

# Óravázlatok: Számítógépes Fizika (*Computational Physics*)

Numerikus módszerek összefoglaló

Bartha Ferenc\*

Szegedi Tudományegyetem, Elméleti Fizikai Tanszék

készültség: May 9, 2003

(<http://www.jate.u-szeged.hu/~barthaf/oktatas.htm>)

A jegyzet folyamatosan változik. A javított (rontott) és kibővített részek képezik az aktuális anyagot. Az "Eddig..... ál-fejezetcím" után, a miheztartás végett közlöm az átdolgozott rész után következő anyag korábbi változatát.

Kritikai megjegyzéseket, hibajegyzéket örömmel fogadok.

## Tartalom

|                                                                                    |    |
|------------------------------------------------------------------------------------|----|
| <b>1. Bevezetés: Numerikus differenciálás és integrálás</b>                        | 3  |
| 1.1. Numerikus differenciálás véges differenciák használatával                     | 3  |
| 1.2. Praktikus megfontolások                                                       | 4  |
| 1.3. Milyen kicsi legyen $h$ ?                                                     | 4  |
| 1.4. Nem lehetne akkor inkább másképp eljárni?                                     | 6  |
| 1.5. Csebisev-approximáció                                                         | 6  |
| 1.6. Numerikus integrálás                                                          | 7  |
| 1.7. Klasszikus módszerek                                                          | 8  |
| 1.8. Egy praktikus algoritmus                                                      | 9  |
| 1.9. Gauss–Legendre-kvadratura                                                     | 10 |
| <b>2. Közönséges differenciálegyenletek (ODE: Ordinary Differential Equations)</b> | 11 |
| 2.1. ODE redukálása egyenletrendszerre                                             | 11 |
| 2.2. Euler-módszer                                                                 | 12 |
| 2.3. Pontosság                                                                     | 13 |
| 2.4. Stabilitás                                                                    | 13 |
| 2.5. Többlépéses prediktor módszerek                                               | 14 |
| 2.6. Implicit módszerek                                                            | 14 |
| 2.7. Runge-Kutta-módszer                                                           | 15 |
| 2.8. Adaptív lépések a differenciálegyenletek integrálásakor                       | 16 |
| 2.9. Hiányos másodrendű differenciálegyenletek                                     | 18 |
| <b>3. Kezdetiérték- versus peremérték-feladat</b>                                  | 18 |
| 3.1. Lövöldözős (SHOOTING) módszer                                                 | 19 |
| 3.2. Relaxációs módszerek                                                          | 19 |
| 3.3. Egy BVP: Az egydimenziós Schrödinger-egyenlet kötött állapotai                | 20 |
| <b>4. Parciális differenciálegyenletek</b>                                         | 21 |
| 4.1. A lineáris egyenletek osztályozása                                            | 21 |
| 4.2. Hiperbolikus egyenletek                                                       | 22 |
| 4.2.1. A bébi hullámegyenlet                                                       | 22 |
| 4.2.2. Numerikus diszperzió és disszipáció                                         | 22 |
| 4.2.3. Teljesen diszkretizált explicit módszerek                                   | 24 |
| 4.3. Parabolikus egyenletek                                                        | 27 |
| 4.3.1. Példa: az időfüggő Schrödinger-egyenlet                                     | 30 |
| 4.3.2. Tridiagonális egyenletrendszer megoldása előre-hátra helyettesítéssel       | 31 |

---

\*Electronic address: [barthaf@physx.u-szeged.hu](mailto:barthaf@physx.u-szeged.hu)

|                                                                                     |           |
|-------------------------------------------------------------------------------------|-----------|
| 4.3.3. A diffúziós probléma több dimenzióban                                        | 32        |
| 4.3.4. Kiegészítés: $\mathbf{A}\vec{v}=\vec{b}$ lineáris egyenletrendszer megoldása | 33        |
| 4.4. Elliptikus egyenletek: peremérték-probléma                                     | 35        |
| 4.4.1. Relaxáció                                                                    | 36        |
| 4.4.2. Váltakozó irányú implicit módszer: ADI                                       | 36        |
| 4.4.3. Ciklikus redukció                                                            | 37        |
| 4.4.4. Fourier-módszer                                                              | 37        |
| <b>5. ----Eddig átdolgozva: May 9, 2003. Folytatás másik félévben..</b>             | <b>38</b> |
| <b>6. Fourier-transzformáció</b>                                                    | <b>39</b> |
| 6.1. A Fourier-transzformációról általában                                          | 39        |
| 6.2. Mintavételezés                                                                 | 39        |
| 6.3. Diszkrét Fourier-transzformáció (DFT)                                          | 40        |
| 6.4. Gyors Fourier-Transzformáció (FFT)                                             | 41        |
| 6.5. FFT valós függvényekre                                                         | 42        |
| 6.6. sinFFT szinusz és cosFFT koszinusz Fourier-transzformáltak                     | 43        |
| 6.7. Többdimenziós FFT                                                              | 44        |
| 6.8. Konvolúciós egyenletek                                                         | 45        |

---

<sup>0</sup> Ajánlott irodalom:

Numerical Recipes Books On-Line, [http://www.ulib.org/webRoot/Books/Numerical\\_Recipes/](http://www.ulib.org/webRoot/Books/Numerical_Recipes/)

# 1. BEVEZETÉS: NUMERIKUS DIFFERENCIÁLÁS ÉS INTEGRÁLÁS

## 1.1. Numerikus differenciálás véges differenciák használatával

Tegyük fel, hogy ismerjük az  $f(x)$  függvény  $\{f_k\}$  értékeit az  $x = x_0$  pont körüli ekvidisztáns beosztáson:

$$f_k = f(x_k) ; x_k = x_0 + kh ; k = 0, \pm 1, \pm 2, \dots \quad (1.1)$$

Ha a függvény az  $x = x_0$  körül konvergens hatványsorba fejthető, az

$$f(x) = f(x_0) + \frac{(x - x_0)}{1!} f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0) + \frac{(x - x_0)^3}{3!} f'''(x_0) + \dots \quad (1.2)$$

Taylor-sor alapján írhatjuk, hogy

$$f_k = \sum_{i=0}^n \frac{(kh)^i}{i!} f^{(i)}(x_0) + O(h^{n+1}) \quad (1.3)$$

ahol  $O(h^{n+1})$  a  $h$ -nak  $n$ -nél magasabb hatványait tartalmaz. Ha  $h$  elegendően kicsiny, akkor az előbbi kifejezésből az  $O(h^{n+1})$  maradékot elhagyjuk, az ezzel elkövetett hibát **csonkolási hibának** hívjuk. A csonkolás után az ismeretlen  $f^{(i)}(x_0)$ ,  $i = 1, \dots, n$  deriváltakra egy lineáris egyenletet kapunk minden  $k$  értéknél. Felírva  $n$  darab különböző  $k \neq 0$  értékre az egyenleteket a

$$f_k = \sum_{i=0}^n \frac{(kh)^i}{i!} f^{(i)}(x_0) , \quad k = k_1, \dots, k_n , \quad f^{(0)}(x_0) = f(x_0) = f_0 , \quad (1.4)$$

lineáris egyenletrendszer kell megoldanunk a keresett deriváltakra. A leggyakoribb így kapott kifejezések az első deriváltakra (a csonkolási hiba feltüntetésével):

- **kétpontos**

$$\begin{aligned} f_{\pm 1} &= f(x_0) \pm hf'(x_0) + O(h^2) \\ f'(x_0) &= \pm \frac{f_{\pm 1} - f_0}{h} + O(h) \end{aligned} \quad (1.5)$$

- **hárompontos, vagy szimmetrikus**

$$\begin{aligned} f_{\pm 1} &= f(x_0) \pm hf'(x_0) + \frac{h^2}{2} f''(x_0) + O(h^3) \\ f'(x_0) &= \frac{f_{+1} - f_{-1}}{2h} + O(h^2) \end{aligned} \quad (1.6)$$

- **5-pontos**

$$f'(x_0) = \frac{1}{12h} [f_{-2} - 8f_{-1} + 8f_{+1} + f_{+2}] + O(h^4) \quad (1.7)$$

Három pont ( $f_{-1}, f_0, f_{+1}$ ) felhasználásával a második derivált is kifejezhető. Az ebből kapható formula csonkolási hibája a várt  $O(h)$  helyett szerencsénkre csak  $O(h^2)$  lesz. Az alábbi számolás mutatja, hogy a  $h^3$ -ös tag valóban kiesik a rendezés során:

$$\begin{aligned} f_{\pm 1} &= f(x_0) \pm hf'(x_0) + \frac{h^2}{2} f''(x_0) \pm \frac{h^3}{6} f'''(x_0) + O(h^4) \Rightarrow f_{+1} + f_{-1} = 2f(x_0) + \frac{h^2}{2} f''(x_0) + O(h^4) \\ f''(x_0) &= \frac{f_{+1} - 2f_0 + f_{-1}}{h^2} + O(h^2) \end{aligned} \quad (1.8)$$

Néhány további többpontos és magasabb rendű derivált formula:

|               | 4-pont                                                          | 5-pont                                                          |
|---------------|-----------------------------------------------------------------|-----------------------------------------------------------------|
| $hf^{(1)}$    | $\pm \frac{1}{6} [-2f_{\mp 1} - 3f_0 + 6f_{\pm 1} - f_{\pm 2}]$ | $\frac{1}{12} [f_{-2} - 8f_{-1} + 8f_{+1} + f_{+2}]$            |
| $h^2 f^{(2)}$ | $[4f_{-1} - 2f_0 + f_{+1}]$                                     | $\frac{1}{12} [-f_{-2} + 16f_{-1} - 30f_0 + 16f_{+1} - f_{+2}]$ |
| $h^3 f^{(3)}$ | $\pm [-f_{\mp 1} + 3f_0 - 3f_{\pm 1} + f_{\pm 2}]$              | $\frac{1}{2} [-f_{-2} + 2f_{-1} - 2f_{+1} + f_{+2}]$            |
| $h^4 f^{(4)}$ | .....                                                           | $[f_{-2} - 4f_{-1} + 6f_0 - 4f_{+1} + f_{+2}]$                  |

(1.9)

Parciális deriváltakat értelemszerűen hasonló módon közelíthetünk. Például

$$\begin{aligned}\frac{\partial^2}{\partial x \partial y} f &= \frac{\partial}{\partial x} \left( \frac{\partial}{\partial y} f \right) = \frac{\partial}{\partial x} \left( \frac{f(x, y+h) - f(x, y-h)}{2h} \right) \\ \frac{\partial^2}{\partial x \partial y} f &= \frac{[f(x+h, y+h) - f(x+h, y-h)] - [f(x-h, y+h) - f(x-h, y-h)]}{4h^2}\end{aligned}\quad (1.10)$$

ahol kétszer alkalmaztuk az első deriváltak (1.6) szimmetrikus közelítését.

## 1.2. Praktikus megfontolások

- Előfordulhat, hogy a differenciálni kívánt függvény nem ekvidisztáns (1.1) beosztáson adott. A csonkolt Taylor-polinom használatával ekkor is felírható a deriváltak meghatározására szolgáló egyenletrendszer, a megfelelő kifejezések elemi úton meghatározhatók.
- A tabulált függvényértékek 'zajosak' lehetnek. Különösen így van ez, ha azok mérési adatokból származnak, vagy éppen valamely bonyolult számítás eredményeként álltak elő. Ezekben az esetekben nem célszerű az előbbi formulák közvetlen alkalmazása. A legkézenfekvőbb eljárás az lehet, hogy a kérdéses pont körül vett megfelelő számú  $\{f_k\}$  értékre előbb valamilyen görbét (pl. polinomot) illesztünk (pl. legkisebb négyzetek módszerével), majd az így kapott analitikus görbét differenciáljuk. A konkrét feladat ismeretében (mi a pontok tipikus várt viselkedése, hanyadrendű deriváltak számolására van szükség,...) dől el, hogy hány pontra és milyen görbét illesztünk. Az illesztés és az illesztett függvény differenciálása egyetlen eljárásba gazdaságosan összevonható.

## 1.3. Milyen kicsi legyen $h$ ?

A deriváltak véges differenciákkal való helyettesítése sima függvényekre elvileg határértékben egzakt eljárás, hiszen például

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = f'(x) \quad (1.11)$$

Azonban a  $h \rightarrow 0$  határátmenet praktikusán csak közelítőleg végezhető el. Numerikusan úgy járunk el, hogy megpróbálunk valamely "nagyon kicsi"  $h$ -t választani. Milyen kicsit?

Az első probléma, hogy a számítógépek a valós számokat véges pontosságú (többnyire) *lebegőpontos szám* formában tárolják. A lebegőpontos ábrázolás általános alakja

$$x \approx s * M * B^e \quad \text{például} \quad -0.94654676 * 2^{37} \quad (1.12)$$

Itt  $x$  az ábrázolni kívánt valós szám,  $s$  az előjel,  $M$  a mantissza,  $B$  az "alap" és  $e$  az egész exponens.  $B$  a géptől és a programozási nyelvtől függően rögzített, általában  $B = 2$ , de előfordul gyakran, hogy  $B = 16$ . Az  $s$ ,  $M$  és  $e$  számokat valahány bit segítségével (kettes számrendszerben) írjuk fel és tároljuk a

| s | e(-E)     | M                 |
|---|-----------|-------------------|
| 1 | 1 0 ... 0 | 1 0 0 1 ... 1 0 0 |

sematikus példa szerint. Az  $s$  előjel egzaktul ábrázolható egyetlen bit segítségével. Az (esetleg eltolt)  $e$  exponens néhány bites egész számként egzakt az ábrázolható intervallumban, például 8 bites exponens esetén és  $E = 128$  eltolást választva  $-128 \leq e \leq 127$  egész exponensek lehetségesek. A kritikus pont a mantissza ábrázolása. A tipikusan néhány tíz bit segítségével lehetetlen lenne az egységnyi nagyságrendű kontinuum sok valós számot különbözőként ábrázolni. A számábrázolás (~számítógép) pontosságát az

$$\epsilon_m \quad (\text{epmach : machine epsilon, kerekítési hiba})$$

valós számmal szokás jellemezni.  $\epsilon_m$  az a legkisebb valós szám, amit az 1.0 számhoz hozzáadva 1.0-tól különböző számot kapunk. Ez a szám egyúttal jellemzi azt a tipikus **relatív hibát**, amivel tetszőleges valós szám ábrázolható:

$$\left| \frac{\Delta x}{x} \right| \approx \epsilon_m \quad (1.13)$$

Jellemző értékek: 4 byte-on való ábrázolásnál  $\epsilon_m \sim 10^{-8}$ , 8 byte esetén  $\epsilon_m \sim 10^{-16}$ .

A kerekítési hiba egyik hatása a véges differenciákkal való számolásokban, hogy az "állítólag"  $x$  és az "állítólag"  $x + h$  számok tényleges különbsége nem  $h$ , hanem attól lényegesen eltérő lehet. Mivel az  $x$  és az  $x + h$  számok a számítógép belső ábrázolásában  $\sim \epsilon_m x$  hibával terheltek, így a  $h$  relatív hibája  $\sim \epsilon_m x / h$  lesz. Ez azt jelenti, hogy pl.  $x \sim 100$ ,  $\epsilon_m \sim 10^{-6}$  és  $h \sim 10^{-3}$  mellett 10% hibával számolhatunk. Szerencsére ezt a kerekítési hibát kis odafigyeléssel kikerülhetjük a következő trükkel

$$\begin{aligned} h &= 10^{-3} \\ x &= 100 \\ y &= x + h \\ h &= y - x \end{aligned}$$

mely esetben  $h$  a gépben ábrázolt  $x$  és  $y$  "egzakt" különbsége lesz. Így már csak arra kell vigyáznunk, hogy  $h \gg \epsilon_m x$  legyen, mert ellenkező esetben könnyen lehet, hogy  $x$  és  $y$  ugyanaz a szám, és az "igazi"  $h$  nulla lesz.

A kerekítési hiba  $h$ -ra való hatását az előbbiek szerint kiküszöbölve, a számlálóban levő függvényértékekről kell szólni. A függvényértékek előállításuk módjától ("egyszerű" belső függvényhívással, méréssel, bonyolult számítás eredményeként,...) függően  $\epsilon_f$  relatív hibával állnak rendelkezésre. Egy

$$f'(0) = \frac{f_1 - f_0}{h} \quad (1.14)$$

első derivált véges differenciákkal való számításakor a kerekítési hiba

$$E_r \approx \epsilon_f \frac{|f(x)|}{h} \quad (1.15)$$

Az  $O(h)$  csonkolási hiba pedig

$$E_t \approx |f''(x)| h \quad (1.16)$$

Az (1.14) kifejezés teljes relatív hibája tehát

$$\epsilon(h) = \frac{E_r + E_t}{|f'|} = \frac{\epsilon_f}{h} \left| \frac{f}{f'} \right| + \left| \frac{f''}{f'} \right| h \quad (1.17)$$

A legkisebb hibát annál a  $h$ -nál kapjuk, amikor

$$\frac{\epsilon(h)}{dh} = -\frac{\epsilon_f}{h^2} \left| \frac{f}{f'} \right| + \left| \frac{f''}{f'} \right| = 0$$

A szélsőértékre ebből

$$h_{opt} = \sqrt{\epsilon_f \left| \frac{f}{f''} \right|} \quad \text{és} \quad \epsilon_{min} = \frac{\epsilon_f}{h} \left| \frac{f}{f'} \right| + \left| \frac{f''}{f'} \right| h = 2\sqrt{\epsilon_f} \sqrt{\left| \frac{f * f''}{f' * f'} \right|} \quad (1.18)$$

Általában  $f$ ,  $f'$  és  $f''$  olyan, hogy az utóbbi gyökjel alatti kifejezés egységnyi nagyságrendű, tehát

$$\epsilon_{min} \approx \sqrt{\epsilon_f} \quad \text{és} \quad h_{opt} = \sqrt{\epsilon_f} * x_c \quad (1.19)$$

ahol az  $x_c = \sqrt{f/f''}$  a függvény viselkedésére jellemző karakterisztikus hossz. A függvényről való egyéb információ hiányában  $x_c \sim x$  vehető (kivéve az  $x \approx 0$  környéket). Általában  $\epsilon_f \gg \epsilon_m$ , most tekintsük a legjobb esetet, amikor  $\epsilon_f \approx \epsilon_m$ . Azt látjuk, hogy a megfelelő  $h$  választásával a legjobb esetben is csak

$$\epsilon_{min} \approx \sqrt{\epsilon_m} \quad (1.20)$$

érhető el. Ez intő jel.

Többspontos formulát használva az elérhető pontosság javul, így pl. a szimmetrikus (1.6) derivált hibája hasonló számítással megmutathatóan

$$\epsilon_{min} \approx (\epsilon_f)^{2/3} \quad \text{ha} \quad h_{opt} = (\epsilon_f)^{1/3} * x_c \quad (1.21)$$

Ez  $\epsilon_m \sim 10^{-6}$  mellett egy nagyságrenddel pontosabb eredményt tesz lehetővé, mint az előző formulánk, miközben egy nagyságrenddel nagyobb  $h$ -val (ugyanakkora intervallumon tizedannyi osztóponttal) számolhatunk.

#### 1.4. Nem lehetne akkor inkább másképp eljárni?

Szükség lehet arra, hogy az előbbieknél nagyobb pontossággal, vagy esetleg kevesebb függvényszámolással állítsuk elő a deriváltakat. Az egyik kézenfekvő eljárás a nagyobb pontosság eléréséhez, hogy numerikusan szimuláljuk a (1.11) határátmenetet. Ez úgy történik, hogy valamely  $h$  mellett kiszámoljuk a differenciahányadost, majd egyre kisebb  $h$  választással figyeljük, hogy a számolt értékek tartanak-e valahová. Természetesen ez az eljárás is könnyen csődöt mondhat, különösen, ha ügyetlenül írjuk meg. Nem részletezzük most, hogy milyen trükkök szem előtt tartása célszerű egy ilyen eljárás megírásához, az érdeklődők a szakirodalomban megkereshetik a jótanácsokat.

Egy alternatív eljárást már korábban említettünk, amikor a függvényértékekre illesztett könnyen differenciálható görbe használatát javasoltuk. Egy ilyen jellegű eljárást részletesebben is kidolgozunk, tekintettel arra, hogy az a numerikus kalkulus egyéb vonatkozásaiban is igen hasznos. Ha a véges intervallumon vizsgálandó függvényről joggal feltételezzük, hogy az elegendően sima, akkor a függvényt célszerűen a Csebisev-polinomok segítségével közelíthetjük.

#### 1.5. Csebisev-approximáció

A  $[-1, 1]$  intervallumon értelmezett  $T_n$ ,  $n = 0, 1, 2, \dots$  Csebisev-polinomok általános alakja

$$T_n(\cos\vartheta) = \cos n\vartheta, \text{ vagy } T_n(x) = \cos(n \arccos x) \quad (1.22)$$

Néhány alacsonyabbrendű polinom

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x, \quad T_4(x) = 8x^4 - 8x^2 + 1 \quad (1.23)$$

és a továbbiak is egyszerűen generálhatók a

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0 \quad (1.24)$$

rekurzióval. A polinomok ortogonálisak az

$$\int_{-1}^{+1} T_n(x)T_m(x)(1-x^2)^{-1/2}dx = \begin{cases} 0 & m \neq n \\ \pi/2 & m = n \neq 0 \\ \pi & m = n = 0 \end{cases} \quad (1.25)$$

értelemben. A függvények Csebisev-sorfejtése szempontjából fontos, hogy  $T_n(x)$ -nek a  $(-1, 1)$  intervallumban pontosan  $n$  darab zéróhelye van, nevezetesen az

$$x_k = \cos\left(\frac{\pi(k - \frac{1}{2})}{n}\right) \quad k = 1, 2, \dots, n \quad (1.26)$$

helyeken. Az is látható, hogy ugyanezen az intervallumon  $T_n(x)$  pontosan  $n + 1$  lokális szélsőértékkel rendelkezik az

$$x_k = \cos\left(\frac{\pi k}{n}\right) \quad k = 1, 2, \dots, n \quad (1.27)$$

pontokban. A maximumok mindegyikénél  $T_n(x_k) = 1$ , míg a minimumoknál  $T_n(x_k) = -1$ . A polinomok zéróhelyeik szerint egy diszkrét ortogonalitási relációt is kielégítenek: Ha  $x_k$ ,  $k = 1, 2, \dots, n$  a  $T_n$  zéróhelyei és  $i, j < n$ , akkor

$$\sum_{k=1}^n T_i(x_k)T_j(x_k) = \begin{cases} 0 & i \neq j \\ n/2 & i = j \neq 0 \\ n & i = j = 0 \end{cases} \quad (1.28)$$

Ezen összefüggés segítségével megmutatható: Ha  $f(x)$  a  $[-1, 1]$  intervallumon értelmezett függvény és az  $N$  darab  $c_j$  együtthatót úgy választjuk, hogy

$$c_j = \frac{2}{N} \sum_{k=1}^N f(x_k)T_{j-1}(x_k) \quad (1.29)$$

legyen, akkor az

$$f(x) \approx \left[ \sum_{k=1}^N c_k T_{k-1}(x) \right] - \frac{1}{2} c_1 \quad (1.30)$$

közelítés *egzakt egyenlőséget* jelent minden olyan  $x_k$  helyen, ahol  $T_N(x) = 0$ . Ez a függvény egy polinomiális közelítése, semmivel sem jobb (pontosabb), mint más, hasonló fokszámú polinommal készített egyéb közelítés. A nagy előnye abban rejlik, hogy ezt az esetleg magas fokszámú polinomot "visszavághatjuk" egy kezelhetően alacsony  $m \ll N$  fokszámúra

$$f(x) \approx \left[ \sum_{k=1}^m c_k T_{k-1}(x) \right] - \frac{1}{2} c_1 \quad (1.31)$$

ahol  $c_k$  az előbbieken meghatározott változatlan együtttható. Egy sima függvénytől joggal elvárhatjuk, hogy a Csebisev-sorában a  $c_k$  együttthatók gyorsan csökkennek nagy  $k$ -ra, azt meg tudjuk, hogy  $|T_n| \leq 1$ , minek következtében a csonkolt sor és az eredeti (1.30) közötti különbség dominánsan a  $c_{m+1}T_m(x)$  tagban van. Amellett, hogy ez kicsiny, külön előny, hogy a  $T_m(x)$  gyorsan oszcillál, így ez a kis hiba is egyenletesen "szétkenődik" az egész intervallumon.

