Számítógépes Fizika *(Computational Physics) :*Numerikus módszerek összefoglaló.

Bartha Ferenc

e-mail: barthaf@physx.u-szeged.hu, http : www.jate.u-szeged.hu/˜barthaf/
Szegedi Tudományegyetem, Elméleti Fizikai Tanszék
(készültségi fok: April 24, 2002)

Folyamatosan bővítve, javítva, ahogy az anyagban haladunk. Kérem hogy az elgépeléseket, helyesírási hibákat, hibás és téves képleteket, kijelentéseket keressétek és segítsetek kijavítani azokat!

Contents

I. KÖZÖNSÉGES DIFFERENCIÁLEGYENLETEK (ODE: ORDINARY DIFFERENTIAL EQUATIONS)

A. Numerikus differenciálás

Keressük $f(x)$ deriváltját az $x=0$ pontban, $f'(0)$-t miközben ismerjük $f(x)$ értékét az $x=0$ pont körüli ekvidisztáns beosztáson

$$f_k = f(x_k)\ ;\ x_k = kh\ ;\ k = 0, \pm 1, \pm 2, ... \tag{1.1}$$

A függvényt $x=0$ körül Taylor sorral közelítjük:

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + ... \implies f_k = \sum_{i=0}^{n} \frac{(kh)^i}{i!} f^{(i)}(0) + O\left(h^{n+1}\right) \tag{1.2}$$

Ebből különböző kifejezéseket származtathatunk $f'(0)$-ra. Így **kétpontos** formula

$$f_0 = f(0)\ \ ;\ \ f_{\pm 1} = f(0) \pm hf'(0) + O\left(h^2\right) \implies f'(0) = \pm \frac{f_{\pm 1} - f_0}{h} + O\left(h\right) \tag{1.3}$$

illetve **hárompontos**

$$f_{\pm 1} = f(0) \pm hf'(0) + \frac{h^2}{2}f''(0) + O\left(h^3\right) \implies f'(0) = \frac{f_{+1} - f_{-1}}{2h} + O\left(h^2\right) \tag{1.4}$$

vagy **5-pontos**

$$f'(0) = \frac{1}{12h}\left[f_{-2} - 8f_{-1} + 8f_{+1} + f_{+2}\right] + O\left(h^4\right) \tag{1.5}$$

Magasabb rendű deriváltakhoz jutunk például

$$f_{\pm 1} = f(0) \pm hf'(0) + \frac{h^2}{2}f''(0) \pm \frac{h^3}{6}f'''(0) + O\left(h^4\right) \tag{1.6}$$

$$f_{+1} + f_{-1} = 2f(0) + \frac{h^2}{2}f''(0) + O\left(h^4\right) \implies f''(0) = \frac{f_{+1} - 2f_0 + f_{-1}}{h^2} + O\left(h^2\right) \tag{1.7}$$

három ponttal. További hasznos formulák

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

(1.8)

B. Numerikus integrálás

Szükség lehet arra, hogy közelítőleg meghatározzuk az

$$\int_a^b f(t)dt \quad , \quad a > b \tag{1.9}$$

határozott integrált, amikor $f(t)$ primitív függvénye nem ismert. (Esetleg nem is létezik , mint pl. $f(t) = \frac{\sin(x)}{x}$, vagy $f(t) = \sqrt{1+x^2}$ esetben. Vagy az integrandus nem is adott analitikus formában, csak pl. tabulálva, stb.)

Felosztjuk az integrációs tartományt részintervallumokra, az egyes részintervallumokon valamilyen egyszerű függvénnyel (pl. polinommal) illesztjük (helyettesítjük) a függvényt és integráljuk az illesztő függvényt. Minél kisebb részintervallumokon illesztünk és integrálunk, annál pontosabb értéket várunk.
Trapéz szabály: Az $[a, b]$ intervallumot n egyenlő részre vágjuk. Ezek hossza $h = \frac{b-a}{n}$. A függvényt az egyes részintervallumkonon első fokú polinommal közelítjük

$$f(t) \approx f_0 + t \cdot f'(0) = f_0 + t \cdot \frac{f_1 - f_0}{h} \tag{1.10}$$

$$\int_0^h f(t)dt \approx h \cdot f_0 + \frac{h^2}{2} \cdot \frac{f_1 - f_0}{h} = \frac{h}{2}\left[f_0 + f_1\right] + O(h^3) \tag{1.11}$$

illetve

$$\int_a^b f(t)dt \approx \frac{h}{2}\left[f_0 + 2f_1 + ... + 2f_{n-1} + f_n\right] \tag{1.12}$$

A közelítés hibája: Ha $f^{(2)}$ (második derivált) folytonos az $[a, b]$-n és létezik olyan M_2 felső korlát, hogy $\left|f^{(2)}\right| \leq M_2$ 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

$$f(t) \approx f_0 + t \cdot f'(0) + \frac{t^2}{2} \cdot f''(0) = f_0 + t \cdot \frac{f_1 - f_{-1}}{2h} + \frac{t^2}{2} \cdot \frac{f_{+1} - 2f_0 + f_{-1}}{h^2} + O\left(h^3\right) \tag{1.13}$$

$$\int_{-h}^h f(t)dt \approx 2h \cdot f_0 + \frac{h^3}{6} \cdot \frac{f_{+1} - 2f_0 + f_{-1}}{h^2} = \frac{h}{3}\left[f_{-1} + 4f_0 + f_1\right] + O(h^5) \tag{1.14}$$

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

$$\int_a^b f(t)dt \approx \frac{h}{3}\left[f_0 + 4f_1 + 2f_2 + ... + 2f_{n-2} + 4f_{n-1} + f_n\right] \tag{1.15}$$

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

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

Simpson 3/8 szabály: harmadfokú Taylor polinommal és megfelelő numerikus deriváltakkal

$$\int_{x_0}^{x_3} f(t)dt \approx \frac{3h}{8}\left[f_0 + 3f_1 + 3f_2 + f_3\right] + O(h^5) \tag{1.16}$$

Bode szabály: negyedfokú polinommal

$$\int_{x_0}^{x_4} f(t)dt \approx \frac{2h}{45}\left[7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4\right] + O(h^7) \tag{1.17}$$

C. ODE redukálása egyenletrendszerre

A közönséges N-edrendű differenciálegyenletek elsőrendű differenciélegyenlet rendszerré alakíthatók.
Példa: A másodrendű

$$\frac{d^2y}{dx^2} + q(x)\frac{dy}{dx} = r(x) \tag{1.18}$$

egyenlet elsőrendű egyenletek rendszerévé átírható, mint

$$\frac{dy}{dx} = z(x) \quad \text{és} \quad \frac{dz}{dx} = r(x) - q(x)z(x) \tag{1.19}$$

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 őrült dolgokat csinálnak, akkor célszerű a dolognak utánajárni és más változókat választani.

Az általános feladat tehát N csatolt elsőrendű differenciálegyenlet,

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x, \mathbf{y}) \tag{1.20}$$

megoldása. Itt a 'vektor' jelöléseben $\mathbf{y}(x) = \{y_1(x), ..., y_N(x)\}$ az ismeretlenek, míg $\mathbf{f} = \{f_1, ..., f_N\}$ adott függvények $f_i = f_i(x, y_1(x), ..., y_N(x))$.

Egy jól határozott probléma nemcsak a 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 legegyszerűbb esetben megkövetelhetjük, hogy egy változó adott pontban adott értékrt vegyen fel, $y_i(x_a) = c_{i,a}$, de igen bonyolultak is lehetnek, mint például nemlineáris egyenlet rendszer a változók valamely pontban felvett értéke között.

A továbbiakban egyetlen (skalár)

$$\frac{dy}{dx} = f(x, y) \tag{1.21}$$

differenciálegyenlet megoldásával foglalkozunk. A csatolt egyenletrendszer hasonlón oldható meg természetes mátrix-algebra segítségével.

Az $y(x = x_0) = y_0$ típusú kezdőfeltételt fogjuk használni. Az egyszerűség kedvéért vizsgáljuk a feladatot az $x \in [0, 1]$ intervallumon, a kezdőfeltétel legyen $y(0) = y_0$. Osszuk be az intervallumot N egyenlő részere, legyen $h = 1/N$, $x_n = nh$ és $y_n = y(x_n)$; $n = 0, 1, ..., N$.

D. Euler módszer

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

$$\left.\frac{dy}{dx}\right|_{x=x_n} \approx \frac{y_{n+1} - y_n}{h} = f(x_n, y_n) \tag{1.22}$$

kapjuk a rekurzióra alkalmas

$$y_{n+1} = y_n + hf(x_n, y_n) + O(h^2) \tag{1.23}$$

egyenletet. Az aszimmetrikus formula h intervallumot léptet előre x_n-ről $x_{n+1} = x_n + h$-ra. , miközben csak az intervallum kezdetén levő derivált információt használja fel. A lokális hiba csak egy nagyságrenddel kisebb a $dy = y_{n+1} - y_n \sim O(h)$ korrekciónál.

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

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2}y''_n + O(h^3) \tag{1.24}$$

Behelyettesítve a

$$y' = \frac{dy}{dx} = f \quad \text{és} \quad y'' = \frac{df}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = \frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y} \tag{1.25}$$

deriváltakat

$$y_{n+1} = y_n + hf_n + \frac{h^2}{2}\left[\frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}\right]_n + O(h^3) \tag{1.26}$$

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

E. Pontoság

A közelítő eljárásaink csonkolási lépéshibái a különböző sorfejtésekből tipikusan $O(h^{m+1})$ vislkedésűek. A globális hiba, miközben $N \sim O(1/h)$ lépést tettünk $NO(h^{m+1}) \sim O(h^m)$ lesz. A gyakorlati használat során valójában a pontatlanságnak csak egyik oka származik a csonkolásból. A numerikus hibák másik forrása a számítógépek pontatlan aritmetikája. A véges számábrázolásból eredően az elemi aritmetikai műveletek eredménye egy - a géptől és a programnyelvtől függő - tipikus hibával rendelkezik. Ha egy művelet 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

$$\epsilon = O(\varepsilon/h) + O(h^m) \tag{1.27}$$

ami akkor a legkisebb, ha

$$h \sim \varepsilon^{\frac{1}{m+1}} \quad \epsilon \sim \varepsilon^{\frac{m}{m+1}} \tag{1.28}$$

Ennél nagyobb lépésköznél a sorfejtés hibája, kisebbnél a számítógép hibája dominál. Tipikus értékek: 8 byte-on ábrázolt lebegőpontos számoknál $\varepsilon \sim 10^{-16}$ míg 4 byte egyszeres pontosságnál ez csak $\varepsilon \sim 10^{-7}$ Tájékozódásul egy táblázat $\varepsilon \sim 10^{-16}$ gépi adattal

m	h	ϵ
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	$2.2{\cdot}10^{-3}$	$4.6{\cdot}10^{-14}$

(1.29)

F. Stabilitás

Az Euler módszeren bemutatva.....

