1}^{k+1} - 2h_j^{k+1} + h_{j+1}^{k+1}}{\Delta x^2}. \tag{8.10}$$

Równanie to można uporządkować, zapisując niewiadome po jednej, zaś wiadome po drugiej stronie znaku równości. Otrzymujemy

$$-C_d \cdot h_{j-1}^{k+1} + (1+2C_d) h_j^{k+1} - C_d h_{j+1}^{k+1} = h_j^k, \tag{8.11}$$

gdzie C_d jest dyfuzyjną liczbą Couranta zdefiniowaną zależnością (8.6).



Rys. 8.2. Niejawny schemat aproksymacji różnicowej równania (8.9)

Jak widać, pochodna II rzędu względem x została tutaj aproksymowana – nie jak poprzednio na poziomie k, lecz na poziomie $k + 1$. W równaniu (8.11) występują zatem trzy niewiadome reprezentujące wartości poszukiwanej funkcji w trzech sąsiadujących ze sobą węzłach. Podobnych równań możemy napisać tyle, ile jest węzłów wewnętrznych, czyli $N - 2$ (rys. 8.2), zatem otrzymujemy w rezultacie układ równań algebraicznych liniowych, w którym występuje N wartości węzłowych. Układ zamykamy, wprowadzając dwa dodatkowe równania wynikające z zadanych warunków brzegowych:

$$h_1^{k+1} = h_0(t_{k+1}), \quad h_N^{k+1} = h_L(t_{k+1}).$$

Układ ten można zapisać w notacji macierzowej

$$\mathbf{A}\,\mathbf{X} = \mathbf{W}, \tag{8.12}$$

gdzie: **A** – macierz współczynników,
X – wektor niewiadomych,
W – wektor prawych stron.

$$\mathbf{A} = \begin{bmatrix} b_1 & c_1 & & & & & & \\ a_2 & b_2 & c_2 & & & & & \\ & a_3 & b_3 & c_3 & & & & \\ & & \cdot & \cdot & \cdot & & & \\ & & & \cdot & \cdot & \cdot & & \\ & & & & \cdot & \cdot & \cdot & \\ & & & & & a_{N-1} & b_{N-1} & c_{N-1} \\ & & & & & & a_N & b_N \end{bmatrix}, \quad \mathbf{X} = \begin{Bmatrix} h_1^{k+1} \\ h_2^{k+1} \\ h_3^{k+1} \\ \vdots \\ h_{N-1}^{k+1} \\ h_N^{k+1} \end{Bmatrix}, \quad \mathbf{W} = \begin{Bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_{N-1} \\ w_N \end{Bmatrix}$$

przy czym: $b_1 = 1, \quad c_1 = 0, \quad w_1 = h_0(t_{k+1})$,

$a_j = -C_d, \quad b_j = 1 + 2C_d, \quad c_j = -C_d, \quad w_j = h_j^k \quad (j = 2, 3, \ldots, N-1)$,

$a_N = 0, \quad b_N = 1, \quad w_N = h_L(t_{k+1})$.

Jak widać, otrzymany układ równań liniowych ma trójdiagonalną macierz współczynników. Można go rozwiązać, stosując bardzo efektywny algorytm rozwiązania, przedstawiony w rozdziale 1.

Ten sposób rozwiązania równania dyfuzji jest bardziej skomplikowany niż przedstawiony wcześniej, w którym zastosowano schemat jawny. W tym wypadku nie można wprost wyznaczyć wartości funkcji h w kolejnych węzłach j na poziomie czasu $k + 1$. Należy to zrobić jednocześnie, rozwiązując układ równań. Ten dodatkowy nakład pracy jest rekompensowany bardzo cenną własnością schematu niejawnego: jest on bezwarunkowo stabilny, a to oznacza możliwość całkowania równania z dużym krokiem Δt, gdyż ze względu na stabilność numeryczną rozwiązania nie ma tutaj ograniczeń na stosunek dobranych wymiarów siatki Δx i Δt.

Zwykle badanie stabilności schematu numerycznego wykonuje się metodą Neumanna, analizując rozwinięcie funkcji w szereg Fouriera w sposób opisany w podrozdziale 7.4. O stabilności można wnioskować również na podstawie właściwości macierzy współczynników układu równań, gdyż stabilność lub niestabilność schematu numerycznego implikuje cechy tej macierzy. Mianowicie, układ równań liniowych ma jednoznaczne rozwiązanie, gdy w jego macierzy współczynników dominuje główna przekątna, tzn. gdy:

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{ij}| \quad (i = 1, 2, \ldots, N), \tag{8.13}$$

gdzie: a_{ii} – element leżący na głównej przekątnej,
N – wymiar układu.

W przypadku układu (8.12) będziemy mieli

$$|1 + 2C_d| \geq |-C_d| + |-C_d|. \tag{8.14}$$

Ponieważ D, Δt i Δx są zawsze większe od zera, to również liczba Couranta C_d będzie zawsze większa od zera. Zatem dla dowolnych wartości Δt i Δx warunek (8.13) będzie spełniony zawsze, gdyż

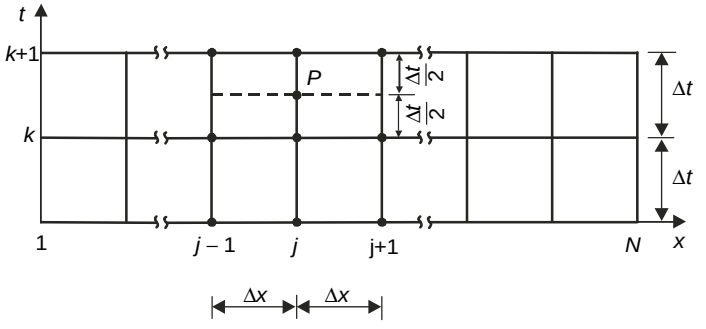
$$1 + 2\,C_d > 2C_d. \tag{8.15}$$

Schemat niejawny, mimo że dla przejścia z poziomu k do $k + 1$ wymaga nieco większego nakładu pracy obliczeniowej niż schemat jawny, jest chętnie stosowany w praktyce. Nie występuje tu, jak wyżej wykazano, żadne ograniczenie na stosunek kroku czasowego i przestrzennego. Ma to istotne znaczenie. Obliczenia można bowiem prowadzić z dużym krokiem czasowym, czyli wykonać znacznie mniej kroków, a w rezultacie końcowym – w krótszym czasie.

Ad c) ***Schemat Cranka-Nicolsona***

Omówione wcześniej schematy rozwiązania równania dyfuzji dawały błąd aproksymacji rzędu $O(\Delta t, \Delta x^2)$. Równanie dyfuzji można rozwiązać schematem bardziej dokładnym. Jest nim schemat Cranka-Nicolsona. W schemacie tym aproksymację równania wykonuje się w punkcie P leżącym w połowie przedziału czasowego, czyli na poziomie

czasu $t + \Delta t/2$ (rys. 8.3). Warunki początkowe i brzegowe są identyczne jak w poprzednich schematach.



Rys. 8.3. Siatka węzłów stosowana w schemacie Cranka-Nicolsona

Zatem pochodną względem czasu aproksymuje się ilorazem różnicowym centralnym

$$\left.\frac{\partial h}{\partial t}\right|_p = \frac{h_j^{k+1} - h_j^k}{\Delta t}. \tag{8.16}$$

Natomiast pochodną II rzędu względem x aproksymuje się za pomocą średniej arytmetycznej z aproksymacji na poziomie czasowym k oraz $k + 1$

$$\left.\frac{\partial^2 h}{\partial x^2}\right|_p \approx \frac{1}{2}\left(\left.\frac{\partial^2 h}{\partial x^2}\right|_j^k + \left.\frac{\partial^2 h}{\partial x^2}\right|_j^{k+1}\right) \approx \frac{1}{2}\left(\frac{h_{j-1}^k - 2h_j^k + h_{j+1}^k}{\Delta x^2} + \frac{h_{j-1}^{k+1} - 2h_j^{k+1} + h_{j+1}^{k+1}}{\Delta x^2}\right). \tag{8.17}$$

Podstawiając powyższe aproksymacje do równania filtracji (8.3), otrzymujemy równanie algebraiczne

$$\frac{h_j^{k+1} - h_j^k}{\Delta t} = \frac{D}{2}\left(\frac{h_{j-1}^k - 2h_j^k + h_{j+1}^k}{\Delta x^2} + \frac{h_{j-1}^{k+1} - 2h_j^{k+1} + h_{j+1}^{k+1}}{\Delta x^2}\right), \tag{8.18}$$

które po wprowadzeniu dyfuzyjnej liczby Couranta (8.6) i uporządkowaniu można zapisać następująco:

$$-\frac{C_d}{2}h_{j-1}^{k+1} + (1 + C_d)h_j^{k+1} - \frac{C_d}{2}h_{j+1}^{k+1} = h_j^k + \frac{C_d}{2}(h_{j-1}^k - 2h_j^k + h_{j+1}^k). \tag{8.19}$$

Równanie tego typu można napisać dla każdego węzła wewnętrznego, czyli dla $j = 2, 3, ..., N - 1$. Otrzymujemy układ równań, który zamykamy wprowadzając dwa dodatkowe równania wynikające z zadanych warunków brzegowych. Ostatecznie powstaje układ równań podobny do układu otrzymanego w przypadku schematu całkowicie niejawnego. Ma on postać (8.12), przy czym współczynniki macierzy **A** oraz wektora wyrazów wolnych **W** są zdefiniowane następująco:

$$b_1 = 1, \quad c_1 = 0, \quad w_1 = h_0(t_{k+1}),$$

$$a_j = -\frac{C_d}{2}, \quad b_j = 1 + C_d, \quad c_j = -\frac{C_d}{2} \quad (j = 2, 3, \ldots, N - 1),$$

$$w_j = h_j^k + \frac{C_d}{2}(h_{j-1}^k - 2h_j^k + h_{j+1}^k) \quad (j = 2,3,\ldots,N-1),$$

$$a_N = 0, \quad b_N = 1, \quad w_N = h_L(t_{k+1}).$$

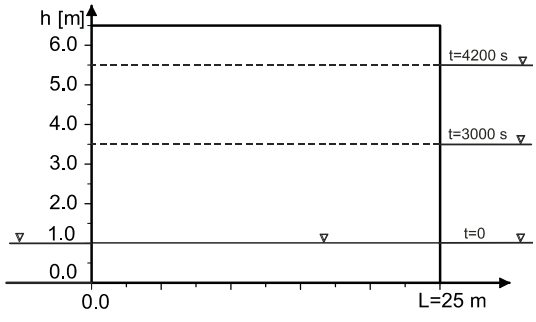
Rozwiązując ten układ znaną metodą Thomasa, otrzymujemy poszukiwane wartości funkcji h we wszystkich węzłach na poziomie czasu $k + 1$.

W sposób zupełnie podobny jak w przypadku schematu niejawnego można wykazać, że macierz współczynników układu ma dominującą główną przekątną. Zatem układ zawsze będzie miał jednoznaczne rozwiązanie, a to oznacza absolutną stabilność schematu Cranka-Nicolsona. Schemat zapewnia aproksymację rzędu $O(\Delta t^2, \Delta x^2)$. Jest on najczęściej stosowanym schematem do rozwiązania równania dyfuzji.

■ *Przykład 8.1*

Filtracja przez groblę prostokątną

Prostokątna grobla wykonana z gruntu przepuszczalnego i posadowiona na gruncie nieprzepuszczalnym oddziela 2 zbiorniki (rys. 8.1.1).



Rys. 8.1.1. Schemat grobli oddzielającej dwa zbiorniki

W chwili początkowej poziom wody w obu zbiornikach jest jednakowy i wynosi 1,0 m ponad spąg warstwy nieprzepuszczalnej. W zbiorniku lewym ($x = 0$) poziom wody nie zmienia się w czasie natomiast poziom wody w zbiorniku ($x = L$) prawym zmienia się następująco:

$$h_L(t) = \begin{cases} 1{,}0 \text{ m dla } t \le 0 \\ 1{,}0 + 2{,}5t/3000 \text{ m} \quad \text{dla } 0 < t \le 3000 \text{ s} \\ 3{,}5 + 2{,}0(t - 3000)/(4200 - 3000) \text{ m} \quad \text{dla } 3000 \text{ s} < t \le 4200 \text{ s} \\ 5{,}5 \text{ m} \quad \text{dla } t > 4200 \text{ s}. \end{cases} \tag{8.1.1}$$

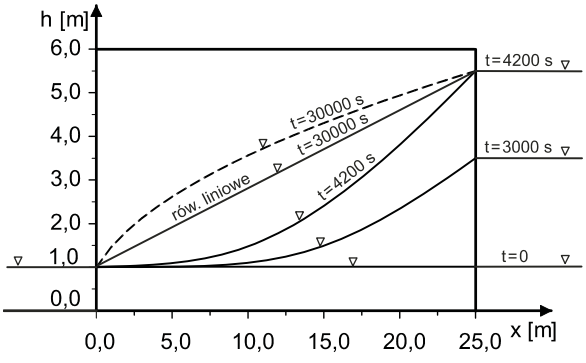
Wyznaczyć chwilowe położenia zwierciadła wody w grobli wywołane zmianami napełnienia prawego zbiornika, jeżeli wiadomo, że:
— współczynnik filtracji wynosi $k = 0{,}002$ m/s,
— porowatość efektywna wynosi $\mu = 0{,}20$,

— rzędna spągu warstwy wodonośnej $z(x) = 0{,}0$ m,
— szerokość grobli wynosi $L = 25{,}0$ m.

Zmieniające się położenie zwierciadła w grobli otrzymamy rozwiązując równanie filtracji (8.3) dla następujących warunków początkowo-brzegowych:
— *warunek początkowy*
w chwili $t = 0$ poziom wody w grobli jest równy poziomowi wody w zbiornikach czyli $h(t = 0, x) = 1{,}0$ m dla $0 \le x \le L$;

— *warunki brzegowe*
– na brzegu $x = 0$ poziom wody jest stały i wynosi $h(t, x = 0) = 1{,}0$ m dla $t \ge 0$,
– na brzegu $x = L$ poziom wody zmienia się zgodnie z równaniem (7.1.1) czyli $h(t, x = L) = h_L(t)$ dla $t \ge 0$.

Równanie (8.3) należy rozwiązać metodą różnic skończonych – schematem Cranka-Nicolsona. Obliczone chwilowe zwierciadła wody w grobli przedstawione na rys. 8.1.2 otrzymano rozwiązując liniowe równanie dyfuzji (8.3) przyjmując krok przestrzenny $\Delta x = 1{,}0$ m oraz krok czasowy $\Delta t = 300$ s.



Rys. 8.1.2. Obliczone chwilowe położenia zwierciadła wody w grobli oddzielającej dwa zbiorniki

Dla porównania, na tym samym rysunku naniesiono końcowy układ zwierciadła wody w grobli otrzymany dla równania nieliniowego (8.1). Jak można zauważyć, linearyzacja równania ma poważne konsekwencje. ■

8.2. Rozwiązanie układu równań de Saint-Venanta

W praktyce inżynierskiej bardzo często spotykamy się z potrzebą rozwiązania równań de Saint-Venanta opisujących wolnozmienny przepływ nieustalony w kanale otwartym. Z tego powodu zagadnieniem tym interesowano się od dłuższego czasu. Efektem jest szereg algorytmów numerycznego rozwiązania opracowanych w ciągu ostatnich 30 lat.

Schematem najczęściej i najchętniej stosowanym jest schemat czteropunktowy niejawny, a dokładniej jego szczególna wersja – schemat Preissmanna. Można w zasadzie powiedzieć, że jest on podstawowym narzędziem rozwiązywania równań przepływu nieustalonego. O jego powodzeniu zadecydowały następujące cenne cechy (Cunge, Holly i Verwey, 1980):

— schemat wykorzystuje nieprzesuniętą siatkę węzłów, co pozwala obliczać obydwie niewiadome w każdym węźle siatki (w przypadku kanałów naturalnych ma to duże znaczenie);
— schemat wiąże zmienne wyłącznie z dwóch sąsiednich węzłów x_j oraz x_{j+1}, co umożliwia łatwe uwzględnienie zmiennego kroku przestrzennego bez zakłócania dokładności aproksymacji;
— schemat jest aproksymacją pierwszego rzędu, zaś w pewnym szczególnym przypadku – aproksymacją drugiego rzędu;
— schemat zapewnia dokładne rozwiązanie liniowych równań falowych (5.3) dla odpowiednio dobranych wartości Δx i Δt, co pozwala na porównanie rozwiązań numerycznych z rozwiązaniem dokładnym;
— schemat jest niejawny i absolutnie stabilny, a zatem nie wymaga ograniczenia wartości kroku całkowania po czasie.

W praktycznych zastosowaniach wygodniej jest stosować układ równań de Saint Venanta w postaci, w której zmiennymi zależnymi są natężenie przepływu Q oraz rzędna zwierciadła wody h. W celu otrzymania tej wersji równań, układ (7.10) i (7.11) przekształcamy wyrażając głębokość i prędkość przepływu odpowiednio przez rzędną zwierciadła wody i natężenie przepływu. Po odpowiednich przekształceniach (Szymkiewicz, 2000; 2010) otrzymuje się:

$$\frac{\partial Q}{\partial t}+\frac{\partial}{\partial x}\left(\frac{Q^2}{A}\right)+g\cdot A\frac{\partial h}{\partial x}+\frac{g\cdot n^2|Q|Q}{A\cdot R^{4/3}}=0 \tag{8.20}$$

$$\frac{\partial h}{\partial t}+\frac{1}{B}\frac{\partial Q}{\partial x}=0 \tag{8.21}$$

gdzie: Q – natężenie przepływu,
h – rzędna zwierciadła wody liczona od przyjętego poziomu porównawczego,
A – pole powierzchni przekroju czynnego kanału,
g – przyspieszenie grawitacyjne,
n – współczynnik szorstkości według Manninga,
R – promień hydrauliczny przekroju kanału,
B – szerokość kanału na poziomie zwierciadła wody.

Rozwiążmy układ równań de Saint-Venanta opisujący przepływ w kanale o zmiennym kształcie przekroju poprzecznego i o długości L. Przyjmujemy, że w kanale panuje przepływ nadkrytyczny (spokojny). Zgodnie z informacjami podanymi w rozdziale 7.1, w takim przypadku na każdym brzegu kanału $x = 0$ oraz $x = L$ należy zadać po jednym warunku brzegowym: funkcję $h(x = 0, t) = h_0(t)$ albo $Q(x = 0, t) = Q_0(t)$ oraz funkcję $h(x = L, t) = h_L(t)$ albo $Q(x = L, t) = Q_L(t)$.

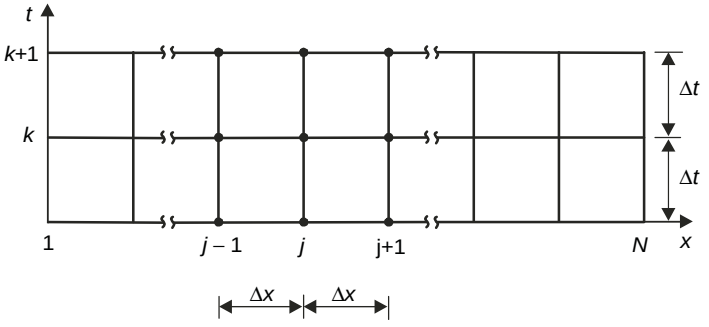
Obszar rozwiązania $0 \le x \le L$ i $t \ge 0$ zastępujemy obszarem dyskretnym. W tym celu konstruujemy siatkę węzłów, jak na rys. 8.4. Wybierzmy jedno dowolne oczko tej siatki utworzone przez węzły (j, k), $(j + 1, k)$, $(j + 1, k + 1)$ i $(j, k + 1)$ (rys. 8.5). Wewnątrz oczka, w połowie odległości pomiędzy przekrojami j oraz $j + 1$, wybierzmy dowolny punkt P, w którym wykonamy aproksymację pochodnych. Położenie tego punktu względem osi czasu jest zmienne. Jego lokalizacja zdefiniowana jest przez parametr θ. Stosowane w tym przypadku formuły aproksymacyjne mają postać:

$$\left.\frac{\partial f}{\partial t}\right|_p \approx \frac{1}{2}\left(\frac{f_j^{k+1}-f_j^k}{\Delta t}+\frac{f_{j+1}^{k+1}-f_{j+1}^k}{\Delta t}\right), \tag{8.22a}$$

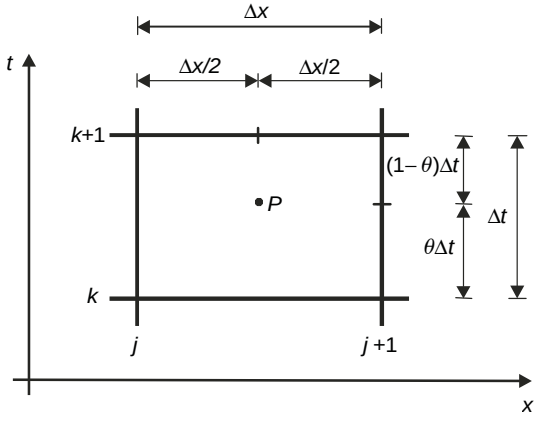
$$\left.\frac{\partial f}{\partial x}\right|_p \approx (1-\theta)\frac{f_{j+1}^k - f_j^k}{\Delta x} + \theta\frac{f_{j+1}^{k+1} - f_j^{k+1}}{\Delta x}, \tag{8.22b}$$

$$f_p \approx \frac{1}{2}\left[\left(\theta f_j^{k+1} + (1-\theta) f_j^k\right) + \left(\theta f_{j+1}^{k+1} + (1-\theta) f_{j+1}^k\right)\right], \tag{8.22c}$$

gdzie: θ – parametr wagowy z przedziału $\langle 0, 1\rangle$,
j – indeks węzła,
k – indeks poziomu czasowego.



Rys. 8.4. Siatka węzłów stosowana w schemacie Preissmanna



Rys. 8.5. Oczko siatki stosowanej w schemacie Preissmanna

Zastosujmy powyższe formuły do aproksymacji pochodnych i funkcji występujących w układzie równań de Saint-Venanta. Równanie dynamiczne (8.20) przyjmie postać następującego równania algebraicznego:

$$\frac{1}{2}\frac{Q_j^{k+1}-Q_j^k}{\Delta t}+\frac{1}{2}\frac{Q_{j+1}^{k+1}-Q_{j+1}^k}{\Delta t}+$$
$$+\frac{(1-\theta)}{\Delta x}\left(\left(\frac{Q^2}{A}\right)_{j+1}^{k}-\left(\frac{Q^2}{A}\right)_j^{k}\right)+\frac{\theta}{\Delta x}\left(\left(\frac{Q^2}{A}\right)_{j+1}^{k+1}-\left(\frac{Q^2}{A}\right)_j^{k+1}\right)+ \tag{8.23a}$$
$$+gA_p\left((1-\theta)\frac{h_{j+1}^k-h_j^k}{\Delta x}+\theta\frac{h_{j+1}^{k+1}-h_j^{k+1}}{\Delta x}\right)+\left(\frac{gn^2}{R^{4/3}}\frac{|Q|Q}{A}\right)_P=0\,.$$

Podobnie wykonana aproksymacja równania ciągłości (8.21) pozwala zapisać je w postaci

$$\frac{1}{2}\frac{h_j^{k+1}-h_j^k}{\Delta t}+\frac{1}{2}\frac{h_{j+1}^{k+1}-h_{j+1}^k}{\Delta t}+\frac{1}{B_P}\left((1-\theta)\frac{Q_{j+1}^k-Q_j^k}{\Delta x}+\theta\frac{Q_{j+1}^{k+1}-Q_j^{k+1}}{\Delta x}\right)=0\,. \tag{8.23b}$$

W obu powyższych równaniach indeks *P* oznacza aproksymację funkcji lub wyrażenia arytmetycznego w punkcie *P* zgodnie z formułą (8.22). Zatem będzie

$$A_p=\frac{1}{2}\left[\theta A_j^{k+1}+(1-\theta)A_j^k\right]+\frac{1}{2}\left[\theta A_{j+1}^{k+1}+(1-\theta)A_{j+1}^k\right], \tag{8.24a}$$

$$B_p=\frac{1}{2}\left[\theta B_j^{k+1}+(1-\theta)B_j^k\right]+\frac{1}{2}\left[\theta B_{j+1}^{k+1}+(1-\theta)B_{j+1}^k\right], \tag{8.24b}$$

$$\left(\frac{gn^2}{R^{4/3}}\frac{|Q|Q}{A}\right)_P=\frac{1}{2}\left[\theta\left(\frac{gn^2}{R^{4/3}}\frac{|Q|Q}{A}\right)_j^{k+1}+(1-\theta)\left(\frac{gn^2}{R^{4/3}}\frac{|Q|Q}{A}\right)_j^{k}\right]+$$
$$+\frac{1}{2}\left[\theta\left(\frac{gn^2}{R^{4/3}}\frac{|Q|Q}{A}\right)_{j+1}^{k+1}+(1-\theta)\left(\frac{gn^2}{R^{4/3}}\frac{|Q|Q}{A}\right)_{j+1}^{k}\right]. \tag{8.24c}$$

W równaniach (8.23a, b) nieznane są wartości rzędnej zwierciadła wody h_j^{k+1} i h_{j+1}^{k+1} oraz natężenie przepływu Q_j^{k+1} i Q_{j+1}^{k+1}, czyli wartości obu funkcji na następnym poziomie czasowym $k+1$. Jeśli *N*, zgodnie z przyjętym oznaczeniem, jest liczbą przekrojów poprzecznych na długości *L* analizowanego odcinka rzeki, to pisząc dla każdego oczka siatki węzłów analogiczny zestaw dwóch równań algebraicznych, otrzymamy układ $2(N-1)$ równań. W równaniach tych występuje $2N$ węzłowych wartości funkcji *h* oraz *Q*. Brakujące dwa równania znane są z zadanych warunków brzegowych. Można je zapisać w następującej postaci:

— dla węzła $j=1$

$$\delta h_1^{k+1}+(1-\delta)Q_1^{k+1}=\delta h_0(t_{k+1})+(1-\delta)Q_0(t_{k+1})\,, \tag{8.25}$$

— dla węzła $j=N$

$$\delta h_N^{k+1}+(1-\delta)Q_N^{k+1}=\delta h_L(t_{k+1})+(1-\delta)Q_L(t_{k+1})\,. \tag{8.26}$$

W obu przypadkach parametr δ może przyjmować wartość 0 lub 1, przy czym $\delta = 1$ oznacza warunek w postaci limnigramu ($h_0(t)$ lub $h_L(t)$), natomiast $\delta = 0$ oznacza warunki brzegowe w postaci hydrogramów ($Q_0(t)$ lub $Q_L(t)$). Powyższe równania zamykają układ otrzymany w wyniku aproksymacji równań de Saint-Venanta. Jest to układ równań algebraicznych nieliniowych. Po wprowadzeniu warunków brzegowych (8.25) i (8.26) można go zapisać w postaci wektorowej

$$\mathbf{F}(\mathbf{X}) = \mathbf{0}, \tag{8.27}$$

gdzie: $\mathbf{X} = (h_1, Q_1, ..., h_j, Q_j, ..., h_N, Q_N)^T$ – wektor niewiadomych,
$\mathbf{F} = (F_1, F_2, ..., F_j, F_{j+1}, ..., F_{2N-1}, F_{2N})^T$ – wektor równań.

Kolejne równania wektora **F** mają postać:

$$F_1(h_1^{k+1}, Q_1^{k+1}) = \delta h_1^{k+1} + (1-\delta)Q_1^{k+1} - \delta h_0(t_{k+1}) - (1-\delta)Q_0(t_{k+1}) = 0\,, \tag{8.28a}$$

$$\vdots$$

$$\left.\begin{aligned} F_{2j}(h_j^{k+1}, Q_j^{k+1}, h_{j+1}^{k+1}, Q_{j+1}^{k+1}) = 0 \\ F_{2j+1}(h_j^{k+1}, Q_j^{k+1}, h_{j+1}^{k+1}, Q_{j+1}^{k+1}) = 0 \end{aligned}\right\} \; j = 1, 2, \ldots, N-1 \tag{8.28b,c}$$

$$\vdots$$

$$F_{2N}(h_N^{k+1}, Q_N^{k+1}) = \delta h_N^{k+1} + (1-\delta)Q_N^{k+1} - \delta h_L(t_{k+1}) - (1-\delta)Q_L(t_{k+1}) = 0. \tag{8.28d}$$

Pierwsze i ostatnie równanie wynika z zadanych warunków brzegowych. Pozostałe równania układu są symbolicznie zapisanymi równaniami odpowiednio (8.23a) oraz (8.23b) dla kolejnych oczek siatki.

Przyjęty sposób dyskretyzacji powoduje, że w każdym równaniu występują jedynie cztery niewiadome, gdyż równania wiążą z sobą wartości funkcji tylko w węzłach tworzących analizowane oczko siatki. Do rozwiązania układu (8.27) należy zastosować metodę iteracyjną. W tym przypadku bardzo skuteczną metodą jest metoda Newtona opisana w punkcie 2.7.2. Przyjmijmy pierwsze przybliżenie wektora **X**

$$\mathbf{X}^{(0)} = \left((h_1^{k+1})^{(0)}, (Q_1^{k+1})^{(0)}, \ldots, (h_j^{k+1})^{(0)}, (Q_j^{k+1})^{(0)}, \ldots, (h_N^{k+1})^{(0)}, (Q_N^{k+1})^{(0)}\right)^T. \tag{8.29}$$

Ponieważ wektor ten nie jest dokładnym rozwiązaniem układu (8.27), jego prawa strona będzie różna od zera

$$\mathbf{F}(\mathbf{X}^{(0)}) = \mathbf{F}^{(0)} \neq \mathbf{0}\,. \tag{8.30}$$

Następne przybliżenie zgodnie z metodą Newtona obliczamy następująco (Dahlquist i Bjorck, 1974):

$$\mathbf{J}^{(0)}(\mathbf{X}^{(1)} - \mathbf{X}^{(0)}) = -\mathbf{F}^{(0)}\,. \tag{8.31}$$

Formułę tę można zapisać w ogólniejszej postaci

$$\mathbf{J}^{(l)}\Delta\mathbf{X}^{(l+1)} = -\mathbf{F}^{(l)}\,, \tag{8.32}$$

gdzie: $\Delta\mathbf{X}^{(l+1)} = \mathbf{X}^{(l+1)} - \mathbf{X}^{(l)}$,
l – indeks iteracji.

Występująca w (8.31) i (8.32) macierz **J** jest jakobianem układu (8.27). Jej elementami są pierwsze pochodne kolejnych równań układu (8.27) liczone względem kolejnych składowych wektora **X**

$$\mathbf{J} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_i} & \cdots & \frac{\partial F_1}{\partial x_M} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_i} & \cdots & \frac{\partial F_2}{\partial x_M} \\ \vdots & \vdots & & \vdots & & \vdots \\ \frac{\partial F_i}{\partial x_1} & \frac{\partial F_i}{\partial x_2} & \cdots & \frac{\partial F_i}{\partial x_i} & \cdots & \frac{\partial F_i}{\partial x_M} \\ \vdots & \vdots & & \vdots & & \vdots \\ \frac{\partial F_M}{\partial x_1} & \frac{\partial F_M}{\partial x_2} & \cdots & \frac{\partial F_M}{\partial x_i} & \cdots & \frac{\partial F_M}{\partial x_M} \end{bmatrix}, \tag{8.33}$$

gdzie $M = 2N$ jest rozmiarem układu (8.27).

Jak już wspomniano wcześniej, w każdym równaniu układu (8.27) wystąpią co najwyżej cztery niewiadome. Oznacza to, że współczynniki przy pozostałych niewiadomych są zerowe. Konsekwencją tego faktu jest pasmowość macierzy Jacobiego (8.33). W pierwszym wierszu wystąpi jeden element niezerowy, w wierszach od drugiego do przedostatniego wystąpią cztery niezerowe elementy, zaś w ostatnim wierszu ponownie wystąpi tylko jeden element różny od zera. Strukturę macierzy **J**, otrzymaną dla odcinka kanału składającego się z pięciu węzłów ($N = 5$), przedstawiono na rysunku 8.6 dla różnych wariantów przyjętych warunków brzegowych na brzegu górnym $x = 0$ oraz brzegu dolnym $x = L$. Macierz **J** jest macierzą pasmową okołoprzekątniową, przy czym szerokość pasma wynosi 5. W przypadku (a) w punkcie $x = 0$ przyjęto warunek $h_1 = h_0(t)$, co odpowiada $\delta = 1$ w równaniu (8.25), zaś w punkcie $x = L$ warunek $Q_N = Q_L(t)$, co odpowiada $\delta = 0$ w równaniu (8.26). Natomiast w przypadku (b) sytuacja jest odwrotna. Na początku kanału warunek brzegowy określa hydrogram $Q_0(t)$, zaś na końcu kanału – limnigram $h_L(t)$.

Proces iteracyjny (8.32) prowadzi się tak długo, aż przyjęte kryterium dokładności rozwiązania zostanie spełnione. Jako kryterium można przyjąć różnicę wartości składowych wektora niewiadomych w kolejnych iteracjach, czyli warunek

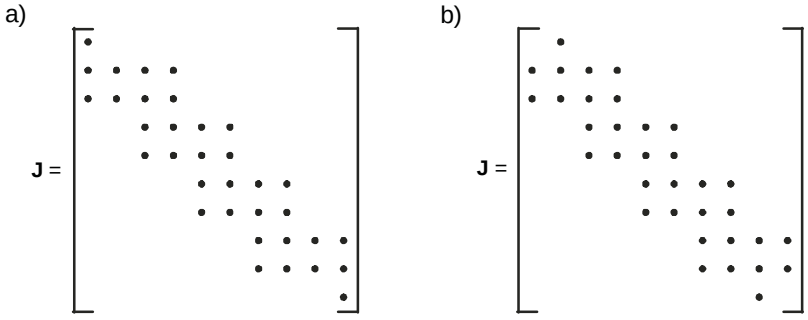
$$\left|(Q_j^{k+1})^{(l+1)} - (Q_j^{k+1})^{(l)}\right| \le \varepsilon_Q \text{ i } \left|(h_j^{k+1})^{(l+1)} - (h_j^{k+1})^{(l)}\right| \le \varepsilon_h \quad (j = 1, 2, ..., N), \tag{8.34}$$

gdzie: ε_Q – dokładność obliczenia natężenia przepływu Q,
ε_h – dokładność obliczenia rzędnej zwierciadła h,
l – indeks iteracji.

Inną miarą dokładności rozwiązania może być odpowiednia norma wektora residualnego $\mathbf{F}^{(l)}$. Jeśli jej wartość będzie mniejsza od przyjętego parametru ε, określającego dokładność rozwiązania, proces iteracyjny kończy się.

Istotną sprawą w każdym procesie iteracyjnym jest przyjęcie początkowego przybliżenia ($l = 0$) wektora niewiadomych. W tym przypadku jako pierwsze przybliżenie można przyjąć wynik obliczeń z poprzedniego kroku czasowego, czyli

$$(Q_j^{k+1})^{(l=0)} = Q_j^k \text{ i } (h_j^{k+1})^{(l=0)} = h_j^k \quad (j = 1, 2, ..., N). \tag{8.35}$$



Rys. 8.6. Struktura macierzy Jacobiego układu (8.27) przy $N = 5$ i różnych warunkach brzegowych

W większości przypadków przyjęcie początkowego przybliżenia zgodnie z (8.35) i wykonanie 2÷3 iteracji zapewnia rozwiązanie z wystarczającą dla celów praktycznych dokładnością.

Z równania (8.32) wynika, że w każdej iteracji nowe przybliżenie wektora niewiadomych otrzymuje się w wyniku rozwiązania układu algebraicznych równań liniowych. Do tego celu zwykle wykorzystuje się dokładną metodę rozwiązania, jak metoda Gaussa lub metoda rozkładu macierzy współczynników. Dla uzyskania efektywnego algorytmu należy stosować warianty wymienionych metod uwzględniające pasmowość macierzy **J**. Pozwala to na istotną oszczędność zarówno pamięci, jak i czasu pracy komputera.

Zastosowany do rozwiązania układu równań de Saint-Venanta schemat zawiera parametr wagowy definiujący lokalizację punktu aproksymacji P wewnątrz oczka siatki węzłów (rys. 8.5). Parametr ten decyduje o właściwościach numerycznych schematu. W celu zbadania tych właściwości wykonuje się analizę stabilności oraz dokładności rozwiązania numerycznego liniowego układu równań falowych, które otrzymujemy upraszczając układ równań de Saint-Venanta:

$$\frac{\partial H}{\partial t} + \overline{H}\frac{\partial U}{\partial x} = 0\,, \tag{8.36}$$

$$\frac{\partial U}{\partial t} + g\frac{\partial H}{\partial x} = 0\,, \tag{8.37}$$

gdzie $\overline{H}$ jest uśrednioną stałą głębokością.