Az eljárás tehát a következő:

- Válasszuk meg úgy az  $N$ -et a (1.29) számoláshoz, hogy az  $N$  darab függvényhívás még elviselhető (értelmes) feladat legyen. Ha az összes  $c_k$  együttthatót meghatároznánk, akkor a (1.30) polinomunk egzaktul megegyezne az  $N$  sűrűn elhelyezkedő pontokban a függvénnyel.
- Ne határozzuk meg az összes  $c_k$ -t, hanem figyeljük, hogy azok hogyan csökkennek. Álljunk meg egy olyan  $m \ll N$  indexnél, ahol  $c_m$  már elegendően kicsiny és az  $m$ -edfokú polinommal való további számolás praktikusnak látszik. Ha szerencsénk van, akkor ezzel megkaptuk az  $f(x)$  egy jól használható közelítését.
- Hogyan számoljuk ki a Csebisev-polinom ismeretében a függvény  $f(x)$  értékét valamely  $x$  pontban? Megtehetnénk, hogy kiszámoljuk az  $m$  darab  $T_0(x), T_1(x), \dots, T_{m-1}(x)$  értéket a rekúzió segítségével, majd ezeket a megfelelő  $c$ -vel megszorozzuk és összeadjuk a szorzatokat. Ez kevésbé gazdaságos, a két műveletet össze lehet vonni, a rekúziót közvetlenül a Csebisev-polinomra is felírhatjuk az alábbi módon:

$$\begin{aligned} d_{m+1} &= d_{m+2} = 0 \\ d_j &= 2xd_{j+1} - d_{j+2} + c_j \quad j = m, m-1, \dots, 2 \\ f(x) &= xd_2 - d_3 + \frac{1}{2}c_1 \end{aligned} \quad (1.32)$$

- A Csebisev-polinom  $c_k$  koefficienseinek az ismeretében a függvény deriváltját is egyszerűen számolhatjuk, ugyanis a derivált Csebisev-polinomja

$$f'(x) \approx \left[ \sum_{k=1}^{m-1} \tilde{c}_k T_{k-1}(x) \right] - \frac{1}{2} \tilde{c}_1 \quad (1.33)$$

ahol

$$\begin{aligned} \tilde{c}_m &= \tilde{c}_{m+1} = 0 \\ \tilde{c}_{k-1} &= \tilde{c}_{k+1} - 2(k-1)c_k \quad k = m, m-1, \dots, 2 \end{aligned} \quad (1.34)$$

## 1.6. Numerikus integrálás

Míg az elemi függvények differenciálása többnyire analitikusan is elvégezhető, ugyanez nem mondható el az integrálásukra. Például már az olyan "egyszerű" függvényeknek, mint  $f(x) = \frac{\sin(x)}{x}$ , vagy  $f(x) = \sqrt{1+x^2}$  nincs primitív függvénye. Továbbá a gyakorlatban az integrálandó függvény gyakran nem is áll rendelkezésre explicit formában, csak mintavételezése lehetséges. Így nem csoda, hogy az

$$\int_a^b f(t)dt \quad , \quad a > b \quad (1.35)$$

határozott integrál közelítő kiszámítása, a **numerikus kvadratúra** több évszázada az érdeklődés homlokterében van.

A következő alfejezetben felsorolunk néhány klasszikus eljárást, melyek mindegyike az integrálok közelítő összeggel való felírásán alapul. A modern számítógépek megjelenésével ezek a híres módszerek tulajdonképpen elavultak.

Korszerűbbnek tekinthető, ha az integrál direkt kiszámolása helyett a

$$\frac{d}{dx}F(x) = f(x) \text{ , } F(a) = 0 \quad (1.36)$$

kezdetiérték-problémát oldjuk meg a közönséges differenciálegyenletek megoldására ajánlott nagy hatékonyságú eljárásokkal. A későbbiekben foglalkozunk részletesen ezekkel differenciálegyenletekkel.

A Csebisev-sorfejtés szintén lehetővé teszi, hogy az integrálokat kiszámítsuk. Ugyanis, ha az  $f(x)$  Csebisev-sorfejtése

$$f(x) \approx \left[ \sum_{k=1}^m c_k T_{k-1}(x) \right] - \frac{1}{2}c_1 \quad (1.37)$$

akkor

$$\int^x f \approx \left[ \sum_{k=1}^m C_k T_{k-1}(x) \right] - \frac{1}{2}C_1 \text{ ahol } C_k = \frac{c_{k-1} - c_{k+1}}{2(k-1)} \text{ } k > 1 \quad (1.38)$$

és  $C_1$  tetszőleges (integrációs) állandó.

### 1.7. Klasszikus módszerek

Az általános eljárás szerint ekvidisztáns

$$f_k = f(x_k) \text{ ; } x_k = a + k \frac{b-a}{n} \text{ ; } k = 0, 1, 2, \dots, n \Rightarrow f_0 = f(a) \text{ , } f_n = f(b) \quad (1.39)$$

módon felosztjuk az integrációs  $[a, b]$  tartományt  $n$  darab  $h = (b-a)/n$  hosszúságú részintervallumra. Az egyes részintervallumokon valamilyen egyszerű függvénnyel (pl. polinommal) illesztjük (helyettesítjük) a függvényt és az illesztő függvényt integráljuk. Minél kisebb részintervallumokon (nagy  $n$ , illetve kis  $h$ ) illesztünk és integrálunk, annál pontosabb értéket várunk.

**Trapéz-szabály:** A függvényt az egyes részintervallumokon első fokú

$$f(t) = f_0 + t \cdot f'_0 + O(h^2 f'') = f_0 + t \cdot \frac{f_1 - f_0}{h} + O(h^2 f'') \quad (1.40)$$

polinommal közelítjük. Ekkor a *kétpontos* formula:

$$\int_{x_0}^{x_1} f(t) dt = h \cdot f_0 + \frac{h^2}{2} \cdot \frac{f_1 - f_0}{h} + O(h^3 f'') = \frac{h}{2} [f_0 + f_1] + O(h^3 f'') \quad (1.41)$$

illetve az összes részintervallumra összeadva:

$$\int_a^b f(t) dt = \frac{h}{2} [f_0 + 2f_1 + \dots + 2f_{n-1} + f_n] + O\left(\frac{1}{n^2}\right) \quad (1.42)$$

Becslés a közelítés hibájára: Ha  $f''$  (második derivált) folytonos az  $[a, b]$ -n és létezik olyan  $M_2$  felső korlát, hogy  $|f''| \leq M_2$  az egész intervallumon, akkor

$$E_T \leq \frac{b-a}{12} h^2 M_2$$

Azt látjuk, hogy ez esetben valóban nagyon kis  $h$ -t választva nagy pontosságot érhetnénk el. A gyakorlatban azonban nem lehet tetszőlegesen kis  $h$ -t használni a numerikus bizonytalanságok miatt. Ha a trapéz szabály "értelmes"  $h$ -val nem ad kielégítő pontosságot, akkor más módszert kell használni.

**Simpson-szabály:** Az  $[a, b]$  intervallumot *páros számú*, egyenlő hosszúságú részintervallumra osztjuk. Minden részintervallum *páron* másodfokú polinommal közelítve a függvényt

$$f(t) = f_0 + t \cdot f'_0 + \frac{t^2}{2} \cdot f''_0 + O(h^3 f''') = f_0 + t \cdot \frac{f_{+1} - f_{-1}}{2h} + \frac{t^2}{2} \cdot \frac{f_{+1} - 2f_0 + f_{-1}}{h^2} + O(h^3 f''') \quad (1.43)$$



kapjuk a *hárompontos* formulát, hogy

$$\int_{-h}^h f(t)dt = 2h \cdot f_0 + \frac{2h^3}{6} \cdot \frac{f_{+1} - 2f_0 + f_{-1}}{h^2} + O(h^5 f^{(4)}) = \frac{h}{3} [f_{-1} + 4f_0 + f_{+1}] + O(h^5 f^{(4)})$$

Vegyük észre, hogy a  $t^3$  tag integrálja nulla lenne, így a vártnál jobb közelítést kaptunk, ezért az  $O(h^5)$  hiba. A teljes integrál tehát

$$\int_a^b f(t)dt = \frac{h}{3} [f_0 + 4f_1 + 2f_2 + \dots + 2f_{n-2} + 4f_{n-1} + f_n] + O(\frac{1}{n^4}) \quad (1.44)$$

A közelítés hibája: Ha  $f^{(4)}$  (negyedik derivált) folytonos az  $[a, b]$ -n és létezik olyan  $M_4$  felső korlát, hogy  $|f^{(4)}| \leq M_4$  az egész intervallumon, akkor

$$E_S \leq \frac{b-a}{180} h^4 M_4$$

**Simpson 3/8-os szabálya:** harmadfokú Taylor polinommal és megfelelő numerikus deriváltakkal *négypontos* formula

$$\int_{x_0}^{x_3} f(t)dt = \frac{3h}{8} [f_0 + 3f_1 + 3f_2 + f_3] + O(h^5 f^{(4)}) \quad (1.45)$$

**Bode-szabály:** negyedfokú polinommal illesztett 5-pontos formula

$$\int_{x_0}^{x_4} f(t)dt = \frac{2h}{45} [7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4] + O(h^7 f^{(6)}) \quad (1.46)$$

ahol a polinom páros fokszáma miatt a pontosság újra egy renddel jobb a vártnál.

### 1.8. Egy praktikus algoritmus

A trapéz-szabály példáján néhány szót ejtünk a gyakorlati programozásról is. Nem tudhatjuk előre, hogy hány pontra is osszuk fel az integrációs tartományt a kívánt pontosságú integrál eléréséhez. Értelmes stratégia a következő:

- Az első lépésben csak a két végpont segítségével becsüljük az integrált

$$\int_a^b f(t)dt \approx s_1 = \frac{f(a) + f(b)}{2} * (b - a)$$

- A következő tennivaló, hogy felvesszük az intervallumot felező  $c$  osztópontot, majd az előbbi becslés eredményét felhasználva írjuk, hogy

$$\int_a^b f(t)dt \approx s_2 = \frac{1}{2} s_1 + \frac{b-a}{2} f(c)$$

- Többször megismételjük az intervallumfelezéseket és a

$$\int_a^b f(t)dt \approx s_k = \frac{1}{2} s_{k-1} + \frac{b-a}{2^{k-1}} \sum_i f(x_i) \quad (1.47)$$

formula segítségével javítjuk az integrál értékét. Itt az összegzés az újonnan keletkezett  $x_i$  osztópontokon fut végig.

- Vizsgáljuk, hogy az előre kijelölt  $\epsilon$  hibát sikerült-e már elérni, azaz addig folytatjuk, amíg

$$|s_k - s_{k-1}| < \epsilon$$

nem lesz. Természetesen vigyázni kell arra, hogy  $\epsilon$  olyan legyen, hogy a konvergencia még értelmes számú osztóponttal bekövetkezzen. Azaz  $k$  ne nőjön túl nagyra, ami nemcsak azért lenne kellemetlen, mert nagyon sok függvényértéket kellene kiszámítani, hanem a túl sűrű osztópontok a (nagyon kis  $h$ ) miatt a számábrázolási határba ütközünk.

- Valójában a trapéz szabály itt vázolt algoritmussal, ugyanakkora munkával gyorsabban/pontosabb eredményre vezethet kis trükk segítségével. Ellenőrizhető, hogy az

$$S_k = \frac{4}{3}s_k - \frac{1}{3}s_{k-1} \quad (1.48)$$

kombináció éppen a Simpson-szabálynak megfelelő összeget adja. Így minden további erőlködés nélkül bízhatunk az  $O(\frac{1}{n^4})$  konvergenciában.

### 1.9. Gauss–Legendre-kvadrátúra

Az előbbiekben megismert eljárás során a határozott integrálokat úgy számoltuk ki, hogy az integrálási tartományt egyenlő hosszúságú részintervallumokra bontottuk, majd azokon a függvényt valamilyen (alacsony) fokszámú polinommal helyettesítettük. A trapéz-szabálytól (elsőfokú) a Bode- (negyedfokú) integrálásig megismert formulák magasabb rendű polinomokra elvileg könnyen felírhatók, mégis ritkán használják azokat. Két oka is van a mellőzésnek:

- A magasabbrendű képletekben az együtthatók váltakozó előjelűek, a numerikus hibák nagyon felnőhetnek a sok taggal végzett műveletekben.
- A Gauss-Legendre integrálás a polinommal jól közelíthető függvényekre viszonylag kevés pontban elvégzett függvényszámolással is nagy pontosságú eredményt ad, ami különösen fontos több dimenziós integrálok számolásánál.

A módszer bemutatásához tekintsük az

$$I = \int_{-1}^1 f(x)dx \quad (1.49)$$

integrált. Alkalmas koordináta-transzformációval minden integrálunk ilyen alakra hozható. Az elemi kvadrátúra formulák szerint az integrál

$$I \approx \sum_{i=1}^N w_i f(x_i) \quad , \quad x_i = -1 + 2 \cdot \frac{i-1}{N-1} \quad (1.50)$$

közelítő összeggel számolható, ahol csak a  $w_i$  súlyok változnak a különböző eljárásokban. Nyilvánvaló, hogy ha az  $f(x)$  függvény  $N-1$ -edfokú polinom, akkor alkalmas súlyokkal egzaktul integrálhatjuk ezzel a formulával, hiszen az

$$\int_{-1}^1 x^p dx = \sum_{i=1}^N w_i x_i^p \quad , \quad p = 0, 1, \dots, N-1 \quad (1.51)$$

egyenletrendszer  $w_i$ -kre megoldható. Ha  $N$  páratlan, akkor ez igaz  $N$ -edfokú polinomra is.

Ha elengedjük a megszorítást, hogy  $x_i$ -k egy ekvidisztáns beosztás pontjai, akkor az  $N$  darab súly mellett az  $N$  darab  $x_i$  osztópont is szabadon variálható paraméter lesz. A súlyok és az osztópontok optimális megválasztásával elérhetjük, hogy  $2N-1$ -nél nem magasabb fokszámú polinomokra

$$I = \sum_{i=1}^N w_i f(x_i) \quad (1.52)$$

egzakt legyen. A feladat megoldásához most a  $P_i(x)$ ,  $i = 0, 1, 2, \dots$  Legendre-polinomokat használjuk.  $P_i(x)$  egy  $i$ -edfokú polinom,  $i$  darab zéróhelye van a  $[-1, 1]$  intervallumon. A Legendre-polinomok teljes ortogonális rendszert alkotnak a  $[-1, 1]$  intervallumon, azaz

$$\int_{-1}^1 P_i(x)P_j(x)dx = \frac{2}{2i+1}\delta_{ij} \quad (1.53)$$

Ha  $f(x) = p_{2N-1}(x)$  egy  $2N-1$ -nél nem magasabb fokszámú polinom (illetve ezzel jól közelíthető), akkor polinomosztással felírhatjuk, hogy

$$f(x) = p_{2N-1}(x) = p_{N-1}(x) \cdot P_N(x) + q_{N-1}(x) \quad (1.54)$$

ahol  $p_{N-1}(x)$  és  $q_{N-1}(x)$  polinomok fokszáma maximum  $N - 1$ . Az integrál ekkor

$$\int_{-1}^1 f(x)dx = \int_{-1}^1 \{p_{N-1}(x) \cdot P_N(x) + q_{N-1}(x)\} dx = \int_{-1}^1 q_{N-1}(x)dx \quad (1.55)$$

mert  $P_N(x)$  ortogonális minden  $N$ -nél kisebb fokszámú polinomra, így  $p_{N-1}(x)$ -re is. Ezzel a feladatot egy durván feleakkora fokszámú polinom integrálására redukáltuk, ami egy rögzített beosztás mellett a súlyok alkalmas megválasztásával egzaktul elvégezhető. A probléma csak az, hogy el kellene elvileg végeznünk a polinomosztást, hogy  $q_{N-1}(x)$  előálljon. Ha azonban előírjuk, hogy az  $x_i$ -k  $P_N(x)$  zéróhelyeire essenek, akkor

$$f(x_i) = q_{N-1}(x_i) \quad (1.56)$$

és

$$\int_{-1}^1 f(x)dx = \sum_{i=1}^N w_i q_{N-1}(x_i) = \sum_{i=1}^N w_i f(x_i) \quad (1.57)$$

tehát valójában nem is kell ismernünk  $q_{N-1}(x)$ -et.

A Legendre polinomok zéróhelyei megadhatók, az ehhez tartozó súlyok pedig megmutathatóan

$$w_i = \frac{2}{(1 - x_i^2) [P'_N(x_i)]^2} \quad (1.58)$$

módon választandók. Sima függvények, melyek véges intervallumon hatványsorba fejthetők jól közelíthetők polinomokkal, így azokra ez az integrálási séma nagy pontosságot ad.

Hasonló kvadratura formulák vezethetők le más típusú integrálokra egyéb ortogonális függvényekkel is, ilyenek például

$$\int_0^\infty e^{-x} f(x)dx \approx \sum_{i=1}^N w_i f(x_i) \quad , \quad L_N(x_i) = 0 \quad \text{ahol} \quad L_N(x) \text{ Laguerre-polinom} \quad (1.59)$$

$$\int_{-\infty}^\infty e^{-x^2} f(x)dx \approx \sum_{i=1}^N w_i f(x_i) \quad , \quad H_N(x_i) = 0 \quad \text{ahol} \quad H_N(x) \text{ Hermite-polinom} \quad (1.60)$$

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \approx \sum_{i=1}^N w_i f(x_i) \quad , \quad T_N(x_i) = 0 \quad \text{ahol} \quad T_N(x) \text{ Csebisev-polinom} \quad (1.61)$$

## 2. KÖZÖNSÉGES DIFFERENCIÁLEGYENLETEK (ODE: ORDINARY DIFFERENTIAL EQUATIONS)

### 2.1. ODE redukálása egyenletrendszerre

A közönséges  $N$ -edrendű differenciálegyenletek elsőrendű differenciálegyenlet rendszerre alakíthatók.

*Példa:* A másodrendű

$$\frac{d^2 y}{dx^2} + q(x) \frac{dy}{dx} = r(x) \quad (2.1)$$

egyenlet elsőrendű egyenletek rendszerévé átirható, mint

$$\frac{dy}{dx} = z(x) \quad \text{és} \quad \frac{dz}{dx} = r(x) - q(x)z(x) \quad (2.2)$$

ahol  $z(x)$  egy új változó.

Szokásos és nyilvánvaló választás olyan új változókat bevezetni, melyek egymás deriváljai, ezzel minden ODE átalakítható a kívánt alakra. Néha hasznos, hogy egyéb faktorokat beépítsünk az új változókba, ezzel nemcsak egyszerűbb egyenletekre juthatunk, de a megoldások várható szinguláris viselkedéséből származó numerikus problémákat is megelőzhetjük. Ha azt látjuk, hogy az eredeti egyenlet megoldása sima, de az önkényesen bevezett segédfüggvények örült dolgokat csinálnak, akkor célszerű a dolognak utánajárni és más változókat választani.

Az általános feladat tehát  $N$  csatolt elsőrendű differenciálegyenlet

$$\frac{d}{dx}y_i(x) = f_i(x, y_1(x), \dots, y_N(x)) \quad , \quad i = 1, \dots, N \quad (2.3)$$

megoldása. Az ismert  $f_i$  függvényeket és az ismeretlen  $y_i$ -ket 'vektor'-ba rendezve írhatjuk, hogy

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x, \mathbf{y}) \quad (2.4)$$

Egy jól határozott probléma nemcsak a vonatkozó differenciálegyenlet megadását jelenti, hanem ú.n. **kezdeti feltételek** (vagy határfeltételek) kijelölését is. A határfeltételek algebrai feltételek lehetnek  $y_i(x_a)$  értékeire bizonyos  $x_a$  pontokban. A legegyszerűbben kezelhető esetben megköveteljük, hogy egy változó adott pontban adott értéket vegyen fel:

$$y_i(x_a) = c_{i,a} \quad (2.5)$$

A modellezni kívánt probléma függvényében a feltételek igen bonyolultak is lehetnek, mint például a változók valamely pontban felvett értékei közötti nemlineáris egyenletrendszer.

A továbbiakban elsőrendű differenciálegyenletre és  $y(x = x_0) = y_0$  típusú kezdőfeltételre ismertetjük az alapvető numerikus módszereket. Magasabbrendű egyenletekkel és/vagy más típusú kezdő-, vagy határfeltételekkel elsősorban a példák tárgyalásakor fogunk találkozni.

A

$$\frac{dy}{dx} = f(x, y) \quad , \quad y(x_0) = y_0 \quad , \quad x_0 \leq x \quad (2.6)$$

közönséges differenciálegyenlet megoldását előbb ekvidisztáns beosztáson tárgyaljuk:

$$x_n = x_0 + nh \quad , \quad y_n = y(x_n) \quad , \quad n = 0, 1, \dots, M \quad (2.7)$$

majd hatékonyabb (adaptív) módszereket mutatunk be, ahol a beosztást (lépésközt) rugalmasan módosítjuk a számolás közben.

## 2.2. Euler-módszer

Az  $n$ -edik osztópontban a deriváltat a kétpontos előre-formulával helyettesítve

$$\frac{y_{n+1} - y_n}{h} + O(h) = \left. \frac{dy}{dx} \right|_{x=x_n} = f(x_n, y_n) \quad (2.8)$$

kapjuk a rekurzióra alkalmas

$$y_{n+1} = y_n + hf(x_n, y_n) + O(h^2) \quad (2.9)$$

egyenletet. Az aszimmetrikus formula  $h$  intervallumot léptet előre  $x_n$ -ről  $x_{n+1} = x_n + h$ -ra, eközben csak az intervallum kezdőpontjában levő  $f_n$  információt használja fel a derivált értékéhez. A lokális csomólási hiba csak egy renddel kisebb a megtenni kívánt  $y_{n+1} - y_n = \Delta y (\approx O(h))$  lépésnél. Ez gyakorlatilag azt jelenti, hogy a módszer használhatatlan.

Javíthat a helyzeten, ha a Taylor sorból több tagot veszünk figyelembe, pl.

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2}y''_n + O(h^3) \quad (2.10)$$

Behelyettesítve a

$$y' = \frac{dy}{dx} = f \quad \text{és} \quad y'' = \frac{df}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \frac{dy}{dx} = \frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y} \quad (2.11)$$

deriváltakat

$$y_{n+1} = y_n + hf_n + \frac{h^2}{2} \left[ \frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y} \right]_n + O(h^3) \quad (2.12)$$

jobb rekurziós egyenletet kaptunk. Persze ez csak akkor hasznos, ha  $f(x, y)$  parciális deriváltjai könnyen (analitikusan) számolhatók.

### 2.3. Pontosság

Az iménti közelítő eljárásokban egyetlen  $y_n \rightarrow y_{n+1}$  lépés során a csonkolási hiba tipikusan  $O(h^{m+1})$  viselkedésű. Miközben  $M \sim O(1/h)$  lépést teszünk a globális csonkolási hiba  $MO(h^{m+1}) \sim O(h^m)$  lesz. A megoldási módszert  **$m$ -edrendű módszernek** hívjuk, ha a globális csonkolási hiba  $O(h^m)$ .

Ha osztópontban a kijelölt műveletek elvégzésekor az aritmetikai hiba  $O(\varepsilon)$ , akkor az  $O(1/h)$  lépés során felhalmozódó összes aritmetikai hiba  $O(\varepsilon/h)$ . A teljes eljárás hibája tehát

$$\epsilon = O(\varepsilon/h) + O(h^m) \quad (2.13)$$

ami akkor a legkisebb, ha

$$h \sim \varepsilon^{\frac{1}{m+1}} \quad \epsilon \sim \varepsilon^{\frac{m}{m+1}} \quad (2.14)$$

Ennél nagyobb lépésköznél a sorfejtés hibája, kisebbnél az aritmetikai hiba dominál. Tegyük fel, hogy az aritmetikai hiba a lehető legkisebb, azaz csak a számítógép véges számábrázolásából ered: 8 byte esetén  $\varepsilon = \epsilon_m \sim 10^{-16}$ . Ekkor az optimális lépésköz és az elérhető pontosság a különböző rendű módszerekben a táblázat szerint alakul:

| m | $h$                 | $\epsilon$           |
|---|---------------------|----------------------|
| 1 | $1.0 \cdot 10^{-8}$ | $1.0 \cdot 10^{-8}$  |
| 2 | $4.6 \cdot 10^{-6}$ | $2.2 \cdot 10^{-11}$ |
| 3 | $1.0 \cdot 10^{-4}$ | $1.0 \cdot 10^{-12}$ |
| 4 | $6.3 \cdot 10^{-4}$ | $1.6 \cdot 10^{-13}$ |
| 5 | $2.2 \cdot 10^{-3}$ | $4.6 \cdot 10^{-14}$ |

(2.15)

### 2.4. Stabilitás

A differenciálegyenletek megoldásakor a numerikus hibák jelenléte nemcsak azért nem hagyható figyelmen kívül, mert a megoldásunk "kicsit" eltér a valótól. Az egymást követő lépések során a korábban begyűjtött hibák önálló életre kelnek és alapvetően hamis eredményre vezethetnek. Az eljárást úgy kell megterveznünk, hogy az stabilis legyen az elkerülhetetlen lépésenkénti kis kis hibák mellett is.