A numerikus eljárás során y_n értékét több okból is hibával kapjuk meg. Fontos kérdés, hogy valamely korábbi pontban begyüjtött hiba hogyan 'fejlődik' a további lépések során. Tegyük fel, hogy y_n helyett $y_n + \delta y_n$ hibával terhelt értékről kívánunk továbblépni az

$$y_{n+1} + \delta y_{n+1} = y_n + \delta y_n + h\left[f(x_n, y_n) + \left.\frac{\partial f}{\partial y}\right|_n \delta y_n\right] \tag{1.30}$$

Euler módszerrel. A hiba ebből

$$\delta y_{n+1} = \left[1 + h \left.\frac{\partial f}{\partial y}\right|_n\right] \delta y_n \tag{1.31}$$

ami csökkenő görgetett hibát jelent, ha

$$\left|1 + h\frac{\partial f}{\partial y}\right|^2 < 1 \tag{1.32}$$

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

Példa: $\alpha > 0$

$$\frac{dy}{dx} = -\alpha y \;\; \frac{\partial f}{\partial y} = -\alpha \quad \text{stabil} \quad h < \frac{2}{\alpha} \tag{1.33}$$

$$\frac{dy}{dx} = +\alpha y \;\; ; \;\; \frac{dy}{dx} = i\alpha y \quad \text{instabil} \tag{1.34}$$

G. Többlépéses módszerek

Az Euler módszerrel az n-edik pontbeli értékekből léptünk az x_{n+1}-be. Több korábbi pont figyelembel vételével további rekurziós formulát készíthetünk. Integráljuk formálisan a differenciálegyenletet

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y)dx \tag{1.35}$$

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

$$f \approx \frac{x - x_{n-1}}{h} f_n - \frac{x - x_n}{h} f_{n-1} + O(h^2) \tag{1.36}$$

Beírva és integrálva kapjuk az **Adam-Bashforth két lépéses** (prediktor) módszert

$$y_{n+1} = y_n + h\left[\frac{3}{2}f_n - \frac{1}{2}f_{n-1}\right] + O(h^3) \tag{1.37}$$

Több lépéses formulákat kaphatunk, ha magasabbrendű extrapolációt használunk. Ilyen pl. az **Adam-Bashforth négy lépéses** (prediktor) módszer, amikor harmadfokú polinómmal illesztjük a függvény f_n, f_{n-1}, f_{n-2} és f_{n-3} értékét. Az eredmény

$$y_{n+1} = y_n + \frac{h}{24}\left[55f_n - 59f_{n-1} - 37f_{n-2} - 9f_{n-3}\right] + O(h^4) \tag{1.38}$$

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

H. Implicit módszerek

Eleddig y_{n+1} értékét explicite kifejeztük. Az implicit módszerekben y_{n+1} értékére egy egyenletet állítunk fel, ezt az egyenletet még külön meg kell oldani y_{n+1} explicit előállításához.

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

$$\left.\frac{dy}{dx}\right|_{x_{n+\frac{1}{2}}} = f(x_{n+\frac{1}{2}}\ ,\ y_{n+\frac{1}{2}}) \tag{1.39}$$

Használjunk szimmetrikus formulát az $x_{n+\frac{1}{2}}$ pontban a deriváltra a jobboldalon, a baloldalon pedig helyettesítsük $f_{n+\frac{1}{2}}$ értékét az osztópontbeli értékek számtani közepével:

$$\left.\frac{dy}{dx}\right|_{x_{n+\frac{1}{2}}} \approx \frac{y_{n+1}-y_n}{h} + O(h^2) = \frac{f_n+f_{n+1}}{2} + O(h^2) \tag{1.40}$$

Mindkét oldalon közelítés hibája $O(h^2)$. A **Heun** rekurziós összefüggés tehát

$$y_{n+1} = y_n + \frac{h}{2}\left[f(x_n\ ,\ y_n) + f(x_{n+1}\ ,\ y_{n+1})\right] + O(h^3) \tag{1.41}$$

ami 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

$$y_{n+1} = y_n + \frac{h}{12}\left[5f_{n+1} + 8f_n - f_{n-1}\right] + O(h^4) \tag{1.42}$$

kétlépéses formulát. Harmadfokú illesztéssel a három lépéses formula

$$y_{n+1} = y_n + \frac{h}{24}\left[9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\right] + O(h^5) \tag{1.43}$$

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.

I. Runge-Kutta módszer

Az implicit módszer mintájára írjuk fel, hogy

$$\left.\frac{dy}{dx}\right|_{x_{n+\frac{1}{2}}} \approx \frac{y_{n+1}-y_n}{h} + O(h^2) = f(x_{n+\frac{1}{2}}\ ,\ y_{n+\frac{1}{2}}) \tag{1.44}$$

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

$$y_{n+\frac{1}{2}} \approx y_n + \frac{h}{2}f(x_n, y_n) + O(h^2) \implies f(x_{n+\frac{1}{2}}\ ,\ y_{n+\frac{1}{2}}) \approx f(x_{n+\frac{1}{2}}\ ,\ y_n + \frac{h}{2}f(x_n, y_n)) + O(h^2) \tag{1.45}$$

és így jutottunk el a **másodrendű RK** eljáráshoz:

$$k_1 = hf(x_n, y_n) \quad , \quad k_2 = hf(x_n + \frac{h}{2}\ ,\ y_n + \frac{k_1}{2}) \tag{1.46}$$

$$y_{n+1} = y_n + k_2 + O(h^3) \tag{1.47}$$

Hasonlóan magasabb rendű formulák készíthetők. Példaként a **negyedrendű RK** megoldás formulái:

$$k_1 = hf(x_n, y_n) \quad , \quad k_2 = hf(x_n + \frac{h}{2}\ ,\ y_n + \frac{k_1}{2}) \tag{1.48}$$

$$k_3 = hf(x_n + \frac{h}{2}\ ,\ y_n + \frac{k_2}{2}) \quad , \quad k_4 = hf(x_n + h\ ,\ y_n + k_3) \tag{1.49}$$

$$y_{n+1} = y_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} + O(h^5) \tag{1.50}$$

J. Kiegészítések

1. Gauss-kvadratúra

A korábbiakban megismert eljárások 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

$$I = \int_{-1}^{1} f(x)dx \tag{1.51}$$

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

$$I \approx \sum_{i=1}^{N} w_i f(x_i) \quad , \quad x_i = -1 + 2 \cdot \frac{i-1}{N-1} \tag{1.52}$$

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

$$\int_{-1}^{1} x^p dx = \sum_{i=1}^{N} w_i x_i^p \quad , \quad p = 0, 1, ..., N-1 \tag{1.53}$$

egyenletrendszer w_i-re 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

$$I = \sum_{i=1}^{N} w_i f(x_i) \tag{1.54}$$

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

$$\int_{-1}^{1} P_i(x) P_j(x) dx = \frac{2}{2i+1} \delta_{ij} \tag{1.55}$$

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

$$p_{2N-1}(x) = p_{N-1}(x) \cdot P_N(x) + q_{N-1}(x) \tag{1.56}$$

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

$$\int_{-1}^{1} f(x)dx = \int_{-1}^{1} \{p_{N-1}(x) \cdot P_N(x) + q_{N-1}(x)\}\, dx = \int_{-1}^{1} q_{N-1}(x)dx \tag{1.57}$$

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

$$\int_{-1}^{1} f(x)dx = \sum_{i=1}^{N} w_i q_{N-1}(x_i) = \sum_{i=1}^{N} w_i f(x_i) \tag{1.58}$$

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

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

$$w_i = \frac{2}{(1-x_i^2)\left[P_N'(x)\right]^2} \tag{1.59}$$

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

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

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

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{i=1}^{N} w_i f(x_i) \quad , \quad H_N(x_i) = 0 \quad \text{ahol} \quad H_N(x) \quad \text{Hermite polinom} \tag{1.61}$$

2. Adaptív lépések a differenciálegyenletek integrálásakor

Eddigi módszereinkben a megoldást a teljes megcélzott intervallumon egyetlen előre kiválasztott 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 a függvény gyorsan változik, finomabb beosztást kellene 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 munkaigényes feladat.

Univerzálisabb megoldás, ha a lépésközt valamely tervezett hibához igazítjuk. Tegyük fel, hogy a lépésenkénti hibát szeretnénk valamilyen optimális ϵ_{opt} értéken tartani. Ha sikerülne az egyes lépések során kis többletmunka árán becslést kapni arra, hogy mekkora a lépés ϵ_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

$$y(x+2h) = y_a + 2\,(h)^5 \cdot \Phi + O(h^6) \tag{1.62}$$

$$y(x+2h) = y_b + (2h)^5 \cdot \Phi + O(h^6) \tag{1.63}$$

ahol $\Phi \approx y^{(5)}(x)/5!$ nagyjából ugyannak vehető a két egyenletben. A két hiba vezető rendben tehát

$$\epsilon_a \approx 2h^5 \cdot \Phi \approx 2\epsilon_h \quad \text{és} \quad \epsilon_b \approx 32h^5 \cdot \Phi \approx 32\epsilon_h \tag{1.64}$$

Ezekből az aktuális lépésköz melletti tipikus hiba hatodrendben előáll:

$$\epsilon_h \approx \frac{1}{30} |\epsilon_b - \epsilon_a| \approx \frac{1}{30} |y_a - y_b| + O(h^6) \tag{1.65}$$

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} lépésközre érdemes áttérni, hiszen ötödrendű hibatagról lévén szó:

$$\left(\frac{h_{opt}}{h}\right)^5 = \frac{\epsilon_{opt}}{\epsilon_h} \qquad h_{opt} = h\left(\frac{\epsilon_{opt}}{\epsilon_h}\right)^{0.2} \tag{1.66}$$

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

$$k_1 = hf(x_n, y_n) \tag{1.67}$$
$$k_2 = hf(x_n + a_2 h\ ,\ y_n + b_{21}k_1) \tag{1.68}$$
$$\vdots \tag{1.69}$$
$$k_M = hf(x_n + a_M h\ ,\ y_n + b_{M1}k_1 + ... + b_{M,M-1}k_{M-1}) \tag{1.70}$$
$$y_{n+1} = y_n + \sum_{i=1}^{M} c_i k_i + O(h^{M+1}) \tag{1.71}$$

Ha $M > 4$, akkor lehetőség van arra, hogy másképp is elvégezzük a lépést ugyanazokkal a k_i számokkal de más együtthatókkal úgy, hogy

$$\tilde{y}_{n+1} = y_n + \sum_{i=1}^{M} \tilde{c}_i k_i + O(h^M) \tag{1.72}$$

legyen. Megfelelő együtthatókat (táblázatosan adott) választva a lépés tipikus hibája megbecsülhető:

$$\epsilon_h = |y_{n+1} - \tilde{y}_{n+1}| \tag{1.73}$$

Az egyik leggyakoribb ilyen beágyazott RK eljárás a Fehlberg 4-5 rendű Runge-Kutta módszer, ahol az ötödrendű RK mellett ugyanazon függvényértékekkel egy negyedredű lépést is teszünk és ebből becsüljük a hibát. Ilyen a MAPLE **rkf45** rutinja.