Pochodne w obu równaniach zastępujemy ich aproksymacjami (8.22), otrzymując następujące równania algebraiczne:

$$\frac{1}{2}\frac{H_j^{k+1}-H_j^k}{\Delta t} + \frac{1}{2}\frac{H_{j+1}^{k+1}-H_{j+1}^k}{\Delta t} + \overline{H}(1-\theta)\frac{U_{j+1}^k-U_j^k}{\Delta x} + \overline{H}\theta\frac{U_{j+1}^{k+1}-U_j^{k+1}}{\Delta x} = 0\,, \tag{8.38}$$

$$\frac{1}{2}\frac{U_j^{k+1}-U_j^k}{\Delta t} + \frac{1}{2}\frac{U_{j+1}^{k+1}-U_{j+1}^k}{\Delta t} + g(1-\theta)\frac{H_{j+1}^k-H_j^k}{\Delta x} + g\theta\frac{H_{j+1}^{k+1}-H_j^{k+1}}{\Delta x} = 0\,. \tag{8.39}$$

Badanie stabilności numerycznej niejawnego schematu Preissmanna wykonane metodą Neumanna (Szymkiewicz, 2000) wykazuje, że schemat jest absolutnie stabilny, gdy

$$\theta \geq \frac{1}{2}\,. \tag{8.40}$$

Nawiązując do rysunku (8.5), można powiedzieć, że schemat czteropunktowy będzie absolutnie stabilny, jeśli punkt aproksymacji P będzie położony w górnej połówce oczka siatki. Dla $\theta = 1/2$ punkt P położony jest w środku oczka siatki. W tym przypadku aproksymacja formułami (8.20) i (8.21) jest aproksymacją za pomocą ilorazów różnicowych centralnych, czyli o maksymalnej możliwej w tym przypadku dokładności.

Znajomość rzędu aproksymacji pochodnych w równaniach różniczkowych jest informacją, z której można wysnuć ogólne wnioski o własnościach schematu. Niestety, nie dostarcza ona szczegółowych informacji o błędach generowanych przez schemat, czyli o dokładności rozwiązania. Informacje takie można uzyskać, wykonując analizę dokładności metodą równania zmodyfikowanego. Korzystając z szeregu Taylora wartości węzłowe w równaniach (8.38) i (8.39) wyrażamy przez wartości w węźle ($j + 1$, $k + 1$) (rys. 8.5). Po uwzględnieniu w szeregu Taylora wyrazów z trzecią pochodną włącznie, otrzymujemy oszacowania wartości węzłowych, które wstawiamy do równań różnicowych (8.38) i (8.39). Po uporządkowaniu otrzymujemy zmodyfikowane w wyniku aproksymacji pochodnych równania (8.36) i (8.37). Mają one następującą postać:

$$\begin{aligned}
&\frac{\partial H}{\partial t} + \overline{H}\frac{\partial U}{\partial x} = \frac{\Delta t}{2}\frac{\partial^2 H}{\partial t^2} - \frac{\Delta t^2}{6}\frac{\partial^3 H}{\partial t^3} + \frac{\Delta x}{2}\frac{\partial^2 H}{\partial x \partial t} + \\
&- \frac{\Delta x^2}{4}\frac{\partial^3 H}{\partial x^2 \partial t} - \frac{\Delta x \Delta t}{4}\frac{\partial^3 H}{\partial x \partial t^2} + \overline{H}\frac{\Delta x}{2}\frac{\partial^2 U}{\partial x^2} - \overline{H}\frac{\Delta x^2}{6}\frac{\partial^3 U}{\partial x^3} + \\
&- \overline{H}(1-\theta)\Delta t\frac{\partial^2 U}{\partial x \partial t} - \overline{H}(1-\theta)\frac{\Delta x \Delta t}{2}\frac{\partial^3 U}{\partial x^2 \partial t} - \overline{H}(1-\theta)\frac{\Delta t^2}{2}\frac{\partial^3 U}{\partial x \partial t^2},
\end{aligned} \tag{8.41}$$

$$\begin{aligned}
&\frac{\partial U}{\partial t} + g\frac{\partial H}{\partial x} = \frac{\Delta t}{2}\frac{\partial^2 U}{\partial t^2} - \frac{\Delta t^2}{6}\frac{\partial^3 U}{\partial t^3} + \frac{\Delta x}{2}\frac{\partial^2 U}{\partial x \partial t} - \frac{\Delta x^2}{4}\frac{\partial^3 U}{\partial x^2 \partial t} + \\
&- \frac{\Delta x \Delta t}{4}\frac{\partial^3 U}{\partial x \partial t^2} + g\frac{\Delta x}{2}\frac{\partial^2 H}{\partial x^2} - g\frac{\Delta x^2}{6}\frac{\partial^3 H}{\partial x^3} + g(1-\theta)\Delta t\frac{\partial^2 H}{\partial x \partial t} + \\
&- g(1-\theta)\Delta t\frac{\Delta x \Delta t}{2}\frac{\partial^3 H}{\partial x^2 \partial t} - g(1-\theta)\frac{\Delta t^2}{2}\frac{\partial^3 H}{\partial x \partial t^2}.
\end{aligned} \tag{8.42}$$

W powyższych równaniach dla uproszczenia zapisów pominięto indeksy węzła ($j + 1$, $k + 1$), jednak należy pamiętać, że są one ważne w tym węźle. Po wyeliminowaniu z równań (8.41) i (8.42) pochodnych wyższych rzędów względem czasu, równania zmodyfikowane przyjmą postać

$$\frac{\partial H}{\partial t} + \overline{H}\frac{\partial U}{\partial x} = \nu\frac{\partial^2 H}{\partial x^2} + \varepsilon_1\frac{\partial^3 H}{\partial x^3} + \varepsilon_2\frac{\partial^3 U}{\partial x^3}\ldots, \tag{8.43}$$

$$\frac{\partial U}{\partial t} + g\frac{\partial H}{\partial x} = \nu\frac{\partial^2 U}{\partial x^2} + \varepsilon_1\frac{\partial^3 U}{\partial x^3} + \varepsilon_2\frac{\partial^3 H}{\partial x^3} + \ldots, \tag{8.44}$$

gdzie

$$\nu = \left(\theta - \frac{1}{2}\right)g\overline{H}\Delta t, \tag{8.45}$$

$$\varepsilon_1 = \frac{g\overline{H}\Delta x\Delta t}{2}\left(\frac{1}{2}-\theta\right), \tag{8.46}$$

$$\varepsilon_2 = \frac{g\Delta x^2}{6}\left[(3\theta-2)C_r^2+\frac{1}{2}\right]. \tag{8.47}$$

Występującą w powyższych zależnościach liczbę Couranta C_r definiuje równanie:

$$C_r = \frac{\sqrt{g\overline{H}}\,\Delta t}{\Delta x}. \tag{8.48}$$

Jak widzimy, aproksymacja równań (8.36) i (8.37) schematem czteropunktowym w punkcie P powoduje ich modyfikację. Po prawej stronie równań pojawiły się człony z wyższymi pochodnymi pominiętymi w trakcie aproksymacji pochodnych. Występujący z pochodnymi drugiego rzędu współczynnik ν jest współczynnikiem dyfuzji numerycznej. Natomiast w członach z pochodnymi trzeciego rzędu występują współczynniki ε_1 oraz ε_2 nazywane współczynnikami dyspersji numerycznej.

Z postaci równań zmodyfikowanych (8.43) i (8.44) można wyciągnąć szereg wniosków na temat właściwości numerycznych schematu.

Po pierwsze, łatwo można stwierdzić, że schemat ten daje równania algebraiczne zgodne z równaniami różniczkowymi. Dla Δx, $\Delta t \to 0$ wszystkie człony prawych stron równań zmodyfikowanych zmierzają do zera. W konsekwencji w granicy równania zmodyfikowane (8.43) i (8.44) stają się równaniami falowymi (8.36) i (8.37). Zatem schemat spełnia warunek zgodności. Ponieważ jest on również absolutnie stabilny przy $\theta \geq 1/2$, zatem jest on zbieżny, gdy θ spełnia warunek (8.40).

Po drugie, człon dyfuzji numerycznej znika, gdy przyjmiemy $\theta = 1/2$. W tej sytuacji schemat nie będzie generował dyfuzji numerycznej. Ogólniej mówiąc, schemat będzie niedyssypatywny. Wynika to z faktu uwzględnienia w szeregu Taylora członów z drugimi pochodnymi. Przy tej wartości θ pochodne są aproksymowane ilorazami różnicowymi centralnymi. W błędzie obcięcia zaczynają dominować człony zawierające pochodne trzeciego rzędu. Efektem tego są człony dyspersji numerycznej w równaniach zmodyfikowanych (8.43) i (8.44). Rzeczywiście, o ile przy $\theta = 1/2$ mamy $\nu = 0$, to jednocześnie otrzymujemy

$$\varepsilon_1 = 0, \tag{8.49a}$$

$$\varepsilon_2 = \frac{g\Delta x^2}{12}(1-C_r^2). \tag{8.49b}$$

Obecność członów dyspersji, jak wiemy, modyfikuje prędkości fazowe fal i w konsekwencji, przy nieobecności mechanizmu dyssypacji ($\nu = 0$), wywołują one oscylacje rozwiązania. Zauważmy jednak, że w szczególnym przypadku, gdy liczba Couranta równa jest jedności ($C_r = 1$), człony dyspersji numerycznej zanikają. W tym przypadku równania zmodyfikowane (8.43) i (8.44) są identyczne z równaniami falowymi (8.36) i (8.37). Konsekwencją tego jest dokładne rozwiązanie układu. Zatem przy $\theta = 1/2$ i $C_r = 1$ schemat Preissmanna nie generuje, ani błędu dyssypacji ani błędu dyspersji numerycznej i zapewnia dokładne rozwiązanie równań falowych. Inaczej mówiąc, dla wymienionych wartości θ i C_r schemat jest niedyssypatywny i niedyspersyjny.

Po trzecie, równania zmodyfikowane (8.43) i (8.44) dostarczają identycznych informacji na temat stabilności schematu jak analiza stabilności. Równania (8.43) i (8.44) stanowią układ równań parabolicznych. Dla układu tego formułuje się tzw. problem początkowo-brzegowy, którego rozwiązanie istnieje tylko wtedy, gdy współczynnik dyfuzji jest dodatni: $\nu > 0$. Tymczasem zauważmy, że będzie on ujemny przy $\theta < 1/2$. Zatem dla tej wartości parametru θ problem początkowo-brzegowy dla układu (8.43) i (8.44) jest postawiony niepoprawnie, a jego rozwiązanie nie istnieje. Niemożność uzyskania rozwiązania na gruncie numerycznym traktuje się jako niestabilność metody numerycznej. Zatem dla $\theta < 1/2$ schemat Preissmanna jest niestabilny. Wniosek ten, jak widzimy, całkowicie pokrywa się z warunkiem stabilności (8.40).

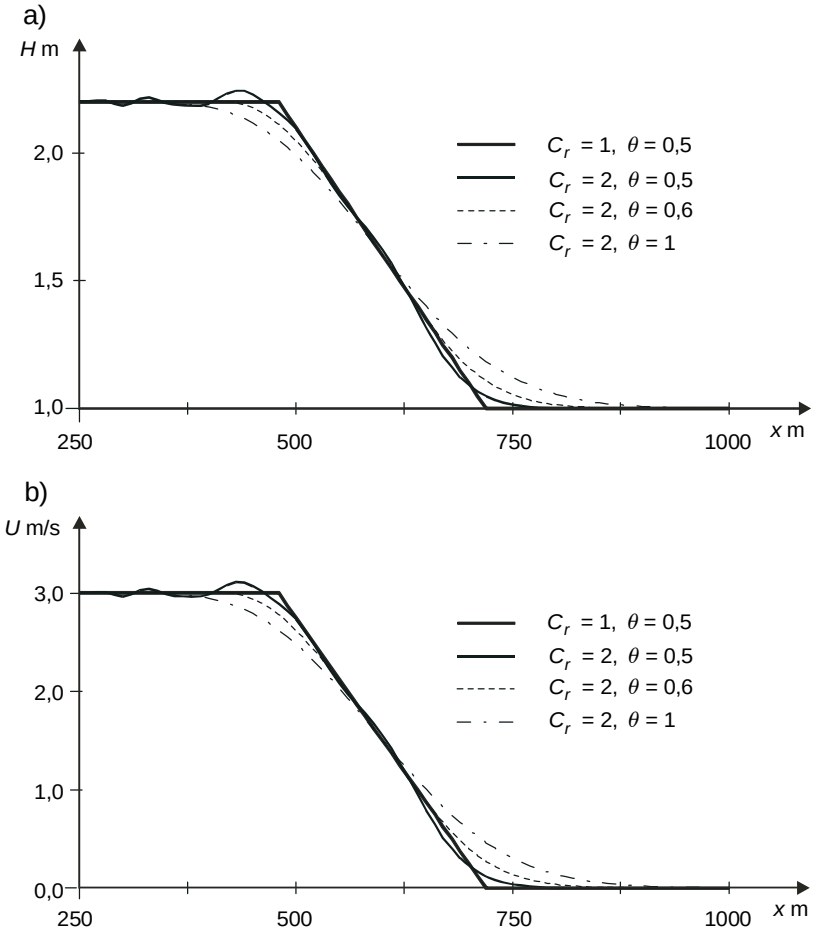
Po czwarte, dyfuzja numeryczna wystąpi w rozwiązaniu równań (8.36) i (8.37), jeśli przyjmiemy $\theta > 1/2$. Przy tych warunkach $\nu > 0$, co oznacza, że problem początkowo-brzegowy jest postawiony poprawnie i jego rozwiązanie istnieje zawsze. Schemat jest więc absolutnie stabilny, niezależnie od przyjętych wartości wymiarów siatki Δx i Δt oraz głębokości przepływu $\overline{H}$. Jednak schemat generuje dyfuzję numeryczną, której wielkość zależy od wymiarów siatki, głębokości $\overline{H}$ oraz wartości parametru θ. Maksymalna dyfuzja numeryczna wystąpi, gdy $\theta = 1$. W tym przypadku schemat zapewnia aproksymację pierwszych pochodnych względem czasu z dokładnością pierwszego rzędu. W błędzie obcięcia dominują teraz człony z drugimi pochodnymi, czyli człony dyfuzyjne. Konsekwencją tego jest wygładzanie rozwiązania w sposób charakterystyczny dla procesu dyfuzji. Schemat jest teraz schematem dyssypatywnym. Eliminuje on oscylacje wywołane dyspersyjnością, lecz jednocześnie zakłóca wyniki obliczeń, powodując wygładzanie rozwiązania i w konsekwencji redukując gradienty. Jeśli rozwiązanie równań (8.36) i (8.37) opisuje propagację gwałtownego wezbrania, to można spodziewać się, że schemat generujący tak dużą dyfuzję numeryczną da wyniki o małej dokładności.

Wszystkie powyższe wnioski potwierdzają obliczenia (Szymkiewicz, 2000). Dobrą ilustracją omówionych cech schematu jest rozwiązanie równań falowych w przypadku nagłej zmiany warunków przepływu w kanale. W tym celu rozpatruje się kanał o długości L = 1000 m, w którym woda pozostaje w spoczynku, wypełniając go do głębokości H = 1 m. Na brzegu x = 0 zadaje się funkcję

$$H(x=0,t) = \begin{cases} 1\ \text{m} & \text{dla} \quad t = 0, \\ 2{,}2\,\dfrac{t}{60}\ \text{m} & \text{dla} \quad 0 < t \le 60\ \text{s}, \\ 2{,}2\ \text{m} & \text{dla} \quad t > 60\ \text{s}. \end{cases}$$

Przyjęty warunek brzegowy wywołuje propagację fali wezbraniowej. Na brzegu $x = L$ przyjęto stałą głębokość równą początkowej: $H(x = L, t)$ = 1 m. Pozostałe dane są następujące: $\overline{H}$ = 1,6 m, g = 10 m/s^2, co oznacza, że fala przemieszcza się z prędkością $(g\,\overline{H}\,)^{1/2}$ = 4 m/s. Kanał podzielono na odcinki o równej długości Δx = 10 m. Obliczenia wykonano dla różnych wartości kroku czasowego Δt oraz różnych wartości parametru θ, a wyniki w postaci funkcji $H(x)$ i $U(x)$ po czasie t = 180 s przedstawione są na rysunku 8.7. Zgodnie z wnioskami wynikającymi z analizy dokładności, schemat zapewnia dokładne rozwiązanie przy C_r = 1 i θ = 1/2. Wyniki uzyskane dla tych wartości parametrów (Δt = 2,5 s) przedstawiają krzywe 1 na rys. 8.7a i 8.7b, reprezentujące odpowiednio głębokości i prędkości wzdłuż osi kanału w chwili t = 180s. Zwiększenie kroku czasowego do wartości Δt = 5 s, co odpowiada liczbie Couranta C_r = 2, spowodowało pojawienie się w rozwiązaniu oscylacji. Wynik

taki jest zgodny z oczekiwaniem. Przy $\theta = 1/2$ schemat jest dyspersyjny, a efektem tego przy braku dyfuzji numerycznej są właśnie oscylacje. Systematyczne zwiększanie wartości parametru θ powoduje wzrost błędu dyssypacji generowanego przez schemat. Coraz większa dyfuzja numeryczna powoduje eliminację oscylacji i coraz silniejsze wygładzanie rozwiązania. Tendencję tę zauważa się porównując kolejne krzywe na rys. 8.7. Podobna sytuacja występuje w przypadku, gdy $C_r < 1$.



Rys. 8.7. Porównanie głębokości i prędkości dla liniowych równań falowych przy różnych wartościach C_r i θ (Szymkiewicz, 2000)

■ *Przykład 8.2*

Transformacja fali wezbraniowej w kanale otwartym

W kanale otwartym o stały spadku dna i przekroju trapezowym przemieszcza się fala wezbraniowa o kształcie danym równaniem:

$$Q_0(t) = q_0 + (q_{\max} - q_0)\left(\frac{t}{t_{\max}}\right)^2 \exp\left(1 - \left(\frac{t}{t_{\max}}\right)^2\right) \tag{8.2.1}$$

gdzie: t – czas,
Q_0 – chwilowe natężenie przepływu w początkowym przekroju kanału,
q_0 – bazowe natężenie przepływu (przed nadejściem fali wezbraniowej),
$q_{\max}$ – kulminacyjne natężenie przepływu,
$t_{\max}$ – czas wystąpienia kulminacji fali.

Wyznaczyć kształt fali wezbraniowej w przekrojach kanału położonych w odległości x = 25,0 km, x = 50,0 km oraz x = 75,0 km od przekroju początkowego.

Kształt transformującej się w kanale fali wezbraniowej otrzymamy rozwiązując numerycznie układ równań de Saint Venanta (8.20) i (8.21) dla następujących warunków początkowych i brzegowych:

— war*unki początkowe*
w chwili $t = 0$ woda w kanale płynie ruchem ustalonym jednostajnym z natężeniem q_0 co oznacza, że $Q(t = 0, x) = q_0$ dla $0 \le x \le L$ zaś głębokość w każdym przekroju jest równa głębokości normalnej wynikającej ze wzoru Manninga czyli $h(t = 0, x) = h_n$ dla $0 \le x \le L$;

— *warunki brzegowe*
na brzegu x=0 zadane jest natężenie przepływu, które opisuje funkcja (8.2.1) czyli $Q(t, x = 0) = Q_0(t)$ dla $t \ge 0$, ponieważ na brzegu $x = L$ nie jest znana ani funkcja $Q_L(t)$ ani funkcja $h_L(t)$ dla $t \ge 0$ przyjmuje się, że zależność pomiędzy nimi opisuje równanie przepływu ustalonego jednostajnego czyli równanie Manninga

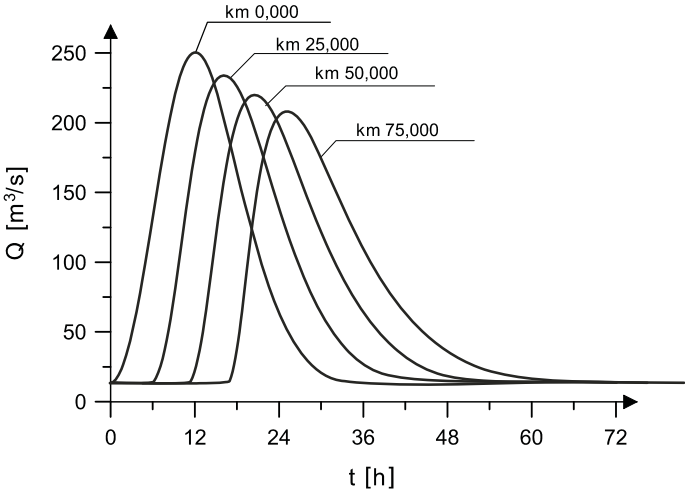
$$Q_L(t) = \frac{1}{n} R\left(h_L(t)\right)^{2/3} \cdot s^{1/2} \cdot A\left(h_L(t)\right) \tag{8.2.2}$$

gdzie: n – współczynnik szorstkości według Manninga,
R – promień hydrauliczny przekroju kanału,
s – podłużny spadek dna kanału,
A – pole powierzchni przekroju czynnego kanału.

Dla wyżej sformułowanych warunków początkowo-brzegowych równania de Saint Venanta rozwiązuje się metodą różnic skończonych – schematem Preissmanna. Do obliczeń przyjęto następujące dane:
— przepływ bazowy q_0 = 13,20 m^3/s,
— przepływ kulminacyjny $q_{\max}$ = 250 m^3/s,
— czas wystąpienia kulminacji fali $t_{\max}$ = 12 h,
— szerokość dna kanału B = 10,0 m,
— pochylenie skarp kanału m = 1,50,
— spadek podłużny dna kanału s = 0,0005,
— współczynnik szorstkości kanału według Manninga n = 0.050,
— długość kanału L = 80 km,
— odległość pomiędzy przekrojami obliczeniowymi Δx = 500 m,
— krok całkowania w czasie Δt = 600 s,
— parametr wagowy θ = 0,525.

Kolejne krzywe na rys. 8.2.1 reprezentują kształt fali wezbraniowej w wybranych przekrojach kanału, która pojawiła się w jego przekroju początkowym (równanie 8.2.1). Jak widać, propagująca fala wezbraniowa ulega transformacji polegającej na systematycznej redukcji

jej kulminacji i wydłużaniu podstawy. Ponieważ w układzie równań de Saint Venanta równanie ciągłości (8.21) reprezentuje zasadę zachowania masy, to przy braku dopływu bocznego, jak w rozwiązywanym przykładzie, powierzchnie zawarte pod kolejnymi krzywymi są jednakowe.



Rys. 8.2.1. Obliczone natężenia przepływu w kolejnych przekrojach kanału ■

8.3. Rozwiązanie jednowymiarowego równania adwekcji-dyfuzji

Rozpatrzmy jednowymiarowe równanie adwekcji-dyfuzji (7.16):

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} - D\frac{\partial^2 C}{\partial x^2} = 0\,, \tag{8.49}$$

gdzie: C – koncentracja domieszki przenoszonej przez płynącą wodę,
U – prędkość przepływu,
D – współczynnik dyfuzji.

Równanie tego typu stosowane jest bardzo często w inżynierii wodnej. Opisuje ono przenoszenie przez płynącą wodę rozpuszczonych w niej domieszek, przenoszenie energii cieplnej itd. Równaniem tego typu jest również równanie (7.19), reprezentujące uproszczony model transformacji fal wezbraniowych w rzece – tzw. model fali dyfuzyjnej. Chociaż równania opisujące rzeczywiste sytuacje przedstawione w podrozdziale 7.1 są nieco bardziej skomplikowane niż (8.49), to zasady rozwiązywania numerycznego oraz problemy występujące w jego trakcie można wyjaśnić na przykładzie najprostszej wersji równania, jaką jest równanie (8.49) ze stałymi współczynnikami. Zakładamy więc, że U = const > 0 i D = const.

W ogólnym przypadku z rozwiązaniem równania adwekcji-dyfuzji mogą się wiązać trudności typowe dla równania czystej adwekcji, omówione w podrozdziale 5.4. Przyczyną trudności jest dwoista natura równania (8.49). Jest ono superpozycją procesu przenoszenia adwekcyjnego, które w równaniu reprezentuje człon $U\partial C/\partial x$, oraz przenoszenia dyfuzyjnego reprezentowanego przez człon $D\partial^2 C/\partial x^2$. Człon adwekcyjny ma naturę hiperboliczną, natomiast człon dyfuzyjny – paraboliczną. Trudności w trakcie numerycznego rozwiązywania wystąpią, gdy w procesie przenoszenia dominującą rolę będzie odgrywało przeno-

szenie adwekcyjne. W przypadku przeciwnym, przy dominacji przenoszenia dyfuzyjnego, rozwiązanie równania (8.49) nie nastręcza większych trudności. Można je otrzymać, stosując w zasadzie dowolny schemat metody różnic skończonych.

Z wymienionych wyżej powodów zasadnicze znaczenie ma ocena wzajemnej proporcji przenoszenia adwekcyjnego i dyfuzyjnego. W tym celu można wykorzystać tzw. liczbę Pecleta, która w przypadku rozwiązań na siatkach węzłów ma postać (Patankar, 1980)

$$P = \frac{U\Delta x}{D}, \tag{8.50}$$

gdzie: P – numeryczna liczba Pecleta,
Δx – wymiar przestrzenny siatki węzłów,
U – prędkość przepływu,
D – współczynnik dyfuzji.

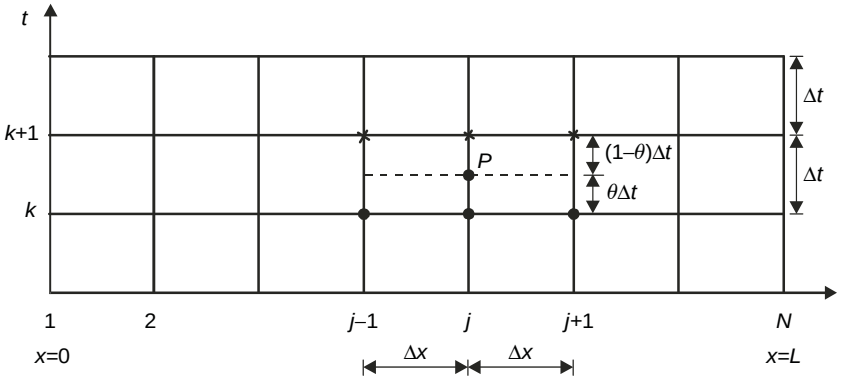
Liczbę tę można traktować jako stosunek adwekcyjnej liczby Couranta do dyfuzyjnej liczby Couranta. Ze względu na podobieństwo do znanej z mechaniki płynów liczby Reynoldsa, często nazywa się ją komórkową liczbą Reynoldsa (w języku angielskim: cell Reynolds number (Fletcher, 1991)).

Liczba Pecleta jest dodatnią liczbą z przedziału $\langle 0, \infty)$. Z definicji (8.50) wynika, że:
— $P = 0$ dla równania dyfuzji,
— $P = \infty$ dla równania adwekcji.

Numeryczne rozwiązanie równania (8.49) staje się kłopotliwe przy dużych liczbach Pecleta. Powód trudności i ich skutki omówiono w podrozdziale 7.4, przy okazji omawiania rozwiązywania równania adwekcji oraz związanych z tym zjawisk dyssypacji i dyspersji numerycznej.

Rozpatrzmy problem rozwiązania równania adwekcji-dyfuzji (8.49) w obszarze $0 \le x \le L$ i $t \ge 0$, gdzie L jest długością badanego odcinka kanału. Na granicach obszaru rozwiązania zadajemy warunki dodatkowe, stosowne dla równania typu parabolicznego. Obszar rozwiązania pokrywamy siatką węzłów o wymiarach $\Delta x \cdot \Delta t$, jak na rysunku 8.8. Równanie (8.49) aproksymujemy w punkcie P. Pochodną względem czasu zastępujemy ilorazem różnicowym w następujący sposób:

$$\left.\frac{\partial C}{\partial t}\right|_P \approx \frac{C_j^{k+1} - C_j^k}{\Delta t}. \tag{8.51}$$



Rys. 8.8. Siatka węzłów stosowana w schemacie różnicowym dla równania adwekcji-dyfuzji

Pochodne względem *x* aproksymujemy w punkcie *P*, interpolując liniowo ich aproksymacje nas poziomie czasu *k* oraz *k* + 1.

$$\left.\frac{\partial^2 C}{\partial x^2}\right|_P = \theta \left.\frac{\partial^2 C}{\partial x^2}\right|_j^{k+1} + (1-\theta)\left.\frac{\partial^2 C}{\partial x^2}\right|_j^{k}, \tag{8.52}$$

$$\left.\frac{\partial C}{\partial x}\right|_P = \theta \left.\frac{\partial C}{\partial x}\right|_j^{k+1} + (1-\theta)\left.\frac{\partial C}{\partial x}\right|_j^{k}, \tag{8.53}$$

gdzie θ jest parametrem wagowym z przedziału ⟨0, 1⟩.
Pochodną II rzędu aproksymujemy różnicowym ilorazem symetrycznym

$$\frac{\partial^2 C}{\partial x^2} \approx \frac{C_{j-1} - 2C_j + C_{j+1}}{\Delta x^2}. \tag{8.54}$$

Do aproksymacji pochodnej I rzędu zastosujemy szczególny sposób umożliwiający sterowanie sposobem aproksymacji (Kot i Szymkiewicz, 2002). Przyjmijmy, że aproksymacja pochodnej w węźle *j* będzie średnią ważoną z aproksymacji w tym węźle ilorazem różnicowym przednim i wstecznym. Ma ona następującą postać:

$$\left.\frac{\partial C}{\partial x}\right|_j = \sigma \frac{C_j - C_{j-1}}{\Delta x} + (1-\sigma)\frac{C_{j+1} - C_j}{\Delta x}, \tag{8.55}$$

gdzie σ jest parametrem wagowym z przedziału ⟨0, 1⟩. Powyższą formułę można przepisać w nieco zmienionej postaci

$$\left.\frac{\partial C}{\partial x}\right|_j = \frac{-\sigma C_{j-1} - (1-2\sigma)C_j + (1-\sigma)C_{j+1}}{\Delta x}. \tag{8.56}$$

Zauważmy, że jej szczególnymi przypadkami są znane ilorazy różnicowe aproksymujące pochodną I rzędu. Przyjmując $\sigma = 0$, otrzymujemy iloraz różnicowy przedni. Przy $\sigma = 1/2$ jest to iloraz różnicowy centralny, zaś przy $\sigma = 1$ – iloraz różnicowy wsteczny. Zaletą zastosowanego podejścia jest możliwość stosowania dowolnych wartości σ z przedziału ⟨0, 1⟩ i tym samym sterowanie dokładnością aproksymacji pochodnej I rzędu względem *x*.

Podstawiając kolejno (8.54) i (8.56) odpowiednio do (8.52) i (8.53), a następnie (8.51), (8.52) i (8.53) do równania adwekcji-dyfuzji (8.49), otrzymujemy następującą zależność:

$$\begin{aligned}\frac{C_j^{k+1} - C_j^k}{\Delta t} + U\Bigg[\theta\frac{-\sigma C_{j-1}^{k+1} - (1-2\sigma)C_j^{k+1} + (1-\sigma)C_{j+1}^{k+1}}{\Delta x} + \\ + (1-\theta)\frac{-\sigma C_{j-1}^{k} - (1-2\sigma)C_j^{k} + (1-\sigma)C_{j+1}^{k}}{\Delta x}\Bigg] + \\ - D\left[\theta\frac{C_{j-1}^{k+1} - 2C_j^{k+1} + C_{j+1}^{k+1}}{\Delta x^2} + (1-\theta)\frac{C_{j-1}^{k} - 2C_j^{k} + C_{j+1}^{k}}{\Delta x^2}\right] = 0\ .\end{aligned} \tag{8.57}$$

Ponieważ wartości funkcji C na poziomie czasu k są znane albo z warunku początkowego, albo z obliczeń w poprzednim kroku czasowym, niewiadomymi w powyższym równaniu są wartości funkcji C na poziomie $k + 1$. Wprowadzając znane z rozdziału 7 definicje adwekcyjnej i dyfuzyjnej liczby Couranta

$$C_a = \frac{U\Delta t}{\Delta x}, \quad C_d = \frac{D\Delta t}{\Delta x^2}, \tag{8.58, 8.59}$$

równanie (8.57) można uporządkować następująco:

$$\begin{aligned} &-\theta(\sigma C_a + C_d)C_{j-1}^{k+1} + [1 - \theta C_a(1-2\sigma) + 2\theta C_d]C_j^{k+1} + \\ &+ \theta[(1-\sigma)C_a - C_d]C_{j+1}^{k+1} = C_j^k - (1-\theta)C_a[-\sigma C_{j-1}^k - \\ &-(1-2\sigma)C_j^k + (1-\sigma)C_{j+1}^k] + (1-\theta)C_d[C_{j-1}^k - 2C_j^k + C_{j+1}^k]\,. \end{aligned} \tag{8.60}$$

Równanie typu (8.60) można napisać dla każdego węzła wewnętrznego, tzn. dla $j = 2, 3, ..., N - 1$ (rys. 8.8). Do otrzymanych w ten sposób $N - 2$ równań dodajemy 2 równania wynikające z zadanych warunków na brzegach $x = 0$ oraz $x = L$. W ten sposób powstaje zamknięty układ algebraicznych równań liniowych o wymiarze $N \times N$, który zapisujemy w postaci analogicznej do (8.12):

$$\mathbf{A}\,\mathbf{C} = \mathbf{W}, \tag{8.61}$$

gdzie: $\mathbf{A}$ – trójdiagonalna macierz, której elementy pod główną przekątną (a_j), na głównej przekątnej (b_j) i nad główną przekątną (c_j) są równe:

$b_1 = 1$, $c_1 = 0$,

$$\left.\begin{aligned} a_j &= -\theta(\sigma \cdot C_a + C_d), \\ b_j &= 1 - \theta C_a(1-2\sigma) + 2\theta C_d, \\ c_j &= \theta((1-\sigma)C_a - C_d), \end{aligned}\right\} \quad (j = 2, 3, \cdots, N-1),$$

$a_N = 0, \quad b_N = 1$,

$\mathbf{C} = (C_1^{k+1}, C_2^{k+1}, \cdots C_{N-1}^{k+1}, C_N^{k+1})^T$ – wektor niewiadomych utworzony z wartości węzłowych funkcji C na poziomie czasu $k + 1$,

$\mathbf{W} = (w_1, w_2, \ldots w_{N-1}, w_N)^T$ – wektor prawych stron, którego składowe, przy założeniu warunków typu Dirichleta na obu brzegach, są równe:

$$\begin{aligned} w_j &= C_j^k - (1-\theta)C_a[-\sigma C_{j-1}^k - (1-2\sigma)C_j^k + (1-\sigma)\,C_{j+1}^k] + \\ &\quad + (1-\theta)C_a[C_{j-1}^k - 2C_j^k + C_{j+1}^k] \quad (j = 2, 3, \ldots, N-1), \end{aligned}$$

$$w_1 = C_0(t_{k+1}), \quad w_N = C_L(t_{k+1}),$$

przy czym $C_0(t)$ i $C_L(t)$ są zadanymi funkcjami zależnymi od czasu, reprezentującymi zmiany wymuszane na brzegach.

Po rozwiązaniu układu (8.61) metodą Thomasa otrzymujemy przybliżone wartości funkcji C w węzłach na poziomie czasu $k + 1$, tzn. C_j^{k+1} ($j = 1, 2, ..., N$).

Zastosowany do rozwiązania równania adwekcji-dyfuzji (8.49) schemat numeryczny modyfikuje je. Mianowicie, wyrażając w równaniu (8.60) wartości węzłowe funkcji C przez jej rozwinięcia w szereg Taylora wokół węzła (j, $k + 1$), otrzymujemy zamiast (8.49) następujące równanie:

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} - D\frac{\partial^2 C}{\partial x^2} = D_n\frac{\partial^2 C}{\partial x^2} + E_n\frac{\partial^3 C}{\partial x^3} + \dots, \tag{8.62}$$

w którym współczynniki dyfuzji i dyspersji numerycznej D_n oraz E_n zdefiniowane są następująco:

$$D_n = \frac{U\Delta x}{2}[(2\theta - 1)C_a + (2\sigma - 1)], \tag{8.63}$$

$$E_n = \frac{U\Delta x^2}{2}\left[\left(\theta - \frac{2}{3}\right)C_a^2 - (1-\theta)(1-2\sigma)C_a - \frac{1}{3}\right]. \tag{8.64}$$

Z powyższych równań wynika ważna cecha zastosowanej metody rozwiązania, a mianowicie zgodność. Widzimy, że przy redukcji wymiarów siatki węzłów, gdy $\Delta x \to 0$ oraz $\Delta t \to 0$, równanie (8.62) zmierza do równania (8.49), gdyż $D_n \to 0$ oraz $E_n \to 0$.