Az Euler-integrálás példáján megmutatjuk, hogy valamely korábbi pontban begyűjtött hiba hogyan 'fejlődik' a további lépések során. Tegyük fel, hogy az "igazi"  $y_n$  helyett  $y_n + \delta y_n$  hibával terhelt értékről kívánunk továbblépni Euler-módszerrel. Ekkor

$$y_{n+1} + \delta y_{n+1} = y_n + \delta y_n + h \left[ f(x_n, y_n) + \frac{\partial f}{\partial y} \Big|_n \delta y_n \right] \quad (2.16)$$

miközben

$$y_{n+1} = y_n + hf(x_n, y_n) \quad (2.17)$$

A hiba ebből

$$\delta y_{n+1} = \left[ 1 + h \frac{\partial f}{\partial y} \Big|_n \right] \delta y_n \quad (2.18)$$

ami csökkenő görgetett hibát jelent, ha

$$\left| 1 + h \frac{\partial f}{\partial y} \right|^2 < 1 \quad (2.19)$$

egyébként a hiba nő. Az  $f(x, y)$  parciális derivált nagyságának ( $f$  általában komplex függvény) ismeretében eldönthetjük, hogy lehet-e stabil az eljárás, illetve, hogy mekkora lépésközzel dolgozzunk ennek elérésére.

Példa:  $\alpha > 0$  valós szám esetén

$$\frac{dy}{dx} = -\alpha y \quad \frac{\partial f}{\partial y} = -\alpha \quad \text{stabil, ha} \quad h < \frac{2}{\alpha} \quad (2.20)$$

$$\frac{dy}{dx} = +\alpha y \quad ; \quad \frac{dy}{dx} = \pm i\alpha y \quad \text{instabil} \quad (2.21)$$

## 2.5. Többlépéses prediktor módszerek

Az Euler-módszerrel csak az  $n$ -edik pontbeli értékeket felhasználva léptünk az  $x_{n+1}$  helyre. Az  $x_n \rightarrow x_{n+1}$  lépéskor persze ismerjük már a korábbi  $(\dots, y_{n-2}, y_{n-1}, y_n)$  értékeket, feltéve, hogy nem az első lépés(ek)nél tartunk. Célszerűnek tűnik, hogy ezt az információt felhasználjuk az aktuális lépésben. Az ilyen típusú eljárásokat **prediktor** módszereknek nevezzük, ugyanis a korábbi pontokban megismert információ alapján jósoljuk meg a következő pontbeli értéket.

A többlépéses eljárások bemutatásához integráljuk formálisan a differenciálegyenletet

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y) dx \quad (2.22)$$

Igaz, hogy nem ismerjük  $y(x)$ -et (így  $f(x, y)$ -t sem) az integrációs tartományon, de a korábbi lépéseinkből az  $y_{n-1}$  és  $y_n$  értékét tudjuk. Ezen két érték segítségével használunk lineáris extrapolációt az integrandusra:

$$f(x) = \frac{x - x_{n-1}}{h} \cdot f_n - \frac{x - x_n}{h} \cdot f_{n-1} + O(h^2) \Rightarrow f_{n+1} \approx 2f_n - f_{n-1} \quad (2.23)$$

és a Simpson-szabály alapján integráljunk:

$$\int_{x_n}^{x_{n+1}} f(x, y) dx = h \frac{f_n + f_{n+1}}{2} = \frac{h}{2} [3f_n - f_{n-1}] \quad (2.24)$$

Így kapjuk az **Adam-Bashforth-féle kétlépéses** (prediktor) módszert

$$y_{n+1} = y_n + \frac{h}{2} [3f_n - f_{n-1}] + O(h^3) \quad (2.25)$$

**Háromlépéses** módszerhez jutunk, ha az

$$f(x) = f_{n-1} + (x - x_{n-1}) \cdot \frac{f_n - f_{n-2}}{2h} + \frac{(x - x_{n-1})^2}{2} \cdot \frac{f_n - 2f_{n-1} + f_{n-2}}{h^2} + O(h^3) \quad (2.26)$$

extrapolációt használjuk és a másodrendű görbét integráljuk a jelzett intervallumon

$$\int_{x_n}^{x_{n+1}} f(x, y) dx = \frac{h}{12} [5f_{n-2} - 16f_{n-1} + 23f_n] \quad (2.27)$$

azaz

$$y_{n+1} = y_n + \frac{h}{12} [5f_{n-2} - 16f_{n-1} + 23f_n] + O(h^4) \quad (2.28)$$

Végül a megfelelő **néglépéses** formula

$$y_{n+1} = y_n + \frac{h}{24} [55f_n - 59f_{n-1} - 37f_{n-2} - 9f_{n-3}] + O(h^5) \quad (2.29)$$

A többlépéses eljárások indításához nem elég egyetlen pont (kezdőfeltétel). Egyéb módszerrel, pl. az Euler módszerrel előbb le kell gyártani a szükséges számú induló pontot.

## 2.6. Implicit módszerek

Eddig  $y_{n+1}$  értékét explicite kifejeztük a korábbi adatokból. Az implicit módszerekben  $y_{n+1}$  értékére egy (nemlineáris) implicit egyenletet állítunk fel, ezt az egyenletet még külön meg kell oldani  $y_{n+1}$  explicit előállításához.

Tekintsük egy  $x_{n+\frac{1}{2}} = x_0 + (n + \frac{1}{2})h$  pontot az osztópontok között félúton, ahol is

$$\left. \frac{dy}{dx} \right|_{x_{n+\frac{1}{2}}} = f(x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}}) = f_{n+\frac{1}{2}} \quad (2.30)$$

Használjunk szimmetrikus formulát az  $x_{n+\frac{1}{2}}$  pontban a deriváltra a baloldalon

$$\left. \frac{dy}{dx} \right|_{x_{n+\frac{1}{2}}} = \frac{y_{n+1} - y_n}{h} + O(h^2) = f_{n+\frac{1}{2}} \quad (2.31)$$

a baloldalon pedig helyettesítsük  $f_{n+\frac{1}{2}}$  értékét az osztópontbeli értékek számtani közepével

$$f_{n+\frac{1}{2}} = \frac{f_n + f_{n+1}}{2} + O(h^2) \quad (2.32)$$

Mindkét oldalon közelítés hibája  $O(h^2)$ . Az így kapott **Heun**-rekurzió

$$y_{n+1} = y_n + \frac{h}{2} [f(x_n, y_n) + f(x_{n+1}, y_{n+1})] + O(h^3) \quad (2.33)$$

az ismeretlen  $y_{n+1}$  értéket mindkét oldalon tartalmazza. Konkrét differenciálegyenletekre (adott  $f$  mellett) az egyenlet analitikusan, vagy numerikusan (gyökkereső algoritmusokkal) megoldható.

Az **Adams-Bashforth-Moulton** módszer egyszerre többlépéses és implicit. Másodfokú polinomot illesztünk a függvény  $f_{n-1}, f_n$  és  $f_{n+1}$  értékeire majd kapjuk az

$$y_{n+1} = y_n + \frac{h}{12} [5f_{n+1} + 8f_n - f_{n-1}] + O(h^4) \quad (2.34)$$

kétlépéses formulát. Harmadfokú illesztéssel a három lépéses formula

$$y_{n+1} = y_n + \frac{h}{24} [9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}] + O(h^5) \quad (2.35)$$

Az implicit egyenletek megoldásának egyik módszere a **prediktor-korrektor** eljárás. Előbb valamely 'prediktor' eljárással megbecsüljük  $y_{n+1}$  értékét, majd ezt az értéket helyettesítjük az egyenlet jobb oldalán  $f_{n+1} = f(x_{n+1}, y_{n+1})$ -be, hogy ezzel a bal oldalon 'korrigált'  $y_{n+1}$ -et állítsunk elő. Konkrétan például a

$$y_{n+1} = y_n + \frac{h}{12} [5f_{n-2} - 16f_{n-1} + 23f_n] + O(h^4) \quad (2.36)$$

háromlépéses (2.28) explicit Adams-Bashforth prediktor képlettel kiszámoljuk  $y_{n+1}$ , jelöljük ezt a műveletet **P**-vel, majd az **F**-fel jelzett lépésben kiszámítjuk  $f_{n+1}$ -et és **C**-vel jelölve alkalmazzuk az Adams-Moulton korrektor (2.34) formulát. Az utóbbi két lépést  $m$ -szer végrehajtva az eljárás szimbolikusan

$$\mathbf{P}(\mathbf{FC})^m \quad (2.37)$$

alakba írható. A korrektor lépésekben figyeljük a konvergenciát, ha  $y_{n+1}$  már nem változik a hibahatárnál nagyobb mértékben, akkor leállunk. Megjegyzés: Ha  $m$ -et nem a hiba figyelésével állapítjuk meg minden pontban, hanem valamely előre megadott értéken rögzítjük, akkor valójában explicit prediktor eljárást csináltunk, hiszen a végső  $y_{n+1}$  szám a korábbi pontokban számolt értékek (esetleg kicsit bonyolult) kifejezéseként elvileg felírható.

## 2.7. Runge-Kutta-módszer

A Runge-Kutta-módszer sem nem gyorsabb, sem nem pontosabb, mint pl. a prediktor-korrektor módszer, de egyszerűsége és viszonylagos stabilitása miatt előszeretettel használják.

Az implicit módszer mintájára írjuk fel, hogy

$$\left. \frac{dy}{dx} \right|_{x_{n+\frac{1}{2}}} = \frac{y_{n+1} - y_n}{h} + O(h^2) = f(x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}}) \quad (2.38)$$

majd a baloldalon helyettesítsük  $f_{n+\frac{1}{2}}$  értékét

$$y_{n+\frac{1}{2}} = y_n + \frac{h}{2} f(x_n, y_n) + O(h^2) \implies f(x_{n+\frac{1}{2}}, y_{n+\frac{1}{2}}) = f(x_{n+\frac{1}{2}}, y_n + \frac{h}{2} f(x_n, y_n)) + O(h^2) \quad (2.39)$$

amivel

$$y_{n+1} = y_n + hf(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}) + O(h^3) \quad (2.40)$$

Ez a **másodrendű RK** eljárás, amit célszerűen három utasításra bontva írunk fel:

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}) \\ y_{n+1} &= y_n + k_2 + O(h^3) \end{aligned} \quad (2.41)$$

Teljesen hasonlóan kapható a **harmadrendű RK**

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}) \\ k_3 &= hf(x_n + h, y_n - k_1 + 2k_2) \\ y_{n+1} &= y_n + \frac{1}{6}(k_1 + 4k_2 + k_3) + O(h^4) \end{aligned} \quad (2.42)$$

És végül a **negyedrendű RK** négy logikai lépése:

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}) \\ k_3 &= hf(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}) \\ k_4 &= hf(x_n + h, y_n + k_3) \\ y_{n+1} &= y_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} + O(h^5) \end{aligned} \quad (2.43)$$

Magasabb rendet nem célszerű felírni, ugyanis látható, hogy egy  $n$ -edrendű RK-ban  $n$ -szer kell  $f(x, y)$  függvényértéket számolni, ami nagyon költséges lehet. Ennek csak akkor van haszna, ha egyúttal nagyobb lépésközt ( $h$ -t) is választhatunk.

A korábban bevezetett elnevezések szerint az RK módszer: **egylépéses, prediktor és explicit**. Igaz, hogy RK lépésekről beszélünk, többesszámban, ezek azonban csak az algoritmusban értendők. Az előző fejezet "többlépéses" módszerivel szemben, illetve abban az értelemben az RK eljárás egylépéses, hiszen nem használjuk fel a korábbi osztópontokban begyűjtött információt.

## 2.8. Adaptív lépések a differenciálegyenletek integrálásakor

Eddig a megoldást a teljes megcélzott intervallumon egyetlen előre kiválasztott  $h$  lépésközzel állítottuk elő. A meghatározandó függvény viselkedése viszont nagyon különbözhet az egyes résztartományokon. Olyan tartományban, ahol a függvény sima a kis lépésköz feleslegesen sok számolási munkát, indokolatlanul halmozódó kerekítési hibát okoz. Ahol meg a függvény gyorsan változik, finomabb beosztást illene használni.

Kézenfekvő megoldás minden lépésnél megvizsgálni az  $f(x_n, y_n)$  elsőrendű (és esetleg magasabb) deriváltjainak a nagyságát és ennek megfelelően menet közben változtatni lépésközt. Kvalitatíve jó lehet ez az eljárás, de nehéz általános kvantitatív szabályt a derivált alapján felállítani, arról nem is szólva, hogy ezen deriváltaknak az előállítása (ha egyáltalán lehetséges) nagyon munkáigényes feladat.

Univerzálisabb megoldás, ha a lépésközt valamely tervezett hibához igazítjuk. Tegyük fel, hogy a lépésenkénti hibát szeretnénk valamilyen optimális  $\epsilon_{opt}$  értéken tartani. Ha sikerülne az egyes lépések során kis többletmunka árán becslést kapni arra, hogy mekkora a lépés  $\epsilon_h$  hibája, akkor  $h$ -t menet közben úgy növelhetnénk/csökkenthetnénk, hogy  $\epsilon_h \approx \epsilon_{opt}$  legyen.

*Példa:* Lépésduplázós negyedrendű Runge-Kutta-eljárás

Az aktuális RK lépésben  $x_n$ -ről indulva kétféleképpen is ellépünk  $x_n + 2h$ -ra:

- **a)** az aktuális  $h$ -val kétszer lépünk



- b) majd  $2h$ -val egyetlen lépéssel.

Az egzakt  $y(x + 2h)$  értékről és az iménti lépésekben kapott két RK  $O(h^5)$  rendben pontos eredményről feltehetjük, hogy

$$y(x + 2h) = y_a + 2(h)^5 \cdot \Phi + O(h^6) \quad (2.44)$$

$$y(x + 2h) = y_b + (2h)^5 \cdot \Phi + O(h^6) \quad (2.45)$$

ahol  $\Phi$  nagyjából ugyanakkor vehető a két egyenletben, egyébként nagyságrendben  $\Phi \approx y^{(5)}(x)/5!$ . A három szóbanforgó lépés hibája vezető rendben tehát

$$\epsilon_h = h^5 \cdot \Phi, \quad \epsilon_a = 2h^5 \cdot \Phi = 2\epsilon_h \quad \text{és} \quad \epsilon_b = 32h^5 \cdot \Phi = 32\epsilon_h \quad (2.46)$$

Ezekből az aktuális lépésköz melletti tipikus hiba hatodrendben előáll:

$$\epsilon_h = \frac{1}{30} |\epsilon_b - \epsilon_a| = \frac{1}{30} |y_a - y_b| + O(h^6) \quad (2.47)$$

Ugyan a hatodrendű tag nem feltétlenül jelent elhanyagolható mennyiséget, mégis általában elfogadhatjuk, hogy az  $O(h^6)$  korrekciótól eltekintsünk. Mostmár megvizsgálhatjuk, hogy mekkora  $h_{opt}$  lépésköze érdemes áttérni, hiszen ötödrendű hibagról lévén szó:

$$\left(\frac{h_{opt}}{h}\right)^5 = \frac{\epsilon_{opt}}{\epsilon_h} \quad h_{opt} = h \left(\frac{\epsilon_{opt}}{\epsilon_h}\right)^{1/5} \quad (2.48)$$

Ha  $h_{opt}$  nem tér el lényegesen  $h$ -tól, akkor megtartjuk a kiszámolt  $y_{n+2} = y_a$  számot, ha lényegesen kisebb, akkor újra számolunk kisebb lépésközzel, ha lényegesen nagyobb, akkor a következő lépésekben növelt lépéshosszal próbálkozunk. Természetesen vigyázzunk arra, hogy  $h_{opt}$  ne legyen sem túl nagy, sem túl kicsiny.

A vázolt eljárás túlmunkával jár, hiszen egy RK lépés során 4 alkalommal kell függvényértéket számolni. A két  $h$  nagyságú lépéshez szükséges 8 függvényhívás helyett most a  $2h$  hosszú extra lépés miatt további három közbülső pontban, azaz összesen 11 alkalommal számoltuk ki  $f(x, y)$ -t. Ez mintegy 1.3-szoros számolási túlmunkát jelent. A 'rázós' intervallumokon azonban nem nő meg ellenőrizetlenül a hiba, míg a 'sima' szakaszokon olyan nagy lépésekkel haladhatunk, hogy összességében nagyságrendekkel csökkenhet a számolási munka.

*Példa:* Beágyazott Runge-Kutta

Egy  $M$ -edrendű Runge Kutta eljárásban a számolási séma

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + a_2 h, y_n + b_{2,1} k_1) \\ &\vdots \\ k_{\tilde{M}} &= hf(x_n + a_{\tilde{M}} h, y_n + b_{\tilde{M},1} k_1 + \dots + b_{\tilde{M},\tilde{M}-1} k_{\tilde{M}-1}) \\ y_{n+1} &= y_n + \sum_{i=1}^{\tilde{M}} c_i k_i + O(h^{M+1}) \end{aligned} \quad (2.49)$$

ahol  $M \leq \tilde{M} \leq M + 2$ . Ha  $M > 4$ , akkor lehetőség van arra, hogy másképp is elvégezzük a lépést ugyanazokkal a  $k_i$  számokkal de más együttthatókkal úgy, hogy egy "beágyazott"

$$\tilde{y}_{n+1} = y_n + \sum_{i=1}^M \tilde{c}_i k_i + O(h^M) \quad (2.50)$$

alacsonyabb rendű formulát felírhatunk. Megfelelő együttthatókat (táblázatosan adott) választva az

$$O(h^M) \approx |y_{n+1} - \tilde{y}_{n+1}| \quad (2.51)$$

hiba kiszámolható, ebből  $\epsilon_h$  és  $h_{opt}$  számolható. Az egyik leggyakoribb ilyen beágyazott RK eljárás a Fehlberg 4-5 rendű Runge-Kutta módszere, ahol az ötödrendű RK mellett ugyanazon függvényértékekkel egy negyedrendű lépést is teszünk és ebből becsljük a hibát. Ilyen a MAPLE **rkf45** rutinja.

### 2.9. Hiányos másodrendű differenciálegyenletek

A magasabb rendű DE-k valóban redukálhatók elsőrendű DE-k rendszerére, néhány speciális, de gyakori esetben az mégsem célszerű. Ilyen a másodrendű

$$y'' = f(x, y) \quad , \quad y(x_0) = y_0 \quad , \quad y'(x_0) = z_0 \quad (2.52)$$

kezdetiérték-probléma, ahol az első derivált hiányzik az egyenletből. A szokásos beosztást véve a második deriváltat a hárompontos formulával felírhatjuk, ezzel látszólag készen is lennénk

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f_n + O(h^4) \quad (2.53)$$

Láthatóan van egy kis probléma, az adott kezdőfeltételek mellett  $y_{-1}$  hiányában nem tudjuk elindítani a lépéseket. A jelenlegi kezdőfeltétel mellett

$$y_1 = y_0 + y'(x_0)h + y''(x_0)\frac{h^2}{2} + O(h^3)$$

miatt az eljárás egyszerűen megadható

$$y_1 = y_0 + z_0 h + f_n \frac{h^2}{2} + O(h^3) \quad (2.54)$$

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f_n + O(h^4) \quad , \quad n = 1, 2, \dots, M-1 \quad (2.55)$$

$$y'_M = \frac{y_M - y_{M-1}}{h} + f_n \frac{h}{2} + O(h^2) \quad (2.56)$$

Az utolsó sor az

$$y_{M-1} = y_M - y'_M h + y''_M \frac{h^2}{2} + O(h^3) \quad (2.57)$$

sorfejtés következménye, amire nyilván csak akkor van szükség, ha a megoldásban az  $y'(x)$  függvényt is elő kell állítanunk. A kerekítési hibák csökkentése végett célszerű az előbbi egyenleteket átírni:

$$\Delta_0 = h \left[ z_0 + f_n \frac{h}{2} \right] \quad , \quad y_1 = y_0 + \Delta_0 \quad (2.58)$$

$$\Delta_n = \Delta_{n-1} + h^2 f_n \quad , \quad y_{n+1} = y_n + \Delta_n \quad , \quad n = 1, 2, \dots, M-1 \quad (2.59)$$

$$y'_M = \frac{\Delta_{M-1}}{h} + f_n \frac{h}{2} \quad (2.60)$$

### 3. KEZDETIÉRTÉK- VERSUS PEREMÉRTÉK-FELADAT

Az eddig megismert ODE megoldó módszerek alkalmazása magától értetődő, ha kezdetiérték-problémákról (**IVP**: Initial Value Problem) van szó. Ha kezdőfeltételek adottak valamilyen pontban, onnan a rekurziók segítségével a megfelelő partikuláris megoldást lépcsőről-lépésre elő tudjuk állítani a kívánt intervallumon.

Sokkal nehezebb feladat az úgynevezett peremérték-probléma (**BVP**: Boundary Value Problem). Ezekben az esetekben nem adjuk meg a független változó valamely értékénél a partikuláris megoldáshoz szükséges összes határfeltételt, hanem a keresett megoldásnak több pontban kirótt feltételeket kell kielégítenie. A legegyszerűbb(?) esetben az  $N$ -edrendű differenciálegyenlethez

$$\frac{d}{dx} y_i(x) = f_i(x, y_1, \dots, y_N) \quad , \quad i = 1, \dots, N \quad (3.1)$$

$n_1$  kezdőfeltételt adunk egy  $x_a$  pontban:

$$A_k(x_a ; y_1, \dots, y_N) = 0 \quad k = 1, \dots, n_1 \quad (3.2)$$

és  $n_2 = N - n_1$  feltételt egy másik  $x_b$  pontban (gyakran a végpontban)

$$B_k(x_b ; y_1, \dots, y_N) = 0 \quad k = 1, \dots, n_2 \quad (3.3)$$

Ha a megoldás az  $[x_a, x_b]$  intervallumon keresendő, akkor annak az intervallum peremén kell kielégítenie bizonyos feltételeket, innen származik a probléma elnevezése. Valójában a partikuláris megoldástól megkövetelt viselkedést ennél bonyolultabb összefüggések határozhatják meg. Ha nem kezdetiérték-feladatról van szó, akkor az általános esetben is peremérték-problémáról beszélünk.

A mostani rövid tárgyalásban csak ízelítőt adunk speciális BVP esetekre.

### 3.1. Lövdőzős (SHOOTING) módszer

Az egyik végpontban adott hiányos kezdeti feltételeket a szükséges számú "értelmes" de amúgy ízlés szerinti további  $V_1, \dots, V_{n_2}$  kezdőértékekkel kiegészítjük, hogy a differenciálegyenlethez az összes kezdeti érték

$$y_i^V(x_a) = y_i(x_a; V_1, \dots, V_{n_2}) \quad (3.4)$$

előálljon. Az IVP így megoldható, de megoldása nyilván a  $V$ -k paraméteres függvénye. Csaknem biztosak lehetünk abban, hogy a másik határon kirótt feltételeket ez a próba-megoldás nem fogja teljesíteni, azaz valószínűleg

$$E_k(V_1, \dots, V_{n_2}) = B_k(x_b; y_1^V, \dots, y_N^V) \neq 0 \quad (3.5)$$

Változtatunk a találmányra felvett kezdőfeltételeken, és figyeljük, hogy a megoldás hogyan viselkedik a másik végen. Alkalmas (pl. többdimenziós gyökkereső) algoritmussal addig próbálkozunk, míg az eredeti feltételek ki nem elégülnek, azaz

$$|E_k(V_1, \dots, V_{n_2})| < \epsilon \quad k = 1, \dots, n_2 \quad (3.6)$$

nem teljesül. Az  $n_2$  **számú ismeretlen**  $V_1, \dots, V_{n_2}$  paraméter így meghatározható.

Gyakran nem célszerű (vagy nem is lehet) az intervallum egyik végéből a másikba integrálni az önkényes kezdőfeltételekkel. Szokásos eljárás, hogy mindkét végen kiegészítjük önkényesen a feltételeket

$$y_i^V(x_a) = y_i(x_a; V_1, \dots, V_{n_2}) \quad (3.7)$$

$$y_i^W(x_b) = y_i(x_b; W_1, \dots, W_{n_1}) \quad (3.8)$$

majd balról jobbra és jobbról balra egyaránt integrálunk valamilyen köztes  $x_c$  pontig, ahol a két megoldás találkozik. Itt meg kell megvizsgálni, hogy azok megegyeznek-e

$$E_i(V_1, \dots, V_{n_2}, W_1, \dots, W_{n_1}) = y_i^V(x_c) - y_i^W(x_c) \stackrel{?}{\approx} 0 \quad i = 1, \dots, N \quad (3.9)$$

Addig keresgetjük az  $N$  **számú ismeretlen**  $V_1, \dots, V_{n_2}, W_1, \dots, W_{n_1}$  paramétert, míg az illeszkedés kielégítő nem lesz.

### 3.2. Relaxációs módszerek

A lövdőzős módszerekben a numerikus megoldásunk (szakaszonként) kielégíti a differenciálegyenletet, de eleinte nem tesz eleget a peremfeltételeknek. A relaxációs módszerek más stratégiával dolgoznak. Előbb megtippeljük a "megoldást", ez a tipp általában sem a differenciálegyenletet, sem a peremfeltételeket nem elégíti ki. Ezután egy iteratív eljárás során deformáljuk a próbafüggvényeket, miközben azok egyre inkább megoldásai lesznek a kitűzött feladatnak. Részletesebben a parciális differenciálegyenletek (PDE) tárgyalásakor fogjuk a módszert elemezni, ott ugyanis a lövdőzés már nem vezet eredményre.

