DOKTORANDSKÉ DNY 2008

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

7. a 21. listopadu 2008

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2008
sborník workshopu doktorandů FJFI oboru Matematické inženýrství

P. Ambrož, Z. Masáková (editoři)
Kontakt petr.ambroz@fjfi.cvut.cz / 224 358 544

Vydalo České vysoké učení technické v Praze
Zpracovala Fakulta jaderná a fyzikálně inženýrská
Vytisklo Nakladatelství ČVUT-výroba, Zikova 4, Praha 6
Počet stran 224, Vydání 1.

ISBN 978-80-01-04195-6

Seznam příspěvků

Předmluva

Smyslem doktorského studia je vychovat nastupující vědeckou generaci. Vědeckou přípravu na samotný výzkum však musí nezbytně doprovázet i příprava na prezentování výsledků srozumitelným způsobem před nejširším odborným publikem. Doktorandské dny, které se na katedře matematiky FJFI stávají již tradicí, jsou k tomu nejlepší příležitostí. Letos se konaly již po třetí, a to ve dnech 7. a 21. listopadu 2008.

Témata pokrytá přednáškami na workshopu Doktorandské dny sahaly od ryze teoretických problémů matematické fyziky, přes matematické modely přírodních procesů, až po zpracování obrazu, či databázové systémy. Hlavními přednášejícími byli studenti v prezenční formě doktorského studijního programu Aplikace přírodních věd oboru Matematické inženýrství. Jejich příspěvky předkládáme v tomto sborníku.

O pozitivním ohlasu Doktorandských dnů z minulých let svědčí i to, že k prezentaci na workshopu se hlásí i doktorandi z jiných kateder, a na přednášky přicházejí hosté z řad odborné veřejnosti. Přejme si, aby tato každoroční akce pro doktorandy získávala stále více příznivců.

Editoři

Paralelní algoritmy pro numerické řešení hydrodynamiky laserového plazmatu

Ľuboš Bednárik

1. ročník PGS, email: Lbs@centrum.sk
Katedra matematiky, Fakulta jadrová a fyzikálne inžinierska, ČVUT
školiteľ: Richard Liska, Katedra fyzikální elektroniky, Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. For solution of laser plasma hydrodynamic we introduce model of Lagrangian equations, which includes heat conductivity and laser absorption. We show us the discretization of hydrodynamical equations as well as heat conductivity equation and describe one step of the difference schema. Further we introduce the paralelization and by obtained results we determine its efficiency.

Abstrakt. Pre riešenie hydrodynamiky laserovej plazmy sa v úvode zoznámime s modelom Lagrangeovských rovníc, ktorý v sebe zahŕňa aj tepelnú vodivosť a laserovú absorpciu. Ukážeme si diskretizáciu jak hydrodynamických rovníc tak aj rovnice vedenia tepla a popíšeme jeden cyklus diferenčnej schémy. Ďalej si predstavíme prostriedky pre paralelizáciu a získanými výsledkami určíme jej efektivitu.

1 Formulácia úlohy

Laserová plazma, ktorá vzniká pri interakcii laserového žiarenia s hmotou, je typicky modelovaná ako stlačitelná kvapalina prostredníctvom Eulerových rovníc s tepelnou vodivosťou a laserovou absorpciou. Simuláciou vznikajú oblasti, ktoré sa vyznačujú vysokou expanziou resp. kompresiou. Popis systému v Lagrangeovských súradniciach je preto vhodnejší než klasický Eulerovský popis, ktorý nie je vhodný pre problémy, kde nastávajú veľké zmeny vo výpočtovej doméne (podrobný popis transformácie môžeme nájsť v [6, 7]). Budeme sa teda venovať problému, ktorý v Lagrangeovských súradniciach (S, t) má tvar

$$\frac{d\eta}{dt} = v_S \tag{1}$$

$$\frac{dv}{dt} = -p_S \tag{2}$$

$$\frac{d\varepsilon}{dt} = -pv_S - W_S - L_S \tag{3}$$

kde $\eta = 1/\rho$, ρ je hustota, v rýchlosť, p tlak, ε vnútorná energia, W je tepelný tok a L je hustota toku energie (intenzita) laserového žiarenia. Jednotlivé rovnice vyjadrujú postupne zákon zachovania hmotnosti (1), zákon zachovania hybnosti (2) a zákon zachovania energie (3). Systém doplňujeme ďalej ešte o stavové rovnice $p = p(\varepsilon, \rho)$, $T = T(\varepsilon, \rho)$,

ktoré pre ideálny plyn uvažujeme v tvare:

$$p = \varepsilon\rho(\gamma - 1) \tag{4}$$

$$T = \frac{A}{Z+1}\frac{p}{c_p\rho},\ c_p = \frac{k_B}{m_u} \tag{5}$$

kde $\gamma = 5/3$ je plynová konštanta, Z stupeň ionizácie, A atómové číslo, k_B Boltzmanova konštanta a $m_u = 1,6605.10^{-24}g$ atómová hmotnostná jednotka.

Systém rovníc (1), (2), (3) riešime v dvoch krokoch. V prvom kroku riešime samostatne systém hydrodynamických rovníc

$$\frac{d\eta}{dt} = v_S \tag{6}$$

$$\frac{dv}{dt} = -p_S \tag{7}$$

$$\frac{d\varepsilon}{dt} = -pv_S \tag{8}$$

V druhom kroku riešime samostatne rovnicu vedenia tepla so zahrnutým členom pre laserove žiarenie

$$\frac{d\varepsilon}{dt} = -W_S - L_S \tag{9}$$

2 Diskretizácia

Systém riešime numericky diskretizáciou v čase aj v priestore, pričom parciálne derivácie nahrádzame diferenciami. Nech teda daná oblasť, v našom prípade interval $\langle a, b\rangle$, je lubovoľne rozdelena bodmi x_1 až x_{m+1} na m subintervalov, kde $x_1 = a$ a $x_{m+1} = b$. Tieto subintervaly budeme nazývať *primárna sieťka*. *Primárne body* definujeme ako stredy týchto subintervalov a značíme postupne $x_{3/2}$, $x_{5/2}$ až $x_{m-1/2}$, $x_{m+1/2}$. Vrcholy primárnej sieťky tvoria tzv. *duálne body*, ktoré označujeme indexom s celočíselným argumentom. *Duálna sieťka* bude obsahovať duálne body vnútri svojich buniek, a teda jej vrcholmi sú primárne body. Označme ďalej Δt časový krok a $t^n = n\Delta t$, $n = 0, 1, 2, ...$

2.1 Diskretizácia hydrodynamických rovníc

Najskôr zdiskretizujeme systém hydrodynamických rovníc (6),(7),(8). Časové derivácie nahradíme jednoduchými diferenciami:

$$\frac{df}{dt} \to \frac{f_i^{n+1} - f_i^n}{\Delta t} \tag{10}$$

kde $f_i^n = f(x_i, t^n)$. V rovniciach pre zákony zachovania hybnosti a energie navyše použijeme člen pre umelú viskozitu q, definovanú vzťahom

$$q_{i+1/2}^n = \begin{cases} 0 & \text{pre } v_{i+1}^n - v_i^n \geq 0 \\ -\frac{3}{2}\rho_{i+1/2}^n (v_{i+1}^n - v_i^n)\sqrt{(\gamma-1)\gamma\varepsilon_{i+1/2}^n} & \text{pre } v_{i+1}^n - v_i^n < 0 \end{cases} \tag{11}$$

Zákon zachovania hybnosti diskretizujeme podľa schémy

$$\frac{v_i^{n+1} - v_i^n}{\Delta t} = -\frac{p_{i+1/2}^n + q_{i+1/2}^n - p_{i-1/2}^n - q_{i-1/2}^n}{m_i}, \text{ pre } i = 2, .., m-1 \tag{12}$$

Rýchlosti v_1 a v_J sú dané okrajovými podmienkami. Ak poznáme rýchlosti vo všetkých bodoch sieťky, potom podľa nasledujúceho vzťahu stanovíme pohyb sieťky:

$$\frac{x_i^{n+1} - x_i^n}{\Delta t} = \frac{v_i^{n+1} + v_i^n}{2} \tag{13}$$

Zákon zachovania energie diskretizujeme podľa schémy

$$\frac{\varepsilon_{i+1/2}^{n+1} - \varepsilon_{i+1/2}^n}{\Delta t} = -(p_{i+1/2}^n + q_{i+1/2}^n)\frac{\frac{1}{2}(v_{i+1}^{n+1} + v_{i+1}^n) - \frac{1}{2}(v_i^{n+1} + v_i^n)}{m_{i+1/2}}, \text{ pre } i = 1, .., m \tag{14}$$

Hustota je daná pohybom sieťky, pretože hmotnosť zostáva v každom čase pre každú bunku konštantná:

$$\rho_{i+1/2}^{n+1} = \frac{m_{i+1/2}}{x_{i+1}^{n+1} - x_i^{n+1}} \tag{15}$$

2.2 Diskretizácia rovnice vedenia tepla

Rovnicu vedenia tepla (9) zdiskretizujeme po prechode od systému (S, t) k (x, t), a tak podľa [5] má táto rovnica po tejto transformácii tvar:

$$\rho\frac{d\varepsilon}{dt} = -W_x - L_x \tag{16}$$

kde $W = -\kappa T_x$ a κ je koeficient tepelnej vodivosti. Z rovníc (4) a (5) dostávame

$$\varepsilon(T, \rho) = \frac{T(Z+1)c_p}{A(\gamma - 1)} \tag{17}$$

kde vidíme, že vnútorna energia je pre ideálny plyn funkciou iba teploty $\varepsilon = \varepsilon(T)$. Keďže pre totálnu časovú deriváciu vnútornej energie platí

$$\frac{d\varepsilon}{dt}(T, \rho) = \frac{\partial\varepsilon}{\partial T}\frac{\partial T}{\partial t} + \frac{\partial\varepsilon}{\partial\rho}\frac{\partial\rho}{\partial t} \tag{18}$$

a z (17) vyplýva $\partial\varepsilon/\partial\rho = 0$, potom môžeme vzťah (16) prepísať do tvaru

$$T_t = \frac{1}{\rho\varepsilon_T}(\kappa T_x)_x - \frac{1}{\rho\varepsilon_T}L_x \tag{19}$$