Dodatkowe człony występujące po prawej stronie równania (6.62) są wynikiem aproksymacji pochodnych I rzędu w rozwiązywanym równaniu (8.49). Zauważmy, że przy $\theta = 1/2$ i $\sigma = 1/2$, tzn. gdy pochodne te aproksymujemy z dokładnością II rzędu (ilorazami różnicowymi centralnymi), zjawisko dyfuzji numerycznej nie wystąpi, gdyż $D_n = 0$. Dla innych wartości θ oraz σ schemat generuje dyfuzję numeryczną. Zatem w rozwiązaniu wystąpi zarówno efekt dyfuzji fizycznej obecnej w równaniu procesu adwekcji-dyfuzji (8.49), jak i dodatkowy efekt dyfuzji numerycznej wywołanej zastosowanym schematem rozwiązania. Ostatecznie otrzymane rozwiązanie będzie silniej spłaszczone niż wynika to z dyfuzji fizycznej zdefiniowanej współczynnikiem D, gdyż łączny efekt będzie wynikiem obu typów dyfuzji. Otrzymane rozwiązanie będzie odpowiadało rozwiązaniu równania adwekcji-dyfuzji ze współczynnikiem dyfuzji efektywnej

$$D_{ef} = D + D_n. \tag{8.65}$$

Z tego względu naturalne jest dążenie do eliminacji dyfuzji numerycznej w rozwiązaniu. Jeśli dyfuzja fizyczna jest znaczna, to przyjmując $\theta = 1/2$ i $\sigma = 1/2$, otrzymujemy schemat niedyssypatywny, który mimo to zapewni gładkie rozwiązanie. Jeśli natomiast dyfuzja fizyczna jest mała, w celu zapewnienia gładkiego rozwiązania konieczne jest wprowadzenie pewnej dyfuzji numerycznej. Można to zrobić, przyjmując odpowiednie dla danej sytuacji wartości parametrów wagowych θ i σ.

Analiza stabilności schematu wykonana metodą Neumanna pozwala wykazać, że schemat jest absolutnie stabilny, gdy spełnione są następujące warunki (Kot i Szymkiewicz, 2002):

$$\theta \geq \frac{1}{2}, \quad \sigma \geq \frac{1}{2}. \tag{8.66a,b}$$

Dla innych wartości parametrów schemat może być niestabilny lub warunkowo stabilny.

Zastosowany schemat można potraktować jako uogólnioną postać innych znanych schematów różnicowych. Otrzymujemy je, przyjmując szczególne wartości parametrów wagowych. Rozpatrzmy niektóre przypadki szczególne.

- $\theta = 0$, $\sigma = 1$.

Przyjęcie powyższych wartości pozwala zapisać równanie (8.60) w postaci

$$C_j^{k+1} = C_a \cdot C_{j-1}^k + (1 - C_a)C_j^k + C_d(C_{j-1}^k - 2C_j^k + C_{j+1}^k). \tag{8.67}$$

Jest to równanie odpowiadające schematowi jawnemu z aproksymacją „pod prąd" członu adwekcyjnego. Przyjmując $j = 2, 3, ..., N - 1$, obliczamy wartości na poziomie $k + 1$. Natomiast w węzłach skrajnych determinują je zadane warunki brzegowe: $C_1^{k+1} = C_0(t_{k+1})$ i $C_N^{k+1} = C_L(t_{k+1})$. Zauważmy, że w tym przypadku macierz **A** układu równań (8.61) stała się macierzą jednostkową. Pozwala to rozszczepić układ na niezależne równania typu (8.67) i rozwiązać je w dowolnej kolejności. Schemat jest warunkowo stabilny, a kryterium stabilności ma postać

$$\Delta t \le \min\left(\frac{\Delta x}{U}, \frac{\Delta x^2}{2D}\right). \tag{8.68}$$

Pierwszy wyraz prawej strony powyższej nierówności odnosi się do stabilności części adwekcyjnej ($C_a \le 1$), zaś drugi – do stabilności części dyfuzyjnej ($C_d \le 1/2$).

- $\theta = 1/2$, $\sigma = 1/2$

Równania układu (8.61) przyjmą postać

$$\begin{aligned} -\left(\frac{C_a}{4} + \frac{C_d}{2}\right)C_{j-1}^{k+1} + (1 + C_d)C_j^{k+1} + \left(\frac{C_a}{4} - \frac{C_d}{2}\right)C_{j+1}^{k+1} = \\ = C_j^k - \frac{C_a}{4}(C_{j+1}^k - C_{j-1}^k) + \frac{C_d}{2}(C_{j-1}^k - 2C_j^k + C_{j+1}^k) \\ (j = 2, 3, ..., N - 1)\,. \end{aligned} \tag{8.69}$$

Jest to doskonale znany i powszechnie stosowany schemat Cranka-Nicolsona. Fletcher (1991) wykazuje, że schemat ten zapewnia gładkie, bezoscylacyjne rozwiązanie, gdy iloczyn współczynników równania leżących poza główną przekątną macierzy jest nieujemny, tzn. gdy $a_j \cdot c_j \ge 0$. Warunek ten ma postać

$$-\left(\frac{C_a}{4} + \frac{C_d}{2}\right)\left(\frac{C_a}{4} - \frac{C_d}{2}\right) \ge 0\,. \tag{8.70}$$

Relację powyższą przekształcamy następująco:

$$\left(\frac{C_a}{2} + C_d\right)\left(\frac{C_a}{2} - C_d\right) \le 0\,. \tag{8.71}$$

Ponieważ zarówno adwekcyjna liczba Couranta C_a, jak i dyfuzyjna liczba Couranta są dodatnie, powyższa relacja będzie spełniona, gdy

$$\frac{C_a}{2} - C_d \le 0\,. \tag{8.72}$$

Jeśli weźmiemy pod uwagę definicję liczby Pecleta (8.50) oraz definicje obu liczb Couranta (8.58) i (8.59), z powyższej relacji otrzymamy warunek gładkiego rozwiązania równania adwekcji-dyfuzji

$$P \le 2\,. \tag{8.73}$$

Zatem w przypadku równania adwekcji-dyfuzji, jeśli tylko liczba Pecleta nie jest większa od 2, gładkie rozwiązanie zapewni nam schemat niedyssypatywny. Oznacza to, że dyfuzja fizyczna w równaniu (8.49) jest na tyle duża, iż wygładzi oscylacje wywołane dyspersyjnością spowodowaną aproksymacją pochodnych I rzędu.

Nawiązując do ogólnej postaci omawianego schematu rozwiązania równania adwekcji-dyfuzji, można stwierdzić, że w konkretnych zadaniach parametry wagowe θ i σ mogą przyjmować inne niż wymienione wyżej wartości. Jednak jeśli chcemy zapewnić absolutną stabilność schematu, ich wartości powinny spełniać następujące relacje:

$$1/2 \le \theta \le 1, \quad 1/2 \le \sigma \le 1. \tag{8.74a,b}$$

Zaletą zastosowanego schematu jest możliwość płynnego włączania, w miarę potrzeby, efektu „pod prąd". Ma to znaczenie szczególnie, gdy w poszukiwanym rozwiązaniu równania adwekcji-dyfuzji występują duże gradienty.

Podsumowując można stwierdzić, że rozwiązanie numeryczne równania adwekcji-dyfuzji z dominującą adwekcją może być kłopotliwe. Przyczyną są błędy dyfuzji i dyspersji numerycznej wynikające z aproksymacji pochodnych I rzędu. Z tego powodu, mimo że opracowano już szereg metod, w dalszym ciągu poszukuje się specjalnych technik rozwiązywania tego równania.

8.4. Rozwiązanie dwuwymiarowego równania filtracji nieustalonej

Zagadnienie dwuwymiarowego przepływu w gruncie rozpatrzymy na przykładzie równania (7.2) ze stałym współczynnikiem dyfuzji D

$$\frac{\partial h}{\partial t} = D\left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right) + w\,. \tag{8.75}$$

Ponieważ jest to równanie typu parabolicznego, problem jego rozwiązania formułujemy jako zagadnienie początkowo-brzegowe. Poszukujemy funkcji $h(x, y, t)$, która w obszarze rozwiązania spełni równanie (8.75) oraz warunek początkowy

$$h(x, y, t = 0) = h_p(x, y)$$

i następujące warunki brzegowe:

$$h(x_B, y_B, t) = h_B(t),$$

gdzie B oznacza brzeg obszaru.

Załóżmy dla uproszczenia zapisów, że w płaszczyźnie (x, y) rozważany obszar ma kształt prostokąta o wymiarach L_x i L_y i przyjmijmy ponadto w równaniu (8.75) $w = 0$. Rozwiązanie równania poszukuje się więc w obszarze

$$0 \le x \le L_x, \quad 0 \le y \le L_y, \quad t \ge 0.$$

Obszar ograniczony brzegiem B pokrywamy siatką węzłów o wymiarach $\Delta x \times \Delta y$. Będzie ona liczyła $N \times M$ węzłów (gdzie: N – liczba wierszy, M – liczba kolumn). Równanie to rozwiążemy, stosując kolejno następujące schematy:

a) schemat jawny,
b) schemat niejawny,
c) schemat naprzemiennych kierunków.

Ad a) ***Schemat jawny***

Aproksymację pochodnej względem czasu wykonujemy analogicznie, jak w wypadku równania jednowymiarowego. Pochodne II rzędu aproksymujemy na poziomie t. Ponieważ odległości Δx i Δy są stałe, dowolny węzeł siatki można zidentyfikować poprzez indeksy i

oraz j (rys. 8.9a). Podobnie indeks k będzie oznaczał poziom czasowy t. W efekcie otrzymujemy w węźle (i, j) równanie algebraiczne

$$\frac{h_{i,j}^{k+1} - h_{i,j}^{k}}{\Delta t} = D\left(\frac{h_{i,j+1}^{k} - 2h_{i,j}^{k} + h_{i,j-1}^{k}}{(\Delta x)^2} + \frac{h_{i+1,j}^{k} - 2h_{i,j}^{k} + h_{i-1,j}^{k}}{(\Delta y)^2} \right), \tag{8.76}$$

gdzie: i, j – indeksy węzła określające jego położenie względem osi x, y,
k – indeks poziomu czasowego.

W praktycznych obliczeniach bardzo często stosuje się siatkę kwadratową. Przyjmując jednakowy krok przestrzenny siatki, tj. $\Delta x = \Delta y = \Delta$, otrzymujemy

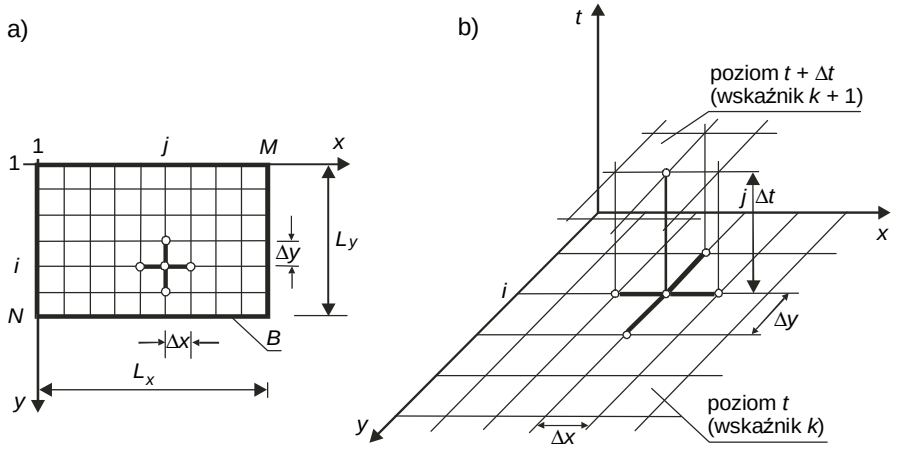
$$h_{i,j}^{k+1} = h_{i,j}^{k} + C_d \left(h_{i,j+1}^{k} + h_{i,j-1}^{k} + h_{i+1,j}^{k} + h_{i-1,j}^{k} - 4h_{i,j}^{k} \right), \tag{8.77}$$

gdzie

$$C_d = \frac{D\Delta t}{\Delta^2} \tag{8.78}$$

jest dyfuzyjną liczbą Couranta.
W ten sposób uzyskano formułę, która w jawny sposób pozwala obliczyć wartość poszukiwanej funkcji h w węźle (i, j) na poziomie $k + 1$, czyli w chwili $t + \Delta t$ (rys. 8.9b). Jak widać, formuła ta, bardzo prosta, obowiązuje we wszystkich węzłach wewnętrznych obszaru, tzn. dla $i = 2, 3, ..., N - 1$; $j = 2, 3, ..., M - 1$. W węzłach brzegowych wartości funkcji h określają zadane warunki brzegowe. Schemat ten, jak każdy schemat jawny, jest warunkowo stabilny. Dla zachowania stabilności musi być spełniona relacja $C_d \le 0{,}25$, z której wynika warunek (Fletcher, 1991)

$$\Delta t \le \frac{\Delta^2}{4D}. \tag{8.79}$$



Rys. 8.9. Jawny schemat aproksymacji różnicowej równania (8.75)

Ad b) ***Schemat niejawny***

W tym przypadku aproksymację pochodnych II rzędu w równaniu (8.75) wykonujemy na poziomie czasu $k + 1$, czyli w chwili $t + \Delta t$. Jest to więc poziom czasowy, na którym poszukujemy wartości funkcji h. Będziemy mieli odpowiednio (rys. 8.10b):

$$\left.\frac{\partial h}{\partial t}\right|_{i,j}^{k+1} \approx \frac{h_{i,j}^{k+1} - h_{i,j}^{k}}{\Delta t}, \tag{8.80}$$

$$\left.\frac{\partial^2 h}{\partial x^2}\right|_{i,j}^{k+1} \approx \frac{h_{i,j-1}^{k+1} - 2h_{i,j}^{k+1} + h_{i,j+1}^{k+1}}{\Delta x^2}, \tag{8.81}$$

$$\left.\frac{\partial^2 h}{\partial y^2}\right|_{i,j}^{k+1} \approx \frac{h_{i-1,j}^{k+1} - 2h_{i,j}^{k+1} + h_{i+1,j}^{k+1}}{\Delta y^2}. \tag{8.82}$$

Przy założeniu, że siatka węzłów jest kwadratowa, o wymiarach $\Delta x = \Delta y = \Delta$, po wstawieniu powyższych aproksymacji do równania (8.75) otrzymujemy

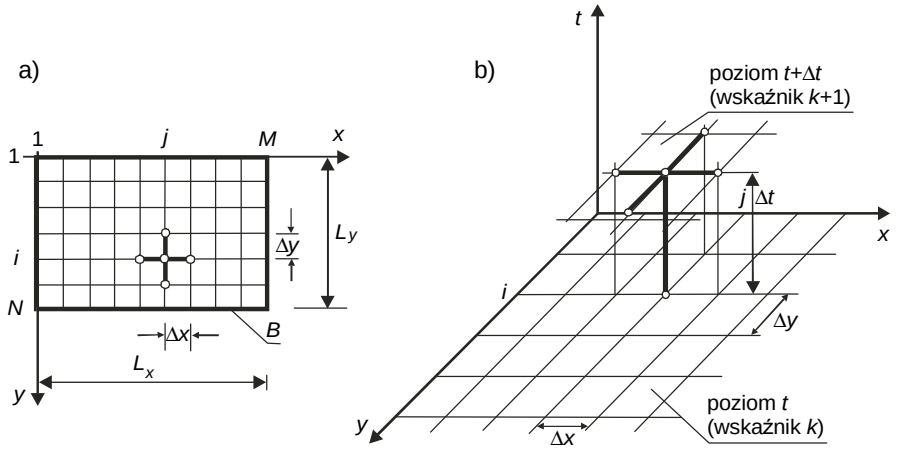
$$-h_{i,j-1}^{k+1} - h_{i,j+1}^{k+1} - h_{i-1,j}^{k+1} - h_{i+1,j}^{k+1} + (4+s)h_{i,j}^{k+1} = sh_{i,j}^{k}, \tag{8.83}$$

gdzie

$$s = \frac{\Delta^2}{D\Delta t} \tag{8.84}$$

jest odwrotnością dyfuzyjnej liczby Couranta (8.78).

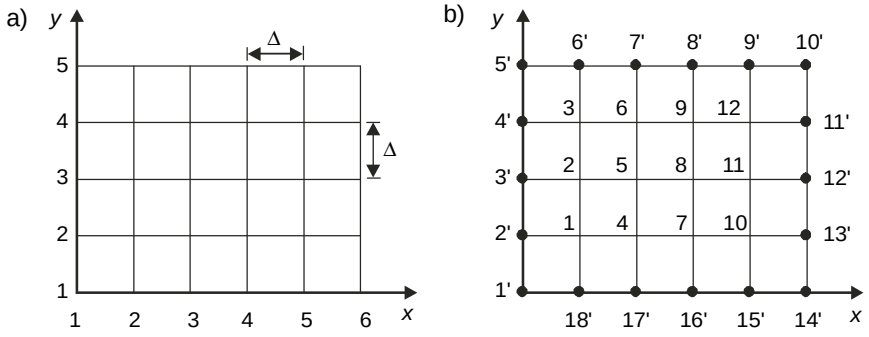
W równaniu (8.83) wartości z górnym indeksem $k + 1$ są nieznane. Jednak równania tego typu można napisać dla każdego węzła wewnętrznego. Utworzą one układ równań o liczbie równań równej liczbie niewiadomych.



Rys. 8.10. Niejawny schemat aproksymacji różnicowej równania (6.75)

Dla zilustrowania struktury układu utwórzmy go dla siatki węzłów przedstawionej na rys. 8.11a. Obszar ma kształt prostokąta, w którym występuje $N = 5$ wierszy i $M = 6$ ko-

lumn, co razem daje 30 węzłów. Wymiary oczek są równe $\Delta\times\Delta$. Załóżmy warunki brzegowe typu Dirichleta.



Rys. 8.11. Numeracja węzłów w przyjętej siatce

Oznacza to, że w każdym węźle na obwodzie obszaru znane są wartości funkcji *h*. Definiują je zadane warunki brzegowe. Dla ułatwienia konstrukcji układu zamieńmy dwuindeksowy system oznaczeń węzłów jednym indeksem. Określa on kolejny numer węzła w przyjętym sposobie numeracji. Numerację tę wykonamy wzdłuż pionowych linii, kontynuując ją od dołu w górę. Wydzielmy węzły brzegowe, oznaczając ich numery dodatkowym symbolem „prim" (rys. 8.11b). W węzłach tych, jak wcześniej wspomniano, wartość funkcji określają zadane warunki brzegowe. Wewnątrz obszaru występuje 12 węzłów, w których będziemy poszukiwali wartości *h*. Dla każdego z nich można napisać równanie (8.83). Równania utworzą układ, który można zapisać w postaci

$$\mathbf{A}\,\mathbf{h}^{k+1} = \mathbf{W}. \tag{8.85}$$

Uwzględniając strukturę $\mathbf{A}$, $\mathbf{h}^{k+1}$ oraz $\mathbf{W}$, układ ten zapisujemy następująco:

4+s	–1		–1									h_1^{k+1}		$sh_1^k + h_{2'} + h_{18'}$
–1	4+s	–1		–1								h_2^{k+1}		$sh_2^k + h_{3'}$
	–1	4+s			–1							h_3^{k+1}		$sh_3^k + h_{4'} + h_{6'}$
–1			4+s	–1		–1						h_4^{k+1}		$sh_4^k + h_{17'}$
	–1		–1	4+s	–1		–1					h_5^{k+1}		sh_5^k
		–1		–1	4+s			–1				h_6^{k+1}	* =	$sh_6^k + h_{7'}$
			–1			4+s	–1		–1			h_7^{k+1}		$sh_7^k + h_{16'}$
				–1		–1	4+s	–1		–1		h_8^{k+1}		sh_8^k
					–1		–1	4+s			–1	h_9^{k+1}		$sh_9^k + h_{8'}$
						–1			4+s	–1		h_{10}^{k+1}		$sh_{10}^k + h_{13'} + h_{15'}$
							–1		–1	4+s	–1	h_{11}^{k+1}		$sh_{11}^k + h_{12'}$
								–1		–1	4+s	h_{12}^{k+1}		$sh_{12}^k + h_{9'} + h_{11'}$

Jak widać, macierz współczynników układu jest macierzą pasmową rzadką. Szerokość pasma zależy od przyjętego sposobu numerowania węzłów. Numeracja wzdłuż linii transwersalnych do dłuższej osi daje szerokość minimalną. Ponieważ w praktycznych zagadnieniach układ ma duże rozmiary (rzędu 1000), do jego rozwiązania stosuje się iteracyjne metody rozwiązania. Zastosowany schemat jest absolutnie stabilny. Macierz współczynników ma dominującą główną przekątną. Wadą schematu jest konieczność rozwiązywania układu równań o bardzo dużych wymiarach.

Ad c) ***Metoda naprzemiennych kierunków***

W celu uniknięcia wady schematu jawnego, którą jest warunkowa stabilność, i niejawnego, którą są wielkie rozmiary rozwiązywanych w każdym kroku czasowym układów równań algebraicznych, Peaceman i Rachford (Potter, 1977) zaproponowali technikę rozwiązania dwuwymiarowego równania dyfuzji, zapewniającą absolutną stabilność numeryczną oraz minimalizującą wymiary powstających układów równań. Poprzednio przedstawiane schematy umożliwiały bezpośrednie przejście z poziomu czasu t na poziom $t + \Delta t$. Metoda naprzemiennych kierunków umożliwia przejście na poziom czasu $t + \Delta t$ po wykonaniu obliczeń na poziomie pośrednim $t + \Delta t/2$. Pomimo konieczności prowadzenia dwóch etapów obliczeń w celu przejścia z poziomu t na $t + \Delta t$, technika ta okazuje się najbardziej ekonomiczną pod względem kosztów obliczeń. Jest ona jednym z wariantów tzw. techniki dekompozycji, która umożliwia sprowadzenie rozwiązania równania wielowymiarowego do rozwiązania ciągu równań jednowymiarowych (Marczuk, 1983).

Metoda ta wykorzystuje znaną własność całkowania. Całka oznaczona w przedziale czasu równa jest sumie całek w podprzedziałach tworzących przedział. Scałkujmy równanie (8.75) w przedziale czasu $\langle t, t + \Delta t\rangle$. Otrzymamy ogólną zależność

$$h(x,y,t+\Delta t) - h(x,y,t) = \int_{t}^{t+\Delta t} D\left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right) dt \tag{8.86}$$

Wykorzystując wspomnianą wyżej własność całki, można napisać

$$h(x,y,t+\Delta t) = h(x,y,t) + \int_{t}^{t+\Delta t/2}\left(D\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right) dt + \int_{t+\Delta t/2}^{t+\Delta t}\left(D\frac{\partial^2 h}{\partial x^2} + D\frac{\partial^2 h}{\partial y^2}\right) dt \tag{8.87}$$

Oznaczając sumę dwóch pierwszych członów prawej strony powyższego równania symbolem $h(x, y, t + \Delta t/2)$, można zapisać dwa równania równoważne równaniu (8.86):

$$h(x,y,t+\Delta t/2) = h(x,y,t) + \int_{t}^{t+\Delta t/2}\left(D\frac{\partial^2 h}{\partial x^2} + D\frac{\partial^2 h}{\partial y^2}\right) dt \tag{8.88}$$

$$h(x,y,t+\Delta t) = h(x,y,t+\Delta t/2) + \int_{t+\Delta t/2}^{t+\Delta t}\left(D\frac{\partial^2 h}{\partial x^2} + D\frac{\partial^2 h}{\partial y^2}\right) dt \tag{8.89}$$

W konsekwencji przejście z poziomu czasowego t na poziom $t + \Delta t$ odbywa się dwuetapowo. Najpierw przechodzimy z poziomu t na $t + \Delta t/2$, wykorzystując równanie (8.88), a następnie z poziomu $t + \Delta t/2$ na poziom $t + \Delta t$, wykorzystując równanie (8.89).

Jeśli zastąpić zmienne ciągłe odpowiednimi indeksami, pochodne pod całką – ilorazami różnicowymi na poziomie $t + \Delta t/2$ względem x oraz na poziomie t względem y, oraz obliczyć całkę metodą prostokątów, równanie (8.88) w węźle (i, j) przyjmie postać

$$\frac{h_{i,j}^{k+\frac{1}{2}} - h_{i,j}^{k}}{\Delta t} = \frac{D}{2}\left(\frac{h_{i,j+1}^{k+\frac{1}{2}} - 2h_{i,j}^{k+\frac{1}{2}} + h_{i,j-1}^{k+\frac{1}{2}}}{\Delta x^2} + \frac{h_{i+1,j}^{k} - 2h_{i,j}^{k} + h_{i-1,j}^{k}}{\Delta y^2} \right), \tag{8.90}$$

przy czym wskaźnik $k + 1/2$ stosowany jest dla wartości obliczanych na poziomie $t + \Delta t/2$ (rys. 8.12). Jak widać, druga pochodna względem x została aproksymowana na poziomie $t + \Delta t/2$, natomiast druga pochodna względem y – na poziomie t. Przyjmując siatkę kwadratową $\Delta x = \Delta y = \Delta$ oraz oznaczając

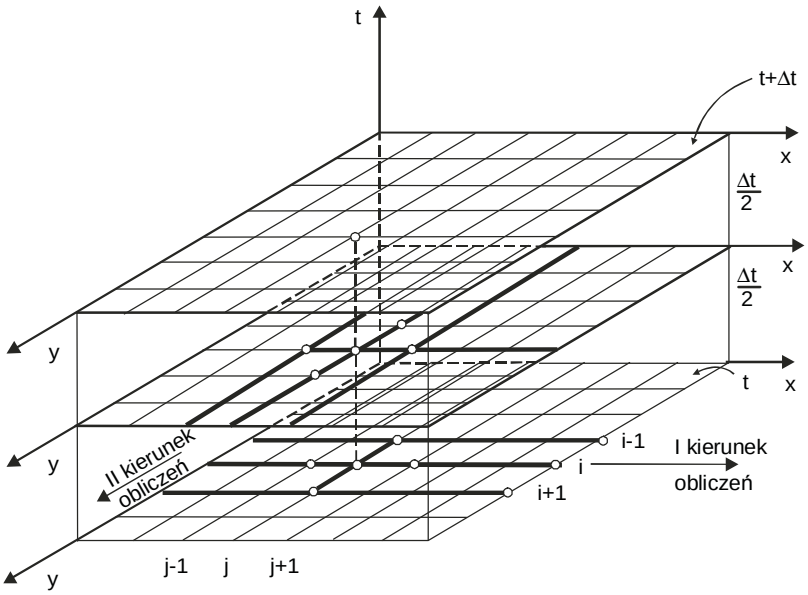
$$s = \frac{2\Delta^2}{D\Delta t}, \tag{8.91}$$

możemy równanie (8.90) przekształcić do postaci:

$$h_{i,j-1}^{k+\frac{1}{2}} - (2+s)h_{i,j}^{k+\frac{1}{2}} + h_{i,j+1}^{k+\frac{1}{2}} = P^k, \tag{8.92}$$

gdzie:

$$P^k = -h_{i+1,j}^{k} + (2-s)h_{i,j}^{k} - h_{i-1,j}^{k}. \tag{8.93}$$



Rys. 8.12. Przebieg obliczeń wg metody Peacemana i Rachforta

Powyższe równanie zawiera trzy niewiadome, którymi są wartości h w trzech kolejnych węzłach danego wiersza. Jeżeli wiersz o wskaźniku i składa się z M węzłów, wówczas możemy dla niego napisać $M - 2$ takich równań z $M - 2$ niewiadomymi. W węzłach $j = 1$

oraz $j = M$ wartości h są zadane. Układ ten rozwiązujemy metodą Thomasa, opisaną w podrozdziale 1.2. Postępując analogicznie we wszystkich wierszach, otrzymujemy wartości $h_{i,j}^{k+1/2}$ we wszystkich węzłach. Należy zaznaczyć, że wartości te nie mają fizycznego sensu. Są one jedynie wielkościami pomocniczymi, umożliwiającymi dokonanie następnego kroku, prowadzącego do znalezienia rozwiązania na poziomie $t + \Delta t$.

W drugim kroku wyjściowe równanie różniczkowe (8.89) aproksymujemy następująco:

$$\frac{h_{i,j}^{k+1} - h_{i,j}^{k+\frac{1}{2}}}{\Delta t} = \frac{D}{2}\left(\frac{h_{i,j+1}^{k+\frac{1}{2}} - 2h_{i,j}^{k+\frac{1}{2}} + h_{i,j-1}^{k+\frac{1}{2}}}{\Delta^2} + \frac{h_{i+1,j}^{k+1} - 2h_{i,j}^{k+1} + h_{i-1,j}^{k+1}}{\Delta^2} \right). \tag{8.94}$$

Postępując jak poprzednio, możemy powyższe równanie doprowadzić do postaci

$$h_{i-1,j}^{k+1} - (2+s)h_{i,j}^{k+1} + h_{i+1,j}^{k+1} = P^{k+\frac{1}{2}}, \tag{8.95}$$

gdzie:

$$P^{k+\frac{1}{2}} = -h_{i,j+1}^{k+\frac{1}{2}} + (2-s)h_{i,j}^{k+\frac{1}{2}} - h_{i,j-1}^{k+\frac{1}{2}}. \tag{8.96}$$

Ponieważ $1 \le i \le N$, to dla każdej kolumny można napisać $N - 2$ takich równań z $N - 2$ niewiadomymi oraz dodać równania wynikające z zadanych warunków brzegowych. Rozwiązując otrzymane układy dla wszystkich kolumn $j (1 < j < M)$, uzyskujemy poszukiwane rozwiązanie na poziomie $t + \Delta t$. W następnym kroku czasowym cały cykl powtarza się.

Opisana metoda, nazywana w skrócie ADI (od angielskiej nazwy *Alternating Direction Implicit method*), jest bezwarunkowo stabilna, czyli nie ma ograniczenia na wartość s. W praktyce oznacza to, że możemy stosować dowolnie duży krok czasowy Δt. Metoda jest ekonomiczna, chociaż przejście z poziomu t na poziom $t + \Delta t$ odbywa się w dwóch etapach. Jednak w każdym etapie rozwiązuje się sekwencje zagadnień jednowymiarowych, których dyskretyzacja prowadzi do układów algebraicznych równań liniowych z trójdiagonalnymi macierzami współczynników. Taki sposób postępowania jest zdecydowanie bardziej ekonomiczny niż rozwiązanie schematem całkowicie niejawnym.

Załóżmy, że równanie (8.75) rozwiązujemy w obszarze prostokątnym pokrytym siatką węzłów o wymiarach $N \times M$. Wiadomo, że czas rozwiązania układu równań z pasmową macierzą współczynników jest proporcjonalny do liczby równań i kwadratu szerokości pasma. Po zastosowaniu zatem schematu niejawnego (8.83), rozwiązanie układu równań (8.85) w każdym kroku czasowym będzie wymagało czasu

$$T_1 \sim (N \times M) \cdot n^2, \tag{8.97}$$

gdzie: N – liczba wierszy,
M – liczba kolumn,
n – szerokość pasma.

Z kolei wykonanie obliczeń metodą naprzemiennych kierunków wymaga na każdym etapie rozwiązywania odpowiednio N oraz M układów z macierzami trójdiagonalnymi, czyli o szerokości pasma $n = 3$. Czas obliczeń wyniesie więc

$$T_2 \sim N \cdot (M \cdot 3^2) + M \cdot (N \cdot 3^2) = 18 \cdot N \cdot M. \tag{8.98}$$

Stosunek czasów obliczeń wykonywanych w jednym kroku czasowym wymienionymi metodami będzie w rezultacie równy

$$\frac{T_2}{T_1} = \frac{18 \cdot N \cdot M}{N \cdot M \cdot n^2} = \frac{18}{n^2}. \tag{8.99}$$

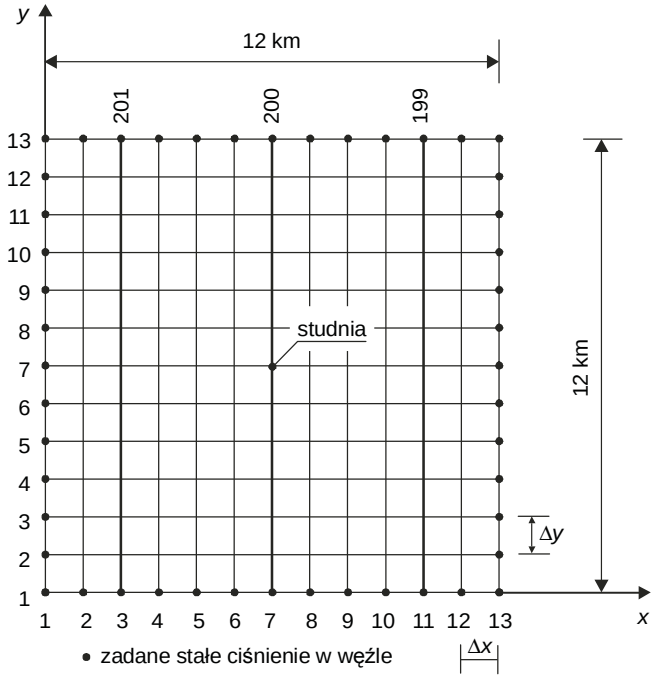
Jeśli weźmiemy pod uwagę fakt, że szerokość pasma n macierzy współczynników otrzymanej w schemacie niejawnym w zadaniach praktycznych może być rzędu 100, to wyrażenie (8.99) dowodzi radykalnego skrócenia czasu obliczeń dzięki zastosowaniu metody ADI.

Algorytm metody naprzemiennych kierunków jest bardzo łatwy w realizacji komputerowej, gdy obszar rozwiązania jest regularny, najlepiej prostokątny. W przypadku obszaru o dużej nieregularności zaprogramowanie algorytmu obliczeń jest kłopotliwe.

■ *Przykład 8.3*

Nieustalona filtracja w warstwie artezyjskiej

Rozpatrzmy przepływ wody w jednorodnej warstwie artezyjskiej o ograniczonym zasięgu. W widoku z góry obszar rozwiązania ma kształt kwadratu o wymiarach 13×13 km. Obszar ten przedstawiono na rys. 8.3.1, zaś przekrój pionowy na rys. 8.3.2.



Rys. 8.3.1. Obszar filtracji, jego dyskretyzacja oraz początkowy rozkład ciśnienia w warstwie artezyjskiej

Jak wiadomo z podrozdziału 7.1, przepływ wody w takich warunkach opisany jest równaniem (7.3)

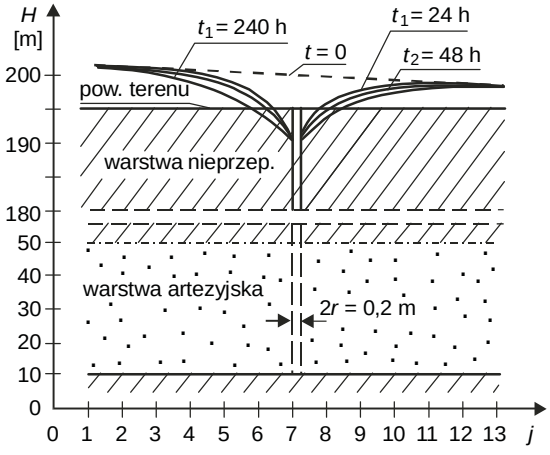
$$s\frac{\partial h}{\partial t}=\frac{\partial}{\partial x}\left(T\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(T\frac{\partial h}{\partial y}\right)+w\,, \tag{8.3.1}$$

gdzie: $h = h(x, y, t)$ – rzędna linii ciśnień,

$T = T(x, y) = k \cdot m$ – przewodność warstwy równa iloczynowi współczynnika filtracji $k(x, y)$ i miąższości warstwy $m(x, y)$,

$s = s(x, y)$ – współczynnik zasobności sprężystej,

$w = w(x, y, t)$ – zasilanie zewnętrzne warstwy.



Rys. 8.3.2. Przekrój pionowy przez warstwę artezyjską i obliczone chwilowe układy powierzchni reprezentujących ciśnienie piezometryczne

Dla uproszczenia problemu załóżmy, że warstwa artezyjska ma stałą miąższość m, jest jednorodna – co oznacza stałą wartość współczynnika filtracji k oraz współczynnika zasobności sprężystej s – oraz że nie istnieje zasilanie zewnętrzne, czyli $w = 0$. Powyższe założenia pozwalają zapisać równanie (8.3.1) w następującej prostszej formie:

$$\frac{\partial h}{\partial t}=D\left(\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}\right), \tag{8.3.2}$$

gdzie

$$D=\frac{k\cdot m}{s}=\text{const}$$

jest współczynnikiem dyfuzji.