Itt csak megemlítjük, hogy nagyjából hogyan kellene eljárni az ODE relaxációs módszerben. Írjuk át a

$$\frac{dy}{dx} = \mathbf{f}(x, \mathbf{y}) \quad (3.10)$$

differenciálegyenletet véges differenciákra valamilyen  $\{x_1, x_2, \dots, x_M\}$  beosztást tekintve a független változóban. Az egyenleteink az

$$\mathbf{y}_n = \begin{pmatrix} y_1(x_n) \\ y_2(x_n) \\ \vdots \\ y_N(x_n) \end{pmatrix} \quad (3.11)$$

oszlopvektorral például az

$$\frac{\mathbf{y}_n - \mathbf{y}_{n-1}}{x_n - x_{n-1}} = \tilde{\mathbf{f}}\left(\frac{x_n + x_{n-1}}{2}, \frac{\mathbf{y}_n + \mathbf{y}_{n-1}}{2}\right) \quad (3.12)$$

implicit alakba írhatók. Azért  $\tilde{\mathbf{f}}$ , mert a differenciálegyenletet csak a belső osztópontokban vehetjük komolyan. A peremen, amikor  $n = 1$ , vagy  $n = M$  az  $\mathbf{f}(x, \mathbf{y})$  helyett a peremfeltételeket kell figyelembe venni.

Ezt az egyenletet (vagy ennek célszerűen átalakított változatát) most úgy fogjuk fel, mint az ismeretlen  $\{y_i(x_n) = y_{i,n} \mid i = 1..N, n = 1..M\}$  számokra vonatkozó algebrai egyenletrendszer. Az egyenletrendszer elvileg az  $\approx N \times M$  **számú ismeretlen**  $y_{i,n}$  értékre megoldható. Az ismeretlenek óriási száma miatt ez praktikusán többnyire csak iteratív módon végezhető el. A részletekkel most nem foglalkozunk.

### 3.3. Egy BVP: Az egydimenziós Schrödinger-egyenlet kötött állapotai

Az egydimenziós Schrödinger-féle sajátértékegyenlet

$$H\varphi_\mu(x) = E_\mu\varphi_\mu(x) \quad , \quad \mu = 0, 1, 2, \dots \quad (3.13)$$

ahol

$$H = -\frac{\partial^2}{\partial x^2} + V(x) \quad \left( \frac{\hbar}{2m} = 1 \text{ egységekben} \right) \quad (3.14)$$

A feladat az, hogy adott  $V(x)$  potenciál mellett keressük meg, hogy milyen  $E_\mu$  energia-sajátértékekkel léteznek az adott peremfeltételt kielégítő (saját-)megoldások. A feladat  $x$ -ben végtelen tartományon is érdekes, a numerikus eljárásunkhoz mégis szükséges, hogy egy véges intervallumra szorítkozzunk. Alkalmas koordináta-transzformációval elérhető, hogy a vizsgált tartomány az  $x \in [-1, 1]$  intervallum legyen.

A kötött állapotokra a

$$\varphi_\mu(\pm 1) = 0 \quad (3.15)$$

peremfeltételt írjuk elő. A megoldandó BVP tehát

$$\varphi''(x) = f(x)\varphi(x) \quad , \quad f(x) = V(x) - E \quad (3.16)$$

$$\varphi(-1) = 0 = \varphi(1) \quad (3.17)$$

Ez egy közönséges másodrendű differenciálegyenlet, elvileg a korábban használt ODE megoldók bármelyikét elővehetnénk a tárgyaláshoz. Ha figyelembe vesszük, hogy az egyenlet hiányos, hiszen  $\varphi'$  nem szerepel benne, akkor azt is tudjuk, hogy nem szükséges egyenletrendszerre alakítani, hanem közvetlenül diszkrétizálható. Mivel a jobb oldalon az ismeretlen függvény speciálisan (szorzóként) szerepel, célszerű a feladatra szabott integrálási eljárást felírni. Ilyen a **Numerov**-rekurzió, amit az alábbiakban vezetünk le.

A Taylor-sorból felírható, hogy

$$\varphi_{n+1} - 2\varphi_n + \varphi_{n-1} = [\varphi'']_n h^2 + \frac{h^4}{12} [\varphi''']_n + O(h^6) \quad (3.18)$$

Figyelembe véve a második deriváltat meghatározó differenciálegyenletet

$$[\varphi'']_n = [f\varphi]_n \quad \text{és} \quad [\varphi''']_n = [(f\varphi)']_n \approx \left[ \frac{(f\varphi)_{n+1} - 2(f\varphi)_n + (f\varphi)_{n-1}}{h^2} \right] \quad (3.19)$$

kapjuk, a rekurzióra alkalmas

$$\varphi_{n+1} - 2\varphi_n + \varphi_{n-1} = [f\varphi]_n h^2 + \frac{h^4}{12} \left[ \frac{(f\varphi)_{n+1} - 2(f\varphi)_n + (f\varphi)_{n-1}}{h^2} \right] + O(h^6) \quad (3.20)$$

hatodrendben pontos formulát. Némi átrendezéssel

$$\left[ 1 - \frac{h^2}{12} f_{n+1} \right] \varphi_{n+1} - \left[ 2 + \frac{10h^2}{12} f_n \right] \varphi_n + \left[ 1 - \frac{h^2}{12} f_{n-1} \right] \varphi_{n-1} \approx 0 \quad (3.21)$$

Tömörebb írással ez

$$[1 - T_{n+1}] \varphi_{n+1} - [2 + 10T_n] \varphi_n + [1 - T_{n-1}] \varphi_{n-1} \approx 0 \quad \text{ahol} \quad T(x) = \frac{h^2}{12} f(x) \quad (3.22)$$

Az egyenlet a

$$\frac{[1 - T_{n+1}] \varphi_{n+1}}{[1 - T_n] \varphi_n} + \frac{[1 - T_{n-1}] \varphi_{n-1}}{[1 - T_n] \varphi_n} = \frac{[2 + 10T_n]}{[1 - T_n]} \quad (3.23)$$

alakban alkalmasabb a probléma tárgyalására, hiszen a következő rekurziókra vezet

$$\text{előre:} \quad R_{n+1}^+ = U_n - 1/R_n^+ \quad \text{vagy} \quad \text{hátra:} \quad R_{n-1}^- = U_n - 1/R_n^- \quad (3.24)$$

ahol

$$R_{n+1}^+ = \frac{[1 - T_{n+1}] \varphi_{n+1}}{[1 - T_n] \varphi_n} \quad , \quad R_{n-1}^- = \frac{1}{R_n^+} = \frac{[1 - T_{n-1}] \varphi_{n-1}}{[1 - T_n] \varphi_n} \quad , \quad U = \frac{[2 + 10T]}{[1 - T]} \quad (3.25)$$

Ha a beosztást úgy választottuk, hogy  $x_1 = -1$  és  $x_M = 1$ , akkor a peremfeltételek a meghatározandó  $R^\pm$  mennyiségre

$$\varphi_1 = 0 \rightarrow R_2^+ = \infty \quad \text{és} \quad \varphi_M = 0 \rightarrow R_{M-1}^- = \infty \quad (3.26)$$

Balról indulva az előre rekurziót tudjuk hajtani ( $[R_2^+] \rightarrow R_3^+ \rightarrow R_4^+ \rightarrow \dots$ ) és a jobb oldali peremtől pedig a hátra rekurziót ( $[R_{M-1}^+] \rightarrow R_{M-2}^+ \rightarrow R_{M-3}^+ \rightarrow \dots$ ) használhatjuk. A példaprogramban balról haladunk mindaddig, míg  $R_n^+ \leq 1$ . Az  $R_K^+ \approx 1$  osztópontnál megállunk az előre rekurzióval és a másik szélről belépegetünk ehhez a találkozási ponthoz. A találkozásnál kiszámolt

$$|R_{K-}^+ - 1/R_{K-}^-| = \Delta(E) \quad (3.27)$$

mennyiség alkalmas annak az eldöntésére, hogy az aktuális energia sajátenergia-e, hiszen sajátállapotban

$$\Delta(E_\mu) \approx 0 \quad (3.28)$$

- Az alapállapot (és így minden sajátállapot) energiájára alsó korlátot adhatunk, hiszen (*miért?*)

$$E_0 \geq V_{min} \frac{\pi^2}{4} \quad \text{ahol} \quad V_{min} = \min \{V(x) \mid x \in [-1, 1]\} \quad (3.29)$$

- Az 1-dimenziós Schrödinger-egyenlet kötött állapotairól tudjuk, hogy az energiájuk szerint növekvő sorba rendezett  $\{\varphi_\mu(x) \mid \mu = 0, 1, 2, \dots\}$  függvényekre igaz, hogy  $\varphi_\mu(x)$ -nek pontosan  $\mu$  darab belső zéróhelye van. A Numerov-integrálás során megszámláljuk, hogy az aktuális  $E$  érték mellett  $R$  hány osztópontnál negatív, ez adja a zéróhelyek számát. Tekintettel arra, hogy "normális" körülmények között

$$T(x) = \frac{h^2}{12} [V(x) - E] < 1$$

$R_{n-1}^-$  negatív értéke valóban azzal kapcsolatos, hogy  $\varphi_{n-1}$  és  $\varphi_n$  ellenkező előjelű.

- Az  $m = 0, 1, 2, \dots, NodesMax$  adott számú zéróhellyel rendelkező megoldás keresésekor az energiát növeljük, ha a próba során kevesebb zéróhelyet és csökkentjük, ha többet találtunk. Ha a zéróhelyek száma megfelelő (és bízunk abban, hogy közel járunk az igazi sajátértékhez) lineáris extrapolációval keressük a  $\Delta(E) = 0$  egyenlet gyökét. Ez az eljárás alaposan eltévedhet, de nem volt most cél, hogy univerzális (idiótabiztos) eljárást adjunk.

## 4. PARCIÁLIS DIFFERENCIÁLEGYENLETEK

### 4.1. A lineáris egyenletek osztályozása

A fizikában leggyakrabban másodrendű parciális differenciálegyenletek megoldása válik szükségessé. A lineáris 2 dimenziós egyenletek általános alakja

$$\left[ a \frac{\partial^2}{\partial x^2} + 2b \frac{\partial^2}{\partial x \partial y} + c \frac{\partial^2}{\partial y^2} + d \frac{\partial}{\partial x} + e \frac{\partial}{\partial y} + f \right] V(x, y) = g \quad (4.1)$$

Az  $a, b, c, d, e, f, g$  együtthatók maguk is  $(x, y)$  adott függvényei lehetnek. Ha az együtthatók függenek az ismeretlen  $V(x, y)$ -től is, akkor kvázilineáris egyenletről beszélünk. Konstans együtthatók esetén a három másodrendű parciális derivált együtthatói alapján a következő táblázat szerint csoportosítjuk az egyenleteket:

| feltétel   | egyenlet típusa | jellegzetes fizikai problémakör | tipikus egyenlet                                                                          | tipikus határfeltétel |
|------------|-----------------|---------------------------------|-------------------------------------------------------------------------------------------|-----------------------|
| $b^2 < ac$ | elliptikus      | Laplace/(Poisson)-egyenlet      | $\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0$               | peremérték            |
| $b^2 > ac$ | hiperbolikus    | Hullámgörbe                     | $\frac{\partial^2 V}{\partial x^2} - \frac{1}{v^2} \frac{\partial^2 V}{\partial t^2} = 0$ | kezdetiérték          |
| $b^2 = ac$ | parabolikus     | Diffúzió/(Schrödinger)-egyenlet | $\lambda \frac{\partial^2 V}{\partial x^2} - \frac{\partial V}{\partial t} = 0$           | kezdetiérték          |

A csoportosítás nem konstans együtthatók mellett is tartható, de akkor csak lokálisan értendő. A fizikában gyakori egyenletcsoportok jól osztályozhatók a táblázat alapján, ezek típusfeladatok a megoldási módszer szempontjából is nagyon elkülönülnek.

- A hiperbolikus és a parabolikus egyenleteink egyik változója általában az idő. Többnyire kezdetiérték-problémával állunk szemben: adott valamely  $t = t_0$ -kor a  $V(x, t_0)$  kezdemegoldás és ki kell számolnunk, hogy egy későbbi időpontban mi lesz  $V(x, t)$ . Lényegében ez nem különbözik a már megismert integrálási eljárásoktól, az időbeli lépéseket diszkretizálva integrálhatunk előre.
- Az elliptikus egyenletek többnyire csak térszerű változókat tartalmaznak. Valamely zárt határfelületen adott peremfeltételek mellett kell megoldanunk az egyenletet az ismeretlen  $V(x, y)$  előállítására végett. A peremérték-feladatoknál nem lehet egyszerűen 'beintegrálni' a peremről, a megoldást a tér minden pontját egyformán figyelve globális módszerrel kell megközelíteni.

## 4.2. Hiperbolikus egyenletek

### 4.2.1. A bébi hullámeqyenlet

A (homogén) hullámeqyenlet

$$\left[ \frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} \right] u(x, t) = 0 \quad (4.3)$$

(konstans sebesség mellett) átírható a

$$\left[ \frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right] \left[ \frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right] u(x, t) = 0 \quad (4.4)$$

alakra, ahonnan látható, hogy a feladat a

$$\left[ \frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right] z(x, t) = 0 \quad , \quad \left[ \frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right] u(x, t) = z(x, t) \quad (4.5)$$

elsőrendű egyenletek rendszerére redukálható. Az általános hullámeqyenlet (inhomogén, nem-konstans együtthatók) tárgyalása helyett a numerikus megoldási módszerek alapötletét és a felmerülő problémákat egy egyszerűbb modellen vizsgáljuk meg. A következőkben a "bébi hullámeqyenletet"

$$\left[ \frac{\partial}{\partial t} + v \frac{\partial}{\partial x} \right] z(x, t) = 0 \quad , \quad z(x, t_0) = z_0(x) \quad (4.6)$$

mint (időben) kezdetiérték-problémát fogjuk tanulmányozni. Ha a  $v = c$  sebesség valóban állandó, akkor ennek a problémának ismerjük a megoldását, hiszen közvetett differenciálással látható, hogy az differenciálegyenletet minden olyan függvény kielégíti, ami az  $x$  és a  $t$  változóktól csak az  $x - ct$  kombináción keresztül függ. A kezdetiértéket is kielégítő megoldás tehát

$$z(x, t) = z_0(x - ct) \quad (4.7)$$

### 4.2.2. Numerikus diszperzió és disszipáció

Valamelyik (célszerűen most az  $x$ ) változóban átírjuk véges differenciákra a differenciálegyenletet. Ekvidisztáns beosztást véve

$$x_j = x_0 + j\delta x \quad , \quad z_j(t) = z(x_j, t) \quad , \quad j = 0, \dots, M \quad (4.8)$$

a térbeli deriválást különböző módon átírhatjuk véges differenciákra. A legegyszerűbben például az alábbi módokon:

$$\begin{aligned} 1) \quad \dot{z}_j(t) &= -\frac{v}{\delta x} [z_{j+1}(t) - z_j(t)] && 2 \text{ pontos előre} \\ 2) \quad \dot{z}_j(t) &= -\frac{v}{\delta x} [z_j(t) - z_{j-1}(t)] && 2 \text{ pontos hátra} \\ 3) \quad \dot{z}_j(t) &= -\frac{v}{2\delta x} [z_{j+1}(t) - z_{j-1}(t)] && 3 \text{ pontos szimmetrikus} \end{aligned} \quad (4.9)$$

Mindhárom előbb felírt egyenletrendszer felfogható úgy, mint egy közöséges differenciálegyenlet-rendszer

$$\frac{dz_j}{dt} = f_j(v \mid z_0, z_1, \dots, z_M) \quad (4.10)$$

a  $\{z_0(t), z_1(t), \dots, z_M(t)\}$  függvényekre. Az ismeretlen függvények nagy számától eltekintve elvileg ez az egyenletrendszer az ODE integráló módszerekkel már megoldható (sőt, egyszerű esetekben akár analitikusan is). Érdekes azonban előbb megvizsgálni, hogy milyen mesterséges jelenségekre számíthatunk pusztán amiatt, hogy az egyik deriváltat lecseréltük véges differenciákra.

**Diszperziós hiba:** A

$$\frac{\partial}{\partial t} z(x, t) = -c \frac{\partial}{\partial x} z(x, t) \quad (4.11)$$

differenciálegyenlet egy alapmegoldása a

$$z(x, t | k) = e^{-i k c t} e^{i k x} \quad (4.12)$$

síkhullám. Általánosabb egyenletek egy általánosabb megoldása

$$z(x, t | k) = e^{i \omega(k) t} e^{i k x} \quad (4.13)$$

ahol a hullám

$$\frac{d\omega(k)}{dk} \quad (4.14)$$

(csoport)sebssége függhet a  $k$  hullámszámtól. Ha

$$\frac{d\omega(k)}{dk} = c \text{ állandó} \quad (4.15)$$

akkor nincs diszperzió, ha a csoportsebesség nem állandó, akkor pedig diszperzióról beszélünk. Nem diszperzív az előbbi (4.11) egyenlettel leírt rendszer, míg például a

$$\frac{\partial}{\partial t} z(x, t) = -c \frac{\partial}{\partial x} z(x, t) - \lambda \frac{\partial^3}{\partial x^3} z(x, t) \quad (4.16)$$

diszperzív. Általában diszperzióval akkor találkozunk, ha a térváltozóban magasabb páratlan rendű parciális deriváltak is megjelennek.

Tekintsük a nem diszperzív (4.11) egyenletet a

$$z(x, 0) = e^{i k x} \quad (4.17)$$

kezdőfeltétellel. A véges differenciákkal felírt egyenlet megoldását keressük

$$z(x, t) = \Psi(t) e^{i k x} \quad (4.18)$$

alakban. Ekkor a beosztáson

$$\dot{z}_j = \dot{\Psi}(t) e^{i k x_j} \text{ és } z_{j\pm 1} = e^{\pm i k \delta x} z_j \quad (4.19)$$

és a szimmetrikus deriváltat tartalmazó 3) formulába beírva

$$\dot{\Psi}(t) e^{i k x_j} = -\frac{c}{2\delta x} [e^{+i k \delta x} - e^{-i k \delta x}] z_j \Rightarrow \dot{\Psi}(t) = -\frac{c}{\delta x} \sin(k \delta x) \Psi(t) \quad (4.20)$$

azaz

$$\Psi(t) \sim e^{i \omega(k) t} \text{ ahol } \omega(k) = -\frac{c}{\delta x} \sin(k \delta x) \quad (4.21)$$

Nyilvánvaló, hogy

$$\frac{d\omega(k)}{dk} = -c k \cos(k \delta x) \neq \text{állandó} \quad (4.22)$$

tehát a megoldásunk diszperzív lett. Persze ezen nem kell csodálkozni, hiszen a szimmetrikus derivált felírásakor az elkövetett hiba

$$\frac{z_{j+1}(t) - z_{j-1}(t)}{2\delta x} = \left( \frac{\partial}{\partial x} z \right)_j + \frac{(\delta x)^2}{6} \left( \frac{\partial^3}{\partial x^3} z \right)_j + O((\delta x)^4) \quad (4.23)$$

és így azt is joggal mondhatnánk, hogy a

$$\frac{\partial}{\partial t} z(x, t) = -c \left( \frac{\partial}{\partial x} z(x, t) + \frac{(\delta x)^2}{6} \frac{\partial^3}{\partial x^3} z(x, t) \right) \quad (4.24)$$

diszperzív egyenletet oldottuk meg.

Ez előbbi számolásból azt is látjuk, hogy a fázishiba

$$|(\omega(k) + kc)t| = |ckt| \left| \frac{\sin(k \delta x)}{k \delta x} - 1 \right| \approx |ckt| \frac{(k \delta x)^2}{6} \quad (4.25)$$

Ha a megoldás  $|t| < T$  időintervallumban érdekel bennünket és a tipikus térbeli

$$\lambda \sim \frac{2\pi}{k} \quad (4.26)$$

hullámhossz ismert, akkor ezt a formulát használhatjuk a megfelelő beosztás kiválasztásához.

**Disszipációs hiba:** Ha a kétpontos formulákat használjuk, akkor a véges differenciával felírt egyenletünk joggal azonosítható a

$$\frac{\partial}{\partial t} z(x, t) = -c \left( \frac{\partial}{\partial x} z(x, t) \pm \frac{(\delta x)}{2} \frac{\partial^2}{\partial x^2} z(x, t) \right) \quad (4.27)$$

a differenciálegyenlettel. A páros rendű magasabb deriváltak megjelenése arra utal, hogy a rendszer disszipatív. A véges differenciákkal felírt egyenlet megoldását most is

$$z(x, t) = \rho(t) e^{i k x} \quad (4.28)$$

alakban keresve az 1) előre formulával kapjuk, hogy

$$\dot{\rho} = -\frac{c}{\delta x} [e^{+i k \delta x} - 1] \rho \Rightarrow \rho(t) \sim e^{-i \frac{\sin(k \delta x)}{\delta x} ct} e^{-\frac{\cos(k \delta x) - 1}{\delta x} ct} = \Psi(t) \cdot \Phi(t) \quad (4.29)$$

Elegendően finom beosztáson

$$\frac{\cos(k \delta x) - 1}{k \delta x} \approx \frac{1 - \frac{1}{2}(k \delta x)^2 - 1}{k \delta x} = -\frac{k \delta x}{2} \quad (4.30)$$

azaz a megoldásunkban a fázishibás  $\Psi(t)$  mellett megjelenik az időben exponenciálisan növvő

$$\Phi(t) \approx e^{\frac{k \delta x}{2} ct} \quad (4.31)$$

amplitudó faktor. Ez azt eredményezi, hogy a kis hullámhosszú (nagy  $k$ ) megoldások "elszállnak" (ha  $c > 0$ ). Hasonlóan megnézhetjük a 2) hátra formulánkat is, ekkor az exponenciálisok előjele megfordul, azaz

$$\Phi(t) \approx e^{-\frac{k \delta x}{2} ct} \quad (4.32)$$

ami az egyes hullámok "kihalásához" vezet.

#### 4.2.3. Teljesen diszkrétizált explicit módszerek

A véges differenciákra való teljes áttérés végett mindkét ( $x$  és  $t$ ) tengelyen ekvidisztáns

$$x_j = x_0 + j \delta x \quad , \quad t_n = t_0 + n \delta t \quad , \quad z_j^n \equiv z(x_j, t_n) \quad (4.33)$$

beosztást veszünk. A két tengelyen természetesen a beosztások különbözőek. Miután a  $t_n$ -ről való időben előre történő lépésünkör az összes  $z_1^n, z_2^n, \dots$  rendelkezésre áll, az  $x$  szerinti deriváltat tetszőlegesen sok pontból felírhatjuk. Az idő szerinti deriváltra azonban csak korábbi időpontokról van információ, így explicit rekurzióhoz kézenfekvő egyszerű előre deriváltat használni. A kétpontos időbeli első derivált mellett nincs szükség arra, hogy a hárompontos



szimmetrikus formulánál pontosabban használjunk az  $x$  változóban. Kompromisszumként legyen a diszkrétizált egyenlet tehát

$$\left[ \frac{\partial}{\partial t} + v \frac{\partial}{\partial x} \right] z(x, t) \Big|_{j,n} = \frac{z_j^{n+1} - z_j^n}{\delta t} + O(\delta t) + v_j^n \cdot \frac{z_{j+1}^n - z_{j-1}^n}{2\delta x} + O(\delta x^2) = 0 \quad (4.34)$$

Ebből az FTCS (Forward Time, Centered Space) rekurzió

$$z_j^{n+1} = z_j^n - \frac{1}{2} v_j^n (z_{j+1}^n - z_{j-1}^n) \frac{\delta t}{\delta x} \quad (4.35)$$

Sajnos ez az egyszerű formula nemcsak pontatlan, de használhatatlanul instabil, nagyon 'rövid' idő alatt a megoldás 'elszáll'. A *von Neumann* stabilitásvizsgálathoz tekintsük a  $v \equiv c$  speciális eset

$$z(x, t) = e^{ik(x-ct)} \quad (4.36)$$

síkhullám jellegű alapmegoldásait. Ekkor

$$z_j^n = w(k)^n \cdot e^{ikx_j} \quad (4.37)$$

ahol a  $w(k)$  egységnyi modulusú komplex szám  $n$ -edik hatványa szerepel. A rekurziós egyenletbe beírva

$$w(k)^{n+1} = \left[ 1 - i \frac{c\delta t}{\delta x} \sin(k\delta x) \right] w(k)^n \implies w(k) = \left[ 1 - i \frac{c\delta t}{\delta x} \sin(k\delta x) \right] \implies \|w(k)\| > 1 \quad (4.38)$$

azt látjuk, hogy minden  $k$ -val vett alapmegoldásra a formulánk feltételenül instabil. Az iteráció a parciális hullámok amplitudóját különböző mértékben felnöveli az idő haladtával. Megfontolásunkat a homogén hullámegyenletek közül is csak a konstans sebességgel felírtakra végeztük el. Mégis, lassan változó  $v(x, t)$  sebesség mellett és akár inhomogén egyenletek esetén is elfogadhatjuk, hogy ez az eljárás nem stabil.