Parciálna derivácia ε_T je pre ideálny plyn nezávislá na teplote a zo vzťahu (17) ju dokážeme vyjadriť:

$$\varepsilon_T = \frac{(Z+1)c_p}{A(\gamma - 1)} \tag{20}$$

Označme symbolom V_i objem bunky, ktorá obsahuje vnútri duálny bod x_i, tj.

$$V_i = x_{i+1/2} - x_{i-1/2} \tag{21}$$

a symbolom $V_{i+1/2}$ objem bunky obsahujúcej primárny bod $x_{i+1/2}$, tj.

$$V_{i+1/2} = x_{i+1} - x_i \tag{22}$$

Stavové veličiny T, p,ρ, ε, κ sú dané na primárnej sieťke, tj. v bodoch $x_{i+1/2}$, veličiny v, L na duálnej sieťke v bodoch x_i. Potom časovú deriváciu teploty nahradíme

$$[T_t]_{i+1/2} = \frac{T^{n+1}_{i+1/2} - T^n_{i+1/2}}{\Delta_n t} \tag{23}$$

kde symbol $\Delta_n t = t_{n+1} - t_n$. Indexy pri hranatých zátvorkách označujú body sieťky. Deriváciu podľa x nahradíme podobným spôsobom

$$[T_x]_{i+1/2} = \frac{T^{n+1}_{i+1} - T^{n+1}_i}{V_{i+1/2}} \tag{24}$$

Analogicky nahradíme aj druhú deriváciu, a tak následným dosadením do (19) dostávame výslednú schému:

$$\rho^n_{i+1/2}\varepsilon_T \frac{T^{n+1}_{i+1/2} - T^n_{i+1/2}}{\Delta_n t} = \frac{1}{V_{i+1/2}} \left[\frac{\kappa^n_{i+1}}{V_{i+1}}(T^{n+1}_{i+3/2} - T^{n+1}_{i+1/2}) - \frac{\kappa^n_i}{V_i}(T^{n+1}_{i+1/2} - T^{n+1}_{i-1/2}) - (L^n_{i+1} - L^n_i) \right] \tag{25}$$

Táto implicitná schéma predstavuje systém $m-2$ rovníc pre m neznámych$T^{n+1}_{i+1/2}$ pre $i = 1, .., m$ a má tvar

$$a_i T^{n+1}_{i-1/2} + b_i T^{n+1}_{i+1/2} + c_i T^{n+1}_{i+3/2} = R_i, \ i = 2, .., m-1$$

kde koeficienty a_i,b_i, c_i a R_i vyplývajú z (25)

$$\begin{aligned} a_i &= -\frac{\kappa_i}{V_i} \frac{\Delta_n t}{V_{i+1/2}\rho_{i+1/2}\varepsilon_T} \\ b_i &= 1 + \left(\frac{\kappa_{i+1}}{V_{i+1}} + \frac{\kappa_i}{V_i} \right) \frac{\Delta_n t}{V_{i+1/2}\rho_{i+1/2}\varepsilon_T} \\ c_i &= -\frac{\kappa_{i+1}}{V_{i+1}} \frac{\Delta_n t}{V_{i+1/2}\rho_{i+1/2}\varepsilon_T} \\ R_i &= T^n_{i+1/2} - (L^n_{i+1} - L^n_i) \frac{\Delta_n t}{V_{i+1/2}\rho_{i+1/2}\varepsilon_T} \end{aligned} \tag{26}$$

Zostávajúce dve rovnice

$$b_1 T^{n+1}_{3/2} + c_1 T^{n+1}_{5/2} = R_1 \tag{27}$$

$$a_N T^{n+1}_{m-1/2} + b_N T^{n+1}_{m+1/2} = R_m \tag{28}$$

vyjadrujú okravojé podmienky. Systém sa dá zapísať maticovo ako

$$\begin{pmatrix} b_1 & c_1 & 0 & ... & 0 & 0 & 0 \\ a_2 & b_2 & c_2 & ... & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & ... & a_{m-1} & b_{m-1} & c_{m-1} \\ 0 & 0 & 0 & ... & 0 & a_m & b_m \end{pmatrix} \begin{pmatrix} T^{n+1}_{3/2} \\ T^{n+1}_{5/2} \\ \vdots \\ T^{n+1}_{m-1/2} \\ T^{n+1}_{m+1/2} \end{pmatrix} = \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_{m-1} \\ R_m \end{pmatrix} \tag{29}$$

kde ako vidno má matica systému tridiagonálny tvar, kde nenulové prvky sa nachádzajú na diagonále a nad a pod diagonálou.

2.3 Diferenčná schéma - cyklus

Predpokladajme ďalej, že na začiatku prvého kroku diferenčnej schémy máme dané stavové veličiny ρ_0, v_0, ε_0 a $p_0 = p(\rho_0, \varepsilon_0)$,$T_0 = T(\rho_0, \varepsilon_0)$. Vyriešením systému hydrodynamických rovníc, kde uvažujeme okrajovú podmienku pre rýchlosť alebo tlak, získame nové hodnoty veličín, ktoré označíme ρ_1, v_1, ε_1. Zo stavovej rovnice dopočítame

$$T_1 = T(\rho_1, \varepsilon_1)$$

ako počiatočné riešenie pre rovnicu vedenia tepla. Pre ideálny plyn je dokonca teplota funkciou iba vnútornej energie. Nasleduje vyriešenie rovnice vedenia tepla, čím získame novú teplotu T_2. Časový krok musí byť pre obidva subkroky rovnaký. Ak sa líšia, zvolíme obidva podľa menšieho z nich. Finálne už iba stačí, keď zaktualizujeme vnútornú energiu

$$\varepsilon_2 = \varepsilon(T_2, \rho_1)$$

a ako nové počíatočné podmienky do ďalšieho kroku vezmeme hodnoty ρ_1, v_1, ε_2.

2.4 Absorpcia laseru

V našej rovnici vedenia tepla (9) sa vyskytuje člen L, ktorý v sebe zahrňuje energiu predávanú systému v dôsledku absorpcie laserového žiarenia. Hodnotu tohto členu spočítame zo vzťahu

$$L = \begin{cases} 0 & \text{pre } \rho \geq \rho_c \\ I^L & \text{pre } \rho < \rho_c \end{cases}$$

kde $I^L = I^L(t)$ je intenzita laserového žiarenia popísaná ďalej a ρ_c je tzv. kritická hustota, pre ktorú platí

$$\rho_c = 1,86 \times 10^{-3} \frac{A}{Z} \frac{1}{\lambda_\mu^2}$$

kde λ_μ je vlnová dĺžka laseru v μm. Uvažujeme pritom dopad laserového žiarenia s profilom Gaussovského pulsu, tzn. pre intenzitu použijeme vzťah

$$I^L(t) = I^L_{max} e^{\frac{-(t-t_0)^2 4ln2}{\tau^2}}$$

kde I^L_{max} je maximálna intenzita žiarenia, t_0 je posunutie maxima vzhľadom k času $t = 0$ a τ je šírka pulsu v polovici maximálnej výšky (FWHM).

3 Paralelizácia

Pre urýchlenie výpočtu sme sa rozhodli náš program sparalelizovať, a to prostredníctvom OpenMP. V tomto prípade sa bežiaci proces rozdelí na niekoľko vlákien a na nich prebieha paralelne výpočet. V jazyku C je OpenMP implementované prostredníctvom direktív prekladača riadiacich samotnú paralelizáciu a menšou skupinkou pomocných funkcií, ktoré umožnujú kontrolovať a riadiť jednotlivé vlákna. Direktívy majú tvar `#pragma omp`

a k tým dôležitejším patria `#pragma omp parallel for` a `#pragma omp sections`. Vo fortrane je to takmer take isté, rozdiel je v tom, že direktívy a pomocné funkcie majú odlišnú syntax. Napríklad spomenuté dve direktívy majú vo fortrane tvar `!$omp do` a `!$omp sections`.

Prvá z uvedených direktív sa používa pre paralelizáciu cyklov for. Direktíva s príslušnými parametrami vraví prekladaču, že proces sa má rozdeliť na viac vlákien, nasledujúci cyklus rozdelíť na zodpovedajúci počet častí a potom sa opäť spojíť. Dôležité je, ako sa prideľuje práca jednotlivým vláknam. Máme niekoľko možnosti:

- Staticky, kde sa cyklus rozdelí na niekoľko blokov konštantnej veľkosti a tieto bloky sa hneď na začiatku pridelia jednotlivým vláknam.
- Dynamicky, kde sa cyklus rozdelí opäť na niekoľko blokov konštantnej veľkosti, ale tieto sa prideľujú vláknam podľa potreby. Ktoré vlákno dokončí svoj blok, dostane ďalší.
- Riadene, kde sa mení aj veľkosť prideľovaných blokov.

Je dôležite tiež určiť, ktoré premenné majú byť zdielané (napríklad dáta, na ktorých pracujeme), a ktoré súkromné (napríklad iteračná premenná, ktorá má pre každé vlakno inú hodnotu).

Niekedy je nutné vykonať danú operáciu napríklad len jedným vláknom, alebo viacerými ale odlišne. K tomuto účelu slúži druhá spomínaná direktíva `#pragma omp sections`, kde pre každú sekciu kódu sa dá nastaviť, ako sa má spracovávať. Zo základných spôsobov môžme uviesť:

- `Jednotlivá sekcia` určuje časť kódu, ktorý sa vykoná len jedným vláknom (napr. vstupné a výstupné operácie).
- `Kritická sekcia` určuje časť kódu, ktorá sa smie vykonať maximálne jedným vláknom v tom istom čase (prístup k hardwaru).
- `Zarážka` nastavuje miesto, kam musia všetky vlákna dospieť a počkať na seba.
- `Zoradený kód`, ktorý je vykonávaný v rovnakom poradí ako pri sekvenčnom algoritme.

V niektorých prípadoch potrebujeme poznať aktuálny počet vlákien pripadne číslo vlákna a k tomu nám OpenMP poskytuje vstavané funkcie `omp_get_num_threads` (vráti počet vlákien) a `omp_get_thread_num` (vráti číslo vlákna).

Napísaný program kompilujeme vybranými prekladačmi podporujúci štandard OpenMP s použitím príslušných prepínačov. Pre intelovské prekladače je to prepínač `-openmp`, pre štandardné GNU prekladače, voľne dostupné v každej distribúcii linuxu, prepínač `-fopenmp`.

