SZÁMÍTÓGÉPES MODELLEZÉS

Tóth Gábor

ELTE TTK, Atomfizikai Tanszék
1117 Budapest, Pázmány sétány 1/A, 3.132
gtoth@hermes.elte.hu,
http://hermes.elte.hu/~gtoth/

2001. szeptember – december

Tartalomjegyzék

1. fejezet

Bevezetés

1.1. Szakirodalom és jegyzet

- C.A.J. Fletcher "Computational Techniques for Fluid Dynamics" 1. kötet
- R. LeVeque: "Numerical Methods for Conservation Laws"
- C. Hirsch: "Numerical Computation of Internal and External Flows" (fizikus könyvtár)
- Press, Teukolsky, Vettering, Flannery: "Numerical Recipes"
- G. Tóth: Computational Hydrodynamics and Magnetohydrodynamics, http://hermes.elte.hu/~gtoth/
- G. Tóth: Hidrodinamika, http://galahad.elte.hu/~csabai/compPhys/
- Journal of Computational Physics

1.2. A számítógépes modellezés célja

A fizikai valóságot lényegében kísérletekkel illetve matematikai modellek tanulmányozásával próbáljuk megismerni. Laboratóriumi kísérletek végzése nem mindig lehetséges, illetve nagyon költséges lehet. A matematikai modelleket megadó egyenletek analitikus megoldása gyakran nagy nehézségekbe ütközik. A számítógépes modellezés célja az egyenletek numerikus megoldása.

1.3. Néhány alkalmazás

Az alábbi felsorolás a számítógépes modellezés néhány területét jelzi, és azt is, hogy a kísérleti megközelítés milyen nehézségekbe ütközik:

- Biológia (pl. érrendszer) – nehéz mérni, változó geometria
- Csillagászat – túl messze van
- Geofizika – túl nagy
- Időjárás – túl késő
- Mérnöki alkalmazások – túl drága

1.4. Előnyök és hátrányok

+ Könnyen változtatható
+ Kísérletileg el nem érhető tartományokra is jó
+ Teljes információ
+ Olcsóbb mint a kísérlet
– Véges pontosság
– Az eredmény nem általános

1.5. A számítógépes modellezés módszere

A számítógép csak véges sok mennyiséggel tud számolni, ezért a folytonos változókra felírt (parciális) (differenciál) egyenleteket véges sok változóra vonatkozó algebrai egyenletekkel kell közelíteni. Ezt a lépést nevezzük **diszkretizációnak**:

- parciális differenciál egyenletek $\rightarrow$ algebrai egyenletek
- folytonos független változók $\rightarrow$ diszkrét változók

A véges sok **diszkrét változó** típusai:

- lokálisak (rácspont, cella, elem)
- globálisak (amplitúdó és frekvencia)

Példaképpen tekintsünk a konvekciót és diffúziót leíró parciális differenciálegyenletet (PDE):

$$\frac{\partial T}{\partial t} + v\frac{\partial T}{\partial x} - \alpha\frac{\partial^2 T}{\partial x^2} = 0 \tag{1.1}$$

ahol T a hőmérséklet, x és t a tér és idő változók, v a közeg sebessége és α a diffúziós együttható.

Ebből az úgynevezett véges differenciálás módszerével a következő algebrai egyenlet nyerhető:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + u\frac{T_{j+1}^n - T_{j-1}^n}{\Delta x} + \alpha\frac{T_{j-1}^n - 2T_j^n + T_{j+1}^n}{\Delta x^2} = 0 \tag{1.2}$$

ahol a diszkrét T_j^n változók n felső indexe az időlépést j alsó indexe a rácspontot jelöli, továbbá Δt az időlépés hossza és Δx a rácsállandó. A diszkrét és folytonos változók közti kapcsolat a következő

$$T_j^n \approx T(x_j, t_n) \tag{1.3}$$

$$x_j = x_0 + j\Delta x \quad (j = 1, 2, \ldots, J) \tag{1.4}$$

$$t_n = t_0 + n\Delta t \quad (n = 1, 2, \ldots, N) \tag{1.5}$$

Természetesen a diszkrét megoldás csak közelíti az egzakt analitikus megoldást. A tér és idő koordináták egy véges téridő tartományon belül helyezkednek el. Ennek határán a diszkretizált peremfeltételeknek megfelelően kell módosítani a diszkrét egyenleteket.

1.6. Számítási igények

Ahhoz, hogy a diszkrét megoldás pontos közelítése legyen a PDE analitikus megoldásának, viszonylag sok diszkrét változóra van szükség. Tipikusan egy térbeli változóhoz $100 - 1000$ rácspontra van szükség, 2 illetve 3 térbeli dimenzió esetén ennek négyzetével illetve köbével kell számolni. Egy-egy rácsponton $1 - 100$ függő változót kell tárolni. A diszkretizált egyenletekben egy-egy változóval tipikusan $10 - 100$ műveletet kell elvégezni egy időlépésben. Az időlépések száma néhány száztól a százezres nagyságrendig terjed. Egy időfüggő szimulációban tipikusan el kell menteni a lemezre $10 - 100$ időlépés teljes adatát. Összefoglalva:

- változók száma egy időlépésre: $10^3 - 10^8$
- műveletek száma egy változóra: $10^1 - 10^2$
- időlépések száma egy szimulációban: $10^2 - 10^5$

- lemezen tárolandó időlépések száma: $10^1 - 10^2$

Ennek alapján megbecsülhetjük, hogy egy szimulációhoz milyen hardverre van szükség:

- Memória igény: néhányszor $10^3 - 10^8$ valós szám
- Számítási igény: $10^4 - 10^{10}$ FLOP/lépés (floating point operation)
- Számítási igény: $10^6 - 10^{15}$ FLOP/szimuláció
- Tárolási igény: $10^5 - 10^{12}$ valós szám/szimuláció

1.7. Hardver

A számítógépek sebessége kb. 8 évenként 10-szeres faktorral nő! A memória és tárolási kapacitás is követik ezt az ütemet.

A számítógépek kapacitása a sebességen túl a párhuzamosan végezhető műveletekkel növelhető. Ma már egy személyi számítógép numerikus processzora is képes egy összeadás és egy szorzás egyidejű elvégzésére, és vehetünk személyi számítógépet két processzorral.

A szuperszámítógépek első generációja **vektor processzort** használt, mely ugyanazt a műveletet tudja elvégezni sok, tipikusan néhány száz számra. Később több vektorprocesszort kapcsoltak össze, ezeken lehet független programokat is futtatni, de párhuzamosan is működhetnek egyazon memóriát használva. A közös memória miatt a vektor processzorok száma korlátozott, tipikusan $8 - 16$ körül.

A szuperszámítógépek következő generációjában minden processzor külön memóriát használ, ezt **osztott memóriájú parallel számítógépnek** nevezzük. Az egyes egységeket nagy sebességű kommunikációs hardver köti össze. Az egyes processzorokon futó programoknak ezen a hardveren keresztül kell működésüket szinkronizálni, és a szükséges adatokat átvinni. Az osztott memória lehetővé teszi, hogy több ezer processzor működjön párhuzamosan, ugyanakkor a processzorok közti kommunikáció programozási szempontból nagyon kényelmetlen.

A jelenleg legnépszerűbb **közös memóriájú parallel számítógépek** megpróbálják egyesíteni a többprocesszoros vektor számítógépek és az osztott memóriájú parallel gépek előnyeit. Fizikailag minden processzor saját memóriával rendelkezik, de szoftveresen bármelyik processzor közvetlenül hozzáférhet az egész memóriához. A közvetlen elérést a nagysebességű kommunikációs hardver teszi elfogadhatóan gyorssá. Ezek a gépek szintén több

száz, sőt ezer processzort használnak, ugyanakkor a párhuzamos programozásuk jóval egyszerűbb mint az osztott memóriás gépeké.

Néhány jelenleg használt szuperszámítógép típus:

- Vektor számítógépek (pl. Cray YMP, NEC SX4)
- Parallel vektor számítógépek (pl. Cray C90)
- Osztott memóriájú parallel számítógépek (pl. Cray T3E, IBM SP, PC klaszter)
- Közös memóriájú parallel számítógépek (pl. SGI IRIX 2000, Sun E10000)

1.8. Parallel sebesség

Naivul azt remélhetnénk, hogy kellően sok processzoron futtatva a programot tetszőlegesen gyorsan megkapható a szimuláció eredménye. Sajnos **Amdahl törvényéből** következik, hogy ez a várakozásunk nem realisztikus. A G sebesség növekedés ugyanis függ attól, hogy a program futás hanyad része párhuzamosítható. Ha ezt $P \in [0,1]$ jelöli, és N a processzorok száma, akkor könnyen látható, hogy

$$G = \frac{1}{(1-P) + P/N} \tag{1.6}$$

Ha $N \approx 1/(1-P)$, akkor $G \approx N/2$, azaz a maximálisan várható N-szeres sebességnövekedésnek csak a felét sikerül elérni. A helyzet tovább romlik, ha a processzorok közötti kommunikációs időt is figyelembe kell vennünk.

A párhuzamos futtatás hatékonyságát kétféleképpen szokták mérni:

- sebesség növekedés (speed up) – rögzített összméretű feladat
- skálázás (scaling) – processzoronként konstans méretű feladat

Nyilván jóval könnyebb elérni, hogy egy program jól skálázzon, mint hogy lineáris, a processzorok számával arányos sebesség növekedést mutasson.

1.9. Szoftver

A számítógépes modellezésben használatos programnyelvek többé kevésbé idő és elterjedtség sorrendjében:

- Fortran 77, Fortran 90/95
- C, C++

- Java

Fontos megjegyezni, hogy míg a Fortran 77 egy mára elavult nyelv, addig a vele felülről kompatibilis Fortran 90/95 egy modern, kényelmes és rendkívül hatékony nyelv. A Fortran 90 körülbelül úgy viszonylik a Fortran 77-hez, mint a C++ a C-hez. A Fortran 90 ugyan nem objektum orientált, de tartalmaz modulokat, definiálható változó típusokat, dinamikus változókat, mutatókat, opcionális paramétereket, paraméterek típusától és számától függő szubrutinokat, műveletek átdefiniálását, azaz egy modern nyelv. Előnye a C++ nyelvvel szemben, hogy nagyon sok támogatást ad tömbműveletekhez és matematikai függvényekhez, jóval egyszerűbb nyelv, és általában valamivel gyorsabb programot lehet írni benne. További szempont, hogy a már meglévő gyakran hatalmas szoftverek jelentős része Fortran 77-ben íródott, így ezeket könnyebben lehet Fortran 90-re átírni mint mondjuk C-be.

Természetesen a C++ előnye, hogy objektumorientált, és hogy közvetlen hozzáférést ad a számítógép operációsrendszeréhez. Számítógépes modellezésnél azonban erre általában nincs szükség, hiszen számok és paraméterek alapján kell új számokat előállítani. Az objektumorientáltságnak nagyon nagy méretű szoftverfejlesztésnél van jelentősége. Gyakran ötvözik össze a C++ és Fortran programokat úgy, hogy a számolási rész Fortranban, míg a magas szintű vezérlés C++ -ban íródik. Ez utóbbi szerepre az új objektumorientált nyelv, a Java is rendkívül alkalmas.

Párhuzamos számítógépeken szükség van valamilyen kommunikációs szoftverre. Nagyjából elterjedtségi sorrendben a ma használt szoftverek:

- MPI – Message Passing Interface
- OpenMP – Multithreaded Parallel Execution
- HPF – High Performance Fortran

Az MPI egy általános kommunikációs könyvtár osztott memóriájú gépekre, de természetesen közös memóriájú parallel gépeken is használható. A legalapvetőbb utasításai: üzenetek küldése és fogadása, szinkronizáció, valamint redukciós műveletek (összeg, maximum). Az MPI könyvtár függvényeken és szubrutinokon keresztül érhető el C és Fortran forráskódból.

Az OpenMP egy egészen új nyelv, mely kifejezetten közösmemóriájú gépekhez készült. Alapvetően a C vagy Fortran nyelvű forráskódba írt fordító direktívákkal kell kijelölni, hogy a program mely részei futtathatók párhuzamosan, illetve mely változók közösek és melyek kezelendők privát változókként. Az OpenMP még gyermekbetegségeit éli, de jelenleg dinamikusan fejlődik.

Az HPF a Fortran 90 egy kiterjesztése osztott memóriájú parallel számítógépekre. Szintén fordító direktívákat használ. Lényegében az adattömbök processzorok közti felosztását kell definiálni, a szükséges kommunikációt a fordító hozza létre. Viszonylag egyszerű adatstruktúrákra a HPF nagyon hatékony és programozási szempontból sokkal kényelmesebb, mint az MPI. Sajnos a jelenlegi trendek arra mutatnak, hogy az osztott memóriás gépeken az MPI, a közös memóriájú gépeken pedig a fordítók autoparallelizációja illetve az OpenMP ki fogja szorítani.

2. fejezet

Egyenletek

2.1. A feladat korrekt kitűzése

Egy PDE-re vonatkozó probléma akkor van korrektül kitűzve, ha

- Létezik megoldás
- Pontosan egy megoldás létezik
- A megoldás a peremfeltételektől folytonosan függ

Számítógépes modellezésnél a perem- (kezdeti- és határ-) feltételek és a megoldás is közelítőek, így a folytonos függés különösen fontos. Ellentétben az analitikus problémákkal, a numerikusan modellezett számítási tartomány mindig véges.

Tipikus peremfeltétel típusok:

- Dirichlet: $U = f$
- Neumann: $\frac{\partial U}{\partial n} = f$ és/vagy $\frac{\partial U}{\partial s} = g$
- Kevert: $\frac{\partial U}{\partial n} + kU = f$

ahol $\partial U/\partial n$ a normál irányú, $\partial U/\partial s$ az érintőleges irányú deriváltat jelöli, f, g és k pedig tetszőleges függvények.

2.2. Parciális differenciál egyenletek levezetése

A tömegmegmaradás egy tetszőleges V térfogatra így írható:

$$\frac{d}{dt}\int_V \rho dV = -\int_{\partial V} \rho \mathbf{v} \cdot d\mathbf{A} \tag{2.1}$$

ahol ρ a sűrűség, $\mathbf{v}$ a sebesség, ∂V a V térfogatot körülvevő zárt felület, $d\mathbf{A}$ pedig a felület elem kifelé mutató normál vektora.

Folytonos $\rho(\mathbf{x}, t)$ esetén a bal oldalon az időderivált és az integrál felcserélhető, míg a jobb oldalon alkalmazhatjuk a Gauss tételt:

$$\int_V \frac{\partial \rho}{\partial t} dV = -\int_V \operatorname{div}(\rho \mathbf{v}) dV \tag{2.2}$$

A tetszőleges V-re vett integrál elhagyásával kapjuk a kontinuitási egyenlet differenciális formáját.

2.3. Tipikus parciális differenciál egyenletek

Tömegmegmaradás/konvekció:

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0 \tag{2.3}$$

Diffúzió:

$$\frac{\partial \rho}{\partial t} - \operatorname{div}(\alpha \mathbf{grad}\, \rho) = 0 \tag{2.4}$$

ahol α a diffúziós együttható, mely függhet a tértől és időtől.

Gravitációs potenciál:

$$\operatorname{div} \mathbf{grad} \varphi = 4\pi G \rho \tag{2.5}$$

ahol φ a gravitációs potenciál és G a gravitációs állandó.

Hidrodinamikai momentum egyenlet(rendszer):

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \operatorname{div}(\rho \mathbf{v} \circ \mathbf{v}) + \mathbf{grad} p = 0 \tag{2.6}$$

ahol p a nyomás.

Hullámegyenlet:

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \tag{2.7}$$

ahol c a hullámsebesség.

2.4. Parciális differenciál egyenletek osztályozása

A parciális differenciálegyenleteknek nincsen teljesen általános osztályozása. Bizonyos típusú PDE-k besorolhatóak a **elliptikus, parabolikus, hiperbo-**

likus osztályokba, melyeknek matematikai és numerikus megoldása más-más megközelítést igényel.

2.4.1. Elsőrendű PDE 2 független változóval

$$A\frac{\partial u}{\partial t} + B\frac{\partial u}{\partial x} = C \tag{2.8}$$

A karakterisztikus görbe definíciója:

$$\left.\frac{dx}{dt}\right|_k = \frac{B}{A} \tag{2.9}$$

azaz a görbe iránytangense minden téridő pontban B/A. A karakterisztikus görbe mentén

$$\left.\frac{du}{dt}\right|_k = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\left.\frac{\partial x}{\partial t}\right|_k = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{B}{A} = \frac{C}{A} \tag{2.10}$$

azaz az egyenlet integrálható! Hasonlóan, a görbe mentén

$$\left.\frac{du}{dx}\right|_k = \frac{C}{B} \tag{2.11}$$

Ha az A és B együtthatók konstansok és $C = 0$, akkor a

$$u = f\left(x - \frac{B}{A}t\right) \tag{2.12}$$

megoldás tetszőleges folytonos f függvényre, azaz a homogén egyenletnek van egy B/A sebességgel haladó hullámmegoldása.

Ha A és B nem konstans, akkor a karakterisztikák görbék az $x - t$ síkon!

2.4.2. Elsőrendű PDE rendszer 2 független változóval

Általában az $u_1, u_2, \ldots u_N$ függő változókra vonatkozó elsőrendű két független változójú PDE rendszer a

$$\sum_{j=1}^{N} A_{i,j}\frac{\partial u_j}{\partial x} + \sum_{j=1}^{N} B_{i,j}\frac{\partial u_j}{\partial y} = C_i \tag{2.13}$$

formában írható, ahol $i = 1, \ldots, N$.

Ismét keressünk olyan $(dy/dx)_k$ karakterisztikus irányokat melyek mentén az egyenletek integrálhatóak. Ez az eredeti egyenletekkel általában nem

tehető meg, mert az x és y deriváltak együtthatóinak hányadosa az egyes változókra más és más. Ezért keressünk olyan $L^(k)_i$ szorzókat az i-dik egyenlethez és olyan $(dy/dx)_k$ karakterisztikus irányokat (itt k a karakterisztikus görbék indexe), hogy a felösszegzett egyenletekben

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left[L_i^{(k)}A_{i,j}\frac{\partial u_j}{\partial x}+L_i^{(k)}B_{i,j}\frac{\partial u_j}{\partial y}\right]=\sum_{i=1}^{N}L_i^{(k)}C_i \tag{2.14}$$

az együtthatók hányadosa minden változóra azonos legyen, azaz minden $j = 1 \ldots N$-re

$$\left.\frac{dy}{dx}\right|_k=\frac{\sum_{i=1}^{N}L_i^{(k)}B_{i,j}}{\sum_{i=1}^{N}L_i^{(k)}A_{i,j}} \tag{2.15}$$

Ha ezt sikerül megvalósítani, akkor a parciális deriváltak átírhatók a görbe menti deriváltakra:

$$\sum_{i=1}^{N}L_i^{(k)}A_{i,j}\frac{\partial u_j}{\partial x}+\sum_{i=1}^{N}L_i^{(k)}B_{i,j}\frac{\partial u_j}{\partial y}=\sum_{i=1}^{N}L_i^{(k)}A_{i,j}\left.\frac{du_j}{dx}\right|_k \tag{2.16}$$

Ezután új karakterisztikus változókat vezethetünk be

$$w_k=\sum_{j=1}^{N}\sum_{i=1}^{N}L_i^{(k)}A_{i,j}u_j \tag{2.17}$$

amire nézve a felösszegzett egyenlet egészen egyszerűen

$$\left.\frac{dw_k}{dx}\right|_k=C'_k \tag{2.18}$$

amit ki lehet integrálni. Ha sikerül N karakterisztikus görbét találnunk, akkor a w_k $(k = 1 \ldots N)$ karakterisztikus változókból ki lehet számítani az eredeti u_j $(j = 1 \ldots N)$ változókat.

Ahhoz, hogy találjunk egy karakterisztikus görbét, a

$$\sum_{i=1}^{N}L_i^{(k)}\left[(dy/dx)_kA_{i,j}-B_{i,j}\right]=0 \tag{2.19}$$

lineáris egyenletrendszert $(j = 1, \ldots, N)$ kell megoldani az $L_i^{(k)}$ szorzókra és a $(dy/dx)_k$ karakterisztikus irányra. Az egyenletrendszernek akkor van nem-triviális megoldása, ha

$$\det\left[(dy/dx)_k\mathbf{A}-\mathbf{B}\right]=0 \tag{2.20}$$

Az egyenletrendszer osztályozása a determináns (ami egy N-ed fokú polinomot ad $(dy/dx)_k$-ra) gyökei alapján

- hiperbolikus, ha N valós gyököt találunk
- parabolikus, ha minden gyök valós, de kevesebb mint N különböző van
- elliptikus, ha van komplex gyök

2.4.3. Magasabb rendű PDE rendszer 2 független változóval

Magasabb rendű egyenletrendszer visszavezethető egy első rendűre, ha megfelelő új változókat vezetünk be. Például az általános 2 független változós másodfokú PDE

$$au_{xx} + bu_{xy} + cu_{yy} = h \tag{2.21}$$

átírható a $v = u_x$ és $w = u_y$ változók bevezetésével az

$$\begin{aligned} u_x &= v & (2.22)\\ w_x - v_y &= 0 & (2.23)\\ av_x + bv_y + cw_y &= h & (2.24) \end{aligned}$$

elsőrendű egyenletrendszerre. Ez a

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & a & 0 \end{pmatrix} \begin{pmatrix} u_x \\ v_x \\ w_x \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & b & c \end{pmatrix} \begin{pmatrix} u_y \\ v_y \\ w_y \end{pmatrix} = \begin{pmatrix} v \\ 0 \\ h \end{pmatrix} \tag{2.25}$$

formában is írható, amiből az $\mathbf{A}\lambda - \mathbf{B}$ determinánsa

$$\det \begin{pmatrix} \lambda & 0 & 0 \\ 0 & 1 & \lambda \\ 0 & \lambda a - b & -c \end{pmatrix} = -\lambda(\lambda^2 a - b\lambda + c) \tag{2.26}$$

Ennek egy valós gyöke a $\lambda_1 = 0$. További 2 valós és nem nulla gyöke akkor van, ha a diszkrimináns

$$d = b^2 - 4ac \tag{2.27}$$

pozitív. Ekkor a PDE hiperbolikus. Ha a diszkrimináns 0 akkor minden gyök 0, azaz 3-nál kevesebb valós gyököt találtunk és a PDE parabolikus. Végül, ha a diszkrimináns negatív, akkor két komplex gyököt találunk, azaz a PDE elliptikus.

Megjegyzés: ha a koordinátákat transzformáljuk, az egyenlet típusa nem változik!

2.4.4. Másodrendű PDE sok független változóval

Ha a független változók $x_1, x_2, \ldots x_N$, és

$$\sum_{j=1}^{N}\sum_{k=1}^{N} a_{j,k}\frac{\partial^2 u}{\partial x_j \partial x_k} + H = 0 \tag{2.28}$$

A kereszt deriváltak eltüntethetők megfelelő koordináta-transzformációval:

$$\sum_{k=1}^{N} \lambda_k \frac{\partial^2 u}{\partial \xi_k^2} + H' = 0 \tag{2.29}$$

Ehhez az $A = a_{j,k}$ mátrix sajátvektorait és λ_k sajátértékeit kell megkeresni. Ezek mindig léteznek és valósak, hiszen az A mátrix $\partial^2 u/\partial x_j \partial x_k = \partial^2 u/\partial x_k \partial x_j$ azonosság miatt mindig szimmetrikussá tehető.

A PDE osztályozása a sajátértékek alapján a következő:

- ultraparabolikus, ha több 0 sajátérték van
- parabolikus, ha pontosan egy 0 sajátérték van
- elliptikus, ha minden sajátértéknek azonos az előjele
- hiperbolikus, ha egy kivételével mindnek ugyanaz az előjele
- ultrahiperbolikus, ha több mint egy ellentétes előjelű

2.4.5. Első rendű PDE rendszer sok független változóval

Ez az általános eset (hiszen magasabb rendű PDE rendszer mindig visszavezethető elsőrendűre), de erre nincs is általános osztályozási eljárás. Részeredmények azonban vannak.

Például 3 független változó és N függő változó esetén így írható a PDE rendszer:

$$\sum_{j=1}^{N} A_{i,j}\frac{\partial u_j}{\partial x} + \sum_{j=1}^{N} B_{i,j}\frac{\partial u_j}{\partial y} + \sum_{j=1}^{N} C_{i,j}\frac{\partial u_j}{\partial z} = E_i \tag{2.30}$$

Ekkor

$$\det\left(\lambda_x \mathbf{A} + \lambda_y \mathbf{B} + \lambda_z \mathbf{C}\right) = 0 \tag{2.31}$$

a karakterisztikus polinom. Itt $\lambda_x, \lambda_y, \lambda_z$ egy felület normálisát adják egy adott pontban. Ha $\lambda_x, \lambda_y, \lambda_z$ valósak, akkor a felület egy karakterisztikus felület. Ha N valós gyököt találunk, akkor a PDE rendszer hiperbolikus.

2.5. Példák

2.5.1. Hiperbolikus egyenlet

A legegyszerűbb hiperbolikus PDE a hullám egyenlet

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \tag{2.32}$$

A hullámegyenlet típusa $A = 1$, $B = 0$, $C = -c^2$, így $B^2 - 4AC = 4c^2 > 0$ alapján valóban hiperbolikus. A karakterisztikus görbék merededeksége a

$$\left(\frac{dx}{dt}\right)^2 - c^2 = 0 \tag{2.33}$$

megoldásai:

$$\frac{dx}{dt} = \pm c \tag{2.34}$$

A karakterisztikus változókát a karakterisztikus görbék mentén kiintegrálva megkaphatóak a hullámegyenlet megoldásai, melyek

$$u = f(x - ct) + g(x + ct) \tag{2.35}$$

alakban írhatók, ahol f és g függvények a peremfeltételektől függnek. Mint a megoldásból látható, **a hullámok nem disszipálódnak, és nem folytonos megoldás is létezik.**

A hullámegyenlet tisztán kezdeti érték problémaként is megoldható. Ekkor két feltételt kell megadni a $t = 0$-ra, például $u_0 = u(x, 0)$-t és $v_0 = \partial u(x, t = 0)/\partial t$. Ekkor analitikusan megadható a megoldás

$$u(x,t) = \frac{1}{2}\left[u_0(x + ct, 0) + u_0(x - ct, 0) + \int_{x-ct}^{x+ct} v_0(r)dr\right] \tag{2.36}$$

Egy hiperbolikus PDE-nél a peremfeltétel megadható tisztán kezdeti és kezdeti plusz határfeltételek formájában is. Általában annyi feltételt kell megadni, ahány karakterisztika halad a tartomány belseje felé!

2.5.2. Parabolikus egyenlet

A legegyszerűbb parabolikus egyenlet a diffúziós egyenlet, ami egy dimenzióban konstans diffúziós együttható esetén a következő

$$\frac{\partial \rho}{\partial t} - \alpha \frac{\partial^2 \rho}{\partial x^2} = 0 \tag{2.37}$$

Könnyen leolvasható, hogy $A = -\alpha$ és $B = C = 0$, így $B^2 - 4AC = 0$, azaz a PDE valóban parabolikus (az $A = 0$, $C = -alpha$ választás nem ad értelmes eredményt a karakterisztikus görbére). A karakterisztikus görbék irányára

$$-\alpha \left(\frac{dt}{dx}\right)^2 = 0 \tag{2.38}$$

adódik, azaz csak egy karakterisztikus görbe van, aminek a meredeksége $dt/dx = 0$, azaz az x tengellyel párhuzamos. A megoldás időben exponenciális csökkenést mutat, azaz **a hullámok amplitúdója csökken, a peremfeltételben esetleg jelenlevő diszkontinuitások az értelmezési tartomány belsejében kisimulnak.**

A parabolikus egyenleteknél a peremfeltételekhez a kezdeti és határfeltételeket egyaránt meg kell adni. A kezdeti feltétel tipikusan Dirichlet, a határfeltételek tetszőleges típusúak.

2.5.3. Elliptikus egyenlet

A legegyszerűbb példa elliptikus egyenletre a Laplace egyenlet. Üres tér esetén a gravitációs potenciálra vonatkozó egyenlet két dimenziós Descartes koordinátarendszerben a következőképpen írható:

$$\operatorname{div} \mathbf{grad}\varphi = \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0 \tag{2.39}$$

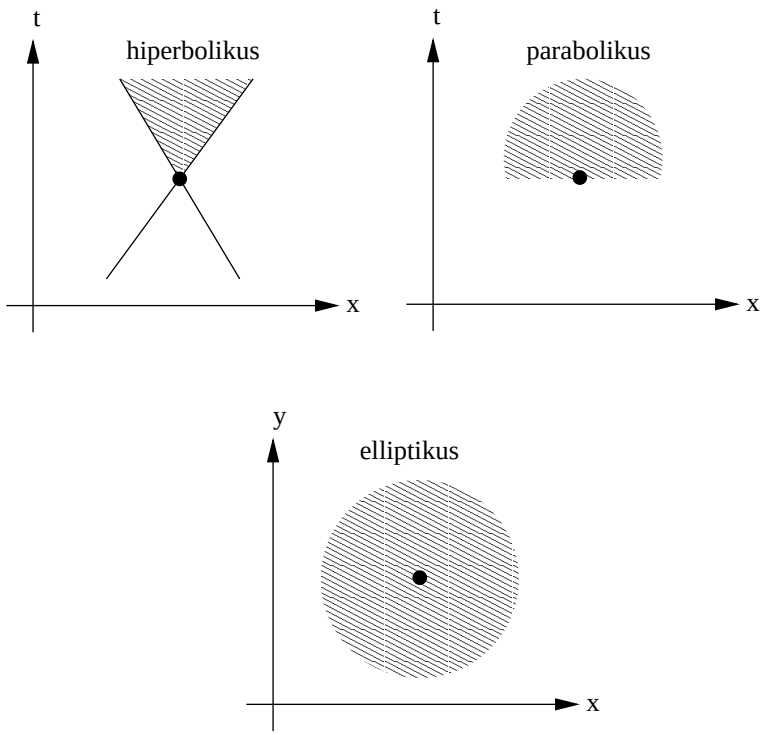
Erre az egyenletre $A = C = 1$ és $B = 0$, azaz $B^2 - 4AC = -4 < 0$, vagyis a Laplace egyenlet valóban elliptikus. A karakterisztikus görbék meredeksége a

$$\left(\frac{dy}{dx}\right)^2 + 1 = 0 \tag{2.40}$$

egyenlet megoldásai, amik imaginárius ak.

Az elliptikus PDE-nek nincs időszerű független változója, ezért a peremfeltételeket egy zárt görbén illetve felületen kell megadni. Bármilyen típusú határfeltétel használható, de ha csak Neumann-t használunk, akkor kell egy extra megkötés, hiszen különben a megoldást egy tetszőleges additív konstanssal el lehetne tolni.

Megjegyezzük, hogy speciálisan a Laplace egyenletre érvényes a maximum tétel: a megoldás maximuma a tartomány határán helyezkedik el, azaz hullámokról nemigen lehet beszélni.



2.1. ábra.
A három PDE osztály téridő diagramja. A satírozott rész függ a ponttal jelölt eseménytől. Az elliptikus esetben mindkét koordináta tér jellegű.