II. PARCIÁLIS DIFFERENCIÁLEGYENLETEK

A. A lineáris egyenletek osztályozása

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

$$\left[a\frac{\partial^2}{\partial x^2}+2b\frac{\partial^2}{\partial x\partial y}+c\frac{\partial^2}{\partial y^2}+d\frac{\partial}{\partial x}+e\frac{\partial}{\partial y}+f\right]V(x,y)=g \tag{2.1}$$

Az a,b,c,d,e,f,g együtthatók maguk is (x,y) adott függvényei lehetnek. Ha az együtthatók függenek az ismeretlen $V(x,y)-$tól 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	$\frac{\partial^2V}{\partial x^2}+\frac{\partial^2V}{\partial y^2}=0$	*peremérték*
$b^2>ac$	hiperbolikus	Hullámegyenlet	$\frac{\partial^2V}{\partial x^2}-\frac{1}{v^2}\frac{\partial^2V}{\partial t^2}=0$	*kezdetiérték*
$b^2=ac$	parabolikus	Diffúzió/(*Schrödinger*)-egyenlet	$\lambda\frac{\partial^2V}{\partial x^2}-\frac{\partial V}{\partial t}=0$	*kezdetiérték*

(2.2)

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 $V(x,t_0)$ kezdőmegoldás 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 $V(x,y)$ előállítása végett. A peremérték-feladatoknál nem lehet egyszerűen 'beintegrálni' a peremről, a megoldást a tér minden pontját egyformán figyelve globális módszerrel kell megközelíteni.

B. Hiperbolikus egyenletek

A (homogén) hullámegyenlet (konstans sebesség mellet) átírható

$$\left[\frac{\partial^2}{\partial t^2}-c^2\frac{\partial^2}{\partial x^2}\right]u(x,t)=0\quad\sim\quad\left[\frac{\partial}{\partial t}+c\frac{\partial}{\partial x}\right]\left[\frac{\partial}{\partial t}-c\frac{\partial}{\partial x}\right]u(x,t)=0 \tag{2.3}$$

alakra, ahonnan látható, hogy a feladat az

$$\left[\frac{\partial}{\partial t}+c\frac{\partial}{\partial x}\right]z(x,t)=0\quad,\quad\left[\frac{\partial}{\partial t}-c\frac{\partial}{\partial x}\right]u(x,t)=z(x,t) \tag{2.4}$$

elsőrendű egyenletek rendszerére redukálható. Másik lehetséges felbontás

$$s(x,t)=c\frac{\partial u}{\partial x}\quad,\quad r(x,t)=\frac{\partial u}{\partial t}\quad,\quad\frac{\partial r}{\partial t}=c\frac{\partial s}{\partial x}\quad,\quad\frac{\partial s}{\partial t}=c\frac{\partial r}{\partial x} \tag{2.5}$$

Az általános hullámegyenlet (inhomogén, nem-konstans együtthatók) tárgyalása helyett a numerikus megoldási módszerek alapötletét és a felmerülő problémákat egy egyszerűbb modellen vizsgáljuk meg. A következőkben a

$$\left[\frac{\partial}{\partial t}+v\frac{\partial}{\partial x}\right]z(x,t)=0\quad,\quad z(x,t_0)=z_0(x) \tag{2.6}$$

kezdetiérték problémát fogjuk tanulmányozni.

A véges differenciákra való áttérés végett ekvidisztáns

$$x_j = x_0 + j\delta x \quad , \quad t_n = t_0 + n\delta t \quad , \quad z_j^n \equiv z(x_j, t_n) \tag{2.7}$$

beosztást veszünk mindkét (x *és* t) tengelyen. Miután a t_n-ről való időben előre történő lépésünkkör az összes $z_1^n, z_2^n, ...$ rendelkezésre áll, az x szerinti deriváltat tetszőlegesen sok pontból felírhatjuk. Az idő szerinti deriváltra azonban csak korábbi időpontokról van információnk, így explicit rekurzióhoz kézenfekvő egyszerű előre deriváltat használni. A kétpontos időbeli első derivált mellett nincs szükség arra, hogy a hárompontos szimmetrikus formulánál pontosabbat használjunk az x változóban. Kompromisszumként legyen a disztkretizált egyenlet tehát

$$\left|\left[\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\right] z(x,t)\right|_{j,n} = \frac{z_j^{n+1} - z_j^n}{\delta t} + O(\delta t) + v_j^n \cdot \frac{z_{j+1}^n - z_{j-1}^n}{2\delta x} + O(\delta x^2) = 0 \tag{2.8}$$

Ebből az FTCS (Forward Time, Centered Space) rekurzió

$$z_j^{n+1} = z_j^n - \frac{1}{2} v_j^n \left(z_{j+1}^n - z_{j-1}^n\right) \frac{\delta t}{\delta x} \tag{2.9}$$

Sajnos ez az egyszerű formula nemcsak pontatlan, de használhatatlanul instabil, nagyon 'rövid' idő alatt a megoldás 'elszáll'. A *von Neumann* stabilitásvizsgálthoz tekintsük a $v(x,t) \equiv c$ speciális hullámegyenlet

$$z(x,t) = e^{ik(x-ct)} \tag{2.10}$$

síkhullám alap-megoldásait. Ekkor

$$z_j^n = w(k)^n \cdot e^{ikx_j} \tag{2.11}$$

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

$$w(k)^{n+1} = \left[1 - i\frac{c\delta t}{\delta x} sin(k\delta x)\right] w(k)^n \quad \Longrightarrow \quad w(k) = \left[1 - i\frac{c\delta t}{\delta x} sin(k\delta x)\right] \implies \|w(k)\| > 1 \tag{2.12}$$

mégis 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 módszer** segítségével. A rekuziónkban a jobboldalon végezzük el a

$$z_j^n \longrightarrow \frac{z_{j+1}^n + z_{j-1}^n}{2} \tag{2.13}$$

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

$$z_j^{n+1} = \frac{z_{j+1}^n + z_{j-1}^n}{2} - \frac{v}{2}\frac{\delta t}{\delta x}\left(z_{j+1}^n - z_{j-1}^n\right) \tag{2.14}$$

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

$$w(k) = \left[cos(k\delta x) - i\frac{c\delta t}{\delta x} sin(k\delta x)\right] \tag{2.15}$$

$$\|w(k)\|^2 = cos^2(k\delta x) + \left(c\frac{\delta t}{\delta x}\right)^2 sin^2(k\delta x) = 1 - \left[1 - \left(c\frac{\delta t}{\delta x}\right)^2\right] sin^2(k\delta x) \tag{2.16}$$

Minden k-ra stabil az eljárásunk, ha

$$|v|\frac{\delta t}{\delta x} \leq 1 \quad \text{vagy} \quad |v|\,\delta t \leq \delta x \tag{2.17}$$

teljesül. Ez a *Courant-Friedrichs-Lewy* stabilitási feltétel. Fizikai magyarázat: a hullámnak $(x_{j\pm1}, t_n)$-ből oda kell érnie x_j, t_{n+1}-be. Matematikai magyarázat: A Lax formulát kicsit átírva

$$\frac{z_j^{n+1} - z_j^n}{\delta t} = \frac{(\delta x)^2}{2\delta t} \cdot \frac{z_{j+1}^n - 2z_j^n + z_{j-1}^n}{(\delta x)^2} - v\left(\frac{z_{j+1}^n - z_{j-1}^n}{2\delta x}\right) \tag{2.18}$$

észrevehetjük, hogy ez a

$$\left[\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\right] z(x,t) = \lambda \frac{\partial^2}{\partial x^2} z(x,t) \quad , \quad \lambda = \frac{(\delta x)^2}{2\delta t} \tag{2.19}$$

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

$$\lambda = \frac{\delta x}{\delta t}\frac{\delta x}{2} \geq \frac{\delta x}{2\,|v|} \tag{2.20}$$

tehát azt jelenti, hogy elegendően erős *numerikus disszipációt, numerikus viszkozitást* imitáltunk. A súrlódásra szükségünk volt, hogy stabil legyen az eljárás, ugyanakkor a nagy súrlódás 'elemészti' a megoldásainkat. Ha olyan problémákat vizsgálunk, ahol a lényegesnek vélt komponensekre teljesül, hogy $k\delta x \gg 1$, 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 felnövekszenek és uralkodni kezdenek, míg az utóbbiban kihalnak.