Jest to znane z podrozdziału 8.4 dwuwymiarowe liniowe równanie dyfuzji. Przyjmijmy następujące warunki początkowo-brzegowe:

— warunek początkowy: $h(x, y, t = 0) = h_p(x, y)$,

— warunki brzegowe: $h(x_B, y_B, t) = h_B(t)$.

Do rozwiązania (8.3.2) zastosujmy opisany wcześniej schemat jawny. W tym celu obszar rozwiązania pokrywamy siatką węzłów jak na rys. 8.3.1 o wymiarach $\Delta x \times \Delta y$. Pochodną względem czasu aproksymujemy ilorazem różnicowym przednim, a pochodne względem x i y – ilorazami różnicowymi centralnymi. W efekcie otrzymujemy równanie (8.76)

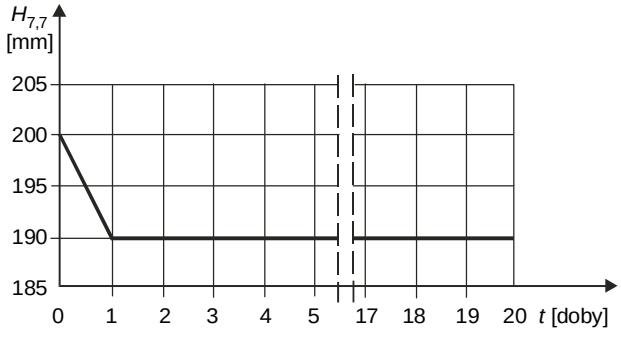
$$\frac{h_{i,j}^{k+1} - h_{i,j}^{k}}{\Delta t} = D\left(\frac{h_{i,j+1}^{k} - 2h_{i,j}^{k} + h_{i,j-1}^{k}}{(\Delta x)^2} + \frac{h_{i+1,j}^{k} - 2h_{i,j}^{k} + h_{i-1,j}^{k}}{(\Delta y)^2}\right). \tag{8.3.3}$$

W powyższym równaniu różnicowym, będącym aproksymacją wyjściowego równania różniczkowego, występuje tylko jedna niewiadoma $h_{i,j}^{k+1}$, gdyż wartości funkcji h z poziomu czasowego k są znane. Można ją obliczyć według wzoru

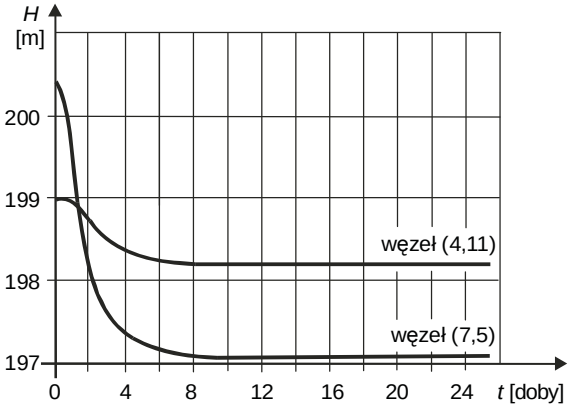
$$h_{i,j}^{k+1} = h_{i,j}^{k} + D \cdot \Delta t\left(\frac{h_{i,j+1}^{k} - 2h_{i,j}^{k} + h_{i,j-1}^{k}}{(\Delta x)^2} + \frac{h_{i+1,j}^{k} - 2h_{i,j}^{k} + h_{i-1,j}^{k}}{(\Delta y)^2}\right). \tag{8.3.4}$$

W chwili początkowej $t = 0$ węzłowe wartości h określa przyjęty warunek początkowy. Równanie różnicowe (8.3.4) umożliwia obliczenie h we wszystkich węzłach wewnętrznych w chwili $t = \Delta t$. Natomiast na brzegu B ograniczającym obszar, funkcja h jest znana z warunku brzegowego.

Obliczenia przepływu wody wykonano dla warstwy artezyjskiej przedstawionej na rys. 8.3.1 i 8.3.2. Przyjęty obszar filtracji pokryto siatką węzłów o wymiarach $\Delta x \times \Delta y = 100 \times 100$ m. Warunek początkowy wynika z przyjętego początkowego rozkładu ciśnień w warstwie w chwili $t = 0$. Zaznaczono go na rys. 8.3.1 i 8.3.2. Na brzegach obszaru przyjęto, że ciśnienia nie zmieniają się w czasie i pozostają takie, jak w chwili początkowej. W węźle (7, 7) istnieje studnia, która rozpoczynając pracę w chwili $t = 0$, w ciągu 24 godzin wytwarza depresję o wartości 10 m, a następnie utrzymuje ją na stałym poziomie (rys. 8.3.3). Przyjęta stała miąższość warstwy artezyjskiej wynosi $m = 50$ m (rys. 8.3.1), a jej własności filtracyjne definiuje współczynnik Darcy'ego $k = 0{,}3$ m/h. Ponadto przyjęto, że współczynnik zasobności sprężystej warstwy wynosi $s = 10^{-4}$. W węzłach leżących na brzegu lewym ($x = 0$) ciśnienie jest jednakowe i wynosi 201,5 m npp. Natomiast w węzłach leżących na brzegu prawym ($x = 12$ km) jest ono równe 198,5 m npp. Pomiędzy wymienionymi brzegami ciśnienie zmienia się liniowo, co ilustrują zaznaczone na rys. 8.3.1 izolinie: 201 m, 200 m, 199 m. Przyjęty krok całkowania w czasie spełniał kryterium stabilności (8.79). Otrzymane w wyniku obliczeń rezultaty przedstawiono na rys. 8.3.1 i 8.3.4. Na pierwszym z nich zilustrowano ewolucję w czasie leja depresji, zaznaczając położenie powierzchni $h(x, y, t)$. Z kolei na rys. 8.1.4 przedstawiono ewolucję ciśnienia w czasie w dwóch wybranych węzłach obszaru, a mianowicie w węźle (7, 5) i (4, 11).



Rys. 8.3.3. Wymuszone zmiany poziomu zwierciadła wody w studni zlokalizowanej w węźle (7, 7)

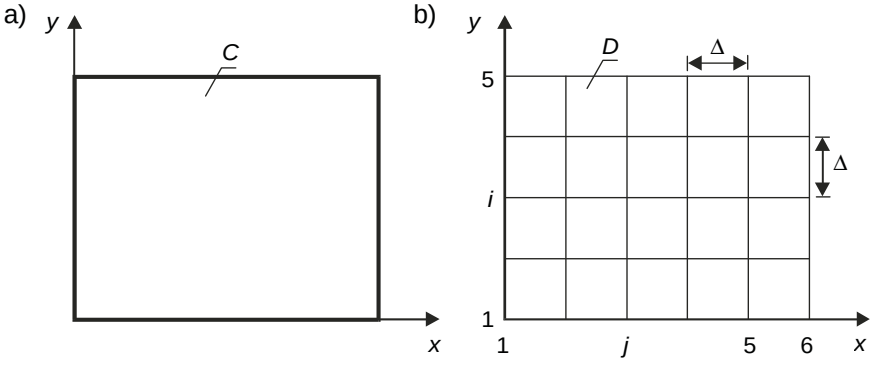


Rys. 8.3.4. Ewolucja ciśnienia w wybranych węzłach obszaru filtracji ■

8.5. Rozwiązanie dwuwymiarowego równania filtracji ustalonej pod ciśnieniem

Ze względu na liczne zastosowania, równanie Laplace'a odgrywa szczególną rolę w inżynierii wodnej, hydromechanice i hydrogeologii. Typowym zagadnieniem opisywanym tym równaniem jest filtracja pod budowlą piętrzącą (rys. 7.3). Na tym przykładzie omówimy metodę numerycznego rozwiązywania równania Laplace'a. Rozważać będziemy zagadnienie dwuwymiarowe opisywane równaniem (7.5)

$$\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}=0 \quad . \tag{8.100}$$



Rys. 8.13. Obszar ciągły i dyskretny, w którym poszukuje się rozwiązania równania filtracji ustalonej

Podobnie jak to uczyniliśmy poprzednio dla równania filtracji nieustalonej, również i teraz ciągły obszar C zastąpimy obszarem dyskretnym D, złożonym z węzłów siatki, jaką pokryty został obszar C. Założymy ponadto dla uproszczenia, że obszar jest prostokątny,

a siatka jest kwadratowa o boku $\Delta x = \Delta y = \Delta$ (rys. 8.13b). Rozpatrzmy zagadnienie brzegowe Dirichleta, tzn. przypadek rozwiązania równania (8.100) w obszarze, na którego brzegu zadane są wartości funkcji h (patrz podrozdział 7.1).

W każdym węźle siatki (rys. 8.13b) równanie (8.100) zastępujemy równaniem algebraicznym. W typowym węźle wewnętrznym (i, j), mającym otoczenie innych węzłów, pochodne II rzędu aproksymujemy, stosując różnicowe ilorazy symetryczne

$$\left.\frac{\partial^2 h}{\partial x^2}\right|_{i,j} \approx \frac{h_{i,j-1} - 2h_{i,j} + h_{i,j+1}}{\Delta^2}, \tag{8.101}$$

$$\left.\frac{\partial^2 h}{\partial y^2}\right|_{i,j} \approx \frac{h_{i-1,j} - 2h_{i,j} + h_{i+1,j}}{\Delta^2}. \tag{8.102}$$

Po podstawieniu do równania Laplace'a (8.100) otrzymujemy

$$h_{i+1,j} + h_{i,j+1} + h_{i-1,j} + h_{i,j-1} - 4h_{i,j} = 0. \tag{8.103}$$

Jest to równanie algebraiczne aproksymujące równanie Laplace'a w typowym węźle mającym pełne czteropunktowe otoczenie. Zauważmy, że przy założeniu warunków brzegowych typu Dirichleta w zadaniu wystąpią tylko takie węzły. W węzłach brzegowych zadane są wartości funkcji, a zatem nie ma potrzeby wykonywania w nich jakiejkolwiek aproksymacji. Zatem równania typu (8.103) można napisać dla każdego węzła wewnętrznego. Jeśli prostokątny obszar dyskretny liczy N wierszy i M kolumn, to otrzymamy układ równań o wymiarze $(N - 2) \times (M - 2)$, w którym liczba niewiadomych wartości węzłowych funkcji h równa jest liczbie równań. W zadaniach praktycznych układ ten zawiera setki, a nawet tysiące równań i niewiadomych. Ponieważ w każdym równaniu występuje tylko 5 niewiadomych, cechą charakterystyczną układu jest rzadka macierz współczynników. Jest to również macierz pasmowa o szerokości pasma zależnej od sposobu numerowania węzłów. Z tego powodu do rozwiązania układu równań, powstającego w trakcie rozwiązywania równania Laplace'a, niecelowe jest stosowanie metod dokładnych. Zdecydowanie skuteczniejsze będą metody iteracyjne.

Dla lepszej ilustracji problemu zbudujmy taki układ równań. W tym celu załóżmy obszar rozwiązania, jak na rys. 8.11. Składa się on z $N = 5$ wierszy i $M = 6$ kolumn. Odległości między węzłami są równe $\Delta x = \Delta y = \Delta$. W węzłach leżących na brzegu i oznaczonych numerami z indeksem „prim" zadane są wartości funkcji h. Należy znaleźć przybliżone wartości h w węzłach wewnętrznych, ponumerowanych jak na rys. 8.11b, od 1 do 12. Jak wspomniano wcześniej, dla każdego węzła wewnętrznego piszemy równanie typu (8.103). W rezultacie otrzymujemy układ 12 równań algebraicznych zawierających 12 niewiadomych. Układ ten zapisany w notacji macierzowej ma postać:

$$\mathbf{A}\,\mathbf{h} = \mathbf{W} \tag{8.104}$$

gdzie: **A** – macierz współczynników układu,
h – wektor niewiadomych utworzony z węzłowych wartości funkcji h,
W – wektor wyrazów wolnych.

Uwzględniając strukturę elementów układu wynikającą z równania (8.103), układ (8.104) można zapisać następująco:

–4	1		1									h_1		$-h_{2'} - h_{18'}$
1	–4	1		1								h_2		$-h_{3'}$
	1	–4			1							h_3		$-h_{4'} - h_{6'}$
1			–4	1		1						h_4		$-h_{17'}$
	1		1	–4	1		1					h_5		0
		1		1	–4			1				h_6	=	$-h_{7'}$
			1			–4	1		1			h_7		$-h_{16'}$
				1		1	–4	1		1		h_8		0
					1		1	–4			1	h_9		$-h_{8'}$
						1			–4	1		h_{10}		$-h_{13'} - h_{15'}$
							1		1	–4	1	h_{11}		$-h_{12'}$
								1		1	–4	h_{12}		$-h_{9'} - h_{11'}$

Jak widzimy, macierz układu rzeczywiście jest stała, rzadka i pasmowa. Ponieważ na jej głównej przekątnej leżą elementy niezerowe, z kolejnych równań układu można wyznaczyć kolejne niewiadome. Prowadzi to do znanej z podrozdziału 1.3 formuły

$$h_l = \frac{-\sum_{k=1}^{l-1} a_{lk} \cdot h_k - \sum_{k=l+1}^{K} a_{lk} \cdot h_k + w_l}{a_{ll}} \quad (l = 1, 2, \cdots, K), \tag{8.105}$$

gdzie: $K = (N-2) \times (M-2)$ – wymiar układu równań,
$a_{l\,k}$ – element macierzy współczynników **A**,
h_l – element wektora niewiadomych **h**,
w_l – element wektora prawych stron **W**.

Jak wiadomo również z podrozdziału 1.3, najprostszy proces iteracyjny (metoda Jacobiego – równanie (1.49)) ma postać

$$h_l^{(m+1)} = \frac{-\sum_{k=1}^{l-1} a_{lk} h_k^{(m)} - \sum_{k=1+1}^{K} a_{lk} h_k^{(m)} + b_l}{a_{ll}}, \quad (l = 1, 2, \ldots, K), \tag{8.106}$$

gdzie: m – indeks iteracji.

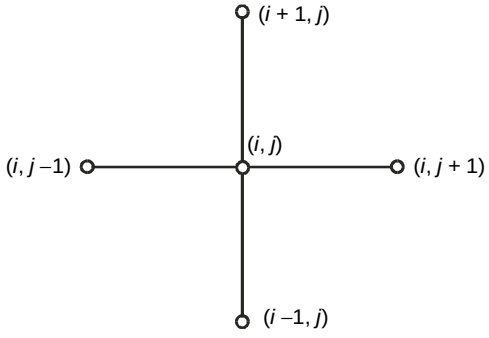
Typowe równanie układu (8.104) dla węzła przedstawionego na rys. 8.14 ma postać (8.103).

Wprowadźmy pojedyncze indeksy niewiadomych zgodnie z oznaczeniami przyjętymi na rysunku (8.11b). Równanie (8.103) dla węzła 8, czyli 8 równanie układu (8.104), ma postać:

$$h_5 + h_7 - 4h_8 + h_9 + h_{11} = 0, \tag{8.107}$$

zaś formuła Jacobiego (8.106) dla $l = 8$ będzie następująca:

$$h_8^{(m+1)} = \frac{1}{4}\left(h_5^{(m)} + h_7^{(m)} + h_9^{(m)} + h_{11}^{(m)}\right). \tag{8.108}$$



Rys. 8.14. Typowy węzeł wewnętrzny

Jeśli teraz ponownie powrócimy do dwuindeksowego oznaczenia wartości funkcji h w węzłach jak na rys. 8.14 oraz na rysunku 8.11a, to powyższe równanie zapiszemy następująco:

$$h_{i,j}^{(m+1)} = \frac{1}{4}\left(h_{i,j-1}^{(m)} + h_{i-1,j}^{(m)} + h_{i,j+1}^{(m)} + h_{i+1,j}^{(m)}\right). \tag{8.109}$$

Jest ona ważna dla $i = 2, 3, \ldots N - 1$; $j = 2, 3, \ldots, M - 1$.

Jak widać, w przypadku równania Laplace'a rozwiązywanego metodą różnic skończonych, rozwiązanie finalnego układu równań algebraicznych (8.104) iteracyjną metodą Jacobiego sprowadza się do bardzo prostego algorytmu obliczeń. Wystarczy w dowolnej kolejności przemierzyć obszar rozwiązania i według formuły (8.109) policzyć nowe przybliżenie poszukiwanych wartości funkcji h we wszystkich węzłach wewnętrznych.

Do rozwiązania układu równań (8.104) można zastosować również metodę Gaussa-Seidela (1.51). Sposób przemieszczania się po siatce węzłów w metodzie Jacobiego, czyli kolejność obliczania nowego przybliżenia, jest dowolny. Jeśli jednak przemieszczanie się będzie systematyczne, na przykład kolejnymi wierszami lub kolumnami, to jak łatwo można zauważyć, przystępując do obliczeń w węźle (i, j), w węzłach $(i, j - 1)$ oraz $(i - 1, j)$ będą już znane nowe przybliżenia funkcji h. Mogą one być wykorzystane do obliczeń. Formuła (8.109), uwzględniająca ten fakt, przyjmie postać

$$h_{i,j}^{(m+1)} = \frac{1}{4}\left(h_{i,j-1}^{(m+1)} + h_{i-1,j}^{(m+1)} + h_{i+1,j}^{(m)} + h_{i,j+1}^{(m)}\right). \tag{8.110}$$

Jest to znana metoda Gaussa-Seidela. Algorytm obliczeń jest tutaj równie prosty, jak w metodzie Jacobiego.

Ponieważ macierz współczynników układu (8.104) ma dominującą główną przekątną, zarówno metoda Jacobiego, jak i metoda Gaussa-Seidela są bezwzględnie zbieżne niezależnie od przyjętego przybliżenia początkowego ($m = 0$). Wadą obu metod jest wolna zbieżność procesu iteracyjnego.

Metodą najszybciej zbieżną jest metoda nadrelaksacji (1.55)

$$h_{i,j}^{(m+1)} = h_{i,j}^{(m)} + \frac{\omega}{4} \cdot r_{i,j}^{(m)}, \tag{8.111}$$

gdzie

$$r_{i,j}^{(m)} = h_{i,j-1}^{(m+1)} + h_{i-1,j}^{(m+1)} + h_{i+1,j}^{(m)} + h_{i,j+1}^{(m)} - 4h_{i,j}^{(m)}. \tag{8.112}$$

Podstawiając (8.112) do (8.111), otrzymuje się

$$h_{i,j}^{(m+1)} = h_{i,j}^{(m)} + \omega\frac{1}{4}\left(h_{i,j-1}^{(m+1)} + h_{i-1,j}^{(m+1)} + h_{i+1,j}^{(m)} + h_{i,j+1}^{(m)}\right) - \omega h_{i,j}^{(m)}, \tag{8.113}$$

lub inaczej zapisując

$$h_{i,j}^{(m+1)} = (1-\omega)h_{i,j}^{(m)} + \omega\bar{h}_{i,j}^{(m+1)}, \tag{8.114}$$

gdzie:

$$\bar{h}_{i,j}^{(m+1)} = \frac{1}{4}\left(h_{i,j-1}^{(m+1)} + h_{i-1,j}^{(m+1)} + h_{i+1,j}^{(m)} + h_{i,j+1}^{(m)}\right). \tag{8.115}$$

Współczynnik nadrelaksacji ω zawiera się w przedziale $\langle 1, 2)$. Dla $\omega = 1$ metoda nadrelaksacji staje się metodą Gaussa-Seidela. Dla $\omega = 2$ proces iteracyjny jest niezbieżny. Optymalną wartością ω jest taka wartość, dla której proces iteracyjny ma najszybszą zbieżność, tzn. taki, który prowadzi do rozwiązania z zadaną dokładnością w najmniejszej liczbie iteracji. Istnieje szereg propozycji oszacowania optymalnej wartości ω. Na przykład, wg Younga dla zagadnienia Dirichleta w prostokątnym obszarze o M kolumnach i N wierszach ω określa się następująco (Remson, Hornberger i Molz, 1971):

$$\omega = 1 + \left(\frac{\mu}{1+\sqrt{1+\mu^2}}\right)^2, \tag{8.116}$$

gdzie

$$\mu = \frac{1}{2}\left(\cos\frac{\pi}{M-1} + \cos\frac{\pi}{N-1}\right). \tag{8.117}$$

Mimo że wzór (8.116) dotyczy tylko obszaru prostokątnego i warunków Dirichleta, stosuje się go czasem do oszacowania wartości ω dla innych przypadków. Inny sposób oceny współczynnika ω podał Moller

$$\omega = \frac{2}{1+1.701\frac{\Delta}{r}}, \tag{8.118}$$

gdzie Δ jest krokiem siatki, natomiast r oznacza promień koła o powierzchni równej powierzchni obszaru S, w którym poszukuje się rozwiązania, czyli $r = (S/\pi)^{1/2}$

W kolei Piwecki i Sokólski (1966) podają wzór określony na drodze eksperymentu numerycznego

$$W = \frac{4}{\omega} = \frac{3{,}28}{\sqrt{MN}}\left(\frac{M}{N}\right)^q, \tag{8.119}$$

gdzie: $q = 0{,}3\, M\, N^{0,1}$,
M – ilość wierszy,
N – ilość kolumn.

Ponieważ omówione metody charakteryzują się bezwzględną zbieżnością, prowadzą one zawsze do rozwiązania, niezależnie od początkowego przybliżenia wektora niewiadomych. Jest sprawą oczywistą, że tym szybciej uzyskamy rozwiązanie równania Laplace'a, im będzie ono lepsze. Wybierając pierwsze przybliżenie rozwiązania, warto pamiętać o zasadzie ekstremum, którą spełnia rozwiązanie równania Laplace'a (Fletcher, 1991). Zasada ta mówi, że funkcja spełniająca w dowolnym obszarze *C* równanie Laplace'a w żadnym jego punkcie nie może przyjąć wartości większej od największej wartości zadanej na brzegu i wartości mniejszej od najmniejszej wartości zadanej na brzegu. Oznacza to, że dobrym przybliżeniem rozwiązania są wartości funkcji *h* zawarte pomiędzy zadaną na brzegu maksymalną i minimalną wartością tej funkcji.

Innym istotnym problemem związanym z rozwiązywaniem iteracyjnym układu równań jest test końca obliczeń. O zakończeniu procesu iteracyjnego decyduje założona dokładność rozwiązania. Jak wiadomo z podrozdziału 1.3, jeden z możliwych testów zakończenia obliczeń polega na porównywaniu wartości funkcji w kolejnych przybliżeniach we wszystkich węzłach. Test ten ma postać

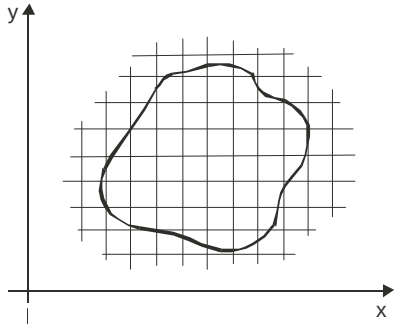
$$\left|h_{i,j}^{(m+1)} - h_{i,j}^{(m)}\right| \le \varepsilon \quad \text{dla} \quad i = 1, 2, 3, ..., N, j = 1, 2, 3, ..., M. \tag{8.120}$$

gdzie: $\varepsilon > 0$ jest przyjętą dokładnością rozwiązania.

Wadą tego testu jest to, że przy wolnej zbieżności różnice między wartościami funkcji obliczonymi w kolejnych iteracjach mogą być małe, mimo że rozwiązanie może być odległe od rozwiązania dokładnego. Innym testem jest badanie wartości residuum kolejnych równań:

$$r_{i,j} = \left|h_{i,j-1}^{(m+1)} + h_{i-1,j}^{(m+1)} - 4h_{i,j}^{(m+1)} + h_{i+1,j}^{(m+1)} + h_{i,j+1}^{(m+1)}\right| \le \varepsilon' , \tag{8.121}$$

gdzie: $\varepsilon' > 0$ jest przyjętą dokładnością rozwiązania.



Rys. 8.15. Dyskretyzacja nieregularnego obszaru całkowania

Opisany wyżej sposób postępowania dotyczył rozwiązania równania Laplace'a w najprostszym przypadku, tzn. w przypadku obszaru prostokątnego z warunkami brzegowymi Dirichleta. Rozwiązując zagadnienia praktyczne, zwykle mamy do czynienia z obszarem o nieregularnym kształcie. Pokrycie go prostokątną lub kwadratową siatką węzłów powoduje, że brzeg obszaru nie pokrywa się z bokami oczek siatki węzłów. Typową sytuacją będzie więc położenie brzegu pomiędzy węzłami, jak na rysunku 8.15.

Rozważmy sytuację pokazaną na rys. 8.16. Brzeg przecina ramiona siatki w punktach A i B, w których zadane są wartości ciśnienia *h*. Rozwijając funkcję *h* w szereg Taylora w otoczeniu węzła (*i*, *j*) wzdłuż kierunku osi *y* możemy napisać, pomijając wyrazy z pochodnymi rzędu wyższego niż drugi

$$h_A \approx h_{i,j} + \frac{\xi\Delta}{1!}\frac{\partial h}{\partial y} + \frac{(\xi\Delta)^2}{2!}\frac{\partial^2 h}{\partial y^2} , \tag{8.122}$$

oraz

$$h_{i-1,j} \approx h_{i,j} + \frac{\Delta}{1!}\frac{\partial h}{\partial y} + \frac{\Delta^2}{2!}\frac{\partial^2 h}{\partial y^2} . \tag{8.123}$$

Eliminując z powyższych równań pierwszą pochodną, otrzymujemy następującą aproksymację drugiej pochodnej względem y w węźle (i, j):

$$\left.\frac{\partial^2 h}{\partial y^2}\right|_{i,j} = \frac{2}{\xi(1+\xi)\Delta^2} h_A + \frac{2}{(1+\xi)\Delta^2} h_{i-1,j} - \frac{2}{\xi\Delta^2} h_{i,j} \tag{8.124}$$

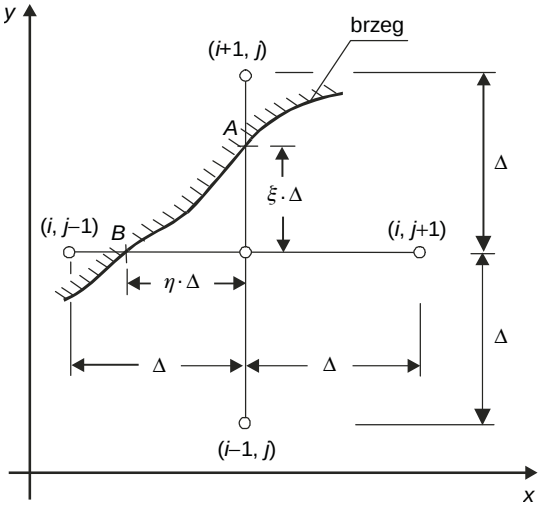
Analogicznie, dla kierunku x otrzymamy

$$\left.\frac{\partial^2 h}{\partial x^2}\right|_{i,j} = \frac{2}{\eta(1+\eta)\Delta^2} h_B + \frac{2}{(1+\eta)\Delta^2} h_{i,j+1} - \frac{2}{\eta\Delta^2} h_{i,j} . \tag{8.125}$$

Różnicowym odpowiednikiem równania Laplace'a dla rozważanego węzła będzie więc wyrażenie

$$\frac{2}{\xi(1+\xi)} h_A + \frac{2}{\eta(1+\eta)} h_B + \frac{2}{1+\eta} h_{j,j+1} + \frac{2}{1+\xi} h_{i-1,j} - \left(\frac{2}{\xi} + \frac{2}{\eta}\right) h_{i,j} = 0 , \tag{8.126}$$

w którym $0 < \xi \le 1$ i $0 < \eta \le 1$ oznaczają skrócenie ramion siatki, zaś h_A, h_B, $h_{i,j+1}$, $h_{i-1,j}$, $h_{i,j}$ są wartościami funkcji odpowiednio w punktach A, B, $(i-1, j)$, $(i, j+1)$, (i, j). Warto zauważyć, że dla $\xi = \eta = 1{,}0$ wyrażenie (8.126) przechodzi natychmiast w (8.103).



Rys. 8.16. Położenie brzegu obszaru w stosunku do węzłów

Przypadek, kiedy w punktach A i B (rys. 8.16) zadana jest nie funkcja h, lecz jej pochodna, tzn. warunek typu Neumanna, jest bardziej skomplikowany i nie będzie tu omawiany.

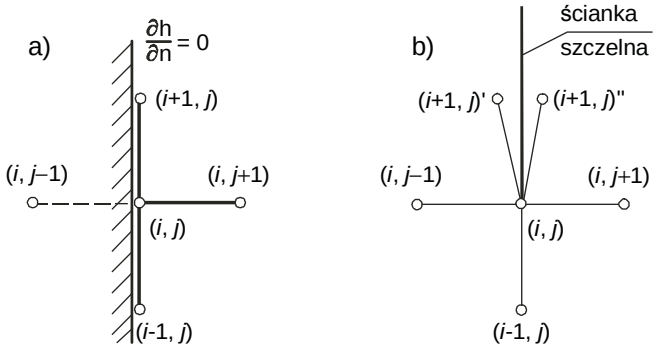
Uwzględnienie warunku Neumanna jest natomiast stosunkowo proste w sytuacji, gdy brzeg obszaru przechodzi przez węzły siatki. Jest to szczególnie łatwe, gdy brzeg

przebiega równolegle do jednej z osi współrzędnych i jest to brzeg nieprzepuszczalny, tzn. gdy

$$\partial h/\partial n = 0. \tag{8.127}$$

Rozpatrzmy sytuację w węźle leżącym na lewym pionowym brzegu (rys. 8.17a). Dla węzła tego piszemy równanie (8.103), będące aproksymacją równania Laplace'a

$$h_{i+1,j} + h_{i,j+1} + h_{i-1,j} + h_{i,j-1} - 4h_{i,j} = 0\,. \tag{8.128}$$



Rys. 8.17. Układ węzłów na brzegu nieprzepuszczalnym i ostrzu ścianki szczelnej

Dokonajmy również aproksymacji warunku (8.127), stosując iloraz różnicowy centralny. Będziemy mieli

$$\left.\frac{\partial h}{\partial n}\right|_{i,j} = \left.\frac{\partial h}{\partial x}\right|_{i,j} \approx \frac{h_{i,j+1} - h_{i,j-1}}{2\Delta} = 0. \tag{8.129}$$

Ponieważ węzeł $(i, j-1)$ leży poza obszarem, więc wartość $h_{i,j-1}$ jest fikcyjna i nie może wystąpić w równaniu różnicowym dla węzła (i, j). Uwzględniając jednak na podstawie równania (8.129), że $h_{i,j-1} = h_{i,j+1}$, możemy napisać

$$h_{i-1,j} + 2h_{i,j+1} + h_{i+1,j} - 4h_{i,j} = 0\,. \tag{8.130}$$

Jest to różnicowy odpowiednik równania Laplace'a dla węzła na lewym brzegu, z uwzględnieniem warunku nieprzepuszczalności brzegu. Utworzenie analogicznych równań różnicowych dla węzłów leżących na brzegu górnym, dolnym lub prawym oraz w narożach nie przedstawia trudności. Postępuje się analogicznie.

Jeśli w obszarze całkowania istnieje ścianka szczelna, wówczas należy wprowadzić jeszcze jeden rodzaj węzła, a zatem jeszcze jedną postać równania różnicowego. Ścianka szczelna ma grubość bardzo małą w stosunku do wymiaru siatki węzłów. Należy ją więc traktować jako nieskończenie cienki obiekt, na którego obu powierzchniach, lewej i prawej (od strony górnej i dolnej wody), znajdują się węzły, w których spełniony jest warunek (8.127). Przyjmując, że koniec ścianki szczelnej znajduje się w węźle i, j, możemy dla tego węzła napisać (rys. 8.17b)

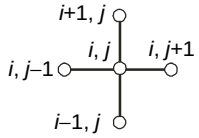
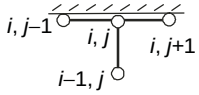
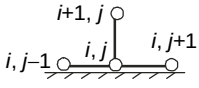
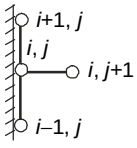
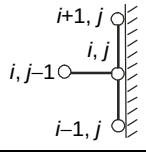
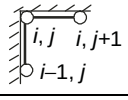
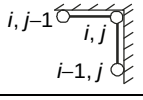
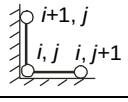
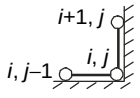
$$h'_{i+1,j} + h_{i,j+1} + h_{i-1,j} + h_{i,j-1} - 4h_{i,j} = 0\,, \tag{8.131}$$

oraz

$$h''_{i+1,j} + h_{i,j+1} + h_{i-1,j} + h_{i,j-1} - 4h_{i,j} = 0\,. \tag{8.132}$$

Tabela 8.1

Różnicowe odpowiedniki równania Laplace'a

Nr	Położenie węzła	Równanie różnicowe	Uwagi
1		$h_{i+1,j} + h_{i,j+1} + h_{i-1,j} + h_{i,j-1} - 4h_{i,j} = 0$	węzeł wewnętrzny
2		$h_{i,j+1} + 2h_{i-1,j} + h_{i,j-1} - 4h_{i,j} = 0$	węzeł na brzegu nieprzepuszczalnym
3		$2h_{i+1,j} + h_{i,j+1} + h_{i,j-1} - 4h_{i,j} = 0$	-"-
4		$h_{i+1,j} + 2h_{i,j+1} + h_{i-1,j} - 4h_{i,j} = 0$	-"-
5		$h_{i+1,j} + h_{i-1,j} + 2h_{i,j-1} - 4h_{i,j} = 0$	-"-
6		$2h_{i,j+1} + 2h_{i-1,j} - 4h_{i,j} = 0$	-"-
7		$2h_{i-1,j} + 2h_{i,j-1} - 4h_{i,j} = 0$	-"-
8		$2h_{i+1,j} + 2h_{i,j+1} - 4h_{i,j} = 0$	-"-
9		$2h_{i+1,j} + 2h_{i,j-1} - 4h_{i,j} = 0$	-"-

Dodając powyższe równania do siebie i dzieląc przez 2, otrzymujemy

$$\frac{1}{2}(h'_{i+1,j} + h''_{i+1,j}) + h_{i,j+1} + h_{i-1,j} + h_{i,j-1} - 4h_{i,j} = 0\,, \tag{8.133}$$

przy czym $h'_{i+1,j}$ i $h''_{i+1,j}$ oznaczają ciśnienia piezometryczne w węzłach leżących na lewej i prawej powierzchni ścianki, bezpośrednio nad węzłem centralnym.

Równania różnicowe dla omówionych wyżej warunków zestawiono w tabeli 8.1.

Rozwiązując równanie filtracji ustalonej pod ciśnieniem, każdemu węzłowi przypisujemy odpowiednie równanie różnicowe, zależnie od położenia węzła w obszarze. Następnie, po przyjęciu początkowego przybliżenia funkcji h we wszystkich węzłach, obliczamy kolejne przybliżenia, stosując metodę iteracyjną (8.109), (8.110) lub (8.114). Należy pamiętać, że formuły te są realizacją metod iteracyjnego rozwiązania układu algebraicznych równań liniowych, chociaż układu tego w postaci (8.104) nie budujemy w sposób jawny.

Omówione wyżej rodzaje warunków brzegowych i odpowiadające im różnicowe odpowiedniki równania Laplace'a nie wyczerpują wszystkich możliwości, umożliwiają jednak modelowanie szerokiej klasy zagadnień praktycznych, nie tylko dotyczących filtracji, lecz również szeregu innych zjawisk, opisywanych tym równaniem.

■ *Przykład 8.4*

Filtracja pod budowlą piętrzącą

Rozwiążmy zagadnienie filtracji pod budowlą piętrzącą wynikającą z różnicy poziomów wody górnej i dolnej, wywołanej przez budowlę. Jak wiadomo, zwykle zagadnienie filtracji można traktować jako płaskie, co uzasadnia stosowanie równania Laplace'a.

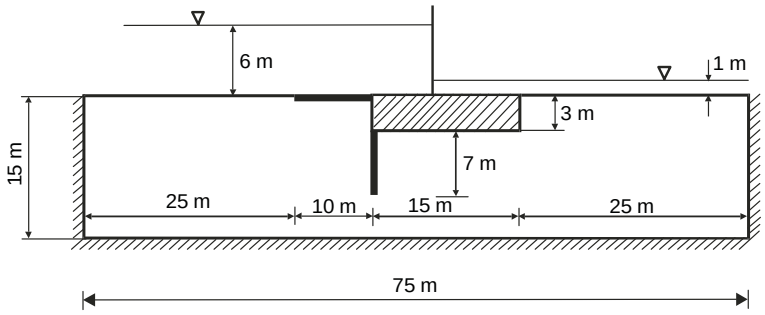
Załóżmy, że warunki posadowienia budowli są następujące:

— grunt jest jednorodny,
— warstwa nieprzepuszczalna jest pozioma, zaś miąższość warstwy przepuszczalnej jest stała i wynosi 15 m,
— aktywny obszar filtracji ma długość 75 m,
— od strony wody górnej zastosowano uszczelnienie w postaci fartucha iłowego o długości 10 m i grubości 1 m,
— na krawędzi budowli wykonano przesłonę do głębokości 7 m poniżej spodu fundamentu.