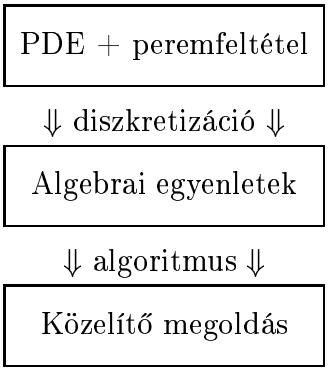
2.6. Összefoglaló

Típus	hiperbolikus	parabolikus	elliptikus
példa	hullám egy.	diffúzió egy.	Laplace egy.
karakterisztikák	csupa különböző valós	kevesebb valós	komplex
hullámok	disszipáció mentes	disszipálódnak	nincsenek
szakadás	terjed	szétdiffundál	csak határon
időszerű vált.	van	van	nincs
függés	karakterisztikák között	időben korábbi	teljes tér
peremfeltételek	kezdeti v. kezdeti+határ	kezdeti+határ	zárt határ

3. fejezet

Diszkretizáció

A PDE numerikus megoldásához vezető lépéseket az alábbi diagramm szemlélteti:



3.1. Példák

Tekintsük a diffúziós egyenlet

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} \tag{3.1}$$

néhány lehetséges diszkretizációját.

3.1.1. Térbeli diszkretizáció

Véges differencia módszerrel diszkretizálva:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \alpha \frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2} \tag{3.2}$$

$$T_j^{n+1} = T_j^n + \frac{\alpha \Delta t}{\Delta x^2} \left(T_{j+1}^n - 2T_j^n + T_{j-1}^n \right) \tag{3.3}$$

Véges elem módszerrel:

$$\frac{1}{6}\frac{T_{j-1}^{n+1}-T_{j-1}^{n}}{\Delta t}+\frac{2}{3}\frac{T_{j}^{n+1}-T_{j}^{n}}{\Delta t}+\frac{1}{6}\frac{T_{j+1}^{n+1}-T_{j+1}^{n}}{\Delta t}=\alpha\frac{T_{j+1}^{n}-2T_{j}^{n}+T_{j-1}^{n}}{\Delta x^2} \quad (3.4)$$

Spektrális módszerrel:

$$T = \sum_{j=1}^{J} a_j(t)\varphi_j(x) \quad (3.5)$$

$$a_j^{n+1} = a_j + \Delta t \sum_{j=1}^{J} p_j a_j^n \quad (3.6)$$

ahol p_j ismert algebrai együtthatók.

3.1.2. Időbeli diszkretizáció

Az eddigi példákban eddig mindig **explicit** idődiszkretizációt alkalmaztunk, azaz az $n+1$-dik időlépéshez szükséges térderiváltakat az n-dik lépésből becsültük meg. Az explicit diszkretizáció általában egyszerű algebrai egyenletekhez vezet.

Az **implicit** idődiszkretizáció esetén a tér deriváltban felhasználjuk az $n+1$-dik lépés változóit:

$$\frac{T_j^{n+1}-T_j^n}{\Delta t}=\alpha\frac{T_{j+1}^{n+1}-2T_j^{n+1}+T_{j-1}^{n+1}}{\Delta x^2} \quad (3.7)$$

Ez egy egyenletrendszert ad T^{n+1}-re, aminek a megoldását valamilyen nem triviális algoritmussal kell megkeresni.

Az sem szükségszerű, hogy az $n+1$-dik lépéshez tartozó megoldás kiszámításához csak az n-dik és $n+1$-dik lépéseket használjuk. Egy példa **háromszintű** idődiszkretizációra az időben centrális diszkretizáció:

$$\frac{T_j^{n+1}-T_j^{n-1}}{2\Delta t} = \alpha\frac{T_{j+1}^{n}-2T_j^{n}+T_{j-1}^{n}}{\Delta x^2} \quad (3.8)$$

$$T_j^{n+1} = T_j^{n-1} + \frac{2\alpha\Delta t}{\Delta x^2}\left(T_{j+1}^{n}-2T_j^{n}+T_{j-1}^{n}\right) \quad (3.9)$$

Mint látni fogjuk, ez a diszkretizació a diffúziós egyenletre instabil. A többszintű idődiszkretizációhoz több megoldás vektort kell eltárolni, valamint az első lépésben, amikor még csak egy vektor adott, más idődiszkretizációt kell használni.

Végül az idődiszkretizációt szét lehet választani a térbeli diszkretizációtól. Ehhez a PDE-t csak térben diszkretizálva egy közönséges differenciál egyenletrendszert kapunk:

$$\frac{\partial T_j}{\partial t} = L(T_1, \ldots, T_J) \tag{3.10}$$

ahol L a térbeli derivált diszkretizációját adja meg. Erre az egyenletrendszerre alkalmazhatóak a közönséges differenciál egyenletek megoldási módszerei. Azonban nem biztos, hogy érdemes nagyon precízen integrálni időben, mert L már tartalmazza a térbeli diszkretizáció hibáját.

3.2. Véges differencia formulák ad hoc konstrukciója

Tegyük fel, hogy a megoldás folytonos és sokszor deriválható! Egy adott x_j, t_n téridő pont körüli rácspontokban a megoldás felírható Taylor sorok segítségével. Például az $x_j + \Delta x, t_n$ pontban:

$$T^n_{j+1} = T^n_j + \Delta x \left[\frac{\partial T}{\partial x}\right]^n_j + \frac{\Delta x^2}{2}\left[\frac{\partial^2 T}{\partial x^2}\right]^n_j + O(\Delta x^3) \tag{3.11}$$

Hasonlóan az $x_j, t_n + \Delta t$ pontban

$$T^{n+1}_j = T^n_j + \Delta t \left[\frac{\partial T}{\partial t}\right]^n_j + \frac{\Delta t^2}{2}\left[\frac{\partial^2 T}{\partial t^2}\right]^n_j + O(\Delta t^3) \tag{3.12}$$

A fenti formulákból látható, hogy a térbeli és időbeli deriváltak hogyan becsülhetők:

$$\frac{T^n_{j+1} - T^n_j}{\Delta x} = \left[\frac{\partial T}{\partial x}\right]^n_j + \frac{\Delta x}{2}\left[\frac{\partial^2 T}{\partial x^2}\right]^n_j + O(\Delta x^3) \tag{3.13}$$

$$\frac{T^{n+1}_j - T^n_j}{\Delta t} = \left[\frac{\partial T}{\partial t}\right]^n_j + \frac{\Delta t}{2}\left[\frac{\partial^2 T}{\partial t^2}\right]^n_j + O(\Delta t^3) \tag{3.14}$$

3.3. Véges differencia formulák szisztematikus konstrukciója

Tegyük fel, hogy a $\partial T/\partial x$-t kívánjuk közelíteni az x_j, t_n pontban a szomszédos két pontbeli érték felhasználásával. Írjuk fel a közelítést egy általános

lineáris kombináció formájában:

$$\begin{aligned}\left[\frac{\partial T}{\partial x}\right]_j^n &= aT_{j-1}^n + bT_j^n + cT_{j+1}^n + O(\Delta x^m) \qquad (3.15)\\ &= (a+b+c)T_j^n + (-a+c)\Delta x\left[\frac{\partial T}{\partial x}\right]_j^n + (a+c)\frac{\Delta x^2}{2}\left[\frac{\partial^2 T}{\partial x^2}\right]_j^n \\ &\quad +(-a+c)O(\Delta x^3)\end{aligned}$$

Ennek alapján a következő egyenleteket írhatjuk fel az a, b, c együtthatókra:

$$a+b+c = 0 \qquad (3.16)$$

$$-a+c = \frac{1}{\Delta x} \qquad (3.17)$$

$$a+c = 0 \qquad (3.18)$$

Az utolsó egyenlet a másodrendű hibatag kiejtésére szolgál. Megoldás:

$$a = -\frac{1}{2\Delta x} \quad b = 0 \quad c = \frac{1}{2\Delta x} \qquad (3.19)$$

Az így nyert szimmetrikus 3 pont formula

$$\left[\frac{\partial T}{\partial x}\right]_j^n = \frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} + O(\Delta x^2) \qquad (3.20)$$

másodrendű ($m = 2$) pontos közelítés.

Természetesen nem kötelező az $a + c = 0$ választás, így kapható az előre ($a = 0$) ill. visszafele ($c = 0$) féloldalas közelítések:

$$\left[\frac{\partial T}{\partial x}\right]_j^n = \frac{T_{j+1}^n - T_j^n}{\Delta x} + O(\Delta x) \qquad (3.21)$$

$$\left[\frac{\partial T}{\partial x}\right]_j^n = \frac{T_j^n - T_{j-1}^n}{\Delta x} + O(\Delta x) \qquad (3.22)$$

$$(3.23)$$

Ha több mint 3 pontot használunk fel, magasabb rendű közelítő formulákat is lehet tervezni. Például az 5 pont szimmetrikus formula

$$\left[\frac{\partial T}{\partial x}\right]_j^n = \frac{T_{j-2}^n - 8T_{j-1}^n + 8T_{j+1} - T_{j+2}}{12\Delta x} + O(\Delta x^4) \qquad (3.24)$$

negyedrendig pontos.

3.4. Alacsony és magas rendű differencia formulák

Naivan azt gondolhatnánk, hogy minél magasabb rendű formulákat használunk, annál pontosabb lesz a megoldás. Azonban

- Magas rendű formula több számítást igényel
- Véges felbontásra nem mindig a magasabb rendű a pontosabb
- Szakadások és éles gradiensek esetén a sorfejtés nem érvényes
- A magasabb rendű formulák gyakran kevésbé stabilak

Általában legalább 2-od rendű formulákra van szükség, ennél magasabb rend csak bizonyos esetekben fizetődik ki.

3.5. Hullám reprezentáció

Egy Δx rácsállandójú rácson legfeljebb $2\Delta x$ hosszú hullámokat lehet reprezentálni (és azt se valami pontosan!). Tehát a diszkrét megoldás egy hosszú hullámhosszú közelítése az egzakt megoldásnak. Nézzük meg, hogy egy differencia formula milyen pontosan reprezentálja egy adott λ hullámhosszú, $k = 2\pi/\lambda$ hullámszámú

$$T = \sin(kx + \varphi) \tag{3.25}$$

megoldás deriváltjait. Az egzakt deriváltak

$$\frac{\partial T}{\partial x} = k\cos(kx + \varphi) \tag{3.26}$$

$$\frac{\partial^2 T}{\partial x^2} = -k^2 \sin(kx + \varphi) \tag{3.27}$$

Helyettesítsünk az első deriváltat közelítő szimmetrikus 3-pont formulába

$$\begin{aligned} \frac{T_{j+1} - T_{j-1}}{2\Delta x} &= \frac{\sin(kx + k\Delta x + \varphi) - \sin(kx - k\Delta x + \varphi)}{2\Delta x} \\ &= \frac{1}{2\Delta x} 2\sin(k\Delta x)\cos(kx + \varphi) \\ &= \left[\frac{\sin(k\Delta x)}{k\Delta x}\right] k\cos(kx + \varphi) \end{aligned} \tag{3.28}$$

A szögletes zárójelben lévő együttható tartalmazza a közelítő és az egzakt amplitúdók hányadosát. Látható, hogy a hosszú hullámhosszú ($k \to 0$) limitben az amplitúdók aránya 1-hez tart, viszont a minimális hullámhosszra

($k = \pi/\Delta x$) a diszkrét formula a deriváltra mindig 0-t ad a fázistól függetlenül.

Nézzük meg a féloldalas differenciál formula hibáját is:

$$\begin{aligned}\frac{T_{j+1} - T_j}{\Delta x} &= \frac{\sin(kx + k\Delta x + \varphi) - \sin(kx + \varphi)}{\Delta x} \\ &= \frac{1}{\Delta x} 2\sin(k\Delta x/2)\cos(kx + \varphi + k\Delta x/2) \\ &= \left[\frac{\sin(k\Delta x/2)}{k\Delta x/2}\right] k\cos(kx + \varphi + k\Delta x/2) \qquad (3.29)\end{aligned}$$

Az amplitúdó hiba mellett fellép egy fázis hiba is!

A másodrendű derivált hibája szimmetrikus 3-pont formulára

$$\begin{aligned}\frac{T_{j+1} - 2T_j + T_{j-1}}{\Delta x^2} &= \frac{\sin(kx + \Delta x + \varphi) - 2\sin(kx + \varphi) + \sin(kx - k\Delta x + \varphi)}{\Delta x^2} \\ &= -\left[\frac{\sin(k\Delta x/2)}{k\Delta x/2}\right]^2 k^2 \sin(kx + \varphi) \qquad (3.30)\end{aligned}$$

Itt ismét a szögletes zárójelben látható az amplitúdóhiba.

4. fejezet

Konvergencia

A numerikus modellezés célja, hogy kellően finom rácsot alkalmazva az algebrai egyenletek numerikus megoldása a PDE-k analitikus megoldáshoz tartson minden rácspontban:

$$T_j^n \to T^a(x_j, t_n) \qquad \text{ha } \Delta x, \Delta t \to 0 \tag{4.1}$$

ahol T a diszkrét, míg T^a a folytonos analitikus megoldást jelöli. Ezt a feltételt nevezzük konvergenciának. Azonban a konvergencia bizonyítása általános esetben rendkívül nehéz.

4.1. Konzisztencia fogalma

A konvergenciának nyilvánvalóan szükséges feltétele, hogy az algebrai egyenletek a $\Delta x, \Delta t \to 0$ limitben pontosan közelítsék a PDE-t. Ezt úgy ellenőrizhetjük, hogy **az analitikus egyenlet megoldását az algebrai egyenletekbe helyettesítjük**, és felső becslést adunk a PDE-től való eltérésre, azaz a diszkretizációs hibára. Amennyiben a diszkretizációs hiba $\Delta x, \Delta t \to 0$ esetén nullához tart, az algebrai egyenlet konzisztens az eredeti parciális differenciál egyenletekkel.

A konzisztencia azonban önmagában nem elegendő a konvergenciához, mivel a hibák lépések során fokozatosan összeadódnak.

4.2. Stabilitás fogalma

Az algebrai egyenletek megoldása sohasem egzakt, mivel a számítógép a valós számokat csak véges sok tizedesjegyig tárolja. Így a megoldás során mindig fellép a kerekítési hiba. A diszkretizáció stabilitása azt követeli meg, hogy **az**

algebrai egyenletek megoldásában a hiba nem nőhet határok nélkül. Nyilvánvaló, hogy egy instabil megoldási módszerrel nyert megoldás nem fog az analitikus megoldáshoz konvergálni, ugyanis a felbontás növelésével egyre több időlépésre van szükségünk, hogy a megoldást egy adott t időpontra vonatkozóan megkapjuk, és így a felgyűlt hiba egyre nagyobb lesz.

4.3. Lax ekvivalencia tétele

Mint láttuk, a konzisztencia és a stabilitás a konvergencia szükséges feltételei. Lax bebizonyította, hogy **lineáris kezdetiérték problémákra a véges differencia módszer akkor és csak akkor konvergál, ha stabil és konzisztens**. A bizonyítás kiterjeszthető minden olyan módszerre, ami valamilyen pontszerű változókat használ, így a véges térfogat és a véges elem módszerre is általánosítható.

A Lax ekvivalencia tétel ugyan nagyon fontos, de a gyakorlati problémák általában (kezdeti+)határ feltételekkel adottak, és gyakran nemlineárisak. Ebben az esetben a stabilitás és konzisztencia szükséges, de nem elegendő feltételei a konvergenciának.

4.4. Numerikus konvergencia

A gyakorlatban a konzisztencia és a stabilitás mellett a numerikus konvergenciát követeljük meg, ami azt jelenti, hogy **a numerikus megoldás egyre finomabb rácsokon egyre közelebb kerül egy határértékhez**. Általában feltehető, hogy ez a határérték az analitikus megoldással egyezik meg. A numerikus konvergenciának természetesen feltétele a stabilitás, azonban a konzisztencia egy független és ellenőrizendő feltétel!

1. Egyszerű problémákon meggyőződünk arról, hogy a diszkretizáció és annak implementációja (beprogramozott megvalósítása) konzisztens a PDE-vel.

2. A valódi problémát legalább három különböző rácsfelbontással megoldjuk, és megvizsgáljuk a numerikus konvergenciát.

Amikor az analitikus megoldás ismert, a hibát a T^a analitikus megoldáshoz képest az

$$E_1 = \frac{1}{N} \sum_{j=1}^{N} |T_j^n - T^a| \tag{4.2}$$

úgynevezett 1-es normában, vagy az

$$E_2 = \left[\frac{1}{N} \sum_{j=1}^{N} (T_j^n - T^a)^2 \right]^{1/2} \tag{4.3}$$

2-es normában szokás mérni. Kétféle rácsfelbontással elvégezve a szimulációt megállapíthatjuk, hogy az

$$E \propto (\Delta x)^m \tag{4.4}$$

relációban mi az m kitevő, azaz a numerikus konvergencia exponens értéke. Minél finomabb rácsokat alkalmazunk, annál közelebb fog esni m a diszkretizációs hiba elméletileg megállápított rendjéhez.

Amikor nem ismert az analitikus megoldás, akkor legalább háromféle felbontásra van szükség, ugyanis két numerikus megoldás közötti eltérés nagyságából nem lehet következtetni konvergenciára. Ha a három megoldást D (durva), K (közepes) és F (finom) betűkkel jelöljük, akkor a következő eltéréseket kell kiszámítani

$$E_{KD}^2 = \frac{1}{N_D} \sum_{j=1}^{N_D} (T_j^K - T_j^D)^2 \tag{4.5}$$

$$E_{FK}^2 = \frac{1}{N_D} \sum_{j=1}^{N_D} (T_j^F - T_j^K)^2 \tag{4.6}$$

ahol az összegzést a legdurvább rács N_D rácspontjára végezzük el. Ha a durva rácspontokkal nem esnek egybe a finomabb rácsok rácspontjai, akkor megfelelő rendű interpolációt kell alkalmazni. Tegyük fel, hogy a megoldás m exponenssel konvergál az ismeretlen T^a analitikus megoldáshoz minden pontban, azaz

$$T_j^{D,K,F} = T_j^a + (\Delta x_{D,K,F})^m E_j \tag{4.7}$$

ahol E_j a lokális hiba $\Delta x = 1$-re. Ha ezt a modellt behelyettesítjük a (4.5) és (4.6) egyenletekbe, akkor két egyenletet kapunk az m konvergencia kitevőre, valamint a $\sum_j E_j^2$ globális hibára. A gyakorlatban általában megelégszünk azzal a kvalitatív kritériummal, hogy E_{FK}-nak jóval kisebbnek kell lennie mint E_{KD}, valamint E_{FK} legyen elegendően kicsi a változók értékéhez illetve a konkrét szimulációnál megkövetelt pontossághoz képest.

4.5. Konzisztencia vizsgálata

Az algebrai egyenlet konzisztens a PDE-vel, ha az analitikus megoldást az algebrai egyenletbe helyettesítve a maradéktagok nullához tartanak a rácsfelbontás növelésével. Vizsgáljuk meg pl. a hődiffúzió egyenletét az egyszerű

FTCS (forward in time, centered in space) véges differencia diszkretizációra:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \alpha \frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2} \tag{4.8}$$

Taylor sorba fejtve a T_j^{n+1}, T_{j+1}^n és T_{j-1}^n tagokat

$$\frac{\partial T}{\partial t} + \frac{\Delta t}{2}\frac{\partial^2 T}{\partial t^2} + O(\Delta t^2) = \alpha \left(\frac{\partial^2 T}{\partial x^2} + \frac{\Delta x^2}{12}\frac{\partial^4 T}{\partial x^4} \right) + O(\Delta x^4) \tag{4.9}$$

A két vezető tag a diffúziós PDE-t adja, míg a maradék tagok $O(\Delta t, \Delta x^2)$ rendűek. Vegyük észre, hogy

$$\frac{\partial^2 T}{\partial t^2} = \partial_t \left(\alpha \partial_{xx} T \right) = \alpha \partial_{xx} \partial_t T = \alpha^2 \frac{\partial^4 T}{\partial x^4} \tag{4.10}$$

A hiba tag így átírható a

$$E^{FTCS} = \left(\frac{\alpha^2 \Delta t}{2} - \frac{\alpha \Delta x^2}{12} \right) \frac{\partial^4 T}{\partial x^4} + O(\Delta t^2, \Delta x^4) \tag{4.11}$$

formába, amiből a vezető tag eltüntethető, ha olyan időlépést illetve rácsfelbontást választunk, amire

$$\Delta t = \frac{\Delta x^2}{6\alpha} \tag{4.12}$$

Ezzel a választással a diszkretizáció rendje és pontossága nagy mértékben nő!

Természetesen nem mindig lehet kiejteni a hiba tagokat, pl. a teljesen implicit

$$\frac{T_j^n - T_j^{n-1}}{\Delta t} = \alpha \frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2} \tag{4.13}$$

diszkretizációra a hiba tag

$$E^{impl} = \left(\frac{\alpha^2 \Delta t}{2} + \frac{\alpha \Delta x^2}{12} \right) \frac{\partial^4 T}{\partial x^4} + O(\Delta t^2, \Delta x^4) \tag{4.14}$$

Mivel $\Delta x, \Delta t, \alpha > 0$, nincs mód a vezető tag eltűntetésére.

Némely modern kompakt diszkretizációban hasonló trükökkel növelik a diszkretizáció rendjét.

4.6. Stabilitás vizsgálat

4.6.1. Mátrix módszer

Lineáris PDE-re a ξ hiba ugyanazt az egyenletet elégíti ki, mint a megoldás. Ha az algebrai egyenletek is lineárisak, akkor

$$\xi^{n+1} = A\xi^n \tag{4.15}$$

formában írhatók, ahol az A mátrix az összes (nem rögzített értékű) rácspontban definiált ξ vektorra hat. A hibák akkor nem nőnek korlátok nélkül, ha A legnagyobb abszolút értékű sajátértékére

$$|\lambda| < 1 \tag{4.16}$$

Például a diffúziós egyenlet

$$T_j^{n+1} = T_j^n + \frac{\alpha \Delta t}{\Delta x^2} \left(T_{j+1}^n - 2T_j^n + T_{j-1}^n \right) \tag{4.17}$$

diszkretizációhoz tartozó $N \times N$-es mátrix, ahol N az ismeretlenek száma, a következő lesz

$$A = \begin{pmatrix} 1-2s & s & & & & & \\ s & 1-2s & s & & & & \\ & s & 1-2s & s & & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \cdot \\ & & & & & s & 1-2s \end{pmatrix} \tag{4.18}$$

Itt bevezettük az

$$s = \alpha \frac{\Delta t}{\Delta x^2} \tag{4.19}$$

dimenziótlan diffúziós együttható. A határfeltételek a mátrix első és utolsó soraiban vannak elrejtve. Ha T a hibát jelenti, akkor ez a mátrix megfelel annak, hogy a hiba 0 a két határon. Az A mátrix sajátértékei analitikusan meghatározhatóak:

$$\lambda_j = 1 - 4s \sin^2 \left(\frac{j\pi}{2(N+1)} \right) \tag{4.20}$$

Látható, hogy $|\lambda_j| < 1$ ha

$$s \leq \frac{1}{2} \quad \text{azaz} \quad \Delta t \leq \frac{\Delta x^2}{2\alpha} \tag{4.21}$$

A mátrix módszer tehát tartalmazza a határfeltételek hatását is, viszont a sajátértékek meghatározása általános esetben nagyon nehéz. A legnagyobb sajátértéket a hatványozás módszerével lehet megkapni, ami annyit tesz, hogy A-t újra és újra alkalmazzuk, a megoldást normáljuk, amíg a leginstabilabb módushoz tartozó sajátvektort és a hozzátartozó sajátértéket meg nem kapjuk. Ez azonban lényegében azonos a diszkretizált egyenletek megoldásával.

4.6.2. Von Neumann módszer

Elvileg csak lineáris, konstans együtthatós kezdeti érték problémákra ad szükséges és elégséges feltételt. Nem lineáris problémákra a nem linearitás befagyasztását alkalmazzuk. Ilyenkos szükséges, de nem elégséges stabilitási feltételt kapunk.

Fejtsük a hibát, vagy lineáris egyenletnél magát a megoldást, Fourier sorba. A stabilitás feltétele, hogy minden egyes Fourier komponens stabil legyen, így elegendő egyet vizsgálni:

$$\xi_j^0 = \exp(ikx_j) \qquad j = 1, \ldots, N \tag{4.22}$$

ahol $k = 2\pi m/(N\Delta x)$ a hullámszám. Lineáris konstans együtthatós algebrai egyenletek esetén az n-ik lépésben

$$\xi_j^n = (G)^n \exp(ikx_j) \tag{4.23}$$

ahol G a *komplex* erősítési faktor. Például a hővezetési egyenlet FTCS diszkretizációjára

$$\begin{aligned}
(G)^{n+1}e^{ikx_j} &= (1-2s)(G)^n e^{ikx_j} + s(G)^n e^{ik(x_j+\Delta x)} + s(G)^n e^{ik(x_j-\Delta x)} \\
G &= 1 - 2s + se^{ik\Delta x} + se^{-ik\Delta x} \\
&= 1 - 2s + 2s\cos(k\Delta x) \\
&= 1 - 4s\sin^2(k\Delta x/2)
\end{aligned} \tag{4.24}$$

amiből $|G| < 1$, ha $s \leq \frac{1}{2}$ egyezésben a mátrix módszer eredményével. Lényegében expliciten megoldottuk a sajátérték problémát, hiszen konstans együtthatós (periodikus határfeltételekhez tartozó) mátrixok sajátvektorai exponenciális alakban kereshetők.

A diffúziós egyenletet impliciten diszkretizálva a von Neumann stabilitás vizsgálat a következőt adja

$$\begin{aligned}
G &= 1 - 2Gs + sGe^{ik\Delta x} + sGe^{-ik\Delta x} \\
G &= \frac{1}{1 + 2s - 2s\cos(k\Delta x)} \\
&= \frac{1}{1 + 4s\sin^2(k\Delta x/2)}
\end{aligned} \tag{4.25}$$

ami minden s-re $|G| < 1$-t ad, azaz az implicit diszkretizáció *feltétel nélkül stabil.*

Tekintsük most a konvekciós egyenlet FTCS diszkretizációját:

$$\rho_j^{n+1} = \rho_j^n - v\Delta t\frac{\rho_{j+1}^n - \rho_{j-1}^n}{2\Delta x} \tag{4.26}$$

Helyettesítsük be a Fourier komponenst, és egyszerűsítsünk:

$$\begin{aligned} G &= 1 - \frac{v\Delta t}{2\Delta x}\left(e^{ik\Delta x} - e^{-ik\Delta x}\right) \\ &= 1 - iC\sin(k\Delta x) \end{aligned} \tag{4.27}$$

ahol $C = v\Delta t/\Delta x$ a Courant szám. A G erősítési faktor abszolútértéke minden $C > 0$-ra nagyobb mint 1, azaz az FTCS diszkretizáció a kontinuitási egyenletre *feltétel nélkül instabil*!

Ha maga a fizikai probléma instabil, akkor a megoldás, és így a hiba is nőni fog. Azonban ez a növekedés a fizikai idővel, és nem a lépések számával arányos, azaz a stabilitási feltétel

$$|G| \leq 1 + O(\Delta t) \tag{4.28}$$

Több dimenzióra **k** egy hullámvektor lesz. Több szintű diszkretizáció esetén magasabb rendű egyenletet kapunk G-re. Egyenletrendszerek esetén M változóra G egy $M \times M$ mátrix lesz. A diszkretizáció akkor stabil, ha az erősítési mátrix λ_m sajátértékeire fennáll, hogy

$$|\lambda_m| \leq 1 + O(\Delta t) \tag{4.29}$$

Az erősítési mátrix sajátvektorai a módusokban az egyes változók összetételét adják meg.

4.7. Konvergencia nem folytonos megoldások esetén

A parciális differenciál egyenleteknek csak folytonos és differenciálható függvények lehetnek a megoldásai. Azonban bevezethető a **gyenge megoldás** fogalma, mely lehetővé teszi a nem folytonos megoldások matematikailag konzisztens kezelését. Ilyen megoldások hiperbolikus PDE-kben léteznek. A valóságban a lökéshullámok és a kontakt diszkontinuitások nagyon éles gradiensekként jelentkeznek, de ezeket nagyon jól lehet szakadásokkal modellezni.

4.7.1. Gyenge megoldás

Tekintsük a

$$\frac{\partial U}{\partial t} + \mathrm{div}\mathbf{F} = S \tag{4.30}$$

PDE rendszert, ahol U a függővváltozók vektora, F a fluxust és S a forrástagokat jelöli. Ha ezt a PDE-t kiintegráljuk egy V térfogatra, visszakapjuk a megmaradási törvényt kifejező

$$\frac{d}{dt}\int_V UdV + \int_{\partial V} \mathbf{F}\cdot d\mathbf{A} = \int_V SdV \tag{4.31}$$

integrálegyenletet. A gyenge megoldás ennek az integrálegyenletnek egy nem (feltétlenül) folytonos megoldása minden V tartományra. Ez tovább integrálható időben a t_1, t_2 intervallumra:

$$\int_V U(t_2)dV - \int_V U(t_1)dV + \int_{t_1}^{t_2}\int_{\partial V} \mathbf{F}\cdot d\mathbf{A}dt = \int_{t_1}^{t_2}\int_V SdVdt \tag{4.32}$$

Egy másik ekvivalens megközelítésben a PDE-t megszorozzuk egy folytonosan differenciálható kompakt tartójú ϕ teszt függvénnyel és integrálunk a teljes térre és az időre

$$\int_0^\infty dt \int dV \left(\phi\frac{\partial U}{\partial t} + \phi \mathrm{div}\mathbf{F} - \phi S\right) = 0 \tag{4.33}$$

Részenkénti integrálással átírható

$$\int_0^\infty dt \int dV \left(\frac{\partial \phi}{\partial t}U + \mathbf{grad}\phi\cdot\mathbf{F} + \phi S\right) = -\int dV\phi(t=0)U(t=0) \tag{4.34}$$

A többi határon vett tag eltűnik, mert ϕ kompakt tartójú. A PDE gyenge megoldásai a fenti differenciál integrál egyenlet megoldásai tetszőleges ϕ-re.

Fontos megjegyezni, hogy általában egy PDE-nek több gyenge megoldása is létezhet. Ezek között szerepel a fizikailag értelmes megoldás is, mely a nullához tartó viszkozitás limitben kapott megoldás például a gáz dinamikában. A fizikai gyenge megoldást az entrópia növekedési elve is kiválasztja, ezért szokás **entrópia megoldásnak** is nevezni.

4.7.2. Lax-Wendroff tétel

A konzisztencia és stabilitási vizsgálatokban kihasználtuk, hogy a megoldás folytonosan differenciálható. Lax ekvivalencia tétele megmenthető, ha a gyenge megoldást mint folytonos megoldások határértékét tekintjük. Ugyanakkor a konvergencia rendje általában kisebb mint amit a konzisztencia vizsgálat alapján folytonos megoldásokra várhatunk.

Nemlineáris PDE-k gyenge megoldásaira Lax tétele nyilván nem érvényes, sőt előfordulhat, hogy

- a lineárisan stabil módszer nem-lineárisan instabil
- a numerikus megoldás egy olyan megoldáshoz konvergál, ami nem gyenge megoldás
- a numerikus megoldás nem az entrópia megoldáshoz konvergál

Az első és harmadik problémával később fogunk foglalkozni. A másodikkal kapcsolatban Lax és Wendroff megmutatták, hogy **egy konzervatív diszkretizáció mindig jó gyenge megoldáshoz konvergál, ha konvergál**.