Amiatt, hogy az időbeli deriváltat kétpontos formulával írtuk fel, míg a térbelit hárompontossal, a pontosság miatt a δt beosztás finomságára jobban kell ügyelnünk, mint δx-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:

$$\left|\left[\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\right] z(x,t)\right|_{j,n} = \frac{z_j^{n+1} - z_j^{n-1}}{2\delta t} + O(\delta t^2) + v_j^n \cdot \frac{z_{j+1}^n - z_{j-1}^n}{2\delta x} + O(\delta x^2) = 0 \tag{2.21}$$

$$z_j^{n+1} = z_j^{n-1} - \frac{\delta t}{\delta x} v_j^n \left(z_{j+1}^n - z_{j-1}^n\right) \tag{2.22}$$

Ebben a rekurzióban szükség van arra, hogy a korábbi időpillantra kiszámolt z^{n-1} értékeket megint elővegyük. A stabilitásvizsgálatot a korábbiakhoz hasonlóan elvégezve kapjuk, hogy

$$w^2 = 1 - w2i\frac{c\delta t}{\delta x} sin(k\delta x) \tag{2.23}$$

Kis kézimunkával megmutathatjuk, hogy

$$\frac{c\delta t}{\delta x} \leq 1 \quad \Longrightarrow \quad \|w(k)\| = 1 \tag{2.24}$$

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

$$\left[\frac{\partial^2 u}{\partial t^2} - v^2\frac{\partial^2 u}{\partial x^2}\right]_{j,n} = \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\delta t)^2} - v^2\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\delta x)^2} = 0 \tag{2.25}$$

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) tejesen szétcsatoltak az egyenletben. A megoldás a **kétlépéses Lax-Wendroff módszer** Az első lépésben Lax módszerrel csak fele akkor beosztású rácson lépünk a centrális pontba

$$z_{j+1/2}^{n+1/2} = \frac{z_{j+1}^n + z_j^n}{2} - \frac{v}{2}\frac{\delta t}{\delta x}\left(z_{j+1}^n - z_j^n\right) \tag{2.26}$$

majd a centrális értékkel bakugrás következik

$$z_j^{n+1} = z_j^n - \frac{\delta t}{\delta x} v_{j+1/2}^{n+1/2}\left(z_{j+1/2}^{n+1/2} - z_{j-1/2}^{n+1/2}\right) \tag{2.27}$$

Az eljárás stabilitásának feltétele megintcsak a $|v|\,\delta t \leq \delta x$ Courant szabály. Csillapítás itt is előfordul a nagy hullámszámokra, de ez a disszipáció kisebb, mint az eredeti Lax módszernél.

C. Parabolikus egyenletek

A diffúziós egyenlet

$$\frac{\partial V}{\partial t} = \lambda \frac{\partial^2 V}{\partial x^2} \tag{2.28}$$

megoldásával foglalkozunk először, miközben a $\lambda > 0$ diffúziós együtthatót állandónak vesszük. Diszkretizálva a feladatot $V_j^n \equiv V(x_j, t_n)$ jelöléssel az **FTCS** egyenletünk

$$\frac{V_j^{n+1} - V_j^n}{\delta t} = \lambda \frac{V_{j+1}^n - 2V_j^n + V_{j-1}^n}{(\delta x)^2} \tag{2.29}$$

$$V_j^{n+1} = V_j^n + \alpha \left[V_{j+1}^n - 2V_j^n + V_{j-1}^n\right] \qquad , \qquad \alpha = \frac{\lambda \delta t}{(\delta x)^2} \tag{2.30}$$

A stabilitásvizsgálathoz vegyük a diffúziós egyenletnek egy a térben síkhullám alapmegoldását

$$V(x,t) = w(t) \cdot e^{ikx} \quad , \quad w(t) = e^{-\lambda k^2 t} \tag{2.31}$$

és vizsgáljuk meg, hogy a

$$V_j^n = w^n \cdot e^{ikx_j} \tag{2.32}$$

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

$$w = 1 + \alpha \left[e^{ik\delta x} - 2 + e^{-ik\delta x}\right] = 1 + 2\alpha \left[cos(k\delta x) - 1\right] = 1 - 4\alpha sin^2(\frac{k\delta x}{2}) \tag{2.33}$$

amiből

$$4\alpha \; sin^2(\frac{k\delta x}{2}) \leq 2 \tag{2.34}$$

A stabilitás feltétele ezek szerint, hogy

$$\alpha \leq 1/2 \quad \text{azaz} \;\; \delta t \leq \frac{1}{2}\frac{(\delta x)^2}{\lambda} \sim \tau \tag{2.35}$$

ahol τ a δx távolságra való diffúzió karakterisztikus ideje. A feltételes stabilitás ugyan biztató,.mégis a módszer alkalmazása a gyakorlatban nem cálszerű. A reális problémákra jellemző L tipikus hosszúság mellett a diffúziós karakterisztikus idő

$$\tau \sim \frac{L^2}{\lambda} = \frac{L^2}{(\delta x)^2}\frac{(\delta x)^2}{\lambda} \tag{2.36}$$

Miközben értelemszerűen a pontosabb számoláshoz $L/\delta x \gg 1$ térbeli beosztás szükséges, a τ ideig való szimulációhoz $L^2/(\delta x)^2$ nagyságrendben kell időbeli lépést tenni.

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:

$$\frac{V_j^{n+1} - V_j^n}{\delta t} = \lambda \frac{V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}}{(\delta x)^2} \tag{2.37}$$

Így egy **teljesen implicit** egyenletrendszert kapunk:

$$-\alpha V_{j+1}^{n+1} + (1+2\alpha)V_j^{n+1} - \alpha V_{j-1}^{n+1} = V_j^n \quad , \quad j = 1, 2, ..., J-1 \tag{2.38}$$

A stabilitás vizsgálatánál kapjuk, hogy

$$w\left[1 + 2\alpha - 2\alpha cos(k\delta x)\right] = 1 \quad w = \frac{1}{1 + 4\alpha sin^2(\frac{k\delta x}{2})} \tag{2.39}$$

azaz az eljárás minden lépésközzel stabil, $|w| \leq 1$. Az implicit egyeletrendszer vektor jelölésben

$$\mathbf{A} \cdot \mathbf{V}^{n+1} = \mathbf{V}^n \tag{2.40}$$

ahol $\mathbf{A}$ tridiagonális mátrix:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ -\alpha & 1+2\alpha & -\alpha & 0 & \cdot & 0 \\ 0 & -\alpha & 1+2\alpha & -\alpha & 0 & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -\alpha & 1+2\alpha & -\alpha \\ \cdot & \cdot & \cdot & 0 & 0 & 1 \end{pmatrix} \tag{2.41}$$

Az ilyen egyenletet például rekurzióval minden időbeli lépésre meg tudjuk oldani.

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

$$\left.\frac{\partial V}{\partial t}\right|_{j,n+1/2} = \lambda \left.\frac{\partial^2 V}{\partial x^2}\right|_{j,n+1/2} \approx \frac{\lambda}{2}\left[\left.\frac{\partial^2 V}{\partial x^2}\right|_{j,n+1} + \left.\frac{\partial^2 V}{\partial x^2}\right|_{j,n}\right] \tag{2.42}$$

A deriváltakat diszkretizálva kapjuk

$$\frac{V_j^{n+1} - V_j^n}{\delta t} = \frac{\lambda}{2\left(\delta x\right)^2}\left[\left(V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}\right) + \left(V_{j+1}^n - 2V_j^n + V_{j-1}^n\right)\right] \tag{2.43}$$

ami átrendezve a **Crank-Nicholson** formula

$$V_j^{n+1} - \frac{\alpha}{2}\left(V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}\right) = V_j^n + \frac{\alpha}{2}\left(V_{j+1}^n - 2V_j^n + V_{j-1}^n\right) \tag{2.44}$$

$$-\alpha V_{j+1}^{n+1} + (2+2\alpha)V_j^{n+1} - \alpha V_{j-1}^{n+1} = \alpha V_{j+1}^n + (2-2\alpha)V_j^n + \alpha V_{j-1}^n \tag{2.45}$$

vagy mátrix alakban

$$\mathbf{A} \cdot \mathbf{V}^{n+1} = \mathbf{B} \cdot \mathbf{V}^n \tag{2.46}$$

rősítése

$$w = \frac{1-\alpha+\alpha cos(k\delta x)}{1+\alpha-\alpha cos(k\delta x)} = \frac{1-2\alpha sin^2(\frac{k\delta x}{2})}{1+2\alpha sin^2(\frac{k\delta x}{2})} \leq 1 \tag{2.47}$$

tehát a módszer 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

$$2V_j^n \longrightarrow V_j^{n+1} + V_j^{n-1} \tag{2.48}$$

helyettesítéssel élünk. Ekkor a

$$\frac{V_j^{n+1} - V_j^{n-1}}{2\delta t} = \lambda \frac{V_{j+1}^n - V_j^{n+1} - V_j^{n-1} + V_{j-1}^n}{\left(\delta x\right)^2} \tag{2.49}$$

$$V_j^{n+1} = V_j^{n-1} + 2\alpha\left[V_{j+1}^n - V_j^{n+1} - V_j^{n-1} + V_{j-1}^n\right] \tag{2.50}$$

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

$$V_j^{n+1} = \left(\frac{1-2\alpha}{1+2\alpha}\right)V_j^{n-1} + \left(\frac{2\alpha}{1+2\alpha}\right)\left[V_{j+1}^n + V_{j-1}^n\right] \tag{2.51}$$

A stabilitásvizsgálakor másodfokú egyenletet kapunk w-re

$$w^2 = \left(\frac{1-2\alpha}{1+2\alpha}\right) + 2w\left(\frac{2\alpha}{1+2\alpha}\right) cos(k\delta x \tag{2.52}$$

amiből

$$w = \frac{1}{1+2\alpha}\left(2\alpha cos(k\delta x) \pm \sqrt{1-4\alpha^2 sin^2(k\delta x)}\right) \tag{2.53}$$

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

$$-(2\alpha+1) \leq cos(k\delta x) \pm \sqrt{1-4\alpha^2 sin^2(k\delta x)} \leq (2\alpha+1) \quad \text{miatt} \quad |w| \leq 1 \tag{2.54}$$

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

$$\|w\|^2 = \frac{1}{(1+2\alpha)^2}\left(cos^2(k\delta x) + 4\alpha^2 sin^2(k\delta x) - 1\right) = \frac{4\alpha^2-1}{(1+2\alpha)^2} = \frac{2\alpha-1}{2\alpha+1} \text{ miatt } \quad \|w\|^2 \leq 1 \tag{2.55}$$

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 térbeli beosztáson, ami a V_j^{n-1} értékek tárolását kívánja meg. Ez jelentős tárolókapcitást igényelhet. Továbbá kellemetlen, hogy kezdőfeltételként is két időpontban is meg kell adnunk a V_j^{n-1} értékeket, ami indokolatlan egy időben elsőredű differenciálegyenletnél.