4 Výsledky

Väčšina simulácii prebiehala na počítači pozostávajúceho zo 4 dvojjadrových procesorov Intel Xeon s frekvenciou 2667 MHz a 24 GB RAM. Mohli sme tak spustiť výpočet až na

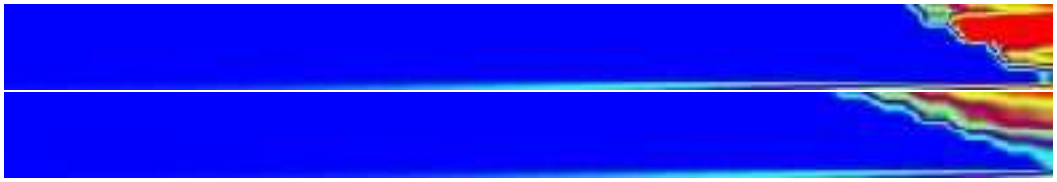
8 procesoch. Účinnosť paralelizácie vypočítame podľa nasledujúceho vzťahu

$$\eta = \frac{t_1}{t_n.n}, \tag{30}$$

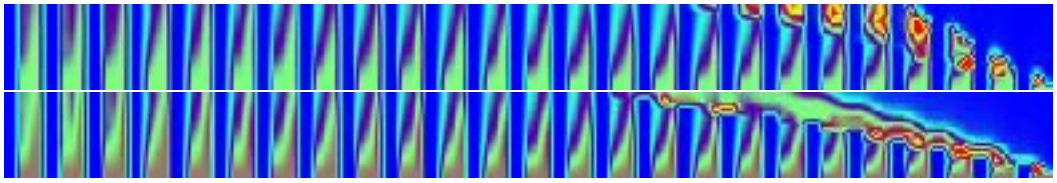
kde t_1 je doba výpočtu jedným procesom a t_n doba výpočtu n procesmi. Pomer $\frac{t_1}{t_n}$ pritom určuje dosiahnuté zrýchlenie.

V simulácii uvažujeme penový terčík, modelovaný sériou stien a medzier s danými vlastnosťami, na ktorý sprava dopadá laserový zväzok. Na obrázkoch y-ová os zodpovedá určitej časovej hladine a na x-ovej osi sú vynášané hodnoty konkretnej fyzikálnej veličiny. Uvažeme pokročilejší model absorpcie laseru, kde laser je absorbovaný nielen v tzv. kritickom mieste, tj. mieste s kritickou hustotou, ale aj v jeho okolí. V každej dvojíci obrázkov vrchný zobrazuje simuláciu bez vedenia tepla, dolný zobrazuje simuláciu s tepelnou vodivosťou, pričom môžeme vidieť postupne zobrazenie laseru (obrázok 1), zobrazenie hustoty (obrázok 2) a zobrazenie teploty (obrázok 3).

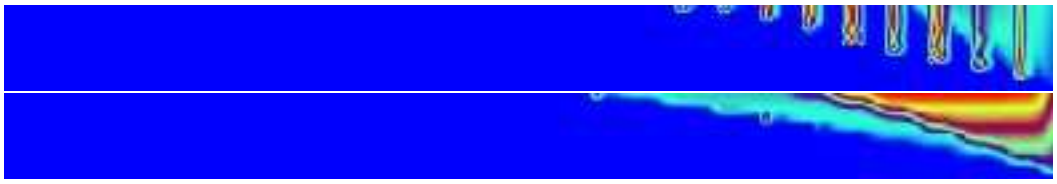
Porovnaním prvých dvoch dvojíc obrázkov vidíme, že laser nie je absorbovaný len v mieste dopadu, ale aj hlbšie v materiáli. Na tretej dvojíci obrázkov (obrázok 3) potom vidíme vplyv rovnice vedenia tepla.



Obrázok 1: Simulácia absorpcie laseru, zobrazenie laseru, horný obrazok bez vedenia tepla, dolný s vedením tepla.



Obrázok 2: Simulácia absorpcie laseru, zobrazenie hustoty, horný obrazok bez vedenia tepla, dolný s vedením tepla.



Obrázok 3: Simulácia absorpcie laseru, zobrazenie teploty, horný obrazok bez vedenia tepla, dolný s vedením tepla.

Dobu výpočtu ako aj zrýchlenie a efektivitu pri použití sieťky zloženej z 250 buniek [1] vidíme v tabuľke 2. V prvom stĺpci máme počet procesov (resp. vlákien), na ktorých bol výpočet spustený. V druhom a treťom stĺpci sú zaznamenané namerané doby výpočtu. Druhý stĺpec zobrazuje skutočnú dobu výpočtu, tj. odkedy sa výpočet spustil až po jeho skončenie, zatiaľ čo v treťom stĺpci sa nachádza čas procesu, ktorý je súčtom časov na jednotlivých vláknach. V posledných dvoch stĺpcoch sa nachádza zrýchlenie a dosiahnutá efektivita, ktorá je počítaná pomocou vzťahu (30). Nízka hodnota tejto efektivity je spôsobená okrem nízkeho počtu buniek aj tým, že v celkovej dobe výpočtu je zahrnutá aj doba neparalelizovaných výpočtov.

Procesy	Čas	Čas procesu	Zrýchlenie	Efektivita
1	320	319	1	1
2	235	333	1,36	68.1 %
4	209	382	1,53	38.2 %
8	193	416	1,65	20.7 %

Tabuľka 1: Výsledok paralelizácie na sieťke s 250 bunkami - celková doba výpočtu.

Ak sa zameriame iba na dobu paralelizovanej časti výpočtu, konkrétne o hydrodynamickú časť, môžeme pozorovať lepšie výsledky (viď. tabuľka 2). Výpočet opeť prebiehal na oblasti s 250 bunkami. Výsledky obidvoch tabuliek spolu s ďalšími simuláciami pre 125, 500 a 1000 buniek vidíme na obrázku 4. Môžeme pozorovať, že zvyšujúcim sa počtom buniek sa zvyšuje aj efektivita paralelizácie.

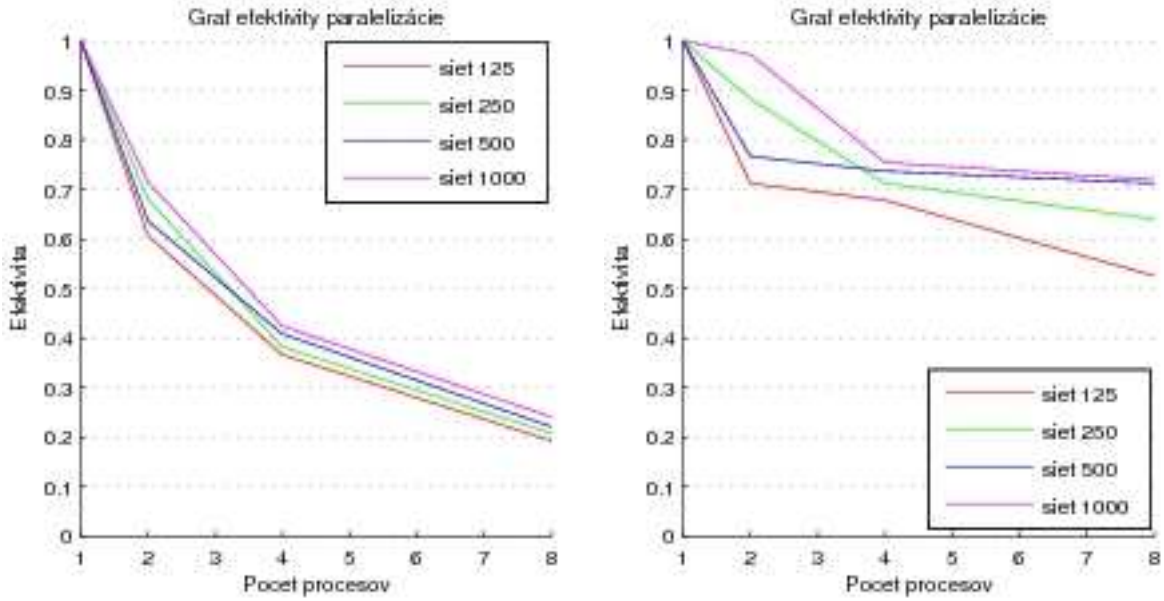
Procesy	Čas	Čas procesu	Zrýchlenie	Efektivita
1	199	199	1	100 %
2	113	211	1.76	88,3 %
4	69,8	241	2.85	71.3 %
8	38,8	258	5.12	64.0 %

Tabuľka 2: Výsledok paralelizácie na sietke s 250 bunkami - doba výpočtu paralelizovanej hydrodynamickej časti výpočtu.

5 Záver

Zoznámili sme sa s modelom Lagrangeovských rovnic pre riešenie hydrodynamiky laserovej plazmy, ktorý v sebe zahrňuje aj tepelnú vodivosť a laserovú absorpciu. Ďalej sme si ukázali diskretizáciu a predstavili prostriedky pre paralelizáciu tohto problému. Získanými výsledkami sme skúmali efektivitu paralelizácie a overili, že pre daný počet procesov (resp. vlákien) sa efektivita zvyšuje so zvyšujúcim sa počtom buniek na sieti.

[1] Nízky počet buniek je zvolený zámerne. Pre vysoké počty buniek a malý počet procesov (vlákien) vychádza efektivita veľmi vysoka. My sme však chceli ukázať, kde sa nachádza spodná hranica počtu buniek, tak aby mala paralelizácia ešte zmysel.



Obrázok 4: Graf efektivity paralelizácie postupne na sieťkach s 125, 250, 500 a 1000 bunkami. Vľavo celá simulácia, vpravo iba hydrodynamická časť.

Literatúra

[1] E.J. Caramana, D.E. Burton, M.J. Shashkov., P.P. Whalen. *The Construction of Compatible Hydrodynamics Algorithms Utilizing Conservation of Total Energy*, J. of Com. Phys. (1998), **146**: 227-262.

[2] T. Kapin, M. Kuchařík, J. Limpouch, R. Liska, P. Váchal. *Arbitrary Lagrangian Eulerian method for laser plasma simulations.* Int. J. Numer. Meth. Fluids (2008), **56**: 1337-1342.

[3] T. Kapin, M. Kuchařík, J. Limpouch, R. Liska. *Hydrodynamic simulations of laser interactions with low-density foams.* Czechoslovak Journal of Physics (2006), **56**: B493-B499.

[4] R. Liska, M. Kuchařík. *Arbitrary Lagrangian Eulerian Method for Compressible Plasma Simulations*, Proceedings of Equadiff-11, (2005), pp. 1-10.