4.7.3. Konzervatív diszkretizáció

Egy diszkretizációt akkor nevezünk konzervatívnak, ha a PDE-t konzervatív formájában úgy diszkretizálja, hogy a megmaradó változók diszkrét értelemben is megmaradnak. Például szabályos rácson kiválasztva egy tetszőleges részt

$$\sum_j U_j^{n+1} = \sum_j U_j^n + \frac{\Delta t}{\Delta x} \sum_{\text{határ}} \mathbf{F} \cdot d\mathbf{A} \tag{4.35}$$

azaz a diszkrét változók összege csak a határon átmenő diszkrét fluxus miatt változhat. Több dimenziós nem szabályos rácsra a feltétel

$$\sum_j V_j U_j^{n+1} = \sum_j V_j U_j^n + \Delta t \sum_{\text{határ}} \mathbf{F} \cdot d\mathbf{A} \tag{4.36}$$

alakban írható, ahol V_j a j-ik cella térfogata.

A konzervatív diszkretizáció garantálja, hogy a szakadásokra (pl. lökéshullámok) vonatkozó ugrás feltételek numerikusan is teljesülnek.

5. fejezet

Súlyozott Reziduum Módszerek

A véges differencia módszernél a megoldás csak a rácspontokban adott, a rácspontok között nem definiált. A súlyozott reziduum módszereknél viszont a numerikus megoldás mindenütt definiált, és a

$$T(\mathbf{x},t) = T_0(\mathbf{x},t) + \sum_{j=1}^{J} a_j(t)\varphi_j(\mathbf{x}) \tag{5.1}$$

formában írható, ahol $\varphi_j(\mathbf{x})$ ismert alakú függvények. Ha a numerikus T megoldásra hattatjuk az $L(T^a) = 0$ alakú PDE (ahol T^a az analitikus megoldás) L differenciáloperátorát, akkor egy maradvány tagot, ún. reziduumot kapunk:

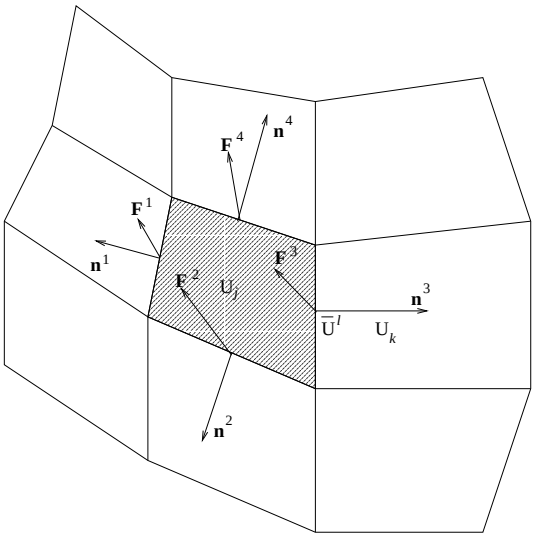
$$R(\mathbf{x},t) = L(T) \tag{5.2}$$

Célunk az, hogy R minél kisebb legyen, azaz a numerikus megoldás T minél pontosabban kielégítse a PDE-t. A súlyozott reziduum módszerek esetén R-nek a W_m súlyfüggvényekkel szorzott, a V számítási tartományra vett integrálját tesszük nullává:

$$\int_V W_m(\mathbf{x})R(\mathbf{x},t)d\mathbf{x} = 0 \tag{5.3}$$

A W_m súlyfüggvény alakjától függően az alábbi módszereket kapjuk

- Kollokációs módszer: $W_m(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_m)$
- Véges térfogat módszer: $W_m(\mathbf{x}) = 1$ ha $\mathbf{x} \in D_m$ és 0 egyébként
- Galerkin módszer: $W_m(\mathbf{x}) = \phi_m(\mathbf{x})$
- Legkisebb négyzetek: $W_m(\mathbf{x}) = \partial R/\partial a_m$



5.1. ábra.
Véges térfogat diszkretizáció 2 dimenzióban. Az n^l a normálvektorokat, míg F^l a fluxusokat jelöli, amit például a határra átlagolt $\bar{U}$ alapján lehet kiszámítani.

A kollokációs módszer lényegében a véges differenciához hasonlít, hiszen a numerikus megoldásra egy-egy pontban kapunk $R(\mathbf{x}_m) = 0$ feltételt. Azonban a véges differencia módszerben nincs közelítő megoldás.

A Galerkin módszert a **véges elem** és a **spektrális** módszerekben alkalmazzák. Lényeges, hogy a súlyfüggvények komplett rendszert alkossanak, így $M \to \infty$ esetén $R \to 0$.

Végül a legkisebb négyzetek módszere az $R^2(\mathbf{x}, t)$-t minimalizálja az $a_m(t)$ együtthatók függvényében.

5.1. Véges térfogat módszer

Ha a súlyozott reziduum módszerben a súlyfüggvényt a rácscellák karakterisztikus függvényeinek választjuk, akkor az m-ik cellára

$$\int_{D_m} L(T) = 0 \tag{5.4}$$

Ha $L(T)$-ben csak első deriváltak szerepelnek és felírható

$$\frac{\partial T}{\partial t} + \mathrm{div}\mathbf{F} = 0 \tag{5.5}$$

formában, akkor

$$\frac{d}{dt}\int_{D_m} T dV + \int_{\partial D_m} \mathbf{F}\cdot d\mathbf{A} = 0 \tag{5.6}$$

Ezt közelíthetjük úgy, hogy a fluxusokat a cella határoló élek/lapok közepén értékeljük ki:

$$\frac{d}{dt}\int_{D_m} T dV + \sum_l F_{m,l} d\mathbf{A}_{m,l} = 0 \tag{5.7}$$

ahol l az m-ik cella éleit indexeli.

Általában a gradiens, rotáció és divergencia operátorokat

$$\int_{D_m} dV \mathbf{grad} T = \sum_l d\mathbf{A}_{m,l} T_{m,l} \tag{5.8}$$

$$\int_{D_m} dV \mathrm{div}\mathbf{u} = \sum_l d\mathbf{A}_{m,l}\cdot \mathbf{u}_{m,l} \tag{5.9}$$

$$\int_{D_m} dV \mathbf{rot u} = \sum_l d\mathbf{A}_{m,l}\times \mathbf{u}_{m,l} \tag{5.10}$$

formában diszkretizálhatjuk.

Magasabb rendű differenciáloperátorok, pl. a Laplace operátor több féleképpen is diszkretizálható véges térfogat módszerrel. A legegyszerűbb a gradiens és a divergencia operátorok egymás utáni alkalmazása. Ennél bonyolultabb, de valamivel kompaktabb tartót ad, ha a gradienst **duális rácson** diszkretizáljuk.

5.2. Véges Elem Módszer

$$T = \sum_j T_j \phi_j(\mathbf{x}) \tag{5.11}$$

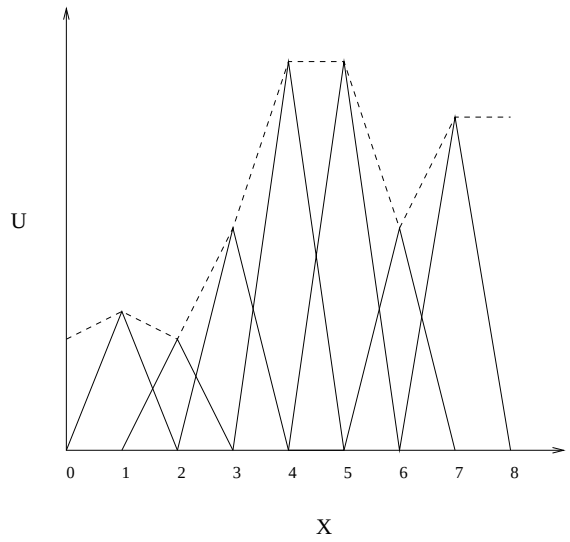
ahol T_j egy csomóponti változó, azaz T értéke egy adott helyen, míg ϕ_j egy szakaszonként alacsony rendű kompakt tartójú polinom, mely néhány szomszédos elemen vesz fel csak 0-tól különböző értéket.

Egy dimenzióban a lineáris elemeket definiáló bázis függvényeket így írhatjuk fel:

$$\phi_j = \begin{cases} \frac{x-x_{j-1}}{x_j-x_{j-1}} & x_{j-1}\le x\le x_j \\ \frac{x_{j+1}-x}{x_{j+1}-x_j} & x_j \le x \le x_{j+1} \\ 0 & \text{máshol} \end{cases} \tag{5.12}$$

Vegyük észre, hogy $\phi_j(x_j) = 1$ és $\phi_j(x_{j+1}) = \phi_j(x_{j-1}) = 0$. Mivel a ϕ függvények egy adott x pontban csak két elemre adnak nullától különböző értéket

$$T(x) = \phi_{j-1}T_{j-1} + \phi_j T_j \quad x_{j-1}\le x\le x_j \tag{5.13}$$



5.2. ábra.
Véges elem módszer 1 dimenzióban lineáris elemekkel. A csomóponti értékekkel megszorzott lineáris elemeket a folytonos vonallal rajzolt háromszögek, ezek összegét, azaz a közelítő megoldást szaggatott vonal jelzi.

ami egy lineáris interpolációt jelent a T_{j-1} és T_j csomóponti értékek között. Nézzük meg mit kapunk egy egyszerű időderiváltra a súlyozott reziduum módszerben. Legyen a súlyfüggvény ϕ_j, így elegendő az $[x_{j-1}, x_{j+1}]$ intervallumra integrálni. A közelítő megoldásban szereplő elemekből csak a ϕ_{j-1}, ϕ_j és ϕ_{j+1} különböznek 0-tól ebben az intervallumban, tehát

$$\int dx\phi_j \frac{\partial T}{\partial t} = \int_{x_{j-1}}^{x_{j+1}} dx\phi_j \left(\phi_{j-1}\frac{\partial T_{j-1}}{\partial t} + \phi_j \frac{\partial T_j}{\partial t} + \phi_{j+1}\frac{\partial T_{j+1}}{\partial t}\right) = \frac{\Delta x_{j-1/2}}{6}\frac{\partial T_{j-1}}{\partial t} + \frac{\Delta x_{j-1/2} + \Delta x_{j+1/2}}{3}\frac{\partial T_j}{\partial t} + \frac{\Delta x_{j+1/2}}{6}\frac{\partial T_{j+1}}{\partial t} \quad (5.14)$$

ahol a $\phi_j\phi_{j-1}$, ϕ_j^2, illetve $\phi_j\phi_{j+1}$ integrálokat expliciten kiszámítottuk.

Elsőrendű térbeli derivált esetén a ugyanennek a három tagnak az x szerinti deriváltját kell kiintegrálni

$$\int dx\phi_j \left(T_{j-1}\frac{\partial \phi_{j-1}}{\partial x} + T_j \frac{\partial \phi_j}{\partial x} + T_{j+1}\frac{\partial \phi_{j+1}}{\partial x}\right) = -\frac{1}{2}T_{j-1} + \frac{1}{2}T_{j+1} \quad (5.15)$$

Másodrendű térbeli deriváltra nem használhatjuk ugyanezt a módszert, mivel $\partial^2\phi/\partial x^2 = 0$-t kapnánk függetlenül T_j-től. Ezért részenkénti integrálást kell alkalmazni

$$\int dx\phi_j \frac{\partial^2 T}{\partial x^2} = -\int dx \frac{\partial \phi_j}{\partial x}\frac{\partial T}{\partial x} \quad (5.16)$$

$$
\begin{aligned}
&= -\int dx \frac{\partial \phi_j}{\partial x}\left(T_{j-1}\frac{\partial \phi_{j-1}}{\partial x} + T_j \frac{\partial \phi_j}{\partial x} + T_{j+1}\frac{\partial \phi_{j+1}}{\partial x}\right) \\
&= \frac{1}{\Delta x_{j-1/2}} T_{j-1} - \left(\frac{1}{\Delta x_{j-1/2}} + \frac{1}{\Delta x_{j+1/2}}\right) T_j + \frac{1}{\Delta x_{j+1/2}} T_{j+1} \qquad (5.17)
\end{aligned}
$$

A részenkénti integrálásnál a határon vett tagok eltűnnek, mivel az összes ϕ függvény tartója kompakt.

Egy teljes példaként diszkretizáljuk a

$$\frac{\partial T}{\partial t} + v\frac{\partial T}{\partial x} - \alpha \frac{\partial^2 T}{\partial x^2} = 0 \qquad (5.18)$$

konvekció-diffúzió egyenletet a véges elem módszerrel úgy, hogy az időderiváltakat egyszerű véges differencia módszerrel becsüljük:

$$
\begin{aligned}
\frac{\Delta x_{j-1/2}}{6}\frac{T_{j-1}^{n+1} - T_{j-1}^n}{\Delta t} &+ \frac{\Delta x_{j-1/2} + \Delta x_{j+1/2}}{3}\frac{T_j^{n+1} - T_j^n}{\Delta t} + \frac{\Delta x_{j+1/2}}{6}\frac{T_{j+1}^{n+1} - T_{j+1}^n}{\Delta t} \\
&+ v\frac{T_{j+1}^n - T_{j-1}^n}{2} \qquad (5.19) \\
&- \alpha\left[\frac{T_{j-1}^n}{\Delta x_{j-1/2}} - \frac{T_j^n}{\Delta x_{j-1/2}} - \frac{T_j^n}{\Delta x_{j+1/2}} + \frac{T_{j+1}^n}{\Delta x_{j+1/2}}\right] = 0
\end{aligned}
$$

Érdemes megfigyelni, hogy egy egyenletrendszert kaptunk T^{n+1}-re, mivel az időderivált tartalmaz egy térbeli operátort. A véges elem módszernél tehát akkor is egyenletrendszert kell megoldani, ha az idődiszkretizáció explicit.

5.3. Spektrális Módszer

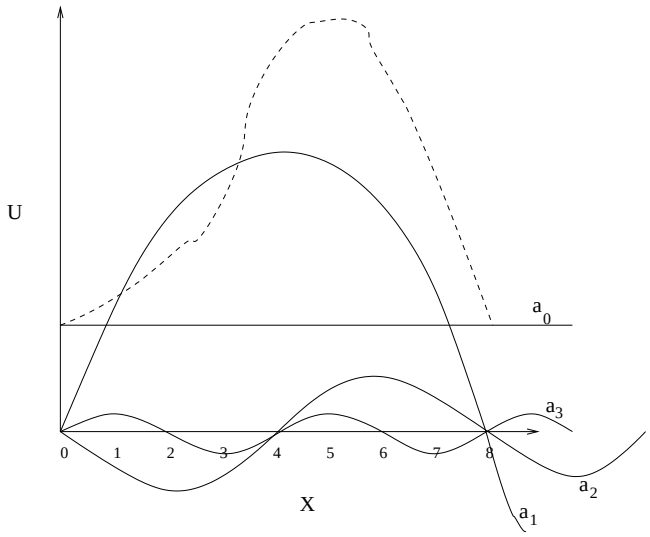
A spektrális módszerben a ϕ_j bázis függvények az egész térre kiterjedő ortogonális függvények a_j amplitúdóval:

$$T = T_0 + \sum_{j=1}^{J} a_j(t)\phi_j(\mathbf{x}) \qquad (5.20)$$

Ha lehetséges, T_0-t úgy választjuk meg, hogy kielégítse a kezdeti és határfeltételeket.

Oldjuk meg például a $\partial_t T - \alpha \partial_{xx} T = 0$ diffúziós egyenletet $T(x,0) = 5x - 4x^2$ kezdeti és $T(0,t) = 0$, $T(1,t) = 1$ határfeltételekkel. Ekkor

$$T = 5x - 4x^2 + \sum_{j=1}^{J} a_j(t)\sin(j\pi x) \qquad (5.21)$$



5.3. ábra.
Spektrális módszer 1 dimenzióban. A szaggatott vonallal ábrázolt közelítő megoldás a 4 folytonos vonallal jelölt a_j amplitúdókkal megszorzott bázisfüggvény összege.

közelítő megoldás automatikusan kielégíti a peremfeltételeket. A reziduum

$$R = L(T) = 8\alpha + \sum_{j=1}^{J} \left[\partial_t a_j + \alpha a_j (j\pi)^2\right] \sin(j\pi x) \tag{5.22}$$

A súlyozott reziduum módszer szerint

$$\begin{aligned} 0 &= \int_0^1 dx \sin(m\pi x) R(x,t) \\ &= \frac{8\alpha}{m\pi} \left[-\cos(m\pi x)\right]_0^1 + \frac{1}{2}\left[\partial_t a_m + \alpha a_m (m\pi)^2\right] \end{aligned} \tag{5.23}$$

Ha az időderiváltat egyszerű véges differencia módszerrel diszkretizáljuk, akkor

$$a_m^{n+1} = a_m^n - \Delta t \left[\alpha (m\pi)^2 + \frac{32\alpha}{m\pi} \mathrm{mod}(m,2)\right] \tag{5.24}$$

ahol $\mathrm{mod}(m,2)$ az m kettő szerinti maradéka, azaz 0 páros m-re és 1 páratlan m-re. Ez nagyon hatékony módszert ad, mivel a pontosság nagyon gyorsan nő a bázis függvények számával, ugyanakkor a diszkrét egyenlet még egyszerűbb mint a többi tárgyalt módszernél.

Azonban itt erősen kihasználtuk, hogy a kezdeti és határfeltételek egyszerű analitikus formában adottak, valamint a PDE linearitását. A nem

lineáris tagok konvolúció(ka)t adnak, míg a bonyolult határfeltételek az amplitúdók között jelentenek komplikált kapcsolatot. Nem lineáris problémákra a spektrális módszer túl költségessé válik.

5.3.1. Pszeudospektrális Módszer

A megoldás mind spektráltérben, mind fizikai térben diszkretizált, az utóbbinál kollokációs módszerrel, azaz rácspontokban adott a megoldás. A két diszkretizáció között valamilyen FFT típusú algoritmussal lehet konvertálni. A lineáris térbeli differenciaoperátorokat spektrális térben számoljuk ki, a nemlineáris tagokat és a határfeltételeket a fizikai térben.

Például a

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \tag{5.25}$$

Burgers egyenlet esetén a pszeudospektrális módszer az alábbi lépésekből áll

1. FFT-vel u_j^n-ből határozzuk meg a_k^n amplitúdókat
2. Spektrális térben számítsuk ki a $\partial_x u$ amplitúdóit
3. FFT-vel határozzuk meg $(\partial_x u)_j^n$-t a fizikai térben
4. Fizika térben $u_j^{n+1} = u_j^n + \Delta t u_j^n (\partial_x u)_j^n$

5.4. Összefoglaló

Módszer	véges diff.	véges térf.	véges elem	pszeudo-spektr.
tartó	lokális	lokális	lokális	globális
rend	1, 2,...	1, 2, ?	2, 3,...	sima fv-re jó
kozervatív	?	igen	?	???
görbevonalú rács	igen	igen	igen	nem
szabálytalan rács	?	igen	igen	nem
komplexitás	legkisebb	kicsi	közepes	nagy

6. fejezet

Rács Típusok

6.1. Statikus rácsok

A statikus rácsok az időlépéstől függetlenek.

6.1.1. Szabályos rácsok

Descartes rács

$$x = x_0 + i\Delta x \tag{6.1}$$
$$y = y_0 + j\Delta y \tag{6.2}$$
$$z = z_0 + k\Delta z \tag{6.3}$$

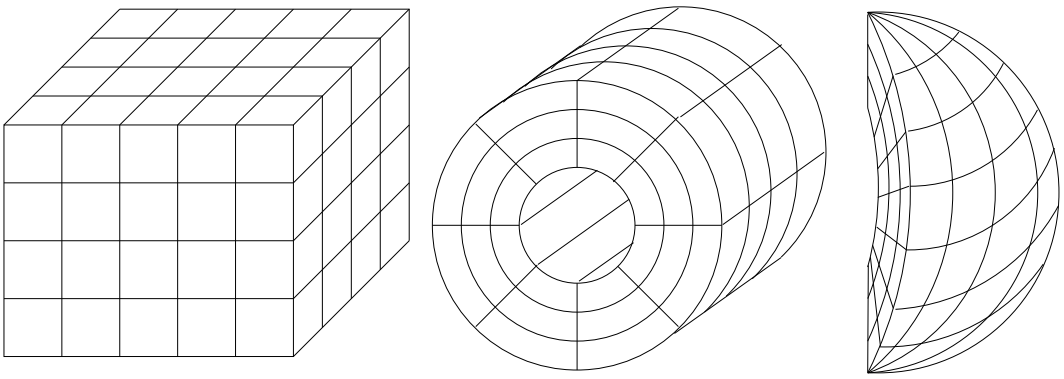
Polár koordinátás rács

$$r = r_0 + i\Delta r \tag{6.4}$$
$$\phi = \phi_0 + j\Delta\phi \tag{6.5}$$
$$z = z_0 + k\Delta z \tag{6.6}$$

a független térbeli változók, és a PDE-t ezekben oldjuk meg. Az $r = 0$ tengely mentén a polár koordináták szingulárisak. Véges felbontás esetén a tengely körüli rácscellák nagyon megnyúlnak, ami stabilitási szempontból nem szerencsés.

Gömb koordinátás rács

$$r = r_0 + i\Delta r \tag{6.7}$$



6.1. ábra. Descartes, polár és gömbkoordinátás szabalyos rácsok

$$\phi = \phi_0 + j\Delta\phi \tag{6.8}$$
$$\theta = \theta_0 + k\Delta\theta \tag{6.9}$$

a független térbeli változók, és a PDE-t ezekben oldjuk meg. A $\theta = 0$, $\theta = \pi$ tengely mentén a rács szinguláris. Véges felbontásnál a cellák elnyúltakká válnak.

6.1.2. Strukturált rács

Nem egyenletes rács

A rácsállandó

$$\Delta x_{i+1/2} = x_{i+1} - x_i \tag{6.10}$$

nem konstans. Ez alkalmas arra, hogy a rácsfelbontást növeljük egy adott helyen, de több dimenzióban nem feltétlenül hatékony.

Általánosított koordináták

$$x = x(i\Delta\xi, j\Delta\eta, k\Delta\zeta) \tag{6.11}$$
$$y = y(i\Delta\xi, j\Delta\eta, k\Delta\zeta) \tag{6.12}$$
$$z = z(i\Delta\xi, j\Delta\eta, k\Delta\zeta) \tag{6.13}$$

ahol x, y, z folytonos függvényei a ξ, η, ζ általánosított koordinátáknak. A PDE-t is ξ, η, ζ függvényében írjuk fel. A ξ, η, ζ koordinátákban a számítási tartomány egy szabályos rácson helyezkedik el.

Görbevonalú rács

$$x = x(i,j,k) \tag{6.14}$$
$$y = y(i,j,k) \tag{6.15}$$
$$z = z(i,j,k) \tag{6.16}$$

Az x, y, z rácsban az (i,j,k) ponthoz az $(i \pm 1, j, k)$, $(i, j \pm 1, k)$, $(i, j, k \pm 1)$ rácspontok esnek a legközelebb. A PDE-t az x, y, z változókban oldjuk meg.

6.1.3. Strukturálatlan rács

A rácspontok sorrendje tetszőleges, koordinátáikat a $\mathbf{x} = \mathbf{x}(i)$ adja meg. A rácspontok szomszédait valamilyen adatstruktúrával kell leírni. Két dimenzióban háromszög rácsot szokás alkalmazni, három dimenzióban tetraéderest, de más rácstípusok is léteznek. Strukturálatlan rácsoknál a rács generálás bonyolult feladatot jelent.

6.2. Dinamikus rácsok

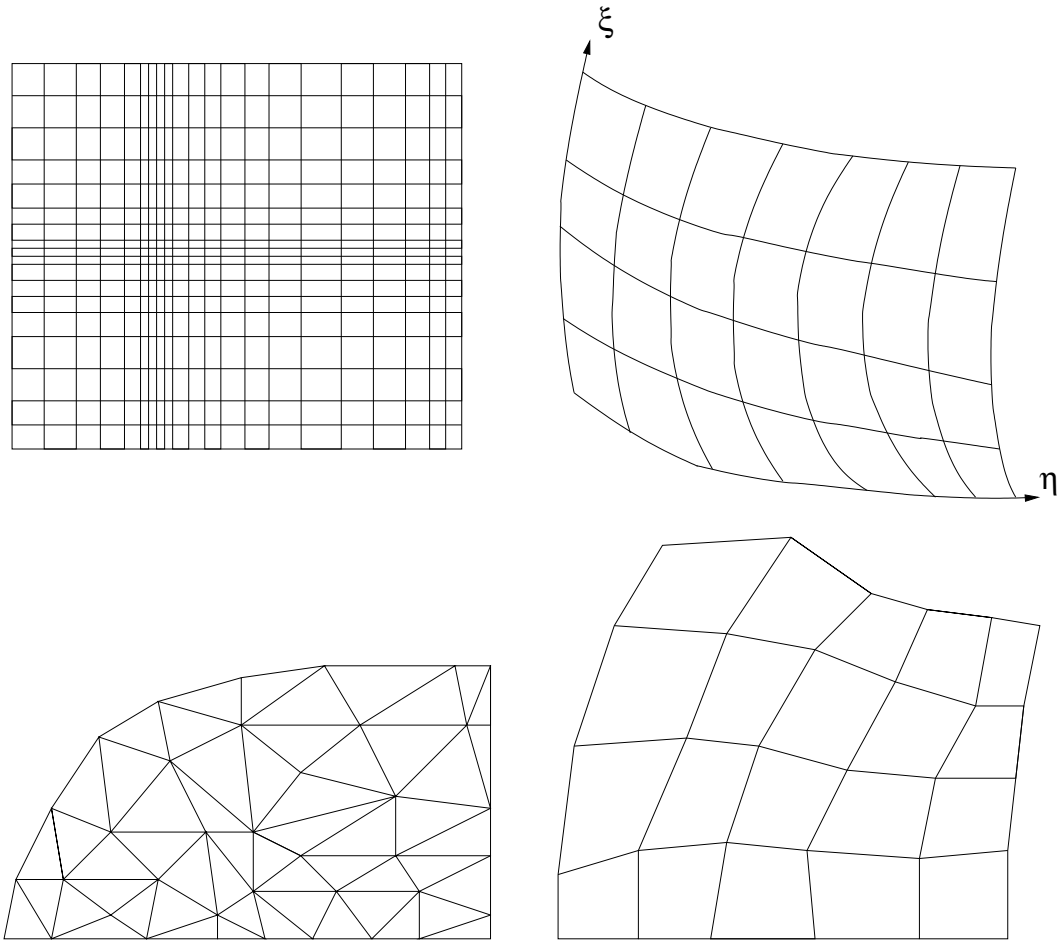
Egyes problémék megkívánják, hogy a rács változzon a szimuláció folyamán. Ez szükséges lehet a határok mozgása miatt, vagy azért, hogy a rácsfelbontást dinamikusan tudjuk változtatni.

6.2.1. Mozgó rács

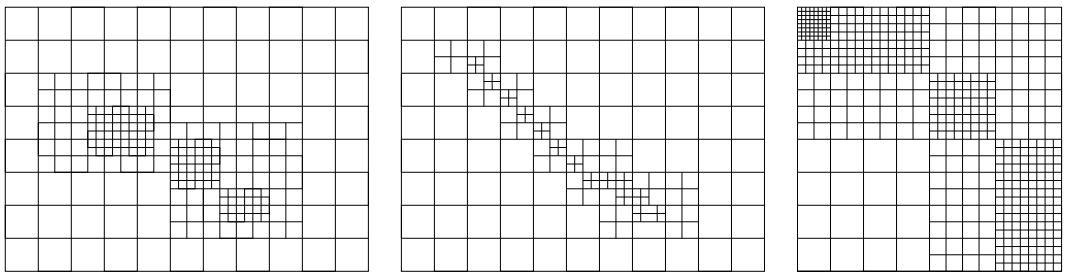
Egy dimenzióban nem konstans rácsállandójú, több dimenzióban általánosított koordinátás vagy görbevonalú rácsot használunk. A rácspontok időben folytonosan mozognak $v_R(i,j,k)$ sebességgel. Konvekció esetén például Lagrange leírást használhatunk. Az egyenletek a rács mozgása miatt extra tagokkal bővülnek. Gyakran a rácspontok sebességét is valamilyen PDE megoldásaként határozzuk meg. Több dimenzióban nehéz a rács feltekeredését elkerülni.

6.2.2. Adaptív rács finomítás

Az adaptív rácsok finomításnál (Adaptive Mesh Refinement, AMR) a rács szerkezete változik időben, maguk a rácspontok nem mozognak.



6.2. ábra.
Strukturált rácsok: nem egyenletes rácsállandó (bal fent), általánosított koordináták (jobb fent), görbevonalú rács (jobb lent). Strukturálatlan háromszögrács (bal lent).



6.3. ábra.
Hierarchikus rács (bal oldalt), cellánként finomított AMR rács (középen), blokk AMR rács (jobb oldalt)

Hierarchikus rács

A hierarchikus rácsban több különböző rácsállandójú – általában szabályos – rács fekszik egymáson. A PDE-t mindegyik rácson megoldjuk. Egy-egy rács határfeltételeit a nála durvább illetve a szomszédos azonos finomságú rácsok biztosítják. Az időlépés minden rácson különböző (lehet). Általában az időlépések és a térbeli felbontások aránya is kis egész páros számok, pl. 2, 4 stb.

A megmaradási tételek teljesítésére vigyázni kell, azaz a különböző szinten számolt fluxusoknak meg kell egyezniük a finom és durva rács határán.

Strukturálatltan adaptív rács

Ha minden egyes cellát finomíthatunk és durvíthatunk, akkor egy strukturálatlan rácsot kapunk. Ez nagyon bonyolult tértartományok leírására is alkalmas. Viszonylag bonyolult adatszerkezetet igényel, és párhuzamos futtatásnál nem triviális a processzorok közti ideális munkamegosztást kialalkítani.

Blokk adaptív rács

A blokk adaptív rácsban egyes blokkokat finomítunk és durvítunk tipikusan egy 2-es faktorral. A blokkok nem fednek át, hanem pontosan kitöltik a teret. Az időlépés lehet különböző, de jóval könnyebb megcsinálni egyformára. Párhuzamos programokban nagyon hatékony, mert a sok egyforma számú cellából álló blokkot könnyű szétosztani a processzorok között.