Példa: **Időfüggő Scrödinger egyenlet 1 dimenzióban**

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

$$i\frac{\partial\psi}{\partial t} = -\frac{\partial^2\psi}{\partial x^2} + U(x)\psi \quad , \quad \psi(x, t=t_0) = \psi_0(x) \quad , \quad \psi(\pm\infty, t) = 0 \tag{2.56}$$

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

$$\frac{\psi_j^{n+1} - \psi_j^n}{\delta t} = \frac{i}{2}\left[\frac{\psi_{j+1}^{n+1} - 2\psi_j^{n+1} + \psi_{j-1}^{n+1}}{(\delta x)^2} - U_j(x)\psi_j^{n+1} + \frac{\psi_{j+1}^n - 2\psi_j^n + \psi_{j-1}^n}{(\delta x)^2} - U_j(x)\psi_j^n\right] \tag{2.57}$$

vagy rendezve

$$-a\psi_{j+1}^{n+1} + (1+2a+b_j)\,\psi_j^{n+1} - a\psi_{j-1}^{n+1} = a\psi_{j+1}^n + (1-2a-b_j)\,\psi_j^n + a\psi_{j-1}^n \; ; \quad a = \frac{i\delta t}{2\,(\delta x)^2} \; ; \; b_j = \frac{i\delta t}{2}U(x_j) \tag{2.58}$$

A tridiagonális egyenletrendszer numerikusan hatékonyan invertálható.

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

$$e^{-iH\delta t}\psi(x,t_n) = \psi(x,t_{n+1}) \tag{2.59}$$

Ha ebben az egyenletben az

$$e^{-iH\delta t} \approx 1 - iH\delta t \quad \Longrightarrow \quad \psi_j^{n+1} = (1 - iH\delta t)\,\psi_j^n \tag{2.60}$$

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

$$(1 - iH\delta t)^{-1}\,\psi_j^{n+1} = \psi_j^n \tag{2.61}$$

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

$$e^{-iH\delta t} \approx \frac{1 - \frac{1}{2}iH\delta t}{1 + \frac{1}{2}iH\delta t} \quad \Longrightarrow \quad \left(1 + \frac{1}{2}iH\delta t\right)\psi_j^{n+1} = \left(1 - \frac{1}{2}iH\delta t\right)\psi_j^n \tag{2.62}$$

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

D. Kiegészítés: Tridiagonális egyenletrendszer megoldása rekurzióval

Tekintsük az

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{d} \tag{2.63}$$

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

$$\begin{pmatrix} \beta_1 & \gamma_1 & 0 & \cdot & \cdot & \cdot \\ \alpha_2 & \beta_2 & \gamma_2 & 0 & \cdot & \cdot \\ 0 & \alpha_3 & \beta_3 & \gamma_3 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \alpha_{n-1} & \beta_{n-1} & \gamma_{n-1} \\ \cdot & \cdot & \cdot & 0 & \alpha_n & \beta_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \cdot \\ d_{n-1} \\ d_n \end{pmatrix} \tag{2.64}$$

vagy

$$\alpha_i x_{i-1} + \beta_i x_i + \gamma_i x_{i+1} = d_i \quad , \quad i = 1, ..., n \quad ; \quad \alpha_1 = \gamma_n = 0 \tag{2.65}$$

Vezessük be a g_i és h_i segédváltozókat úgy, hogy

$$x_{i+1} = g_i x_i + h_i \quad , \quad i = 1, ..., n-1 \tag{2.66}$$

legyen. Ezekkel

$$\begin{aligned} d_i &= \alpha_i x_{i-1} + \beta_i x_i + \gamma_i x_{i+1} && (2.67)\\ &= \alpha_i x_{i-1} + \beta_i x_i + \gamma_i \left(g_i x_i + h_i\right) && (2.68)\\ &= \alpha_i x_{i-1} + \left(\beta_i + \gamma_i g_i\right) x_i + \gamma_i h_i && (2.69)\\ &= \alpha_i x_{i-1} + \left(\beta_i + \gamma_i g_i\right) \left(g_{i-1} x_{i-1} + h_{i-1}\right) + \gamma_i h_i && (2.70)\\ &= \left(\alpha_i + g_{i-1} \left(\beta_i + \gamma_i g_i\right)\right) x_{i-1} + \left(\beta_i + \gamma_i g_i\right) h_{i-1} + \gamma_i h_i && (2.71) \end{aligned}$$

ami kielégíthető, ha

$$0 = \alpha_i + g_{i-1} \left(\beta_i + \gamma_i g_i\right) \quad \Longrightarrow \quad g_{i-1} = -\frac{\alpha_i}{\beta_i + g_i} \tag{2.72}$$

$$d_i = \left(\beta_i + \gamma_i g_i\right) h_{i-1} + \gamma_i h_i \quad \Longrightarrow \quad h_{i-1} = \frac{d_i - \gamma_i h_i}{\beta_i + \gamma_i g_i} \tag{2.73}$$

Láthatóan g_{n-1} és h_{n-1} ismeretében 'lelfelé' rekurzív módón előállítható az összes g és h. Ellenőrízhető közvetlen kiírással, hogy az induló értékek:

$$g_{n-1} = -\frac{\alpha_n}{\beta_n} \qquad \text{és} \qquad h_{n-1} = \frac{d_n}{\beta_n}$$

Az eredeti egyenletrendszer első egyenletéből

$$\beta_1 x_1 + \gamma_1 x_2 = d_1 \quad \Longrightarrow \quad \beta_1 x_1 + \gamma_1 \left(g_1 x_1 + h_1\right) = d_1 \tag{2.74}$$

azaz g_1 és h_1 segitségével x_1 kifejezhető:

$$x_1 = \frac{d_1 - \gamma_1 h_1}{\beta_1 + \gamma_1 g_1} \tag{2.75}$$

Erről az értékről most 'felfele' használva a

$$x_{i+1} = g_i x_i + h_i \tag{2.76}$$

rekurziót minden keresett x_i kiszámolható.

E. Kiegészítés: a diffúziós probléma több dimenzióban

A diffúziós egyenlet 2-dimenzióban

$$\frac{\partial V}{\partial t} = \lambda \left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \right) \tag{2.77}$$

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

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

vagy

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

ahol

$$\delta x = \delta y = \Delta \ \ , \ \ \alpha = \frac{\lambda \delta t}{\Delta^2} \ \ , \ \ \mathcal{L}_x V_{j,l} = \left[V_{j+1,l} - 2V_{j,l} + V_{j-1,l}\right] \ \ , \ \ \mathcal{L}_y V_{j,l} = \left[V_{j,l+1} - 2V_{j,l} + V_{j,l-1}\right] \tag{2.78}$$

Az eljárás térben és időben egyaránt másodrendben pontos továbbá feltétel nélkül stabil. Rendezzük a kétindexes mennyiségeket egyetlen vektorba valamely $j, l \longrightarrow s$ megfeleltetéssel. Így az ismeretlen $\vec{v} \sim \left\{ V_{j,l}^{n+1} \right\}$ vektorra az egyenletrendszer

$$\mathbf{A} \cdot \vec{v} = \vec{b} \tag{2.79}$$

alakú. Sajnos, az 1-dimenziós problémával szemben a $\mathbf{A}$ most nem tridiagonális, szerencsére azonban ritka mátrix. Az ilyen típusú egyenletek numerikusan standard módszerekkel (Jacobi, Gauss-Seidel és SOR) megoldhatók, erről külön kiegészítésben szólunk.

Kissé átalakítva a CN egyenletet alternatív rokonszenvesebb 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

$$\begin{aligned} V_{j,l}^{n+1/2} - V_{j,l}^n &= \frac{\alpha}{2} \left(\mathcal{L}_x V_{j,l}^{n+1/2} + \mathcal{L}_y V_{j,l}^n \right) \\ V_{j,l}^{n+1} - V_{j,l}^{n+1/2} &= \frac{\alpha}{2} \left(\mathcal{L}_x V_{j,l}^{n+1/2} + \mathcal{L}_y V_{j,l}^{n+1} \right) \end{aligned} \tag{2.80}$$

Az eljárás pontossága és stabilitása az előző direkt formulával azonos, de az egyenletek megoldása könyebb. 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:

$$\frac{\partial V}{\partial t} = \lambda \mathcal{D} V = \lambda \left(\mathcal{D}_1 V + ... + \mathcal{D}_m V \right) \tag{2.81}$$

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

$$\frac{\partial V}{\partial t} = \lambda \mathcal{D}_i V \tag{2.82}$$

egyenleből viszonylag egyszerű lenne diszkretizált eljárással

$$\vec{v}^{[n+1]} = F_i(\vec{v}^{[n]}, \delta t) \tag{2.83}$$

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

$$\vec{v}^{\left[n+\frac{1}{m}\right]} = F_1(\vec{v}^{[n]}, \frac{\delta t}{m}) \tag{2.84}$$

$$\vec{v}^{\left[n+\frac{2}{m}\right]} = F_2(\vec{v}^{\left[n+\frac{1}{m}\right]}, \frac{\delta t}{m}) \tag{2.85}$$

$$\vdots \tag{2.86}$$

$$\vec{v}^{[n+1]} = F_m(\vec{v}^{\left[n+\frac{m-1}{m}\right]}, \frac{\delta t}{m}) \tag{2.87}$$

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 az $\mathcal{D}_2$-ben stabil lépés, és így tovább. Így az előbbi OS egyenlet a korábbi ADI sémát adja.