[5] P. Havlík. *Diferenční schémata pro hydrodynamiku na nerovnoměrých a Lagrangeovských sítích.* Výzkumní úkol, České vysoké učení technicke v Praze, (2004).

[6] M. Shashkov. *Conservative Finite-Difference Methods on General Grids.* CRC Press, Boca Raton, (1996).

[7] M. Shashkov, B. Wendroff. *A Composite Scheme for Gas Dynamics in Lagrangian Coordinates*, J. of Comp. Phys. (1999), **150**: 502-517.

Transport of Colloids through Porous Media

Pavel Beneš

1st year of PGS, email: benespa1@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Jiří Mikyška, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. The goal of this contribution is to describe the transport of colloids in the porous media. This work includes equation describing the flow field, transport of colloids and deposition of colloids in the porous media. Then there is a numerical discretisation of the system of equations describing the colloid transport with known flow field by means of the upwind scheme.

Abstrakt. Hlavním cílem tohto příspěvku je popis transportu koloidů v porézním prostředí. Tato práce obsahuje rovnici popisující proudové pole, transport koloidů a jejich ukládání v porézním prostředí. Dále je v práci obsažena numerická diskretizace tohoto systému rovnic popisujícího transport koloidů při známém proudovém poli za použití upwindového schématu.

1 Introduction

This contribution is a review of colloidal transport in porous media. The Contribution contains equations describing this complicated but important system. Colloids are small particles with at least one dimension less than 100 nm. Examples of colloids are bacteria or viruses. Colloids have many usages. One of reasons for studying colloidal the transport in the porous media is that colloids are able under certain conditions to make contaminant transport in porous media faster. One case was measured for example in Los Alamos, where Pu transport was measured and Pu particles reached 1200 times farther than was predicted by classical transport model. Colloids stimulated this phenomenon [5].

2 The physical model

This section presents equations describing the colloidal transport in porous media [1].

2.1 Flow field

To describe the transport of colloidal particles in the porous media three equations are necessary. The first is describing the flow field and we will call it the flow equation. We can write it in this form:

$$S_s \frac{\partial h}{\partial t} = \nabla \cdot (K \nabla h) - Q, \tag{1}$$

where S_S is the specific storage, t is time, h is the hydraulic head, K is the hydraulic conductivity and Q is the pumping or recharge rate. After time the water flow in the

porous media comes to the steady state. Then is possible to measure the hydraulic head and use the Darcy law to obtain the Darcy velocity $\mathbf{q}$:

$$\mathbf{q} = -K\nabla h. \tag{2}$$

2.2 Colloid transport equation

This equation can be derived from the mass balance of colloids over the REV (representative element volume). There are three main mechanisms controlling the colloidal transport: hydrodynamic dispersion, advection and colloid deposition and release. This can be described by the generalized advection dispersion equation:

$$\frac{\partial C}{\partial t} = \nabla \cdot (D\nabla c) - \nabla \cdot (\mathbf{V}C) - \frac{\varrho_b}{\varepsilon}\frac{\partial S}{\partial t}, \tag{3}$$

where C is the mass concentration of colloids in aqueous phase,

$$S = \frac{\text{colloid mass captured by solid matrix}}{\text{total mass of solid matrix}},$$

D is particle hydrodynamic dispersion tensor, $\mathbf{V}$ is particle velocity tensor, ε is the porosity and ϱ_b is the bulk density of porous media. Because the colloid particle size is much smaller than the pore size, it is possible to take $\mathbf{V}$ like interstitial fluid velocity. In two dimensions it is possible to take

$$D_{ij} = \alpha_T \bar{V}\delta_{ij} + (\alpha_L - \alpha_T)\frac{\bar{V}_i\bar{V}_j}{\bar{V}} + D_d T\delta_{ij},$$

where D_d is the Stokes-Einstein diffusivity, $\bar{V}_i\bar{V}_j$ are components of the interstitial velocity, α_L is the longitudinal dispersivity, α_T is the transverse dispersivity and T is the tortuosity of porous medium. Further we will use the equivalent of (1), where the unknown is the particle number concentration n:

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n) - \nabla \cdot (\mathbf{V} \cdot n) - \frac{f}{\pi a_p^2}\frac{\partial \theta}{\partial t}, \tag{4}$$

where θ is the specific surface coverage, defined as

$$\theta = \frac{\text{total cross-section area of deposited colloids}}{\text{interstitial surface area of the porous media solid matrix}},$$

f is specific surface area

$$f = \frac{\text{interstitial surface area}}{\text{porous medium pore volume}},$$

and a_p is radius of colloidal particles.

2.3 Colloid deposition and release

Let λ be the percentage part of the solid matrix with favorable conditions for colloid deposition. This can be for example areas with iron oxides on its surface. These surfaces are typically positive charged and colloids are typically negatively charged. Deposition on the surfaces is usually irreversible. On the rest $(1-\lambda)$ of the solid matrix surface are unfavorable conditions for the colloidal deposition. Deposition takes place on both parts, but difference in rates can be huge. For particle surface coverage rate we can adopt this patch wise model:

$$\frac{\partial\theta}{\partial t} = \lambda\frac{\partial\theta_f}{\partial t} + (1-\lambda)\frac{\partial\theta_u}{\partial t}, \tag{5}$$

where θ_f is favorable surface fraction and θ_u is unfavorable surface fraction. For there rates exists partial differential equations:

$$\frac{\partial\theta_f}{\partial t} = \pi a_p^2 k_{dep,f} n B(\theta_f) - k_{det,f}\theta_f R(\theta_f), \tag{6}$$

$$\frac{\partial\theta_u}{\partial t} = \pi a_p^2 k_{dep,u} n B(\theta_u) - k_{det,u}\theta_u R(\theta_u), \tag{7}$$

where k_{dep} is the colloid deposition rate constant, k_{det} is the colloid release rate constant, $B(\theta)$ is the dynamic blocking function and $R(\theta)$ is dynamic release function. Colloid deposition rate coefficient k_{dep} can be expressed by means of single collector efficiency η:

$$k_{dep} = \frac{\eta\varepsilon V}{4} = \frac{\alpha\eta_0\varepsilon V}{4}, \tag{8}$$

where V is the fluid advection velocity, ε is porosity and η_0 is the favorable single collector removal efficiency.

2.4 Dynamic blocking and release functions $B(\theta)$, $R(\theta)$

Dynamic blocking functions characterize particle deposition [4]. When is the collector at the beginning particle free has blocking function value $B(\theta) = 1$. As deposited particles blocking the surface more and more $B(\theta)$ decreases and when at maximum attainable surface coverage $\theta = \theta_{max}$ (jamming limit) it is $B(\theta) = 0$. We will present two models of this function here.

2.4.1 Langmuirian dynamic blocking function

This blocking function is a linear approximation:

$$B(\theta) = 1 - \frac{1}{\theta_{max}}\theta.$$

This model was made for point size particles. For larger (finite size) particles linear description is not sufficient. For this reason we will show non-linear blocking function here: the RSA model.

2.4.2 RSA dynamic blocking function

For colloidal particles depositing on the oppositely charged collector surface these conditions for use of RSA model are given:

- attachment is irreversible as long as conditions do not change
- surface diffusion is negligible
- particle-particle contact is prohibited

For low and moderate surface coverage the function has this form:

$$B(\theta) = 1 - 4\theta_\infty \frac{\theta}{\theta_{max}} + \frac{6\sqrt{3}}{\pi}\left(\theta_\infty \frac{\theta}{\theta_{max}}\right)^2 + \left(\frac{40}{\sqrt{3}\pi} - \frac{176}{3\pi^2}\right)\left(\theta_\infty \frac{\theta}{\theta_{max}}\right)^3,$$

where θ_∞ is the hard sphere jamming limit. For coverage approaching θ_{max} ($\theta > 0,8\theta_{max}$)

$$B(\theta) = \frac{(1 - \frac{\theta}{\theta_{max}})^3}{2m^2\left(\frac{1}{\theta_{max}}\right)^3},$$

where m is the jamming limit slope.

2.4.3 Dynamic release function

Dynamic release function describes the probability of colloid release from the porous media surface covered by retained colloids [1]. This function should in general depend on colloid the residence time and the retained colloid concentration. Because the colloid release is not well understood we will use $R(\theta) = 1$. Then equations (6) and (7) represent first order kinetics release function.

3 Solved equation

This section shows solved equation, initial and boundary conditions. By substituting equations describing the colloid deposition and release (5), (6) and (7) to (4), we obtain the following expression:

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n) - \nabla \cdot (\mathbf{V} \cdot n) - \frac{f}{\pi a_p^2}((\lambda \pi a_p^2 k_{dep,f} B(\theta_f) + (1-\lambda)\pi a_p^2 k_{dep,u} B(\theta_u))n - ((\lambda \pi k_{det,f}\theta_f R(\theta_f) + (1-\lambda)k_{dep,u}\theta_u R(\theta_u)). \quad (9)$$

Now we assume that $K(\theta) = 1$ (first order kinetics release mechanism) and use the following notations:

$$\gamma = \frac{f}{\pi a_p^2}, \quad (10)$$

$$K_a(\theta_f, \theta_u) = \pi a_p^2[\lambda k_{dep,f} B(\theta_f) + (1-\lambda)k_{dep,u} B(\theta_u)], \quad (11)$$

$$K_r(\theta_f, \theta_u) = \lambda \pi k_{det,f}\theta_f + (1-\lambda)k_{dep,u}\theta_u. \quad (12)$$

After application of these assumptions the following equation is obtained:

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n) - \nabla \cdot (\mathbf{V} \cdot n) - \frac{K_a(\theta_f, \theta_u)}{\gamma} n + \frac{K_r(\theta_f, \theta_u)}{\gamma}. \tag{13}$$

Now we will complete this system of equations by means of equations:(1), (2), (6) and (7):

$$S_s \frac{\partial h}{\partial t} = \nabla \cdot (K\nabla h) - Q, \tag{14}$$
$$\mathbf{q} = -K\nabla h,$$
$$\mathbf{V} = \frac{\mathbf{q}}{n},$$

$$\frac{\partial \theta_f}{\partial t} = \pi a_p^2 k_{dep,f} n B(\theta_f) - k_{det,f}\theta_f, \tag{15}$$
$$\frac{\partial \theta_u}{\partial t} = \pi a_p^2 k_{dep,u} n B(\theta_u) - k_{det,u}\theta_f. \tag{16}$$