6.3. Határfeltételek

6.3.1. Szellem cellák

A határfeltételeket legegyszerűbb szellemcellákkal leírni. Valamilyen módon meghatározzuk, hogy mennyi lenne a függőváltozók értéke a szellem cellákban. A szellem cellák biztosítják, hogy a diszkretizáció tartója minden fizikai cellára ismert legyen. Például ha két szellemcellára van szükség, az $i = 0$ és $i = -1$ indexű cellákban a következő értékeket írjuk elő a határfeltétel típusától függően:

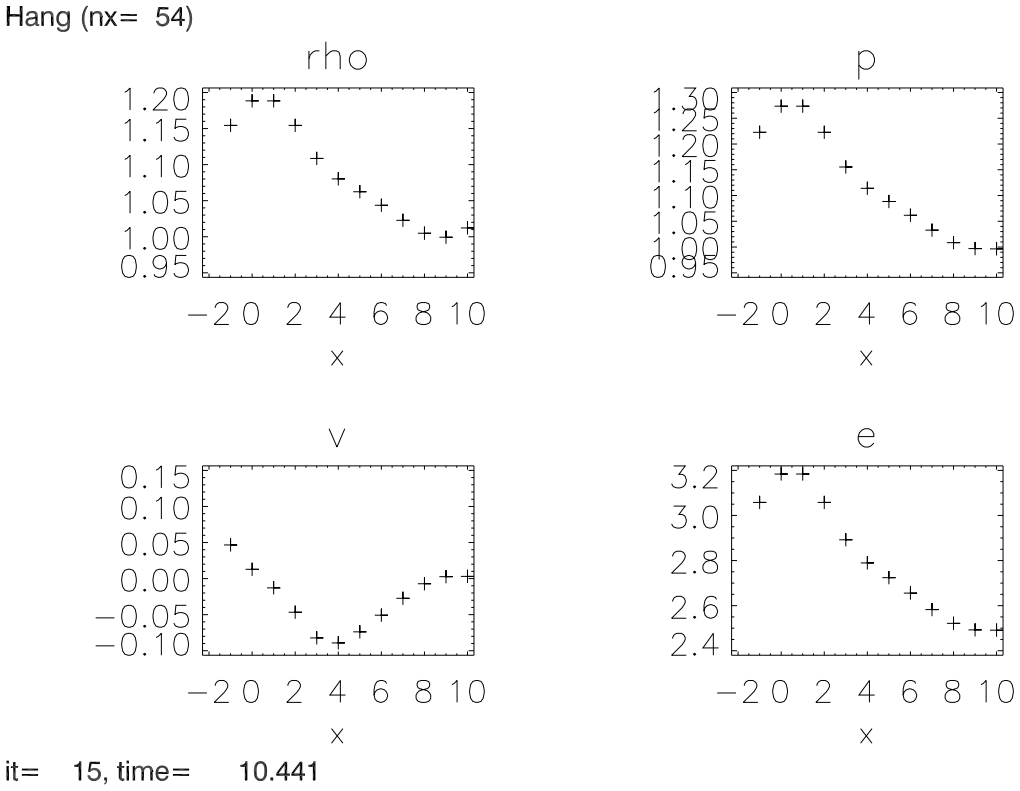
- periodikus: $T_0 = T_J$, $T_{-1} = T_{J-1}$
- rögzített (Dirichlet): $T_0 = T_{-1} = T_h$
- folytonos (Neumann): $T_0 = T_{-1} = T_1$
- szimmetrikus: $T_0 = T_1$, $T_{-1} = T_2$
- antiszimmetrikus: $T_0 = -T_1$, $T_{-1} = -T_2$

6.3.2. Fluxus határfeltételek

Véges térfogat módszernél a cellahatáron vett fluxusokat használjuk. A határfeltételeket úgy is megadhatjuk, hogy a határon lévő cellákhoz megadjuk a fluxusokat. A szellem cellákkal a fluxus határfeltételek egy részhalmaza valósítható meg.

6.3.3. Speciális határdiszkretizáció

Elvileg a határon használhatunk teljesen más diszkretizációt mint a számítási tartomány belsejében. Ez nagyon rugalmas megoldás egy konkrét problémára, de elég bonyolult programozási szempontból, és nem használható általános programokban. Példa ilyen speciális diszkretizációra a karakterisztikus határfeltételek alkalmazása, ahol a határon az egyenlet karakterisztikus hullámait számítjuk ki, és ezek alapján számítjuk ki a határfeltételeket.



6.4. ábra.
Reflektív határfeltétel megvalósítása egy hidrodinamikai szimulációban szellemcellák segítségével. A bal szélső két rácspont tartozik a szellemcellákhoz. A ρ sűrűségre és a p nyomásra szimmetrikus, a határra merőleges v sebességre antiszimmetrikus határfeltételeket alkalmazunk.

7. fejezet

Explicit időintegrálási módszerek

Eddig szinte mindig a legegyszerűbb időben elsőrendű időben előrehaladó (forward in time) diszkretizációt használtuk az időderivált explicit diszkretizálásához. Ezt Euler lépésnek is nevezik. Ebben a részben a magasabb rendű pontosság elérésére alkalmas módszereket tekintjük át. A jelölés egyszerűsítése miatt bevezetjük a

$$\frac{\partial U}{\partial t} = S - \mathrm{div} F = R \tag{7.1}$$

jelölést, ahol R a jobb oldalon lévő fluxusok és forrás tagok összege. Ebben a jelölésben az Euler lépés

$$U^{n+1} = U^n + \Delta t R(U^n) \tag{7.2}$$

7.1. Runge-Kutta

A Runge-Kutta mószerekben több részlépésben becsüljük R-t, és ezeknek a becsléseknek valamilyen lineáris kombinációját alkalmazzuk.

Másodrendű Runge-Kutta:

$$U^{n+1/2} = U^n + \frac{1}{2}\Delta t R(U^n) \tag{7.3}$$

$$U^{n+1} = U^n + \Delta t R(U^{n+1/2}) \tag{7.4}$$

Negyedrendű Runge-Kutta:

$$U_1 = U^n + \frac{1}{2}\Delta t R(U^n) \tag{7.5}$$

$$U_2 = U^n + \frac{1}{2}\Delta t R(U_1) \tag{7.6}$$

$$U_3 = U^n + \Delta t R(U_2) \tag{7.7}$$

$$U^{n+1} = U^n + \Delta t \left[\frac{1}{6}R(U^n) + \frac{1}{3}R(U_1) + \frac{1}{3}R(U_2) + \frac{1}{6}R(U_3)\right] \tag{7.8}$$

Programozási szempontból valamivel célszerűbb, ha csak az U-kat kell tárolni, az R-ket nem. A fenti utolsó egyenlet helyett ezért érdemes az alábbi ekvivalens képleteket használni

$$U_4 = U^n + \frac{1}{6}\Delta t R(U_3) \tag{7.9}$$

$$U^{n+1} = U_4 + \frac{1}{3}\left(U_1 + 2U_2 + U_3 - 4U^n\right) \tag{7.10}$$

7.2. Prediktor-korrektor

A másodrendű prediktor-korrektor módszer lényegében egy másodrendű Runge-Kutta, de más (egyszerűbb) diszkretizációt használ az első prediktor lépésben, mint a második korrektor lépésben.

7.3. Több szintű idődiszkretizáció

Ha nem csak az n-dik és $n+1$-dik lépéseket használjuk fel, akkor az idődiszkretizáció rendje növelhető. Peldául egy háromszintű másodrendű idődiszkretizáció a következő:

$$U^{n+1} = U^{n-1} + 2\Delta t R(U^n) \tag{7.11}$$

Használhatjuk a korábbi időlépés jobb oldalát is

$$U^{n+1} = U^n + \frac{3}{2}\Delta t R(U^n) - \frac{1}{2}\Delta t R(U^{n-1}) \tag{7.12}$$

Mint a Taylor sorfejtésből látható, ez szintén időben másodrendű diszkretizáció.

A többszintü módszerek hátránya, hogy el kell tárolni a korábbi időlépéseket, és az első időlépésben 2 szintű formulát kell alkalmazni. Ha az időlépések hossza nem állandó, akkor kissé elbonyolódnak a formulák.

7.4. Operátor bontás

Tegyük fel, hogy van egy jól működő módszerünk/programunk egy 1 térdimenziós PDE megoldására. Hogyan általánosíthatjuk a programot/módszert több dimenzióra?

Legyen L_x az x irányú másodrendű diszkretizáció ami megoldja a

$$\partial_t U = R_x(U) \tag{7.13}$$

egyenletrendszert. Az egyszerűség kedvéért tegyük fel továbbá, hogy R_x egy lineáris operátor. A Taylor sorfejtés miatt

$$\begin{aligned} U^{n+1} &= L_x(\Delta t)U^n = U^n + \Delta t\partial_t U + \frac{1}{2}\Delta t^2\partial_{tt}U + O(\Delta t^3) \\ &= (I + \Delta t R_x + \frac{1}{2}\Delta t^2 R_x^2)U^n + O(\Delta t^3) \end{aligned} \tag{7.14}$$

Hasonlóan L_y a $\partial_t U = R_y$ másodrendű diszkretizációja. Ekkor

$$\begin{aligned} U^{n+1} &= L_x(\Delta t)L_y(\Delta t)U^n \\ &= (I + \Delta t R_x + \frac{1}{2}\Delta t^2 R_x^2)(I + \Delta t R_y + \frac{1}{2}\Delta t^2 R_y^2)U^n + O(\Delta t^3) \\ &= \left[I + \Delta t(R_x + R_y) + \frac{1}{2}\Delta t^2(R_x^2 + R_y^2 + 2R_xR_y)\right]U^n + O(\Delta t^3) \end{aligned} \tag{7.15}$$

ami nem másodrendű, mert R_x és R_y általában nem felcserélhető operátorok, és így

$$(R_x + R_y)^2 = R_x^2 + R_xR_y + R_yR_x + R_y^2 \neq R_x^2 + 2R_xR_y + R_y^2 \tag{7.16}$$

Viszont, ha alkalmazzuk a **Godunov féle operátor bontást**, és minden második lépésben megcseréljük a sorrendet, akkor már másodrendű lesz a diszkretizáció:

$$\begin{aligned} U^{n+2} &= L_x(\Delta t)L_y(\Delta t)L_y(\Delta t)L_x(\Delta t)U^n = \dots \\ &= \left[I + 2\Delta t(R_x + R_y) + \frac{1}{2}(2\Delta t)^2(R_x + R_y)^2\right]U^n + O(\Delta t^3) \end{aligned} \tag{7.17}$$

Természetesen itt feltételeztük, hogy az időlépés nem, vagy nem sokat változik az n-dik és az $n+1$-dik lépések között.

A **Strang féle operátor bontás** a Godunov-hoz hasonlóan működik, azonban egyetlen időlépést bont fel:

$$\begin{aligned} U^{n+1} &= L_x(\Delta t/2)L_y(\Delta t)L_x(\Delta t/2)U^n = \dots \\ &= \left[I + \Delta t(R_x + R_y) + \frac{1}{2}\Delta t^2(R_x + R_y)^2\right]U^n + O(\Delta t^3) \end{aligned} \tag{7.18}$$

A Strang féle operátor bontásban három részoperátort kell alkalmazni időlépésenként, ami 50%-kal drágább mint a Godunov féle operátor bontás.

Rekurzív módszerrel megmutatható, hogy az operátor bontás $m > 2$ operátor esetén is alkalmazható. Legyen $R_y = R_v + R_w$, és ezek diszkrét megfelelői az L_v és L_w másodrendű operátorok. Ekkor a Godunov operátor bontást kétszer alkalmazva

$$\begin{aligned} U^{n+2} &= L_x(\Delta t)L_y(\Delta t)L_y(\Delta t)L_x(\Delta t)U^n \\ &= L_x(\Delta t)L_y(2\Delta t)L_x(\Delta t)U^n + O(\Delta t^3) \\ &= L_x(\Delta t)L_v(\Delta t)L_w(\Delta t)L_w(\Delta t)L_v(\Delta t)L_x(\Delta t)U^n + O(\Delta t^3) \end{aligned} \tag{7.19}$$

illetve hasonlóan a Strang féle operátor bontásnál

$$U^{n+1} = L_x(\Delta t/2)L_v(\Delta t/2)L_w(\Delta t)L_v(\Delta t/2)L_x(\Delta t/2)U^n + O(\Delta t^3) \tag{7.20}$$

Ez a lépés tetszőlegesen sok operátorra megismételhető. Fontos megjegyezni, hogy az operátor bontás módszere nem szorítkozik a különböző irányokban vett fluxusokra, de éppúgy alkalmazható például forrástagok és fluxusok bontott diszkretizációjára is.

Gyenge megoldásokra vonatkozóan csak skalár egyenletekre és úgynevezett monoton módszerekre sikerült bebizonyítani [2], hogy az operátor bontással kapott diszkretizáció helyes gyenge megoldást ad két dimenzióban is.

8. fejezet

Implicit időintegrálási módszerek

Implicit időintegrálásra akkor van szükség, amikor az explicit módszer nem gazdaságos a numerikus stabilitás által limitált kis időlépés miatt. Természetesen, ha a kis időlépést a pontosság követeli meg, akkor nem érdemes implicit időintegrálást használni.

8.1. Implicit diszkretizációk

Ismét a

$$\frac{\partial U}{\partial t} = R \tag{8.1}$$

egyenletet oldjuk meg. Elsőrendű **fordított (backwards) Euler lépés**

$$U^{n+1} = U^n + \Delta t R(U^{n+1}) \tag{8.2}$$

Másodrendű **trapéz módszer**

$$U^{n+1} = U^n + \Delta t \left[\frac{1}{2}R(U^{n+1}) + \frac{1}{2}R(U^n)\right] \tag{8.3}$$

A trapéz módszer lineáris PDE-re kiválóan működik, nem lineáris PDE-re viszont általában instabil, ha az időlépés nagy. Ez javítható, ha növeljük az implicit rész arányát:

$$U^{n+1} = U^n + \Delta t \left[\beta R(U^{n+1}) + (1-\beta)R(U^n)\right] \tag{8.4}$$

ahol $\beta > 1/2$ esetén a módszer stabillá válik. Formálisan elérhető a másodrendű pontosság, ha $\beta = 1/2 + \kappa\Delta t$, azaz végtelen kis időlépésre a $\beta \to 1/2$. Véges felbontásnál azonban ez inkább csak önmagunk megnyugtatására jó, hiszen κ-t úgy kell megválasztanunk, hogy a módszer stabil maradjon.

Három időszint felhasználásával konstruálható nem lineárisan is stabil és másodrendű módszer. Ilyen például a BDF2 (backwards difference formula 2):

$$U^{n+1} = U^n + \Delta t_n \left[\beta R(U^{n+1}) + (1-\beta) R(U^n)\right] + \Delta t_n \alpha \left[\frac{U^n - U^{n-1}}{\Delta t_{n-1}} - R(U^n)\right] \tag{8.5}$$

Itt $\alpha = \Delta t_n / (\Delta t_n + 2\Delta t_{n-1})$ és $\beta = 1 - \alpha$ választással a BDF2 másodrendű diszkretizációt kapjuk. Ha az időlépés konstans, akkor $\alpha = 1/3$ és $\beta = 2/3$, azaz az implicit rész dominál, ami stabilabb mint a trapéz módszer ($\beta = 1/2, \alpha = 0$).

A háromszintű módszerek hátránya, hogy több adatot kell tárolni, illetve az első időlépésben nem áll rendelkezésre mindhárom szint. Ez utóbbi problémán könnyű segíteni: az első lépésben egy kétszintű módszert, pl. a trapéz módszert lehet használni.

8.1.1. Alacsonyabb rendű implicit diszkretizáció

A megoldandó egyenletrendszer egyszerűsítése érdekében célszerű az $R(U^{n+1})$ tagot minél egyszerűbb formában diszkretizálni. Megmutatható, hogy ha például a trapéz módszert átírjuk a

$$U^{n+1} = U^n + \Delta t R_2(U^n) + \frac{\Delta t}{2}\left[R_1(U^{n+1}) - R_1(U^n)\right] \tag{8.6}$$

alakba, ahol R_2 egy térben másodrendű, míg R_1 egy térben első rendű diszkretizáció. Megmutatjuk, hogy az egész diszkretizáció térben és időben másodrendű lesz, ugyanis a hiba

$$\begin{aligned} U^{n+1} &= U^n + \Delta t R_2(U^n) + \frac{\Delta t}{2}\left[\frac{dR_1}{dt}\Delta t + \mathcal{O}(\Delta t^2)\right] \\ &= U^n + \Delta t R_2(U^n) + \frac{\Delta t^2}{2}\frac{dR_2}{dt} + \mathcal{O}(\Delta x \Delta t^2) + \mathcal{O}(\Delta t^3), \end{aligned} \tag{8.7}$$

ahol kihasználtuk, hogy $\dot{R}_2 = \dot{R}_1 + O(\Delta x)$.

8.2. Szemi-implicit módszerek

Nagyon sokféle módszert neveznek szemi-implicitnek:

- A PDE egyes tagjai implicitek
- A PDE egyes változói implicitek

- A számítási tartomány egyes részeiben használunk implicit diszkretizációt

A szemi-implicit módszerek az explicit és az implicit módszerek előnyeit próbálják egyesíteni: ahol lehet az olcsóbb explicit módszert használják, és a drágább de stabil implicit módszert csak akkor és ott használják ahol szükséges. Persze a szemi-implicit módszerek erősen egyenlet és problémafüggők, a stabilitási kritériumokat is nehéz megállapítani.

Például tegyük fel, hogy a

$$\frac{\partial U}{\partial t} = R = R_{\text{expl}} + R_{\text{impl}} \tag{8.8}$$

PDE-ben csak az R_{impl} tagokat kívájuk implicit módon kezelni. Ekkor például a következő diszkretizáció másodrendű lesz:

$$U^{n+1} = U^n + \Delta t R_{\text{expl}} \left(U^n + \frac{\Delta t}{2} R^n \right) + \frac{\Delta t}{2} \left[R_{\text{impl}}(U^n) + R_{\text{impl}}(U^{n+1}) \right] \tag{8.9}$$

Egy másik megközelítés a Godunov vagy Strang féle operátor bontás lehet, ahol az explicit és implicit rész diszkretizációja teljesen szétválik.

8.3. Nem lineáris egyenletrendszer megoldása

Keressük az

$$F(U) = 0 \tag{8.10}$$

egyenletrendszer U megoldását, ahol F tetszőleges nemlineáris függvénye az U vektornak. Általában nagyon nehéz megtalálni egy nem lineáris egyenletrendszer gyökeit [1]. Az egyik leghatékonyabb megoldási módszer a **Newton-Raphson** iteráció, illetve ennek javított változatai.

Legyen a kezdeti vektor U_0, és az iteráció a k-dik lépésben

$$U_{k+1} = U_k - \left(\frac{\partial F}{\partial U} \right)^{-1} F(U_k) \tag{8.11}$$

Itt $\partial F / \partial U$ a **Jacobi mátrix**. Ha a módszer konvergál, és már eléggé közel vagyunk a megoldáshoz, akkor megmutatható, hogy a hiba minden lépésben négyzetesen csökken, azaz

$$||U_{k+1} - U_\infty|| \propto ||U_k - U_\infty||^2 \tag{8.12}$$

ahol U_∞ a konvergált megoldás. Sajnos egyáltalán nincs rá garancia, hogy a Newton-Raphson módszer konvergálna.

A konvergencia sugarat **vonalmenti kereséssel** lehet növelni. Az eljárás lényege a következő: egy Newton lépésben előfordulhat, hogy $||F(U_{k+1})||$ nem kisebb $||F(U_k)||$ kezdeti hibánál. Ugyanakkor megmutatható, hogy az $f = F^2/2$ hibafüggvény csökken a $-J^{-1}F(U_k)$ irányban, hiszen

$$\mathbf{grad}f \cdot (-J^{-1}F) = -(F.J)(J^{-1}F) = -F^2 < 0 \tag{8.13}$$

Tehát ha megfelelően választott kisebb lépést teszünk ebben az irányban, akkor a hiba garantáltan csökkenni fog:

$$U_{k+1} = U_k - \lambda \left(\frac{\partial F}{\partial U}\right)^{-1} F(U_k) \tag{8.14}$$

A $0 < \lambda < 1$ együttható megfelelő megválasztásával a konvergencia lényegében biztossá tehető.

8.4. Implicit diszkretizáció linearizálása

Mint látható, a Newton-Raphson módszerben minden iterációjában egy lineáris egyenletrendszert kell megoldani. Célszerű a nemlineáris egyenletrendszert úgy linearizálni, hogy a diszkretizáció rendje ne csökkenjen, mert így csak egy lineáris egyenletrendszert kell megoldani. Ezt könnyen megtehetjük, ha az ismeretlen R^{n+1}-t

$$\begin{aligned} R(U^{n+1}) &= R(U^n) + \Delta t \frac{\partial R}{\partial t} + O(\Delta t^2) \\ &= R(U^n) + \Delta t \frac{\partial R}{\partial U}\frac{\partial U}{\partial t} + O(\Delta t^2) \\ &= R(U^n) + \frac{\partial R}{\partial U}\left(U^{n+1} - U^n\right) + O(\Delta t^2) \end{aligned} \tag{8.15}$$

alakban írjuk fel. Ezután tekintsük ismeretlennek a

$$\Delta U = U^{n+1} - U^n \tag{8.16}$$

vektort, és ebben írjuk fel az impliciten diszkretizált egyenletet. Például a trapéz módszer

$$U^{n+1} = U^n + \Delta t R_2(U^n) + \frac{\Delta t}{2}\left[R_1(U^{n+1}) - R_1(U^n)\right] \tag{8.17}$$

linearizált formában

$$\Delta U = \Delta t R_2(U^n) + \frac{\Delta t}{2}\frac{\partial R_1}{\partial U}\Delta U + O(\Delta t^2) \tag{8.18}$$

ami tehát másodrendű pontos. Ez átrendezhető az

$$\left(I - \frac{1}{2}\Delta t \frac{\partial R_1}{\partial U}\right) \Delta U = \Delta t R_2(U^n) \tag{8.19}$$

alakra, ami egy lineáris egyenletrendszer ΔU-ra nézve. Ez az egyenletrendszer megegyezik a Newton-Raphson eljárás első iterációjával, ha a kezdeti vektor $\Delta U_0 = 0$.

Ha a linearizált implicit diszkretizáció stabilitása kielégítő, akkor mindig hatékonyabb mint a nem-lineáris diszkretizáció pontos megoldása, mert csak egy lineáris egyenletrendszert kell megoldani, és a megoldás rendje ugyanolyan mindkét esetben. Előfordulhat, hogy a nem-lineáris rendszer stabilabb, és érdemes egynél több Newton iterációt végrehajtani.

8.5. Jacobi mátrix meghatározása

Akár a nemlineáris, akár a lineáris problémát tekintjük, ki kell számítani a

$$J = I - \beta \Delta t \frac{\partial R_1}{\partial U} \tag{8.20}$$

mátrix elemeit, vagy legalábbis a J mátrix inverzének és egy tetszőleges vektornak a szorzatatát.

8.5.1. Jacobi mátrix analitikusan

A legnyilvánvalóbb, de nem feltétlenül a legegyszerűbb vagy éppen leghatékonyabb megközelítés a Jacobi mátrix elemeinek analitikus kiszámítása. Fontos megjegyezni, hogy itt a diszkrétizált R illetve R_1 függvények parciális deriváltjaira van szükség és nem az analitikus PDE-ben szereplő jobb oldalra. Ha a diszkretizáció nagyon egyszerű, pl. centrális differencia, akkor az elemek analitikusan kiszámíthatóak. Például

$$R_i = S_i - \frac{F_{i+1} - F_{i-1}}{2\Delta x} \tag{8.21}$$

esetén

$$\begin{aligned} \frac{\partial R_i}{\partial U_i} &= \left(\frac{\partial S}{\partial U}\right)_i \\ \frac{\partial R_i}{\partial U_{i+1}} &= -\frac{1}{2\Delta x}\left(\frac{\partial F}{\partial U}\right)_{i+1} \\ \frac{\partial R_i}{\partial U_{i-1}} &= +\frac{1}{2\Delta x}\left(\frac{\partial F}{\partial U}\right)_{i-1} \end{aligned} \tag{8.22}$$

Bonyolultabb diszkretizáció esetén sokkal több tag lép fel. Az egyes parciális deriváltakat kiszámíthatjuk analitikusan vagy numerikusan:

$$\frac{\partial f}{\partial U_w} = \frac{f(U + \epsilon\delta^w) - f(U)}{\epsilon}. \tag{8.23}$$

ahol f az F fluxus vagy az S forrás vektorok valamelyik komponense, míg w az U vektor egy komponensét jelöli. Az $\epsilon\delta^w$ perturbáció csak az U_w komponenst perturbálja ϵ-nal.

A perturbáció mértékét úgy kell megválasztani, hogy a κ kerekítési (számábrázolási) hiba hatását minimalizáljuk, ugyanakkor a deriváltat minél pontosabban közelítsük, azaz az $O(\epsilon)$ diszkretizációs hiba minél kisebb legyen. A számábrázolási hiba miatt legyen a perturbálatlan $f(U)$-ban egy $-f(U)\kappa$ hiba, így a hiba vezető rendben

$$\frac{f(U + \epsilon\delta^w) - f(U)(1-\kappa)}{\epsilon} - \frac{\partial f}{\partial U_w} = \frac{\epsilon}{2}\frac{\partial^2 f}{\partial U_w^2} + f(U)\frac{\kappa}{\epsilon} \tag{8.24}$$

aminek a minimum helye az

$$\epsilon = \sqrt{\frac{2\kappa f}{f''}} \approx \sqrt{\kappa}||U_w|| \tag{8.25}$$

Természetesen f/f''-t nem ismerjük, csupán nagyságrendileg becsültük U_w^2-tel. Dupla pontos számábrázolásnál például $\kappa \approx 10^{-12}$.

8.5.2. Jacobi mátrix numerikusan

Ha a diszkretizáció bonyolult, akkor az előző részben tárgyalt képletek nagyon elbonyolódhatnak. A mátrix elemeit numerikusan is meg tudjuk becsülni, ha az ismeretleneket egyesével perturbáljuk:

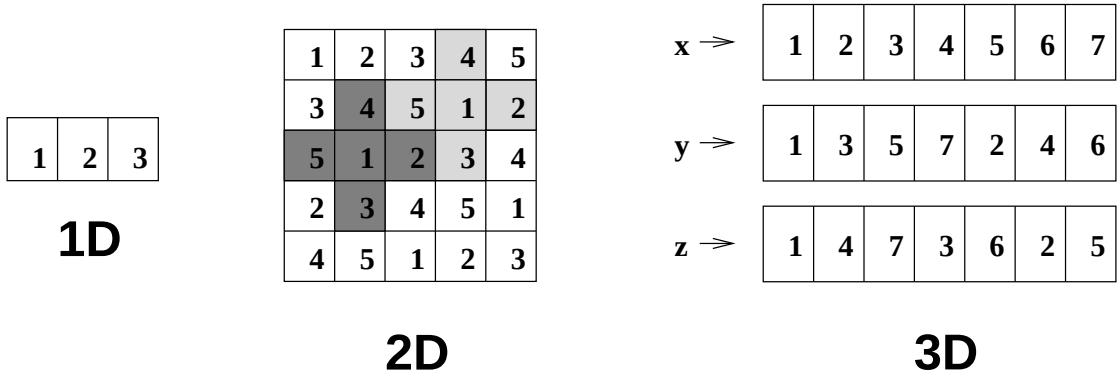
$$\frac{\partial R_i^u}{\partial U_j^w} = \frac{R_i^u(U + \epsilon\delta_j^w) - R_i^u(U)}{\epsilon}. \tag{8.26}$$

ahol u és w az R illetve U megfelelő komponenseit jelölik. A $\epsilon\delta_j^w$ perturbáció csak a j-dik rácspontban hat az U-nak a w komponensére.

8.5.3. Mátrix mentes módszer

Bizonyos iteratív módszerek nem igénylik az invertálandó mátrix elemeit, elegendő ha a mátrixszal meg tudunk szorozni egy tetszőleges vektort. Mivel az invertálandó mátrix speciális alakú, ez megtehető a következő módon:

$$\left(I - \beta\Delta t\frac{\partial \vec{R}}{\partial \vec{U}}\right) a = a - \beta\Delta t\frac{R(U^n + \epsilon a) - R(U^n)}{\epsilon} \tag{8.27}$$



8.1. ábra.
A Jacobi mátrix numerikus kiszámításánál úgy kell az egyes cellákban a változókat perturbálni, hogy a tartók ne fedjenek át. Az ábrán az 1, 2 és 3 dimenziós perturbációs minták láthatók arra az esetre, amikor a tartók a legközelebbi cellákra terjednek ki. Az azonos számmal jelölt cellák perturbálhatók egyszerre.

Mint látható a mátrix helyett csak az R-t kell kiszámolni. Az $R(u^n)$ független a-tól, így tulajdonképpen minden mátrix-vektor szorzáshoz az R-t egyszer kell kiértékelni. Az ϵ paraméterre egy célszerű érték a $\sqrt{\kappa}||U||/||a||$.

9. fejezet

Lineáris egyenletrendszerek megoldása

Mint a korábbi fejezetekben láttuk, egyenletrendszerek megoldására szükség van az alábbi esetekben:

- elliptikus PDE diszkretizációja
- véges elem módszer
- implicit időintegrálás

A nem-lineáris egyenletrendszereket vagy linearizáljuk, vagy Newton-Raphson iterációval oldjuk meg. Ez utóbbinál minden egyes iterációban egy lineáris egyenletrendszert kell megoldani.

A lineáris egyenletrendszert

$$A \cdot V = B \tag{9.1}$$

alakban írhatjuk, ahol V az ismeretlenek vektora, B a jobb oldal vektora, A pedig egy mátrix. A legcélszerűbb megoldási módszer megválasztása leginkább az A mátrix tulajdonságaitól függ:

- A elemeinek nagy része nem 0 – sűrű mátrix
- A elemeinek nagy része 0 – ritka (sparse) mátrix
- A nem 0 elemei a fő diagonálishoz közel esnek – ritka sávos mátrix

A megoldási módszer kiválasztásánál a módszer tulajdonságai is fontosak lehetnek, így például

- számítási igény

- memóriaigény
- parallelizálhatóság
- mátrix elemek szükséges/szükségtelen volta
- adott megoldási pontossághoz szükséges munka

9.1. Direkt módszerek

9.1.1. Gauss elimináció

A Gauss eliminációt kizárólag sűrű mátrixok esetén érdemes alkalmazni. Ilyen mátrixok például a spektrális módszernél léphetnek fel. A Gauss eliminációt a gyakorlatban szinte mindig két lépésben végezzük el

1. LU dekompozíció
2. megoldás

Az LU dekompozícióban az eredeti A mátrixot egy U felső és egy L alsó háromszög mátrix szorzatára bontjuk, azaz $L \cdot U = A$. Az L és U mátrixokat a *Crout* algoritmussal állítjuk elő:

$$L_{i,i} = 1 \tag{9.2}$$

minden $i = 1 \ldots N$-re, majd $j = 1, \ldots, N$ sorrendben

$$U_{i,j} = A_{i,j} - \sum_{k=1}^{i-1} L_{i,k} U_{k,j} \qquad i = 1, 2, \ldots, j \tag{9.3}$$

$$L_{i,j} = \frac{1}{U_{j,j}} \left(A_{i,j} - \sum_{k=1}^{j-1} L_{i,k} U_{k,j} \right) \qquad i = j+1, j+2, \ldots, N \tag{9.4}$$

A két háromszögmátrix éppen elfér az eredeti A mátrix helyén, hiszen az L diagonálisán csupa 1-es áll, amit nem kell tárolni. A tridiagonális mátrixok előállítása $O(N^3)$ műveletet igényel, míg a tridiagonális mátrixok inverzével való szorzás $O(N^2)$-t. Ez akkor lényeges, ha sok egyenletet kell ugyanazzal az A mátrixszal, de különböző B jobb oldalakkal megoldani. A 0-val, vagy kis számokkal való osztást a sorok cseréjével (pivoting) kell elkerülni.