Viszonylag könnyen segíthetünk az imént feltárt instabilitáson a **Lax-Friedrichs-módszer** segítségével. A rekurzióinkban a jobboldalon végezzük el a

$$z_j^n \longrightarrow \frac{z_{j+1}^n + z_{j-1}^n}{2} \quad (4.39)$$

helyettesítést, azaz  $z_j^n$ -t két térbeli szomszédja átlagára cseréljük le. Ezzel kapjuk a Lax-formulát

$$z_j^{n+1} = \frac{z_{j+1}^n + z_{j-1}^n}{2} - \frac{v}{2} \frac{\delta t}{\delta x} (z_{j+1}^n - z_{j-1}^n) \quad (4.40)$$

A stabilitás vizsgálatát ismét a *von Neumann* módszerrel vizsgálva most azt kapjuk, hogy

$$w(k) = \left[ \cos(k\delta x) - i \frac{c\delta t}{\delta x} \sin(k\delta x) \right] \quad (4.41)$$

$$\|w(k)\|^2 = \cos^2(k\delta x) + \left( c \frac{\delta t}{\delta x} \right)^2 \sin^2(k\delta x) = 1 - \left[ 1 - \left( c \frac{\delta t}{\delta x} \right)^2 \right] \sin^2(k\delta x) \quad (4.42)$$

Minden  $k$ -ra stabil az eljárásunk, ha

$$|v| \frac{\delta t}{\delta x} \leq 1 \quad \text{vagy} \quad |v| \delta t \leq \delta x \quad (4.43)$$

teljesül. Ez a *Courant-Friedrichs-Lewy* stabilitási feltétel. Fizikai magyarázat: a hullámnak  $(x_{j\pm 1}, t_n)$ -ből oda kell érnie  $x_j, t_{n+1}$ -be. Matematikai magyarázat: A Lax-formulát kicsit átírva

$$\frac{z_j^{n+1} - z_j^n}{\delta t} = \frac{(\delta x)^2}{2\delta t} \cdot \frac{z_{j+1}^n - 2z_j^n + z_{j-1}^n}{(\delta x)^2} - v \left( \frac{z_{j+1}^n - z_{j-1}^n}{2\delta x} \right) \quad (4.44)$$

észrevehetjük, hogy ez a

$$\left[ \frac{\partial}{\partial t} + v \frac{\partial}{\partial x} \right] z(x, t) = \lambda \frac{\partial^2}{\partial x^2} z(x, t) \quad , \quad \lambda = \frac{(\delta x)^2}{2\delta t} \quad (4.45)$$

differentiálegyenlet diszkretizációjából is jöhetett volna. A jobboldali diffúziós tagot jelentő második derivált például sűrűlő folyadékok áramlásánál jelenik meg a fizikai egyenletekben, lényegében disszipációs tagot jelent. A stabilitás feltétele

$$\lambda = \frac{\delta x}{\delta t} \frac{\delta x}{2} \geq \frac{\delta x}{2|v|} \quad (4.46)$$

tehát azt jelenti, hogy elegendően erős *numerikus disszipációt*, *numerikus viszkozitást* imitáltunk. A sűrűlődsra szükségünk volt, hogy stabil legyen az eljárás, ugyanakkor a nagy sűrűlőds 'elemésztí' a megoldásainkat. Ha olyan problémákat vizsgálunk, ahol a lényegesnek vélt komponensekre teljesül, hogy  $k\delta x \ll 1$ , akkor ezekre a komponensekre mindkét eljárás használható:  $\|w(k)\|^2 \approx 1$ . A rövid hullámhosszú, azaz  $k\delta x \sim 1$  komponensek azonban az első eljárásban felnővekszenek és uralkodni kezdenek, míg az utóbbiban kihálnak.

Amiatt, hogy az időbeli deriváltat kétpontos formulával írtuk fel, míg a térbelit hárompontosal, a pontosság miatt a  $\delta t$  beosztás finomságára jobban kell ügyelnünk, mint  $\delta x$ -re. Ez alkalmasint azzal járhat, hogy  $|v|\delta t \ll \delta x$  lesz, ami a Courant-feltételnél indokolatlanul erősebb megszorítás.

A **leapfrog** (bakugrásos) módszerrel segíthetünk ezen, olyan egyenletet írunk fel, hogy az mindkét változóban másodrendű hibájú legyen:

$$\left[ \left[ \frac{\partial}{\partial t} + v \frac{\partial}{\partial x} \right] z(x, t) \right]_{j,n} = \frac{z_j^{n+1} - z_j^{n-1}}{2\delta t} + O(\delta t^2) + v_j^n \cdot \frac{z_{j+1}^n - z_{j-1}^n}{2\delta x} + O(\delta x^2) = 0 \quad (4.47)$$

$$z_j^{n+1} = z_j^{n-1} - \frac{\delta t}{\delta x} v_j^n (z_{j+1}^n - z_{j-1}^n) \quad (4.48)$$

A formula alkalmazásával  $t_{n-1}$ -ről közvetlenül  $t_{n+1}$ -re ugrunk, miközben felhasználjuk a kihagyott  $t_n$  időbeli értékeket. Ez azzal jár együtt, hogy az időben páros és a páratlan osztópontokon külön rekurzióval kell haladnunk. Ehhez persze nem elegendő a szokásos kezdőfeltétel. Valamilyen egyéb módszerrel (nagyon óvatos kis időlépéseket választva, mert az itt begyűjtött hiba nem hal ki a disszipáció hiánya miatt) előbb ki kell számolnunk a  $t_0$  kezdőidőtől a  $t_1$  időpontig a megoldást. Az sem mellékes, hogy mindeközben a korábbi időpillantra kiszámolt  $z^{n-1}$  értékeket meg kellett őrizni. A stabilitásvizsgálatot a korábbiakhoz hasonlóan elvégezve kapjuk, hogy

$$w^2 = 1 - w2i \frac{c\delta t}{\delta x} \sin(k\delta x) \quad (4.49)$$

Kis kézimunkával megmutathatjuk, hogy

$$\frac{c\delta t}{\delta x} \leq 1 \implies \|w(k)\| = 1 \quad (4.50)$$

tehát a Courant-feltétel teljesülése szükséges, de ilyenkor sem erősítés, sem disszipáció nincs (diszperzió azért természetesen van). Az egyenlet annál is inkább érdekes, mert az eredeti másodrendű hullámegyenletünkre visszavezetve annak közvetlen diszkretizálásával ekvivalens

$$\left[ \frac{\partial^2 u}{\partial t^2} - v^2 \frac{\partial^2 u}{\partial x^2} \right]_{j,n} = \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\delta t)^2} - v^2 \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\delta x)^2} = 0 \quad (4.51)$$

A bakugrásos módszerrel bakot lőhetünk bonyolultabb egyenletekre. A hiba egyik forrása éppen a bakugrás, hiszen a páros és a páratlan osztópontok (akár a sakktábla világos és sötét mezői) teljesen szétcsatoltak az egyenletben.

Lehetséges megoldás a **kétlépéses Lax-Wendroff-módszer**. Az első lépésben Lax-módszerrel (képzeletben) csak fele akkor beosztású rácson lépünk a centrális pontba

$$z_{j+1/2}^{n+1/2} = \frac{z_{j+1}^n + z_j^n}{2} - \frac{v}{2} \frac{\delta t}{\delta x} (z_{j+1}^n - z_j^n) \quad (4.52)$$

majd a centrális értékkel bakugrás következik

$$z_j^{n+1} = z_j^n - \frac{\delta t}{\delta x} v_{j+1/2}^{n+1/2} (z_{j+1/2}^{n+1/2} - z_{j-1/2}^{n+1/2}) \quad (4.53)$$

A két egymásutáni lépést egyetlen lépéssel is helyettesíthetjük esetleges  $v$  változásától eltekintve. Ekkor

$$z_j^{n+1} = z_j^n - \frac{\delta t}{\delta x} v (z_{j+1/2}^{n+1/2} - z_{j-1/2}^{n+1/2}) = z_j^n - v \frac{\delta t}{\delta x} \left( \frac{z_{j+1}^n + z_{j-1}^n}{2} \right) + \left( v \frac{\delta t}{\delta x} \right)^2 \left( \frac{z_{j+1}^n - 2z_j^n + z_{j-1}^n}{2} \right) \quad (4.54)$$

lesz. Ezt az egyenletet másképp is megkaphatjuk, hiszen az időbeli differenciáláshoz

$$z(x, t + \delta t) = z(x, t) + \frac{\partial}{\partial t} z(x, t) \cdot \delta t + \frac{\partial^2}{\partial t^2} z(x, t) \cdot \frac{(\delta t)^2}{2} + O((\delta t)^3) \quad (4.55)$$

Figyelembe véve a differenciálegyenletet, hogy

$$\frac{\partial}{\partial t} z(x, t) = -v \frac{\partial}{\partial x} z(x, t) \quad \text{és következésképp} \quad \frac{\partial^2}{\partial t^2} z(x, t) = v^2 \frac{\partial^2}{\partial x^2} z(x, t) \quad (4.56)$$

írhatjuk, hogy

$$z(x, t + \delta t) = z(x, t) - v \delta t \cdot \frac{\partial}{\partial x} z(x, t) + v^2 \frac{(\delta t)^2}{2} \cdot \frac{\partial^2}{\partial x^2} z(x, t) + O((\delta t)^3) \quad (4.57)$$

Értelmesen átírva véges differenciákra

$$z_j^{n+1} = z_j^n - v \delta t \cdot \left[ \frac{z_{j+1}^n + z_{j-1}^n}{2\delta x} \right] + v^2 \frac{(\delta t)^2}{2} \cdot \left[ \frac{z_{j+1}^n - 2z_j^n + z_{j-1}^n}{\delta x \cdot \delta x} \right] \quad (4.58)$$

ami éppen az előbbi egyenlet.

Az eljárás stabilitásának feltétele megintcsak a  $|v| \delta t \leq \delta x$  Courant-szabály. Csillapítás mutatható ki minden  $k \neq 0$  hullámszám mellett ez a disszipáció azonban kisebb, mint az eredeti Lax-módszernél.

#### 4.3. Parabolikus egyenletek

A egydimenziós

$$\frac{\partial V}{\partial t} = \lambda \frac{\partial^2 V}{\partial x^2} \quad (4.59)$$

diffúziós egyenlet numerikus megoldását vizsgáljuk. Valódi diffúziós problémákban  $V(x, t)$  a koncentráció és  $\lambda > 0$  a **diffúziós állandó**. Ugyanilyen differenciálegyenlet írja le a hővezetést, ekkor  $V$  a hőmérséklet és  $\lambda = \frac{\sigma}{\rho c}$  szintén pozitív. Itt  $\sigma$  a fajlagos **hővezetési képesség**,  $\rho$  az anyag sűrűsége és  $c$  az állandó nyomáson vett fajhő. Az időfüggő Schrödinger-egyenlet is ebbe a családba tartozik, ott azonban  $\lambda$  imaginárius. Ezzel később külön példában foglalkozunk, most feltesszük, hogy  $\lambda > 0$ .

Véges differenciákra teljesen átírva a feladatot  $V_j^n \equiv V(x_j, t_n)$  jelöléssel az **FTCS** egyenletünk

$$\frac{V_j^{n+1} - V_j^n}{\delta t} = \lambda \frac{V_{j+1}^n - 2V_j^n + V_{j-1}^n}{(\delta x)^2} \quad (4.60)$$

$$V_j^{n+1} = V_j^n + \alpha [V_{j+1}^n - 2V_j^n + V_{j-1}^n] \quad , \quad \alpha = \frac{\lambda \delta t}{(\delta x)^2} \quad (4.61)$$

A stabilitásvizsgálathoz vegyük a diffúziós egyenletnek egy alapmegoldását

$$V(x, t) = w(t) \cdot e^{ikx} \quad , \quad w(t) = e^{-\lambda k^2 t} \quad (4.62)$$

Vizsgáljuk meg, hogy a

$$V_j^n = w^n \cdot e^{ikx_j} \quad (4.63)$$

helyettesítéssel  $w(t)$ -re a differencia-egyenlet milyen viselkedést ad. Azt kapjuk, hogy

$$w = 1 + \alpha [e^{ik\delta x} - 2 + e^{-ik\delta x}] = 1 + 2\alpha [\cos(k\delta x) - 1] = 1 - 4\alpha \sin^2\left(\frac{k\delta x}{2}\right) \quad (4.64)$$

amiből látjuk, hogy

$$|w| \leq 1 \quad \text{minden } k\text{-ra, ha} \quad 4\alpha \sin^2\left(\frac{k\delta x}{2}\right) \leq 2 \quad (4.65)$$

A stabilitás feltétele ezek szerint, hogy

$$\alpha \leq 1/2 \quad \text{azaz} \quad \delta t \leq \frac{1}{2} \frac{(\delta x)^2}{\lambda} \sim \tau \quad (4.66)$$

ahol  $\tau$  a  $\delta x$  távolságra való diffúzió karakterisztikus ideje. A feltételes stabilitás ugyan biztató, mégis a módszer alkalmazása a gyakorlatban nem célszerű. A reális problémákra jellemző  $L$  tipikus diffúziós hosszúság mellett a diffúziós karakterisztikus idő

$$T \sim \frac{L^2}{\lambda} = \frac{L^2}{(\delta x)^2} \frac{(\delta x)^2}{\lambda} \sim \frac{L^2}{(\delta x)^2} \tau \quad (4.67)$$

Miközben értelemszerűen a pontosabb számoláshoz  $L/\delta x \gg 1$  térbeli osztópont szükséges, a  $T$  ideig való szimulációhoz  $L^2/(\delta x)^2$  nagyságrendben kell időbeli lépéseket tenni. A stabilitás követelte nagyon nagy számú lépés nemcsak számolási igénye miatt kényelmetlen, de a felhalmozódó numerikus hibák miatt is veszélyes.

Lépéshossztól függetlenül stabil egyenletekhez jutunk, ha a bal oldalon a térváltozó szerinti második deriváltat nem az iménti módon a  $t_n$ -ben, hanem a  $t_{n+1}$ -ben írjuk fel:

$$\frac{V_j^{n+1} - V_j^n}{\delta t} = \lambda \frac{V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}}{(\delta x)^2} \quad (4.68)$$

Így egy **teljesen implicit** egyenletrendszer kapunk:

$$-\alpha V_{j+1}^{n+1} + (1 + 2\alpha)V_j^{n+1} - \alpha V_{j-1}^{n+1} = V_j^n \quad , \quad j = 1, 2, \dots, J-1 \quad (4.69)$$

A stabilitás vizsgálatánál kapjuk, hogy

$$w [1 + 2\alpha - 2\alpha \cos(k\delta x)] = 1 \quad \Rightarrow \quad w = \frac{1}{1 + 4\alpha \sin^2(\frac{k\delta x}{2})} \quad (4.70)$$

azaz az eljárás minden térbeli  $\delta x$  és időbeli  $\delta t$  lépésköz választással és minden  $k$ -ra stabil, azaz feltétel nélkül teljesül, hogy  $|w| \leq 1$ . Az implicit egyenletrendszer vektor jelölésben

$$\mathbf{A} \cdot \mathbf{V}^{n+1} = \mathbf{V}^n \quad (4.71)$$

ahol  $\mathbf{A}$  tridiagonális mátrix:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ -\alpha & 1+2\alpha & -\alpha & 0 & \cdot & 0 \\ 0 & -\alpha & 1+2\alpha & -\alpha & 0 & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -\alpha & 1+2\alpha & -\alpha \\ \cdot & \cdot & \cdot & 0 & 0 & 1 \end{pmatrix} \quad (4.72)$$

Az ilyen egyenletet például rekurzióval minden időbeli lépésre meg tudjuk oldani. Lásd a megfelelő alfejezetet!

Az eddigi két eljárás elsőrendben volt pontos. Figyelembe véve, hogy a kétpontos első derivált másodrendben pontos szimmetrikus formulaként is felfogható a köztes  $t_{n+1/2}$  pillanatra, *másodrendben pontos* egyenletet kapunk, ha a jobboldalt is így fogjuk fel

$$\left. \frac{\partial V}{\partial t} \right|_{j,n+1/2} = \lambda \left. \frac{\partial^2 V}{\partial x^2} \right|_{j,n+1/2} \approx \frac{\lambda}{2} \left[ \left. \frac{\partial^2 V}{\partial x^2} \right|_{j,n+1} + \left. \frac{\partial^2 V}{\partial x^2} \right|_{j,n} \right] \quad (4.73)$$

A deriváltakat diszkretizálva kapjuk

$$\frac{V_j^{n+1} - V_j^n}{\delta t} = \frac{\lambda}{2(\delta x)^2} [(V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}) + (V_{j+1}^n - 2V_j^n + V_{j-1}^n)] \quad (4.74)$$

ami átrendezve a **Crank-Nicholson**-formula

$$V_j^{n+1} - \frac{\alpha}{2} (V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}) = V_j^n + \frac{\alpha}{2} (V_{j+1}^n - 2V_j^n + V_{j-1}^n) \quad (4.75)$$

$$-\alpha V_{j+1}^{n+1} + (2 + 2\alpha)V_j^{n+1} - \alpha V_{j-1}^{n+1} = \alpha V_{j+1}^n + (2 - 2\alpha)V_j^n + \alpha V_{j-1}^n \quad (4.76)$$

vagy mátrix alakban

$$\mathbf{A} \cdot \mathbf{V}^{n+1} = \mathbf{B} \cdot \mathbf{V}^n \quad (4.77)$$

ahol  $\mathbf{A}$  és  $\mathbf{B}$  tridiagonális mátrix. Az eljárás erősítése

$$w = \frac{1 - \alpha + \alpha \cos(k\delta x)}{1 + \alpha - \alpha \cos(k\delta x)} = \frac{1 - 2\alpha \sin^2(\frac{k\delta x}{2})}{1 + 2\alpha \sin^2(\frac{k\delta x}{2})} \leq 1 \quad (4.78)$$

tehát a módszer feltétel nélkül stabil.

Az implicit módszerekben a tridiagonális mátrixokkal való munka időigényes lehet. Kis módosítással olyan egyenletet is felírhatunk, ahol hasonló pontossággal (és stabilitással) explicit formulát kapunk. Ilyen a **Dufort-Frankel** módszer. Továbbra is másodrendben korrekt idő szerinti deriváltat írunk fel, a jobb oldalon a tér szerinti második deriváltban pedig a

$$2V_j^n \longrightarrow V_j^{n+1} + V_j^{n-1} \quad (4.79)$$

helyettesítéssel élünk. Ekkor a

$$\frac{V_j^{n+1} - V_j^{n-1}}{2\delta t} = \lambda \frac{V_{j+1}^n - V_j^{n+1} - V_j^{n-1} + V_{j-1}^n}{(\delta x)^2} \quad (4.80)$$

$$V_j^{n+1} = V_j^{n-1} + 2\alpha [V_{j+1}^n - V_j^{n+1} - V_j^{n-1} + V_{j-1}^n] \quad (4.81)$$

látszólag implicit egyenletet kapjuk. Az ismeretlen  $V_j^{n+1}$  azonban triviális módon explicite kifejezhető:

$$V_j^{n+1} = \left( \frac{1 - 2\alpha}{1 + 2\alpha} \right) V_j^{n-1} + \left( \frac{2\alpha}{1 + 2\alpha} \right) [V_{j+1}^n + V_{j-1}^n] \quad (4.82)$$

A stabilitásvizsgálakor másodfokú egyenletet kapunk  $w$ -re

$$w^2 = \left( \frac{1 - 2\alpha}{1 + 2\alpha} \right) + 2w \left( \frac{2\alpha}{1 + 2\alpha} \right) \cos(k\delta x) \quad (4.83)$$

amiből

$$w = \frac{1}{1 + 2\alpha} \left( 2\alpha \cos(k\delta x) \pm \sqrt{1 - 4\alpha^2 \sin^2(k\delta x)} \right) \quad (4.84)$$

Ha  $4\alpha^2 \sin^2(k\delta x) \leq 1$ , akkor a gyök valós, és

$$-(2\alpha + 1) \leq \cos(k\delta x) \pm \sqrt{1 - 4\alpha^2 \sin^2(k\delta x)} \leq (2\alpha + 1) \quad \text{miatt} \quad |w| \leq 1 \quad (4.85)$$

Komplex gyökre, amikor  $4\alpha^2 \sin^2(k\delta x) > 1$

$$\|w\|^2 = \frac{1}{(1 + 2\alpha)^2} (\cos^2(k\delta x) + 4\alpha^2 \sin^2(k\delta x) - 1) = \frac{4\alpha^2 - 1}{(1 + 2\alpha)^2} = \frac{2\alpha - 1}{2\alpha + 1} \quad \text{miatt} \quad \|w\|^2 \leq 1 \quad (4.86)$$

Tehát az eljárás minden beosztás mellett stabil. A **Dufort-Frankel** módszernek is vannak hátrányai. Egyrészt az eljárás során két megelőző időpontban kell ismernünk a megoldást a teljes térbeli beosztáson, ami a  $V_j^{n-1}$  értékek tárolását kívánja meg. Ez jelentős tárolókapacitást igényelhet. Továbbá kellemetlen, hogy kezdőfeltételként is két időpontban is meg kell adnunk a  $V_j^{n-1}$  értékeket, ami indokolatlan egy időben elsődredű differenciálegyenletnél.

A stabilitás és a pontosság ömagában nem jelent elég körültekintő eljárást. A parciális differenciálegyenletek diszkretizációja teljesen meghamisíthatja az eredeti problémát. Láttuk, a hullámeqyenletnél, hogy numerikus disszipáció és diszperzió jelent meg a megoldásban, mintha egy másik (fizikai) feladatot tűztünk volna ki. Mindaddig, amíg csak a térváltozóban diszkretizáltunk, ez a hamis tag a differenciálegyenletben a térrács beosztásának a finomításával tetszés szerint kicsívé tehető. Ha az egyenleteket teljesen (mindkét változójában) átírjuk véges differenciákra, akkor külön figyelniük kell arra, hogy a beosztások finomításával a hamis tagok valóban eltűnjenek, azaz az eljárás nemcsak stabil és pontos legyen, de legyen **hű** is. A Dufort-Frankel- egyenletet példáján kis átalakítással

$$\frac{V_j^{n+1} - V_j^{n-1}}{2\delta t} = \lambda \frac{V_{j+1}^n - V_j^{n+1} - V_j^{n-1} + V_{j-1}^n}{(\delta x)^2} = \lambda \left[ \frac{V_{j+1}^n - 2V_j^n + V_{j-1}^n}{(\delta x)^2} - \left( \frac{\delta t}{\delta x} \right)^2 \frac{V_j^{n+1} - 2V_j^n + V_j^{n-1}}{(\delta t)^2} \right] \quad (4.87)$$

Mármost megmutattuk, hogy  $\delta t$  és  $\delta x$  bármilyen választása mellett az eljárás stabil és az is nyilvánvaló, hogy  $\delta t \rightarrow 0$  és  $\delta x \rightarrow 0$  finomítással egyre pontosabbak a megfelelő deriváltak. Igen, de most finomítsuk a beosztást úgy, hogy

$$\delta t \rightarrow 0 \quad \text{és} \quad \delta x \rightarrow 0 \quad \text{miközben} \quad \frac{\delta t}{\delta x} = \alpha \quad \text{konstans} \quad (4.88)$$

Nyilvánvaló, hogy a

$$\frac{\partial V}{\partial t} = \lambda \left[ \frac{\partial^2 V}{\partial x^2} - \alpha \frac{\partial^2 V}{\partial t^2} \right] \quad (4.89)$$

**hamis** differenciálegyenletet oldjuk meg ilyenkor. Akkor konvergál a megoldás az eredeti egyenlet igazi megoldásához, ha például

$$\delta t \rightarrow 0 \quad \text{és} \quad \delta x \rightarrow 0 \quad \text{miközben} \quad \frac{\delta t}{(\delta x)^2} = \beta \quad \text{konstans} \quad (4.90)$$

ugyanis ekkor

$$\frac{\partial V}{\partial t} = \lambda \left[ \frac{\partial^2 V}{\partial x^2} - \beta \cdot \delta t \cdot \frac{\partial^2 V}{\partial t^2} \right] \rightarrow \lambda \frac{\partial^2 V}{\partial x^2} \quad (4.91)$$

#### 4.3.1. Példa: az időfüggő Schrödinger-egyenlet

Egy  $\psi(x, t)$  'hullámcsomag' szóródását a  $U(x)$  potenciálon a

$$i \frac{\partial \psi}{\partial t} = -\frac{\partial^2 \psi}{\partial x^2} + U(x)\psi \quad , \quad \psi(x, t = t_0) = \psi_0(x) \quad , \quad \psi(\pm\infty, t) = 0 \quad (4.92)$$

kezdetiérték probléma írja le. Az előzőekben tárgyalt egyenletekhez képest ez az egyenlet alapvetően két dologban új: a) a  $\lambda$  együttható imaginárius, b) a potenciállal kapcsolatos tag eddig nem szerepelt. Az egyenletek természetesen általánosíthatók erre az esetre is. Tekintsük a **Crank-Nicholson** eljárást, miszerint

$$\frac{\psi_j^{n+1} - \psi_j^n}{\delta t} = \frac{i}{2} \left[ \frac{\psi_{j+1}^{n+1} - 2\psi_j^{n+1} + \psi_{j-1}^{n+1}}{(\delta x)^2} - U_j(x)\psi_j^{n+1} + \frac{\psi_{j+1}^n - 2\psi_j^n + \psi_{j-1}^n}{(\delta x)^2} - U_j(x)\psi_j^n \right] \quad (4.93)$$

vagy rendezve

$$-a\psi_{j+1}^{n+1} + (1 + 2a + b_j)\psi_j^{n+1} - a\psi_{j-1}^{n+1} = a\psi_{j+1}^n + (1 - 2a - b_j)\psi_j^n + a\psi_{j-1}^n ; \quad a = \frac{i\delta t}{2(\delta x)^2} ; \quad b_j = \frac{i\delta t}{2}U(x_j) \quad (4.94)$$

A tridiagonális egyenletrendszer numerikusan hatékonyan invertálható.