To solve this system, we will need boundary and initial conditions for each equation (13),(14),(15) and (16).
Let us have rectangular domain oriented in directions of axis x, where lower boundary is denoted Γ_1, right Γ_2, upper Γ_3 and left Γ_4 (fig. (1)).
For concentration equation (13) will have initial condition given by some function $n_0(\mathbf{x})$ and boundary conditions will describe sources of colloids, so there are some functions $n_i(\mathbf{x}, t)$, where $i \in 1, \ldots, 4$.
For equation describing the flow field (14), we will have some initial hydrodynamic head e.,g. $h(\mathbf{x}) = h_0$ for $t = 0$, Dirichlet boundary condition on Γ_2 and Γ_4 e.,g. $h(\mathbf{x}, t) = h_i(\mathbf{x})$ on $\mathbf{x} \in \Gamma_i, t > 0$ for $i = 2, 4$ and zero Neumann boundary conditions on Γ_2 and Γ_4 e.,g. $\frac{\partial h(\mathbf{x}}{\partial z} = 0$ on $\mathbf{x} \in \Gamma_{1,3}, t > 0$.
For equations (15) and (16) there are initially no deposited colloids so ($\theta_f = \theta_u = 0$) and then there are zero dispersive flux boundary conditions for $t > 0$ e.,g. $\frac{\partial \theta_j(\mathbf{x}}{\partial z} = 0$ on $\mathbf{x} \in \Gamma_{1,3}$ and $\frac{\partial \theta_j(\mathbf{x}}{\partial x} = 0$ on $\mathbf{x} \in \Gamma_{2,4}$ for $t > 0$ and $j = f, u$.

4 Numerical solution

Now we discuss how numerically solve the system above [1], [2], [3]. Let us suppose that the flow field is time independent and that the flow field is given. In this case it will not have to solve equation (14) and we will know velocity $\mathbf{V}$. If the flow field is not known it is necessary to solve equation (14) in each time step, like first one. For known flow field we have to solve three coupled equations (13),(15) and (16) with initial and boundary conditions given in the previous chapter. First thing to do in numerical solution is the numerical grid. We will use triangulation on our domain Ω. It is called the primary grid. On the primary grid we will construct the dual grid. We will connect midpoints of triangle with all its sites in each triangle from primary grid. In this way we will obtain a polygon around each node from the primary grid (on the boundary of the domain ($\partial\Omega$),

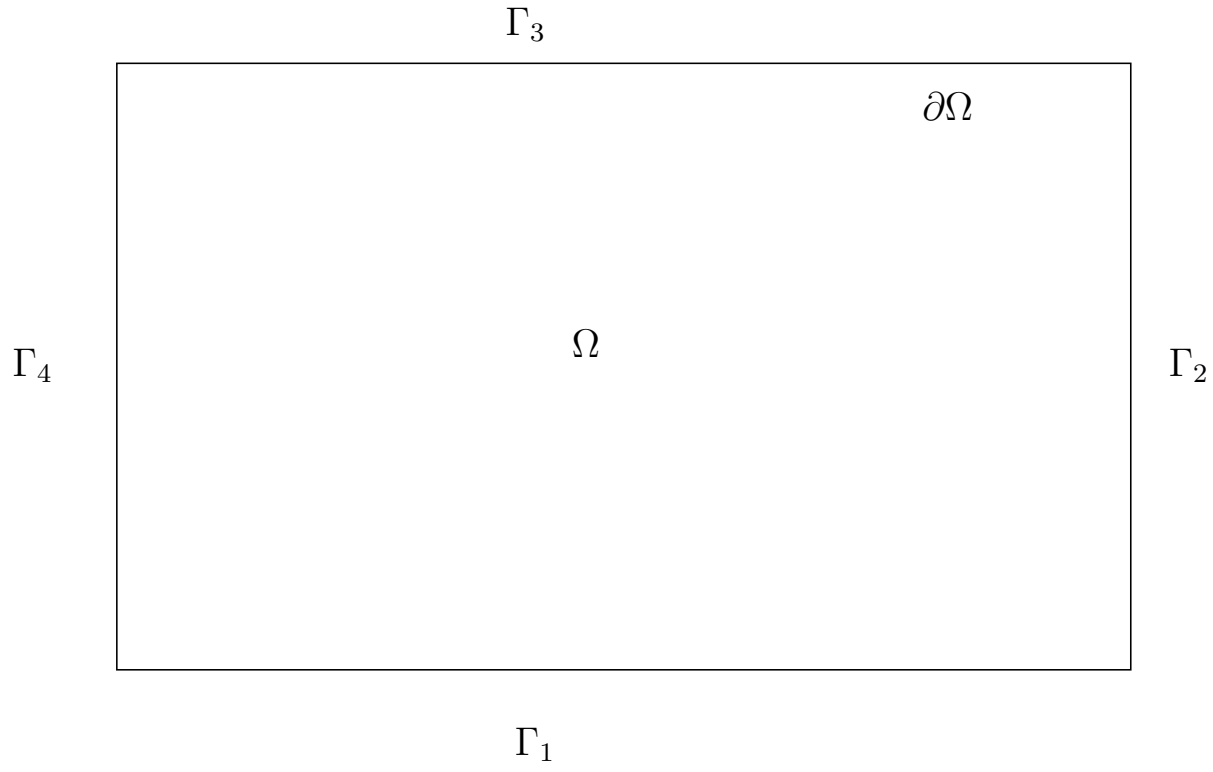


Figure 1: The domain Ω.

polygons are incomplete). For primary mesh node i, we will call this polygon B_i exclusive subdomain of node i. B_i consists of several abscissae and each of abscissa belongs to one abscissa connecting node i with his neighbor m. For each couple i, m there two abscissae, we will denote them $\partial B^l_{i,m}$. The middle point of the abscissa $\partial B^l_{i,m}$ is denoted $\gamma\partial B^l_{i,m}$ (fig. 2). Time step will be denoted by upperscript k. By |"something"| is denoted the area or the length of "something" (for example $|\partial B^l_{i,m}|$ is the length of abscissa $\partial B^l_{i,m}$).

The coupled system of equations is solved as follows. First the number concentration n based on the coverage at old time level is computed. Then new surface coverage is computed.

We will show how to solve equation (13). First we will integrate this equation over domain Ω:

$$\int_\Omega \left[\frac{\partial n}{\partial t} + \frac{K_a(\theta_f,\theta_u)}{\gamma} n - \frac{K_r(\theta_f,\theta_u)}{\gamma}\right] \mathrm{d}S = \int_\Omega \left[\nabla\cdot(D\nabla n) - \nabla\cdot(\mathbf{V}\cdot n)\right] \mathrm{d}S. \qquad (17)$$

Now we will use Gauss formula on the right hand side of the equation (17):

$$\int_\Omega \left[\nabla\cdot(D\nabla n) - \nabla\cdot(\mathbf{V}\cdot n)\right] \mathrm{d}S = \int_{\partial\Omega} (D\nabla n)\cdot\mathbf{n}_{\partial\Omega}\mathrm{d}l - \int_{\partial\Omega} (\mathbf{V}\cdot n)\cdot\mathbf{n}_{\partial\Omega}\mathrm{d}l \qquad (18)$$

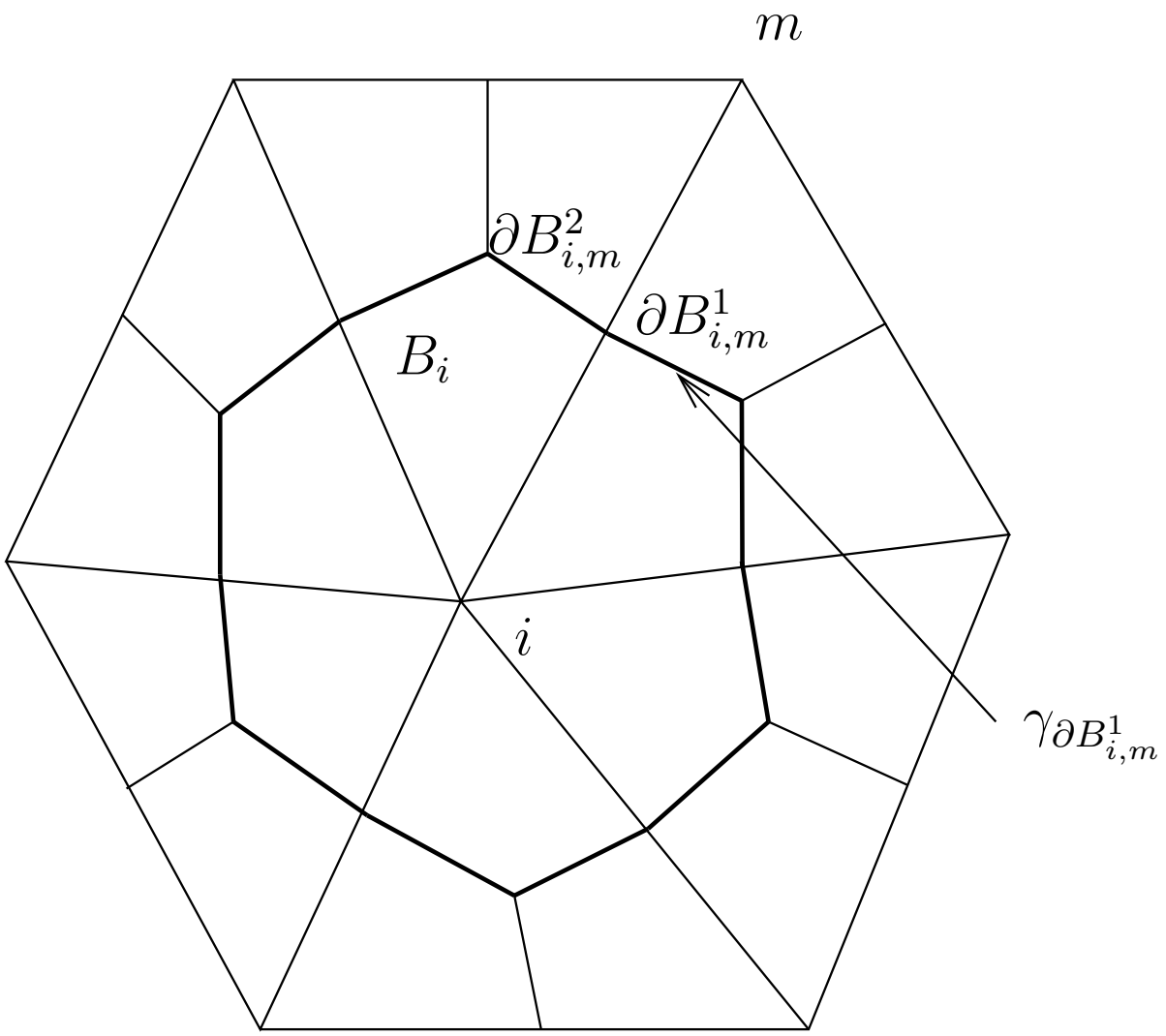


Figure 2: The exclusive subdomain for node i.