9.1.2. Tridiagonális mátrix

Tridiagonális mátrixokban a nullától különböző elemek mind a főátlón ($A_{i,i}$) illetve ez alatt ($A_{i-1,i}$) és efölött helyezkednek el:

$$\begin{pmatrix} b_1 & c_1 & & & & 0 \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \\ 0 & & & & a_N & b_N \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ \\ v_N \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ \\ d_N \end{pmatrix} \tag{9.5}$$

Ilyen mátrixok lépnek fel pl. a véges elem módszernél vagy az implicit véges differencia és véges térfogat módszereknél. Tridiagonális A mátrix esetén az egyenletrendszert legkönnyebb a Thomas algoritmussal megoldani. Először az $i = 1, 2 \ldots N$ sorrendben (forward sweep) eltüntetjük a diagonális alatti elemeket, és magát a diagonálist egységnyire normalizáljuk:

$$c'_1 = \frac{c_1}{b_1} \tag{9.6}$$

$$d'_1 = \frac{d_1}{b_1} \tag{9.7}$$

$$c'_i = \frac{c_i}{b_i - a_i c'_{i-1}} \tag{9.8}$$

$$d'_i = \frac{d_i - a_i d_{i-1}}{b_i - a_i c'_{i-1}} \tag{9.9}$$

Ennek eredményeképpen a transzformált egyenlet a következő lesz

$$\begin{pmatrix} 1 & c'_1 & & & & 0 \\ & 1 & c'_2 & & & \\ & & 1 & c'_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \\ 0 & & & & & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ \\ v_N \end{pmatrix} = \begin{pmatrix} d'_1 \\ d'_2 \\ d'_3 \\ \vdots \\ \\ d'_N \end{pmatrix} \tag{9.10}$$

Ezután $i = N, N-1, \ldots 1$ sorrendben (backward sweep) megkapjuk a megoldást

$$v_N = d'_N \tag{9.11}$$

$$v_i = d'_i - v_{i+1} c'_i \tag{9.12}$$

A Thomas algoritmus műveletigénye mindössze $\approx 5N$. Lényegében a Gauss eliminációnak a tridiagonális mátrixra egyszerűsített változata.

9.1.3. Sávos mátrix

Szélesebb sávos mátrix magasabb rendű véges elem illetve implicit véges differencia módszereknél lép fel. A Thomas algoritmus könnyen általánosítható szélesebb sávos mátrixokra is. Több forward sweep-pel a mátrixot felső háromszög mátrixra transzformáljuk, majd visszahelyettesítéssel megoldjuk. Persze a műveletigény a sáv M szélességével gyorsan $O(M^2N)$ ütemben nő.

9.1.4. Blokk tridiagonális mátrix

Egyenletrendszerek diszkretizálása esetén a mátrix szerkezete tipikusan blokkokból áll, ahol M egyenlet esetén blokkok $M \times M$ méretűek

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{c}_1 & & & & 0 \\ \mathbf{a}_2 & \mathbf{b}_2 & \mathbf{c}_2 & & & \\ & \mathbf{a}_3 & \mathbf{b}_3 & \mathbf{c}_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \\ 0 & & & & \mathbf{a}_N & \mathbf{b}_N \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \vdots \\ \\ \mathbf{v}_N \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \mathbf{d}_3 \\ \vdots \\ \\ \mathbf{b}_N \end{pmatrix} \tag{9.13}$$

Itt pl. $\mathbf{v}_1$ az első rácspontban lévő M ismeretlent tartalmazza lényegében tetszőleges sorrendben. A Thomas algoritmus általánosítható blokk tridiagonális mátrixokra is:

$$\begin{aligned} c'_i &= (\mathbf{b}_i - \mathbf{a}_i \cdot \mathbf{c}'_{i-1})^{-1} \cdot \mathbf{c}_i \\ d'_i &= (\mathbf{b}_i - \mathbf{a}_i \cdot \mathbf{c}'_{i-1})^{-1} \cdot (\mathbf{d}_i - \mathbf{a}_i \cdot \mathbf{d}_{i-1}) \end{aligned} \tag{9.14}$$

Majd $i = N, N-1, \ldots 1$ sorrendben (backward sweep):

$$\mathbf{v}_i = \mathbf{d}'_i - \mathbf{c}'_i \cdot \mathbf{v}_{i+1} \tag{9.15}$$

A 9.14 egyenletekben szereplő mátrix invertálások helyett természetesen LU dekompozíciós egyenletrendszer megoldást végzünk. A blokk tridiagonális Thomas algoritmus műveletigénye $\approx 5NM^3/3$, ami sokkal jobb mint a teljes Gauss $O((NM)^3/3)$, viszont nem sokkal jobb mint a sávos mátrix algoritmus $O(N'M^2)$, ahol $N' = NM$ a mátrix elemekben számolt mérete.

9.1.5. Ciklikus (blokk) tridiagonális mátrix

Periodikus határfeltételek diszkretizációja összeköti az első és az N-dik rácspontokat. Ilyenkor gyakran kapunk ciklikus (blokk) tridiagonális mátrixot:

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{c}_1 & & & & \mathbf{a}_1 \\ \mathbf{a}_2 & \mathbf{b}_2 & \mathbf{c}_2 & & & \\ & \mathbf{a}_3 & \mathbf{b}_3 & \mathbf{c}_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \\ \mathbf{c}_N & & & & \mathbf{a}_N & \mathbf{b}_N \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \vdots \\ \\ \mathbf{v}_N \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \mathbf{d}_3 \\ \vdots \\ \\ \mathbf{b}_N \end{pmatrix} \tag{9.16}$$

Ezt a lineáris egyenletrendszert visszavezethetjük a nem-ciklikus problémára a *Woodbury* formula segítségével:

$$(A + B \cdot C^T)^{-1} = A^{-1} - \left[A^{-1} \cdot B \cdot (1 + C^T A^{-1} \cdot B)^{-1} \cdot C^T \cdot A^{-1}\right] \tag{9.17}$$

ahol B és C mátrixok $NM \times M$ rendűek, és $B \cdot C^T$ adja a ciklikus blokkokat. A konkrét algoritmust a Numerical Recipes [1] tartalmazza. Lényegében ugyanazzal az A mátrixszal $M + 1$ különböző jobb oldalhoz kell megoldást találni.

9.2. Hagyományos iteratív módszerek

Az iteratív módszerekben A-t felbontjuk

$$A = N + P \tag{9.18}$$

mátrixokra, ahol $N \approx A$ könnyen faktorizálható, míg P egy lehetőleg kicsi korrekció. Ekkor az eredeti egyenlet

$$N \cdot V = B - P \cdot V \tag{9.19}$$

alakban írható. Mivel N könnyen invertálható, ezt tovább alakíthatjuk a

$$V = N^{-1} \cdot (B - P \cdot V) \tag{9.20}$$

egyenletre, amit a

$$V^{n+1} = N^{-1} \cdot (B - P \cdot V^n) \tag{9.21}$$

iterációval próbálunk megoldani, ahol n indexeli az iterációkat. Az eljárás konvergál, ha $N^{-1}P$ legnagyobb sajátértékének abszolút értéke 1-nél kisebb. Általában elegendő, ha A diagonálisan dominált, azaz

$$|A_{j,j}| > \sum_{i \neq j} |A_{i,j}| \tag{9.22}$$

9.2.1. Jacobi iteráció

$$N = \mathrm{diag}(A) \quad P = L + U \tag{9.23}$$

azaz N a diagonális, míg P az összes nem diagonális elemet tartalmazza. Ez konkrétan a következő iterációhoz vezet:

$$V_i^{n+1} = \frac{1}{A_{i,i}} \left(B_i - \sum_{j \neq i} A_{i,j} V_j^n \right) \tag{9.24}$$

A Jacobi iteráció nem különösebben hatékony, de egyszerű, és a hatékonyabb módszerek alapja, valamint könnyen párhuzamosítható.

9.2.2. Gauss-Seidel iteráció

$$N = \mathrm{diag}(A) + L \quad P = U \tag{9.25}$$

azaz P a diagonális átló feletti elemeket tartalmazza. Ez konkrétan a következő iterációhoz vezet, amit az $i = 1 \ldots N$ sorrendben kell végezni:

$$V_i^{n+1} = \frac{1}{A_{i,i}} \left(B_i - \sum_{j<i} A_{i,j} V_j^{n+1} - \sum_{j>i} A_{i,j} V_j^n \right) \tag{9.26}$$

Ez a módszer kb. kétszer gyorsabban konvergál a Jacobi iterációnál, viszont nem párhuzamosítható.

9.2.3. Szukcesszív túlrelaxálás (SOR)

A Gauss-Seidel módszerhez hasonló, de a „lépéshosszt” variáljuk:

$$N = \frac{\mathrm{diag}(A)}{\lambda} + L \tag{9.27}$$

ahol λ egy optimálisan választott paraméter.

$$V_i^{n+1} = (1 - \lambda) V_i^n + \frac{\lambda}{A_{i,i}} \left(B_i - \sum_{j<i} A_{i,j} V_j^{n+1} - \sum_{j>i} A_{i,j} V_j^n \right) \tag{9.28}$$

Ez $0 < \lambda < 2$-re konvergál. A konvergencia akkor lehet gyorsabb a Gauss-Seidel módszernél, ha $\lambda > 1$, amit túlrelaxálásnak nevezünk. Az optimális λ egyszerű mátrixokra ismert, de általános esetben nehéz eltalálni. Csak viszonylag közel az optimális értékhez ad az SOR igazán jó konvergenciát.

9.3. Krylov altér típusú iteratív módszerek

Az eddig tárgyalt direkt és iteratív módszerk mind felhasználják az A mátrix elemeit, és átlalában szekvenciálisan haladnak végig a mátrix sorain és oszlopain. A Krylov típusú iteratív módszerekben a mátrixról csak annyit kell tudnunk, hogy egy tetszőleges vektrorral megszorozva milyen értéket ad. Ez egyrészt azt jelenti, hogy a mátrix elemeket nem is kell feltétlenül kiszámolni, illetve hogy az algoritmus könnyen párhuzamosítható.

9.3.1. Konjugált gradiens

A konjugált gradiens módszer csak szimmetrikus pozitív definit mátrixokra alkalmazható. Az alapötlet, hogy keressük az

$$f(V) = \frac{1}{2} V \cdot A \cdot V - B \cdot V \tag{9.29}$$

függvény minimumhelyét. A megoldás

$$\mathbf{grad} f(V) = A \cdot V - B = 0 \tag{9.30}$$

azaz a minimalizálási feladattal éppen az eredeti egyenletrendszert oldjuk meg. Az iteratív algoritmus lényege a következő:

- Válasszunk egy keresési irányt P_k ami „merőleges” ($P_k \cdot A \cdot P_l = 0$ ha $l < k$) a korábbiakra
- Minimalizáljuk az $f(V_k + \alpha_k P_k)$ függvényt α szerint
- $V_{k+1} = V_k + \alpha_k P_k$
- $k \to k + 1$

A konkrét algoritmus:

1. $R_1 = B - A \cdot V_1$ és $P_1 = R_1$
2. $\alpha_k = \frac{R_k \cdot R_k}{P_k \cdot A \cdot P_k}$
3. $R_{k+1} = R_k - \alpha_k A \cdot P_k$
4. $\beta_k = \frac{R_{k+1} \cdot R_{k+1}}{R_k \cdot R_k}$
5. $P_{k+1} = R_k + \beta_k P_k$
6. $V_{k+1} = V_k + \alpha_k P_k$

Megfelelő sorrendcserékkel és ideiglenes változók bevezetésével elérhető, hogy a CG módszer a megoldáson és a jobb oldalon kívül csak két extra vektort tároljon, és iterációnként csak egy mátrix vektor szorzásra van szükség.

9.3.2. BiCG és BiCGstab

A konjugált gradiens módszer általánosítható nem szimmetrikus mátrixokra. A BiCG algoritmus kétszer annyi műveletet és tárolási helyet igényel mint a CG. Elvileg működik, de nem mindig konvergál. Ennek javított változata a BiCGSTAB (stabilizált BiCG) algoritmus, ami sokkal megbízhatóbb. Tárolási igénye $4-7$ vektor az implementációtól és az algoritmus részleteitől függően.

9.3.3. MINRES és GMRES

A MINRES (minimum reziduum) módszer a CG egy általánosítása, amelyik szimmetrikus de nem pozitív definit mátrixokra is működik. Lényegében az

$$f(V) = \frac{1}{2} |A \cdot V - B|^2 \tag{9.31}$$

függvényt minimalizálja a CG-hez hasonló módon. A függvény minimuma

$$\mathbf{grad} f(V) = A^T \cdot (A \cdot V - B) = 0 \tag{9.32}$$

helyen van. Az $A^T \cdot A$ mátrix szimmetrikus és pozitív definit.

A MINRES módszert sikerült általánosítani nem szimmetrikus mátrixokra is, ez a GMRES (generalized minimum reziduum) algoritmus. Ez az egyik legjobb Krylov típusú iteratív módszer. Egyetlen hátránya, hogy az összes korábbi keresési irányt tárolni kell. A gyakorlatban bizonyos számú iteráció (pl. 20) után újra lehet indítani.

9.3.4. Prekondicionálás

A Krylov típusú iteratív módszerek diagonálisan dominált mátrixokra működnek igazán jól. Ezért érdemes az eredeti egyenletet megszorozni balról és/vagy jobbról egy P prekondicionáló mátrixszal, ami A inverzét közelíti:

$$A' \cdot V = (P \cdot A) \cdot V = P \cdot B = B' \tag{9.33}$$

Az A' mátrixra alkalmazva a Krylov módszert nagy mértékben javul a konvergencia. Persze a prekondicionáláshoz általában kellenek az A elemei, és tipikusan szekvenciális művelet, valamilyen közelítő dekompozíció (incomplete LU, ILU).

A Schwarz prekondicionálás megoldja a párhuzamosítást: minden processzor úgy prekondicionálja az A mátrixot, mintha az csak a processzorhoz tartozó rácsrészen hatna.

9.4. Multigrid

Az iteratív módszerek gyorsan megszabadulnak a kis hullámhosszú hibától, de soká tart amíg a globális hibákat lecsökkentik. Ezt használja ki a multigrid módszer. A multigrid módszerek lényege, hogy a lineáris egyenletrendszert különböző rácsfelbontások mellett oldjuk meg. A rácsok között finomító és durvító interpolációkkal adjuk át a megoldást. Így minden szint a maga rövid hullámhosszú hibáitól szabadul meg, ami az eredeti probléma összes hullámhosszát jelenti. A multigrid rendkívül hatékony diagonálisan dominált problémákra, ugyanakkor meglehetősen komplikált. Nem lineáris problémákra is alkalmazható.

9.5. Pszeudo-tranziens módszer

Stacionárius problémák általában elliptikus egyenletet adnak. Ha ezt diszkretizáljuk, akkor egy (nem-lineáris) egyenletrendszert kapunk, amit az eddig tárgyalt egyenletrendszer megoldó algoritmusokkal lehet megoldani. Egy másik megközelítés, hogy az eredeti probléma helyett egy olyan időfüggő problémát oldunk meg, ami a stacionárius probléma megoldásához tart. Ez is lényegében egy iteratív módszer, ahol az időlépések veszik át az iterációk szerepét.

Például a

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0 \tag{9.34}$$

elliptikus egyenlet helyett oldjuk meg a

$$\frac{\partial V}{\partial t} = \alpha \left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \right) \tag{9.35}$$

parabolikus egyenletet. Például véges differencia módszerrel

$$V_{j,k}^{n+1} = (1-4s)V_{j,k}^n + s(V_{j-1,k}^n + V_{j+1,k}^n + V_{j,k-1}^n + V_{j,k+1}^n) \tag{9.36}$$

$$s = \frac{\alpha \Delta t}{\Delta x^2} \tag{9.37}$$

Nézzük meg mi a kapcsolat a pszeudotranziens módszer és a fentebb tárgyalt iteratív módszerek között. Az eredeti elliptikus egyenletet véges differencia módszerrel diszkretizálva

$$0 = \frac{1}{\Delta x^2}(V_{j-1,k}^n + V_{j+1,k}^n + V_{j,k-1}^n + V_{j,k+1}^n - 4V_{j,k}^n) \tag{9.38}$$

amiből egy pentadiagonális mátrixot kapunk:

$$A = \begin{pmatrix} 4 & -1 & & -1 & & 0 \\ -1 & 4 & -1 & & -1 & \\ & -1 & 4 & -1 & & \ddots \\ -1 & & \ddots & \ddots & \ddots & \\ & \ddots & & & & \\ 0 & & -1 & & -1 & 4 \end{pmatrix} \tag{9.39}$$

Ha erre a Jacobi módszert alkalmazzuk, akkor

$$V_{j,k}^{n+1} = \frac{1}{4}(V_{j-1,k}^{n} + V_{j+1,k}^{n} + V_{j,k-1}^{n} + V_{j,k+1}^{n}) \tag{9.40}$$

ami egybeesik a pszeudo-tranziens módszerrel $s = 1/4$ választás esetén.

Természetesen a pszeudo-tranziens módszert nem csak explicit véges differencia módszerrel diszkretizálhatjuk, így a klasszikus iteratív módszerekhez képest rengeteg lehetőségünk van.

Erre egy példa a lokális időlépés alkalmazása: legyen az időlépés mindenhol akkora, amit a lokális stabilitási limit megenged. Az eredmény gyorsabb konvergencia.

10. fejezet

Diffúzió egyenlet

Ebben a fejezetben a lineáris parabolikus diffúziós egyenlet diszkretizációit tekintjük át. Az egyenlet általános formája

$$\frac{\partial T}{\partial t} - \mathrm{div}\alpha\mathbf{grad}T = 0 \tag{10.1}$$

ahol T pl. a hőmérséklet, és $\alpha > 0$ a diffúziós együttható. Az egyszerűség kedvéért feltesszük, hogy α konstans.

10.1. 1 dimenziós diffúzió egyenlet

Egy dimenzióban a diffúziós egyenlet a

$$\frac{\partial T}{\partial t} - \alpha\frac{\partial^2 T}{\partial x^2} = 0 \tag{10.2}$$

formában írható.

10.1.1. FTCS módszer

A legegyszerűbb diszkretizáció

$$T_j^{n+1} = (1-2s)T_j^n + s(T_{j+1}^n + T_{j-1}^n) \tag{10.3}$$

$$s = \frac{\alpha\Delta t}{\Delta x^2} \tag{10.4}$$

Taylor sorfejtéssel megkaphatjuk a diszkretizációs hibát

$$E = \frac{\Delta t}{2}\frac{\partial^2 T}{\partial t^2} - \alpha\frac{\Delta x^2}{12}\frac{\partial^4 T}{\partial x^4} + O(\Delta t, \Delta x^2) \tag{10.5}$$

Megjegyezzük, hogy $s = 1/6$ választással a hiba $O(\Delta t^2, \Delta x^4)$-re redukálódik a parciális deriváltak köti kapcsolat miatt.

A von Neumann stabilitásvizsgálat $T = (G)^n \exp(ikx_j)$ helyettesítéssel

$$G = 1 - 2s + e^{ik\Delta x} + e^{-ik\Delta x} = 1 - 4s\sin^2(k\Delta x/2) \tag{10.6}$$

amire $|G| < 1$, ha $s < 1/2$ stabilitási feltétel adódik.

10.1.2. Richardson módszer

Próbáljuk meg az időbeli diszkretizációt másodrendűvé tenni

$$\frac{T_j^{n+1} - T_j^{n-1}}{2\Delta t} = \alpha\frac{T_{j-1}^n - 2T_j^n + T_{j+1}^n}{\Delta x^2} \tag{10.7}$$

$$T_j^{n+1} = T_j^{n-1} + 2s(T_{j-1}^n - 2T_j^n + T_{j+1}^n) \tag{10.8}$$

ami szimmetriából következően $O(\Delta t^2, \Delta x^2)$ rendű. A stabilitásvizsgálatból

$$G = \frac{1}{G} - 8s\sin^2(k\Delta x/2) \tag{10.9}$$

Ez egy másodfokú egyenlet G-re. Ha bevezetjük a $b = 4s\sin^2(k\Delta x/2)$ jelölést, akkor

$$G_{1,2} = -b \pm \sqrt{b^2 + 1} \tag{10.10}$$

amiből $|G_2| > 1$. Azaz a Richardson módszer a diffúziós egyenletre feltétel nélkül instabil.

10.1.3. DuFort-Frankel módszer

Próbáljuk meg stabillá tenni a Richardson módszert:

$$\frac{T_j^{n+1} - T_j^{n-1}}{2\Delta t} = \alpha\frac{T_{j-1}^n - (T_j^{n-1} + T_j^{n+1}) + T_{j+1}^n}{\Delta x^2} \tag{10.11}$$

$$T_j^{n+1} = \frac{1-2s}{1+2s}T_j^{n-1} + \frac{2s}{1+2s}(T_{j-1}^n + T_{j+1}^n) \tag{10.12}$$

A von Neumann vizsgálattal megmutatható, hogy ez a módszer feltétel nélkül stabil. A hiba tag azonban

$$E = \alpha\frac{\Delta t^2}{\Delta x^2}\frac{\partial^2 T}{\partial t^2} + O(\Delta t^2, \Delta x^2) \tag{10.13}$$

ami csak akkor kicsi, ha $\Delta t \ll \Delta x$. Azaz nem a stabilitás, hanem a konzisztencia követeli meg a kicsi időlépést.

10.1.4. Teljesen implicit módszer

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \alpha \frac{T_{j-1}^{n+1} - 2T_j^{n+1} + T_{j+1}^{n+1}}{\Delta x^2} \tag{10.14}$$

$$T_j^n = (1+2s)T_j^{n+1} - s(T_{j+1}^{n+1} + T_{j-1}^{n+1}) \tag{10.15}$$

Ez egy $O(\Delta t, \Delta x^2)$ rendű feltétel nélkül stabil diszkretizáció, hiszen

$$G = \frac{1}{1 + 4s\sin^2(k\Delta x/2)} \tag{10.16}$$

amiből $|G| < 1$.

A kapott egyenletrendszerben egy tridiagonális mátrix szerepel $(-s, 1+2s, -s)$ elemekkel a három átlóban, így a Thomas módszerrel hatékonyan meg lehet oldani.

10.1.5. Crank-Nicholson módszer

Az időintegrálás másodrendűvé tehető:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \frac{\alpha}{2}\left[\frac{T_{j-1}^n - 2T_j^n + T_{j+1}^n}{\Delta x^2} + \frac{T_{j-1}^{n+1} - 2T_j^{n+1} + T_{j+1}^{n+1}}{\Delta x^2}\right] \tag{10.17}$$

ami az $n + 1/2$ időlépéshez képesti szimmetriából következően $O(\Delta t^2, \Delta x^2)$ rendű. A stabilitásvizsgálatból

$$G = \frac{1 - 2s\sin^2(k\Delta x/2)}{1 + 2s\sin^2(k\Delta x/2)} \tag{10.18}$$

amiből $|G| \leq 1$, ami feltétel nélkül stabil. A megoldandó egyenletrendszer éppolyan bonyolult mint a teljesen implicit módszernél.

10.2. Több dimenziós diffúziós egyenlet

Két dimenzióban Descartes rácson

$$\frac{\partial T}{\partial t} - \alpha_x \frac{\partial^2 T}{\partial x^2} - \alpha_y \frac{\partial^2 T}{\partial y^2} = 0 \tag{10.19}$$

formában írható, ahol megengedtük, hogy a diffúziós együttható különbözzön az x és y irányokban.

10.2.1. FTCS módszer

A legegyszerűbb diszkretizáció

$$\begin{aligned} T_{j,k}^{n+1} &= (1-2s_x-2s_y)T_{j,k}^n + s_x(T_{j+1,k}^n + T_{j-1,k}^n) + s_y(T_{j,k+1}^n + T_{j,k-1}^n) \\ s_x &= \frac{\alpha_x \Delta t}{\Delta x^2} \\ s_x &= \frac{\alpha_y \Delta t}{\Delta y^2} \end{aligned} \tag{10.20}$$

Taylor sorfejtéssel megkaphatjuk a diszkretizációs hibát $O(\Delta t, \Delta x^2, \Delta y^2)$. Itt már általában nincs olyan speciális választás s-re, ami a diszkretizációt magasabbrendűvé tenné. A stabilitásvizsgálatot $T = (G)^n \exp(ik_x x + ik_y y)$ helyettesítéssel végezzük el, és az erősítési faktor

$$G = 1 - 4s_x \sin^2(k_x \Delta x/2) - 4s_y \sin^2(k_y \Delta y/2) \tag{10.21}$$

amiből $|G| < 1$ feltétele $s_x + s_y < 1/2$. Ez feleakkora mint 1 dimenzióban. 3 dimenzióban $s_x + s_y + s_z < 1$ pedig harmadakkora.

10.2.2. Operátor bontott FTCS

Oldjuk meg az x és y irányú diffúziós egyenletet váltakozva:

$$T^{n+1} = (I + \Delta t \alpha_x L_{xx})(I + \Delta t \alpha_y L_{yy}) T^n \tag{10.22}$$

ahol L_{xx} és L_{yy} operátorok a $\partial^2/\partial x^2$ illetve $\partial^2/\partial y^2$ diszkretizációi, míg I az identitás operátor. Az operátor bontásból származó hiba Δt^2-tel arányos, azaz a diszkretizáció rendje továbbra is $O(\Delta t, \Delta x^2)$. Ugyanakkor a stabilitási limit egyszerűen az x és az y irányú egy dimenziós diszkretizációk stabilitási limitjei, azaz $s_x < 1/2$ és $s_y < 1/2$. Az operátorbontott FTCS módszerben tehát ($s_x = s_y$ esetén) kétszer akkora időlépés használható, mint az eredeti 2 dimenziós FTCS módszerben.

10.2.3. Implicit módszerek

A diffúziós egyenlet implicit diszkretizációja könnyen általánosítható 2 vagy 3 dimenzióra:

$$\frac{T^{n+1} - T^n}{\Delta t} = [\alpha_x L_{xx} + \alpha_y L_{yy}] \left[(1-\beta) T^n + \beta T^{n+1} \right] \tag{10.23}$$

ahol $\beta = 1$ az $O(\Delta t, \Delta x^2, \Delta y^2)$ rendű teljesen implicit módszert, míg $\beta = 1/2$ a térben és időben másod $O(\Delta t^2, \Delta x^2, \Delta y^2)$ rendű Crank-Nicholson módszert adja. A diszkretizáció $\beta > 1/2$ esetén feltétel nélkül stabil.

A fenti implicit diszkretizáció átrendezhető

$$[I - \Delta t\beta(\alpha_x L_{xx} + \alpha_y L_{yy})]\,\Delta T^{n+1} = \Delta t(\alpha_x L_{xx} + \alpha_y L_{yy})T^n \qquad (10.24)$$

alakra, ahol az ismeretlen $\Delta T^{n+1} = T^{n+1} - T^n$. Ebben az egyenletrendszerben egy pentadiagonális mátrix szerepel az L_{xx} és L_{yy} operátorok miatt. Az egyenletrendszert a korábban tárgyalt iteratív módszerekkel meg lehet ugyan oldani, de ez mindenképpen elég költséges.

10.2.4. Váltakozó irányban implicit módszer

A pentadiagonális mátrix invertálását a váltakozó irányban implicit (alternating direction implicit, ADI) módszer úgy kerüli el, hogy az időlépést két féllépésre bontjuk: az első féllépésben az x, a másodikban az y irányban implicit a diszkretizáció:

$$\frac{T^{n+1/2} - T^n}{\Delta t/2} = \alpha_x L_{xx} T^{n+1/2} + \alpha_y L_{yy} T^n \qquad (10.25)$$

$$\frac{T^{n+1} - T^{n+1/2}}{\Delta t/2} = \alpha_x L_{xx} T^{n+1/2} + \alpha_y L_{yy} T^{n+1} \qquad (10.26)$$

Mindkét féllépésben egy egyenletrendszert kell megoldanunk a $T^{n+1/2}$ illetve T^{n+1} ismeretlenekre, azonban a mátrix mindkét esetben tridiagonális lesz, így az egyenletrendszereket a Thomas algoritmussal hatékonyan meg tudjuk oldani. Az ADI algoritmus térben és időben másodrendű. Az időben másodrendű pontosság a $(t_n + t_{n+1})/2$ időpont körüli szimmetriából következik. A stabilitásvizsgálatból az erősítési faktorra

$$G = G'G'' = \left[\frac{1 - 2s_y \sin^2(k_y \Delta y/2)}{1 + 2s_x \sin^2(k_x \Delta x/2)}\right]\left[\frac{1 - 2s_x \sin^2(k_x \Delta x/2)}{1 + 2s_y \sin^2(k_y \Delta y/2)}\right] \qquad (10.27)$$

adódik, amiből $|G| < 1$, azaz az ADI módszer két dimenzióban feltétel nélkül stabil. Itt az egyes féllépések erősítési faktora $|G'|$ illetve $|G''|$ lehet egynél nagyobb, de a teljes lépés stabil marad.

Az ADI módszer három dimenzióra is általánosítható: a három harmadlépésben az x, y, majd z irányban implicit, a másik két irányban explicit diszkretizációt használunk. Ez is $O(\Delta t^2, \Delta x^2, \Delta y^2)$ rendű diszkretizáció, azonban a stabilitás már nem feltétel nélküli, ugyanis például az első harmad lépés erősítési faktora

$$G' = \frac{1 - (4/3)s_y \sin^2(k_y \Delta y/2) - (4/3)s_z \sin^2(k_z \Delta z/2)}{1 + (4/3)s_x \sin^2(k_x \Delta x/2)} \qquad (10.28)$$

amiből látható, hogy a $|G'| < 1$ feltétele $s_x, s_y, s_z < 3/2$ lesz, és ugyanezt kapjuk a második és harmadik részlépés erősítési faktorára is, így a teljes $G = G'G''G'''$-re is.

10.2.5. Implicit közelítő faktorizáció

Az explicit időlépésnél használt operátor bontás az implicit időintegrálásra is használható. Itt nem a stabilitási tartomány növelése, hanem a megoldandó egyenletrendszer egyszerűsítése a cél: az eredeti implicit diszkretizációban szereplő pentadiagonális mátrixot két tridiagonális mátrix szorzatával közelítjük:

$$[I - \Delta t\beta\alpha_x L_{xx}]\,[I - \Delta t\beta\alpha_y L_{yy}]\,\Delta T^{n+1} = \Delta t(\alpha_x L_{xx} + \alpha_y L_{yy})T^n \quad (10.29)$$

Ezt az egyenletrendszert két lépésben oldjuk meg

$$[I - \Delta t\beta\alpha_x L_{xx}]\,\Delta T^* = \Delta t(\alpha_x L_{xx} + \alpha_y L_{yy})T^n \quad (10.30)$$

$$[I - \Delta t\beta\alpha_y L_{yy}]\,\Delta T^{n+1} = \Delta T^* \quad (10.31)$$

Mindkét lépésben egy tridiagonális mátrix szerepel, így az egyenletrendszerek hatékonyan megoldhatóak a Thomas algoritmussal. Az operátorbontásból származó extra tag $\Delta t^2 \Delta T^{n+1}$-gyel arányos. Mivel ΔT^{n+1} maga is arányos az időlépéssel, az új tag harmadrendű hibát okoz. Így $\beta = 1/2$-re a módszer $O(\Delta t^2, \Delta x^2, \Delta y^2)$ rendű akárcsak az eredeti Crank-Nicholson diszkretizáció. A stabilitásvizsgálat azt mutatja, hogy az operátorbontott diszkretizáció feltétel nélkül stabil $\beta \geq 1/2$ esetén.

Az operátorbontásos eljárás könnyen általánosítható 3 dimenzióra, és – ellentétben az ADI módszerrel – továbbra is feltétel nélkül stabil marad $\beta \geq 1/2$-re.

11. fejezet

Konvekció dominált problémák

A numerikus hidrodinamikában az áramlással kapcsolatos tagok diszkretizálása az egyik legnehezebb probléma. Szemben a diffúziós tagokkal, melyek a megoldást – és így a diszkretizációs hibákat is – kisímítják, a konvekciós tagok a hullámokat amplitúdó csökkenés nélkül továbbítják. A diszkretizációs hibák tehát amplitúdó és diszperziós hibaként lépnek fel. Ezek együttes minimalizálása nem könnyű feladat.