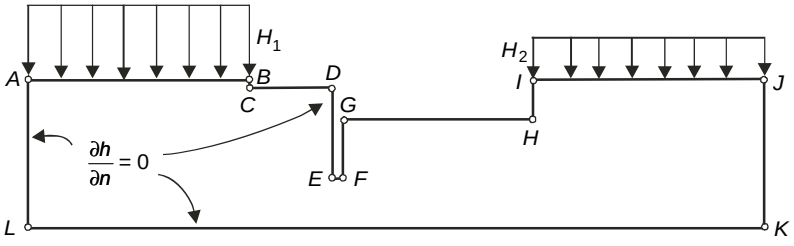
Wymiary fundamentu, elementów uszczelniających oraz obszaru filtracji przedstawiono na rys. 8.4.1. Poziom wody górnej wynosi $H_1 = 6$ m, zaś dolnej – $H_2 = 1$ m. Rozważany obszar filtracji ogranicza linia łamana *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J*, *K*, *L*, pokazana na rys. 8.4.2.

Przyjęto następujące warunki brzegowe:

— na odcinkach *AB* i *IJ* znane są wartości funkcji h, które wynoszą odpowiednio $H_1 = 6$ m i $H_2 = 1$ m,
— na pozostałych odcinkach, tzn. *BC*, *CD*, *DE*, *EF*, *FG*, *GH*, *HI*, *JK*, *KL* i *LA*, znana jest wartość pochodnej normalnej do brzegu: $\partial h/\partial n = 0$.



Rys. 8.4.1. Obszar filtracji pod budowlą piętrzącą

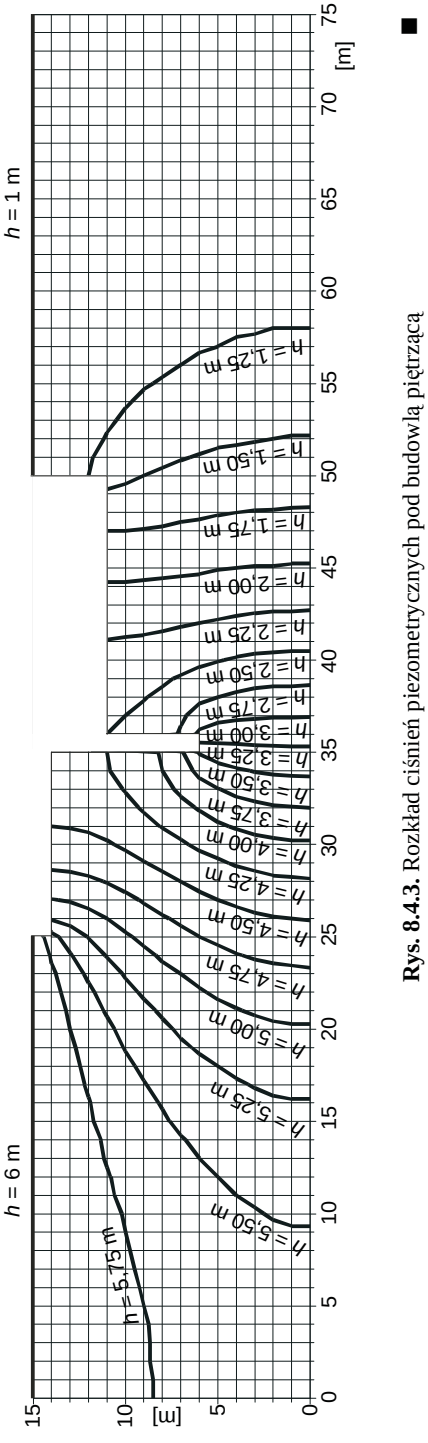


Rys. 8.4.2. Warunki brzegowe dla filtracji pod budowlą piętrzącą

Mamy więc do czynienia z zagadnieniem III rodzaju, czyli z mieszanymi warunkami brzegowymi. Rozwiązaniem równania Laplace'a z powyższymi warunkami brzegowymi jest funkcja $h(x, y)$ reprezentująca ciśnienie piezometryczne. W celu numerycznego rozwiązania przedstawionego zagadnienia, ciągły obszar filtracji zastąpiony został obszarem dyskretnym. Zastosowano siatkę kwadratową o wymiarach oczka $\Delta x = 1$ m i $\Delta y = 1$ m. W obszarze tym mamy do czynienia z różnymi typami węzłów obliczeniowych. Równania algebraiczne obowiązujące w poszczególnych typach węzłów zestawiono w tabeli 8.1. Do rozwiązania otrzymanego układu równań algebraicznych użyto metody Jacobiego. Wynikiem obliczeń są ciśnienia piezometryczne w węzłach siatki. Rozkład ciśnień zmieniających się od $h = H_1 = 6$ m – od strony wody górnej, do $h = H_2 = 1$ m – od strony wody dolnej, przedstawiono na rys. 8.4.3. Znajomość położenia linii jednakowych ciśnień piezometrycznych pozwala wyznaczyć rodzinę linii prądu, a na ich podstawie rozkład prędkości wody wypływającej z gruntu na dolnym stanowisku, rozkład ciśnienia i parcia działającego na fundament budowli, a także ilość wody filtrującej pod budowlą piętrzącą.

Liczba iteracji zależy od przyjęcia pierwszego przybliżenia rozwiązania oraz od jego dokładności. Na przykład, rozwiązanie przybliżone z dokładnością $\varepsilon = 0{,}001$ otrzymano po:
— 1514 iteracjach przy $\mathbf{h}^{(0)} = 0$;
— 1242 iteracjach przy $\mathbf{h}^{(0)} = H_1$;
— 1279 iteracjach przy $\mathbf{h}^{(0)} = H_2$;
— 753 iteracjach przy $\mathbf{h}^{(0)} = (H_1 + H_2)/2$.

Natomiast rozwiązanie z dokładnością $\varepsilon = 0{,}01$ przy $\mathbf{h}^{(0)} = (H_1 + H_2)/2$ otrzymano po 61 iteracjach.



Rys. 8.4.3. Rozkład ciśnień piezometrycznych pod budowlą piętrzącą

8.6. Rozwiązanie układu równań uderzenia hydraulicznego metodą charakterystyk z zastosowaniem metody różnic skończonych

Poszukujemy przybliżonego rozwiązania układu równań różniczkowych cząstkowych typu hiperbolicznego (7.8), (7.9), opisującego nieustalony przepływ w rurociągu

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + g\frac{\partial H}{\partial x} + \frac{f}{2d}U^2 = 0\,, \tag{8.134}$$

$$\frac{\partial H}{\partial t} + U\frac{\partial H}{\partial x} + \frac{c^2}{g}\frac{\partial U}{\partial x} = 0\,, \tag{8.135}$$

gdzie: x – położenie,
t – czas,
U – prędkość,
H – ciśnienie piezometryczne w rurociągu,
f – współczynnik oporów na długości,
d – średnica rurociągu,
g – przyspieszenie ziemskie,
c – prędkość fali ciśnienia.

Prędkość fali ciśnienia jest wyrażona następująco:

$$c = \frac{1}{\sqrt{\rho\left(\frac{1}{K} + \frac{d}{Ee}\right)}}\,, \tag{8.136}$$

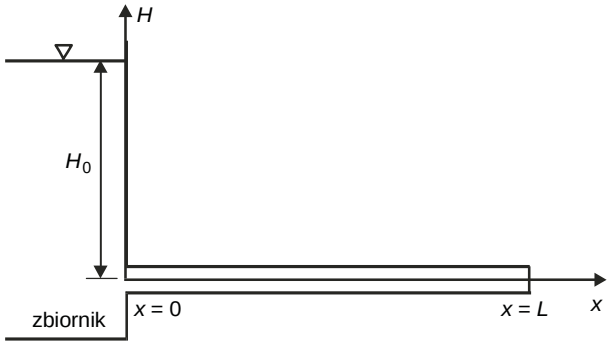
gdzie: ρ – gęstość cieczy,
K – współczynnik ściśliwości cieczy,
E – moduł sprężystości ścian przewodu,
e – grubość ścianki przewodu.

Dla układu tego formułuje się tzw. problem początkowo-brzegowy, który rozwiązuje się w obszarze: $0 \le x \le L$, $t \ge 0$ (rys. 8.18). Warunek początkowy określa się przy założeniu w chwili $t = 0$ w rurociągu ruchu ustalonego. Pozwala to określić funkcje $U(x, t = 0)$ i $H(x, t = 0)$ na całej jego długości. Na brzegach $x = 0$ i $x = L$, tzn. na początku i końcu rurociągu należy zadać albo funkcję $U(t)$, albo $H(t)$, ponieważ – jak pokazano w podrozdziale 7.2 – z każdego brzegu jedna charakterystyka wchodzi do obszaru rozwiązania.

Klasyczną metodą rozwiązania układu (8.134), (8.135) jest tzw. metoda charakterystyk. Szczególnie szeroko stosowany jest jej wariant, w którym wykorzystuje się stałą siatkę węzłów. W efekcie otrzymujemy algorytm rozwiązania silnie związany z metodą różnic skończonych. Ten wariant metody charakterystyk zastosujemy do rozwiązania równań uderzenia hydraulicznego.

Istota metody charakterystyk polega na odpowiedniej zamianie układu współrzędnych. Dzięki temu układ równań różniczkowych cząstkowych przekształca się w układ równań różniczkowych zwyczajnych, który następnie rozwiązuje się znaną metodą różnic skończonych (Ziółko, 2000). W metodzie charakterystyk zamieniamy układ współrzędnych (x, t) na (φ, ξ), $x = x(\varphi, \xi)$, $t = t\,(\varphi, \xi)$, przy czym nowe współrzędne zmieniają się wzdłuż krzywych zwanych charakterystykami. Dokładny opis metody charakterystyk podają np. Legras

(1974), Godunow (1975), Abbott (1979), Abbott i Basco (1989), Ziółko (2000) i inni. Tutaj przytoczymy za Streeterem i Lai (1962) skrócony sposób wyprowadzania równań charakterystyk, równań na charakterystykach oraz metodę ich rozwiązania na stałej siatce węzłów.



Rys. 8.18. Schemat rurociągu prostego

W układzie równań (8.134) i (8.135) należy wyznaczyć pochodne funkcji U i H w wybranym kierunku definiującym charakterystykę. W tym celu przepiszmy wymieniony układ

$$J_1 = \frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + g\frac{\partial H}{\partial x} + \frac{f}{2d}U^2 = 0\,, \tag{8.138}$$

$$J_2 = \frac{\partial H}{\partial t} + U\frac{\partial H}{\partial x} + \frac{c^2}{g}\frac{\partial U}{\partial x} = 0 \tag{8.139}$$

i utwórzmy liniową kombinację równań o współczynniku λ. Otrzymamy:

$$J = J_1 + \lambda J_2 = \frac{\partial U}{\partial t} + \left(U + \lambda\frac{c^2}{g}\right)\frac{\partial U}{\partial x} + \lambda\frac{\partial H}{\partial t} + \lambda\left(\frac{g}{\lambda} + U\right)\frac{\partial H}{\partial x} + \frac{f}{2d}U^2 = 0\,. \tag{8.140}$$

Ponieważ $U = U(x, t)$ i $H = H(x, t)$, ich różniczki zupełne są równe

$$dU = \frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial t}dt\,, \tag{8.141}$$

$$dH = \frac{\partial H}{\partial x}dx + \frac{\partial H}{\partial t}dt\,. \tag{8.142}$$

Porównując odpowiednie człony równania (8.140) z powyższymi zależnościami, można poszukiwać ich identyczności

$$\left(U + \lambda\frac{c^2}{g}\right)\frac{\partial U}{\partial x} + \frac{\partial U}{\partial t} = \frac{\partial U}{\partial x}\cdot\frac{dx}{dt} + \frac{\partial U}{\partial t} = \frac{dU}{dt}\,, \tag{8.143}$$

$$\left(\frac{g}{\lambda} + U\right)\frac{\partial H}{\partial x} + \frac{\partial H}{\partial t} = \frac{\partial H}{\partial x}\cdot\frac{dx}{dt} + \frac{\partial H}{\partial t} = \frac{dH}{dt}\,. \tag{8.144}$$

Równanie (8.140) może być zredukowane do postaci

$$J = \frac{dU}{dt} + \lambda \frac{dH}{dt} + \frac{f}{2d} U^2 = 0 , \tag{8.145}$$

jeśli spełnione są następujące warunki:

$$\frac{dx}{dt} = U + \lambda \frac{c^2}{g} , \tag{8.146}$$

$$\frac{dx}{dt} = \frac{g}{\lambda} + U . \tag{8.147}$$

Z powyższej pary równań wyznaczamy współczynnik λ

$$\lambda = \pm \frac{g}{c} . \tag{8.148}$$

Po podstawieniu powyższego wyniku do równań (8.146), (8.147) i (8.145) otrzymujemy cztery następujące równania związane z rozwiązywanym układem (8.134) i (8.135):

$$\left.\frac{dt}{dx}\right|_{+} = \frac{1}{U + c}, \quad \left.\frac{dt}{dx}\right|_{-} = \frac{1}{U - c} , \tag{8.149, 8.150}$$

$$\frac{dU}{dt} + \frac{g}{c}\frac{dH}{dt} + \frac{f}{2d} U^2 = 0 , \tag{8.151}$$

$$\frac{dU}{dt} - \frac{g}{c}\frac{dH}{dt} + \frac{f}{2d} U^2 = 0 . \tag{8.152}$$

Równania (8.149) i (8.150) definiują dwie rodziny omówionych w podrozdziale 7.2 krzywych charakterystycznych (nazywanych w skrócie charakterystykami) na płaszczyźnie ($x - t$). Natomiast równania (8.151) i (8.152), zależne tylko od czasu, definiują funkcje $U(x, t)$ i $H(x, t)$ na wymienionych charakterystykach. Każda para funkcji $U(x, t)$ i $H(x, t)$, będąca rozwiązaniem układu (8.135) i (8.134), spełnia równania (8.149)÷(8.152).

Z równań (8.149) i (8.150) wynika – znany z podrozdziału 7.2 – fakt, że układ równań uderzenia hydraulicznego ma dwie rodziny charakterystyk różniących się kątem nachylenia stycznych do nich. Konsekwentnie, biorąc pod uwagę znak, mówimy o charakterystyce dodatniej C^+ oraz o charakterystyce ujemnej C^-. Z charakterystyką C^+ związane są równania

$$\frac{dt}{dx} - \frac{1}{U + c} = 0, \quad \frac{dU}{dt} + \frac{g}{c}\frac{dH}{dt} + \frac{f}{2d} U^2 = 0 , \tag{8.153, 8.154}$$

zaś z charakterystyką C^-

$$\frac{dt}{dx} - \frac{1}{U - c} = 0, \quad \frac{dU}{dt} - \frac{g}{c}\frac{dH}{dt} + \frac{f}{2d} U^2 = 0 . \tag{8.155, 8.156}$$

Układ obu charakterystyk na płaszczyźnie ($x - t$) przedstawiono na rysunku 8.19.

Załóżmy, że na przecinających się charakterystykach dane są punkty R oraz S o współrzędnych (x_R, t_R) i (x_S, t_S), w których znane są wartości funkcji U i H, czyli U_R i H_R oraz U_S i H_S. Na podstawie wymienionych informacji można obliczyć współrzędne punktu przecięcia charakterystyk x_P i t_P oraz wartości poszukiwanych funkcji w tym punkcie, czyli U_P

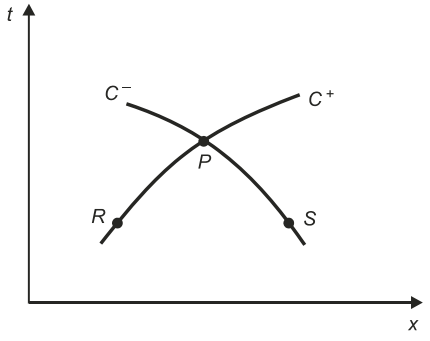
i H_P. W tym celu należy dokonać aproksymacji równań (8.153)÷(8.156) w sposób typowy dla metody różnic skończonych. Otrzymamy kolejno:

$$\frac{t_P - t_R}{x_P - x_R} - \frac{1}{U_R + c} = 0, \tag{8.157}$$

$$\frac{U_P - U_R}{t_P - t_R} + \frac{g}{c}\frac{H_P - H_R}{t_P - t_R} + \frac{f}{2d}U_R^2 = 0, \tag{8.158}$$

$$\frac{t_P - t_S}{x_P - x_S} - \frac{1}{U_S - c} = 0, \tag{8.159}$$

$$\frac{U_P - U_S}{t_P - t_S} - \frac{g}{c}\frac{H_P - H_S}{t_P - t_S} + \frac{f}{2d}U_S^2 = 0. \tag{8.160}$$

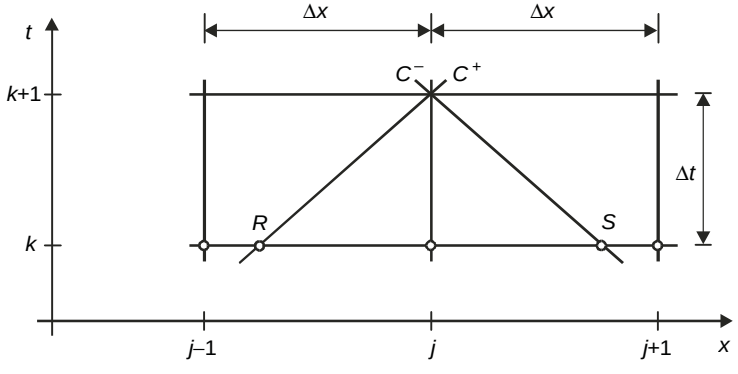


Rys. 8.19. Przecięcie dwóch charakterystyk

Powyższe równania tworzą układ, którego rozwiązaniem są współrzędne punktu P (x_P i t_P) oraz wartości poszukiwanych funkcji U_P i H_P.

Przedstawiony sposób obliczeń, chociaż bardzo prosty, jest jednak niepraktyczny. Chodzi o to, że punkt P, w którym otrzymujemy rozwiązanie, przemieszcza się po płaszczyźnie $(x - t)$ w sposób zależny od funkcji U i H w punktach R i S. W efekcie otrzymane rozwiązania nie będą związane z wybranym punktem rurociągu o danej współrzędnej, co w obliczeniach inżynierskich jest istotną sprawą. Bardziej praktyczny jest taki wariant metody charakterystyk, w którym operuje się stałą siatką węzłów, jak w metodzie różnic skończonych.

Rozpatrzmy fragment siatki węzłów, która pokrywa obszar rozwiązania układu (8.134), (8.135): $0 \le x \le L$, $t \ge 0$. Przedstawiono go na rysunku 8.20.



Rys. 8.20. Siatka węzłów z przecinającymi się charakterystykami

W węzłach siatki na poziomie czasowym k znamy wartości funkcji U oraz H. Jest to zadany warunek początkowy lub wynik obliczeń w poprzednim kroku czasowym. Przez węzeł $(j, k + 1)$ poprowadźmy charakterystyki układu równań C^+ oraz C^-. Przetną one poziom czasowy k odpowiednio w punktach R i S. Jak wiemy, dysponując równaniami (8.157)÷(8.160) i znając współrzędne obu punktów oraz wartości funkcji U i H w tych punktach, można określić zarówno współrzędne węzła $(j, k + 1)$, jak i poszukiwane wartości funkcji w nim, tzn. U_j^{k+1} oraz H_j^{k+1}. Jeśli przyjąć siatkę węzłów, jak na rys. 8.20, współrzędne punktu przecięcia charakterystyk są oczywiście znane, gdyż z góry zakładamy, że przecięcie będzie miało miejsce w węźle $(j, k + 1)$. Nie są natomiast znane wartości U_R, H_R, U_S i H_S konieczne do obliczenia U_j^{k+1} oraz H_j^{k+1}. Można je jednak obliczyć na podstawie wartości obu funkcji w węzłach siatki na poziomie czasu k. Współrzędne obu punktów będą odpowiednio równe:

$$x_R = x_j - (U_j^k + c)\Delta t\,, \tag{8.161}$$

$$x_S = x_j - (U_j^k - c)\Delta t\,, \tag{8.162}$$

gdzie: Δt jest krokiem czasowym.

Ponieważ przestrzenny wymiar siatki Δx jest stosunkowo mały, można założyć, że prędkości przepływu w punktach R i S nieznacznie różnią się od prędkości w węźle (j, k), co pozwala przyjąć

$$(U + c)_R = U_j^k + c \quad \text{oraz} \quad (U - c)_S = U_j^k - c\,.$$

Jeśli znamy współrzędne punktów R i S możemy wyznaczyć wartości U_R, H_R, U_S i H_R. W tym celu dokonujemy interpolacji liniowej pomiędzy węzłami $(j - 1, k)$ i (j, k) oraz (j, k) i $(j + 1, k)$. Efektem tej operacji są następujące zależności:

$$U_R = U_j^k\left(1 - \frac{\Delta t}{\Delta x}\left(U_j^k + c\right)\right) + U_{j-1}^k\left(U_j^k + c\right)\frac{\Delta t}{\Delta x}\,, \tag{8.163}$$

$$U_S = U_j^k\left(1 + \frac{\Delta t}{\Delta x}\left(U_j^k - c\right)\right) - U_{j+1}^k\left(U_j^k - c\right)\frac{\Delta t}{\Delta x}\,, \tag{8.164}$$

$$H_R = H_j^k\left(1 - \frac{\Delta t}{\Delta x}\left(U_j^k + c\right)\right) + H_{j-1}^k\left(U_j^k + c\right)\frac{\Delta t}{\Delta x}\,, \tag{8.165}$$

$$H_S = H_j^k\left(1 + \frac{\Delta t}{\Delta x}\left(U_j^k - c\right)\right) - H_{j+1}^k\left(U_j^k - c\right)\frac{\Delta t}{\Delta x}\,. \tag{8.166}$$

Znając wartości ciśnienia i prędkości w punktach R i S, nieznane prędkości i ciśnienia w węzłach na następnym poziomie oblicza się z układu równań (8.158) i (8.159):

$$\frac{U_j^{k+1} - U_R}{\Delta t} + \frac{g}{c}\cdot\frac{H_j^{k+1} - H_R}{\Delta t} + \frac{f}{2d}(U_j^k)^2 = 0\,, \tag{8.167}$$

$$\frac{U_j^{k+1} - U_S}{\Delta t} - \frac{g}{c}\cdot\frac{H_j^{k+1} - H_S}{\Delta t} + \frac{f}{2d}(U_j^k)^2 = 0\,. \tag{8.168}$$

Rozwiązując powyższy układ równań, otrzymujemy następujące zależności:

$$U_j^{k+1} = \frac{1}{2}\left(U_R + U_S\right) + \frac{g}{2c}\left(H_R - H_S\right) - \left(\frac{f}{2d} U|U|\right)_j^k \Delta t\,, \tag{8.169}$$

$$H_j^{k+1} = \frac{c}{2g}\left(U_R - U_S\right) + \frac{1}{2}\left(H_R + H_S\right) \tag{8.170}$$

$$\text{dla } j = 2, 3, ..., M-1,$$

gdzie: M – liczba węzłów.

W równaniu (8.169) zamiast wyrażenia U^2 wprowadzono $U|U|$, co zapewnia automatyczne uwzględnienie w trakcie obliczeń znaku siły tarcia, która działa zawsze w kierunku przeciwnym do kierunku przepływu.

Równania algebraiczne (8.169) i (8.170) umożliwiają obliczenie przybliżonych wartości prędkości i ciśnienia we wszystkich węzłach wewnętrznych na poziomie czasu $k + 1$. Szczególnego potraktowania wymagają węzły leżące na obu końcach rurociągu.

W węzłach tych zadane są następujące warunki brzegowe:

— na początku rurociągu, od strony zbiornika ($j = 1$) przyjmuje się stałą wartość ciśnienia $H_1 = H_z = \text{const}$,

— na końcu rurociągu ($j = M$) zakłada się, że prędkość wypływu jest funkcją stopnia otwarcia zaworu. Oblicza się ją, rozwiązując następujące równanie nieliniowe (Streeter i Lai, 1962):

$$\frac{U_M^{k+1}}{U_0} = \tau_p \left(\frac{1}{H_0} \left(H_M^k + \frac{c}{g}\left(U_M^k - U_M^{k+1}\right)\right)\right)^{1/2}, \tag{8.171}$$

gdzie: U_0 – początkowa prędkość w stanie ustalonym,
H_0 – początkowe ciśnienie odpowiadające prędkości U_0,
τ_p – stosunek efektywnego otwarcia zaworu w czasie t_{k+1} do otwarcia w chwili $t = 0$, który może być obliczony wg formuły:

$$\tau_p = \left(1 - \frac{t}{T_c}\right)^2, \tag{8.172}$$

T_c – czas zamykania zaworu.

Równanie (1.172) jest ważne, gdy $t \le T_c$. Przy $t > T_c$ $U_M^{k+1} = 0$, co oznacza całkowite zamknięcie przepływu.

W konsekwencji, w węzłach skrajnych $j = 1$ oraz $j = M$, znana jest jedna funkcja określona przez warunek brzegowy H_1 oraz U_M. Do obliczenia brakujących wartości U_1 i H_M w tych węzłach wykorzystuje się równania charakterystyk. I tak, na brzegu lewym, od strony zbiornika, gdzie zadany jest warunek $H_1 = H_z = \text{const}$, prędkość oblicza się według formuły:

$$U_1^{k+1} = U_S + \frac{g}{c}\left(H_1^k - H_S\right) - \left(\frac{f}{2d} U|U|\right)_1^k \Delta t\,, \tag{8.173}$$

natomiast na brzegu prawym, po obliczeniu prędkości wypływu z równania (8.171), brakujące ciśnienie oblicza się według formuły:

$$H_M^{k+1} = H_R - \frac{c}{g}\left(U_M^k - U_R\right) - \left(\frac{cf}{2gd} U^2\right)_M^k \Delta t\,. \tag{8.174}$$

Występujące w powyższych wyrażeniach wartości H_S, U_S, H_R i U_R obliczane są zgodnie z wzorami (8.163)÷(8.166). W ten sposób dla zadanych warunków początkowych i brzegowych można obliczyć wartości ciśnienia i prędkości we wszystkich węzłach na poziomie czasu $k + 1$.

Analizę stabilności i dokładności rozwiązania równań uderzenia hydraulicznego metodą charakterystyk wykonamy dla ich zlinearyzowanej postaci. Pominięcie członów nieliniowych i siły tarcia pozwala zapisać układ (8.134), (8.135) w postaci

$$\frac{\partial U}{\partial t} + g\frac{\partial H}{\partial x} = 0\,, \tag{8.175}$$

$$\frac{\partial H}{\partial t} + \frac{c^2}{g}\frac{\partial U}{\partial x} = 0\,. \tag{8.176}$$

Wynikiem linearyzacji układu (8.134) i (8.135) jest zmiana równań jego charakterystyk. Będą one prostymi o nachyleniu:

$$\left.\frac{dx}{dt}\right|_{+} = \frac{1}{c} \quad \text{oraz} \quad \left.\frac{dx}{dt}\right|_{-} = -\frac{1}{c}\,. \tag{8.177, 8.178}$$

Z kolei konsekwencją zmiany równań charakterystyk, a także pominięcia tarcia, jest zmiana równań aproksymujących (8.169) i (8.170). Będą one następujące:

$$U_j^{k+1} = \frac{1}{2}\left(U_R + U_S\right) + \frac{g}{2c}\left(H_R - H_S\right), \tag{8.179}$$

$$H_j^{k+1} = \frac{c}{2g}\left(U_R - U_S\right) + \frac{1}{2}\left(H_R + H_S\right), \tag{8.180}$$

przy czym równania (8.163)÷(8.166) przyjmą postać:

$$U_R = U_j^k\left(1 - \frac{\Delta t}{\Delta x}c\right) + U_{j-1}^k\frac{\Delta t}{\Delta x}c\,, \tag{8.181}$$

$$U_S = U_j^k\left(1 - \frac{\Delta t}{\Delta x}c\right) + U_{j+1}^k\frac{\Delta t}{\Delta x}c\,, \tag{8.182}$$

$$H_R = H_j^k\left(1 - \frac{\Delta t}{\Delta x}c\right) + H_{j-1}^k\frac{\Delta t}{\Delta x}c\,, \tag{8.183}$$

$$H_S = H_j^k\left(1 - \frac{\Delta t}{\Delta x}c\right) + H_{j+1}^k\frac{\Delta t}{\Delta x}c\,. \tag{8.184}$$

Wstawiając zależności (8.181)÷(8.184) do równań (8.179) i (8.180), otrzymujemy

$$\frac{U_j^{k+1} - U_j^k}{\Delta t} + g\frac{H_{j+1}^k - H_{j-1}^k}{2\Delta x} + \left(-\frac{1}{2}U_{j-1}^k + U_j^k - \frac{1}{2}U_{j+1}^k\right)\frac{c}{\Delta x} = 0\,, \tag{8.185}$$

$$\frac{H_j^{k+1} - H_j^k}{\Delta t} + \frac{c^2}{g}\frac{U_{j+1}^k - U_{j-1}^k}{2\Delta x} + \left(-\frac{1}{2}H_{j-1}^k + H_j^k - \frac{1}{2}H_{j+1}^k\right)\frac{c}{\Delta x} = 0\,. \tag{8.186}$$

Stabilność numeryczną rozwiązania zbadajmy znaną metodą Neumanna, opisaną w podrozdziale 7.5. W jej wyniku otrzymuje się macierz wzmocnienia, której wartość własna jest równa:

$$\lambda = 1 + C_a\left(\cos\left(m\Delta x\right) - 1\right) - i\, C_a \sin\left(m\Delta x\right), \tag{8.187}$$

gdzie: i – jednostka urojona,
m – liczba falowa,
$C_a = c \cdot \Delta t/\Delta x$ – liczba Couranta.

Warunek stabilności numerycznej (Fletcher, 1991) wymaga, aby największa wartość modułu wartości własnej była nie większa od 1, czyli aby

$$\left|\lambda\right| = \left(1 - 4C_a\left(1 - C_a\right)\sin^2\left(\frac{m\Delta x}{2}\right)\right)^{1/2} \le 1. \tag{8.188}$$

Ponieważ $C_a > 0$ oraz $\sin^2(m\Delta x/2) > 0$ dla każdej liczby falowej m, to warunek powyższy będzie spełniony, gdy:

$$C_a \le 1. \tag{8.189}$$

Jest to znany warunek stabilności numerycznej schematów jawnych dla równań hiperbolicznych.

Analiza dokładności rozwiązania równań wyjściowych prowadzi do równań zmodyfikowanych o postaci

$$\frac{\partial U}{\partial t} + g\frac{\partial H}{\partial x} = \nu_n \frac{\partial^2 U}{\partial x^2} + \dots\ , \tag{8.190}$$

$$\frac{\partial H}{\partial t} + \frac{c^2}{g}\frac{\partial U}{\partial x} = \nu_n \frac{\partial^2 H}{\partial x^2} + \dots, \tag{8.191}$$

w których współczynnik dyfuzji numerycznej zdefiniowany jest zależnością

$$\nu_n = \frac{c\Delta x}{2}\left(1 - C_a\right). \tag{8.192}$$

W tym przypadku błąd dyfuzji numerycznej został wprowadzony wskutek liniowej interpolacji pomiędzy węzłami siatki zastosowanej do obliczenia wartości funkcji U i H w punktach przecięcia charakterystyk z poziomem czasu k, w tzw. punktach R i S (rys. 8.20).

Identyczny wynik otrzymuje się, rozwiązując równanie czystej adwekcji różnicowym schematem typu „up–wind", opisanym w podrozdziale 7.5.

Z równania współczynnika dyfuzji numerycznej (8.192) wynika, że:

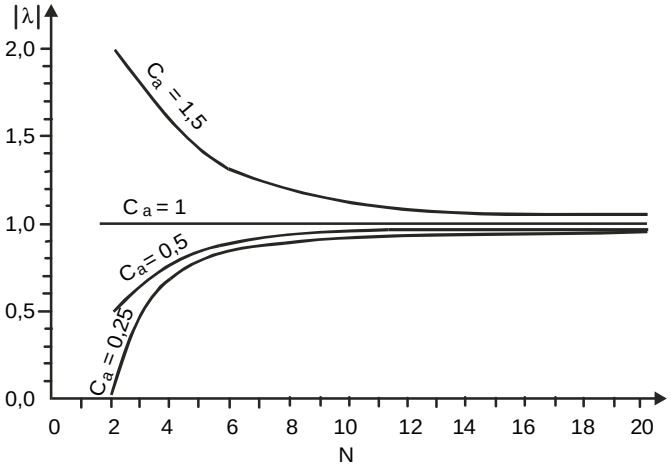
— metoda charakterystyk zapewnia dokładne rozwiązanie przy $C_a = 1$, gdyż w tym przypadku $\nu_n = 0$. W tej sytuacji charakterystyki pokrywają się z przekątnymi oczek siatki i przecinają poziom czasu k w węzłach $(j - 1, k)$ i $(j + 1, k)$ (rys. 8.20);

— metoda charakterystyk jest niestabilna, gdy $C_a > 1$. W tym przypadku $\nu_n < 0$, a to oznacza, że problem początkowo-brzegowy dla układu (8.191), (8.192) jest niepoprawnie postawiony i w konsekwencji jego rozwiązanie nie istnieje;

— metoda charakterystyk zapewnia stabilne rozwiązanie, gdy $C_a < 1$, ponieważ przy $\nu_n > 0$ problem początkowo–brzegowy dla układu (8.191), (8.192) jest poprawnie postawiony, a zatem jego rozwiązanie istnieje zawsze. W tym przypadku metoda generuje dyfuzję numeryczną. Jej wielkość rośnie ze wzrostem Δx i zmniejszeniem liczby Couranta, po-

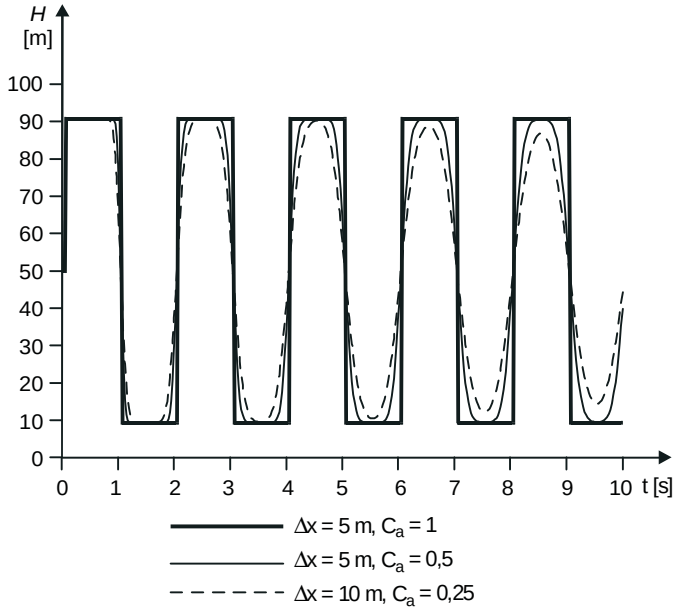
wodując wygładzenie rozwiązania i redukcję gradientów ciśnienia, a także tłumienie amplitudy fali.

Powyższe wnioski znajdują potwierdzenie na wykresie $|\lambda| = f(N)$. Moduł wartości własnej przedstawiono tutaj jako funkcję liczby interwałów Δx przypadających na falę o długości L_f, czyli $N = L_f/\Delta x$. Zgodnie z (7.111) pozwala to zastąpić w równaniu wyrażenie $m \cdot \Delta x$ wyrażeniem $2\pi/N$ (Szymkiewicz, 2000). Na rysunku 8.21 widzimy, że metoda charakterystyk jest warunkowo stabilna, gdyż spełnienie nierówności $|\lambda| \le 1$ otrzymujemy tylko przy $C_a \le 1$. Tłumienie amplitudy, duże dla fal krótkich (małe N), rośnie w miarę zmniejszania liczby Couranta. Przy $C_a = 1$ otrzymuje się $|\lambda| = 1$, czyli fala nie jest ani wzmacniana, ani tłumiona. Dla liczb Couranta większych od jedności otrzymuje się $|\lambda| > 1$, co oznacza wzmocnienie amplitudy fali, czyli niestabilność metody.