where $\partial\Omega$ is boundary of Ω and $\mathbf{n}_{\partial\Omega}$ is the normal vector to $\partial\Omega$ But the mass balance has to be satisfied not only on the whole domain Ω but also on each exclusive subdomain B_i which belongs to the primary mesh node i:

$$\int_{B_i}\left[\frac{\partial n}{\partial t}+\frac{K_a(\theta_f,\theta_u)}{\gamma}n-\frac{K_r(\theta_f,\theta_u)}{\gamma}\right]\mathrm{d}S=\int_{\partial B_i}(D\nabla n)\cdot\mathbf{n}_{\partial B_i}\mathrm{d}l-\int_{\partial B_i}(\mathbf{V}\cdot n)\cdot\mathbf{n}_{\partial B_i}\mathrm{d}l \quad (19)$$

Now we will approximate the left hand side of (19).

$$\begin{aligned}\int_{B_i}\left[\frac{\partial n}{\partial t}+\frac{K_a(\theta_f,\theta_u)}{\gamma}n-\frac{K_r(\theta_f,\theta_u)}{\gamma}\right]\mathrm{d}S\approx\\ \approx\left[\frac{n_i^{k+1}-n_i^k}{\Delta t}+\frac{K_a(\theta_{f,i}^k,\theta_{u,i}^k)}{\gamma}n_i^k-\frac{K_r(\theta_{f,i}^k,\theta_{u,i}^k)}{\gamma}\right]|B_i| \quad (20)\end{aligned}$$

Approximation of the first therm on the right hand side of (19):

$$\begin{aligned}\int_{\partial B_i}(D\nabla n)\cdot\mathbf{n}_{\partial B_i}\mathrm{d}l&=\sum_{m,l}\int_{\partial B_{i,m}^l}(D\nabla n)\cdot\mathbf{n}_{\partial B_{i,m}^l}\mathrm{d}l\\ &\approx\sum_{m,l}\left[(D(\gamma_{\partial B_{i,m}^l})(\nabla n)^k(\gamma_{\partial B_{i,m}^l}))\cdot\mathbf{n}_{\partial B_{i,m}^l}|\partial B_{i,m}^l|\right],\end{aligned} \quad (21)$$

where $(\nabla n)_i^k)$ is the approximation of ∇n from concentration values from time step k. Approximation of the second therm on the right hand side of (19)

$$\int_{\partial B_i}(\mathbf{V}\cdot n)\cdot\mathbf{n}_{\partial B_i}\mathrm{d}l=\sum_{m,l}\int_{\partial B_{i,m}^l}(\mathbf{V}(\gamma_{\partial B_{i,m}^l})\cdot n_{i,m,l}^\star)\cdot\mathbf{n}_{\partial B_{i,m}^l}|\partial B_{i,m}^l|, \quad (22)$$

where upwind value is given as:

$$n_{i,m,l}^\star=\begin{cases}n_i^k & \text{for}\quad \mathbf{n}_{\partial B_{i,m}^l}\cdot\mathbf{V}(\gamma_{\partial B_{i,m}^l})>0\\ n_m^k & \text{for}\quad \mathbf{n}_{\partial B_{i,m}^l}\cdot\mathbf{V}(\gamma_{\partial B_{i,m}^l})\le 0\end{cases} \quad (23)$$

The approximation (22) is called the first order upwind scheme and helps us to avoid oscillations in the solution, but suffers of the numerical diffusion. To obtain smaller numerical diffusion without oscillations higher order upwind scheme with limiter (without limiters, there are small oscillations in the solution) can be used.

Values of B_i^k on the boundary $\partial\Omega$ are taken from the boundary conditions. θ_f and θ_u for the first time step can be θ_f and θ_u taken from the initial condition for them. Than we will give approximations (20), (21) and (22) together, find n_i^{k+1} and obtain the explicit scheme.

We can use explicit scheme for equations (15) and (16) to obtain from known surface coverage from old time step θ_l^k and calculated number concentration n_i^{k+1} to obtain new particle coverage θ_l^{k+1} for favorable case $l=f$ and unfavorable case $l=u$:

$$\theta_{l,i}^{k+1}=\theta_{l,i}^k+\pi a_p k_{dep,l,i}n_i^{k+1}B(\theta_{l,i}^k)-k_{det,l,i}\theta_{l,i}^k\qquad l\in\text{f,u}. \quad (24)$$

where $\theta_{l,i}^k$, $k_{dep,l,i}$, $k_{det,l,i}$ are values of θ_l^k, $k_{dep,l}$, $k_{det,l}$ in the node i.

5 Conclusion

In this contribution a summary of equations describing the colloid transport was presented and discretization of equations by means of first order upwind scheme was derived.

Future work will be focused on behavior of colloids in the porous media and especially of nanocolloids.

6 Acknowledgment

This work was supported by grants:

- Development and Validation of Porous Media Flow and Transport Models for Subsurface Environmental Application, project of Czech Ministry of Education, Youth and Sports Kontakt ME878, 2006-2009, support for cooperation with the Colorado School of Mines, Golden
- Applied Mathematics in Technical and Physical Sciences, Research Direction Project of the Ministry of Education of the Czech Republic No. MSM6840770010, principal investigator Prof. K. Kozel, Faculty of Mechanical Engineering of CTU in Prague, 2005-2009

References

[1] N. Sun, M. Elimelech, N.-Z. Sun *A novel two-dimensional model for colloid transport in physically and geochemically heterogeneous porous media.* 'Journal of Contaminant Hydrology 49', (2001), 173–199.

[2] N.-Z. Sun *Mathematical Modeling of Groundwater Pollution.* Springer-Verlag, New York.

[3] N.-Z. Sun W.W.-G, Yeh, *A proposed upstream weight numerical method for simulating pollutant transport in groundwater.* 'Water Resour. Res. 19 (1983) 1489–1500.

[4] J.N. Ryan, M. Elimelech *Review Colloid mobilization and transport in groundwater.* 'Colloids and Surfaces A: Physicochemical and Engineering Aspects 107 (1996) 1–56.

[5] W.R.Penrose, W.L. Polzer, E.H. Robertson and K.H. Abel Environ. Sci. Technol., 24 (1990) 228

The Q-Buf Kernel Stream Buffering Engine*

Martin Dráb

5th year of PGS, email: martin.drab@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Ladislav Kalvoda, Department of Solid State Engineering, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. Whenever you have to deal with a device that is generating data either based on external triggering events or that is constantly delivering realtively huge amounts of somehow acquired data and has no means to store them locally in itself for longer periods of time or just the possibility to deliver them on request, there are just two things you can do about it to handle the device in a proper way. Either you have to run on a hard-real-time operating system, or you stick with the classical time-sharing OS, but have to have a mechanism that would do the best possible effort to service the data under given circumstances to prevent any losses. And this is the point where the Q-Buf engine takes place.

Abstrakt. Kdykoliv máte co do činění se zařízením, které buď generuje data na základě externě triggerovaných událostí nebo které konstantně produkuje relativně velké množství nějakým způsobem získaných dat a nemá žádnou možnost tato data v sobě po delší dobu skladovat či je zkrátka zaslat vždy jen na požádání, pak jsou jen dvě možnosti, jak zařízení správným způsobem obsloužit. Buď běžet pod hard-real-time operačním systémem nebo zůstat na klasickém time-sharing OS, ale pak je třeba mít k dispozici mechanismus, který se bude snažit obsloužit toto zařízení nejlepším možným způsobem za daných podmínek tak, aby pokud možno předešel jakýmkoliv ztrátám dat. A to je právě to místo, pro které byl konstruován Q-Buf engine.

1 Introduction

In project INDECS [1] we collect raw data signals from the position sensitive detectors (PSD) using an ADLink PCI-9812 data acquisition card [2]. Up until now, the data had to be collected under RTLinux hard-real-time operating system, so that we miss as little of the incoming neutron events as possible. Partially also because of the nowadays relatively slow and old PC hardware where the data acquisition card was installed (PII 400 MHz).

Using the free (or also called Open) variant of the RTLinux [3] does have its advantages in data acquisition, since it provides a true hard-real-time OS below the classical time-sharing variant of Linux kernel. However, it also has its drawbacks.

For a long time there have only been free RTLinux patches for the old 2.4 variant of the Linux kernel, but no recent Linux distribution uses these types of kernel. And even when you do manage to compile such a kernel on any recent distribution, it lacks a lot of features (drivers for recent hardware, security patches, and so on). Recently a free

*This work has been supported by grants MSM6840770040 and MPO contract IMPULS No. FI-IM3/136.

RTLinux patch for the 2.6.9 Linux kernel appeared. But 2.6.9 is also a history today. So, keeping up-to-date with current linux kernels is always a bit of a problem when you want to use RTLinux.

Another task, that was there to face (this time not related to project INDECS), was to write a reliable driver for the DAKEL's DTR devices used for data acquisition from and transmission to multiple ultrasound probes. Unlike the PCI-9812 DAQ card, which uses about 10 times higher sampling rates, but just one short discrete burst on each detected neutron event, this is a true continuously streaming device and loosing samples would be perhaps even more unacceptable than in case of the PSD in project INDECS.

Nonetheless, there was another obstacle related to this particular driver. The DTR device requires very fast and huge hard disks at its disposal to store all the streaming data and that means recent computer hardware and for that also recent drivers. So, going for the free RTLinux solution was not the best option to choose, not withstanding the additional licensing difficulties that might (or migt not) matter in this case.

Q-Buf kernel streaming engine was created to solve these two problems. It is just a module of the standard 2.6 Linux kernel [4] maintained compatible with all of the up-to-date 2.6 kernels (at the time of writing this article, the most recent is the 2.6.26 kernel) and perhaps one day we manage to include it directly into the linux kernel source tree, so that it would be kept compatible with all the kernel changes automatically.