Tanulságos az egyenletet egy másik származtatását is megvizsgálni. Az időfüggő Schrödinger egyenlet formális megoldásával

$$e^{-iH\delta t}\psi(x, t_n) = \psi(x, t_{n+1}) \quad (4.95)$$

Ha ebben az egyenletben az

$$e^{-iH\delta t} \approx 1 - iH\delta t \quad \implies \quad \psi_j^{n+1} = (1 - iH\delta t)\psi_j^n \quad (4.96)$$

sorfejtést helyettesítjük, majd a térváltozóban második deriváltat centrált véges differenciával írjuk fel, akkor az FTCS egyenletet kapjuk. Imaginárius  $\lambda$ -val az FTCS azonban instabil. A teljesen implicit egyenletet akkor kapjuk, ha az alternatív

$$(1 - iH\delta t)^{-1}\psi_j^{n+1} = \psi_j^n \quad (4.97)$$

felírást választjuk. Ez az eljárás stabil imaginárius  $\lambda$ -val is, de nem unitér (ahogy az FTCS sem). Ez azzal jár együtt, hogy az eredetileg normált hullámfüggvény normája elomlik a szimuláció során. Ha a Cayley-sorfejtést alkalmazzuk:

$$e^{-iH\delta t} \approx \frac{1 - \frac{1}{2}iH\delta t}{1 + \frac{1}{2}iH\delta t} \quad \implies \quad \left(1 + \frac{1}{2}iH\delta t\right)\psi_j^{n+1} = \left(1 - \frac{1}{2}iH\delta t\right)\psi_j^n \quad (4.98)$$

akkor másodrendben pontos, stabil és unitér eljárást kapunk. A véges differenciákat behelyettesítve ez az egyenlet éppen a **Crank-Nicholson**-formulára vezet.

4.3.2. Tridiagonális egyenletrendszer megoldása előre-hátra helyettesítéssel

Tekintsük az

$$\mathbf{A} \vec{x} = \vec{d} \quad (4.99)$$

egyenletrendszert, ahol  $\mathbf{A}$  egy tridiagonális mátrix,  $\vec{d}$  adott vektor és  $\vec{x}$  az ismeretlen. Komponensekben kiírva:

$$\begin{pmatrix} \beta_1 & \gamma_1 & 0 & \cdot & \cdot & \cdot \\ \alpha_2 & \beta_2 & \gamma_2 & 0 & \cdot & \cdot \\ 0 & \alpha_3 & \beta_3 & \gamma_3 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \alpha_{n-1} & \beta_{n-1} & \gamma_{n-1} \\ \cdot & \cdot & \cdot & 0 & \alpha_n & \beta_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \cdot \\ d_{n-1} \\ d_n \end{pmatrix} \quad (4.100)$$

vagy

$$\alpha_i x_{i-1} + \beta_i x_i + \gamma_i x_{i+1} = d_i, \quad i = 1, \dots, n \quad ; \quad \alpha_1 = \gamma_n = 0 \quad (4.101)$$

Vezessük be a  $g_i$  és  $h_i$  segédváltozókat úgy, hogy

$$x_{i+1} = g_i x_i + h_i, \quad i = 1, \dots, n-1 \quad (4.102)$$

legyen. Ezekkel

$$d_i = \alpha_i x_{i-1} + \beta_i x_i + \gamma_i x_{i+1} \quad (4.103)$$

$$= \alpha_i x_{i-1} + \beta_i x_i + \gamma_i (g_i x_i + h_i) \quad (4.104)$$

$$= \alpha_i x_{i-1} + (\beta_i + \gamma_i g_i) x_i + \gamma_i h_i \quad (4.105)$$

$$= \alpha_i x_{i-1} + (\beta_i + \gamma_i g_i) (g_{i-1} x_{i-1} + h_{i-1}) + \gamma_i h_i \quad (4.106)$$

$$= (\alpha_i + g_{i-1} (\beta_i + \gamma_i g_i)) x_{i-1} + (\beta_i + \gamma_i g_i) h_{i-1} + \gamma_i h_i \quad (4.107)$$

ami kielégíthető, ha

$$0 = \alpha_i + g_{i-1} (\beta_i + \gamma_i g_i) \implies g_{i-1} = -\frac{\alpha_i}{\beta_i + \gamma_i g_i} \quad (4.108)$$

$$d_i = (\beta_i + \gamma_i g_i) h_{i-1} + \gamma_i h_i \implies h_{i-1} = \frac{d_i - \gamma_i h_i}{\beta_i + \gamma_i g_i} \quad (4.109)$$

Láthatóan  $g_{n-1}$  és  $h_{n-1}$  ismeretében 'hátrafelé' haladva

$$(n-1) \rightarrow (n-2) \rightarrow \dots \rightarrow 2 \rightarrow 1 \quad (4.110)$$

előállítható az összes  $g$  és  $h$ . Az induló értékek  $\gamma_n = 0$  miatt:

$$g_{n-1} = -\frac{\alpha_n}{\beta_n} \quad \text{és} \quad h_{n-1} = \frac{d_n}{\beta_n} \quad (4.111)$$

Az eredeti egyenletrendszer első egyenletéből

$$\beta_1 x_1 + \gamma_1 x_2 = d_1 \implies \beta_1 x_1 + \gamma_1 (g_1 x_1 + h_1) = d_1 \quad (4.112)$$

azaz  $g_1$  és  $h_1$  segítségével  $x_1$  kifejezhető:

$$x_1 = \frac{d_1 - \gamma_1 h_1}{\beta_1 + \gamma_1 g_1} \quad (4.113)$$

Erről az értékről most 'előrefelé' haladva

$$1 \rightarrow 2 \rightarrow \dots \rightarrow (n-1) \rightarrow n \quad (4.114)$$

$$x_{i+1} = g_i x_i + h_i \quad (4.115)$$

alapján minden keresett  $x_i$  kiszámolható.

Az eljárás értelemszerűen fordítva is felírható, lehet előre haladni  $g$  és  $h$  meghatározásakor, és hátra  $x$ -nél.

### 4.3.3. A diffúziós probléma több dimenzióban

A diffúziós egyenlet 2 dimenzióban

$$\frac{\partial V}{\partial t} = \lambda \left( \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \right) \quad (4.116)$$

Crank-Nicholson szerint diszkretizálva a  $V_{j,l}^n \equiv V(x_j, y_l, t_n)$  jelöléssel

$$V_{j,l}^{n+1} - V_{j,l}^n = \frac{\alpha}{2} \left( \mathcal{L}_x V_{j,l}^{n+1} + \mathcal{L}_x V_{j,l}^n + \mathcal{L}_y V_{j,l}^{n+1} + \mathcal{L}_y V_{j,l}^n \right) \quad (4.117)$$

vagy

$$(2 - \alpha [\mathcal{L}_x + \mathcal{L}_y]) V_{j,l}^{n+1} = (2 + \alpha [\mathcal{L}_x + \mathcal{L}_y]) V_{j,l}^n \quad (4.118)$$

ahol

$$\delta x = \delta y = \Delta, \quad \alpha = \frac{\lambda \delta t}{\Delta^2} \quad (4.119)$$

és  $\mathcal{L}$  a Laplace-féle differenciaoperátor

$$\mathcal{L}_x V_{j,l} = [V_{j+1,l} - 2V_{j,l} + V_{j-1,l}] \quad , \quad \mathcal{L}_y V_{j,l} = [V_{j,l+1} - 2V_{j,l} + V_{j,l-1}] \quad (4.120)$$

Az eljárás térben és időben egyaránt másodrendben pontos. Az implicit módszerekre jellemzően az megoldás feltétel nélkül stabil.

Rendezzük a kétindexes mennyiségeket egyetlen vektorba valamely  $j, l \rightarrow s$  megfeleltetéssel. Így az ismeretlen  $\vec{v} \sim \{V_{j,l}^{n+1}\}$  vektorra a megoldandó (4.118) lineáris egyenletrendszer

$$\mathbf{A} \vec{v} = \vec{b} \quad (4.121)$$

alakú. Sajnos, az egydimenziós problémával szemben a  $\mathbf{A}$  most nem tridiagonális, szerencsére azonban **ritka** mátrix. Az ilyen típusú egyenletek numerikusan standard módszerekkel (Jacobi, Gauss-Seidel és SOR) hatékonyan megoldhatók, erről külön kiegészítésben szólnunk.

Kissé átalakítva a CN-egyenletet rokonszenvesebb alternatív eljáráshoz jutunk. A váltakozó irányú implicit: **ADI** (*alternating-direction implicit*) módszerben a teljes időlépést két részre bontjuk és mindkét fél lépésben más térbeli dimenziót kezelünk implicit módszerrel

$$V_{j,l}^{n+1/2} - V_{j,l}^n = \frac{\alpha}{2} \left( \mathcal{L}_x V_{j,l}^{n+1/2} + \mathcal{L}_y V_{j,l}^n \right) \quad (4.122)$$

$$V_{j,l}^{n+1} - V_{j,l}^{n+1/2} = \frac{\alpha}{2} \left( \mathcal{L}_x V_{j,l}^{n+1/2} + \mathcal{L}_y V_{j,l}^{n+1} \right) \quad (4.123)$$

Az eljárás pontossága és stabilitása az előző direkt formulával azonos, de az egyenletek megoldása könnyebb. Látható ugyanis, hogy 'csak' két tridiagonális egyenletrendszert kell minden lépésben megoldanunk.

Az ADI módszerrel rokon az **OS** '*operator splitting*'. Tegyük fel, hogy a kezdetiérték-probléma jobb oldalán álló differenciáloperátor valamely operátorok összegeként írható fel:

$$\frac{\partial V}{\partial t} = \lambda \mathcal{D}V = \lambda (\mathcal{D}_1 V + \dots + \mathcal{D}_m V) \quad (4.124)$$

Tegyük fel továbbá, hogy az itt szereplő mindegyik  $\mathcal{D}_i$  olyan, hogy a

$$\frac{\partial V}{\partial t} = \lambda \mathcal{D}_i V \quad (4.125)$$

egyenlebből viszonylag egyszerű lenne diszkretizált eljárással

$$\vec{v}^{[n+1]} = F_i(\vec{v}^{[n]}, \delta t) \quad (4.126)$$



időbeli lépéseket tenni. Ilyenkor egy teljes időbeli lépést feloszthatunk  $m$  darab kisebb lépésre és a

$$\vec{v}^{[n+\frac{1}{m}]} = F_1(\vec{v}^{[n]}, \frac{\delta t}{m}) \quad (4.127)$$

$$\vec{v}^{[n+\frac{2}{m}]} = F_2(\vec{v}^{[n+\frac{1}{m}]}, \frac{\delta t}{m}) \quad (4.128)$$

$$\vdots \quad (4.129)$$

$$\vec{v}^{[n+1]} = F_m(\vec{v}^{[n+\frac{m-1}{m}]}, \frac{\delta t}{m}) \quad (4.130)$$

módon állítjuk elő az új értékeket.

Az ADI és az OS rokonsága abban áll, hogy az iménti egyenletben az  $F_i$  léptetési utasítást másképp is felfoghatjuk. Például lehet  $F_1$  egy diszkretizált lépés a teljes  $\mathcal{D}$  operátorral, de pl. úgy felírva, hogy az csak  $\mathcal{D}_1$ -ben stabil, míg  $F_2$  a  $\mathcal{D}_2$ -ben stabil lépés, és így tovább. Így az előbbi OS egyenlet a korábbi ADI sémát adja.

#### 4.3.4. Kiegészítés: $\mathbf{A}\vec{v}=\vec{b}$ lineáris egyenletrendszer megoldása

A nagy mátrixokkal felírt lineáris egyenletrendszerek megoldására a numerikus módszerek sokaságát ismerjük. A téma rendkívül szerteágazó, most csak ízelítőt adunk belőle.

*Gauss-Jordan-elimináció:* Ezt a módszert alighanem mindannyian ismerjük a lineáris algebra tanulmányainkból. Az eljárás elvileg biztonságosan elvezet a megoldáshoz, nagy mátrixoknál azonban több okból nem szoktuk használni. A rengeteg elemi aritmetikai művelet során a véges számábrázolásból származó hiba oly mértékben fölhalmazódik, hogy a végeredmény megbízhatatlan lesz. A nagyon sok művelet elvégzése nemcsak a pontosság rovására megy, de az elvégzéséhez szükséges számítási idő is meghaladhatja a kapacitásunkat. További hátrány, hogy az eliminációs lépések során teljes méretű hatalmas mátrixokat kell egyidőben a memóriában tartani, ami erősen korlátozza a maximális dimenziókat.

*LU-felbontás:* Ha sikerül hatékony algoritmussal az  $\mathbf{A}$  mátrixot

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \quad (4.131)$$

szorzat alakra felbontani, ahol  $\mathbf{L}$  egy alsó trianguláris (*háromszög*) és  $\mathbf{U}$  egy felső trianguláris mátrix

$$\begin{pmatrix} A_{11} & A_{12} & A_{1n} \\ A_{21} & A_{22} & A_{2n} \\ & \ddots & \\ A_{n1} & A_{n2} & A_{nn} \end{pmatrix} = \begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ & \ddots & \\ L_{n1} & L_{n2} & L_{nn} \end{pmatrix} \cdot \begin{pmatrix} U_{11} & U_{12} & U_{1n} \\ 0 & U_{22} & U_{2n} \\ & \ddots & \\ 0 & 0 & U_{nn} \end{pmatrix} \quad (4.132)$$

akkor a megoldandó egyenletrendszer két részre választható:

$$\mathbf{A}\vec{v} = \vec{b} \quad \Rightarrow \quad \mathbf{L} \vec{w} = \vec{b} \quad \text{és} \quad \mathbf{U} \vec{v} = \vec{w} \quad (4.133)$$

A trianguláris  $\mathbf{L}$  mátrix ismertében közvetlen "előre" behelyettesítéssel  $\vec{w}$  explicite meghatározható, hiszen

$$L_{11}w_1 = b_1 \quad \Rightarrow \quad w_1 \quad (4.134)$$

$$L_{21}w_1 + L_{22}w_2 = b_2 \quad \Rightarrow \quad w_2 \quad (4.135)$$

és így tovább. A másik egyenlet az  $\mathbf{U}$  trianguláris volta miatt  $\vec{w}$  ismeretében "hátra" behelyettesítéssel közvetlenül megoldható a kersett  $\vec{v}$ -re. Az  $LU$ -felbontás egyéb lineáris algebrával kapcsolatos feladat megoldásában is igen hasznos, ilyenek a mátrix determinánsának a kiszámítása vagy a mátrix invertálása. Ennek megfelelően sok kidolgozott kész eljárás található arra, hogy adott  $\mathbf{A}$  mátrixra az  $LU$ -felbontást elvégezzük. A mostani kérdéskör (PDE-BVP) kapcsán ennek mégis inkább csak elvi érdekessége van, figyelembe kell vennünk az  $LU$ -felbontásra szánt számítási munka terhét és pontatlanságát továbbá az előre-hátra helyettesítések jól határozott nagy számát és hibáját.

A PDE problémákban tipikusan speciális nagy mátrixokkal találkozunk, általában ezek **ritka** mátrixok. A ritka mátrixokkal az egyenletrendszer explicit "egzakt" megoldása helyett iteratív sémákat érdemes követni. Ezekből mutatunk be párat. Bontsuk fel az  $\mathbf{A}$  mátrixot a következőképpen

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U} \quad (4.136)$$

ahol  $\mathbf{D}$  az  $\mathbf{A}$  diagonális része,  $\mathbf{L}$  az alsó triangulárisa és  $\mathbf{U}$  pedig a felső trianguláris.

*Jacobi-iteráció:* Az egyenlet átrendezve

$$\mathbf{D} \vec{v} = -[\mathbf{L} + \mathbf{U}] \vec{v} + \vec{b} = (\mathbf{D} - \mathbf{A}) \vec{v} + \vec{b} \implies \vec{v} = \mathbf{G}_J \vec{v} + \mathbf{D}^{-1} \vec{b} \text{ ahol } \mathbf{G}_J = \mathbf{I} - \mathbf{D}^{-1} \mathbf{A} \quad (4.137)$$

szucesszív módszerrel oldjuk meg: induljunk ki valamely  $\vec{v}^{(0)}$ -ból és iteráljunk

$$\vec{v}^{(p+1)} = \mathbf{G}_J \vec{v}^{(p)} + \mathbf{D}^{-1} \vec{b} \quad (4.138)$$

Az eljárás konvergenciája a  $\mathbf{G}_J$  mátrix sajátértékeitől függ. A legnagyobb sajátérték abszolút értékének egynél kisebbnek kell lennie a konvergenciához. Az iteratív eljárásokban a konvergencia biztosíthatósága mellett annak 'gyorsasága' is fontos, hiszen ettől függ, hogy hányszor kell az iterációt elvégezni a kívánt pontosság eléréséhez. A konvergencia gyorsasága az

$$Err(p) = \vec{v}^{(p)} - \vec{v} \quad (4.139)$$

hiba csökkenésének a mértékétől függ. Az a jó, ha  $Err(p)$  gyorsan csökken az iterációk  $p$  számának függvényében. Célszerűen a

$$q_J = \frac{|Err(p+1) - Err(p)|}{|Err(p)|} = \frac{|(\vec{v}^{(p+1)} - \vec{v}) - (\vec{v}^{(p)} - \vec{v})|}{|\vec{v}^{(p)} - \vec{v}|} = \frac{|\vec{v}^{(p+1)} - \vec{v}^{(p)}|}{|\vec{v}^{(p)} - \vec{v}|} \quad (4.140)$$

hányadossal jellemezzük a konvergencia sebességgel. A Jacobi-iterációnál a leglassabban konvergáló vektorokra asszimptotikusan

$$q_J^* \approx |1 - \lambda_J| \quad (4.141)$$

ahol  $\lambda_J$  a  $\mathbf{G}_J$  legnagyobb sajátértéke (spektrál sugara). Láthatóan az eljárás nagyon lassan konvergál, ha a sajátértékek között van egyhez közeli. Sajnos PDE numerikus megoldásánál ez a patológikus eset a tipikus...

*Gauss-Seidel-iteráció:* A baloldalra átvihetjük az alsó trianguláris részt, ha az algoritmust úgy szervezzük, hogy  $v_1^{(p+1)}, v_2^{(p+1)}, \dots$  sorrendben határozzuk meg a vektor elemeit. Az így kapott iterációs séma

$$[\mathbf{L} + \mathbf{D}] \vec{v}^{(p+1)} = -\mathbf{U} \vec{v}^{(p)} + \vec{b} \quad (4.142)$$

Ez lényegében azt jelenti, hogy a már meghatározott  $v_i^{(p+1)}$  értékeket azonnal felhasználjuk a  $v_j^{(p+1)}$ ,  $j > i$  számolásakor a Jacobi-iterációval szemben, ahol minden új értéket kizárólag a  $\vec{v}^{(p)}$  alapján számoltunk. A konvergencia most a

$$\mathbf{G}_{GS} = -[\mathbf{L} + \mathbf{D}]^{-1} \mathbf{U} \quad (4.143)$$

mátrix sajátértékeitől függ. Megmutatható, hogy

$$\lambda_{GS} = \lambda_J^2 \quad (4.144)$$

így az eljárás az előbbinél gyorsabban konvergál

$$q_{GS}^* \approx |1 - \lambda_J^2| \quad (4.145)$$

*Túlrelaxálás (Successive Over-Relaxation, SOR):* Írjuk át a GS egyenletet

$$[\mathbf{L} + \mathbf{D}] \vec{v}^{(p+1)} = [\mathbf{A} - \mathbf{U}] \vec{v}^{(p)} - [\mathbf{A} \vec{v}^{(p)} - \vec{b}] \quad (4.146)$$

majd 'rontsuk el' a jobboldalt egy  $\omega$  relaxációs paraméterrel az alábbi módon:

$$[\mathbf{A} - \mathbf{U}] \vec{v}^{(p)} - \omega [\mathbf{A} \vec{v}^{(p)} - \vec{b}] \quad (4.147)$$

Az így kapott

$$[\mathbf{L} + \mathbf{D}] \vec{v}^{(p+1)} = -[\mathbf{U} - (1 - \omega)\mathbf{A}] \vec{v}^{(p)} + \omega \vec{b} \quad (4.148)$$

SOR formula konvergál, ha  $0 \leq \omega \leq 2$ . Ha  $\omega < 1$ , akkor alulrelaxálásról, ha  $\omega > 1$  akkor túlrelaxálásról beszélünk. A Gauss-Seidel-iterációt kapjuk vissza, ha  $\omega = 1$ . Az SOR formula végsősoron megfelel egy olyan GS eljárásnak, ahol az egyes lépésekben kapott vektorokat összekeverjük az előzővel

$$\omega \vec{v}_{GS}^{(n+1)} + (1 - \omega) \vec{v}^{(n)} \longrightarrow \vec{v}_{SOR}^{(n+1)} \quad (4.149)$$

Az eljárás konvergenciáját most a

$$\mathbf{G}_{SOR} = -[\mathbf{L} + \mathbf{D}]^{-1} [\mathbf{U} - (\mathbf{I} - \omega)\mathbf{A}] \quad (4.150)$$

mátrix határozza meg. Az optimális  $\omega$

$$\omega_{SOR} = \frac{2}{1 + \sqrt{1 - \lambda_J^2}} \quad (4.151)$$

amikor is a spektrálsugár

$$\lambda_{SOR} = \left( \frac{\lambda_J}{1 + \sqrt{1 - \lambda_J^2}} \right)^2 \quad (4.152)$$

és az asszimptotikus konvergencia hányados

$$q_{SOR}^* \approx |1 - \lambda_{SOR}| \quad (4.153)$$

*Iteratív javítás:* Bármelyik módszerrel oldjuk is meg a méretes lineáris egyenletrendszert, a véges számábrázolásból származó kerekítési hibák úgy felhalmozódnak, hogy az eredményben nem szabad túlzottan megbízni. Az

$$\mathbf{A}\vec{v} = \vec{b} \quad (4.154)$$

egyenlet 'igazi'  $\vec{v}$  megoldása helyett a megtalált  $\vec{v} + \delta\vec{v}$  visszahelyettesíthető az egyenletbe

$$\mathbf{A}(\vec{v} + \delta\vec{v}) = \vec{b} + \delta\vec{b} \quad (4.155)$$

ahonnan a hibára kapjuk, hogy

$$\mathbf{A}\delta\vec{v} = \delta\vec{b} \quad (4.156)$$

Ezt az egyenletrendszert (ha  $\delta\vec{b}$  nem tolerálhatóan kicsiny) megintcsak megoldhatjuk a  $\delta\vec{v}$  hibára és korrigálhatjuk az előző jóslatunkat.

#### 4.4. Elliptikus egyenletek: peremérték-probléma

A kezdetiérték-feladatok tárgyalásánál a fő probléma azzal volt, hogy a kézenfekvő megoldási sémák stabilitását biztosítani kellett. A peremérték-problémák megoldásánál a stabilitás viszonylag könnyen elérhető, a fő szempont most az lesz, hogy mennyire hatékony, gazdaságos algoritmust vagyunk képesek felállítani.

Alapfeladatként a kétdimenziós Laplace-egyelet

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = \rho(x, y) \quad (4.157)$$

véges differenciákkal való megoldását vizsgáljuk az ismeretlen  $V(x, y)$  meghatározására adott  $\rho(x, y)$  mellett. Peremfeltételként előírjuk a vizsgált kétdimenziós tartományt határoló görbén az ismeretlen  $V(x, y)$  függvény viselkedését (pl. Dirichlet-, vagy Neumann-határfeltételeket).

A síktartományban felvett gridpontokon

$$(x_j = x_0 + j\delta x, j = 0, 1, \dots, J) ; (y_l = y_0 + l\delta y, l = 0, 1, \dots, L) \quad (4.158)$$

$V(x, y)$ -t a

$$u_{j,l} = V(x_j, y_l) \quad (4.159)$$

diszkrét értékeivel adjuk meg. Az egyszerűség kedvéért most válasszuk a beosztást úgy, hogy legyen

$$\delta x = \delta y = \Delta \quad (4.160)$$

A differenciaegyenlet ekkor (a belső pontokban  $(j = 1, \dots, J-1)$  ,  $(l = 1, \dots, L-1)$ )

$$\frac{u_{j+1,l} - 2u_{j,l} + u_{j-1,l}}{\Delta^2} + \frac{u_{j,l+1} - 2u_{j,l} + u_{j,l-1}}{\Delta^2} = \rho_{j,l} \quad (4.161)$$

Használjunk most Dirichlet-peremfeltételeket, azaz legyenek az  $\{u_{j,0}, u_{0,l}, u_{j,L}, u_{J,l} \mid j = 0..J, l = 0..L\}$  adottak.