11.1. 1 dimenziós lineáris konvekciós egyenlet

$$\frac{\partial T}{\partial t} + v\frac{\partial T}{\partial x} = 0 \tag{11.1}$$

Ez lényegében hiperbolikus PDE-ként viselkedik, hiszen a hullámok disszipációmentesen véges sebességgel terjednek.

11.1.1. FTCS módszer

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + v\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} = 0 \tag{11.2}$$

ami átírható

$$T_j^{n+1} = T_j^n - \frac{C}{2}(T_{j+1}^n - T_{j-1}^n) \tag{11.3}$$

alakra, ahol

$$C = v\frac{\Delta t}{\Delta x} \tag{11.4}$$

a **Courant szám**. Ennek stabilitásvizsgálata

$$G = 1 - \frac{C}{2}(e^{ik\Delta x} - e^{-ik\Delta x}) = 1 - iC\sin(k\Delta x) \tag{11.5}$$

ami $|G| > 1$ miatt feltétel nélkül instabil.

11.2. Áramlásirányú/upwind módszer

Az FTCS módszer instabilitását meg lehet szüntetni, ha a diszkrét térbeli deriváltat centrális helyett féloldalas deriválttal helyettesítjük:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + v\frac{T_j^n - T_{j-1}^n}{\Delta x} = 0 \tag{11.6}$$

ami átírható

$$T_j^{n+1} = (1 - C)T_j^n + CT_{j-1}^n \tag{11.7}$$

Ez a képlet csak pozitív sebességre jó, negatív előjelű v esetén a másik oldali deriváltat kell venni

$$T_j^{n+1} = (1 - |C|)T_j^n + |C|T_{j+1}^n \tag{11.8}$$

A stabilitásvizsgálatból

$$G = 1 - |C| + |C|e^{\pm ik\Delta x} \tag{11.9}$$

Ha $|C| < 1$ akkor $|G| < 1$, azaz az áramlásirányú módszer feltételesen stabil. A

$$|C| = |v|\frac{\Delta t}{\Delta x} < 1 \tag{11.10}$$

stabilitási feltételt Courant-Friedrichs-Lewy (CFL) feltételnek hívjuk. Általában **a hiperbolikus PDE hullámai egy időlépés alatt nem haladhatnak többet, mint a numerikus módszer tartója**.

A diszkretizációs hiba $O(\Delta t, \Delta x)$ rendű, illetve a $\partial^2 T/\partial t^2 = v^2\partial^2 T/\partial x^2$ összefüggés felhasználásával

$$E = \frac{1}{2}v\Delta x(1 - C)\frac{\partial^2 T}{\partial x^2} \tag{11.11}$$

alakra hozható. A hiba tehát diffúziós tag alakú, ami a hullámok amplitúdócsökkenését eredményezi. A $C = 1$ választás esetén az áramlásirányú módszer egzakttá válik, de ez csak konstans v esetén áll fenn, bonyolultabb egyenletre nem alkalmazható.

11.2.1. Leapfrog módszer

Egy időben másodrendű módszert kapunk, ha az időderiváltat is centrálisan diszkretizáljuk:

$$\frac{T_j^{n+1} - T_j^{n-1}}{2\Delta t} + v\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} = 0 \tag{11.12}$$

Ez nyilvánvalóan $O(\Delta t^2, \Delta x^2)$ rendű módszer. A konkrét algoritmus

$$T_j^{n+1} = T_j^{n-1} - C(T_{j+1}^n - T_{j-1}^n) \tag{11.13}$$

aminek a stabilitásvizsgálata

$$G = \frac{1}{G} - 2iC\sin(k\Delta x) \tag{11.14}$$

egyenletet adja. Ennek gyökei

$$G_{1,2} = -iC\sin(k\Delta x) \pm \sqrt{1 - C^2\sin^2(k\Delta x)} \tag{11.15}$$

Ebből $|G| = 1$ ha $|C| \leq 1$, és $|G| > 1$ ha $|C| > 1$, azaz a leapfrog módszer feltételesen stabil, és ugyanannak a CFL stabilitás feltételnek kell eleget tennie mint az upwind módszernek. A leapfrog módszer amplitúdó hibája nulla, csak diszperziós hiba lép fel.

Sajnos a leapfrog módszer egyik hátránya, hogy a páros és páratlan koordinátájú téridő rácspontok szétcsatolódnak. Ez különösen nem lineáris egyenleteknél probléma. Szokás időnkénti átlagolással kiküszöbölni ezt a hibát.

11.2.2. Lax-Wendroff módszer

Időben másodrendű diszkretizációt úgy is kaphatunk, ha a féloldalas (forward in time) időderivált Taylor sorában lévő másodrendű hiba tagot kiejtjük:

$$\frac{\partial T}{\partial t} = \frac{T_j^{n+1} - T_j^n}{\Delta t} - \frac{\Delta t}{2}\frac{\partial^2 T}{\partial t^2} = \frac{T_j^{n+1} - T_j^n}{\Delta t} - v^2\frac{\Delta t}{2}\frac{\partial^2 T}{\partial x^2} \tag{11.16}$$

Az időderiváltat átírtuk térbeli deriváltra. Ha ezt centrális differenciálással diszkretizáljuk, akkor

$$T_j^{n+1} = T_j^n - \frac{C}{2}(T_{j+1}^n - T_{j-1}^n) + \frac{C^2}{2}(T_{j-1}^n - 2T_j^n + T_{j+1}^n) \tag{11.17}$$

Konstrukciójából eredően a Lax-Wendroff módszer $O(\Delta t^2, \Delta x^2)$ rendű. Meglepő módon stabilitása is sokkal kedvezőbb, mint az FTCS módszeré:

$$G = 1 - iC\sin(k\Delta x) - C^2(1 - \cos(k\Delta x)) \tag{11.18}$$

ami a komplex számsíkon egy $(1 - C^2, 0)$ középpontú $a = C^2$ és $b = C$ féltengelyű ellipszis. Az erősítési faktor abszolútérték négyzete

$$|G|^2 = (1 - C^2(1 - \cos(k\Delta x)))^2 + C^2\sin^2(k\Delta x) = 1 - (C^2 - C^4)(1 - \cos(k\Delta x))^2 \tag{11.19}$$

Mint láthatóró $|G| < 1$, ha $|C| < 1$, azaz a Lax-Wendroff módszer feltételesen stabil.

Bonyolultabb egyenletek esetén az időbeli második derivált nem játszható át egyszerűen térbeli deriváltra. Ilyenkor a két lépéses Lax-Wendroff módszert használjuk:

$$\begin{aligned} T_{j+1/2}^{n+1/2} &= \frac{T_j^n + T_{j+1}^n}{2} - \frac{C}{2}(T_{j+1}^n - T_j^n) \\ T_j^{n+1} &= T_j^n - C(T_{j+1/2}^{n+1/2} - T_{j-1/2}^{n+1/2}) \end{aligned} \tag{11.20}$$

Az első lépés egy FTCS típusú diszkretizáció ami átlagolást tartalmaz a páros és páratlan csomópontok között, míg a második lépés a leapfrog módszer térben és időben felére zsugorított változata. Az első egyenletet a másodikba helyettesítve az egy lépéses Lax-Wendroff módszert kapjuk meg. Bonyolultabb egyenletekre azonban a kétlépéses módszer adja az egyszerűbb diszkretizációt. Érdekes megjegyezni, hogy ebben a formájában a Lax-Wendroff módszer nyilvánvalóan konzervatív diszkretizáció.

Bonyolult rácsokon nem könnyű a $j \pm 1/2$ indexű rácspontok meghatározása.

11.2.3. MacCormack módszer

A MacCormack módszer lineáris konvekciós egyenletre ekvivalens a a Lax-Wendroff módszerrel, azonban bonyolultabb egyenletekre a két módszer más diszkretizációhoz vezet. A MacCormack módszerben az első lépés az upwind módszerhez hasonlóan féloldalas térbeli deriváltat használ, a második lépésben pedig az első lépésben kapott megoldást használva a másik irányú deriváltat képezzük, és a két részeredményt átlagoljuk:

$$T_j^* = T_j^n - C(T_j^n - T_{j-1}^n) \tag{11.21}$$

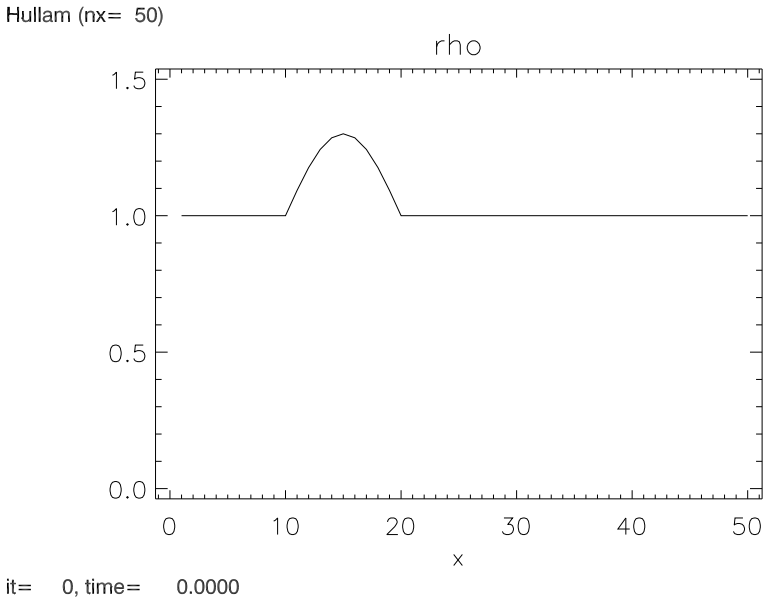
$$T_j^{n+1} = \frac{1}{2}\left[T_j^n + T_j^* - C(T_{j+1}^* - T_j^*)\right] \tag{11.22}$$

Ennek a két lépéses módszernek előnye a két lépéses Lax-Wendroff diszkretizációval szemben, hogy csak az eredeti rácson kell a a térbeli deriváltakat kiértékelni.

11.2.4. Crank-Nicholson módszer

A konvekciós egyenlet egyik lehetséges implicit diszkretizációja a Crank-Nicholson módszer:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + \frac{v}{4\Delta x}(T_{j+1}^n - T_{j-1}^n + T_j^{n+1} - T_{j-1}^{n+1}) = 0 \tag{11.23}$$



11.1. ábra. A félhullám konvekciójának kezdeti feltétele.

Ez egy tridiagonális mátrixot tartalmazó lineáris egyenletrendszer T^{n+1}-re nézve, amit a Thomas módszerrel hatékonyan meg lehet oldani. A módszer időben és térben másodrendű. A stabilitás vizsgálatból az erősítési faktor

$$G = \frac{1 - i(C/2)\sin(k\Delta x)}{1 + i(C/2)\sin(k\Delta x)} \tag{11.24}$$

amiből $|G| = 1$, azaz a Crank-Nicholson módszer feltétel nélkül stabil. Sőt, az is megállapítható, hogy nincs amplitúdó hiba, hiszen az erősítési faktor abszolútértéke 1.

11.3. Félhullám konvekciója

Ebben a részben megnézzük, hogy a fent tárgyalt módszerek mennyire működnek jól a gyakorlatban.

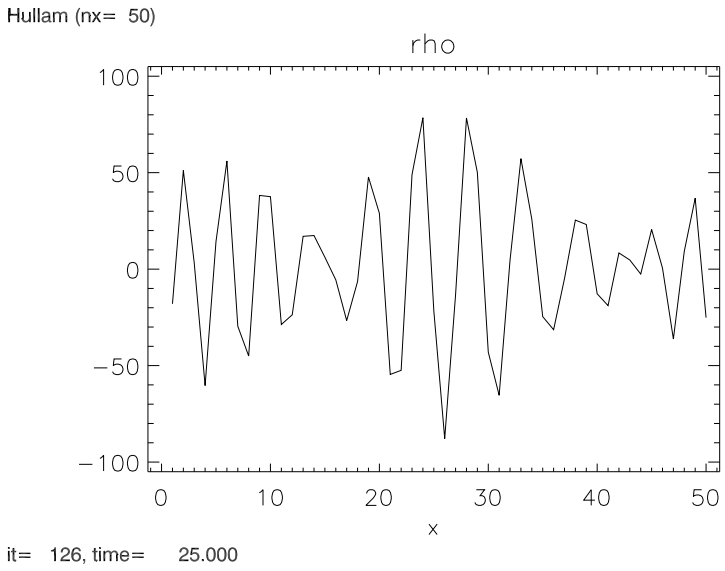
11.3.1. Kezdeti feltétel

A kezdeti feltétel

$$\rho(x) = \begin{cases} 1 + 0.3\sin[0.1\pi(x-10)] & \text{ha } 10 \le x \le 20 \\ 1 & \text{egyébként} \end{cases} \tag{11.25}$$

a $0 < x < 50$ számítási tartományban, amit a 11.1. ábra mutat. A határfeltételek periodikusak. A sebesség $v = 2$, így $t = 25$-re a hullámnak vissza kell térnie a kezdeti pozícióba.

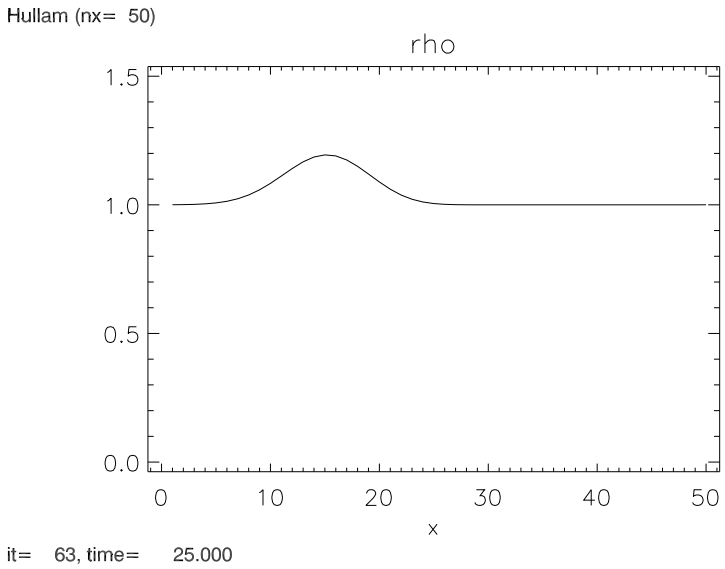
11.3.2. Megoldás FTCS módszerrel



11.2. ábra. FTCS módszer

Az FTCS módszer $\Delta t = 0.2$ értékkel a 11.2 ábrán látható „eredményt" adja, ami a módszer numerikus instabilitását mutatja. Az módszer bármilyen kicsi (pozitív) időlépésre is instabil.

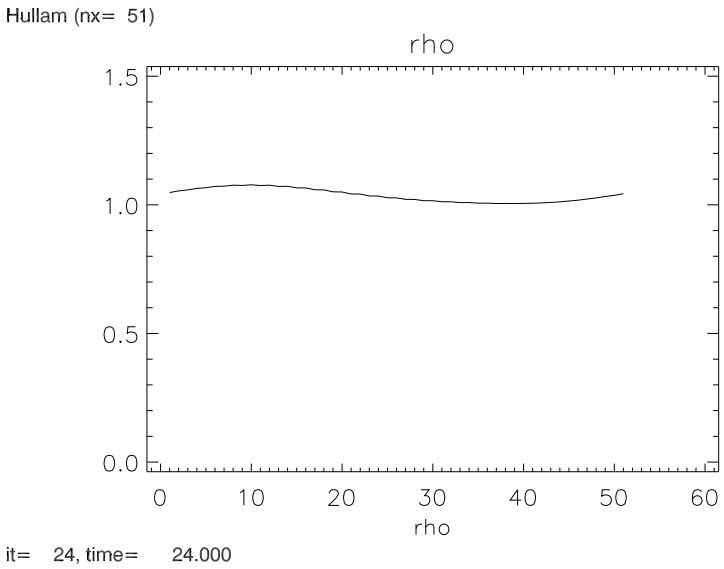
11.3.3. Megoldás upwind módszerrel



11.3. ábra. Upwind módszer

A $\Delta t = 0.4$ esetén a Courant szám $C = 0.8$ és így a megoldás stabil. Az elsőrendű módszer a 11.3 ábrán látható meglehetősen diffúzív eredményt adja, ami következik abból is, hogy az erősítési tényező $|G| < 1$. $\Delta t = 0.5$ esetén $C = 1$ és $|G| = 1$, azaz a megoldás stabil, és nincs diffúzió. Ebben a speciális esetben a megoldás egzakt, minden időlépésben egy rácsszélességgel halad jobbra. Végül $\Delta t = 0.51$ esetén $|G| > 1$ és a megoldás instabillá válik, hasonló ereménnyel mint az FTCS módszerrel kapott 11.2 ábrán látható „megoldás”.

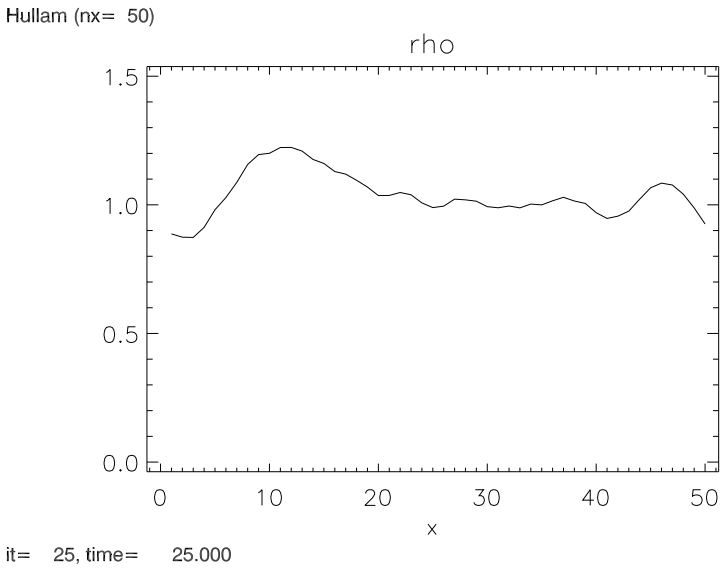
11.3.4. Megoldás implicit időintegrálással



11.4. ábra. Időben elsőrendű implicit módszer

Az implicit módszer feltétel nélkül stabil, így $\Delta t = 1$-t választhatunk időlépésnek, ami $C = 2$-nek felel meg. Az ábrán látható megoldás amplitúdója rendkívül lecsökkent. A numerikus diffúziót az időben elsőrendű módszer és a nagy időlépés kombinációja okozza.

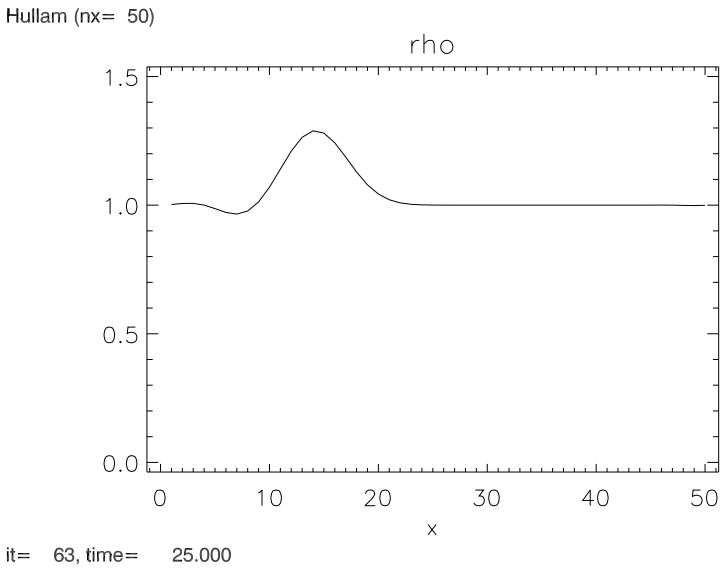
11.3.5. Megoldás Crank-Nicholson módszerrel



11.5. ábra. Trapéz módszer a hullám problémára

A másodrendű trapéz módszerrel kapott megoldásban az amplitúdó hiba kisebb, de erős diszperziós hibát láthatunk.

11.3.6. Megoldás Lax-Wendroff/MacCormack módszerrel



11.6. ábra. MacCormack módszer a hullám problémára

A másodrendű explicit Lax-Wendroff/MacCormack módszerrel kapott megoldás elég pontos. A hullám előtt a sűrűség kis mértékben a kezdeti minimális érték alá süllyed.

11.4. 1 dimenziós transzport egyenlet

$$\frac{\partial T}{\partial t} + v\frac{\partial T}{\partial x} - \alpha\frac{\partial^2 T}{\partial x^2} = 0 \tag{11.26}$$

Ez egy parabolikus PDE. Érdemes bevezetni a diffúziós és konvekciós tag viszonyát jellemző „cella Reynolds számot"

$$R_{cell} = \frac{|C|}{s} = \frac{|v|\Delta x}{\alpha} \tag{11.27}$$

11.4.1. FTCS

A stabilitás feltétele

$$C^2 \leq 2s \leq 1 \tag{11.28}$$

ami azt jelenti, hogy a fizikai diffúzió stabilizálja a konvektív tag instabilitását, ha $C^2 \leq 2s$. Azonban a diszkretizációs hiba tag analíziséből kiderül, hogy a numerikus diffúzió csak akkor lesz elhanyagolható a fizikai diffúzióhoz képest, ha

$$C^2 \ll 2s \qquad \text{azaz} \qquad R_{cell} \ll \frac{2}{|C|} \tag{11.29}$$

11.4.2. Richardson/leapfrog módszer

Feltétel nélkül instabil, kivéve ha nincs diffúzió, azaz $s = 0$.

11.4.3. DuFort-Frankel módszer

A stabilitás feltétele a CFL feltétel, azaz $|C| < 1$, a diffúziós tag s bármekkora lehet. Ugyanakkor a pontosság megköveteli, hogy $C^2 \ll 1$ legyen.

11.4.4. Áramlásoldali módszer

Stabilitás:

$$|C| + 2s \leq 1 \tag{11.30}$$

Pontosság:

$$R_{cell} \ll \frac{2}{1 - |C|} \tag{11.31}$$

11.4.5. Lax-Wendroff

Stabilitás:

$$C^2 + 2s \leq 1 \tag{11.32}$$

Oszcillációmentesség:

$$R_{cell} \leq 2 \tag{11.33}$$

11.4.6. Crank-Nicholson

Feltétel nélkül stabil, de

$$R_{cell} \leq 2 \tag{11.34}$$

az oszcillációmentesség feltétele.

11.5. 2 dimenziós transzport egyenlet

A több dimenziós transzport egyenlet nem sokkal bonyolultabb az 1 dimenziósnál, lényegében csak az implicit módszerek válnak sokkal költségesebbé. Itt is lehet operátor bontást használni: ADI illetve közelítő faktorizáció.

12. fejezet

Teljes Variációt Csökkentő módszerek

Mint az előző fejezet szimulációiból kiderült, a hagyományos lineáris diszkretizációs módszerek elég rosszul működnek a konvekciós egyenletre, és általában a hiperbolikus PDE-kre. Az első rendű módszerek nagy amplitúdó hibát okoznak, míg a másodrendű módszerek nagy részénél a fázis hiba miatt lesz a numerikus megoldás pontatlan. Még a viszonylag elfogadható eredményt adó Lax-Wendroff/MacCormack módszer is kisebb numerikus oszcillációt produkált a fizikai hullám megoldásban. Nemlineáris egyenletekre pedig még a Lax-Wendroff és MacCormack módszerek is nagyon gyengén működnek.

12.1. Teljes Variáció

Olyan módszereket keresünk, melyek a PDE analitikus megoldásainak bizonyos tulajdonságait diszkrét értelemben is megőrzik, és ezáltal a numerikus megoldás pontosabb lesz. Az egyik leghasznosabb ilyen tulajdonság az, hogy a teljes variáció időben előre haladva nem nő.

12.1.1. Teljes variáció definíciója

A teljes variáció matematikai definíciója a következő

$$TV(U) = \sup \sum_{j=1}^{N-1} |U(x_{j+1}) - U(x_j)| \tag{12.1}$$

ahol $0 \leq x_1 < x_2 < \ldots < x_N \leq L$ egy tetszőlegesen választott pont halmaz a $[0, L]$ számítási tartományon belül. A szuprémumot az összes lehetséges

ponthalmazra vesszük. A teljes variáció az egymást követő extrémumok közötti különbségek abszolútértékének összege. Differenciálható $U(x)$ függvény esetén

$$TV(U) = \int_0^L |U'(x)|dx \tag{12.2}$$

A diszkretizált megoldásra a teljes variáció definíciója

$$TV(U) = \sum_j \left|U_{j+1}^n - U_j^n\right| \tag{12.3}$$

Két dimenzióra is általánosítható a teljes variáció definíciója

$$\begin{aligned} TV(U) &= \limsup_{\epsilon\to 0} \frac{1}{\epsilon} \int_0^{L_x} \int_0^{L_y} |U(x+\epsilon, y) - U(x,y)|dxdy \\ &+ \limsup_{\epsilon\to 0} \frac{1}{\epsilon} \int_0^{L_x} \int_0^{L_y} |U(x, y+\epsilon) - U(x,y)|dxdy \end{aligned} \tag{12.4}$$

ami differenciálható függvényekre nyilván

$$TV(U) = \int_0^{L_x} \int_0^{L_y} \left|\frac{\partial U}{\partial x}\right| + \left|\frac{\partial U}{\partial y}\right| dxdy \tag{12.5}$$

Ez a definíció szintén könnyen átvihető a diszkretizált megoldásra:

$$TV(U) = \sum_{j=1}^{N-1} \sum_{k=1}^{M-1} |U(x_{j+1,k}) - U(x_{j,k})| + |U(x_{j,k+1}) - U(x_{j,k})| \tag{12.6}$$

12.1.2. Teljes variáció csökkenése

Skalár hiperbolikus PDE-re bebizonyítható, hogy a megoldás teljes variációja nem nő, azaz

$$TV\left(U(.,t_2)\right) \leq TV\left(U(.,t_1)\right) \qquad \text{ha } t_2 \geq t_1 \tag{12.7}$$

Ha a diszkretizált megoldás is teljesíti ezt a feltételt, azaz

$$TV(U^{n+1}) \leq TV(U^n) \tag{12.8}$$

ami például 1 dimenzióban úgy írható, hogy

$$\sum_j \left|U_{j+1}^{n+1} - U_j^{n+1}\right| \leq \sum_j \left|U_{j+1}^n - U_j^n\right| \tag{12.9}$$

akkor a diszkretizációt **teljes variációt csökkentő** (TVD, total variation diminishing) módszernek nevezzük.

12.1.3. Tételek TVD módszerekre

Korábban láttuk, hogy nemlineáris egyenletekre illetve gyenge megoldásokra Lax ekvivalencia tétele nem alkalmazható, azaz a konzisztenciából és a lineáris stabilitásból nem következik a nemlineáris stabilitás. A Lax-Wendroff tétel is csak annyit mond ki, hogy ha a konzervatív módszer konvergál, akkor az helyes gyenge megoldáshoz konvergál. TVD módszerekre ennél sokkal többet mondhatunk: **Ha egy diszkretizáció konzervatív, konzisztens és TVD, akkor az konvergens, és jó gyenge megoldáshoz konvergál.**

Mivel a TVD tulajdonság nem könnyen bizonyítható, érdemes néhány egyszerűbb tulajdonságot bevezetni, melyek a TVD tulajdonsággal kapcsolatosak.

Kezdjük a **monoton módszerek** definíciójával. Hiperbolikus PDE-k gyenge megoldásai közül az úgynevezett entrópia megoldásokra igaz a monoton tulajdonság, azaz hogy ha $V(x,0) \geq U(x,0)$ minden x-re, akkor $V(x,t) \geq U(x,t)$ minden x-re és t-re. A diszkretizáció monoton tulajdonságú, ha igaz rá, hogy ha $V_j^0 \geq U_j^0$ minden j-re, akkor $V_j^n \geq U_j^n$ minden j-re és n-re. A monotonitásnak egy könnyen ellenőrizhető szükséges és elégséges feltétele, ha be tudjuk látni, hogy $\partial U_j^{n+1}/\partial U_i^n \geq 0$. Ekkor ugyanis ha bármelyik rácspontban növeljük U_i^n-t, akkor a következő lépésben U_j^{n+1} is nagyobb lesz, azaz ha az n-dik lépésben a megoldást mindenhol nagyobbra változtatjuk, akkor az $n+1$-dik lépésben is nagyobb lesz. Például az áramlásirányú deriváltat használó upwind módszerre

$$T_j^{n+1} = (1-C)T_j^n + CT_{j-1}^n \tag{12.10}$$

mindkét együttható pozitív, ha $1 > C > 0$, azaz a módszer monoton.

Másodszor definiáljuk a **monotonitás megőrzését**, ami hiperbolikus PDE-k gyenge megoldásaira általában igaz: ha $U(x,0)$ egy monoton függvény, akkor $U(x,t)$ is az lesz. Ennek diszkrét megfelelője: ha $U_{j+1}^0 > U_j^0$ minden j-re, akkor $U_{j+1}^n > U_j^n$ minden j-re és n-re. A monotonitást megőrző módszerek nem keltenek numerikus oszcillációkat.

A monoton, TVD, és monotonitást megőrző módszerek között fennállnak a következő kapcsolatok:

- A TVD módszerek megőrzik a megoldás monotonitását.
- A monoton módszerek egyben TVD tulajdonságúak is.

Mivel a diszkretizáció monoton voltát viszonylag könnyű ellenőrizni, remélhetnénk, hogy így megfelelő TVD módszereket konstruálhatunk. Azonban figyelembe kell venni a következő megszorító tételeket is

- Csak első rendű módszerek lehetnek monotonak
- Lineáris és monotonitást megőrző módszerek monotonak, és így csak első rendűek lehetnek
- A TVD módszerek pontossága lokális extrémumnál csak első rendű
- 2 dimenziós TVD módszerek csak elsőrendűek lehetnek

Ebből következik, hogy 1 dimenzióban csak nem lineáris módszer lehet másodrendű és TVD, két dimenzióban pedig egyáltalán nem létezik másodrendű TVD módszer. Szerencsére az 1 dimenziós másodrendű TVD módszerek 2 dimenziós másodrendű általánosításai meglehetősen jól működnek, annak ellenére, hogy nem egzaktul TVD módszerek.

12.2. Magasabb rendű TVD módszerek

A magasabbrendű TVD módszerek bevezetését az egy dimenziós lineáris konvekciós egyenlet

$$\frac{\partial U}{\partial t} + v\frac{\partial U}{\partial x} = 0 \tag{12.11}$$

diszkretizációjával kezdjük, sőt először azt is feltesszük, hogy $u > 0$. Később általánosítjuk a módszereket tetszőleges u-ra, nem lineáris skalár egyenletre, egyenletrendszerre, és több dimenzióra.