2 Concept

There are several simple ideas behind the concept of the Q-Buf engine:

- First of all, whenever there are some incoming data available, receive them as soon as possible, without any unnecessary delays.
- Try all available options before giving up on the data.
- Try keeping some extra memory reserves if possible, so that you don't get caught in a situation, where there is no memory available immediatelly, when it is needed the most.
- Try to keep the obtained data as long as necessary before the user-space application processes them.
- If it is about memory consumption, this device has the absolute priority. Meaning that it is allowed to consume all memory available, no matter how much the other processes on the system may suffer because of that. However, to prevent the total consumption of all memory which would probably result in a collaps of the system, this consumption is limited by an arbitrary preset value.

  This idea together with the previous one is generally considered a bad and at least impolite behaviour among kernel-space code. However, as stated above, the data generated by the device that is using the Q-Buf engine are considered so valuable, that they outweight almost every other data with which they may possibly compete for memory space, and thus, such a behaviour is well justified.

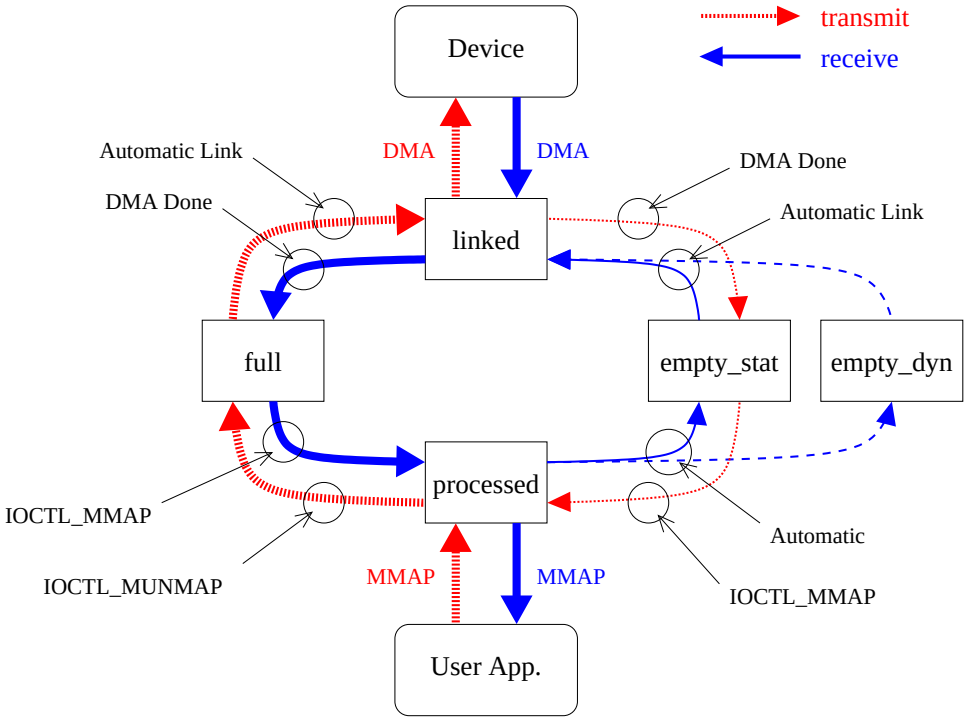


Figure 1: Block schematics of the Q-Buf engine data flow.

- Keep the interface to the driver using the Q-Buf engine simple and let the driver implement only the necessary parts of the whole mechanism, that are different for each device.
- Clean up the claimed resources when the device is not going to need them anymore, so that even if the device consumed a lot of resources, the system may operate smoothly again when the device no longer needs them.
- On the other hand, while the device is still transfering, it may be wise to keep the resources it consumed so far, because there is a high probability that it would need them again soon, and freeing and acquiring them again later would consume the valuable time, and thus, increase the odds of loosing data.

With all these ideas in mind, the concept of static and dynamic streaming buffers was introduced.

2.1 Buffers

The basic structures for data transfer within the Q-Buf engine are referred to as Buffers. Buffers carry reference to the actual data buffer, reference to a custom structure of the driver, should the driver have some information bound to the buffer (usually some structure related to the actual data transfer over the appropriate bus), position of the carried data in the stream, and several other information related to the buffer and the data within.

Careful thought must be taken when allocating the data buffers. We want it to be able to both do the DMA transfers to and/or from the device and be able to do memory mapping into user-space. The first condition requires a consistent continuous block of memory for each buffer, since lots of the devices (such as most USB host controllers, for example) cannot do scatter–gather (or so called *vectored I/O*) DMA. The second condition requires us to have the data buffers alligned to pages in both size and memory position, because memory mapping can be done only by entire pages and exposing any area not dedicated to the actual data buffer would be a potential security risk.

Since most systems use pages of 4 KB[1] in size, the above means that we would have to have transfers quantized by the multiples of that size. However, that is not always the best option. The transfer packets for the appropriate bus usually have different size, not necessarily even size of power of 2. This problem is solved by allocating space for the buffer that would comply to the two constraints mentioned in the previous paragraph, but the actual buffer would use only a subregion of the allocated area.

Considering that that would be a fair waste of memory, had the packet size been significantly smaller than the page size, we allow multiple consequent transfer packets per buffer. The choice of the number of packets per buffer can be chosen by the driver that uses the Q-Buf for the particular device. In fact the usage of the whole buffer area is left up to the driver, the Q-Buf engine just supports slicing the buffers into smaller areas for data exchange and the driver decides what slicing is best suitable for the device. There may even be an unused area at the end of the buffer, if the buffer size is not divisible by the packet size or if there is artificially less packages assigned for one buffer than the maximum possible. And while the bus transfers need not be reliable in some cases, some packets may even deliver no data, so, there may be gaps between some data within a buffer. To make it short, Q-Buf also supports subdividing the occupied buffer space by discrete sparse data blocks.

Another problem of the conditions mentioned three paragraphs before is that allocating memory areas consisting of more than one consequent memory page may stress the memory allocator quite a lot due to the memory fragmentation effect. That means, that allocating a buffer may take relatively a lot of time, if the memory is too fragmented, and it may even not succeed at all because of it. Let alone if we consider that the data buffer should be in the DMA region of that particular device, which for some devices may not even be the entire area of physical memory, but just its subsection. For example devices with a 32-bit (or less) DMA controller on a 64-bit system with more than 4 GB of physical system memory.

2.2 Dynamic Buffers

The one thing, in which the Q-Buf engine even has a slight advantage over the true hard-real-time approach is the fact, that it is allowed to allocate new buffers even after initialization and while the transfer is running.

The defining quality of a hard-real-time operating system is that it guarantees the maximal response time to external events (that means even in the worst case scenario). Simply put, if you have a code that is servicing some external event, let's say an interrupt,

[1] Or more. The page can be as much as 4 MB in size depending on hardware and system configuration.

and you run that code as a real-time thread, then the real-time OS guarantees (and this word is essential) you that it would be scheduled to run within certain very short and well defined period of time from the point where the interrupt occured in time, no matter what the rest of the system does at the moment or is just about to do.

However, that comes at a price. The real-time threads are very limited in what are they allowed to do (at least in the RTLinux, but other real-time OSes have it similarly). One of the major restrictions is, that a real-time thread can not allocate new memory while it is running. Only during the initialization of that thread.

The reason for that is that it is a non-deterministic operation. System needs to run through the page tables and find an appropriate number of free physical pages which it then maps into the virtual space of that thread. That by itself may take some time, if there is a lot of pages and high memory fragmentation. But when there is not enough physical memory available, system has to move some pages into a swap space, suspend the thread, wait until the transfer is done, and then acquire those pages that have been freed that way. This is totally unacceptable to be done in a real-time thread, since it is not possible to predict in advacnce how long shall that take, and it may take a lot.

This results in the necessity to preallocate all the buffers that the real-time thread is going to use for data transfer before the thread is launched and there can be no more added later. If it then turns out that the preallocated buffers are insufficient because the application that does the processing fails to catch up (perhaps because of the hard drive where the data are stored or for whatever other reason), you may possibly loose data.[2] Problem is to determine how many buffers are going to be necessary. If we preallocate too little, we loose data. If we preallocate too much, we may have blocked too much of the physical memory and that may slow down the whole system significantly.

In standard time-sharing OS, we don't have this problem. We can allocate as much as we want to and almost any time we want to. So we use it. But to prevent loosing data for the same reason why real-time thread cannot allocate buffers, we try to preallocate. The Q-Buf engine launches another parallel thread, which allocates several buffers when there is less than some threshold empty buffers available. Both the threshold and the amount of preallocated buffers at a time can be preset. As this runs in parallel, it can do the dirty job for the main thread without delaying it and though increasing the chance of loosing data. Also the higher the threshold is set, the more time the thread has to preallocate the buffers before they run out.

Buffers that the Q-Buf engine allocates this way we refer to as dynamic buffers and they receive special treatment. These buffers can compensate for a lot of throubles that the driver of the device we are talking about might otherwise have. We restrict the amount of physical memory that the buffers can occupy all together to prevent system crashdown by running completely out of physical memory.

2.3 Buffer Queues

Buffers are organized in buffer queues. Each buffer queue is implemented by a standard doubly-linked list with a sentinel node. Every buffer that is put to active duty has to be

[2]Though this situation should also be handled by a real-time thread to be guaranteed to make it in time. But that also brings in other restrictions.

linked in exactly one buffer queue at a time depending on its current status.

2.3.1 Empty Static Buffer Queue

Empty Static buffer queue holds the so called static buffers when they are not used. Static buffers are the buffers that are allocated at the time when the Q-Buf engine is initialized for the specific device. These buffers live in the system for as long as the device exists. They do the most of the work. If an empty buffer is needed, static buffer is always preferred to the dynamic one.

2.3.2 Empty Dynamic Buffer Queue

Empty Dynamic buffer queue contains all the dynamic buffers when they are not used. Unlike Empty Static buffer queue, this buffer queue is initialized empty and is populated with buffers only when there is not enough buffers in the system to do the job. All the dynamic buffers in this queue live only as long as the device is opened by at least one application. As soon as the device is closed, all the dynamic buffers are released.

Reason for this behaviour is in that we do not want to block the memory occupied by the dynamic buffers for longer than necessary, but on the other hand when at some point more buffers were necessary, there is a good chance, that while the transfer is still running, those buffers shall be needed again. Releasing the dynamic buffers after the device is closed seems to be a reasonable compromise between memory usage and complexity of the mechanism that handles the living period of the dynamic buffers.