#### 4.4.1. Relaxáció

Az  $\mathbf{u}_j = (u_{j,1}, u_{j,2}, \dots, u_{j,L-1})$  és az  $\mathbf{r}_j = \Delta^2(\rho_{j,1}, \rho_{j,2}, \dots, \rho_{j,L-1}) - (u_{j,0}, 0, \dots, 0, u_{j,L})$  vektorok bevezetésével, az ismert mennyiségeket a jobboldalra rendezve az egyenletrendszer

$$\mathbf{T} \cdot \mathbf{u}_1 + \mathbf{I} \cdot \mathbf{u}_2 = \mathbf{r}_1 - \mathbf{u}_0 \quad (4.162)$$

$$\mathbf{I} \cdot \mathbf{u}_{j-1} + \mathbf{T} \cdot \mathbf{u}_j + \mathbf{I} \cdot \mathbf{u}_{j+1} = \mathbf{r}_j \quad j = 2, \dots, J-2 \quad (4.163)$$

$$\mathbf{I} \cdot \mathbf{u}_{J-2} + \mathbf{T} \cdot \mathbf{u}_{J-1} = \mathbf{r}_{J-1} - \mathbf{u}_J \quad (4.164)$$

alakba írható, ahol

$$\mathbf{T} = \begin{pmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{pmatrix} \quad \text{és} \quad \mathbf{I} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 1 \\ & & & & & 1 \end{pmatrix} : (L-1) \times (L-1) \quad (4.165)$$

A  $(J-1) \cdot (L-1)$  dimenziós  $\vec{v} = \{\mathbf{u}_1 \oplus \mathbf{u}_2 \oplus \dots \oplus \mathbf{u}_{J-1}\}$  és  $\vec{b} = \{(\mathbf{r}_1 - \mathbf{u}_0) \oplus \mathbf{r}_2 \oplus \dots \oplus (\mathbf{r}_{J-1} - \mathbf{u}_J)\}$  szupervektorokkal a meghatározandó

$$v_r = u_{j,l} \quad , \quad r = (j-1)L + l \quad (4.166)$$

mennyiségek az

$$\mathbf{A} \cdot \vec{v} = \vec{b} \quad \text{ahol} \quad \mathbf{A} = \begin{pmatrix} \mathbf{T} & \mathbf{I} & & & \\ \mathbf{I} & \mathbf{T} & \mathbf{I} & & \\ & \mathbf{I} & \mathbf{T} & \mathbf{I} & \\ & & \ddots & \ddots & \ddots \\ & & & \mathbf{I} & \mathbf{T} & \mathbf{I} \\ & & & & \mathbf{I} & \mathbf{T} \end{pmatrix} \quad (4.167)$$

lineáris egyenletrendszer megoldásával megkaphatók. Az ilyen pentadiagonális (következésképp ritka) mátrixokkal felírt lineáris egyenletek a már megismert (Jacobi, Gauss-Seidel és SQR) módszerekkel megoldhatók.

#### 4.4.2. Váltakozó irányú implicit módszer: ADI

Az  $\mathcal{D}V = \rho$  peremérték-problémával kapcsolatba hozható a

$$\frac{\partial V}{\partial t} = \mathcal{D}V - \rho \quad (4.168)$$

diffúziós probléma. A diffúziós egyenlet  $t \rightarrow \infty$  megoldásai stacionáriusak, ezek éppen a peremértékfeladat keresett megoldásai. Esetünkben kézenfekvő az ADI eljárást felírni

$$\frac{\vec{v}^{[n+1/2]} - \vec{v}^{[n]}}{\delta t/2} = \frac{\mathcal{L}_x \vec{v}^{[n+1/2]} + \mathcal{L}_y \vec{v}^{[n]}}{\Delta^2} - \vec{\rho} \quad \text{és} \quad \frac{\vec{v}^{[n+1]} - \vec{v}^{[n+1/2]}}{\delta t/2} = \frac{\mathcal{L}_x \vec{v}^{[n+1/2]} + \mathcal{L}_y \vec{v}^{[n+1]}}{\Delta^2} - \vec{\rho}$$

Átrendezve

$$\begin{aligned} [q\mathbf{I} - \mathcal{L}_x] \vec{v}^{[n+1/2]} &= [q\mathbf{I} + \mathcal{L}_y] \vec{v}^{[n]} - \Delta^2 \vec{\rho} \\ [q\mathbf{I} - \mathcal{L}_y] \vec{v}^{[n+1]} &= [q\mathbf{I} + \mathcal{L}_x] \vec{v}^{[n+1/2]} - \Delta^2 \vec{\rho}, \quad \text{ahol } q = 2 \frac{\Delta^2}{\delta t} \end{aligned}$$

a bal oldalon álló  $[q\mathbf{I} - \mathcal{L}_i]$  mátrixok tridiagonálisak, így az egyenlet közvetlenül megoldható. Kiindulunk valamely  $\vec{v}^{[0]}$ -ból, ezzel előállítjuk az első sor szerint  $\vec{v}^{[1/2]}$ -et, majd a második sorral  $\vec{v}^{[1]}$ -et. Ezt az első egyenletbe helyettesítjük és addig folytatjuk az eljárást, amíg  $|\vec{v}^{[n+1]} - \vec{v}^{[n]}| < \epsilon$  nem lesz. Az eljárás konvergenciája a képzeletbeli időtengelyen felvett  $\delta t$  lépéshossztól és így  $q$ -tól függ. Ezen értékek dinamikus (iterációnként különböző) megválasztásával elérhető, hogy lényegesen gyorsabban jussunk célhoz, mint pl. az SOR módszerrel.

#### 4.4.3. Ciklikus redukció

Válasszuk a beosztást úgy, hogy  $J = 2^p$  legyen. Tekintsünk három egymás utáni

$$\mathbf{I} \cdot \mathbf{u}_{j-2} + \mathbf{T} \cdot \mathbf{u}_{j-1} + \mathbf{I} \cdot \mathbf{u}_j = \mathbf{r}_{j-1} \quad (4.169)$$

$$\mathbf{I} \cdot \mathbf{u}_{j-1} + \mathbf{T} \cdot \mathbf{u}_j + \mathbf{I} \cdot \mathbf{u}_{j+1} = \mathbf{r}_j \quad (4.170)$$

$$\mathbf{I} \cdot \mathbf{u}_j + \mathbf{T} \cdot \mathbf{u}_{j+1} + \mathbf{I} \cdot \mathbf{u}_{j+2} = \mathbf{r}_{j+1} \quad (4.171)$$

egyenletet. A középsőt megszorozva  $-\mathbf{T}$ -vel és összeadva a hármat

$$\mathbf{I} \cdot \mathbf{u}_{j-2} + \mathbf{T}^{(1)} \cdot \mathbf{u}_j + \mathbf{I} \cdot \mathbf{u}_{j+2} = \mathbf{r}_j^{(1)} \quad (4.172)$$

ahol

$$\mathbf{T}^{(1)} = [\mathbf{2I} - \mathbf{T}^2] \quad \text{és} \quad \mathbf{r}_j^{(1)} = \mathbf{r}_{j-1} - \mathbf{T} \cdot \mathbf{r}_j + \mathbf{r}_{j+1} \quad (4.173)$$

Ezzel a centrális  $\mathbf{u}_j$  meghatározható első szomszédai helyett csupán a második szomszédjaiból, azaz a meghatározandó  $\mathbf{u}$ -k számát a felére csökkentettük (ebben a lépésben kiszórtunk minden páratlan indexű  $\mathbf{u}_j$ -t). Az így kapott egyenlet szerkezetében azonos az eredetivel, így ezt újból és újból redukálhatjuk. A végén

$$\Downarrow \quad (4.174)$$

$$\mathbf{u}_0 + \mathbf{T}^{(p-1)} \cdot \mathbf{u}_{J/4} + \mathbf{u}_{J/2} = \mathbf{r}_{J/4}^{(p-1)} \quad (4.175)$$

$$\mathbf{u}_{J/4} + \mathbf{T}^{(p-1)} \cdot \mathbf{u}_{J/2} + \mathbf{u}_{3J/4} = \mathbf{r}_{J/2}^{(p-1)} \quad (4.176)$$

$$\mathbf{u}_{J/2} + \mathbf{T}^{(p-1)} \cdot \mathbf{u}_{3J/4} + \mathbf{u}_J = \mathbf{r}_{3J/4}^{(p-1)} \quad (4.177)$$

$$\Downarrow \quad (4.178)$$

$$\mathbf{u}_0 + \mathbf{T}^{(p)} \cdot \mathbf{u}_{J/2} + \mathbf{u}_J = \mathbf{r}_{J/2}^{(p)} \quad (4.179)$$

Az utolsó egyenletben  $\mathbf{T}^{(p)}$  és  $\mathbf{r}_{J/2}^{(p)}$  ismert (előállítható) az  $\mathbf{u}_0$  és  $\mathbf{u}_J$  pedig adott a peremfeltételből. Következésképpen  $\mathbf{u}_{J/2}$  kiszámítható. Ebből a  $(p-1)$  lépés egyenleteire visszalépve  $\mathbf{u}_{J/4}$  és  $\mathbf{u}_{3J/4}$  számolható, és így tovább számolva sorra előállíthatjuk a megoldást a többi pontban is.

#### 4.4.4. Fourier-módszer

A Fourier-transzformáltakkal fontosságuk miatt külön fejezetben foglalkozunk részletesen. Megelőlegezve most néhány formulát a diszkrét Fourier-transzformáció témaköréből, megvizsgáljuk, hogy a konstans együttthatós parciális differenciálegyenleteket hogyan oldhatjuk meg ezek segítségével. Példaként a Poisson-egyenletet vesszük elő, ezt véges differenciákkal az (4.161) egyenletben

$$u_{j+1,l} + u_{j-1,l} + u_{j,l+1} + u_{j,l-1} - 4u_{j,l} = \Delta^2 \rho_{j,l} \quad (4.180)$$

alakra hoztuk. Az  $\{u_{j,l}\}$  és  $\{\rho_{j,l}\}$  mennyiségek kétdimenziós diszkrét Fourier-transzformáltja:

$$\tilde{u}_{m,n} = \sum_{j=0}^{J-1} \sum_{l=0}^{L-1} u_{j,l} \cdot e^{i2\pi jm/J} e^{i2\pi ln/L} \quad u_{j,l} = \frac{1}{JL} \sum_{m=0}^{J-1} \sum_{n=0}^{L-1} \tilde{u}_{m,n} \cdot e^{-i2\pi jm/J} e^{-i2\pi ln/L} \quad (4.181)$$

$$\tilde{\rho}_{m,n} = \sum_{j=0}^{J-1} \sum_{l=0}^{L-1} \rho_{j,l} \cdot e^{i2\pi jm/J} e^{i2\pi ln/L} \quad \rho_{j,l} = \frac{1}{JL} \sum_{m=0}^{J-1} \sum_{n=0}^{L-1} \tilde{\rho}_{m,n} \cdot e^{-i2\pi jm/J} e^{-i2\pi ln/L} \quad (4.182)$$

Ezeket a differenciaegyenletbe helyettesítve

$$\tilde{u}_{m,n} \left[ e^{-i2\pi m/J} + e^{i2\pi m/J} + e^{-i2\pi n/L} + e^{i2\pi n/L} - 4 \right] = \Delta^2 \tilde{\rho}_{m,n} \quad (4.183)$$

vagy

$$\tilde{u}_{m,n} = \frac{\Delta^2}{2} \frac{\tilde{\rho}_{m,n}}{\cos(2\pi m/J) + \cos(2\pi n/L) - 2} \quad (4.184)$$

A probléma megoldása ezek után a következő lépésekben történik

- meghatározzuk a  $\tilde{\rho}_{m,n}$  Fourier-transzformáltakat
- az (4.184) egyenlet alapján kiszámítjuk az  $\tilde{u}_{m,n}$  mátrixot
- az inverz formulákkal  $u_{j,l}$  (4.181) alapján előáll

A most vázolt eljárás bármely *periodikus határfeltétel* mellett használható, azaz a megoldásunkra  $u_{j,l} = u_{j+J,l} = u_{j,l+L}$  és különösen a peremen  $u_{0,l} = u_{J,l}$   $u_{j,0} = u_{j,L}$ .

Ha *Dirichlet-peremfeltételünk* van, azaz a megoldás a peremen eltűnik:  $u_{0,l} = u_{j,0} = u_{J,l} = u_{j,L} = 0$ , akkor a **szinusz**-transzformáltakat célszerű felírni:

$$\tilde{u}_{m,n} = \sum_{j=1}^{J-1} \sum_{l=1}^{L-1} u_{j,l} \cdot \sin\left(\frac{\pi jm}{J}\right) \sin\left(\frac{\pi ln}{L}\right) \quad u_{j,l} = \frac{4}{JL} \sum_{m=1}^{J-1} \sum_{n=1}^{L-1} \tilde{u}_{m,n} \cdot \sin\left(\frac{\pi jm}{J}\right) \sin\left(\frac{\pi ln}{L}\right) \quad (4.185)$$

$$\tilde{\rho}_{m,n} = \sum_{j=1}^{J-1} \sum_{l=1}^{L-1} \rho_{j,l} \cdot \sin\left(\frac{\pi jm}{J}\right) \sin\left(\frac{\pi ln}{L}\right) \quad \rho_{j,l} = \frac{4}{JL} \sum_{m=1}^{J-1} \sum_{n=1}^{L-1} \tilde{\rho}_{m,n} \cdot \sin\left(\frac{\pi jm}{J}\right) \sin\left(\frac{\pi ln}{L}\right) \quad (4.186)$$

Ilyen felírással a peremfeltétel automatikusan kielégül. A transzformált egyenlet

$$\tilde{u}_{m,n} = \frac{\Delta^2}{2} \frac{\tilde{\rho}_{m,n}}{\cos(\pi m/J) + \cos(\pi n/L) - 2} \quad (4.187)$$

aminek a megoldási folyamata az előzőével azonos.

*Neumann-peremfeltétel*nél  $\nabla u = 0$  a peremen. Ez esetben a koszinusz transzformáltakat használjuk

$$\tilde{u}_{m,n} = \sum_{j=0}^{J(!)} \sum_{l=0}^{L(!)} u_{j,l} \cdot \cos\left(\frac{\pi jm}{J}\right) \cos\left(\frac{\pi ln}{L}\right) \quad u_{j,l} = \frac{4}{JL} \sum_{m=0}^{J(!)} \sum_{n=0}^{L(!)} \tilde{u}_{m,n} \cdot \cos\left(\frac{\pi jm}{J}\right) \cos\left(\frac{\pi ln}{L}\right) \quad (4.188)$$

$$\tilde{\rho}_{m,n} = \sum_{j=0}^{J(!)} \sum_{l=0}^{L(!)} \rho_{j,l} \cdot \cos\left(\frac{\pi jm}{J}\right) \cos\left(\frac{\pi ln}{L}\right) \quad \rho_{j,l} = \frac{4}{JL} \sum_{m=0}^{J(!)} \sum_{n=0}^{L(!)} \tilde{\rho}_{m,n} \cdot \cos\left(\frac{\pi jm}{J}\right) \cos\left(\frac{\pi ln}{L}\right) \quad (4.189)$$

ahol a (!) jelzés arra figyelmeztet, hogy az összegzéseben, ha az index a határon van, akkor a megfelelő tagnak csak a felét kell számítani.

## 6. FOURIER-TRANSZFORMÁCIÓ

### 6.1. A Fourier-transzformációról általában

Az  $f(x)$  függvény Fourier-transzformáltját a fizikában a

$$H(\omega) = \int_{-\infty}^{\infty} h(t)e^{-i\omega t} dt \quad \text{vagy} \quad H(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(t)e^{-i\omega t} dt \quad (6.1)$$

módon szokás definiálni. Az inverz Fourier-transzformáció szerint (amennyiben az alábbi integrálok léteznek) rendre

$$h(t) \simeq \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega)e^{i\omega t} d\omega \quad \text{vagy} \quad h(t) \simeq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H(\omega)e^{i\omega t} d\omega \quad (6.2)$$

A direkt és az inverz Fourier-transzformációban az exponens előjele ellentétes, de néha szokás fordított sorrendben érteni. A most tárgyalni kívánt numerikus Fourier-transzformációk összefüggésében az  $(1, 2\pi, \sqrt{2\pi})$  faktorokban megjelenő diverzitást a

$$f = \frac{\omega}{2\pi} \quad (6.3)$$

változócserevel oldjuk fel, az exponens előjelét pedig úgy választjuk, hogy Fourier-transzformált  $\{h(x) \Longleftrightarrow H(f)\}$  párról beszélünk, ha

$$H(f) = \int_{-\infty}^{\infty} h(t)e^{i2\pi f t} dt \quad \text{és} \quad h(t) = \int_{-\infty}^{\infty} H(f)e^{-i2\pi f t} df \quad (6.4)$$

Emlékeztetőül néhány összefüggés a Fourier-párookra

$$\begin{aligned} h(at) &\Longleftrightarrow \frac{1}{|a|} H\left(\frac{f}{a}\right) & \frac{1}{|a|} h\left(\frac{t}{a}\right) &\Longleftrightarrow H(af) \\ h(t-a) &\Longleftrightarrow H(f)e^{i2\pi fa} & h(t)e^{-i2\pi ta} &\Longleftrightarrow H(f-a) \\ h^{(n)}(t) &\Longleftrightarrow (-i2\pi f)^n H(f) & (i2\pi t)^n h(t) &\Longleftrightarrow H^{(n)}(f) \end{aligned} \quad (6.5)$$

További hasznos fogalmak, tételek:

- Konvolúciós tétel

$$g * h \equiv \int_{-\infty}^{\infty} h(t-\tau)g(\tau)d\tau = h * g \quad g * h \Longleftrightarrow G \cdot H \quad (6.6)$$

- Korreláció

$$h|g \equiv \int_{-\infty}^{\infty} h(t+\tau)g(\tau)d\tau \quad h|g \Longleftrightarrow H \cdot G^* \quad (* \text{ komplex konjugált}) \quad (6.7)$$

- Autokorreláció

$$h|h \equiv \int_{-\infty}^{\infty} h(t+\tau)h(\tau)d\tau \quad h|h \Longleftrightarrow |H|^2 \quad (6.8)$$

A  $h(t)$  függvényhez tartozó  $P(f) \equiv |H(f)|^2$  más néven a spektrális teljesítménysűrűség.

### 6.2. Mintavételezés

Vizsgáljuk azt az esetet, amikor a transzformálandó függvényünk ekvidisztáns beosztáson az

$$h_n = h(n\Delta) \quad n = \dots, -2, -1, 0, 1, 2, \dots \quad (6.9)$$

értékekkel adott. Ehhez a mintához tartozik egy úgynevezett 'kritikus frekvencia', a *Nyquist-frekvencia*

$$f_c = \frac{1}{2\Delta} \quad (6.10)$$

A kritikus frekvencia jelentősége megmutatkozik az alábbi tételben:

**Mintavételezési tétel:** Ha a  $h(t)$  folytonos függvény *sávhatárolt*, azaz létezik olyan  $f_M$ , hogy

$$H(f) = 0 \quad \text{ha} \quad |f| > f_M \quad (6.11)$$

akkor  $h(t)$  teljesen megadható a

$$h_n = h\left(\frac{n}{2f_M}\right) \quad (6.12)$$

mintavételezéssel és ekkor

$$h(t) = \Delta \sum_{n=-\infty}^{\infty} h_n \frac{\sin[2\pi f_M (n\Delta - t)]}{\pi (n\Delta - t)} \quad (6.13)$$

A tétel nagy jelentőségű sok fizikai/műszaki problémánál, amikor biztosak lehetünk abban, hogy a jel sávhatárolt. Ugyanakkor felhívja a figyelmet arra is, hogy a jelek nem megfelelő mintavételezése ( $f_c < f_M$ ) alapján kiszámolt Fourier-transzformáltak hamisak lesznek. Különösen biztosak lehetünk ebben, ha a függvény nem sávhatárolt, azaz  $f_M \sim \infty$ . Sajnos ez is tipikus eset a gyakorlati problémákban. Ilyenkor a  $P(f)$  teljesítménysűrűség azon része, amire  $|f| > f_M$  hozzáadódik az  $|f| < f_M$  tartományon felvett értékekhez, a mintából készített Fourier-transzformált hamis lesz.

Nem mindig tudjuk előre, hogy a függvényünk sávhatárolt-e, vagy mennyire jó közelítéssel az. Egy mintavételezési gyakoriságról akkor mondhatjuk el, hogy praktikusán kielégítő, ha az eredményül kapott Fourier-transzformáltra teljesül, hogy az kellő mértékben eltűnik miközben a frekvencia felülről közelít valamely  $-f_c$  vagy alulról egy  $f_c$  értéket.

### 6.3. Diszkrét Fourier-transzformáció (DFT)

Tegyük most fel, hogy  $h(t)$  tartója (vagy legalábbis az a tartomány, ahol a függvény viselkedése érdekes) valamely véges intervallumba esik. A vizsgált tartomány a függvény alkalmas eltolásával választható úgy, hogy a  $[0, T]$  intervallum legyen. Valamely

$$h_k = h(k\Delta) \quad , \quad k = 0, 1, \dots, N-1 \quad (6.14)$$

beosztást véve a Fourier-transzformáltat az integrál közelítésével kiszámolhatjuk

$$H(f) = \int_0^T h(t) e^{i2\pi f t} dt \approx \Delta \sum_{k=0}^{N-1} h_k e^{i2\pi f k \Delta} \quad (6.15)$$

Valójában nincs értelme tetszőleges  $f$  frekvenciára ezt kiszámolni, hiszen az  $N$  darab  $h_k$  számból csak  $N$  darab független transzformált szám állítható elő. Legyen  $N$  páros szám és keressük csak a

$$H(f_n) = H\left(\frac{n}{N\Delta}\right) \approx \Delta \sum_{k=0}^{N-1} h_k e^{i2\pi k n / N} \quad , \quad f_n = \frac{n}{N\Delta} = \frac{n}{T} \quad , \quad n = -\frac{N}{2}, \dots, +\frac{N}{2} \quad (6.16)$$

függvényértékeket. Az itt szereplő

$$H_n \equiv \sum_{k=0}^{N-1} h_k e^{i2\pi k n / N} \quad (6.17)$$

számokat az  $N$  darab  $h_k$  pont **diszkrét Fourier-transzformáltjának** nevezzük. Vegyük észre, hogy  $H_{-N/2} = H_{N/2}$ , így pontosan  $N$  független értéket kapunk. Ezekkel a függvény Fourier-transzformáltja

$$H(f_n) \approx \Delta H_n \quad (6.18)$$



A diszkrét Fourier-transzformáltban a negatív indexek kényelmetlenné válhatnak. Látható a definícióból, hogy az indexhatárokat tetszőlegesen kiterjesztve a kifejezés periódikus az indexben, azaz  $H_{-n} = H_{N-n}$ , így kézenfekvő, hogy a negatív indexekkel jellemzett tartományt áthelyezzük egy index-periódussal feljebb. Válasszuk tehát az indexeléshez az  $n = 0, 1, \dots, N-1$  számokat. Így  $n = 0$  a nulla frekvenciához tartozó szám, a pozitív frekvenciás  $0 < f < f_c$  transzformáltak  $1 \leq n \leq N/2 - 1$  indexűek, míg a negatív frekvenciákhoz tartozó értékek az  $N/2 + 1 \leq n \leq N-1$  helyen vannak. Az  $n = N/2$  indexű transzformált  $f_c$  és  $-f_c$ -hez egyaránt hozzátartozik.