12.2.1. Fluxus limitált módszer lineáris konvekcióra

Harten tétele, kimondja, hogy ha a módszer felírható

$$U_j^{n+1} = U_j^n - c_{j-1}(U_j^n - U_{j-1}^n) + d_j(U_{j+1}^n - U_j^n) \tag{12.12}$$

alakban, és igaz az együtthatókra, hogy $c_j, d_j \geq 0$ valamint $c_j + d_j \leq 1$, akkor a módszer TVD tulajdonságú. Fontos megjegyezni, hogy a c és d együtthatók tetszőleges függvényei U-nak, függhetnek több rácspontbeli értéktől is. Ezt a tételt viszonylag könnyű bizonyítani az

$$U_{j+1}^{n+1} - U_j^{n+1} = (1 - c_j - d_j)(U_{j+1}^n - U_j^n) + d_{j+1}(U_{j+2}^n - U_{j+1}^n) + c_{j-1}(U_j^n - U_{j-1}^n) \tag{12.13}$$

egyenlőségből. A bal oldal abszolút értéke kisebb lesz mint a jobb oldalon álló tagok abszolút értékeinek összege

$$|U_{j+1}^{n+1} - U_j^{n+1}| \quad \leq \quad (1 - c_j - d_j)|U_{j+1}^n - U_j^n|$$

$$
\begin{aligned}
& +c_{j-1}|U_j^n - U_{j-1}^n| \\
& +d_{j+1}|U_{j+2}^n - U_{j+1}^n| \qquad (12.14)\\
\leq\ & |U_{j+1}^n - U_j^n| \\
& +c_{j-1}|U_j^n - U_{j-1}^n| - c_j|U_{j+1}^n - U_j^n| \\
& +d_{j+1}|U_{j+2}^n - U_{j+1}^n| - d_j|U_{j+1}^n - U_j^n| \qquad (12.15)
\end{aligned}
$$

ahol az első egyenlőtlenségnél kihasználtuk, hogy az együtthatók mind pozitívak, utána pedig felbontottuk az $(1-c_j-d_j)$ zárójelet. Ha az egyenlőtlenséget felösszegezzük j-re, akkor

$$\sum |U_{j+1}^{n+1} - U_j^{n+1}| \leq \sum |U_{j+1}^n - U_j^n| \qquad (12.16)$$

ami éppen a TVD feltétel. Itt felhasználtuk, hogy a többi összegek kiesnek, mivel csak egy-egy index eltolással különböznek egymástól.

A konvekciós egyenletet upwind módszerrel diszkretizálva

$$U_j^{n+1} = U_j^n - C(U_j^n - U_{j-1}^n) \qquad (12.17)$$

ahol $C = v\Delta t/\Delta x$ a Courant szám. A Harten féle alakkal összehasonlítva, az együtthatókra $c_{j-1} = C$-t és $d_j = 0$-t kapunk. Ha $v > 0$ és a CFL feltételnek eleget teszünk, azaz $0 \leq C \leq 1$, akkor teljesítettük Harten összes feltételét, azaz az upwind módszer TVD, viszont csak elsőrendű.

Írjuk fel a Lax-Wendroff módszert a konvekciós egyenletre

$$U_j^{n+1} = U_j^n - \frac{C}{2}(U_{j+1}^n - U_{j-1}^n) + \frac{C^2}{2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \qquad (12.18)$$

Ezt átírhatjuk az upwind fluxus és egy korrekciós tag összegére

$$U_j^{n+1} = U_j^n - C(U_j^n - U_{j-1}^n) - \frac{1}{2}C(1-C)(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \qquad (12.19)$$

Ha ezt konzervatív módon fluxusok különbségeként írjuk fel, azaz

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}(F_{j+1/2}^n - F_{j-1/2}^n) \qquad (12.20)$$

akkor a numerikus fluxus

$$F_{j+1/2}^n = vU_j^n + \frac{1}{2}v(1-C)(U_{j+1}^n - U_j^n) \qquad (12.21)$$

Itt az első tag az upwind fluxus, a második a korrekció, ami Lax-Wendroff fluxusra egészíti ki. A korrekciós tag limitálását a ϕ fluxus limiterrel végezzük:

$$F_{j+1/2}^n = vU_j^n + \frac{1}{2}v(1-C)(U_{j+1} - U_j)\phi_{j+1/2} \qquad (12.22)$$

Ha $\phi = 0$ akkor az elsőrendű upwind fluxust kapjuk, ha pedig $\phi = 1$, akkor a másodrendű Lax-Wendroff fluxust. Ha ezt visszaírjuk a konzervatív diszkretizációba, akkor az átírható a Harten féle alakba a következő definíciókkal

$$\begin{aligned} c_{j-1} &= C + \frac{1}{2}(1-C)C\left[\frac{\phi_{j+1/2}(U_{j+1}-U_j) - \phi_{j-1/2}(U_j - U_{j-1})}{U_j - U_{j-1}}\right] \\ d_j &= 0 \end{aligned} \tag{12.23}$$

Ha $0 \leq c_j \leq 1$ akkor teljesítettük Harten feltételeit, és a módszer TVD lesz. Vezessük be a meredekségek hányadosát

$$\theta_{j+1/2} = \frac{U_j - U_{j-1}}{U_{j+1} - U_j} \tag{12.24}$$

a CFL stabilitás feltétel alapján feltehetjük, hogy $C < 1$, továbbá $u > 0$ miatt $C > 0$, így a $0 \leq c_j \leq 1$ feltételből

$$\frac{-2}{1-C} \leq \frac{\phi_{j+1/2}}{\theta_{j+1/2}} - \phi_{j-1/2} \leq \frac{2}{C} \tag{12.25}$$

egyenlőtlenségeket kapunk. Célszerű $\phi_{j+1/2}$-t mint $\theta_{j+1/2}$ függvényét definiálni, hiszen, akkor érdemes a másodrendű Lax-Wendroff fluxust használni, ha a megoldás sima, azaz $\theta \approx 1$, míg az elsőrendű upwind fluxust ott érdemes használni, ahol θ távol van 1-től. Függhetne továbbá ϕ a C értékétől is, de a gyakorlatban kiderült, hogy az így nyert limitáló függvények nem különösebben jobbak, mint az egyszerűbb csak θ-tól függőek. Tegyük fel tehát, hogy $\phi_{j+1/2}$ csak $\theta_{j+1/2}$-től függ. Ebben az esetben $0 < C < 1$ miatt a következő C értékétől független feltételt kell kielégíteni

$$\left|\frac{\phi_{j+1/2}}{\theta_{j+1/2}} - \phi_{j-1/2}\right| \leq 2 \tag{12.26}$$

Ahol a meredekség előjelet vált, azaz lokális extrémum van, ott érdemes az upwind módszert használni, azaz $\phi(\theta) = 0$ ha $\theta \leq 0$. Pozitív θ-ra pedig elegendő a következő feltételeket teljesíteni:

$$0 \leq \frac{\phi(\theta)}{\theta}, \phi(\theta) \leq 2 \tag{12.27}$$

hiszen így a két pozitív tag különbségének abszolút értéke biztosan kettőnél kisebb lesz. További feltétel, hogy $\phi(1) = 1$, azaz ha a meredekség nem változik, akkor a Lax-Wendroff fluxust kell venni, különben a módszer nem lesz másodrendű.

Számos olyan függvény van, ami ezeket a feltételeket teljesíti. A gyakorlatban sokat használják a következő fluxus limitáló függvényeket:

- minmod: $\phi(\theta) = \max(0, \min(1, \theta))$
- MC: $\phi(\theta) = \max(0, \min(2, 2\theta, (\theta+1)/2))$
- superbee: $\phi(\theta) = \max(0, \min(1, 2\theta), \min(\theta, 2))$
- beta: $\phi(\theta) = \max(0, \min(1, \beta\theta), \min(\theta, \beta))$
- van Leer: $\phi(\theta) = (|\theta| + \theta)/(|\theta| + 1)$

A minmod limiter a legdiffúzívabb, mert ϕ-t a lehető legkisebbre választja, azaz az elsőrendű upwind fluxust preferálja. Az MC (monoton centrális) limiter sokkal kevésbé diffúzív, míg a superbee mindenhol a maximális ϕ-t használja, ami viszont a megoldás túlzottan sarkossá válásához vezethet. A beta limiterben vagy egy szabad parameter $1 \leq \beta \leq 2$. A $\beta = 1$ a minmod limitert, a $\beta = 2$ a superbee limitert adja. Az utolsó van Leer limiter előnye, hogy differenciálható, ami egyensúlyi megoldás keresésénél javítja a konvergencia esélyét.

12.2.2. Fluxus limitált módszer általánosítása

Ha a v sebesség iránya tetszőleges, akkor az upwind módszert az

$$U_j^{n+1} = U_j^n - \frac{C}{2}(U_{j+1}^n - U_{j-1}^n) + \frac{|C|}{2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \tag{12.28}$$

alakban írhatjuk fel, amihez az

$$F_{j+1/2}^{\text{upwind}} = \frac{v}{2}(U_j + U_{j+1}) - \frac{|v|}{2}(U_{j+1} - U_j) \tag{12.29}$$

numerikus fluxus tartozik. A fluxus limitált Lax-Wendroff fluxus ebből

$$F_{j+1/2} = F_{j+1/2}^{\text{upwind}} + \frac{v}{2}(\text{sgn}(C) - C)(U_{j+1} - U_j)\phi_{j+1/2} \tag{12.30}$$

ahol az sgn függvény $+1$ pozitív és -1 negatív argumentumokra. Felhasználtuk, hogy C és v előjele azonos, így $v\text{sgn}(C) = |v|$. A $\phi_{j+1/2}$ limitert most is a $\theta_{j+1/2}$ meredekség hányadosaként írjuk fel, csakhogy most az upwind irány szerint választjuk a két meredekséget

$$\theta_{j+1/2} = \frac{U_{j'+1} - U_{j'}}{U_{j+1} - U_j} \qquad \text{ahol} \quad j' = j - \text{sgn}(C) \tag{12.31}$$

Ha $C > 0$, akkor $j' = j - 1$ és a korábbi definíciót kapjuk.

12.2.3. Meredekség limitált módszerek

Ezek a módszerek Godunov elsőrendű módszerén alapulnak, mely az alábbi lépésekből áll

- Tekintsük a megoldást konstansnak minden cellában
- Oldjuk meg a cellahatárokon fellépő Riemann problémákat
- Átlagoljuk ki a megoldást minden cellára

A Riemann probléma kezdeti feltétele egy lépcsős függvény, és ennek időfejlődését kívánjuk meghatározni. Konvekciós egyenlet esetén ez a lépcső v sebességgel fog haladni. Hidrodinamikában a kezdeti szakadásból egy lökéshullám, egy kontakt szakadási felület és egy tágulási hullám keletkezik, és ez analitikusan egzaktul kiszámítható. Magnetohidrodinamikában még bonyolultabb a helyzet, analitikusan már nem is lehet megoldani.

Godunov módszeréről könnyű belátni, hogy TVD: az első lépésben, amikor a diszkrét megoldást cellánként konstans függvénnyel helyettesítjük, a teljes variáció nem változik. A második lépésben a hiperbolikus PDE egzakt megoldását számítjuk ki, ami szintén nem növeli a teljes variációt. Végül a harmadik lépés átlagolása szintén nem növelheti az extrémumokat, azaz a teljes variáció ekkor sem nő. Az első lépésben vett cellánként konstans közelítés miatt azonban a Godunov módszer csak elsőrendű.

Megmutatjuk, hogy lineáris konvekciós egyenletre a Godunov módszer azonos az upwind módszerrel. A 2. és 3. lépést helyettesíthetjük azzal, hogy Δt ideig integráljuk az egyes cellákba be- és kiáramló egzakt fluxusokat, és ezzel az értékkel változtatjuk U_j^n-t. A konvekciós egyenletre az egzakt fluxus az áramlásirányból vett U szorozva a v sebességgel. Ha pl. $v > 0$, akkor a $j - 1/2$ határon a j-dik cellába $\Delta t v U_{j-1}$ áramlik be, és a $j + 1/2$ határon $\Delta t v U_j$ áramlik ki, így a cellára integrált U változása

$$\Delta x U_j^{n+1} = \Delta x U_j^n + \Delta t v (U_{j-1}^n - U_j^n) \tag{12.32}$$

ami éppen az upwind módszerrel egyezik meg $v > 0$ esetben.

Természetesen egyenletrendszerek esetében a Godunov módszer már nem ilyen triviális, mivel a cella határon fellépő Riemann probléma megoldása sem az. Ezt a Riemann problémát egzaktul vagy közelítően kell megoldani.

A Godunov módszer pontossága javítható, ha a cellánként konstans függvényeket cellánként lineáris függvényekkel helyettesítjük, így a cella határokon a Riemann problémát az

$$\begin{aligned} U_{j+1/2}^L &= U_j^n + \frac{1}{2}\overline{\Delta U}_j^n \\ U_{j+1/2}^R &= U_{j+1}^n - \frac{1}{2}\overline{\Delta U}_{j+1}^n \end{aligned} \tag{12.33}$$

értékekre kell megoldánunk, ahol $\overline{\Delta U_j}$ a j-dik cellában vett meredekség. A legegyszerűbb választások, például $\Delta U_j = (U_{j+1} - U_{j-1})/2$ vagy $\Delta U_j = U_{j+1} - U_j$, ezek azonban nem vezetnek TVD módszerhez, ugyanis új extrémumok keletkezhetnek!

A $\overline{\Delta U}^n$ meredekségeket tehát úgy kell limitálni, hogy a TVD tulajdonság ne sérüljön. Legegyszerűbb a $\overline{\Delta U}_j^n$ meredekséget a $\Delta U_{j+1/2} = U_{j+1} - U_j$ és a $\Delta U_{j-1/2} = U_j - U_{j-1}$ meredekségek függvényeként felírni. A következő feltételeket kell teljesíteni ahhoz, hogy ne keletkezzen új extrémum:

- Ha $\Delta U_{j+1/2}$ és $\Delta U_{j-1/2}$ előjelei különböznek, akkor $\overline{\Delta U}_j = 0$-t kell választanunk.
- Ha az előjelek megegyeznek (pl. pozitívak), akkor $\overline{\Delta U}_j < 2\Delta U_{j+1/2}$, különben $U^L_{j+1/2} > U_{j+1}$ lenne, hasonlóan $\overline{\Delta U}_j < 2\Delta U_{j-1/2}$, különben $U^R_{j-1/2} < U_{j-1}$ lenne.
- Ha a két oldalon a meredekség megegyezik, akkor a limitált meredekségnek is ennyinek kell lennie, azaz $\Delta U_{j+1/2} = \Delta U_{j-1/2} = \overline{\Delta U}_j$

Ezek a feltételek egy az egyben analógok a $\phi(\theta)$ fluxus limiterre kapott feltételekkel: $\phi(\theta < 0) = 0$, $\phi < 2$, $\phi/\theta < 2$ és $\phi(1) = 1$. Ez könnyen látható, ha a

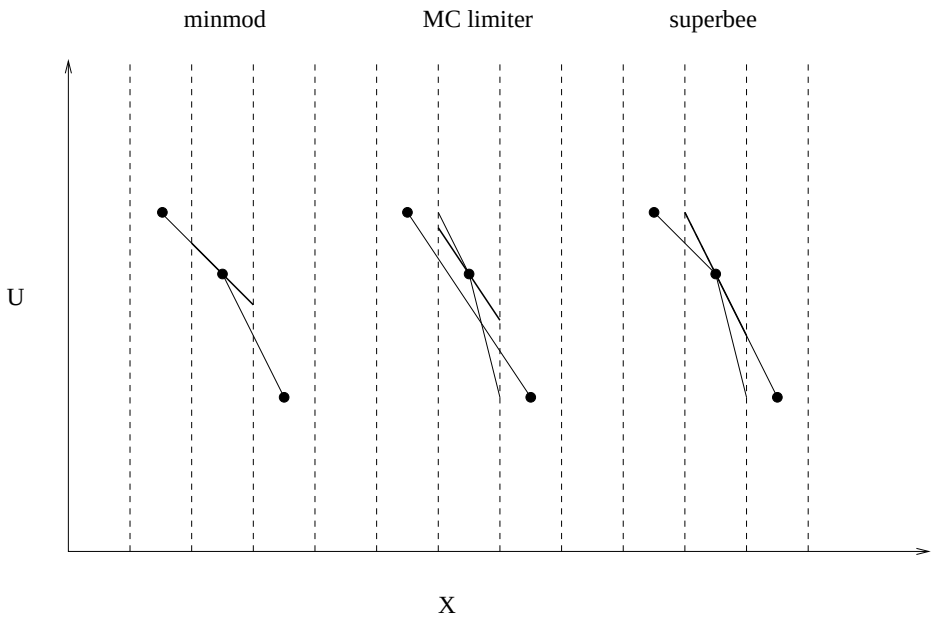
$$\overline{\Delta U}(\Delta U_{j-1/2}, \Delta U_{j+1/2}) = \Delta U_{j+1/2}\phi(\theta) \tag{12.34}$$

megfeleltetést vesszük. Ebben a részben geometriai megfontolásokból kaptuk meg ugyanazokat az egyenlőtlenségeket, mint amit a fluxus limiternél algebrai levezetésből nyertünk.

Íme néhány meredekség limitáló függvény:

- minmod: $\overline{\Delta U}_j = \text{minmod}\left(\Delta U_{j-1/2}, \Delta U_{j+1/2}\right)$
- MC: $\overline{\Delta U}_j = \text{minmod}\left(2\Delta U_{j-1/2}, 2\Delta U_{j+1/2}, \frac{1}{2}\Delta U_{j-1/2} + \frac{1}{2}\Delta U_{j+1/2}\right)$
- superbee: $\overline{\Delta U}_j = s\max[0, \min\left(2|\Delta U_{j+1/2}|, s\Delta U_{j-1/2}\right), \min\left(|\Delta U_{j+1/2}|, 2s\Delta U_{j-1/2}\right)]$

ahol $s = \text{sgn}(\Delta U_{j+1/2})$, továbbá a minmod függvény nulla, ha az argumentumok előjele nem egyezik meg, egyébként pedig a legkisebb abszolút értékű argumentumot adja vissza (előjelesen).



12.1. ábra. Háromféle meredekség limitáló szemléltetése

12.2.4. TVD Lax-Friedrichs módszer

Vegyünk egy tetszőleges konzervatív formában felírt PDE-t egy dimenzióban:

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} = 0 \tag{12.35}$$

A lineáris konvekciós egyenlet ennek speciális esete ($F = vU$). A fenti PDE konzervatív Lax-Friedrichs diszkretizációja a következő:

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}\left(F_{j+1/2} - F_{j-1/2}\right) + \frac{1}{2}\left(\Phi_{j+1/2} - \Phi_{j-1/2}\right) \tag{12.36}$$

ahol

$$F_{j+1/2} = \frac{F_j + F_{j+1}}{2} \tag{12.37}$$

$$\Phi_{j+1/2} = U_{j+1} - U_j \tag{12.38}$$

Itt Φ a stabilitáshoz szükséges numerikus diffúziót biztosítja. A numerikus diffúzió egy $\alpha = \Delta x^2/(2\Delta t)$ diffúziós együtthatójú fizikai diffúzió diszkretizációjának felel meg. A CFL stabilitási feltétel miatt $\Delta t \propto \Delta x$, így a numerikus diffúzió első rendű hibát okoz, azaz a Lax-Friedrichs módszer időben és térben elsőrendű.

Skalár egyenletekre könnyen megmutatható, hogy a Lax-Friedrichs módszer monoton, és így TVD is, ha a CFL stabilitás limiten belül vagyunk, ugyanis

$$\frac{\partial U_j^{n+1}}{\partial U_j^n} = 0 \tag{12.39}$$

$$\frac{\partial U_j^{n+1}}{\partial U_{j\pm1}^n} = \frac{1}{2}\left[1 \mp \frac{\Delta t}{\Delta x}\left(\frac{\partial F}{\partial U}\right)_{j\pm1}\right] > 0 \tag{12.40}$$

A CFL feltétel miatt $c\Delta t/\Delta x < 1$, ahol $c = |\partial F/\partial U|$ a hullám terjedési sebessége.

A diffúzió valamelyest csökkenthető, ha a numerikus fluxust a következőképp definiáljuk:

$$\Phi_{j+1/2} = \frac{\Delta t}{\Delta x} c_{j+1/2}^{\max}(U_{j+1} - U_j) \tag{12.41}$$

ahol $c_{j+1/2}^{\max} = \max(c_j, c_{j+1})$. Így az ekvivalens fizikai diffúziós együttható $\alpha = c^{\max}\Delta x/2$-re csökkent, ami még mindig első rendű hibát okoz. A monotonitás a csökkentett diffúzió mellett is fennáll, hiszen

$$\frac{\partial U_j^{n+1}}{\partial U_j^n} = 1 - \frac{1}{2}\frac{\Delta t}{\Delta x}(c_{j+1/2} + c_{j-1/2}) \tag{12.42}$$

$$\frac{\partial U_j^{n+1}}{\partial U_{j\pm1}^n} = \frac{1}{2}\frac{\Delta t}{\Delta x}\left[c_{j\pm1/2} \mp \left(\frac{\partial F}{\partial U}\right)_{j\pm1}\right] > 0 \tag{12.43}$$

Ha most egy pillanatra visszagondolunk a lineáris konvekciós egyenletre, ahol $F = vU$ és $c^{\max} = |v|$, akkor láthatjuk, hogy a csökkentett diffúziójú Lax-Friedrichs módszer ebben az esetben azonos az upwind módszerrel tetszőleges irányú sebességre, azaz egy Godunov módszernek tekinthető. Alkalmazzuk tehát a Godunov módszernél említett eljárást, hogy magasabbrendű diszkretizációt kapjunk.

Meredekség limiterek segítségével a Lax-Friedrichs diszkretizáció másodrendűvé tehető anélkül, hogy a TVD tulajdonságot elvesztenénk. A másodrendű TVDLF módszerben a fluxusok

$$F_{j+1/2} = \frac{F(U_{j+1/2}^L) + F(U_{j+1/2}^R)}{2} \tag{12.44}$$

$$\Phi_{j+1/2} = \frac{\Delta t}{\Delta x} c_{j+1/2}^{\max}(U_{j+1/2}^R - U_{j+1/2}^L) \tag{12.45}$$

ahol $c_{j+1/2}^{\max}$-t többféleképpen is definiálhatjuk, pl. $c_{j+1/2}^{\max} = \max(c(U^L), c(U^R))$. Ez a diszkretizáció sima megoldásra térben másodrendű, hiszen a diffúzív fluxusban $U_{j+1/2}^R - U_{j+1/2}^L$ $O(\Delta x)$-szel kisebb mint az eredeti $U_{j+1} - U_j$ különbség.

Ugyanakkor a diszkretizáció időben csak első rendű, sőt, a TVD tulajdonság is csak $C < 0.5$-re áll fenn.

Ahhoz, hogy időben is másodrendű legyen a diszkretizáció, egy prediktor lépésre van szükségünk. Az egyik legjobb választás a Hancock prediktor:

$$U_j^{n+1/2} = U_j^n - \frac{\Delta t}{2\Delta x}\left[F(U_j^n + \overline{\Delta U}_j^n/2) - F(U_j^n - \overline{\Delta U}_j^n/2)\right] \tag{12.46}$$

amely $\Delta t/2$-vel lépteti előre a megoldást. Az $U^{n+1/2}$ prediktort a bal és jobb oldali extrapolált értékekhez használjuk fel:

$$\begin{aligned} U_{j+1/2}^L &= U_j^{n+1/2} + \frac{1}{2}\overline{\Delta U}_j^n \\ U_{j+1/2}^R &= U_{j+1}^{n+1/2} - \frac{1}{2}\overline{\Delta U}_{j+1}^n \end{aligned} \tag{12.47}$$

így ezek most már időben is másodrendű pontosak. Fontos megjegyezni, hogy a limitált meredekségeket továbbra is az n-dik lépésből számoljuk, különben a módszer nem lenne TVD. Az új U^L és U^R definíciókkal számolt F és Φ most már térben és időben másodrendű módszert eredményez, sőt a TVD tulajdonság minden $C < 1$-re fennáll.

Ismét tekintsünk a konvekciós egyenletet, mint speciális esetet:

$$U_j^{n+1/2} = U_j^n - \frac{\Delta t}{2\Delta x} v\overline{\Delta U}_j^n = U_j^n - \frac{C}{2}\overline{\Delta U}_j^n \tag{12.48}$$

és így

$$U_{j+1/2}^L = U_j^{n+1/2} + \frac{1}{2}\overline{\Delta U}_j^n = U_j^n + \frac{1}{2}(1-C)\overline{\Delta U}_j^n \tag{12.49}$$

A konvekciós egyenlet és $v > 0$ esetén

$$F_{j+1/2} + \frac{1}{2}\Phi_{j+1/2} = \frac{1}{2}\left[F(U_{j+1/2}^L) + F(U_{j+1/2}^R)\right] + \frac{|v|}{2}(U_{j+1/2}^L - U_{j+1/2}^R) = vU_{j+1/2}^L \tag{12.50}$$

azaz az áramlásirányú fluxust használjuk. Tehát a teljes numerikus fluxus

$$F'_{j+1/2} = vU_j^n + \frac{1}{2}v(1-C)\overline{\Delta U}_j^n \tag{12.51}$$

ami a $\overline{\Delta U}_j^n = (U_{j+1} - U_j)\phi(\theta_{j+1/2})$ helyettesítéssel éppen a limitált Lax-Wendroff fluxust adja. Tehát az itt tárgyalt meredekség és a fluxuslimitált módszerek a konvekciós egyenletre ekvivalensek egymással. Természetesen nemlineáris egyenletre, vagy egyenletrendszerre a két módszerrel kapott diszkretizációk már lényegesen különböznek egymástól.

12.2.5. Általánosítás egyenletrendszerekre

A TVDLF módszer triviálisan általánosítható egyenletrendszerekre. Lényegében csak annyi változik, hogy $\partial F/\partial U$ most nem skalár, hanem mátrix, így a hullám terjedési sebességét a legnagyobb abszolút értékű sajátértékkel kell becsülni

$$c = \max_k(|\lambda_k|) \tag{12.52}$$

ahol λ_k a k-dik sajátértéket jelöli. Ezzel a választással elértük, hogy a TVD tulajdonság lineáris egyenletrendszerre fennmaradjon, illetve nem lineáris egyenletrendszerre is általában oszcillációmentes megoldást kapunk.

A numerikus diffúzió, azaz Φ fluxus csökkenthető, ha az egyenletrendszert karakterisztikus változókra írjuk át, és mindegyik karakterisztikus változóhoz a saját sajátértékével arányos diffúziót rendelünk. Ezt TVD-MUSCL módszernek nevezzük.

A TVD MUSCL módszert először lineáris PDE rendszerre vezetjük le, ami

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} = 0 \tag{12.53}$$

formában írható. Ha a PDE rendszer hiperbolikus, akkor az A mátrix diagonalizálható

$$A = R\Lambda R^{-1} \tag{12.54}$$

ahol R a jobb oldali r^k sajátvektorokat mint oszlopokat, R^{-1} pedig a bal oldali l^k sajátvektorokat mint sorokat, míg a Λ diagonális mátrix a λ^k sajátértékeket tartalmazza. A hiperbolikusság miatt a sajátvektorok és sajátértékek léteznek és valósak. Definiáljuk a karakterisztikus változókat

$$V = R^{-1}U \tag{12.55}$$

formában. Ekkor a lineáris PDE rendszert balról R^{-1}-gyel szorozva

$$R^{-1}\frac{\partial U}{\partial t} + \Lambda R^{-1}\frac{\partial U}{\partial x} = \frac{\partial V}{\partial t} + \Lambda\frac{\partial V}{\partial x} = 0 \tag{12.56}$$

Azaz a karakterisztikus változókban független konvekciós egyenleteket kaptunk, melyekben a sebességet a λ_k sajátértékek adják. Tehát a megoldást a karakterisztikus hullámok terjedésének lineáris kombinációja adja.

Ezekre külön-külön alkalmazhatjuk a skalár advekciós egyenletre kifejlesztett TVD módszert. A TVD-MUSCL módszer numerikus fluxusát ezért

$$\Phi = \frac{\Delta t}{\Delta x}\sum_k r^k\,|\lambda^k|\,l^k\cdot(U^R - U^L) \tag{12.57}$$

formában írhatjuk. Itt az $l^k \cdot (U^R - U^L)$ a k-dik karakterisztikus változó ugrása a cella határon, $|\lambda^k|$ a k-dik karakterisztikus hullám sebessége, végül az r^k-val való szorzás arra szolgál, hogy visszatérjünk a karakterisztikus változókról a normál változókra. Ha a sajátértékeket mind a maximális $c^{\max} = \max_k(|\lambda_k|)$-val helyettesítjük, akkor a TVDLF módszert kapjuk vissza, hiszen a jobb oldali r^k és bal oldali l^k sajátvektorok szorzata az identitás mátrixot adja.

Nemlineáris PDE esetén a $\partial F/\partial U$ mátrix helyettesíti A-t, így ennek jobb oldali r_k, bal oldali l^k sajátvektorait és λ_k sajátértékeit vesszük.

12.2.6. Általánosítás több dimenzióra

Az összes eddig tárgyalt módszer általánosítható 2 vagy 3 dimenzióra az operátor bontás segítségével, ami megtartja a másodrendű pontosságot térben és időben. Természetesen a több dimenziós diszkretizáció már nem lesz TVD, hiszen mint említettük 2 dimenzióban csak elsőrendű módszer lehet TVD.

A másik lehetőség az általánosításra, hogy a különböző irányú (limitált) fluxusokat egyszerűen összeadjuk. Ez a megközelítés a Harten-féle fluxus limitált Lax-Wendroff módszerre nem működik, instabil diszkretizációt kapunk. A meredekség limitált TVDLF és TVD-MUSCL módszerek azonban könnyen általánosíthatók ilyen módon is.

Csupán arra kell vigyázni, hogy a diffúziós fluxusok ne eredményezzenek instabilitást. Ha a limitált meredekségek nullák, akkor a diffúziós együtthatók értéke $\alpha_x = \Delta x c^x/2$ illetve $\alpha_y = \Delta y c^y/2$, ahol c^x és c^y az x illetve y irányban vett maximális hullámsebesség. Ebből a dimenziótlan együtthatók $s_x = \Delta t \alpha_x/\Delta x^2 = \Delta t c^x/(2\Delta x)$ és $s_y = \Delta t c^y/(2\Delta y)$. Mint a 2 dimenziós diffúziós egyenletre alkalmazott FTCS módszernél említettük, a stabilitás feltétele $s_x + s_y < 1/2$, így a 2 dimenziós TVDLF és TVD-MUSCL módszereknél az időlépésre a stabilitás feltétel

$$\Delta t < \min_{j,k} \frac{1}{\frac{c^x_{j+1/2,k}}{\Delta x} + \frac{c^y_{j,k+1/2}}{\Delta y}} \tag{12.58}$$

13. fejezet

A mágneses tér divergenciája

Ebben a fejezetben áttekintünk néhány módszert, ami a numerikus magneto-hidrodinamikában (MHD) a mágneses tér divergenciáját próbálja meg diszkrét értelemben is nullává tartani. Ehhez hasonló módszereket az összenyomhatatlan folyadékoknál a sebességtér divergenciájára, illetve elektro-mágneses egyenletrendszernél az elektromos tér divergenciájára is alkalmaznak.