Rys. 8.21. Wykres $|\lambda| = f(N)$ dla metody charakterystyk reprezentujący tłumienie amplitudy fali przy różnych liczbach Couranta

Dodatkowym potwierdzeniem przedstawionych wniosków o właściwościach metody charakterystyk są wyniki obliczeń testowych. Zlinearyzowane równania (8.175) i (8.176) rozwiązujemy w rurociągu prostym o długości L =500 m. Początkowa prędkość przepływu w stanie ustalonym wynosi U_0 = 0,5 m/s. Poziom wody w zbiorniku jest równy H_0 = 50 m npp i nie zmienia się. Zamknięcie zaworu następuje w sposób natychmiastowy, czyli w czasie $T_c < \Delta t$. Rozwiązanie dokładne układu równań uderzenia hydraulicznego ma postać niegasnącej oscylacji. W tym przypadku charakterystyki pokrywają się z przekątnymi siatki i przechodzą przez jej węzły. Identyczne rozwiązanie otrzymujemy przyjmując wartości Δx i Δt, przy których liczba Couranta równa jest jedności. Dla innych wartości wymienionych parametrów otrzymujemy rozwiązanie stabilne, gdy $C_a < 1$. Jest ono jednak obciążone błędem dyfuzji numerycznej. Na rysunku 8.22 przedstawiono obliczone ciśnienia na końcu rurociągu dla różnych wartości Δx oraz C_a. Zgodnie z (8.192) błąd dyfuzji rośnie w miarę wzrostu Δx i zmniejszenia C_a.



Rys. 8.22. Rozwiązanie zlinearyzowanych równań uderzenia hydraulicznego metodą charakterystyk dla różnych wartości Δx oraz różnych liczb Couranta

9 Algorytmy rozwiązania równań różniczkowych o pochodnych cząstkowych metodą elementów skończonych

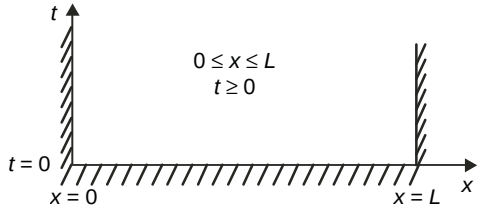
9.1. Rozwiązanie jednowymiarowego równania adwekcji-dyfuzji

Równanie adwekcji-dyfuzji w przypadku jednowymiarowym, przy założeniu stałego współczynnika dyfuzji i braku członu źródłowego ma postać (7.16)

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} - D\frac{\partial^2 C}{\partial x^2} = 0\,. \tag{9.1}$$

Równanie tego typu opisuje szeroką gamę różnych procesów przenoszenia. Jest ono stosowane również do opisu procesu przenoszenia nierozkładalnego czynnika rozpuszczonego w wodzie płynącego cieku (np. chlorki) w strefie dobrego wymieszania w przekroju poprzecznym. W takim przypadku symbolom występującym w równaniu (9.1) nadaje się następujące znaczenie:

x – współrzędna przestrzenna,
t – czas,
$C = C(x, t)$ – koncentracja czynnika rozpuszczonego w wodzie,
$U = U(x, t)$ – uśredniona w przekroju poprzecznym cieku prędkość przepływu wody,
D – współczynnik dyfuzji.



Rys. 9.1. Obszar całkowania równania (9.1) i jego granice

Problem rozwiązania równania (9.1) formułujemy następująco: w obszarze rozwiązania: $0 \le x \le L$, $t \ge 0$ (rys. 9.1) należy znaleźć funkcję $C(x, t)$, która spełnia równanie oraz zadane warunki na granicach obszaru. Warunki te są następujące:

— warunki początkowe
dla $t = 0$: $C(x, t) = C_p(x)$, $\quad (0 \le x \le L)$,

— warunki brzegowe

dla $x = 0$: $C(x, t) = C_0(t)$, $(t \geq 0)$,

dla $x = L$: $C(x, t) = C_L(t)$, $(t \geq 0)$,

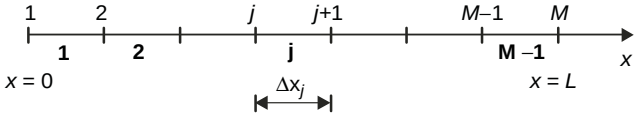
lub

$$\left.\frac{\partial C}{\partial x}\right|_{x=L} = \phi_L(t) \qquad (t \geq 0)$$

gdzie: L – długość odcinka kanału,

$C_p(t)$, $C_o(t)$, $C_L(t)$, $\phi_L(t)$ – znane funkcje.

Zgodnie z wymogami metody elementów skończonych, odcinek kanału o długości całkowitej L podzielono na $M - 1$ elementów liniowych (rys. 9.2) o długości Δx_j ($j = 1, 2, ..., M - 1$).



Rys. 9.2. Podział odcinka cieku o długości L na elementy skończone

W ten sposób otrzymano M węzłów ograniczających elementy, w których poszukiwane będzie rozwiązanie przybliżone. Wewnątrz dowolnego elementu j ograniczonego węzłami j oraz $j + 1$, poszukiwaną funkcję C aproksymuje się zgodnie z (7.60), czyli

$$\widehat{C} = \mathbf{N}\,\mathbf{C} = \ldots N_j(x)\,C_j(t) + N_{j+1}(x)\,C_{j+1}(t)\ldots, \tag{9.2}$$

gdzie funkcje kształtu, w tym przypadku liniowe, opisuje się równaniami (7.67):

$$N_j(x) = \frac{x_{j+1} - x}{x_{j+1} - x_j} = \frac{x_{j+1} - x}{\Delta x_j}; \quad N_{j+1}(x) = \frac{x - x_j}{x_{j+1} - x_j} = \frac{x - x_j}{\Delta x_j}; \quad x \in \left\langle x_j,\, x_{j+1} \right\rangle. \tag{9.3}$$

Pochodne funkcji $\widehat{C}$ są równe:

$$\begin{aligned} \frac{d\widehat{C}}{dx} &= \frac{1}{\Delta x_j}\left(-C_j(t) + C_{j+1}(t)\right), \\ \frac{d\widehat{C}}{dt} &= N_j(x)\frac{dC_j}{dt} + N_{j+1}(x)\frac{dC_{j+1}}{dt}. \end{aligned} \tag{9.4}$$

Procedura Galerkina opisana w podrozdziale 7.4 wymaga, aby poszukiwana funkcja $\widehat{C}$ spełniała warunek

$$\int_0^L \left[\frac{\partial \widehat{C}}{\partial t} + U\frac{\partial \widehat{C}}{\partial x} - D\frac{\partial^2 \widehat{C}}{\partial x^2}\right]\mathbf{N}^T dx = \sum_{j=1}^{M-1}\int_{x_j}^{x_{j+1}} \left[\frac{\partial \widehat{C}}{\partial t} + U\frac{\partial \widehat{C}}{\partial x} - D\frac{\partial^2 \widehat{C}}{\partial x^2}\right]\mathbf{N}^T dx = 0\,. \tag{9.5}$$

Ponieważ w elemencie j tylko $N_j(x) \neq 0$ i $N_{j+1}(x) \neq 0$, należy obliczyć w nim dwie następujące całki:

$$I_j = \int_{x_j}^{x_{j+1}} \left[\frac{\partial \widehat{C}}{\partial t} + U \frac{\partial \widehat{C}}{\partial x} - D \frac{\partial^2 \widehat{C}}{\partial x^2} \right] N_j \, dx , \tag{9.6}$$

$$I_{j+1} = \int_{x_j}^{x_{j+1}} \left[\frac{\partial \widehat{C}}{\partial t} + U \frac{\partial \widehat{C}}{\partial x} - D \frac{\partial^2 \widehat{C}}{\partial x^2} \right] N_{j+1} \, dx . \tag{9.7}$$

Obliczmy najpierw całkę (9.6). Ponieważ można ją przedstawić w postaci sumy całek z kolejnych członów wyrażenia podcałkowego:

$$I_j = I_j^{(1)} + I_j^{(2)} - I_j^{(3)} , \tag{9.8}$$

będziemy obliczali kolejne jej składowe. I tak:

$$\begin{aligned} I_j^{(1)} &= \int_{x_j}^{x_{j+1}} \frac{\partial \widehat{C}}{\partial t} N_j dx = \int_{x_j}^{x_{j+1}} \frac{\partial}{\partial t} (N_j C_j + N_{j+1} C_{j+1}) N_j \, dx = \\ &= \frac{dC_j}{dt} \int_{x_j}^{x_{j+1}} N_j^2 dx + \frac{dC_{j+1}}{dt} \int_{x_j}^{x_{j+1}} N_j N_{j+1} dx = \\ &= \frac{\Delta x_j}{3} \frac{dC_j}{dt} + \frac{\Delta x_j}{6} \frac{dC_{j+1}}{dt} , \end{aligned} \tag{9.9}$$

$$\begin{aligned} I_j^{(2)} &= \int_{x_j}^{x_{j+1}} U \frac{\partial \widehat{C}}{\partial x} N_j dx = U \int_{x_j}^{x_{j+1}} \frac{\partial}{\partial x} (N_j C_j + N_{j+1} C_{j+1}) N_j \, dx = \\ &= U \int_{x_j}^{x_{j+1}} \left(-\frac{C_j}{\Delta x_j} + \frac{C_{j+1}}{\Delta x_j} \right) N_j dx = U \left(-\frac{C_j}{\Delta x_j} + \frac{C_{j+1}}{\Delta x_j} \right) \int_{x_j}^{x_{j+1}} N_j dx = \\ &= \frac{U}{2} (-C_j + C_{j+1}). \end{aligned} \tag{9.10}$$

W całce $I_j^{(3)}$ występuje pochodna funkcji C II rzędu. Tymczasem do aproksymacji C zastosowano funkcje liniowe, dla których nie istnieje druga pochodna. W celu uniknięcia wynikającego stąd konfliktu, można dokonać całkowania przez części, według znanej formuły (Dziubiński i Świątkowski, 1980):

$$\int u dv = uv - \int v du . \tag{9.11}$$

Przyjmując

$$u = N_j \quad \text{oraz} \quad v = \frac{\partial \widehat{C}}{\partial x} ,$$

otrzymamy:

$$I_j^{(3)} = \int_{x_j}^{x_{j+1}} D\frac{\partial^2 \hat{C}}{\partial x^2} N_j dx = D\frac{dC_j}{dx}\bigg|_{x_j}^{x_{j+1}} - D\int_{x_j}^{x_{j+1}} \frac{\partial \hat{C}}{\partial x}\frac{\partial N_j}{\partial x} dx =$$

$$= D\frac{dC_j}{dx}\bigg|_{x_j}^{x_{j+1}} - D\int_{x_j}^{x_{j+1}} \frac{\partial}{\partial x}(N_j C_j + N_{j+1} C_{j+1})\left(-\frac{1}{\Delta x_j}\right) dx = \tag{9.12}$$

$$= -D\frac{dC_j}{dx} + \frac{D}{\Delta x_j}(-C_j + C_{j+1}).$$

Człon $-D\dfrac{dC_j}{dx}$ reprezentuje strumień przez brzeg elementu, to znaczy przez węzeł *j*.

Wprowadzając obliczone całki do (9.8), otrzymujemy równanie:

$$I_j = \frac{\Delta x_j}{6}\left(2\frac{dC_j}{dt} + \frac{dC_{j+1}}{dt}\right) + \frac{U}{2}(-C_j + C_{j+1}) +$$
$$- D\frac{1}{\Delta x_j}(-C_j + C_{j+1}) + D\frac{dC_j}{dx}. \tag{9.13}$$

Obliczenie całki (9.7) przeprowadzamy w sposób analogiczny. Po wykonaniu podobnych obliczeń ostatecznie otrzymujemy wynik, który pozwala zapisać zależność (9.7) w postaci:

$$I_{j+1} = \frac{\Delta x_j}{6}\left(\frac{dC_j}{dt} + 2\frac{dC_{j+1}}{dt}\right) + \frac{U}{2}(-C_j + C_{j+1}) + \frac{D}{\Delta x_j}(-C_j + C_{j+1}) +$$
$$- D\frac{dC_{j+1}}{dx}\ . \tag{9.14}$$

Wynikiem obliczenia całek (9.6) i (9.7) są więc zależności (9.13) i (9.14). Występuje w nich zarówno funkcja *C*, jak i jej pochodne względem czasu w węzłach. Zapiszmy wynik całkowania w elemencie *j*, stosując notację macierzową. Otrzymamy:

$$\int_{x_j}^{x_{j+1}} \left(\frac{\partial \hat{C}}{\partial t} + U\frac{\partial \hat{C}}{\partial x} - D\frac{\partial^2 \hat{C}}{\partial x^2}\right)\mathbf{N}^T dx = \mathbf{S}_j \frac{d\mathbf{C}}{dt} + \mathbf{A}_j\mathbf{C} + \mathbf{F}_j\,, \tag{9.15}$$

gdzie: $\dfrac{d\mathbf{C}}{dt} = \left(\dfrac{dC_1}{dt}, \ldots, \dfrac{dC_j}{dt}, \ldots, \dfrac{dC_M}{dt}\right)^T$,

$\mathbf{C} = (C_1, ..., C_j, C_{j+1}, ..., C_M)^T$,

$\mathbf{F}_j = (0, ..., F_j, F_{j+1}, ..., 0)^T$.

$\mathbf{S}_j =$

	1			j	j+1			M
1								
j				$S_{j,j}$	$S_{j,j+1}$			
j+1				$S_{j,j+1}$	$S_{j+1,j+1}$			
M								

$\mathbf{A}_j =$

	1			j	j+1			M
1								
j				$A_{j,j}$	$A_{j,j+1}$			
j+1				$A_{j,j+1}$	$A_{j+1,j+1}$			
M								

Niezerowe elementy macierzy $\mathbf{S}_j$, $\mathbf{A}_j$ i wektora $\mathbf{F}_j$, zgodnie z zależnościami (9.13) i (9.14), są zdefiniowane następująco:

$$S_{j,j} = \frac{\Delta x_j}{3}, \tag{9.16a}$$

$$S_{j,j+1} = \frac{\Delta x_j}{6}, \tag{9.16b}$$

$$S_{j+1,j} = \frac{\Delta x_j}{6}, \tag{9.16c}$$

$$S_{j+1,j+1} = \frac{\Delta x_j}{3}, \tag{9.16d}$$

$$A_{j,j} = -\frac{U}{2} + \frac{D}{\Delta x_j}, \tag{9.17a}$$

$$A_{j,j+1} = \frac{U}{2} - \frac{D}{\Delta x_j}, \tag{9.17b}$$

$$A_{j+1,j} = -\frac{U}{2} - \frac{D}{\Delta x_j}, \tag{9.17c}$$

$$A_{j+1,j+1} = \frac{U}{2} + \frac{D}{\Delta x_j}, \tag{9.17d}$$

$$F_j = D\frac{dC_j}{dx}, \tag{9.18a}$$

$$F_{j+1} = -D\frac{dC_{j+1}}{dx}. \tag{9.18b}$$

Wynik całkowania w elemencie *j* w postaci (9.15) podstawmy do zależności (9.5) wynikającej z procedury Galerkina. Otrzymamy:

$$\begin{aligned} &\int_0^L \left(\frac{\partial \widehat{C}}{\partial t} + U\frac{\partial \widehat{C}}{\partial x} - D\frac{\partial \widehat{C}}{\partial x^2} \right) \mathbf{N}^T dx = \\ &= \sum_{j=1}^{M-1} \left(\mathbf{S}_j \frac{d\mathbf{C}}{dt} + \mathbf{A}_j \mathbf{C} + \mathbf{F}_j \right) = \mathbf{0}. \end{aligned} \tag{9.19}$$

Rozpisując wyrażenie objęte symbolem sumowania, uzyskujemy zależność

$$\left(\sum_{j=1}^{M-1} \mathbf{S}_j \right) \frac{d\mathbf{C}}{dt} + \left(\sum_{j=1}^{M-1} \mathbf{A}_j \right) \mathbf{C} + \sum_{j=1}^{M-1} \mathbf{F}_j = \mathbf{0}. \tag{9.20}$$

Jeśli wprowadzimy następujące oznaczenia:

$$\mathbf{S} = \sum_{j=1}^{M-1} \mathbf{S}_j, \quad \mathbf{A} = \sum_{j=1}^{M-1} \mathbf{A}_j, \quad \mathbf{F} = \sum_{j=1}^{M-1} \mathbf{F}_j,$$

to wynikiem warunku (9.5) będzie układ równań różniczkowych zwyczajnych o postaci

$$\mathbf{S}\frac{d\mathbf{C}}{dt} + \mathbf{A}\mathbf{C} + \mathbf{F} = \mathbf{0}. \tag{9.21}$$

Jego wymiar jest równy liczbie węzłów występujących w obszarze, tzn. $M \times M$. Zauważmy, że wobec definicji $\mathbf{F}_j$, wektor $\mathbf{F}$ będzie miał składowe zerowe z wyjątkiem pierwszej i ostatniej.

W ten sposób problem rozwiązania równania różniczkowego cząstkowego (9.1) sprowadzono do zagadnienia początkowego układu równań różniczkowych zwyczajnych (9.21). Warunek początkowy jest następujący: $t = 0$, $\mathbf{C} = \mathbf{C}_p$, gdzie $\mathbf{C}_p$ jest wektorem zawierającym początkowe wartości węzłowe funkcji C, tzn. odpowiadające chwili $t = 0$.

Do rozwiązania (9.21) można zastosować dowolną metodę rozwiązywania układów równań różniczkowych zwyczajnych. Dobre rezultaty daje niejawny schemat trapezowy opisany w punkcie 6.2.3:

$$\mathbf{C}_{t+\Delta t} = \mathbf{C}_t + \frac{\Delta t}{2}(\mathbf{C}'_t + \mathbf{C}'_{t+\Delta t}), \tag{9.22}$$

gdzie: Δt – krok czasowy,
$\mathbf{C}' = d\mathbf{C}/dt$.

Zastosowanie tego schematu do układu (9.21) prowadzi do zależności

$$\left(\mathbf{S} + \frac{\Delta t}{2}\mathbf{A}\right)\mathbf{C}_{t+\Delta t} = \left(\mathbf{S} - \frac{\Delta t}{2}\mathbf{A}\right)\mathbf{C}_t - \frac{\Delta t}{2}(\mathbf{F}_t + \mathbf{F}_{t+\Delta t}). \tag{9.23}$$

Jeśli ponadto wprowadzimy dodatkowe oznaczenia:

$$\mathbf{B} = \mathbf{S} + \Delta t/2\ \mathbf{A}, \quad \mathbf{W} = (\mathbf{S} - \Delta t/2\ \mathbf{A}) \cdot \mathbf{C}_t - \frac{\Delta t}{2}(\mathbf{F}_t + \mathbf{F}_{t+\Delta t}),$$

to (9.23) można zapisać w ostatecznej następującej postaci:

$$\mathbf{B}\ \mathbf{C}_{t+\Delta t} = \mathbf{W}. \tag{9.24}$$

W ten sposób otrzymano układ algebraicznych równań liniowych. Jego rozwiązanie w kolejnych krokach czasowych: $t = \Delta t$, $2 \cdot \Delta t$, $3 \cdot \Delta t$, ..., z uwzględnieniem zadanych warunków brzegowych, jest rozwiązaniem równania (9.1). Są nim przybliżone wartości funkcji C w węzłach obszaru.

Macierz współczynników układu (9.24) **B** jest macierzą trójdiagonalną. Do rozwiązania tego układu najkorzystniej jest zastosować znaną metodę Thomasa.

Pewnego wyjaśnienia wymaga sposób uwzględnienia zadanych warunków brzegowych. Formułując zadanie, stwierdzono że na brzegu $x = 0$ zadany jest warunek Dirichleta, czyli funkcja $C_0(t)$. Natomiast na brzegu $x = L$ może być zadany również warunek typu Dirichleta $C_L(t)$ lub warunek Neumanna $\partial C/\partial x = \phi_L\ (t)$. Jeśli na obu brzegach zadane są warunki Dirichleta, oznacza to, że w wektorze niewiadomych $\mathbf{C}_{t+\Delta t}$ w układzie (9.24) znana jest pierwsza i ostatnia składowa. Zatem z układu tego należy wyeliminować pierwsze i ostatnie równanie. Jednakże postępowanie takie jest kłopotliwe. Aby nie przekształcać macierzy układu (9.24), wygodniej jest zastosować inny sposób uwzględnienia warunków brzegowych. Mianowicie pierwsze i ostatnie równanie układu należy zastąpić równaniem wynikającym z zadanych warunków brzegowych, tzn. przyjmując:

$$b_{1,1} = 1, \quad b_{1,2} = 0, \quad w_1 = C_0(t + \Delta t), \tag{9.25a,b,c}$$

$$b_{M,M-1} = 0, \quad b_{M,M} = 1, \quad w_1 = C_L(t + \Delta t). \tag{9.26a,b,c}$$

W tym przypadku wektory $\mathbf{F}_t$ i $\mathbf{F}_{t+\Delta t}$ znikają, bo wszystkie ich składowe są równe zero. Jeśli natomiast na brzegu $x = 0$ zadany jest warunek typu Dirichleta, zaś na brzegu $x = L$ warunek typu Neumanna, pierwsze równanie układu (9.24) modyfikujemy zgodnie z (9.25), natomiast ostatnie równanie pozostawiamy bez zmian. Warunek Neumanna wprowadzamy do układu poprzez wektor $\mathbf{F}_{t+\Delta t}$, którego ostatni element będzie miał wartość równą $\phi_L\ (t + \Delta t)$.

■ *Przykład 9.1*

Rozwiązanie jednowymiarowego równania filtracji nieustalonej metodą elementów skończonych

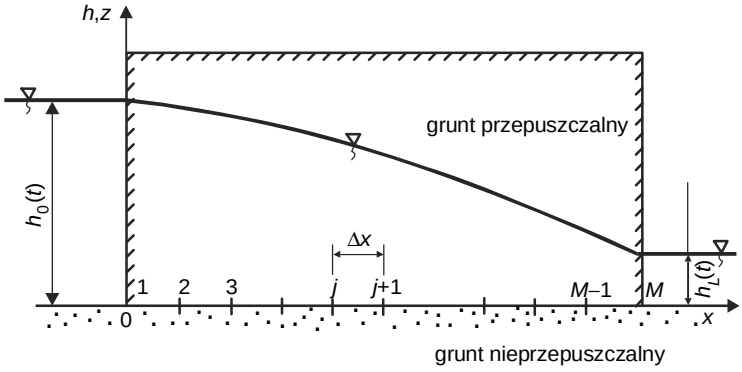
Zastosowanie opisanej metody elementów skończonych zilustrujemy na przykładzie równania jednowymiarowej filtracji nieustalonej ze swobodną powierzchnią, opisanej ogólnym równaniem (7.6). Równanie to, przy pominięciu członu źródłowego ($w = 0$) i po zlinearyzowaniu względem wysokości piezometrycznej h, można zapisać w postaci liniowego równania dyfuzji:

$$\frac{\partial h}{\partial t} - D\frac{\partial^2 h}{\partial x^2} = 0\,, \tag{9.1.1}$$

gdzie: $h = h(x, t)$ – wysokość piezometryczna,
$D = k \cdot (h_s - z)/\mu$,
k – współczynnik filtracji,
h_s – średnia wysokość piezometryczna,
z – rzędna spągu warstwy nieprzepuszczalnej,
μ – współczynnik porowatości efektywnej.

Opisuje ono przepływ wody przez groblę prostokątną wykonaną z przepuszczalnego gruntu i posadowioną na nieprzepuszczalnym poziomym podłożu. Grobla ziemna o wymiarach jak na rys. 9.1.1 oddziela dwa zbiorniki, w których poziomy wody mogą ulegać zmianom w czasie. Zmiany te wywołują zmianę położenia zwierciadła wody w grobli.

Dla równania (9.1.1) formułuje się następujące warunki graniczne:
— w chwili $t = 0$ znane jest położenie zwierciadła wody w grobli, czyli ($h(x, t = 0) = h_p(x)$),
— na brzegu prawym $x = 0$ i lewym $x = L$ zadane są zmiany położenia zwierciadła wody, czyli $h(x = 0, t) = h_0(t)$, oraz $h(x = L, t) = h_L(t)$.



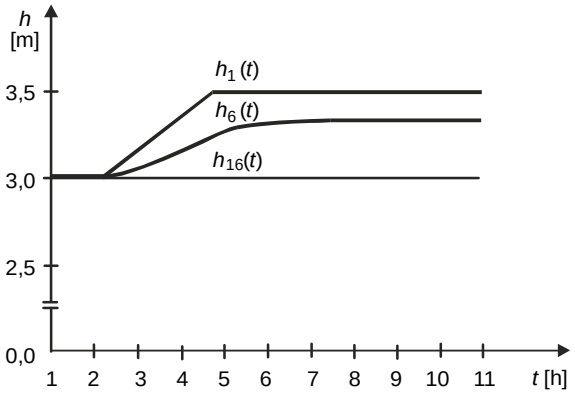
Rys. 9.1.1. Schemat filtracji przez groblę prostokątną

Zagadnienie to rozwiązano metodą elementów skończonych, opisaną w punkcie 9.1. Zauważmy bowiem, że po przyjęciu $U = 0$ i zastąpieniu koncentracji $C(x, t)$ wysokością piezometryczną $h(x, t)$ równanie (9.1.1) staje się tożsame z (9.1). Można więc rozwiązać je tą samą metodą. W rozwiązywanym przykładzie przyjęto następujące dane:

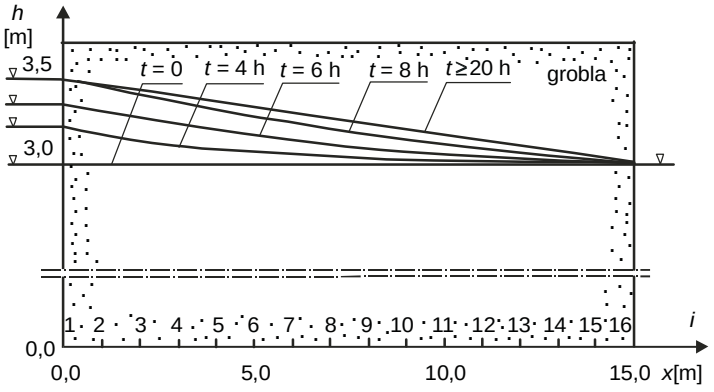
— groblę podzielono na 15 elementów, przyjmując krok przestrzenny $\Delta x = 1$ m = const,
— porowatość gruntu wynosi $\mu = 0{,}2$,
— współczynnik filtracji wynosi $k = 1$ m/h,
— początkowa rzędna zwierciadła wody wynosi $h_p(x) = H_0 = 3$ m = const,
— zmianę zwierciadła wody na brzegu $x = 0$ opisuje funkcja o postaci

$$h_0(t) = \begin{cases} H_0 & \text{dla} \quad t \le t_0 \\ H_0 + \dfrac{H_1 - H_0}{t_1 - t_0} \cdot (t - t_0) & \text{dla} \quad t_0 < t < t_1 \\ H_1 & \text{dla} \quad t \ge t_1 \end{cases},$$

przy czym przyjęto: $t_0 = 2{,}5$ h, $t_1 = 7{,}5$ h, $H_1 = 3{,}5$ m,
— na brzegu $x = L$ poziom wody nie zmienia się i wynosi $h_L(t) = H_0 = \text{const}$,
— średnia głębokość wynosi $h_s = 3{,}25$ m,
— krok całkowania wynosi $\Delta t = 0{,}025$ h,
— zjawisko odtworzono dla $0 \le t \le 20$ h.



Rys. 9.1.2. Przebieg zmian poziomów zwierciadła wody w wybranych węzłach grobli



Rys. 9.1.3. Chwilowe układy zwierciadła wody w grobli

Przebieg obliczeń odbywa się zgodnie z algorytmem przedstawionym w podrozdziale 9.1. Wyniki uzyskane dla przyjętych danych przedstawiono na rys. 9.1.2 i 9.1.3. Na rys. 9.1.2 przedstawiono otrzymane układy zwierciadła wody w grobli dla wybranych czasów. Dla przyjętych wartości parametrów zwierciadło stabilizuje się po ok. 20 h i układa się wzdłuż prostej. Jest to efekt linearyzacji równania 9.1.1. Na rys. 9.1.2 przedstawiono ewolucję poziomu wody w czasie w punkcie 6 (x = 5 m), czyli funkcję $h_6(t)$. Dla porównania naniesiono również funkcje $h_1(t)$ i $h_{16}(t)$, będące przyjętymi warunkami brzegowymi w rozwiązywanym zadaniu. ■

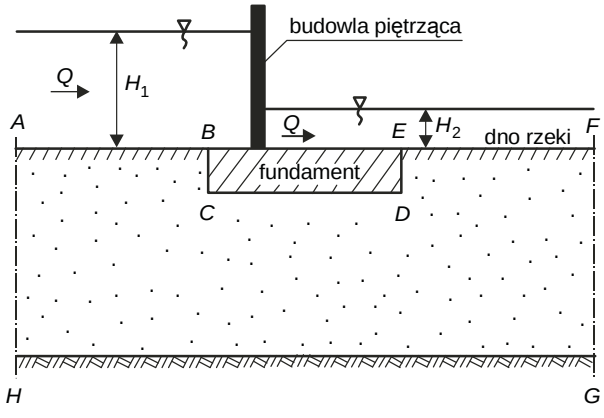
9.2. Rozwiązanie dwuwymiarowego równania filtracji ustalonej pod ciśnieniem

Przykład rozwiązania równania typu eliptycznego metodą elementów skończonych omówimy na przykładzie równania Laplace'a.

$$L(h) = \frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} = 0\,, \tag{9.26}$$

gdzie h oznacza ciśnienie piezometryczne.

Opisuje ono przepływ potencjalny cieczy nieściśliwej w ośrodku porowatym przy założeniu, że jest on ustalony, ośrodek jest jednorodny i izotropowy oraz zasilanie zewnętrzne nie występuje. Typowym przykładem filtracji ustalonej pod ciśnieniem w hydrotechnice jest filtracja pod budowlą piętrzącą, posadowioną na przepuszczalnej warstwie gruntu o skończonej miąższości.



Rys. 9.3. Schemat budowli piętrzącej posadowionej na przepuszczalnej warstwie gruntu

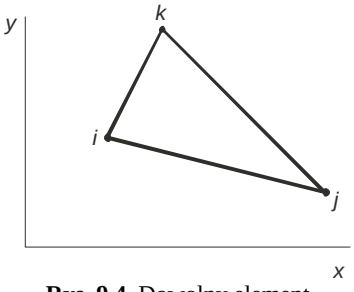
W przypadku przedstawionym na rys. 9.3 dla równania Laplace'a formułuje się następujące zagadnienie brzegowe: znaleźć funkcję $h(x, y)$ spełniającą w obszarze rozwiązania równanie Laplace'a, a na jego brzegach następujące warunki:

— odcinek brzegu: A÷B: $h = H_1$,

— odcinek brzegu E÷F: $h = H_2$,

— odcinki brzegu C÷D, G÷H: $\partial h/\partial y = 0$,
— odcinki brzegu A÷H, B÷C, D÷E, F÷G: $\partial h/\partial x = 0$.

Obszar filtracji pokrywamy siatką trójkątów. Trójkąty mogą mieć dowolną wielkość, jednak w obszarach spodziewanych dużych gradientów funkcji h siatkę należy zagęścić, natomiast w obszarach małych gradientów h siatka może być rzadka, a więc można stosować elementy o dużych rozmiarach. Wybierzmy dowolny element e utworzonej siatki węzłów (rys. 9.4).



Rys. 9.4. Dowolny element skończony o kształcie trójkątnym

Wprowadźmy najpierw ogólne wyrażenia charakterystyczne dla metody, związane z aproksymacją równania (9.26). Przyjmijmy, że wewnątrz trójkątnego elementu skończonego zbudowanego na węzłach i, j, k, potencjał h aproksymowany jest według (7.80) wyrażeniem

$$\hat{h} = \mathbf{Nh} = N_i h_i + N_j h_j + N_k h_k \,, \tag{9.27}$$

gdzie aproksymujące funkcje $N(x, y)$, tj. funkcje kształtu, są funkcjami liniowymi postaci (7.81)

$$N_p(x,y) = \frac{1}{2F_e}(a_p + b_p x + c_p y); \quad p = i, j, k \,, \tag{9.28}$$

w których F_e oznacza pole elementu, natomiast współczynniki a_p, b_p, c_p, zależne od współrzędnych węzłów elementu, obliczone są według formuł (7.83), (7.84) i (7.85).

Zastosujmy teraz do równania (9.26) procedurę Galerkina. Otrzymamy

$$\iint_F L(\hat{h}) N_l \, dxdy = \iint_F \left(\frac{\partial^2 \hat{h}}{\partial x^2} + \frac{\partial^2 \hat{h}}{\partial y^2} \right) N_l \, dxdy = 0, \qquad (l = 1, 2, \ldots, M), \tag{9.29}$$

gdzie: F – obszar rozwiązania,
M – liczba węzłów w obszarze rozwiązania F.

Wyrażenie (9.29) zawiera pochodne drugiego rzędu, co wymaga ciągłości pierwszej pochodnej funkcji kształtu we wszystkich punktach obszaru. Tymczasem do aproksymacji przyjęto liniowe funkcje bazowe. Ograniczenie to możemy ominąć, obniżając stopień operatora różniczkowego, podobnie jak to zrobiliśmy, rozwiązując równanie adwekcji-dyfuzji. W tym wypadku wygodnie jest użyć przekształcenia Greena o ogólnej postaci (Dziubiński i Świątkowski, 1980)

$$\iint_F \left(\psi \frac{\partial^2 \phi}{\partial x^2} + \psi \frac{\partial^2 \phi}{\partial y^2} \right) dxdy = -\iint_F \left(\frac{\partial \psi}{\partial x} \frac{\partial \phi}{\partial x} + \frac{\partial \psi}{\partial y} \frac{\partial \phi}{\partial y} \right) dxdy + \int_B \psi \frac{\partial \phi}{\partial n} dB \,, \tag{9.30}$$

gdzie: B – brzeg obszaru F,
n – kierunek normalny do brzegu B.

Całka krzywoliniowa istnieje tylko dla tych elementów, których jeden z boków leży na brzegu B. Reprezentuje ona strumień przez brzeg obszaru. Dla brzegów nieprzepuszczalnych jej wartość jest równa zeru.

W przypadku rozpatrywanego tutaj zagadnienia filtracji rozwiązywanego metodą elementów skończonych będzie $\Psi = N_l$ oraz $\phi = h$. Zatem warunek Galerkina zapiszemy:

$$\iint_F \left(\frac{\partial^2 \hat{h}}{\partial x^2} + \frac{\partial^2 \hat{h}}{\partial y^2} \right) N_l \, dxdy = \iint_F \left(N_l \frac{\partial^2 \hat{h}}{\partial x^2} + N_l \frac{\partial^2 \hat{h}}{\partial y^2} \right) dxdy =$$
$$= -\iint_F \left(\frac{\partial N_l}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial N_l}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) dxdy + \int_B N_l \frac{\partial \hat{h}}{\partial n} dB = 0, \; (l = 1, \; 2, \cdots, M). \tag{9.31}$$

Warunki graniczne zostały sformułowane tak, że na części brzegu dana jest funkcja *h*, a na części brzegu pochodna tej funkcji normalna do brzegu jest równa zero. Oznacza to, że na odcinkach brzegu równoległych do osi *y* obowiązuje warunek:

$$\frac{\partial h}{\partial x} = 0 = \frac{\partial h}{\partial n},$$

a na części brzegu równoległej do osi *x* obowiązuje warunek:

$$\frac{\partial h}{\partial y} = 0 = \frac{\partial h}{\partial n}.$$

Zatem dla przyjętych warunków brzegowych w tym przypadku na całym obwodzie *B* obszaru *F*, będzie

$$\int_B N_l \frac{\partial h}{\partial n} dB = 0 .$$

Całkę tę można więc pominąć w rozważaniach. Uwzględnienie tego faktu pozwala sprowadzić równanie (9.31) do postaci:

$$\iint_F \left(\frac{\partial N_l}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial N_l}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) dxdy = 0, \quad (l = 1, \; 2, \; \cdots, \; M), . \tag{9.32}$$

Korzystając z zasady sumowania całek w elementach tworzących obszar *F*, możemy napisać

$$\iint_F \left(\frac{\partial \mathbf{N}}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial \mathbf{N}}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) dxdy = \sum_{e=1}^{m} \iint_{F_e} \left(\frac{\partial \mathbf{N}}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial \mathbf{N}}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) dxdy = 0, \tag{9.33}$$

gdzie: F_e – pole elementu *e*,
m – liczba elementów,
$\mathbf{N} = (N_1, ..., N_i, ..., N_j, ..., N_k, ..., N_M)^T$ – wektor funkcji kształtu.