Az **inverz diszkrét Fourier-transzformáció**

$$h_k \equiv \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{-i2\pi kn/N} \quad (6.19)$$

hasznló módon számolható, mint a direkt DFT, a rutinban az exponens előjelét kell csak változtatni, a normálás pedig a rutinon kívül elvégezhető. A formula igazolására helyettesítsünk be

$$h_k \equiv \frac{1}{N} \sum_{n=0}^{N-1} \sum_{l=0}^{N-1} h_l e^{i2\pi ln/N} e^{-i2\pi kn/N} = \sum_{l=0}^{N-1} h_l \sum_{n=0}^{N-1} \left( e^{i2\pi(l-k)/N} \right)^n \quad (6.20)$$

és vegyük észre, hogy

$$\sum_{n=0}^{N-1} \left( e^{i2\pi(l-k)/N} \right)^n = N \quad , \quad \text{ha} \quad l = k \quad (6.21)$$

egyébként pedig

$$\sum_{n=0}^{N-1} \left( e^{i2\pi(l-k)/N} \right)^n = \frac{1 - \left( e^{i2\pi(l-k)/N} \right)^N}{1 - e^{i2\pi(l-k)/N}} = \frac{1 - 1}{1 - e^{i2\pi(l-k)/N}} = 0 \quad (6.22)$$

#### 6.4. Gyors Fourier-Transzformáció (FFT)

A diszkrét Fourier-transzformáltak

$$H_n = \sum_{k=0}^{N-1} h_k w_{n,k} \quad \text{ahol} \quad w_{n,k} = e^{i2\pi kn/N} \quad (6.23)$$

kiszámítása egy komplex mátrixszorzás munkigényének felel meg.  $N \times N$ -es mátrixokra ez  $O(N^2)$  művelet. A gyors Fourier-transzformáció (**FFT**: *Fast Fourier Transform*) speciális esetekben ennél lényegesen kevesebb számolási munkával teszi lehetővé a transzformáció elvégzését. Válasszuk ugyanis a mintavételezésben használt pontok számát úgy, hogy az kettőnek valamely egész hatványa legyen

$$N = 2^p \quad (6.24)$$

Ha valamilyen okból nem áll módunkban így megválasztani a minták számát, akkor egészítsük ki a mintát fiktív nulla értékekkel amíg kettő valamely hatványa nem lesz a kibővített minta. Ugyanis olyan eljárást fogunk adni, ami ilyen esetben csak  $O(pN)$  számítási műveletet igényel. Az FFT algoritmus lényege, hogy az  $N$  pont diszkrét Fourier-transzformációját visszavezetjük két fele akkora méretű, azaz  $N/2$  méretű pontsereg transzformációjára. A dolog menetét a Danielson és Lánzos nevével jegyzett alábbi lemma mutatja.

$$\sum_{k=0}^{N-1} h_k e^{i2\pi kn/N} = \sum_{k=0}^{N/2-1} h_{2k} e^{i2\pi(2k)n/N} + \sum_{k=0}^{N/2-1} h_{2k+1} e^{i2\pi(2k+1)n/N} \quad (6.25)$$

$$= \sum_{k=0}^{N/2-1} h_{2k} e^{i2\pi kn/(N/2)} + W^n \sum_{k=0}^{N/2-1} h_{2k+1} e^{i2\pi kn/(N/2)} \quad (6.26)$$

$$H_n = F_n^e + W^n F_n^o \quad , \quad W = e^{i2\pi/N} \quad (6.27)$$

ahol az utóbbi egyenletben  $F_n^e$  a minta páros sorszámú,  $F_n^o$  pedig csak a páratlan sorszámú pontjaiból készült DFT. Figyeljünk arra, hogy az egyenlet jobb oldalán  $n = 0, 1, \dots, N$ , míg a fele akkora méretű pontthalmazból készült  $F_n^e$  és

$F_n^o$  indexei  $n = 0, 1, \dots, N/2$  Az indexek szerinti periodikusság  $F_n = F_{n \pm N/2}$  miatt azonban egyszerűen kétszer egymás után kell leírunk az  $F_n$  számokat.

Ha az eredeti pontjaink száma  $N = 2^p$ , akkor ezt a 'felezéses' eljárás  $p$ -szer rekurzív módon használhatjuk, míg végül csak egyetlen pontból álló halmaz triviális diszkrét Fourier-transzformációját kell végrehajtunk.

Az FFT rutinok segítségével elvégezhetjük a diszkrét Fourier-transzformációt akkor is, ha a pontjaink száma nem kettő hatványa. Ekkor azonban az előbbi redukció nem, vagy csak részben végezhető el és nem biztos, hogy az "FFT" gyorsabb lesz, mint a közönséges direkt számolás. Néhány szerencsés eset:

- A Danielson-Lánczos teljes faktorizációhoz hasonlóan részleges faktorizációt alkalmazhatunk, ha a pontjaink számának van egész osztója. Jó esetben  $N$  néhány kis prímszám szorzataként előáll. Az iménti módon akkora méretű tömbök transzformációjára vezethetjük vissza a feladatot, mint  $N$  legnagyobb prím osztója.
- Néhány speciális tömbméret mellett lehetőség van arra, hogy nagyon effektív számítógépes rutint írjunk. Így például  $N = 2, 4, 8$  esetén a transzformációban szereplő trigonometrikus függvények értéke speciálisan egyszerű és hatékonyan elvégezhető a transzformáció. Az ekkor a méretű tömbökre jól megírt résztranszformációkat használva nem kell a Lánczos redukciót végigvinni, ami 20-30% nyereséget hozhat a számításban.
- Hatékony transzformációs eljárások léteznek nemcsak az előbbi tömbméretekre, de pl.  $N = 3, 5, 7, 11, 16, \dots$  számú adathalmazra is. Ilyen például a *Winograd-algoritmus*. Előfordulhat, hogy ezzel az eljárással kétszer gyorsabban tudjuk transzformálni az adatainkat, mint a legközelebb eső  $2^p$ -re kiegészített adatsoron FFT-vel. Nagyméretű adatsorokra az algoritmus nagy hátránya, hogy extra ideiglenes tárolási igénye van, szemben a "közönséges" FFT-vel, ahol a számolás nem igényel munkatömböket.

Részletesebben az FFT programozásával nem foglalkozunk, a megfelelő procedurák, szubrutinok általában készen rendelkezésre állnak a különböző programnyelvekhez. Az előzőekből láthatóan indokolt a tanács: Győződjünk meg arról, hogy milyen rutinok állnak rendelkezésre és készítsük melyik problémához melyik rutint ajánlják.

## 6.5. FFT valós függvényekre

Igen gyakori eset, hogy valós függvényt kell Fourier-transzformálni. Nyilván pocséklás lenne, ha a DFT-t úgy végeznénk el, hogy a komplex bemenő függvényre megírt rutinnak átadnánk a valós részben a tényleges mintákat, míg a minta képzetes részét kitöltenénk nullákkal.

- Ha a probléma olyan, hogy több azonos méretű tisztán valós (vagy tisztán képzetes) adatsort kell transzformálni, akkor két mintát összefoglalva egyetlen komplex mintába egy lépésben két transzformációt is elvégezhetünk amivel teljes mértékben kihasználjuk az általános FFT rutinunkat. Legyenek ugyanis  $f(t)$  és  $g(t)$  valós függvények. A  $h(t) = f(t) + i \cdot g(t)$  komplex függvény Fourier-transzformáltjára a linearitás miatt igaz, hogy

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{i2\pi ft} dt = \int_{-\infty}^{\infty} f(t) e^{i2\pi ft} dt + i \int_{-\infty}^{\infty} g(t) e^{i2\pi ft} dt = F(f) + iG(f) \quad (6.28)$$

Vegyük figyelembe, hogy valós függvényekre

$$(F(f))^* = \left( \int_{-\infty}^{\infty} f(t) e^{i2\pi ft} dt \right)^* = \int_{-\infty}^{\infty} f(t) e^{-i2\pi ft} dt = F(-f) \quad (6.29)$$

így

$$(H(f))^* = F(-f) - iG(-f) \quad \text{illetve} \quad (H(-f))^* = F(f) - iG(f) \quad (6.30)$$

A  $H$  komplex Fourier-transzformáltból tehát a keresett két transzformált egyszerűen visszafejthető:

$$F(f) = \frac{(H(-f))^* + H(f)}{2} \quad \text{és} \quad G(f) = i \frac{(H(-f))^* - H(f)}{2} \quad (6.31)$$

Ne felejtjük, hogy a DFT-ben szokásos módon a negatív frekvenciás együtthatókat a tömbben átmásoltuk és a pozitív frekvenciások után tároljuk, azaz

$$H(-f_n) = H_{N-n} \quad , \quad f_n = \frac{1}{N\Delta} \{0, 1, 2, \dots, N/2\} \quad (6.32)$$

- Ha csak egyetlen adatsorunkat kell transzformálni, akkor az  $N$  elemű valós mintát alakítsuk át  $N/2$  elemű komplex mintává a redundancia elkerülésére. Válasszuk két részre a mintát, célszerűen a páros és a páratlan sorszámú pontokra és legyen

$$h_k = f_{2k} + i \cdot f_{2k+1} \quad , \quad k = 0, 1, \dots, N/2 - 1 \quad (6.33)$$

Az DFT-t elvégezzük erre az  $N/2$  méretű tömbre. Eredményül olyan koeficienseket kapunk amiben a páros és a páratlan pontok külön-külön vett transzformáltjai az alábbi módon vannak kombinálva:

$$H_n = F_n^e + iF_n^o \quad (6.34)$$

A Lánczos-lemma kapcsán leírtak szerint

$$F_n = F_n^e + e^{i2\pi n/N} F_n^o \quad (6.35)$$

lenne az eredetileg keresett transzformált. Ezt az iménti  $H_n$ -ekből kifejezhetjük

$$F_n = \frac{1}{2} [H_n + H_{-n}^*] - \frac{i}{2} [H_n - H_{-n}^*] e^{i2\pi n/N} \quad , \quad n = 0, 1, \dots, N-1 \quad (6.36)$$

ahol felhasznátuk, hogy a valós bemenő adatok miatt  $(F_{-n}^x)^* = F_n^x \quad x = e/o$ . A formulában a negatív indexekre továbbra is érvényes a (6.32) tárolási konvenció.

## 6.6. sinFFT szinusz és cosFFT koszinusz Fourier-transzformáltak

Differenciálegyenletekkel kapcsolatos problémákban gyakran teljesül, hogy a peremfeltételek szerint a megoldásnak el kell tűnnie a vizsgált tartomány határain (a peremen). A  $[0, T]$  intervallum határán eltűnő függvények Fourier-színusz sorba fejthetők

$$h(x) = \sum_{n=1}^{\infty} b_n \cdot \sin\left(x \frac{n\pi}{T}\right) \quad b_n = \frac{2}{T} \int_0^T h(x) \sin\left(x \frac{n\pi}{T}\right) dt \quad (6.37)$$

A függvényekről a mintavételezéssel kapott  $\{h_k \mid k = 0, 1, \dots, N-1\}$  értékekre áttérve a  $b_n$  Fourier-együtthatók helyett használjuk a **diszkrét szinusz Fourier-transzformáltakat**, definíció szerint a

$$B_n = \sum_{k=1}^{N-1} h_k \sin\left(\frac{\pi nk}{N}\right) \approx \frac{T}{2\Delta} b_n = \frac{N}{2} b_n \quad n = (0), 1, 2, \dots, N-1, (N) \quad (6.38)$$

számokat. (Feltevésünk szerint  $h_0, B_0, h_N, B_N = 0$ .) A diszkrét Fourier-színusz transzformáció faktortól eltekintve "önmagá" inverze, pontosabban

$$h_k = \frac{2}{N} \sum_{n=1}^{N-1} B_n \sin\left(\frac{\pi nk}{N}\right) \quad (6.39)$$

Ha nem áll rendelkezésre speciálisan az erre a célra használható gyors színusz Fourier-transzformáció (**sinFFT**) rutin, akkor az átlános FFT eljárást használhatjuk a következő megfontolások alapján. Vezessük be a

$$\tilde{h}(x) = \begin{cases} h(x) & , \quad x > 0 \\ -h(x) & , \quad x < 0 \end{cases} \quad (6.40)$$

páratlan segédfüggvényt. A  $[0, T]$  intervallum helyett a  $[-T, T]$  **duplázott** intervallumon

$$\tilde{H}(f) = \int_{-T}^T \tilde{h}(t) e^{i2\pi ft} dt = 2i \int_0^T h(t) \sin(2\pi ft) dt \approx \Delta 2i \sum_{k=1}^{N-1} h_k \sin(2\pi k \Delta f) \quad (6.41)$$

Ha ezek után a  $2T$  intervallum  $2N$  darab  $\{\tilde{h}_k \mid k = 0, 1, \dots, 2N-1\}$  mintájához a szokásos

$$f_n = \frac{n}{2\Delta N} = \frac{n}{2T} \quad (6.42)$$

frekvenciákat választjuk, akkor  $\tilde{H}(f_n) = \Delta \tilde{H}_n$ , ahol  $\{\tilde{H}_n \mid n = 0, 1, \dots, 2N-1\}$  a  $\tilde{h}$  diszkrét Fourier transzformáltja. A megoldás tehát az, hogy az  $N$  számmal adott  $\{h_k \mid k = 0, 1, \dots, N-1\}$  mintát inverzióval megkétszerezzük úgy, hogy  $\tilde{h}_k = h_k$  ha  $k < N$  és  $\tilde{h}_{2N-k} = -h_k$ , ha  $k > N$  és ezen végrehajtunk egy diszkrét Fourier-transzformációt. Utóbbi eredményéből pedig

$$\tilde{H}_n = 2iB_n \quad (6.43)$$

alapján a szinusz transzformáltak az FFT együtthatókból számolhatók.

- Azt vártuk volna, hogy az  $N$  pont szinusz transzformációja fele akkora munkával végezhető el, mint ugyanannyi pont teljes komplex diszkrét transzformációja, ehelyett olyan eljárást adtunk, hogy az egy kétszer akkora méretű komplex transzformációval lett egyenértékű.
- A pazarlás mindaddig többnyire elfogadható, amíg egyszer, vagy csak néhányszor kell elkövetni.
- Súlyos problémává dagadhat a többletszámolás, ha sokszori transzformációban vagyunk kénytelenek lenyelni, mint például egy többdimenziós Fourier-transzformáció számolásakor. Ebben az esetben - és, ha nem áll rendelkezésre megfelelő könyvtári rutin -, szükséges és lehetséges egyéb trükkökkel minimális szinten tartani a számolási munkát. A részletekkel kapcsolatban utalunk a szakirodalomra.

A koszinusz transzformáltak szintén gyakran szükségesek. Tipikus eset, hogy a peremérték-problémában Neumann-határfeltét adunk, azaz a megoldásról előírjuk, hogy annak deriváltja legyen nulla a határon. A peremfeltételeket kielégítő megoldások ilyenkor koszinusz sorba fejthetők:

$$h(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cdot \cos\left(x \frac{n\pi}{T}\right) \quad a_n = \frac{2}{T} \int_0^T h(x) \cos\left(x \frac{n\pi}{T}\right) dt \quad (6.44)$$

A megfelelő diszkrét koszinusz transzformáltak

$$A_n = \frac{1}{2} \{h_0 + (-1)^n h_N\} + \sum_{k=1}^{N-1} h_k \cos\left(\frac{\pi nk}{N}\right) \quad (6.45)$$

Éppúgy, mint a szinusz transzformáltaknál megmutathajuk, hogy a párosan kiterjesztett

$$\tilde{h}(x) = \begin{cases} h(x) & , \quad x > 0 \\ h(x) & , \quad x < 0 \end{cases} \quad (6.46)$$

függvényből vett  $\{\tilde{h}_k \mid k = 0, 1, \dots, 2N-1\}$  minta diszkrét  $\{\tilde{H}_n \mid n = 0, 1, \dots, 2N-1\}$  Fourier-transzformáltjából

$$\tilde{H}_n = 2A_n \quad (6.47)$$

A koszinusz transzformáció megint (faktortól eltekintve) önmaga inverze

$$h_k = \frac{2}{N} \left[ \frac{1}{2} \{A_0 + (-1)^n A_N\} + \sum_{n=1}^{N-1} A_n \cos\left(\frac{\pi nk}{N}\right) \right] \quad (6.48)$$

## 6.7. Többdimenziós FFT

Elsőként a fizikában igen gyakori 3-dimenziós Fourier-transzformálttal foglalkozunk. Egy  $f(\mathbf{r}): \mathbb{R}^3 \rightarrow \mathbb{C}$  függvény Fourier-transzformáltja a konvencionális definíciókkal

$$F(\mathbf{q}) = \int f(\mathbf{r}) e^{-i\mathbf{q}\mathbf{r}} d^3\mathbf{r} \quad \Longleftrightarrow \quad f(\mathbf{r}) = \frac{1}{(2\pi)^3} \int F(\mathbf{q}) e^{i\mathbf{q}\mathbf{r}} d^3\mathbf{q} \quad (6.49)$$

Ha a transzformálandó függvénynek valamilyen szimmetriája van, azt a transzformáció során célszerű kihasználni. A legfontosabb eset, ha a függvény gömbszimmetrikus, azaz  $f(\mathbf{r}) \equiv f(r)$ ,  $r = |\mathbf{r}|$ . Ekkor ugyanis

$$\int f(\mathbf{r}) e^{-i\mathbf{q}\mathbf{r}} d^3\mathbf{r} = 2\pi \int_0^\infty r^2 dr \int_{-1}^1 du f(r) e^{-iqr u} = 2\pi \int_0^\infty r^2 dr f(r) \frac{2 \sin(qr)}{qr} = \frac{4\pi}{q} \int_0^\infty r \cdot f(r) \cdot \sin(qr) dr \quad (6.50)$$

és így a 3-dimenziós Fourier-transzformált helyett a  $g(r) = r \cdot f(r)$  függvény 1-dimenziós szinuszos Fourier-transzformáltját elég kiszámolni. Ha hasonló elvi megfontolásokkal nem sikerül a feladatot alacsonyabb dimenzióra redukálni, akkor kénytelenek leszünk 'gyalog' megoldással élni. Tekintsük például a kétdimenziós

$$H_{n,m} = \sum_{k=0}^{Nk-1} \sum_{l=0}^{Nl-1} h_{k,l} e^{i2\pi kn} e^{i2\pi lm} \quad (6.51)$$

diszkrét transzformáltat, amit a következő módon tudunk kiszámolni. Előbb az egyik dimenzióban, esetünkben minden rögzített  $k$  index mellett  $Nk$ -alkalommal elvégzett DFT-vel kiszámoljuk az összes ideiglenes

$$^{(k)}H_m = \sum_{l=0}^{Nl-1} h_{k,l} e^{i2\pi lm} \quad k = 0, 1, \dots, Nk - 1 \quad (6.52)$$

diszkrét transzformáltat. Az így kapott tömböket átrendezzük

$$^{(k)}H_m \rightarrow ^{(m)}H_k \quad (6.53)$$

alakra, majd  $Nl$  alkalommal használjuk ezekre újból az egydimenziós

$$H_{n,m} = \sum_{k=0}^{Nk-1} ^{(m)}H_k e^{i2\pi kn} \quad l = 0, 1, \dots, Nl - 1 \quad (6.54)$$

transzformációt. Összesen tehát  $Nk \times Nl$  alkalommal kellett az FFT rutint meghívni, jó lesz ügyelni arra, hogy az FFT optimálisan legyen kihasználva. Némi programozási gyakorlattal azt is láthatjuk, hogy a két egydimenziós transzformáció közötti átrendezés (oszlop-sor csere) nagy mennyiségű komplex számra igen forrásigényes lehet. Célszerű tehát meggyőződni, hogy nincs-e a használt programnyelvhez olyan kész rutin, amelyik a többdimenziós transzformáltakat hatékonyabban számolja.

Magasabb dimenziós Fourier-transzformáltak számításakor különösen fontos, hogy minden speciális körülményt figyelembe vegyünk és ne számoljunk redundáns dolgokat. Ilyen gyakori eset az, amikor valós függvényeket kell transzformálni. A fizikában a Poisson-egyenletre vezető problémák többsége ilyen valós függvényekre felírt differenciálegyenlet. Az igencsak nagyon fontos 2 vagy 3 dimenziós képfeldolgozási problémák is valós számokkal adott információk Fourier analízisét-igénylik. Nagyon nagy számú transzformálandó valós adat esetén ne használjuk az előbbi eljárást, hanem keressünk erre a célra írt (remélhetőleg) effektív rutint. Szükség esetén írjunk magunk, segítséget ehhez a szakirodalomban találunk.

## 6.8. Konvolúciós egyenletek

A  $g(t)$  és  $h(t)$  függvények konvolúcióját a

$$g * h \equiv \int_{-\infty}^{\infty} h(t - \tau)g(\tau)d\tau = h * g \quad (6.55)$$

formulával definiáltuk. A konvolúció kommutatív, a két függvény matematikailag azonos szerepet játszik. A gyakorlati problémákban azonban eredetét és természetét illetően lényeges különbség szokott lenni a konvolúciós partnerek között. Míg az egyik függvény valamilyen jel folyam, többnyire végtelen adathalmaz, addig a másik egy többnyire véges tartójú válasz függvény. Ebben az összefüggésben a válaszfüggvénynek szemléletes jelentést adhatunk. Ha a bemenő jel egyetlen tú-impulzus, amit egy Dirac-delta általánosított függvénnyel írhatunk le, akkor a konvolúció eredményeként

$$g * \delta = g \quad (6.56)$$

alapján magát a válaszfüggvényt kapjuk. Másképp mondva: a válaszfüggvény az a mért hatás, amit a készülékbe bemenő rövid impulzus generál.

A két különböző jelentésű függvényt célszerű jelölésben is megkülönböztetni, így a továbbiakban  $s(t)$  jelről (signal) és  $r(t)$  válaszfüggvényről (response) értekezünk. Tegyük fel, hogy a válaszfüggvény véges intervallumon különbözik nullától, azaz valamely  $\{r_k \mid -M/2 < k \leq M/2\}$  mintával megadható. A szokásos elnevezés szerint

$$(s * r)_j = \sum_{k=-M/2+1}^{M/2} s_{j-k} \cdot r_k \quad (6.57)$$

számsor  $s$  és az  $r$  **diszkrét konvolúciója**. Az integrállal adott definícióból láthatjuk, hogy a numerikus diszkrét konvolúcióra

$$(s * r)(t_j) \approx \Delta \cdot (s * r)_j \quad (6.58)$$

A konvolúciós tétel  $s * r \iff S \cdot R$  diszkrét változata a következő: Ha a jel  $N$ -periódikus akkor

$$(s * r)_j = \sum_{k=-N/2+1}^{N/2} s_{j-k} \cdot r_k \iff S_n \cdot R_n \quad (6.59)$$

ahol  $\{s_k \mid 0 \leq k \leq N-1\} \iff \{S_k \mid 0 \leq k \leq N-1\}$  és  $\{r_k \mid 0 \leq k \leq N-1\} \iff \{R_k \mid 0 \leq k \leq N-1\}$ . Értelemszerűen az összegzésben előforduló negatív indexek kezeléséhez használjuk ki, hogy mindegyik számhalmaz indexében periódikus, azaz  $x_{-n} = x_{N-n}$ ,  $x = s, r, S, R$ .

A tétel kellemetlennek tűnő megszorítása az, hogy mindkét mintáról feltételezi, hogy azok periódikusak, ráadásul ugyanazzal a periódussal.

- Ha a vizsgált jel történetesen periódikus és csak az a probléma, hogy a válaszfüggvény tartója rövidebb, mint a jel periódusa, akkor a válasz függvényből vett mintát nullákkal kiegészítjük, azaz

$$r_k = 0 \quad \text{ha} \quad M/2 \leq k \leq N/2 \quad \text{vagy} \quad -N/2 + 1 \leq k \leq -M/2 + 1 \quad (6.60)$$

- Óvatosan kell eljárunk, ha a jel a valóságban nem periódikus. A konvolúciós eljárás a vizsgált jel-intervallumba 'behozza' az intervallumot megelőző és az azt követő tartományból a jel ottani feltételezett értékét. A diszkrét konvolúcióban implicite a jel periódikusan megismételt, így a jelet nullákkal ki kell egészítenünk annyira, hogy a kibővített és periódikus mintában a válaszfüggvény 'hatósugaránál' jobban szeparáltak legyenek a tényleges jel adatai. Ha a véges tartójú válasz olyan, hogy  $r_k = 0$ , ha  $|k| > K$ , akkor a jel mintájának elejére vagy a végére  $K$  darab nullát be kell írunk.

Két függvény konvolúcióját az előbbiek alapján viszonylag veszélytelenül kiszámolhatjuk. A fordított feladat, a **dekonvolúció**. Ilyenkor a konvolúciót ismerjük, és szeretnénk az egyik partner ismeretében a másikat kiszámolni, például  $f * g = h$  konvolúciós egyenletben  $h$  és  $g$  ismeretében meghatározni, hogy mi lehetett az  $f$ . A megoldás *formálisan* egyszerűnek látszik, hiszen

$$\mathcal{F}[h] = \mathcal{F}[f] \cdot \mathcal{F}[g] \implies \mathcal{F}[f] = \mathcal{F}[h] / \mathcal{F}[g] \implies \begin{cases} f = \mathcal{F}^{-1}[\mathcal{F}[h] / \mathcal{F}[g]] \\ f = h * \mathcal{F}^{-1}[1 / \mathcal{F}[g]] \end{cases} \quad (6.61)$$

A probléma ezekkel azonban az, hogy a jelzett műveletek általában nem végezhetők el, vagy megbízhatatlan eredményt adnak. Gondoljunk arra, hogy az  $\mathcal{F}[g]$ -vel való osztás csaknem biztosan problémához vezet amikor a Fourier transzformált kicsi, vagy egyenesen nulla. De, még ha ettől el is tekintünk, a további lépésekben jelzett inverz Fourier transzformáltak létezése, numerikus értékük értelmezése további megfontolásokat igényel.

Két függvény korrelációját korábban a

$$h|g \equiv \int_{-\infty}^{\infty} h(t + \tau)g(\tau)d\tau \quad (6.62)$$

módon definiáltuk. Az  $N$ -periódikus jelekre a **diszkrét korreláció** és a diszkrét Fourier transzformáció képletei:

$$(h|g)_j = \sum_{k=0}^{N-1} h_{j+k} \cdot g_k \iff H_n \cdot G_n^* \quad (6.63)$$

A diszkrét korreláció numerikus számítására a konvolúció kapcsán elmondottak egyszerűen átvihetők.