13.1. Az MHD egyenletek

Az MHD egyenletek megmaradási formában kifejezik a tömeg, a momentum, az energia és a mágneses indukció megmaradását:

$$\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho \mathbf{v}) = 0 \tag{13.1}$$

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \mathrm{div}(\mathbf{v}\rho\mathbf{v} - \frac{1}{\mu_0}\mathbf{B}\mathbf{B}) + \mathbf{grad}(p + \frac{1}{2\mu_0}\mathbf{B}^2) = 0 \tag{13.2}$$

$$\frac{\partial e}{\partial t} + \mathrm{div}[\mathbf{v}(e + p + \frac{1}{2\mu_0}\mathbf{B}^2) - \frac{1}{\mu_0}\mathbf{B}\mathbf{B}\cdot\mathbf{v}] = 0 \tag{13.3}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \mathbf{rot}\mathbf{E} = 0 \tag{13.4}$$

ahol $\mathbf{B}$ a mágneses térerősség, μ_0 a mágneses permeabilitás vákuumban,

$$\mathbf{E} = -\mathbf{v} \times \mathbf{B} \tag{13.5}$$

az elektromos térerősség az ideális MHD közelítésben, valamint e a teljes energiasűrűség. A p nyomás a

$$p = (\gamma - 1)\left(e - \frac{1}{2}\rho\mathbf{v}^2 - \frac{1}{2\mu_0}\mathbf{B}^2\right) \tag{13.6}$$

egyenletből számítható ki. Itt γ az adiabatikus index (a nyomásra és térfogatra vett fajhők hányadosa). A továbbiakban a mágneses tér egységeit úgy választjuk meg, hogy $\mu_0 = 1$ legyen.

A mágneses térre van még egy egyenlet, ami azt fejezi ki, hogy nincsenek mágneses monopólusok:

$$\mathrm{div}\mathbf{B} = 0 \tag{13.7}$$

Ez az egyenlet tulajdonképpen a kezdeti feltételre vonatkozik! Vegyük ugyanis az indukciós egyenlet divergenciáját

$$\mathrm{div}\frac{\partial \mathbf{B}}{\partial t} = \frac{\mathrm{div}\mathbf{B}}{\partial t} = -\mathrm{div}\mathbf{rot}\mathbf{E} = 0 \tag{13.8}$$

Tehát ha $\mathrm{div}\mathbf{B} = 0$ kezdetben, akkor ez így marad a továbbiakban is.

13.2. Divergencia mentesség numerikusan

Egy dimenziós problémákra a divergenciamentesség $\partial B_x/\partial x = 0$-ra egyszerűsödik, azaz B_x térben konstans, és az indukciós egyenletből $\partial B_x/\partial t = -(\mathbf{rot}\mathbf{E})_x$, mivel a rotáció x komponensében szereplő $\partial/\partial y$ és $\partial/\partial z$ operátorok egy dimenzióban azonosan nullát adnak. Tehát egy dimenziós MHD-ben B_x egy egyszerű paraméterré válik, és ezt a diszkretizáció is könnyen meg tudja valósítani.

Két vagy három dimenzióban azonban nincs rá garancia, hogy $\mathrm{div}\mathbf{B} = 0$ numerikusan is megmarad. Természetesen a hiba hasonlóan a többi diszkretizációs hibához 0-hoz tart, ha $\Delta x, \Delta t \to 0$. Azonban a numerikus tesztek eredményei azt mutatják, hogy a hiba felhalmozódik, és a megoldás nagyon pontatlan lesz. Egyes numerikus módszerek instabillá is válhatnak.

Többféle módszer használatos, amely a numerikus $\mathrm{div}\mathbf{B}$ hibát illetve a hiba következményeit csökkenti vagy megszünteti:

- 8 hullám módszer: módosítjuk az MHD egyenleteket úgy, hogy a numerikus hiba jól viselkedjen
- diffúziós kontroll: a numerikus $\mathrm{div}\mathbf{B}$ hibát parabolikus diffúzióval csökkentjük
- projekció: a $\mathrm{div}\mathbf{B}$ hibát elliptikus projekcióval minden időlépésben megszüntetjük
- hiperbolikus Lagrange szorzós módszer: a $\mathrm{div}\mathbf{B}$ hibát elpropagáljuk
- constrained transport: az indukció egyenletet úgy diszkretizáljuk, hogy $\mathrm{div}\mathbf{B}$ ne változzon időben

- vektor potenciál használata: a $\mathbf{B}$ mágneses tér helyett a $\mathbf{A}$ vektor potenciált használjuk. Mivel $\mathbf{B} = \mathbf{rotA}$ a mágneses tér garantáltan divergenciamentes.

13.3. 8-hullám módszer

A 8-hullám módszernél az MHD egyenleteket újra levezetjük úgy, hogy div$\mathbf{B}$ nem feltétlenül nulla:

$$\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho \mathbf{v}) = 0 \tag{13.9}$$

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \mathrm{div}(\mathbf{v}\rho\mathbf{v} - \mathbf{BB}) + \mathbf{grad}(p + \frac{1}{2}\mathbf{B}^2) = -(\mathrm{div}\mathbf{B})\mathbf{B} \tag{13.10}$$

$$\frac{\partial e}{\partial t} + \mathrm{div}[\mathbf{v}(e + p + \frac{1}{2}\mathbf{B}^2) - \mathbf{BB}\cdot\mathbf{v}] = -(\mathrm{div}\mathbf{B})\mathbf{B}\cdot\mathbf{v} \tag{13.11}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \mathbf{rotE} = -(\mathrm{div}\mathbf{B})\mathbf{v} \tag{13.12}$$

Az új forrás tagok a jobb oldalakon találhatók. Ez egy Galilei invariáns egyenletrendszer. Janhunen és Dellar egy Lorentz invariáns megoldást találtak, melyben az energia és a momentum egyenletben nincs extra forrástag.

A módszer onnan kapta nevét, hogy az eredeti MHD egyenletekben 7 karakterisztikus hullám van: 1 entrópia hullám, 2 Alfvén hullám, továbbá 2 lassú és 2 gyors magnetoakusztikus hullám. Ezek a hullámok az egy dimenziós MHD egyenletek linearizálásából nyerhetők. Mint említettük, 1 dimenzióban az MHD egyenletekben a B_x egy konstans paraméterré válik, így valójában csak 7 változó marad: a ρ sűrűség, a $\rho\mathbf{v}$ impulzus 3 komponense, az e energiasűrűség, valamint a mágneses tér B_y és B_z komponensei. Két vagy három dimenzióban azonban már 8 független változo van. Ha megengedjük mágneses monopólusok létezését, akkor megjelenik egy 8. hullám, ami a monopólusok, azaz a div$\mathbf{B}$-nek $\mathbf{v}$ sebességű haladását írja le. Az MHD egyenleteknek ezt a módosított formáját először Godunov javasolta.

Az új egyenletrendszer azonban már nem konzervatív. Ez bizonyos esetekben, igaz meglepően ritkán, problémát okoz: a gyenge megoldás nem helyes.

13.4. Diffúziós kontrol

A diffúziós kontrol módszerben az indukciós egyenletet a következőképpen módosítjuk:

$$\frac{\partial \mathbf{B}}{\partial t} = -\mathbf{rotE} + \eta\mathbf{grad}(\mathrm{div}\mathbf{B}) \tag{13.13}$$

ahol η egy pozitív skalár. Ha vesszük ennek a divergenciáját, akkor

$$\frac{\partial \mathrm{div}\mathbf{B}}{\partial t} = \mathrm{div}\left[\eta \mathbf{grad}(\mathrm{div}\mathbf{B})\right] \tag{13.14}$$

ami egy diffúziós egyenlet div$\mathbf{B}$-re nézve.

Az η együtthatót úgy érdemes megválasztani, hogy a diffúzió maximális legyen, de a diffúziós stabilitási feltétel ne követeljen a CFL stabilitási feltételnél kisebb időlépést. Ebből

$$\eta = D\frac{(\Delta x)^2}{\Delta t} \tag{13.15}$$

ahol $D < 1$ maximális értéke a konkrét diszkretizációtól függ.

13.5. Projekciós módszer

Az MHD egyenleteket két részlépésben oldjuk meg. Az első lépésben egy $\mathbf{B}^*$ megoldást kapunk, melynek divergenciája nem feltétlenül nulla. Egy ilyen vektor mező mindig felírható egy rotáció és egy gradiens összegeként

$$\mathbf{B}^* = \mathbf{rot}\mathbf{A} + \mathbf{grad}\phi \tag{13.16}$$

A megoldás fizikailag értelmes részét a $\mathbf{rotA}$ tag tartalmazza. Ha mindkét oldal divergenciáját vesszük, akkor egy Poisson egyenletet kapunk ϕ-re:

$$\mathrm{div}\mathbf{grad}\phi = \mathrm{div}\mathbf{B}^* \tag{13.17}$$

A Poisson egyenlet megoldásából kapott ϕ skalár függvény gradiensével módosítjuk $\mathbf{B}^*$-t:

$$\mathbf{B}^{n+1} = \mathbf{B}^* - \mathbf{grad}\phi \tag{13.18}$$

Könnyen látható, hogy

$$\mathrm{div}\mathbf{B}^{n+1} = \mathrm{div}\mathbf{B}^* - \mathrm{div}\mathbf{grad}\phi = 0 \tag{13.19}$$

azaz $\mathbf{B}^{n+1}$ divergenciamentes. Azonban nem nyilvánvaló, hogy a korrigált $\mathbf{B}^{n+1}$ mennyire lesz közel a jó megoldáshoz.

13.5.1. Minimális korrekció

Vizsgáljuk a következő problémát. Adott $\mathbf{B}^*$ mellett mi az a $\mathbf{B}$ megoldás, ami divergencia mentes és

$$||\mathbf{B} - \mathbf{B}^*||^2 = \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{B}_i - \mathbf{B}_i^*\right)^2 \tag{13.20}$$

minimális? A válasz Lagrange multiplikátor segítségével található meg. Ha Φ a Lagrange multiplikátora a divergencia mentesség mellékfeltételnek, akkor

$$\frac{\partial[d(\mathbf{B}) + \sum_i \Phi_i (\mathrm{div}\mathbf{B})_i]}{\partial B_j^x} = (B_j^x - B_j^{x,*}) + \sum_i \Phi_i D_{i,j}^x = 0 \tag{13.21}$$

ahol D^x az x szerinti derivált operátora. Hasonló egyenletet kapunk a B_j^y szerinti parciális deriválásból. Az egyenletek megoldása

$$\begin{aligned} B^x &= B^{x,*} - D^{x,T}\Phi \\ B^y &= B^{y,*} - D^{y,T}\Phi \end{aligned} \tag{13.22}$$

ahol T a transzponáltat jelöli. A mellékfeltételből

$$0 = D^x B^{x,*} + D^y B^{y,*} - (D^x D^{x,T} + D^y D^{y,T})\Phi \tag{13.23}$$

Ha figyelembe vesszük, hogy D^x és D^y antiszimmetrikus operátorok, akkor a 13.23 egyenlet a $-\mathrm{div}\mathbf{grad}\Phi = \mathrm{div}\mathbf{B}^*$ Poisson egyenlet diszkretizációja, míg a 13.22 egyenletek a $B = B^* + \mathbf{grad}\Phi$ diszkrét formában. Azaz $\phi = -\Phi$ választással éppen a projekciós módszer 13.17 és 13.18 egyenleteinek diszkretizációit kapjuk.

Tehát a projekciós módszer úgy távolítja el a mágneses tér divergenciáját, hogy közben minimálisan korrigálja a megoldást.

13.5.2. Gyenge megoldás projekcióval

Nem világos, hogyan viselkedik nem folytonos problémákra a projekciós módszer. Elvileg elképzelhető, hogy a Poisson egyenlet megoldásakor a szakadási felületnél lévő hibát az egész számítási tartományra szétszórjuk, és így az ugrásfeltételek elromlanak. Azonban a minimális korrekció alapján be lehet bizonyítani, hogy ez nincs így.

Legyen az alapmódszer (amit a projekció előtt alkalmazunk) térben és időben k illetve m rendű és konzervatív. Ekkor

$$||\mathbf{B}^* - \mathbf{B}^a|| < O(\Delta x^k, \Delta t^m) \tag{13.24}$$

Továbbá legyen $\mathbf{B}^c$ egy tetszőleges divergenciamentes konzervatív diszkretizáció:

$$||\mathbf{B}^c - \mathbf{B}^a|| < O(\Delta x^{k'}, \Delta t^{m'}) \tag{13.25}$$

Mivel a projekció adja a $\mathbf{B}^*$-hoz legközelebbi divergenciamentes megoldást

$$||\mathbf{B}^p - \mathbf{B}^*|| \leq ||\mathbf{B}^c - \mathbf{B}^*|| \tag{13.26}$$

A fenti egyenlőtlenségekből

$$\begin{aligned} ||\mathbf{B}^p - \mathbf{B}^a|| &\leq ||\mathbf{B}^p - \mathbf{B}^*|| + ||\mathbf{B}^* - \mathbf{B}^a|| \leq ||\mathbf{B}^c - \mathbf{B}^*|| + ||\mathbf{B}^* - \mathbf{B}^a|| \\ &\leq ||\mathbf{B}^c - \mathbf{B}^a|| + 2||\mathbf{B}^* - \mathbf{B}^a|| < O(\Delta x^{k''}, \Delta t^{m''}) \end{aligned} \quad (13.27)$$

azaz a projekcios megoldás is konzisztens és a rendje $k'' = \min(k, k') > 0$ illetve $m'' = \min(m, m') > 1$.

13.5.3. Projekció és diffúziós kontroll

Könnyen belátható, hogy a diffúziós kontroll módszer lényegében a projekciós módszer egyetlen iterációra redukálva, illetve úgy is felfogható, mint a projekciós módszer által definiált elliptikus egyenletnek a pszeudo-tranziens módszerrel történő megoldása. Természetesen a mágneses tér dinamikusan változik, így a diffúziós kontroll egyetlen időlépés alatt nem tudja nullára csökkenteni a mágneses tér divergenciáját, de a hibát jelentősen redukálja.

13.6. Hiperbolikus Lagrange multiplikátor

A szokásos indukciós egyenlet helyett oldjuk meg a következő egyenletrendszert:

$$\frac{\partial \mathbf{B}}{\partial t} = \mathbf{rot}\mathbf{E} - \mathbf{grad}\varphi \quad (13.28)$$

$$\mathcal{D}(\varphi) = -\mathrm{div}\mathbf{B} \quad (13.29)$$

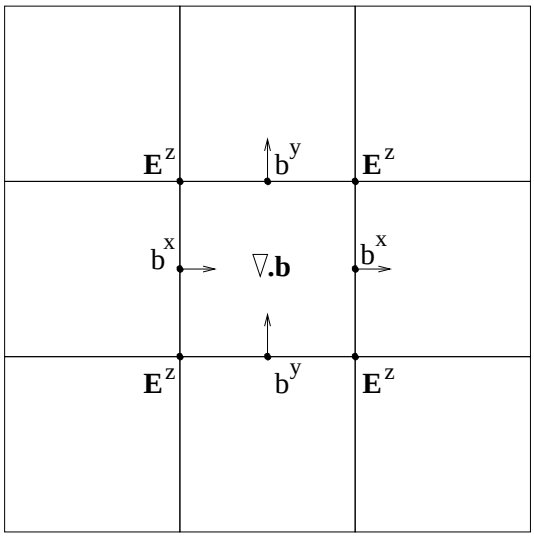
ahol φ egy skalárfüggvény, $\mathcal{D}$ pedig egy lineáris differenciáloperátor. Az operátor megválasztásától függően

- elliptikus: $\mathcal{D}(\varphi) = 0$ (ugyanaz, mint a projekcio)
- parabolikus: $\mathcal{D}(\varphi) = (1/c_p^2)\varphi$ (ugyanaz, mint a diffúziós kontrol)
- hiperbolikus: $\mathcal{D}(\phi) = (1/c_h^2)\partial\varphi/\partial t$ Új!

Megmutatható, hogy a hiperbolikus differenciáloperátorral a módosított egyenletrendszer $\pm c_h$ sebességgel propagálja a div$\mathbf{B}$ hibát.

13.7. Constrained Transport

Ha az elektromos teret a cellasarkokon, a mágneses teret pedig a cellaéleken diszkretizáljuk akkor elérhető, hogy a div**rot** $= 0$ azonosság numerikusan is fennmaradjon. Ezt a diszkretizációt Constrained Transport (CT, kb. limitált transzport) módszernek nevezik. Jelölje b^x és b^y a mágneses teret a



13.1. ábra. Constrained Transport diszkretizáció két dimenzióban

cellaéleken. A CT módszer egyszerű másodrendű véges differencia formulákat használ az indukció egyenlet diszkretizálására:

$$b^{x,n+1}_{j+1/2,k} = b^{x,n}_{j+1/2,k} - \Delta t \frac{E^z_{j+1/2,k+1/2} - E^z_{j+1/2,k-1/2}}{\Delta y} \tag{13.30}$$

$$b^{y,n+1}_{j,k+1/2} = b^{y,n}_{j,k+1/2} + \Delta t \frac{E^z_{j+1/2,k+1/2} - E^z_{j-1/2,k+1/2}}{\Delta x} \tag{13.31}$$

Könnyen meg lehet mutatni, hogy

$$(\mathrm{div}\mathbf{b})_{j,k} = \frac{b^x_{j+1/2,k} - b^x_{j-1/2,k}}{\Delta x} + \frac{b^y_{j,k+1/2} - b^y_{j,k-1/2}}{\Delta y} \tag{13.32}$$

egzaktul megmarad, azaz ha 0 volt kezdetben, akkor az is marad. Az elektromos teret többféleképpen lehet definiálni, pl.

$$\begin{aligned} E^z_{j+1/2,k+1/2} &= \frac{1}{8}(E^{z,n}_{j,k} + E^{z,n}_{j+1,k} + E^{z,n}_{j,k+1} + E^{z,n}_{j+1,k+1} \\ &+ E^{z,*}_{j,k} + E^{z,*}_{j+1,k} + E^{z,*}_{j,k+1} + E^{z,*}_{j+1,k+1}) \end{aligned} \tag{13.33}$$

A mágneses térre persze a cella közepén is szükség van, például a nyomás kiszámításánál. A cella közepén vett mágneses teret egyszerű átlagolással kaphatjuk meg

$$B^x_{j,k} = \frac{b^x_{j-1/2,k} + b^x_{j+1/2,k}}{2} \tag{13.34}$$

$$B^y_{j,k} = \frac{b^y_{j,k-1/2} + b^y_{j,k+1/2}}{2} \tag{13.35}$$

13.7.1. Véges térfogat interpretáció

Ha a 13.30 és 13.31 egyenleteket átlagoljuk a cella közepére a 13.34 illetve 13.35 egyenleteknek megfelelően, akkor a cella közepén definiált mágneses térre

$$B_{j,k}^{x,n+1} = B_{j,k}^{x,n} - \Delta t \frac{\bar{f}_{j,k+1/2}^{y} - \bar{f}_{j,k-1/2}^{y}}{\Delta y} \tag{13.36}$$

$$B_{j,k}^{y,n+1} = B_{j,k}^{y,n} - \Delta t \frac{\bar{f}_{j+1/2,k}^{x} - \bar{f}_{j-1/2,k}^{x}}{\Delta x} \tag{13.37}$$

ahol a fluxusok

$$\bar{f}_{j+1/2,k}^{x} = + \frac{E_{j+1/2,k+1/2}^{z} + E_{j+1/2,k-1/2}^{z}}{2} \tag{13.38}$$

$$\bar{f}_{j,k+1/2}^{y} = - \frac{E_{j+1/2,k+1/2}^{z} + E_{j-1/2,k+1/2}^{z}}{2} \tag{13.39}$$

Az indukciós egyenlet fenti diszkretizációja a mágneses tér

$$\begin{aligned} (\mathrm{div}\mathbf{B})_{j+1/2,k+1/2} &= \frac{B_{j+1,k}^{x} + B_{j+1,k+1}^{x} - B_{j,k}^{x} - B_{j,k+1}^{x}}{2\Delta x} \\ &+ \frac{B_{j,k+1}^{y} + B_{j+1,k+1}^{y} - B_{j,k}^{y} - B_{j+1,k}^{y}}{2\Delta y} \end{aligned} \tag{13.40}$$

szerint definiált divergenciáját megőrzni.

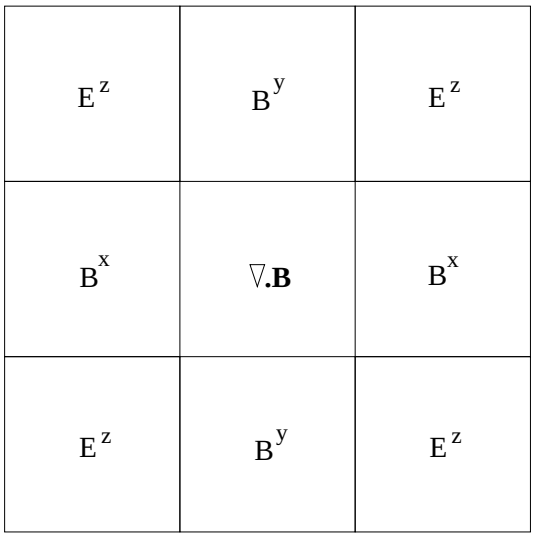
A CT módszer véges térfogat formájának előnye, hogy minden változó a cella közepén van diszkretizálva, ami leegyszerűsíti a határfeltételek megadását és a program adatstruktúráját.

13.8. Centrális differencia indukciós egyenletre

A CT diszkretizáció sémája egyszerű centrális differencián alapul, bár az egyes változók a rácson eltolva (staggered) helyezkednek el. A CT módszer véges térfogat formájából kiderült, hogy ez nem feltétlenül szükséges. Lehet azonban a dolgot még tovább egyszerűsíteni. Ha az indukció egyenletet egyszerű centrális differenciával diszkretizáljuk

$$B_{j,k}^{x,n+1} = B_{j,k}^{x,n} - \Delta t \frac{E_{j,k+1}^{z} - E_{j,k-1}^{z}}{2\Delta y} \tag{13.41}$$

$$B_{j,k}^{y,n+1} = B_{j,k}^{y,n} + \Delta t \frac{E_{j+1,k}^{z} - E_{j-1,k}^{z}}{2\Delta x} \tag{13.42}$$



13.2. ábra.
Divergenciamentes centrális differencia diszkretizáció két dimenzióban

akkor

$$(\mathrm{div}\mathbf{B})_{j,k} = \frac{B^x_{j+1,k} - B^x_{j-1,k}}{2\Delta x} + \frac{B^y_{j,k+1} - B^y_{j,k-1}}{2\Delta y} \tag{13.43}$$

egzaktul megmarad. Ezt a diszkretizációt CD módszernek nevezzük. Ez a 13.1 és 13.2 ábrák összehasonlításából is látszik, hiszen az indukciós egyenlet diszkretizációja egy 2-es faktorral való nagyítástól eltekintve ugyanaz. A nagyításnak köszönhetően azonban a CD módszernél minden változó a cellák közepén helyezkedik el.

Célszerű E^z-t időben átlagolni az n-dik és $n+1$-dik lépés között

$$E^z = \frac{E^{z,n} + E^{z,*}}{2} \tag{13.44}$$

ahol $E^z = -(\mathbf{v}\times\mathbf{B})_z$. Ez az egyik legegyszerűbb diszkretizációja az indukciós egyenletnek, megőrzi a divergenciamentességet, és a tesztek tanúsága szerint a bonyolultabb CT módszerhez hasonlóan, vagy éppen jobban működik.

13.9. Általánosítás nem szabályos rácsokra

A 8-hullám, a diffúziós kontroll és a projekciós eljárás különösebb nehézség nélkül általánosítható tetszőleges rácsokra. Komplikációk csak a CT és CD módszereknél lépnek fel, hiszen itt a diszkretizációnak kell a diffúziómentességet biztosítania.

13.9.1. Általános koordináták

A CT és a CD módszer is kiterjeszthető általánosított koordinátás rácsokra. Ehhez az elektromos és mágneses teret megfelelően transzformálni kell. A megfelelő transzformációk az általános relativitáselméletből vagy a differenciálgeometriából kaphatók meg. Ha az általánosított koordinátákat ξ, η, ζ jelöli, a koordinátatranszformációs Jacobi mátrix

$$J = \begin{pmatrix} \xi_x & \xi_y & \xi_z \\ \eta_x & \eta_y & \eta_z \\ \zeta_x & \zeta_y & \zeta_z \end{pmatrix}, \quad J^{-1,T} = \begin{pmatrix} x_\xi & y_\xi & z_\xi \\ x_\eta & y_\eta & z_\eta \\ x_\zeta & y_\zeta & z_\zeta \end{pmatrix} \tag{13.45}$$

lesz. Ennek segítségével felírható a $\mathcal{B}$ mágneses és $\mathcal{E}$ elektromos tér az általánosított koordinátákban

$$(\mathcal{B}^\xi, \mathcal{B}^\eta, \mathcal{B}^\zeta)^T = \frac{1}{\det J} J \cdot (B^x, B^y, B^z)^T \tag{13.46}$$

$$(\mathcal{E}^\xi, \mathcal{E}^\eta, \mathcal{E}^\zeta)^T = J^{-1,T} \cdot (E^x, E^y, E^z)^T \tag{13.47}$$

Ezekkel a változókkal az indukciós egyenlet éppen olyan egyszerű lesz mint a szabályos Descartes rácson, például a ξ komponensre

$$\frac{\partial \mathcal{B}^\xi}{\partial t} = -\frac{\partial \mathcal{E}^\zeta}{\partial \eta} + \frac{\partial \mathcal{E}^\eta}{\partial \zeta} \tag{13.48}$$

Ezt az egyenletet már akár CT akár CD módszerekkel diszkretizálhatjuk, és így az általános koordinátakban definiált

$$\mathrm{div}\mathbf{B} = \partial_\xi \mathcal{B}^\xi + \partial_\eta \mathcal{B}^\eta + \partial_\zeta \mathcal{B}^\zeta \tag{13.49}$$

meg fog maradni.

13.9.2. Görbevonalú rácsok

Görbevonalú rács esetén is használható az általános koordinátákra megadott módszer. A különbség annyi, hogy a Jacobi mátrix elemeit numerikusan kell meghatározni. Például a J^{-1} mátrix elemei

$$(x_\xi)_{j,k,l} = \frac{x_{j+1,k,l} - x_{j-1,k,l}}{2\Delta\xi}$$

$$(x_\eta)_{j,k,l} = \frac{x_{j,k+1,l} - x_{j,k-1,l}}{2\Delta\eta} \tag{13.50}$$

módon diszkretizálhatók. Miután J adott, az x, y, z komponensekkel adott $\mathbf{E}$ elektromos teret transzformáljuk $\mathcal{E}$-be, kiszámoljuk $\mathcal{B}$ változását a CT vagy CD módszerrel diszkretizált indukciós egyenletből, majd a *változást* a $\Delta\mathbf{B} = \det J J^{-1} \cdot \Delta\mathcal{B}$ segítségével visszatranszformáljuk és hozzáadjuk $\mathbf{B}^n$-hez. Mint látható, ezekhez a lépésekhez a J mátrix elemeire nincs is szükség.

13.9.3. Adaptív rácsok

Adaptív rácsokra csak az eredeti CT módszert lehet általánosítani, a véges térfogat forma illetve CD módszer AMR rácsokon nem használtható. Ennek az az oka, hogy a finom és durva rácsok találkozásánál csak akkor lehet a divergenciamentességet biztosítani, ha a mágneses tér normális komponense azon a felületen adott, ahol a két rács találkozik, azaz a cellahatáron. Az elektromos teret 2 dimenzióban a cella sarkokon, 3 dimenzióban az éleken kell definiálni. A finom és durva rácsok átfedő éleinél biztosítani kell, hogy a finom rács két élén definiált elektromos terek átlaga megegyezzen a durva rács egy élén lévő térrel. Ehhez ugyanolyan korrekciós lépés szükséges mint a konzervatív fluxusokat biztosító korrekció, aholis az egybeeső cellahatárokon a finom rácson definiált 4 fluxus átlagára kell javítáni a durva rácson vett fluxust.

Az AMR durvító és finomító lépéseinél is tekintettel kell lenni arra, hogy a mágneses tér divergenciamentes maradjon. A durvítás triviálisan megoldható. Amikor a 8 finom cellából egy durva cellát képzünk, a mágneses tér normál komponensét a durva rácson a finom rácson vett normál komponensek átlagával kell egyenlővé tenni. Mivel normál komponensnek a 8 finom cellára vett felületi integrálja (összege) nulla volt, ez igaz lesz a durva rácsra is. A divergenciamentes finomítás már kevésbé triviális probléma, de ez is megoldható.

13.9.4. Strukturálatlan rácsok

Struktúrálatlan rácsokra is sikerült a CT módszert általánosítani. A mágneses tér normál komponensei továbbra is a cella éleken (2 dimenzióban) illetve lapokon (3 dimenzióban) adottak. A nehézséget az okozza, hogy az élek nem párhuzamosak a koordinátatengelyekkel (illetve az általánosított koordinátákkal), így nem nyilvánvaló, hogy a normál komponensekből hogyan lehet megkapni az x, y, z komponenseket. A Hans De Sterck által kifejlesztett MUCT módszer úgynevezett felület elemek segítségével határozza meg a mágneses teret a lapokon vett normálkomponens értékek alapján. Ez a véges elem módszerből átvett technika olyan vektor bázis függvényeket használ, melyeknek a normál komponense 1 az egyik élen/lapon, és 0 a többin, továbbá divergenciájuk nulla. A vektor bázisfüggvények lineáris kombinálásával előállítható az a mágneses tér, aminek nulla a divergenciája, és a normál komponense a cella élein/lapjain éppen annyi, amennyit a CT algoritmus alapján kaptunk.

13.10. Összehasonlítás

A mágneses tér divergenciáját a fejezetben tárgyalt számos módszer bármelyikével viszonylag jól lehet kontrollálni. A legköltségesebb a projekciós módszer, mivel meg kell oldani egy Poisson egyenletet minden lépésben. Ha a megoldást csak közelítően végezzük el (pl. egy iterációs módszerrel), akkor div$\mathbf{B}$ nem lesz 0, de lecsökken. Ha csak egyetlen iterációt végzünk, akkor lényegében a diffúziós kontrol módszert kapjuk meg. A 8-hullám és a hiperbolikus Lagrange multiplikátoros módszerek a numerikusan gerjesztett monopólusokat az áramlás, illetve egy rögzített hullámsebességgel próbálják kivinni a szimulációs tartományból. Végül a CT és CD módszerek a diszkretizáció ügyes megválasztásával garantálják, hogy div$\mathbf{B}$ numerikusan megmaradjon. Ugyanakkor az indukció egyenlet speciális diszkretizációja nem feltétlenül szerencsés egyéb szempontokból. Például nem nyilvánvaló, hogyan lehet a megoldás oszcillációmentességét garantálni. Csak numerikus tesztekkel dönthető el, hogy melyik módszer működik a legjobban.

Számos numerikus teszt alapján úgy tűnik, hogy a projekciós módszer az egyik legpontosabb, de ugyanakkor a legköltségesebb is. Meglepő módon a rendkívül egyszerű CD módszer is hasonlóan pontos, de sajnos ez nem általánosítható adaptív rácsokra. A CT módszerek közül azok működnek a legjobban, amelyekben az elektromos teret megfelelő upwind módszerrel diszkretizáljuk. Ha egyszerű interpolációkat alkalmazunk, a megoldásban oszcillációk lépnek fel. Végül a 8-hullám és a diffúziós módszerek ugyan kevésbé pontosak, de rendkívül egyszerűek. A 8-hullám módszer hátránya hogy nem konzervatív, ami ritkán helytelen gyenge megoldást eredményezhet. Ugyanakkor bizonyos alkalmazásoknál, ahol a termikus energia nagyon kicsi a mágneses energiához képest, a többi módszernél jobban biztosítja a nyomás pozitívitását.

Irodalomjegyzék

[1] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran, 2nd edition* (Cambridge University Press, 1992), p. 68.

[2] M. G. Crandall and A. Majda, The method of fractional steps for conservation laws, Math. Comp., **34**, 285-314 (1980)