F. Kiegészítés: Av=b lineáris egyenletrendszer megoldása

Bontsuk fel az $\mathbf{A}$ mátrixot a következőképpen

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U} \tag{2.88}$$

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

$$\mathbf{D}\cdot\mathbf{v} = -\left[\mathbf{L}+\mathbf{U}\right]\cdot\mathbf{v} + \mathbf{B} = \left(\mathbf{D}-\mathbf{A}\right)\cdot\mathbf{v} + \mathbf{B} \quad \Longrightarrow \; \mathbf{v} = \mathbf{G}\cdot\mathbf{v} + \mathbf{D}^{-1}\mathbf{B}\;,\; \mathbf{G}_J = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A} \tag{2.89}$$

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

$$\mathbf{v}^{(n+1)} = \mathbf{G}_J\cdot\mathbf{v}^{(n)} + \mathbf{D}^{-1}\mathbf{B} \tag{2.90}$$

Az eljárás konvergenciája a $\mathbf{G}_J$ mátrix sajátértékeitől függ. A legnagyobb sajátérték abszolut értékének egynél kisebbnek kell lennie a konvergenciához. Ha $\mathbf{G}_J$ legnagyobb sajátértéke (spektrál sugara) λ_J, akkor asszimptotikusan

$$q_J = \frac{\left|\mathbf{v}^{(n+1)} - \mathbf{v}^{(n)}\right|}{\left|\mathbf{v}^{(n)} - \mathbf{v}\right|} \approx \left|1 - \lambda_J\right| \tag{2.91}$$

Az eljárás nagyon lassan kovergál, ha a sajátértékek között van egyhez közeli.
Gauss-Seidel iteráció: A baloldalra átvihetjük az alsó trianguláris részt, ha az algoritmust úgy szervezzük, hogy $\mathbf{v}_1^{(n+1)}, \mathbf{v}_2^{(n+1)}, \ldots$. sorrendben határozzuk meg a vektor elemeit. Az így kapott iterációs séma

$$\left[\mathbf{L}+\mathbf{D}\right]\cdot\mathbf{v}^{(n+1)} = -\mathbf{U}\cdot\mathbf{v}^{(n)} + \mathbf{B} \tag{2.92}$$

A konvergencia most a

$$\mathbf{G}_{GS} = -\left[\mathbf{L}+\mathbf{D}\right]^{-1}\mathbf{U} \tag{2.93}$$

mátrix sajátértékeitől függ. Megmutatható, hogy

$$\lambda_{GS} = \lambda_J^2 \tag{2.94}$$

így az eljárás az előbbinél gyorsabban konvergál

$$q_{GS} \approx \left|1 - \lambda_J^2\right| \tag{2.95}$$

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

$$\left[\mathbf{L}+\mathbf{D}\right]\cdot\mathbf{v}^{(n+1)} = \left[\mathbf{A}-\mathbf{U}\right]\cdot\mathbf{v}^{(n)} - \left[\mathbf{A}\cdot\mathbf{v}^{(n)} - \mathbf{B}\right] \tag{2.96}$$

majd 'rontsuk el' a jobboldalt egy ω relaxációs paraméterrel az alábbi módon:

$$\left[\mathbf{A}-\mathbf{U}\right]\cdot\mathbf{v}^{(n)} - \omega\left[\mathbf{A}\cdot\mathbf{v}^{(n)} - \mathbf{B}\right] \tag{2.97}$$

Az így kapott

$$\left[\mathbf{L}+\mathbf{D}\right]\cdot\mathbf{v}^{(n+1)} = -\left[\mathbf{U} - (1-\omega)\mathbf{A}\right]\cdot\mathbf{v}^{(n)} + \omega\mathbf{B} \tag{2.98}$$

SOR formula konvergál, ha $0 \leq \omega \leq 2$. Ha $\omega < 1$, akkor alul relaxálásról, 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

$$\omega \mathbf{v}_{GS}^{(n+1)} + (1-\omega)\mathbf{v}^{(n)} \longrightarrow \mathbf{v}_{SOR}^{(n+1)} \tag{2.99}$$

Az eljárás konvergenciáját most a

$$\mathbf{G}_{SOR} = -\left[\mathbf{L}+\mathbf{D}\right]^{-\mathbf{1}}\left[\mathbf{U}-(\mathbf{1}-\omega)\mathbf{A}\right] \tag{2.100}$$

mátrix határozza meg. Az optimális ω

$$\omega_{SOR} = \frac{2}{1+\sqrt{1-\lambda_J^2}} \tag{2.101}$$

amikor is a spektrálsugár

$$\lambda_{SOR} = \left(\frac{\lambda_J}{1+\sqrt{1-\lambda_J^2}}\right)^2 \tag{2.102}$$

és az asszimptotikus konvergencia hányados

$$q_{SOR} \approx |1-\lambda_{SOR}| \tag{2.103}$$

Iteratív javítás: Bármelyik előző után jól jön.

G. 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

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = \rho(x,y) \tag{2.104}$$

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 vislekedését (pl. Dirichlet, vagy Neumann határfeltételeket).

A síktartományban felvett diszkrét pontokon

$$(x_j = x_0 + j\delta x\ ,\ j = 0,1,...,J)\ ;\ (y_l = y_0 + l\delta y\ ,\ l = 0,1,...,L) \tag{2.105}$$

$V(x,y)$-t a gridpontokon felvett

$$u_{j,l} = V(x_j, y_l) \tag{2.106}$$

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

$$\delta x = \delta y = \Delta \tag{2.107}$$

A differencia egyenlet ekkor (a belső pontokban $(j = 1,...,J-1)\quad,\quad (l = 1,...,L-1)$)

$$\frac{u_{j+1,l} - 2u_{j,l} + u_{j-1,l}}{\Delta^2} + \frac{u_{j,l+1} - 2u_{j,l} + u_{j,l-1}}{\Delta^2} = \rho_{j,l} \tag{2.108}$$

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

1. Relaxáció

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

$$\mathbf{T} \cdot \mathbf{u}_1 + \mathbf{I} \cdot \mathbf{u}_2 = \mathbf{r}_1 - \mathbf{u}_0 \tag{2.109}$$

$$\mathbf{I} \cdot \mathbf{u}_{j-1} + \mathbf{T} \cdot \mathbf{u}_j + \mathbf{I} \cdot \mathbf{u}_{j+1} = \mathbf{r}_j \quad j = 2, ..., J-2 \tag{2.110}$$

$$\mathbf{I} \cdot \mathbf{u}_{J-2} + \mathbf{T} \cdot \mathbf{u}_{J-1} = \mathbf{r}_{J-1} - \mathbf{u}_J \tag{2.111}$$

alakba írható, ahol

$$\mathbf{T} = \begin{pmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{pmatrix} \text{ és } \mathbf{I} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix} \; : \; (L-1) \times (L-1) \tag{2.112}$$

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

$$v_r = u_{j,l} \quad , \quad r = (j-1)L + l \tag{2.113}$$

mennyiségek az

$$\mathbf{A} \cdot \vec{v} = \vec{b} \quad \text{ahol} \quad \mathbf{A} = \begin{pmatrix} \mathbf{T} & \mathbf{I} & & & & \\ \mathbf{I} & \mathbf{T} & \mathbf{I} & & & \\ & \mathbf{I} & \mathbf{T} & \mathbf{I} & & \\ & & \cdot & \cdot & \cdot & \\ & & & \mathbf{I} & \mathbf{T} & \mathbf{I} \\ & & & & \mathbf{I} & \mathbf{T} \end{pmatrix} \tag{2.114}$$

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

2. Váltakozó irányú implicit módszer: ADI

Az $\mathcal{D}V = \rho$ peremértékproblémával kapcsolatba hozható a

$$\frac{\partial V}{\partial t} = \mathcal{D}V - \rho \tag{2.115}$$

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

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

Átrendezve

$$[q\mathbf{I} - \mathcal{L}_x]\, \vec{v}^{[n+1/2]} = [q\mathbf{I} + \mathcal{L}_y]\, \vec{v}^{[n]} - \Delta^2 \vec{\rho}$$

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

a bal oldalon álló $[q\mathbf{I} \pm \mathcal{L}_i]$ 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 $\left|\vec{v}^{[n+1]} - \vec{v}^{[n]}\right| < \epsilon$ nem lesz. Az eljárás konvergenciája a képzeletbeli időtengelyen felvett δt lépéshossztól és így $q-$tól függ. Ezen értékek dinamikus (iterációnként különböző) megválasztásával elérhető, hogy lényegesen gyorsabban jussunk célhoz, mint pl. az SOR módszerrel.

3. Ciklikus redukció

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

$$\mathbf{I}\cdot\mathbf{u}_{j-2} + \mathbf{T}\cdot\mathbf{u}_{j-1} + \mathbf{I}\cdot\mathbf{u}_j = \mathbf{r}_{j-1} \tag{2.116}$$
$$\mathbf{I}\cdot\mathbf{u}_{j-1} + \mathbf{T}\cdot\mathbf{u}_j + \mathbf{I}\cdot\mathbf{u}_{j+1} = \mathbf{r}_j \tag{2.117}$$
$$\mathbf{I}\cdot\mathbf{u}_j + \mathbf{T}\cdot\mathbf{u}_{j+1} + \mathbf{I}\cdot\mathbf{u}_{j+2} = \mathbf{r}_{j+1} \tag{2.118}$$

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

$$\mathbf{I}\cdot\mathbf{u}_{j-2} + \mathbf{T}^{(1)}\cdot\mathbf{u}_j + \mathbf{I}\cdot\mathbf{u}_{j+2} = \mathbf{r}_j^{(1)} \tag{2.119}$$

ahol

$$\mathbf{T}^{(1)} = \left[\mathbf{2I} - \mathbf{T}^2\right] \quad \text{és} \quad \mathbf{r}_j^{(1)} = \mathbf{r}_{j-1} - \mathbf{T}\cdot\mathbf{r}_j + \mathbf{r}_{j+1} \tag{2.120}$$

Ezzel a centrális $\mathbf{u}_j$ meghatározható első szomszédai helyett csupán a második szomszédaiból, azaz a meghatározandó.**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

$$\Downarrow \tag{2.121}$$
$$\mathbf{u}_0 + \mathbf{T}^{(p-1)}\cdot\mathbf{u}_{J/4} + \mathbf{u}_{J/2} = \mathbf{r}_{J/4}^{(p-1)} \tag{2.122}$$
$$\mathbf{u}_{J/4} + \mathbf{T}^{(p-1)}\cdot\mathbf{u}_{J/2} + \mathbf{u}_{3J/4} = \mathbf{r}_{J/2}^{(p-1)} \tag{2.123}$$
$$\mathbf{u}_{J/2} + \mathbf{T}^{(p-1)}\cdot\mathbf{u}_{3J/4} + \mathbf{u}_J = \mathbf{r}_{3J/4}^{(p-1)} \tag{2.124}$$
$$\Downarrow \tag{2.125}$$
$$\mathbf{u}_0 + \mathbf{T}^{(p)}\cdot\mathbf{u}_{J/2} + \mathbf{u}_J = \mathbf{r}_{J/2}^{(p)} \tag{2.126}$$

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 az $(p-1)$ 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. Fourier módszer

A Fourier transzformáltakkal fontosságuk miatt külön fejezetben foglalkozunk részletesen. Megelőlegezve most néhány formulát a diszkrét Fourier transzformáció témaköréből, megvizsgáljuk, hogy a konstans együtthatós parciális differenciálegyenleteket hogyan oldhatjuk meg ezek segítségével. Példaként a Poisson egyenletet vesszük elő, ezt véges differenciákkal az (2.108) egyenletben

$$u_{j+1,l} + u_{j-1,l} + u_{j,l+1} + u_{j,l-1} - 4u_{j,l} = \Delta^2\rho_{j,l} \tag{2.127}$$