Spośród *m* elementów pokrywających obszar *F* wybierzmy jeden element o wierzchołkach oznaczonych indeksami *i*, *j* oraz *k* przedstawiony na rys. 9.4. W elemencie tym tylko 3 składowe wektora **N**, a mianowicie N_i, N_j oraz N_k, są niezerowe. Wszystkie pozostałe funkcje kształtu w elemencie *e* przyjmują wartości zerowe. Zatem dla elementu *e* należy obliczyć tylko 3 następujące całki:

$$I_l = \iint_{F_e} \left[\frac{\partial N_l}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial N_l}{\partial y} \frac{\partial \hat{h}}{\partial y} \right] dxdy, \quad (l = i, \, j, \, k) . \tag{9.34}$$

Policzmy powyższą całkę dla $l = i$.

$$I_i = \iint_{F_e}\left(\frac{\partial N_i}{\partial x}\frac{\partial \hat{h}}{\partial x}+\frac{\partial N_i}{\partial y}\frac{\partial \hat{h}}{\partial y}\right)dxdy =$$

$$= \iint_{F_e}\left(\frac{\partial N_i}{\partial x}\frac{\partial}{\partial x}(N_ih_i+N_jh_j+N_kh_k)+\frac{\partial N_i}{\partial y}\frac{\partial}{\partial y}(N_ih_i+N_jh_j+N_kh_k)\right)dxdy =$$

$$= \iint_{F_e}\left\{\left(\frac{\partial N_i}{\partial x}\frac{\partial N_i}{\partial x}h_i+\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}h_j+\frac{\partial N_i}{\partial x}\frac{\partial N_k}{\partial x}h_k\right)+\left(\frac{\partial N_i}{\partial y}\frac{\partial N_i}{\partial y}h_i+\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}h_j\right.\right. \tag{9.35}$$

$$\left.\left.+\frac{\partial N_i}{\partial y}\frac{\partial N_k}{\partial y}h_k\right)\right\}dxdy = h_i\iint_{F_e}\left(\frac{\partial N_i}{\partial x}\frac{\partial N_i}{\partial x}+\frac{\partial N_i}{\partial y}\frac{\partial N_i}{\partial y}\right)dxdy+$$

$$+h_j\iint_{F_e}\left(\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}+\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}\right)dxdy+h_k\iint_{F_e}\left(\frac{\partial N_i}{\partial x}\frac{\partial N_k}{\partial x}+\frac{\partial N_i}{\partial y}\frac{\partial N_k}{\partial y}\right)dxdy.$$

Pochodne funkcji bazowych, zgodnie z formułami (7.86) i (7.87), są stałe względem zmiennych całkowania, zatem powyższą całkę można obliczyć następująco:

$$I_i = h_i\iint_{F_e}\left(\frac{1}{2F_e}b_i\frac{1}{2F_e}b_i+\frac{1}{2F_e}c_i\frac{1}{2F_e}c_i\right)dxdy+$$

$$+h_j\iint_{F_e}\left(\frac{1}{2F_e}b_i\frac{1}{2F_e}b_j+\frac{1}{2F_e}c_i\frac{1}{2F_e}c_j\right)dxdy+$$

$$+h_k\iint_{F_e}\left(\frac{1}{2F_e}b_i\frac{1}{2F_e}b_k+\frac{1}{2F_e}c_i\frac{1}{2F_e}c_k\right)dxdy=$$

$$= h_i\left(\frac{1}{4F_e^2}b_ib_i\iint_{F_e}dxdy+\frac{1}{4F_e^2}c_ic_i\iint_{F_e}dxdy\right)+ \tag{9.36}$$

$$+h_j\left(\frac{1}{4F_e^2}b_ib_j\iint_{F_e}dxdy+\frac{1}{4F_e^2}c_ic_j\iint_{F_e}dxdy\right)+$$

$$+h_k\left(\frac{1}{4F_e^2}b_ib_k\iint_{F_e}dxdy+\frac{1}{4F_e^2}c_ic_k\iint_{F_e}dxdy\right)$$

Ponieważ

$$\iint_{F_e}dxdy = F_e\ ,$$

ostatecznie całkę I_i możemy zapisać następująco:

$$I_i = \frac{1}{4F_e}(b_ib_i+c_ic_i)h_i+\frac{1}{4F_e}(b_ib_j+c_ic_j)h_j+\frac{1}{4F_e}(b_ib_k+c_ic_k)h_k\ . \tag{9.37}$$

Podobne wyrażenia otrzymamy, obliczając całkę (9.34) dla pozostałych funkcji kształtu w elemencie *e*. I tak dla N_j uzyskamy

$$I_j = \frac{1}{4F_e}(b_j b_i + c_j c_i)h_i + \frac{1}{4F_e}(b_j b_j + c_j c_j)h_j + \frac{1}{4F_e}(b_j b_k + c_j c_k)h_k , \qquad (9.38)$$

zaś dla N_k będziemy mieli

$$I_k = \frac{1}{4F_e}(b_k b_i + c_k c_i)h_i + \frac{1}{4F_e}(b_k b_j + c_k c_j)h_j + \frac{1}{4F_e}(b_k b_k + c_k c_k)h_k . \qquad (9.39)$$

Zapiszmy otrzymane przez scałkowanie (9.34) zależności (9.37), (9.38) i (9.39), stosując notację macierzową:

$$\iint_{F_e}\left(\frac{\partial \mathbf{N}}{\partial x}\frac{\partial \widehat{h}}{\partial x} + \frac{\partial \mathbf{N}}{\partial y}\frac{\partial \widehat{h}}{\partial y}\right)dxdy = \mathbf{A}_e\mathbf{h} . \qquad (9.40)$$

Prawa strona powyższej zależności ma następującą postać:

	1		*i*		*j*		*k*		*M*		
											h_1
i			$A_{i,i}$		$A_{i,j}$		$A_{i,k}$				h_i
j			$A_{j,i}$		$A_{j,j}$		$A_{j,k}$			x	h_j
k			$A_{k,i}$		$A_{k,j}$		$A_{k,k}$				h_k
M											h_M

przy czym, zgodnie z (9.37), (9.38) i (9.39), elementy macierzy $\mathbf{A}_e$ są zdefiniowane następująco:

$$A_{i,i} = \frac{1}{4F_e}(b_i b_i + c_i c_i) , \qquad (9.41a)$$

$$A_{i,j} = \frac{1}{4F_e}(b_i b_j + c_i c_j) , \qquad (9.41b)$$

$$A_{i,k} = \frac{1}{4F_e}(b_i b_k + c_i c_k) , \qquad (9.41c)$$

$$A_{j,i} = \frac{1}{4F_e}(b_j b_i + c_j c_i) , \qquad (9.41d)$$

$$A_{j,j} = \frac{1}{4F_e}(b_j b_j + c_j c_j)\,, \tag{9.41e}$$

$$A_{j,k} = \frac{1}{4F_e}(b_j b_k + c_j c_k)\,, \tag{9.41f}$$

$$A_{k,i} = \frac{1}{4F_e}(b_k b_i + c_k c_i)\,, \tag{9.41g}$$

$$A_{k,j} = \frac{1}{4F_e}(b_k b_j + c_k c_j)\,, \tag{9.41h}$$

$$A_{k,k} = \frac{1}{4F_e}(b_k b_k + c_k c_k)\,. \tag{9.41i}$$

Zauważmy, że $\mathbf{A}_e$ jest macierzą symetryczną.

Postępując w podobny sposób, obliczamy całki dla każdego elementu $e = 1, 2, ..., m$, otrzymując zależności typu (9.40). Podstawmy wynik obliczeń do równania (9.33). Otrzymamy

$$\sum_{e=1}^{m} \iint_{F_e} \left(\frac{\partial \mathbf{N}}{\partial x}\frac{\partial \hat{h}}{\partial x} + \frac{\partial \mathbf{N}}{\partial y}\frac{\partial \hat{h}}{\partial y} \right) dxdy = \sum_{e=1}^{m} (\mathbf{A}_e \mathbf{h}) = 0\,. \tag{9.42}$$

Sumując rezultaty całkowania w każdym elemencie uzyskujemy następujące wyrażenie:

$$\mathbf{A}\,\mathbf{h} = \mathbf{0}, \tag{9.43}$$

w którym

$$\mathbf{A} = \sum_{e=1}^{m} \mathbf{A}_e\,. \tag{9.44}$$

Jest to układ algebraicznych równań liniowych. Macierz **A**, podobnie jak jej składowe $\mathbf{A}_e$, jest macierzą pasmową i symetryczną.

Do układu (9.43) należy wprowadzić – przyjęte na etapie formułowania problemu rozwiązania – warunki brzegowe. W tym przypadku są to warunki typu Dirichleta. Dodajmy, że warunki typu Neumanna zadane na częściach brzegu B zostały już uwzględnione w całce krzywoliniowej, co pozwoliło zapisać równanie (9.31) w postaci (9.32). Jak wiadomo, warunek typu Dirichleta oznacza, że w węzłach leżących na brzegu obszaru znane są wartości funkcji *h*. Oznacza to, że w wektorze **h** znane są te jego składowe, które odpowiadają węzłom brzegowym z takim warunkiem. Zatem z układu (9.43) należy wyeliminować wymienione składowe, jak i równania odpowiadające im. W rezultacie tego otrzymamy układ (9.43) o wymiarze równym liczbie węzłów, w których nie jest znana funkcja *h*. Postępowanie takie jest jednak kłopotliwe, gdyż zmienia strukturę układu równań. Z tego powodu wygodniej jest zastosować inny sposób uwzględnienia warunków brzegowych, niewymagający zmiany macierzy **A** i wektora **h** (Zienkiewicz, 1972).

Załóżmy, że rozwiązując układ równań o postaci

$$\mathbf{B}\,\mathbf{X} = \mathbf{W} \tag{9.45}$$

i wymiarze $M \times M$, znamy wartość *i*-tej składowej wektora niewiadomych **X**: $x_i = \alpha$. Fakt ten możemy uwzględnić, postępując w następujący sposób:

— element b_{ii} leżący na głównej przekątnej macierzy **B** w równaniu i mnożymy przez bardzo dużą liczbę d (około 8 rzędów większą od elementów macierzy):

$$b_{ii}^{*} = d \cdot b_{ii}\,, \tag{9.46}$$

— składową wektora wyrazów wolnych odpowiadającą równaniu i zamieniamy, zastępując wyrażeniem:

$$w_i^{*} = \alpha \cdot b_{ii}^{*}\,. \tag{9.47}$$

W efekcie rozwiązania tak zmodyfikowanego układu (9.45) otrzymujemy wartość x_i bardzo bliską zadanej wartości α.

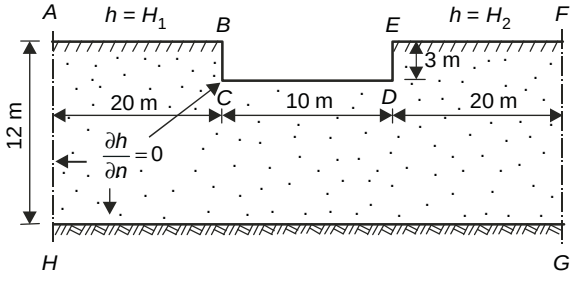
Ze względu na zwykle duże rozmiary układu (9.45) oraz dużą szerokość pasma jego macierzy współczynników, do rozwiązania stosuje się metody iteracyjne. W tym wypadku metody dokładne mogą być również bardzo efektywnym sposobem rozwiązania układu. Jednak wymaga to uwzględnienia symetrii macierzy współczynników oraz takiego algorytmu rozwiązania, który będzie operował jedynie niezerowymi elementami macierzy. Podprogramy realizujące algorytmy o wymienionych właściwościach przedstawia np. Szmelter (1980).

■ *Przykład 9.2*

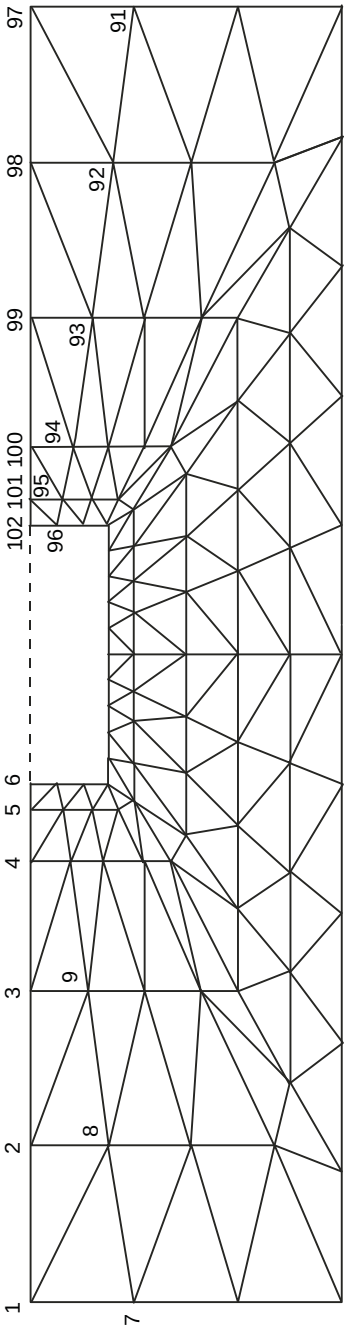
Rozwiązanie równania filtracji ustalonej pod budowlą piętrzącą metodą elementów skończonych

Rozwiążmy ponownie równanie Laplace'a (9.26) opisujące przepływ wody w jednorodnym ośrodku gruntowym pod budowlą piętrzącą, wymuszony znaną różnicą poziomów wody po jej obu stronach. Tym razem zastosujemy metodę elementów skończonych. Przekrój pionowy przez warstwę gruntu oraz budowlę przedstawiono na rys. 9.3. Przyjęto następujące wymiary obszaru (rys. 9.2.1):

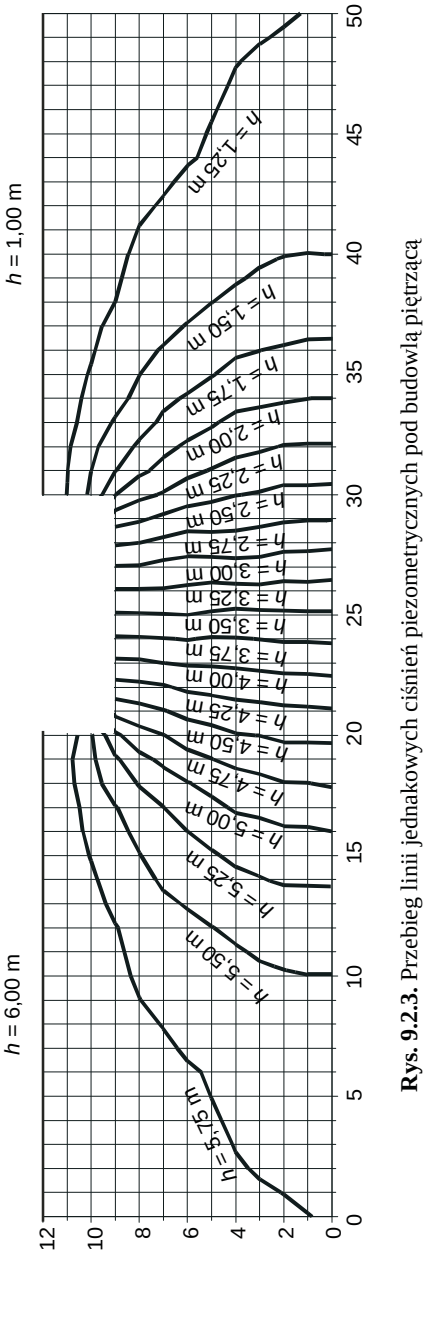
— miąższość warstwy filtracyjnej wynosi 12 m,
— przekrój pionowy fundamentu budowli ma kształt prostokąta o wymiarach 10×3 m,
— obszar aktywnej filtracji ma długość 50 m, przy czym zaczyna się on 20 m przed budowlą i kończy w odległości 20 m za budowlą.



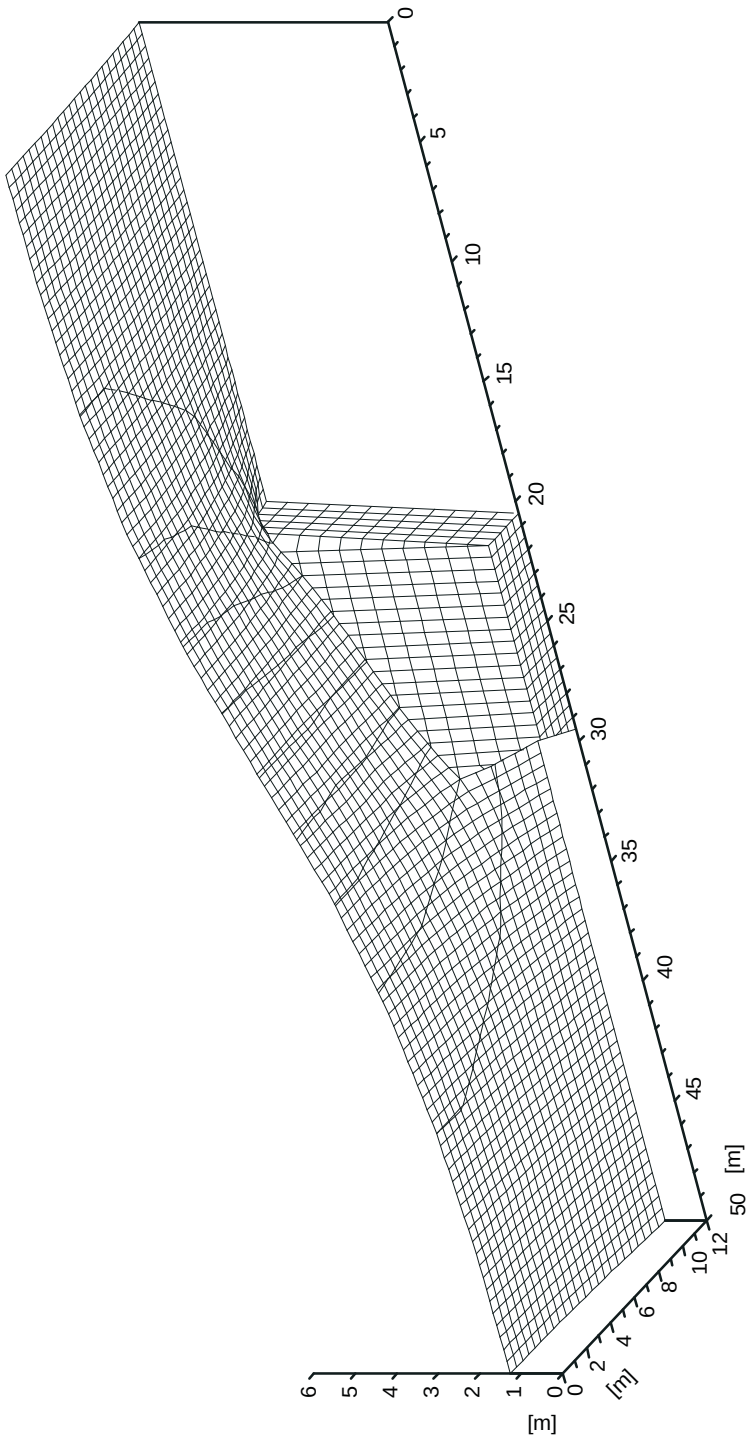
Rys. 9.2.1. Obszar rozwiązania równania Laplace'a i przyjęte warunki brzegowe



Rys. 9.2.2. Podział obszaru filtracji ustalonej pod jazem na elementy skończone



Rys. 9.2.3. Przebieg linii jednakowych ciśnień piezometrycznych pod budowlą piętrzącą



Rys. 9.2.4. Obliczona funkcja $h(x, y)$

Warunki brzegowe zadania są następujące (patrz rys. 9.2.1):
1) wzdłuż linii A – B od strony wody górnej $h = H_1 = 6$ m,
2) wzdłuż linii E – F od strony wody dolnej $h = H_2 = 1$ m,
3) wzdłuż podziemnego konturu budowli, czyli wzdłuż łamanej B – C – D – E: $\partial h/\partial n = 0$,
4) wzdłuż zewnętrznej granicy obszaru, tj. linii A – H – G – F: $\partial h/\partial n = 0$.

Symbol n oznacza normalną do brzegu, skierowaną na zewnątrz obszaru.

Wykonując podział obszaru filtracji na elementy skończone, wykorzystano istotną zaletę metody elementów skończonych w rozwiązywaniu zagadnień brzegowych, którą jest znaczna elastyczność aproksymacji obszaru rozwiązania. Przyjęto niejednorodną siatkę trójkątnych elementów (rys. 9.2.2). Wymiary elementów w rejonie znacznych gradientów ciśnienia piezometrycznego są mniejsze, zaś w rejonie małych gradientów – duże. Obszar rozwiązania tworzy siatka składająca się z $m = 160$ trójkątnych elementów skończonych zbudowanych na $M = 102$ węzłach. Wewnątrz elementów skończonych ciśnienie piezometryczne aproksymowane jest wyrażeniem (9.27). Przyjęte funkcje kształtu są funkcjami liniowymi o postaci (9.28).

Aproksymacja równania Laplace'a metodą elementów skończonych, jak to dokładnie opisano w podrozdziale 9.2, prowadzi do standardowego układu algebraicznych równań liniowych o postaci (9.43). Po wprowadzeniu warunków brzegowych układ ten rozwiązano metodą eliminacji Gaussa. Obliczony przebieg linii ekwipotencjalnych przedstawiono na rys. 9.2.3.

■

9.3. Rozwiązanie dwuwymiarowego równania filtracji nieustalonej ze swobodnym zwierciadłem

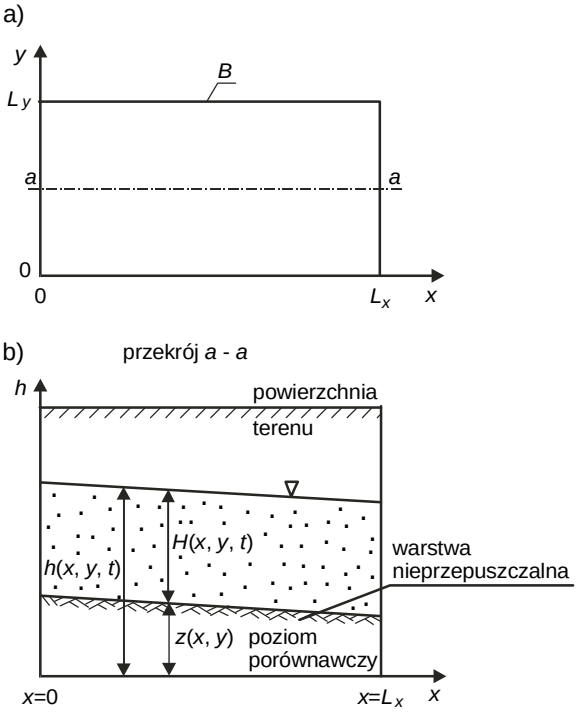
Równaniem, które opisuje przepływ wody w warstwie wodonośnej w sytuacji, gdy zwierciadło wody gruntowej jest zwierciadłem swobodnym, tzn. na jego powierzchni panuje ciśnienie atmosferyczne, jest równanie Boussinesqu'a postaci (7.2)

$$\mu \frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left[k(h-z)\frac{\partial h}{\partial x}\right] + \frac{\partial}{\partial y}\left[k(h-z)\frac{\partial h}{\partial y}\right] + w, \tag{9.48}$$

gdzie: $h(x, y, t)$ – rzędna zwierciadła wody gruntowej,
$z(x, y)$ – rzędna spągu warstwy wodonośnej,
k – współczynnik filtracji,
μ – porowatość efektywna,
w – człon źródłowy reprezentujący zasilanie zewnętrzne.

Na rys. 9.5 przedstawiono przekrój poprzeczny warstwy, wraz ze stosowanymi oznaczeniami.

Ponieważ współczynniki równania, zależne od rozwiązania $h(x, y, t)$, występują pod znakiem pochodnej, równanie to jest równaniem nieliniowym. Rozwiążmy prostszy przypadek filtracji opisanej zlinearyzowanym równaniem (9.48). W tym celu załóżmy, że:
— zmiana miąższości warstwy wodonośnej $H(x, y, t)$ jest na tyle mała, iż możliwe jest założenie $H = h - z = \overline{H} = \text{const}$;
— porowatość efektywna oraz współczynnik filtracji są stałe, czyli $\mu = \overline{\mu} = \text{const}$ oraz $k = \overline{k} = \text{const}$;
— nie ma zasilania zewnętrznego, czyli $w = 0$.



Rys. 9.5. Geometria obszaru filtracji i przekrój pionowy warstwy wodonośnej

Uwzględnienie powyższych założeń pozwala zapisać równanie (9.48) w następującej prostszej postaci:

$$\frac{\partial h}{\partial t} = D\left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right), \tag{9.49}$$

gdzie

$$D = \frac{\overline{k}\overline{H}}{\overline{\mu}} = \text{const} \tag{9.50}$$

jest współczynnikiem dyfuzji. Występujące w (9.50) symbole oznaczają uśrednione w obszarze filtracji wartości parametrów. Powyższe równanie jest dwuwymiarowym liniowym równaniem dyfuzji. Rozwiązanie równania tego typu różnymi wariantami metody różnic skończonych przedstawiono w podrozdziale 8.4.

Chcąc rozwiązać je metodą elementów skończonych, najpierw obszar filtracji pokrywamy siatką trójkątów. Wybierzmy dowolny element *e* utworzonej siatki węzłów przedstawiony na rys. 9.4. Przyjmijmy, że wewnątrz trójkątnego elementu skończonego zbudowanego na węzłach *i*, *j*, *k*, potencjał *h* aproksymowany jest według (7.80) wyrażeniem

$$\widehat{h} = \mathbf{Nh} = N_i(x, y)h_i + N_j(x, y)h_j + N_k(x, y)h_k\,, \tag{9.51}$$

gdzie aproksymujące funkcje kształtu $N(x, y)$ są funkcjami liniowymi postaci (7.81)

$$N_p(x, y) = \frac{1}{2F_e}(a_p + b_p x + c_p y); \quad (p = i, j, k)\,. \tag{9.52}$$

W powyższym wzorze F_e oznacza pole elementu, natomiast współczynniki a_p, b_p, c_p, zależne od współrzędnych węzłów elementu, obliczane są według formuł (7.83), (7.84) i (7.85).

Zastosujmy teraz do równania (9.49) procedurę Galerkina. Otrzymamy

$$\iint_F \left[\frac{\partial h}{\partial t} - D\left(\frac{\partial^2 \hat{h}}{\partial x^2} + \frac{\partial^2 \hat{h}}{\partial y^2} \right) \right] N_l dxdy = 0, \tag{9.53}$$

gdzie: F – obszar rozwiązania,
$l = 1, 2, 3 ..., M$ – indeks węzła,
M – liczba węzłów w obszarze rozwiązania F.

Powyższą zależność przepiszmy w nieco zmienionej postaci:

$$\iint_F \frac{\partial h}{\partial t} N_l dxdy - D \iint_F \left(\frac{\partial^2 \hat{h}}{\partial x^2} + \frac{\partial^2 \hat{h}}{\partial y^2} \right) N_l dxdy = 0, \quad (l = 1,\ 2,\ ...,\ M)\,. \tag{9.54}$$

Wyrażenie (9.54) zawiera pochodne drugiego rzędu, co wymaga ciągłości pierwszej pochodnej funkcji kształtu we wszystkich punktach obszaru, także przy przejściu od elementu do elementu. Tymczasem do aproksymacji przyjęto liniowe funkcje bazowe. Ograniczenie to możemy ominąć, obniżając rząd pochodnych, podobnie jak to zrobiliśmy, rozwiązując równanie Laplace'a. W tym wypadku wygodnie jest użyć przekształcenia Greena o ogólnej postaci (9.30). Zatem warunek Galerkina (9.54) zapiszemy:

$$\begin{aligned} &\iint_F \frac{\partial \hat{h}}{\partial t} N_l\, dxdy - D \iint_F \left(\frac{\partial^2 \hat{h}}{\partial x^2} + \frac{\partial^2 \hat{h}}{\partial y^2} \right) N_l\, dxdy = \iint_F \frac{\partial \hat{h}}{\partial t} N_l\, dxdy + \\ &\quad - D \iint_F \left(N_l \frac{\partial^2 \hat{h}}{\partial x^2} + N_l \frac{\partial^2 \hat{h}}{\partial y^2} \right) dxdy = \iint_F \frac{\partial \hat{h}}{\partial t} N_l dxdy + \\ &+ D \iint_F \left(\frac{\partial N_l}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial N_l}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) dxdy - D \int_B N_l \frac{\partial \hat{h}}{\partial n} dB = 0, (l = 1,\ 2,\ \cdots,\ M). \end{aligned} \tag{9.55}$$

W rozwiązywanym tutaj zagadnieniu warunki zostały sformułowane tak, że na części brzegu dana jest funkcja h, a na części brzegu pochodna tej funkcji normalna do brzegu jest równa zero. Oznacza to, że na odcinkach brzegu równoległych do osi y obowiązuje warunek:

$$\frac{\partial h}{\partial x} = 0 = \frac{\partial h}{\partial n},$$

a na części brzegu równoległej do osi x obowiązuje warunek:

$$\frac{\partial h}{\partial y} = 0 = \frac{\partial h}{\partial n}.$$

Zatem dla przyjętych warunków brzegowych w tym przypadku na całym obwodzie B obszaru F będzie

$$\int_B N_l \frac{\partial h}{\partial n} dB = 0\,.$$

Całkę tę można więc pominąć w rozważaniach. Uwzględnienie tego faktu pozwala sprowadzić równanie (9.55) do postaci:

$$\iint_F \frac{\partial \hat{h}}{\partial t} N_l \, dxdy + D \iint_F \left(\frac{\partial N_l}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial N_l}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) dxdy = 0, \quad (l = 1, 2, \cdots, M) . \tag{9.56}$$

Korzystając z zasady sumowania całek w elementach tworzących obszar F, możemy napisać

$$\begin{aligned} &\iint_F \left[\frac{\partial \hat{h}}{\partial t} \mathbf{N} + D \left(\frac{\partial \mathbf{N}}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial \mathbf{N}}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) \right] dxdy = \\ &= \sum_{e=1}^{m} \iint_{F_e} \left[\frac{\partial \hat{h}}{\partial x} \mathbf{N} + D \left(\frac{\partial \mathbf{N}}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial \mathbf{N}}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) \right] dxdy = 0, \end{aligned} \tag{9.57}$$

gdzie: F_e – pole elementu e,
m – liczba elementów,
$\mathbf{N} = (N_1, ..., N_i, ..., N_j, ..., N_k, ..., N_M)^T$ – wektor funkcji kształtu.

Spośród m elementów pokrywających obszar F wybierzmy jeden element o wierzchołkach oznaczonych indeksami i, j oraz k przedstawiony na rys. 9.4. W elemencie tym tylko 3 składowe wektora $\mathbf{N}$, a mianowicie N_i, N_j oraz N_k są niezerowe. Wszystkie pozostałe funkcje kształtu w elemencie e przyjmują wartości zerowe. Zatem dla elementu e należy obliczyć tylko 3 następujące całki:

$$I_l = \iint_{F_e} \left[\frac{\partial \hat{h}}{\partial t} N_l + D \left(\frac{\partial N_l}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial N_l}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) \right] dxdy, \quad (l = i, j, k) . \tag{9.58}$$

Policzmy powyższą całkę dla $l = i$.

$$\begin{aligned} I_i &= \iint_{F_e} \left[\frac{\partial \hat{h}}{\partial t} N_i + D \left(\frac{\partial N_i}{\partial x} \frac{\partial \hat{h}}{\partial x} + \frac{\partial N_i}{\partial y} \frac{\partial \hat{h}}{\partial y} \right) \right] dxdy = \\ &= \iint_{F_e} \frac{\partial}{\partial t} (N_i h_i + N_j h_j + N_k h_k) N_i dxdy + \\ &+ D \iint_{F_e} \left(\frac{\partial N_i}{\partial x} \frac{\partial}{\partial x} (N_i h_i + N_j h_j + N_k h_k) + \frac{\partial N_i}{\partial y} \frac{\partial}{\partial y} (N_i h_i + N_j h_j + N_k h_k) \right) dxdy = \\ &= \frac{dh_i}{dt} \iint_{F_e} N_i^2 dxdy + \frac{dh_j}{dt} \iint_{F_e} N_j N_i dxdy + \frac{dh_k}{dt} \iint_{F_e} N_k N_i dxdy + \\ &+ Dh_i \iint_{F_e} \left(\frac{\partial N_i}{\partial x} \frac{\partial N_i}{\partial x} + \frac{\partial N_i}{\partial y} \frac{\partial N_i}{\partial y} \right) dxdy + Dh_j \iint_{F_e} \left(\frac{\partial N_i}{\partial x} \frac{\partial N_j}{\partial x} + \frac{\partial N_i}{\partial y} \frac{\partial N_j}{\partial y} \right) dxdy + \\ &+ Dh_k \iint_{F_e} \left(\frac{\partial N_i}{\partial x} \frac{\partial N_k}{\partial x} + \frac{\partial N_i}{\partial y} \frac{\partial N_k}{\partial y} \right) dxdy. \end{aligned} \tag{9.59}$$

Obliczenie kolejnych członów powyższego wyrażenia pozwala zapisać je następująco:

$$I_i = \frac{F_e}{6}\frac{dh_i}{dt} + \frac{F_e}{12}\frac{dh_j}{dt} + \frac{F_e}{12}\frac{dh_k}{dt} + \\ + \frac{D}{4F_e}(b_i b_i + c_i c_i)h_i + \frac{D}{4F_e}(b_i b_j + c_i c_j)h_j + \frac{D}{4F_e}(b_i b_k + c_i c_k)h_k. \tag{9.60}$$

Podobne wyrażenia otrzymamy, obliczając całkę (9.58) dla pozostałych funkcji kształtu w elemencie *e*. I tak dla N_j uzyskamy

$$I_j = \frac{F_e}{12}\frac{dh_i}{dt} + \frac{F_e}{6}\frac{dh_j}{dt} + \frac{F_e}{12}\frac{dh_k}{dt} + \\ + \frac{D}{4F_e}(b_j b_i + c_j c_i)h_i + \frac{D}{4F_e}(b_j b_j + c_j c_j)h_j + \frac{D}{4F_e}(b_j b_k + c_j c_k)h_k, \tag{9.61}$$

zaś dla N_k będziemy mieli

$$I_k = \frac{F_e}{12}\frac{dh_i}{dt} + \frac{F_e}{12}\frac{dh_j}{dt} + \frac{F_e}{6}\frac{dh_k}{dt} + \\ + \frac{D}{4F_e}(b_k b_i + c_k c_i)h_i + \frac{D}{4F_e}(b_k b_j + c_k c_j)h_j + \frac{D}{4F_e}(b_k b_k + c_k c_k)h_k. \tag{9.62}$$

Zapiszmy otrzymane przez scałkowanie (9.58) zależności (9.60), (9.61) i (9.62), stosując notację macierzową:

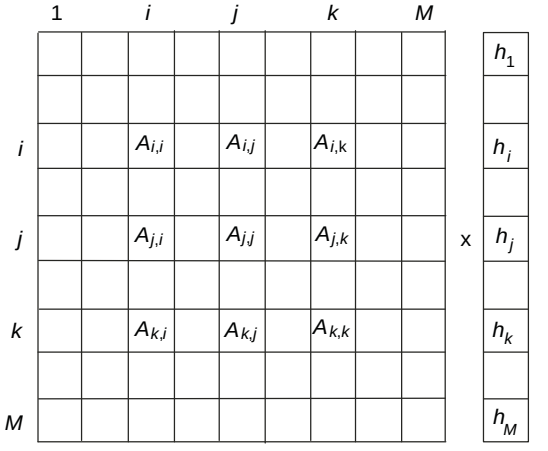
$$\iint_{F_e}\left[\frac{\partial\widehat{h}}{\partial t}\mathbf{N} + D\left(\frac{\partial\mathbf{N}}{\partial x}\frac{\partial\widehat{h}}{\partial x} + \frac{\partial\mathbf{N}}{\partial y}\frac{\partial\widehat{h}}{\partial y}\right)\right]dxdy = \mathbf{S}_e\frac{d\mathbf{h}}{dt} + \mathbf{A}_e\mathbf{h}\,. \tag{9.63}$$

Występujące po prawej stronie powyższej zależności iloczyny mają następującą postać:

— $\mathbf{S}_e d\mathbf{h}/dt$

	1		*i*		*j*		*k*		*M*		
											dh_1/dt
i			$S_{i,i}$		$S_{i,j}$		$S_{i,k}$				dh_i/dt
j			$S_{j,i}$		$S_{j,j}$		$S_{j,k}$			x	dh_j/dt
k			$S_{k,i}$		$S_{k,j}$		$S_{k,k}$				dh_k/dt
M											dh_M/dt

— $\mathbf{A}_e\mathbf{h}$



Zgodnie z (9.60), (9.61) i (9.62) elementy macierzy $\mathbf{S}_e$ oraz $\mathbf{A}_e$ są zdefiniowane następująco:

$$S_{i,j} = S_{j,j} = S_{k,k} = \frac{F_e}{6}, \tag{9.64a}$$

$$S_{j,i} = S_{i,j} = S_{k,i} = S_{i,k} = S_{k,j} = S_{j,k} = \frac{F_e}{12}, \tag{9.64b}$$

$$A_{i,i} = \frac{1}{4F_e}(b_i b_i + c_i c_i), \tag{9.65a}$$

$$A_{i,j} = \frac{1}{4F_e}(b_i b_j + c_i c_j), \tag{9.65b}$$

$$A_{i,k} = \frac{1}{4F_e}(b_i b_k + c_i c_k), \tag{9.65c}$$

$$A_{j,i} = \frac{1}{4F_e}(b_j b_i + c_j c_i), \tag{9.65d}$$

$$A_{j,j} = \frac{1}{4F_e}(b_j b_j + c_j c_j), \tag{9.65e}$$

$$A_{j,k} = \frac{1}{4F_e}(b_j b_k + c_j c_k), \tag{9.65f}$$

$$A_{k,i} = \frac{1}{4F_e}(b_k b_i + c_k c_i), \tag{9.65g}$$

$$A_{k,j} = \frac{1}{4F_e}(b_k b_j + c_k c_j), \tag{9.65h}$$

$$A_{k,k} = \frac{1}{4F_e}(b_k b_k + c_k c_k). \tag{9.65i}$$

Zauważmy, że macierze $\mathbf{S}_e$ oraz $\mathbf{A}_e$ są symetryczne.