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

$$\tilde{u}_{m,n} = \sum_{j=0}^{J-1}\sum_{l=0}^{L-1} u_{j,l}\cdot e^{i2\pi jm/J}e^{i2\pi ln/L} \qquad u_{j,l} = \frac{1}{JL}\sum_{m=0}^{J-1}\sum_{n=0}^{L-1}\tilde{u}_{m,n}\cdot e^{-i2\pi jm/J}e^{-i2\pi ln/L} \tag{2.128}$$
$$\tilde{\rho}_{m,n} = \sum_{j=0}^{J-1}\sum_{l=0}^{L-1} \rho_{j,l}\cdot e^{i2\pi jm/J}e^{i2\pi ln/L} \qquad \rho_{j,l} = \frac{1}{JL}\sum_{m=0}^{J-1}\sum_{n=0}^{L-1}\tilde{\rho}_{m,n}\cdot e^{-i2\pi jm/J}e^{-i2\pi ln/L} \tag{2.129}$$

Ezeket a differencia egyenletbe helyettesítve

$$\tilde{u}_{m,n}\left[e^{-i2\pi m/J} + e^{i2\pi m/J} + e^{-i2\pi n/L} + e^{i2\pi n/L} - 4\right] = \Delta^2\tilde{\rho}_{m,n} \tag{2.130}$$

vagy

$$\tilde{u}_{m,n} = \frac{\Delta^2}{2}\frac{\tilde{\rho}_{m,n}}{cos(2\pi m/J) + cos(2\pi n/L) - 2} \tag{2.131}$$

A probléma megoldása ezek után a következő lépésekben történik

- meghatározzuk a $\tilde{\rho}_{m,n}$ Fourier transzformáltakat
- az (2.131) egyenlet alapján kiszámítjuk az $\tilde{u}_{m,n}$ mátrixot
- az inverz formulákkal $u_{j,l}$ (2.128) 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:

$$\tilde{u}_{m,n} = \sum_{j=1}^{J-1}\sum_{l=1}^{L-1} u_{j,l} \cdot sin\left(\frac{\pi jm}{J}\right) sin\left(\frac{\pi ln}{L}\right) \qquad u_{j,l} = \frac{4}{JL}\sum_{m=1}^{J-1}\sum_{n=1}^{L-1} \tilde{u}_{m,n} \cdot sin\left(\frac{\pi jm}{J}\right) sin\left(\frac{\pi ln}{L}\right) \tag{2.132}$$

$$\tilde{\rho}_{m,n} = \sum_{j=1}^{J-1}\sum_{l=1}^{L-1} \rho_{j,l} \cdot sin\left(\frac{\pi jm}{J}\right) sin\left(\frac{\pi ln}{L}\right) \qquad \rho_{j,l} = \frac{4}{JL}\sum_{m=1}^{J-1}\sum_{n=1}^{L-1} \tilde{\rho}_{m,n} \cdot sin\left(\frac{\pi jm}{J}\right) sin\left(\frac{\pi ln}{L}\right) \tag{2.133}$$

Ilyen felírással a peremfeltétel automatikusan kielégül. A transzformált egyenlet

$$\tilde{u}_{m,n} = \frac{\Delta^2}{2}\frac{\tilde{\rho}_{m,n}}{cos(\pi m/J) + cos(\pi n/L) - 2} \tag{2.134}$$

aminek a megoldási folyamata az előzőével azonos.
Neumann peremfeltételnél $\nabla u = 0$ a peremen. Ez esetben a koszinusz transzformáltat használjuk

$$\tilde{u}_{m,n} = \sum_{j=0}^{J(!)}\sum_{l=0}^{L(!)} u_{j,l} \cdot cos\left(\frac{\pi jm}{J}\right) cos\left(\frac{\pi ln}{L}\right) \qquad u_{j,l} = \frac{4}{JL}\sum_{m=0}^{J(!)}\sum_{n=0}^{L(!)} \tilde{u}_{m,n} \cdot cos\left(\frac{\pi jm}{J}\right) cos\left(\frac{\pi ln}{L}\right) \tag{2.135}$$

$$\tilde{\rho}_{m,n} = \sum_{j=0}^{J(!)}\sum_{l=0}^{L(!)} \rho_{j,l} \cdot cos\left(\frac{\pi jm}{J}\right) cos\left(\frac{\pi ln}{L}\right) \qquad \rho_{j,l} = \frac{4}{JL}\sum_{m=0}^{J(!)}\sum_{n=0}^{L(!)} \tilde{\rho}_{m,n} \cdot cos\left(\frac{\pi jm}{J}\right) cos\left(\frac{\pi ln}{L}\right) \tag{2.136}$$

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

III. FOURIER TRANSZFORMÁCIÓ

A. A Fourier transzformációról általában

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

$$H(\omega) = \int_{-\infty}^{\infty} h(t)e^{-i\omega t}dt \quad \text{vagy} \quad H(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(t)e^{-i\omega t}dt \tag{3.1}$$

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

$$h(t) \simeq \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega)e^{i\omega t}d\omega \quad \text{vagy} \quad h(t) \simeq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H(\omega)e^{i\omega t}d\omega \tag{3.2}$$

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

$$f = \frac{\omega}{2\pi} \tag{3.3}$$

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

$$H(f) = \int_{-\infty}^{\infty} h(t)e^{i2\pi ft}dt \quad \text{és} \quad h(t) = \int_{-\infty}^{\infty} H(f)e^{-i2\pi ft}df \tag{3.4}$$

Emlékeztetőül néhány összefüggés a Fourier párokra

$$\begin{array}{rcl@{\qquad\qquad}rcl} h(at) & \Longleftrightarrow & \frac{1}{|a|}H(\frac{f}{a}) & \frac{1}{|a|}h(\frac{t}{a}) & \Longleftrightarrow & H(af) \\ h(t-a) & \Longleftrightarrow & H(f)e^{i2\pi fa} & h(t)e^{-i2\pi ta} & \Longleftrightarrow & H(f-a) \\ h^{(n)}(t) & \Longleftrightarrow & (-i2\pi f)^n H(f) & (i2\pi t)^n h(t) & \Longleftrightarrow & H^{(n)}(f) \end{array} \tag{3.5}$$

További hasznos fogalmak, tételek:

- Konvolúciós tétel

$$g * h \equiv \int_{-\infty}^{\infty} h(t-\tau)g(\tau)d\tau = h * g \qquad g * h \iff G \cdot H \tag{3.6}$$

- Korreláció

$$h|g \equiv \int_{-\infty}^{\infty} h(t+\tau)g(\tau)d\tau \qquad h|g \iff H \cdot G^* \quad (^* \text{ komplex konjugált}) \tag{3.7}$$

- Autokorreláció

$$h|h \equiv \int_{-\infty}^{\infty} h(t+\tau)h(\tau)d\tau \qquad h|h \iff |H|^2 \tag{3.8}$$

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

B. Mintavételezés

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

$$h_n = h(n\Delta) \qquad n = ..., -2, -1, 0, 1, 2, ... \tag{3.9}$$

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

$$f_c = \frac{1}{2\Delta} \tag{3.10}$$

A kritikus frekvencia jelentősége megmutatkozik az alábbi tételben:
Mintavételezési tétel: Ha a $h(t)$ folytonos függvény *sávhatárolt*, azaz létezik olyan f_M, hogy

$$H(f) = 0 \quad \text{ha} \quad |f| > f_M \tag{3.11}$$

akkor $h(t)$ teljesen megadható a

$$h_n = h(\frac{n}{2f_M}) \tag{3.12}$$

mintavételezéssel és ekkor

$$h(t) = \Delta \sum_{n=-\infty}^{\infty} h_n \frac{sin\left[2\pi f_M \left(n\Delta - t\right)\right]}{\pi\left(n\Delta - t\right)} \tag{3.13}$$

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 $P(f)$ tejesítménysűrűség 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 praktikusan 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.

C. 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

$$h_k = h(k\Delta) \quad , \quad k = 0, 1, ..., N-1 \tag{3.14}$$

beosztást véve a Fourier transzformáltat az integrál közelítésével kiszámolhatjuk

$$H(f) = \int_0^T h(t) e^{i2\pi tf} dt \approx \Delta \sum_{k=0}^{N-1} h_k e^{i2\pi fk\Delta} \tag{3.15}$$

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

$$H(f_n) = H(\frac{n}{N\Delta}) \approx \Delta \sum_{k=0}^{N-1} h_k e^{i2\pi kn/N} \quad , \quad f_n = \frac{n}{N\Delta} = \frac{n}{T} \quad , \quad n = -\frac{N}{2}, ..., +\frac{N}{2} \tag{3.16}$$

függvényértékeket. Az itt szereplő

$$H_n \equiv \sum_{k=0}^{N-1} h_k e^{i2\pi kn/N} \tag{3.17}$$

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

$$H(f_n) \approx \Delta H_n \tag{3.18}$$

A diszkrét Fourier transzformáltban a negatív indexek kényelmetlenné válhatnak. Látható a definícióból, hogy az indexhatárokat tetszőlegesre 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 $n = 0, 1, ..., N-1$ 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**

$$h_k \equiv \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{-i2\pi kn/N} \tag{3.19}$$

hasonló 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

$$h_k \equiv \frac{1}{N} \sum_{n=0}^{N-1} \sum_{l=0}^{N-1} h_l e^{i2\pi ln/N} e^{-i2\pi kn/N} = \sum_{l=0}^{N-1} h_l \sum_{n=0}^{N-1} \left(e^{i2\pi (l-k)/N} \right)^n \tag{3.20}$$

és vegyük észre, hogy

$$\sum_{n=0}^{N-1} \left(e^{i2\pi (l-k)/N} \right)^n = N \quad , \quad \text{ha} \quad l = k \tag{3.21}$$

egyébként pedig

$$\sum_{n=0}^{N-1} \left(e^{i2\pi (l-k)/N} \right)^n = \frac{1 - \left(e^{i2\pi (l-k)/N} \right)^N}{1 - e^{i2\pi (l-k)/N}} = \frac{1-1}{1 - e^{i2\pi (l-k)/N}} = 0 \tag{3.22}$$

D. Gyors Fourier Transzformáció (FFT)

A diszkrét Fourier transzformáltak

$$H_n = \sum_{k=0}^{N-1} h_k w_{n,k} \quad \text{ahol} \quad w_{n,k} = e^{i2\pi kn/N} \tag{3.23}$$

kiszámítása egy komplex mátrixszorzás munkigényének felel meg. NxN-es mátrixokra ez $O(N^2)$ művelet. A gyors Fourier transzformáció (**FFT**: *Fast Fourier Transform)* speciális estekben 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

$$N = 2^p \tag{3.24}$$

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 Lánczos nevével jegyzett alábbi lemma mutatja.

$$\sum_{k=0}^{N-1} h_k e^{i2\pi kn/N} = \sum_{k=0}^{N/2-1} h_{2k} e^{i2\pi(2k)n/N} + \sum_{k=0}^{N/2-1} h_{2k+1} e^{i2\pi(2k+1)n/N} \tag{3.25}$$

$$= \sum_{k=0}^{N/2-1} h_{2k} e^{i2\pi kn/(N/2)} + W^n \sum_{k=0}^{N/2-1} h_{2k+1} e^{i2\pi n/(N/2)} \tag{3.26}$$

$$H_n = F_n^e + W^n F_n^o \quad , \quad W = e^{i2\pi/N} \tag{3.27}$$

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 $n = 0, 1, ..., N$, míg a fele akkora méretű ponthalmazból készült F_n^e és F_n^o indexei $n = 0, 1, ..., N/2$ Az indexek szerinti periodikusság $F_n = F_{n\pm N/2}$ miatt azonban egyszerűen kétszer egymás után kell leírnunk az F_n számokat.