W podobny sposób obliczamy całki dla każdego elementu e = 1, 2, ..., m, otrzymując zależności typu (9.63). Podstawmy wynik obliczeń do równania (9.57). Otrzymamy

$$\sum_{e=1}^{m} \iint_{F_e} \left(\frac{\partial h}{\partial t}\mathbf{N} + D\frac{\partial \mathbf{N}}{\partial x}\frac{\partial \widehat{h}}{\partial x} + \frac{\partial \mathbf{N}}{\partial y}\frac{\partial \widehat{h}}{\partial y} \right) dxdy = \sum_{e=1}^{m} \left(\mathbf{S}_e \frac{\partial \mathbf{h}}{\partial t} + \mathbf{A}_e \mathbf{h} \right) = 0 \,. \tag{9.66}$$

Sumując rezultaty całkowania w każdym elemencie, uzyskujemy następujące wyrażenie:

$$\mathbf{S}\frac{d\,\mathbf{h}}{dt} + \mathbf{A}\mathbf{h} = \mathbf{0} \,, \tag{9.67}$$

w którym

$$\mathbf{S} = \sum_{e=1}^{m} \mathbf{S}_e \,, \quad \mathbf{A} = \sum_{e=1}^{m} \mathbf{A}_e \,. \tag{9.68a,b}$$

Jest to układ równań różniczkowych zwyczajnych. Macierze **S** oraz **A**, podobnie jak ich składowe $\mathbf{S}_e$ i $\mathbf{A}_e$, są pasmowe i symetryczne.

Całkowanie tego układu względem czasu jest zagadnieniem początkowym układu równań różniczkowych zwyczajnych. Do jego rozwiązania zastosujmy opisaną w podrozdziale 6.2.3 niejawną metodę trapezową (6.40). Sposób rozwiązania układu równań różniczkowych tą metodą jest identyczny z opisanym w podrozdziale 9.1 sposobem rozwiązania w odniesieniu do jednowymiarowego równania dyfuzji.

Metoda trapezowa

$$\mathbf{h}_{t+\Delta t} = \mathbf{h}_t + \frac{\Delta t}{2}\left(\frac{d\,\mathbf{h}_t}{dt} + \frac{d\,\mathbf{h}_{t+\Delta t}}{dt} \right) \tag{9.69}$$

gdzie: Δt jest krokiem całkowania,

po zastąpieniu pochodnych wyrażeniem wynikającym z (9.67) prowadzi do układu równań:

$$(\mathbf{S} + 0{,}5\Delta t\mathbf{A})\,\mathbf{h}_{t+\Delta t} = (\mathbf{S} - 0{,}5\Delta t\mathbf{A})\,\mathbf{h}_t \,. \tag{9.70}$$

Jest to układ algebraicznych równań liniowych, którego macierz współczynników jest stała pasmowa i symetryczna.

Do układu (9.70) należy wprowadzić przyjęte na etapie formułowania problemu rozwiązania, warunki brzegowe. W tym przypadku są to warunki mieszane. Dodajmy, że warunki typu Neumanna zadane na częściach nieprzepuszczalnych brzegu B zostały już uwzględnione w całce krzywoliniowej, co pozwoliło zapisać równanie (9.54) w postaci (9.55). Z kolei, warunek typu Dirichleta oznacza, że w węzłach leżących na brzegu obszaru znane są wartości funkcji h. Oznacza to, że w wektorze **h** znane są te jego składowe, które odpowiadają węzłom brzegowym z takim warunkiem. Zatem z układu (9.70) należy wyeliminować wymienione składowe, jak i równania odpowiadające im. W rezultacie tego otrzymamy układ o wymiarze równym liczbie węzłów, w których nie jest znana funkcja h. Ponieważ postępowanie takie jest kłopotliwe, wygodniej jest zastosować opisany w podrozdziale 9.2 sposób uwzględnienia warunków brzegowych, niewymagający zmiany struktury macierzy współczynników i wektora niewiadomych.

W chwili $t = t_0$ znany jest wektor $\mathbf{h}_{t=t_0}$, reprezentujący początkowe położenie swobodnego zwierciadła w całym obszarze rozwiązania. Można więc dobrze obliczyć prawą stronę (9.70), a następnie rozwiązać układ algebraicznych równań liniowych. Powtarzając opisaną

procedurę dla czasu zmieniającego się z interwałem Δt, czyli dla $t = t_0 + \Delta t$, $t_0 + 2\Delta t$, $t_0 + 3\Delta t$, ... aż do momentu, gdy spełniony będzie warunek:

$$t > t_{max}, \tag{9.72}$$

gdzie: t_{max} – zadany czas odtwarzania procesu filtracji,

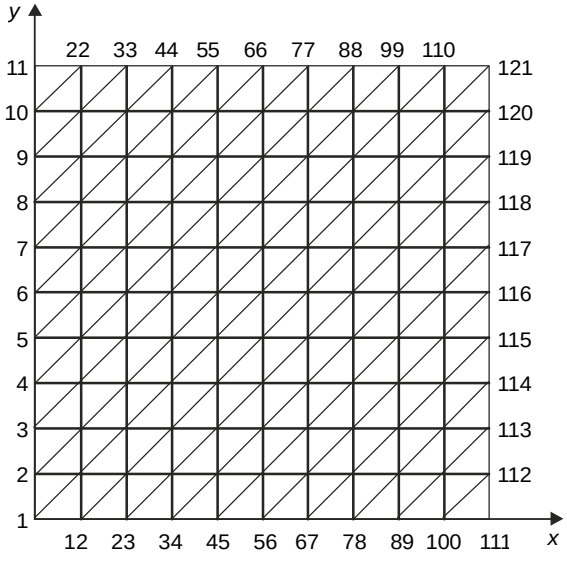
uzyskamy rozkłady ciśnień piezometrycznych **h**, czyli położenia swobodnego zwierciadła w przedziale $\langle 0, t_{max} \rangle$.

Ważnym zagadnieniem w trakcie numerycznego rozwiązywania równań typu parabolicznego jest dobór wielkości dopuszczalnego kroku czasowego Δt decydującego o stabilności metody rozwiązania. Jeśli równanie jest – jak w tym przypadku – liniowe, to dla przyjętego schematu trapezowego (niejawnego) teoretycznie nie ma ograniczeń nałożonych na krok czasowy.

■ *Przykład 9.3*

Filtracja nieustalona wywołana pracą studni

Opisaną wyżej procedurę zastosowano do rozwiązania równania (9.49) w obszarze prostokątnym o wymiarach 3000×3000 m (rys. 9.3.1).



Rys. 9.3.1. Dyskretyzacja obszaru całkowania równania (7.49)

Aproksymację przestrzenną wykonano 200 trójkątnymi elementami skończonymi, zbudowanymi na 121 węzłach siatki kwadratowej o wymiarach 300×300 m. Warstwa wodonośna jest ograniczona od dołu warstwą nieprzepuszczalną, poziomo zalegającą na rzędnej z = 10 m. Warunki graniczne są następujące:

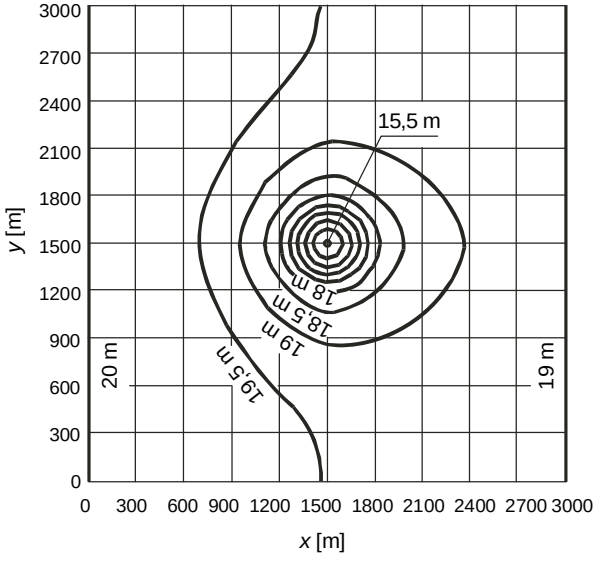
— warunek początkowy: poziom zwierciadła wody w chwili $t_0 = 0$ jest stały i układa się na rzędnej $h = 20$ m w całym obszarze;
— warunki brzegowe (rys. 9.3.1):
  a) na prawym brzegu obszaru (linia 111–121) zadaje się stałą rzędną $h_i = 19$ m ($i = 111, 112, ..., 121$),
  b) na linii 1 – 11 zadaje się stałą rzędną $h_i = 20$ m ($i = 1, 2, ..., 11$),
  c) w węźle nr 61 istnieje studnia zupełna, w której zmiana poziomu zwierciadła wody wywołana poborem wody opisana jest funkcją $f_s(t)$ o postaci

$$f_s(t) = \begin{cases} 19{,}5 \text{ m} & \text{dla} \quad t = 0, \\ 19{,}5\dfrac{4{,}0}{T_s}t & \text{dla} \quad 0 < t \le T_s, \\ 15{,}5 \text{ m} & \text{dla} \quad t > T_s, \end{cases}$$

  gdzie $T_s = 20$ h,

  d) brzegi wzdłuż linii 1–111 i 11–121 są nieprzepuszczalne.
  Warstwa wodonośna jest jednorodna, o następujących własnościach filtracyjnych:
— współczynnik filtracji $k = 25$ m/h,
— współczynnik porowatości $\mu = 0{,}35$.

Proces przepływu odtwarzano z krokiem czasowym $\Delta t = 1{,}0$ h. Stan ustalony przepływu osiągnięty został po około 5 dobach trwania pompowania w studni. Rozkład hydroizohips w obszarze, po osiągnięciu stanu ustalonego, pokazano na rys. 9.3.2.



Rys. 9.3.2. Układ hydroizohips wywołany pracą studni po 120 godzinach ■

Bibliografia

[1] Abbott M. B.: *Computational Hydraulics-Elements of the theory of free surface flow*. London: Pitman 1979.
[2] Abbott M. B., Basco D. R.: *Computational fluid dynamics.* Longmann Scientific and Technical 1989.
[3] Adamski W.: *Modelowanie systemów oczyszczania wód*. Wydawnictwo Naukowe PWN, Warszawa 2002
[4] Burzyński K., Granatowicz J., Piwecki T., Szymkiewicz R.: *Metody numeryczne w hydrotechnice*. Gdańsk: Wydawnictwo Politechniki Gdańskiej 1991.
[5] Chang S. C.: *The method of space-time conservation element and solution element – A new approach for solving the Navier-Stokes and Euler equations*. Journal of Computational Physics, No 119, 1995.
[6] Chapra SC, Canale RP *Numerical methods for engineers*. McGraw-Hill 2006.
[7] Chow V. T.: *Handbook of applied hydrology*. New York: McGraw–Hill 1964.
[8] Chow V. T., Maidment D. R., Mays L. W.: *Applied hydrology*. New York: Mc Graw-Hill 1988.
[9] Cunge J., Holly Jr F. M., Verwey A.: *Practical Aspects of Computational River Hydraulics*. London: Pitman 1980.
[10] Czetwertyński E., Utrysko B.: *Hydraulika i hydromechanika*. Warszawa: PWN 1969.
[11] Dahlquist G., Bjorck A.: *Metody numeryczne*. Warszawa: PWN 1983.
[12] Demidowicz B. P., Maron I. A., Szuwałowa E. Z.: *Metody numeryczne*, cz. I, II. Warszawa: PWN 1965.
[13] Dryja M., Jankowscy J. i M.: *Przegląd metod i algorytmów numerycznych*, cz. 2. Warszawa: WNT 1981.
[14] Dziubiński I., Świątkowski T.: *Poradnik matematyczny*. Warszawa: PWN 1980.
[15] Eagelson P. S.: *Hydrologia dynamiczna*. Warszawa: PWN 1978.
[16] Fletcher C. A.: *Computational techniques for fluid dynamics.* Vol. 1. Berlin: Springer Verlag 1991.
[17] Fortuna Z., Macukow B., Wąsowski J.: *Metody numeryczne*. Warszawa: WNT 1982.
[18] French R. H.: *Open channel hydraulics*. New York: Mc Graw Hill 1985.
[19] Godunow C. K.: *Równania fizyki matematycznej*. Warszawa: WNT 1975.
[20] Jankowscy J. i M.: *Przegląd metod i algorytmów numerycznych*, cz. 1. Warszawa: WNT 1981.
[21] Kacprzyński B.: *Planowanie eksperymentów – podstawy matematyczne*. Warszawa: WNT 1974.
[22] Kącki E.: *Równania różniczkowe cząstkowe w zagadnieniach fizyki i techniki*. Warszawa: WNT 1992.
[23] Korn G.A., Korn T.M.: *Matematyka dla pracowników naukowych i inżynierów*, cz. 1, 2. Warszawa: PWN 1983.
[24] Kot A., Szymkiewicz R.: *Uproszczone liniowe modele transformacji fal w korycie rzecznym*. Warszawa: Monografie Komitetu Gospodarki Wodnej PAN 2002.
[25] Krupowicz A.: *Metody numeryczne zagadnień początkowych równań różniczkowych zwyczajnych*. Warszawa: PWN 1986.
[26] Kusiak J., Danielewska-Tułecka A., Oprocha J.: *Optymalizacja. Wybrane metody z przykładami zastosowań*. Warszawa: Wydawnictwo Naukowe PWN 2009.
[27] Lambor J.: *Hydrologia*. Warszawa: Arkady 1971.
[28] Legras J.: *Praktyczne metody analizy numerycznej*. Warszawa: WNT 1974.

[29] Liggett J. A., Cunge, J. A.: *Numerical method of solution of the unsteady flow equations*. W: K. Mahmood and V. Yevjevich, Unsteady Flow in Open Channels. Fort Collins, Colorado, USA: Water Resources Publ. 1975.
[30] Marczuk G. I.: *Analiza numeryczna zagadnień fizyki matematycznej*. Warszawa: PWN 1983.
[31] Michlin S. G., Smolicki C. L.: *Metody przybliżone rozwiązywania równań różniczkowych i całkowych*. Warszawa: PWN 1972.
[32] Miller W. A., Cunge J. A.: *Simplified equations of unsteady flow*. W: Miller W.A., Yevjewich V.: Unsteady flow in open channels. Fort Collins, Colorado, USA: Water Resources Publishing 1975.
[33] Mitosek M.: *Mechanika płynów w inżynierii środowiska*. Warszawa: Wydawnictwa Naukowe PWN 2001.
[34] Parmakian J.. *Water hammer analysis*. New York: Prentice-Hall, INC 1955.
[35] Patankar S. V.: *Numerical Heat Transfer and Fluid Flow*. Washington: Hemisphere Publishing Corporation, McGraw-Hill 1980.
[36] Piwecki T., Sokólski Z.: *Experimental determination of overrelaxation coefficients for the Laplace equation in cylinder coordinates with mixed boundary conditions*. Algorytmy, no 9, 1966.
[37] Potter D.: *Metody obliczeniowe fizyki*. Warszawa: PWN 1977.
[38] Press W. H., Teukolsky S. A., Veterling W. T., Flannery B. P.: *Numerical Recipes* in C. Cambridge University Press 1992.
[39] Ralston A.: *Wstęp do analizy numerycznej*. Warszawa: PWN 1971.
[40] Remson I., Hornberger G., Molz F.: *Numerical methods in subsurface hydrology*. New York: J. Wiley 1971.
[41] Sawicki J. M.: *Przepływy ze swobodną powierzchnią*. Warszawa: Wydawnictwo Naukowe PWN 1998.
[42] Sawicki A.: *Mechanika kontinuum – wprowadzenie*. Gdańsk: Wydawnictwo IBW PAN 1994.
[43] Stoer J.: *Wstęp do metod numerycznych*, t. 1. Warszawa: PWN 1979.
[44] Stoer J., Bulirsh R.: *Wstęp do metod numerycznych*, t. 2. Warszawa: PWN 1979.
[45] Streeter V. L., Lai Ch.. *Water Hammer Analysis Including Fluid Friction*. Journal of Hydraulics Division ASCE, Vol. 88, HY3, 1962.
[46] Szmelter J.: *Metody komputerowe w mechanice*. Warszawa: PWN 1980.
[47] Szymkiewicz R.: *Modelowanie matematyczne przepływów w rzekach i kanałach*. Warszawa: Wydawnictwo Naukowe PWN 2000.
[48] Szymkiewicz R.: *Numerical methods in open channel hydraulics*. Springer 2010.
[49] Szymkiewicz R.: *Numeryczna dyssypacja w rozwiązaniach równań hiperbolicznych. Zastosowanie mechaniki w budownictwie lądowym i wodnym*. Gdańsk: Wyd. IBW-PAN 2001.
[50] Szymkiewicz R.: *Boundary problem for equations of steady gradually varied flow in open channels*. Archives of Hydro-Engineering and Environmental Mechanics, Vol. 47, no 1–4, 2000.
[51] Szymkiewicz R.: *Numerical modeling in open channal hydraulies*. Springer 2010.
[52] Tan Weiyan: *Shallow water hydrodynamics*. Amsterdam: Elsevier 1992.
[53] Warming R.F., Hyett B.J.: *The modified equation approach to the stability and accuracy analysis of finite – difference method*. Journal of Computational Physics, Vol. 14, 1974.
[54] Wichowski R.: *Wybrane zagadnienia przepływów nieustalonych w sieci wodociągowej pierścieniowej*. Gdańsk: Wyd. Politechniki Gdańskiej, Monografie 27, 2002.
[55] Zienkiewicz O. C.: *Metoda elementów skończonych*. Warszawa: Arkady 1972.
[56] Ziółko M.: *Modelowanie zjawisk falowych*. Kraków: Wydawnictwa Naukowo-Dydaktyczne AGH 2000.

ZAŁĄCZNIK

Błędy w obliczeniach

Jeśli x jest dokładną wartością pewnej wielkości, zaś x' jest jej wartością przybliżoną, to wyrażenie

$$\varepsilon = |x - x'| \tag{Z.1}$$

nazywamy błędem bezwzględnym, zaś wyrażenie

$$\varepsilon' = \left| \frac{x' - x}{x} \right| \tag{Z.2}$$

nazywa się błędem względnym.

W trakcie obliczeń wykonywanych na maszynach cyfrowych występują 3 podstawowe rodzaje błędów:
1) błędy wejściowe,
2) błędy obcięcia,
3) błędy zaokrągleń.

Błędy wejściowe (lub błędy danych wejściowych) występują wówczas, gdy dane liczbowe wprowadzane do pamięci maszyny cyfrowej odbiegają od dokładnych wartości tych danych. Błędy takie pojawiają się np. wtedy, gdy dane wejściowe są wynikiem pomiarów wielkości fizycznych, mierzonych z pewną dokładnością pomiaru. Jednak bardziej typową przyczyną występowania błędów wejściowych jest skończona długość słów binarnych reprezentujących liczby w maszynie.

Pamięć maszyny cyfrowej podzielona jest na fragmenty nazywane słowami. Każde słowo zawiera tę samą liczbę cyfr dziesiętnych oraz znak. W słowie zajmującym D cyfr są one podzielone na dwie oddzielne części. W większości współczesnych maszyn cyfrowych liczby rzeczywiste x są zapamiętywane w tzw. postaci zmiennoprzecinkowej

$$x = M \cdot N^{w}, \tag{Z.3}$$

gdzie: M – mantysa liczby x,
W – wykładnik części potęgowej, czyli tzw. cecha.

Podstawa części potęgowej N jest na ogół, chociaż nie zawsze, równa 2. Cyfry w układzie dwójkowym nazywa się bitami. Zatem w zapisie zmiennoprzecinkowym liczba rzeczywista jest przedstawiona za pomocą dwu grup bitów. Grupa pierwsza, tworząca mantysę M, jest interpretowana jako część ułamkowa. W większości maszyn cyfrowych jest spełniony warunek (Dahlquist i Bjorck, 1983):

$$0{,}1 \le |M| < 1. \tag{Z.4}$$

Natomiast grupa druga, tworząca wykładnik W, jest interpretowana jako liczba całkowita.

W maszynie cyfrowej liczba cyfr służących do reprezentacji *M* oraz *W* jest ograniczona. Wynika stąd skończony zbiór liczb, które mogą być reprezentowane w maszynie. Liczby z tego zbioru nazywa się liczbami zmiennoprzecinkowymi. Ograniczona liczba cyfr cechy implikuje zakres zmienności liczb. Zatem zakres liczb *W* decyduje o zakresie liczb rzeczywistych dopuszczalnych w danej maszynie. Trzeba więc pamiętać o możliwości wystąpienia w trakcie obliczeń takich liczb, w których cecha będzie większa od dopuszczalnej. Wystąpi wtedy tzw. nadmiar.

Rozpatrzy przykład przedstawiony przez Fortunę, Macukowa i Wąsowskiego (1982). Załóżmy, że w zapisie dwójkowym liczbę *M* określa pięć bitów, a liczbę *W* określają trzy bity, przy czym pierwsze bity oznaczają znaki tych liczb. Liczbę ujemną rozpoczyna bit o wartości 1, zaś dodatnią – o wartości 0. W tym wypadku zapis

$$x = \underbrace{(1)1101}_{M}\ \underbrace{(0)10}_{W}$$

w zapisie dwójkowym oznacza liczbę

$$-0.1101 \cdot 2^{+10} = -\left(\frac{1}{2}+\frac{1}{4}+\frac{0}{8}+\frac{1}{16}\right)\cdot 2^{+(1\cdot 2+0\cdot 1)},$$

która w zapisie dziesiętnym jest liczbą –3,25.

Przyjęte powyżej słowo (5 bitów + 3 bity) pozwala utworzyć tylko niektóre liczby dodatnie w zakresie od 0,0625 do 7,5, liczbę 0 oraz liczby ujemne w zakresie –0,0625 do –7,5. Widać z tego, że są liczby rzeczywiste, których za pomocą takiego zapisu nie można dokładnie przedstawić. Na przykład liczba $x = 0.2$ (w zapisie dziesiętnym) ma w zapisie dwójkowym nieskończone rozwinięcie równe 0,0011(0011) ... . Najbliższa jej liczba, którą można przedstawić, stosując przyjęty wzorzec, czyli 5-bitową mantysę i 3-bitowy wykładnik, to liczba

$$x' = (0)1100(1)10 = 0{,}1875.$$

Zatem błąd bezwzględny (Z.1) wynosi 0,0125, zaś błąd względny (Z.2) jest równy $\varepsilon' = -0{,}0625$. Jak wynika z powyższego przykładu, liczby w zapisie dziesiętnym, przy przejściu na reprezentację maszynową (w tym przypadku w systemie dwójkowym) są przedstawiane z pewnym przybliżeniem. Przybliżenie to jest tym lepsze, im więcej bitów zostanie przeznaczonych na mantysę liczby. Działania arytmetyczne na liczbach o długości jednego słowa nazywamy działaniami z pojedynczą precyzją (dokładnością). Niekiedy pojedyncza precyzja jest niewystarczająca i działania arytmetyczne należy wykonywać na liczbach dłuższych. W takiej sytuacji można wykorzystać tzw. arytmetykę w podwójnej precyzji (w języku angielskim: double precision). W większości maszyn cyfrowych możliwość taka istnieje. Powoduje to zwiększenie dokładności, czyli zmniejszenie ε.

Należy dodać, że wstępne zaokrąglenie występuje nie tylko w trakcie wprowadzania do pamięci komputera liczb zapisanych w systemie dziesiętnym, ale także w przypadku wszystkich liczb niewymiernych, takich jak $\sqrt{2}$, π, e itd.

Błędy obcięcia powstają podczas obliczeń na skutek zmniejszenia liczby działań. Zagadnienie to występuje zwykle przy obliczaniu sum nieskończonych, np. wtedy gdy wartość funkcji liczy się na podstawie jej rozwinięcia w szereg. Chcąc obliczyć np. wartość funkcji $f(x_0 + \Delta x)$ na podstawie jej wartości w x_0, korzystamy z szeregu Taylora:

$$f(x_0+\Delta x)=f(x_0)+\frac{\Delta x}{1!}\frac{df}{dx}\bigg|_{x_0}+\frac{\Delta x^2}{2!}\frac{d^2f}{dx^2}\bigg|_{x_0}+\frac{\Delta x^3}{3!}\frac{d^3f}{dx^3}\bigg|_{x_0}+\cdots. \qquad \text{(Z.5)}$$

W powyższym szeregu nieskończonym można uwzględnić dowolną liczbę członów. Wynikający z tego faktu błąd obcięcia szeregu jest zdominowany przez następny człon rozwinięcia. Zatem każda wartość $f(x_0 + \Delta x)$ obliczona na jego podstawie będzie wartością przybliżoną. Dokładność tego przybliżenia będzie zależała od liczby uwzględnionych w obliczeniach wyrazów szeregu. Jeśli w rozwinięciu uwzględnimy wyrazy zawierające pochodne rzędu nie wyższego niż przyjęta liczba N, to wcześniejszy wzór zapiszemy

$$f(x_0+\Delta x)=f(x_0)+\frac{\Delta x}{1!}\frac{df}{dx}\bigg|_{x_0}+\frac{\Delta x^2}{2!}\frac{d^2f}{dx^2}\bigg|_{x_0}+\cdots+\frac{\Delta x^N}{N!}\frac{d^Nf}{dx^N}\bigg|_{x_0}+O(\Delta x^{N+1}), \quad \text{(Z.6)}$$

gdzie: $O(\Delta x^{N+1})$ jest tzw. resztą szeregu Taylora.

Człon $O(\Delta x^{N+1})$ interpretuje się następująco: istnieje tak dodatnia stała K, zależna od funkcji f, że różnica pomiędzy $f(x_0 + \Delta x)$ a sumą $N + 1$ członów prawej strony wzoru (Z.6), obliczonych w punkcie x_0, jest liczbowo mniejsza niż $K \cdot \Delta x^{N+1}$ dla każdej wystarczająco małej wartości Δx. Oznacza to, że każda wartość $f(x_0 + \Delta x)$ obliczona według wzoru (Z.6) jest aproksymacją f z dokładnością rzędu Δx^{N+1}. Interpretacja tego faktu jest następująca: jeśli danej wartości Δx odpowiada błąd obcięcia ε, to dwukrotne zmniejszenie Δx spowoduje błąd $\varepsilon/2^{N+1}$. Zatem człon $O(\Delta x^{N+1})$ mówi o tempie zmiany błędu obcięcia wywołanego zmianą Δx.

Jeśli N będzie dostatecznie dużą liczbą, to obliczona wartość $f(x_0 + \Delta x)$ może być równa wstępnie zaokrąglonej jej wartości, tzn. całkowity błąd popełniony przy jej obliczeniu będzie równy błędowi wejściowemu. Na ogół jednak, z uwagi na czas obliczeń, uwzględnia się niewielką liczbę wyrazów szeregu, co powoduje pojawienie się błędu obcięcia.

Błędy obcięcia pojawiają się również w innych przypadkach, jak na przykład przy obliczaniu całek niewłaściwych

$$I=\int_a^{\infty} f(x)dx, \qquad \text{(Z.7)}$$

w których nieskończoną granicę należy zastąpić wartością skończoną.

Błędy obcięcia odgrywają szczególnie ważną rolę w numerycznym rozwiązywaniu równań różniczkowych. Idea stosowanych metod polega na zastąpieniu pochodnych ilorazami różnicowymi. Do wyznaczenia przybliżonych wartości pochodnych wykorzystuje się szereg Taylora (Z.5). Z szeregu tego można wyznaczyć wartość pochodnej I rzędu w punkcie x_0. Jest ona równa

$$\frac{df}{dx}\bigg|_{x_0}=\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}-\frac{\Delta x}{2}\frac{d^2f}{dx^2}\bigg|_{x_0}-\frac{\Delta x^2}{6}\frac{d^3f}{dx^3}\bigg|_{x_0}+\cdots. \qquad \text{(Z.8)}$$

Jeśli dokonamy teraz obcięcia szeregu i po prawej stronie równości zostawimy tylko pierwszy wyraz, to otrzymamy wyrażenie

$$\left.\frac{df}{dx}\right|_{x_0} \approx \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}, \tag{Z.9}$$

będące aproksymacją pochodnej funkcji f w punkcie x_0. Pierwszy wyraz obciętej części szeregu (Z.8) informuje o dokładności aproksymacji pochodnej. W tym przypadku jest to aproksymacja rzędu $O(\Delta x)$. Oznacza to, że popełniany błąd obcięcia zmienia się liniowo ze zmianą wartości Δx.

Błędy zaokrągleń są efektem skracania liczb. Wiadomo, że w pamięci komputera może być zapisana liczba o skończonej ilości cyfr dziesiętnych. Z tego powodu nie wszystkie liczby mogą być zapisane w pamięci dokładnie. Nie można na przykład dokładnie przedstawić liczb niewymiernych, jak $\sqrt{2}$, π, e. Podobnie iloczyn dwu liczb o skończonym rozwinięciu, mający tyle cyfr, ile mają łącznie oba czynniki, może być zapisany w pamięci po zaokrągleniu.

W przypadku ułamka dziesiętnego można mówić o cyfrach istotnych i cyfrach ułamkowych (Dahlquist, Bjorck, 1983). Mówiąc o cyfrach istotnych nie uwzględnia się zer na początku ułamka, gdyż pomagają one tylko określić pozycję kropki dziesiętnej. Natomiast licząc cyfry ułamkowe, uwzględnia się wszystkie cyfry po kropce dziesiętnej, także ewentualne zera między kropką a pierwszą cyfrą różną od zera. Na przykład liczba 0,00147 ma pięć cyfr ułamkowych, ale tylko trzy cyfry istotne. Natomiast liczba 12,34 ma cztery cyfry istotne i dwie cyfry ułamkowe.

Jeśli moduł błędu wartości x' nie przekracza $1/2 \cdot 10^{-t}$, to mówimy, że x' ma t poprawnych cyfr ułamkowych. Cyfry występujące w x' aż do pozycji t-tej po kropce nazywamy cyframi znaczącymi. Wartość 0,001234 ±0,000004 ma pięć poprawnych cyfr ułamkowych i trzy cyfry znaczące. Natomiast wartość 0,001234 ±0,000006 ma 4 poprawne cyfry ułamkowe i dwie cyfry znaczące. Liczba cyfr poprawnych daje pojęcie o wielkości błędu bezwzględnego, a liczba cyfr znaczących – o wielkości błędu względnego.

Są dwa sposoby skracania liczb do danej długości t cyfr ułamkowych.

Ucinanie polega po prostu na odrzuceniu cyfr na prawo od t-tej. Tego sposobu nie zaleca się stosować, gdyż wprowadzamy błąd na systematycznie przeciwny znak niż sama liczba. Bardzo wiele typów komputerów ucina wyniki wszystkich operacji arytmetycznych. Zwykle nie jest to szkodliwe, gdyż liczba cyfr tworzących wielkości jest znacznie większa niż liczba cyfr znaczących w danych.

Zaokrąglanie polega na wyborze wśród liczb mających tylko t cyfr ułamkowych liczby najbliższej do danej. Jeśli zatem fragment liczby znajdujący się na prawo od t-tej cyfry ułamkowej jest mniejszy co do modułu od $1/2 \cdot 10^{-t}$, to t-tą cyfrę zostawia się bez zmiany. Jeśli ten fragment ma moduł większy od $1/2 \cdot 10^{-t}$, to do t-tej cyfry ułamkowej dodaje się jeden (1). W granicznym przypadku, gdy fragment ten jest dokładnie równy $1/2 \cdot 10^{-t}$, można dodać 1 do t-tej cyfry – gdy jest nieparzysta, lub pozostawić bez zmiany – gdy jest parzysta. Wtedy błędy dodatnie i ujemne będą jednakowo częste. Niezależnie od przyjętej konwencji, błąd zaokrąglenia leży zawsze w przedziale

$$\left[-\frac{1}{2}\cdot 10^{-t},\quad \frac{1}{2}\cdot 10^{-t}\right].$$

Efekty wymienionych wyżej sposobów postępowania można prześledzić, na przykładzie skracania liczb do trzech cyfr ułamkowych, co przedstawiono w tabeli Z.1.

Przenoszenie błędów z danych na wyniki ma miejsce podczas wykonywania operacji arytmetycznych na liczbach obarczonych błędami. Przy wykonywaniu dużej liczby działań zjawisko to może być niebezpieczne, ponieważ w sposób istotny może zmieniać ostateczne wyniki obliczeń. Istotę problemu prześledźmy na prostym przykładzie dotyczącym obliczenia obwodu kwadratu.

Tabela Z.1

Porównanie ucinania i zaokrąglania liczb

Liczba	Zaokrąglenie	Ucięcie
0.2397	0.240	0.239
–0.2397	–0.240	–0.239
0.23750	0.238	0.237
0.23652	0.237	0.236

Załóżmy, że pomiar długości boku kwadratu z dokładnością $\varepsilon = 1$ cm dał wynik $x' = 100$ cm. Oznacza to, że bok kwadratu ma wartość zawartą pomiędzy 99 cm i 101 cm, co można zapisać

$$x = 100 \pm 1 \text{ cm}.$$

Obwód kwadratu L będzie liczba zawartą pomiędzy liczbą 99 + 99 + 99 + 99 = 396 cm a liczbą 101 + 101 + 101 + 101 = 404 cm. Jeśli założymy, ze obwód jest równy średniej arytmetycznej z powyższych liczb, to można powiedzieć, że wynosi on:

$$L = 400 \pm 4 \text{ cm}.$$

Biorąc pod uwagę przyjętą wartość przybliżoną boku kwadratu $x' = 100$ cm oraz błąd bezwzględny $\varepsilon = 1$ cm, zauważamy, że otrzymany wynik sumowania boków kwadratu równy jest

$$L = x' + x' + x' + x' \pm(\varepsilon + \varepsilon + \varepsilon + \varepsilon).$$

Wniosek ten można łatwo uogólnić. Jeśli liczba x_1 oszacowana jest z dokładnością ε_1, zaś liczba x_2 z dokładnością ε_2, czyli $x_1 = x_1' \pm\varepsilon_1$ oraz $x_2 = x_2' \pm\varepsilon_2$, to

$$x_1 + x_2 = x_1' + x_2' \pm(\varepsilon_1 + \varepsilon_2) \tag{Z.10}$$

oraz

$$x_1 - x_2 = x_1' + x_2' \pm(\varepsilon_1 + \varepsilon_2). \tag{Z.11}$$

Oszacowanie błędu bezwzględnego wyniku dodawania lub odejmowania jest więc sumą oszacowań błędów bezwzględnych składników.

Postępując w podobny sposób z powierzchnią analizowanego kwadratu można potwierdzić znaną zasadę (Dahlquist i Bjorck, 1983), że mnożenie dwóch liczb przybliżonych x'_1 i x'_2, których błędy względne wynoszą odpowiednio ε'_1 i ε'_2, daje wynik, którego oszacowanie błędu względnego jest sumą oszacowań błędów względnych czynników. Zatem jeśli zgodnie z (Z.2) przyjmiemy, że

$$x'_1 = x_1(1 + \varepsilon'_1) \quad \text{oraz} \quad x'_2 = x_2(1 + \varepsilon'_2), \tag{Z.12a,b}$$

to

$$x_1' \cdot x_2' = x_1 \cdot x_2\,(1 + \varepsilon_1') \cdot (1 + \varepsilon_2') \tag{Z.13}$$

ma błąd względny ε równy, zgodnie z (Z.2)

$$\varepsilon' = \frac{x_1 x_2 (1+\varepsilon_1')(1+\varepsilon_2') - x_1 x_2}{x_1 x_2} \approx \varepsilon_1' + \varepsilon_2' , \tag{Z.14}$$

dla odpowiednio małych wartości ε'_1 i ε'_2.