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

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. $N = 3, 5, 7, 11, 16....$ 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 készítőik melyik problémához melyik rutint ajánlják.

E. 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

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{i2\pi ft} dt = \int_{-\infty}^{\infty} f(t) e^{i2\pi ft} dt + i \int_{-\infty}^{\infty} g(t) e^{i2\pi ft} dt = F(f) + iG(f) \tag{3.28}$$

  Vegyük figyelembe, hogy valós függvényekre

$$(F(f))^* = \left(\int_{-\infty}^{\infty} f(t)e^{i2\pi ft}dt\right)^* = \int_{-\infty}^{\infty} f(t)e^{-i2\pi ft}dt = F(-f) \tag{3.29}$$

így

$$(H(f))^* = F(-f) - iG(-f) \quad \text{illetve} \quad (H(-f))^* = F(f) - iG(f) \tag{3.30}$$

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

$$F(f) = \frac{(H(-f))^* + H(f)}{2} \quad \text{és} \quad G(f) = i\frac{(H(-f))^* - H(f)}{2} \tag{3.31}$$

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

$$H(-f_n) = H_{N-n} \quad , \quad f_n = \frac{1}{N\Delta}\{0, 1, 2, ..., N/2\} \tag{3.32}$$

- 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

$$h_k = f_{2k} + i \cdot f_{2k+1} \quad , \quad k = 0, 1, ..., N/2 - 1 \tag{3.33}$$

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

$$H_n = F_n^e + iF_n^o \tag{3.34}$$

A Lánczos lemma kapcsán leírtak szerint

$$F_n = F_n^e + e^{i2\pi n/N}F_n^o \tag{3.35}$$

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

$$F_n = \frac{1}{2}\left[H_n + H_{-n}^*\right] - \frac{i}{2}\left[H_n - H_{-n}^*\right]e^{i2\pi n/N} \quad , \quad n = 0, 1, ..., N-1 \tag{3.36}$$

ahol felhasználtuk, hogy a valós bemenő adatok miatt $\left(F_{-n}^x\right)^* = F_n^x \qquad x = e/o$. A formulában a negatív indexekre továbbra is érvényes a (3.32) tárolási konvenció.

F. 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-szinusz sorba fejthetők

$$h(x) = \sum_{n=1}^{\infty} b_n \cdot sin\left(x\frac{n\pi}{T}\right) \qquad b_n = \frac{2}{T}\int_0^T h(x)\ sin\left(x\frac{n\pi}{T}\right)\ dt \tag{3.37}$$

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

$$B_n = \sum_{k=1}^{N-1} h_k \; sin\left(\frac{\pi nk}{N}\right) \;\approx\; \frac{T}{2\Delta} b_n = \frac{N}{2} b_n \qquad n = (0), 1, 2, ..., N-1, (N) \tag{3.38}$$

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

$$h_k = \frac{2}{N} \sum_{n=1}^{N-1} B_n \; sin\left(\frac{\pi nk}{N}\right) \tag{3.39}$$

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

$$\tilde{h}(x) = \left\{ \begin{array}{ccc} h(x) & , & x > 0 \\ -h(x) & , & x < 0 \end{array} \right\} \tag{3.40}$$

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

$$\tilde{H}(f) = \int_{-T}^{T} \tilde{h}(t) e^{i2\pi tf} dt = 2i \int_0^T h(t) \; sin(2\pi tf) dt \approx \Delta 2i \sum_{k=1}^{N-1} h_k \; sin(2\pi k \Delta f) \tag{3.41}$$

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

$$f_n = \frac{n}{2\Delta N} = \frac{n}{2T} \tag{3.42}$$

frekvenciákat választjuk, akkor $\tilde{H}(f_n) = \Delta \tilde{H}_n$, ahol $\left\{ \tilde{H}_n \mid n = 0, 1, .., 2N-1 \right\}$ a $\tilde{h}$ diszkrét Fourier transzformáltja. A megoldás tehát az, hogy az N számmal adott $\{h_k \mid k = 0, 1, ..., N-1\}$ 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

$$\tilde{H}_n = 2iB_n \tag{3.43}$$

alapján a szinusz transzformáltak az FFT együtthatókból számothatók.

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

$$h(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cdot cos\left(x \frac{n\pi}{T}\right) \qquad a_n = \frac{2}{T} \int_0^T h(x) \; cos\left(x \frac{n\pi}{T}\right) \; dt \tag{3.44}$$

A megfelelő diszkrét koszinusz transzformáltak

$$A_n = \frac{1}{2} \left\{ h_0 + (-1)^n \, h_N \right\} + \sum_{k=1}^{N-1} h_k \; cos\left(\frac{\pi nk}{N}\right) \tag{3.45}$$

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

$$\tilde{h}(x) = \left\{ \begin{array}{ll} h(x) & , \ x > 0 \\ h(x) & , \ x < 0 \end{array} \right\} \tag{3.46}$$

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

$$\tilde{H}_n = 2A_n \tag{3.47}$$

A koszinusz transzformáció megint (faktortól eltekintve) önmaga inverze

$$h_k = \frac{2}{N} \left[\frac{1}{2} \left\{ A_0 + (-1)^n A_N \right\} + \sum_{n=1}^{N-1} A_n \, cos \left(\frac{\pi n k}{N} \right) \right] \tag{3.48}$$

G. Többdimenziós FFT

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

$$F(\mathbf{q}) = \int f(\mathbf{r}) e^{-i\mathbf{qr}} d^3\mathbf{r} \quad \Longleftrightarrow \quad f(\mathbf{r}) = \frac{1}{(2\pi)^3} \int F(\mathbf{q}) e^{i\mathbf{qr}} d^3\mathbf{q} \tag{3.49}$$

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

$$\int f(\mathbf{r}) e^{-i\mathbf{qr}} d^3\mathbf{r} = 2\pi \int_0^\infty r^2 dr \int_{-1}^1 du \ f(r) e^{-iqru} = 2\pi \int_0^\infty r^2 dr \ f(r) \frac{2 \ sin(qr)}{qr} = \frac{4\pi}{q} \int_0^\infty r \cdot f(r) \cdot sin(qr) dr \tag{3.50}$$

és így a 3-dimenziós Fourier transzformált helyett a $g(r) = r \cdot f(r)$ függvény 1-dimenziós szinusz 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

$$H_{n,m} = \sum_{k=0}^{Nk-1} \sum_{l=0}^{Nl-1} h_{k,l} \ e^{i2\pi kn} e^{i2\pi lm} \tag{3.51}$$

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

$${}^{(k)}H_m = \sum_{l=0}^{Nl-1} h_{k,l} \ e^{i2\pi lm} \quad k = 0, 1, .., Nk-1 \tag{3.52}$$

diszkrét transzformáltat. Az így kapott tömböket átrendezzük

$${}^{(k)}H_m \to^{(m)} H_k \tag{3.53}$$

alakra, majd Nl alkalommal használjuk ezekre újból az egydimenziós

$$H_{n,m} = \sum_{k=0}^{Nk-1} {}^{(m)}H_k \ e^{i2\pi kn} \quad l = 0, 1, .., Nl-1 \tag{3.54}$$

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.

H. Konvolúciós egyenletek

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

$$g * h \equiv \int_{-\infty}^{\infty} h(t-\tau)g(\tau)d\tau = h * g \tag{3.55}$$

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 tű-impulzus, amit egy Dirac-delta általánosított függvénnyel írhatunk le, akkor a konvolúció eredményeként

$$g * \delta = g \tag{3.56}$$

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

$$(s * r)_j = \sum_{k=-M/2+1}^{M/2} s_{j-k} \cdot r_k \tag{3.57}$$

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

$$(s * r)(t_j) \approx \Delta \cdot (s * r)_j \tag{3.58}$$

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

$$(s * r)_j = \sum_{k=-N/2+1}^{N/2} s_{j-k} \cdot r_k \quad \Longleftrightarrow \quad S_n \cdot R_n \tag{3.59}$$

ahol $\{s_k \mid 0 \leq k \leq N-1\} \quad \Longleftrightarrow \quad \{S_k \mid 0 \leq k \leq N-1\}$ és $\{r_k \mid 0 \leq k \leq N-1\} \quad \Longleftrightarrow \quad \{R_k \mid 0 \leq k \leq N-1\}$. 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

$$r_k = 0 \quad \text{ha} \quad M/2 \leq k \leq N/2 \ \text{ vagy } \ -N/2+1 \leq k \leq -M/2+1 \tag{3.60}$$

- Óvatosan kell eljárnunk, 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 írnunk.

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

$$\mathcal{F}[h] = \mathcal{F}[f] \cdot \mathcal{F}[g] \implies \mathcal{F}[f] = \mathcal{F}[h]/\mathcal{F}[g] \implies \begin{cases} f = \mathcal{F}^{-1}\left[\mathcal{F}[h]/\mathcal{F}[g]\right] \\ f = h * \mathcal{F}^{-1}\left[1/\mathcal{F}[g]\right] \end{cases} \tag{3.61}$$

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

$$h|g \equiv \int_{-\infty}^{\infty} h(t+\tau)g(\tau)d\tau \tag{3.62}$$

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

$$(h|g)_j = \sum_{k=0}^{N-1} h_{j+k} \cdot g_k \quad \Longleftrightarrow \quad H_n \cdot G_n^* \tag{3.63}$$

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