If this prooves to be insufficient, other methods may be introduced. Such as for instance checking the time from the last usage of a dynamic buffer. If one would not be used longer than some threshold period, it shall be freed. This would be even better with respect to the memory consumption, but it has an unnecessary and not completely insignificant processing overhead, which is the reason why it is currently not used. However, at least for the moment, the currently used method seems to be sufficient for the task.

2.3.3 Link Buffer Queue

Link buffer queue contains the buffers that are currently transferring data to or from the device or those that are scheduled for the transfer by the system already. While the transfer is running, this buffer queue must never get empty. If it does then it means that either there is not enough empty buffers available to handle the incoming data (in case of data reception) or the application is unable to deliver data as quickly as needed (in case of data transmission). Either way this means trouble, either we loose data or the transmitted stream will have discontinuities.

2.3.4 Full Buffer Queue

Buffers in the Full buffer queue are filled with data and are waiting for further processing. Either they can go for transfer in case of transmitting direction, or to user application in case of receiving direction.

2.3.5 Processed Buffer Queue

Processed buffer queue holds buffers that are currently directly or indirectly mapped to the user.

2.4 Offsets

Initially buffers were meant to be mapped into user-space directly one by one. Nontheless, Preliminary tests indicated that this would cause very frequent switching between user-space and kernel-space, and thus, bring a significant overhead delaying the user-space processing. The smaller the buffers the higher the overhead. But we should not have very big buffers if we are to pursue the data in something close to real time.[3] This effect was observed when dealing with the streaming DAKEL DTR device.

To make things more flexible, a structure called *offset* was introduced to the process. This structure is used to memory map the data buffers into the user-space and it can contain and thus map more buffers at once. The number of buffers is limited by some upper threshold, which can again be preset, and the consideration on the correct value of this threshold should account for the size of each buffer and the data rate of the device.

The main idea behind this was to let the amount of data be determined by the time when the user-space application asks for it. If the processing of the data does not take too much time relative to the acquired data flow, then the application can allow to ask for the data more often and it gets just few buffers each time, because there would be no more available at the time. However when the processing takes more time either regularly or occasionally due to some external unexpected delays, the time between subsequent data requests from the user-space application would be relatively long and the driver may accumulate more buffers in that period, so the application would get more of them[4] at one request, so that the data are processed quicker.

Buffers within one offset are mapped at the same time, which also means that they are released at the same time as well. Until all of the data are processed, all buffers within the offset remain mapped, and though, occupied. That is also necessary to account for.

Before the buffers within an offset are mapped into the user-space, the data areas of all the contained buffers are concatenated together and can be sliced by the driver into sparse subblocks. The application is then provided with a map of the areas where the data are. The advantage of this approach is that when the data areas are filled completely with acquired data for each buffer, the application sees it as one continuous block of data. But when that is not true for some reason, the application knows where the data can be found, but it has to be aware of that fact and not just blindly read everything.

2.4.1 Offset Queues

All offsets are also dynamically allocated resources, that are stored in special offset queues. There are just two of them. One for the offsets that are currently in use[5] and one for

[3] In this case not refering to the real-time OS, but rather to actual human sensed perspective of watching the data as they are received, which is usually desired.

[4] Up to the upper threshold, of course.

[5] Meaning that they contain buffers and provide a user-space mapping.

the offsets that are currently unused. Each offset structure has its defined offset in the device-space[6] from which it can then be mapped into the user-space memory.

2.5 Application Interface

To make the usage of a device that is using the Q-Buf engine simpler, there are two userspace interfaces that can be used to handle the data.

First there is the mmap(2) interface to fully use the potencial of the Q-Buf engine. This interface allows to directly access the DMA buffers of the device. However to know where to map them from and how the data are organized within these areas, the appliaction has to use the special ioctl(2) call with the IOCTL_MMAP command for the device prior to the mmap(2) to get the specific information. It is a bit more complicated to deal with it, but on the other hand, it is the most effective way, since there is no unnecessary overhead of copying the data in the kernel-space and that may be significant, since there can be great amounts of data transferred.

To let the device be operated easily perhaps by common applications and system commands (like *cp(1)*, *dd(1)*, etc.) a standard read(2)/write(2) interface was also added to the Q-Buf interface. But the luxury of usage by the common applications and continuous data stream is paid by the possibly significant overhead of another copy from the buffers of the Q-Buf engine to the user-space buffers of the application.

3 Conclusion

The described Q-Buf engine seems to fulfill the expectations and it makes the servicing of the DAKEL's DTR devices and the PCI-9812 DAQ card for project INDECS possible even without the hard-real-time OS. The first case is already being successfully used in the real applications of physics for quite some time. The latter case is still in early testing, but the preliminary results look very promissing.

References

[1] M. Dráb. *Project INDECS: The Design of a Superior System of Data Collection and Analysis*. (Diploma Thesis) Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague, (2003).

[2] M. Dráb. *Position Sensitive Detector Data Acquisition Path*. Doktorandské dny 2007, sborník workshopu doktorandu FJFI oboru Matematické inženýrství, FNSPE, Czech Technical University in Prague, ISBN 978-80-01-03913-7, pp. 19 – 26, Nov. (2007)

[3] M. Barabanov, V Yodaiken. *Real-Time Linux*. Linux Journal, Feb. (1997).

[4] J. Corbet, A. Rubini, G. Kroah-Hartman. *Linux Device Drivers, 3rd Edition*. O'Reilly Media, Inc. Sebastopol, CA, U.S.A., ISBN: 0-596-00590-3, Feb. (2005).

[6] The space of the special unix file that represents the actual device and from which the buffers are mapped into the user-space from the perspective of the application.

Implicit Numerical Scheme for Modelling Dynamic Effect in Capillary Pressure

Radek Fučík

2nd year of PGS, email: fucik@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Jiří Mikyška, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. In order to investigate effects of the dynamic capillary pressure-saturation relationship used in the modelling of flow in porous medium with material discontinuities, a one-dimensional fully implicit numerical scheme is proposed and its validity is discussed by means of semi-analytical solutions developed by McWhorter and Sunada and by the authors. The numerical scheme is used to simulate experimental procedure using the measured dataset for the sand and fluid properties. Results of the simulation using different models for dynamic effect term in capillary pressure - saturation relationship are presented and discussed.

Abstrakt. V článku je prezentován jednorozměrný model dvoufázového nemísivého a nestlačitelného proudění který je použit na zkoumání vlivu dynamického efektu pro model kapilárního tlaku v závislosti na saturaci v porézním prostředí. Navržené numerické schéma je plně implicitní v čase a je porovnáno se semi-analytickým řešením McWhortera a Sunady. Takto ověřený numerický model je použit k simulaci laboratorních experimentů s cílem posoudit vliv různých modelů pro koeficient dynamického efektu na řešení jednorozměrné úlohy.

1 Background

This manuscript focuses on the dynamic phenomena in the capillary pressure - saturation relationship that has been examined in various papers in the past decades. The main objective is to propose a numerical scheme that implements the dynamic capillary pressure - saturation relationship for heterogeneous porous media.

Fundamental constitutive quantities used in modelling of flow in porous media are described in the following subsections. Thorough definitions, descriptions, and examples can be found in [7], [17], [1], [16], or [2].

1.1 Wettability

As two immiscible phases are present in the porous media, a meniscus of fluid-fluid interface is formed as a result of the presence of the solid phase (sand grains). The interaction between adhesive and cohesive forces within the fluids leads to the specific angle ϑ between the solid surface and the fluid-fluid interface. The wettability of fluid is then determined as:

$\vartheta = 0$	$\vartheta \in (0, \frac{\pi}{2})$	$\vartheta > \frac{\pi}{2}$
completely wetting,	partially wetting,	non-wetting.

1.2 Saturation

The fluid distribution in immiscible multiphase flow in porous media is described by the saturation S_α $[-]$. It indicates the volumetric portion of void space within the pores occupied by the fluid phase α, hence, S_α is always between 0 and 1. The sum of saturations S_α of all fluids present in the porous media is 1, i.e., $\sum_\alpha S_\alpha = 1$.

Since not all volume of the fluid phase can be displaced in the multiphase flow from a porous medium due to hysteretic effects, the α-phase residual saturation quantity $S_{r\alpha}$ $[-]$ is introduced. It expresses the minimal saturation of the phase α that will retain in the porous medium due to adhesion effects with respect to the solid matrix. Therefore, the effective saturation S_α^e $[-]$ that describes only volumetric portions of displaceable fluid phases is introduced as

$$S_\alpha^e = \frac{S_\alpha - S_{r\alpha}}{1 - \sum_\beta S_{r\beta}}. \tag{1}$$

1.3 Capillary pressure

Following the standard definitions in literature, the capillary pressure p_c $[ML^{-1}]$ on the pore scale is defined as the difference between the non-wetting phase pressure p_n $[ML^{-1}]$ and the wetting phase pressure p_w $[ML^{-1}]$, i.e.,

$$p_c = p_n - p_w. \tag{2}$$

The capillary pressure function has been commonly considered as a function of wetting phase saturation only and it has been widely used in model equations in literature, see for instance [20], [11], [8], or [9].

1.4 Dynamic capillary pressure

The classical capillary pressure - saturation relationships such as [4] or [23] has been used in almost all mathematical studies on modelling of porous media flow in the past decades. Recently, theoretical studies [15], [14], [6], [12], [13], [3], as well as the empirical approach in [22] have produced new aspects in the two-phase flow theories. The most important result is that the classical capillary pressure - saturation relationship holds only in the state of thermodynamic equilibrium. Therefore, the classical approach cannot be used in the modelling of capillarity when the fluid content is in motion. Consequently, a new capillary pressure - saturation relationship is proposed in the following form:

$$p_c := p_n - p_w = p_c^{eq} - \tau \frac{\partial S_w}{\partial t}, \tag{3}$$

where p_c^{eq} is the capillary pressure - saturation relationship in equilibrium and τ $[ML^{-1}T^{-1}]$, the dynamic effect coefficient, is a material property of the system.

Early in 1978, Stauffer [22] proposed a linear dependence in (3) and proposed the following definition of τ:

$$\tau_S = \frac{\alpha_S \mu_w \Phi}{K\lambda} \left(\frac{p_d}{\rho_w g} \right)^2, \tag{4}$$

where $\alpha_S = 0.1$ $[-]$ denotes a scaling parameter, μ_w $[ML^{-1}T^{-1}]$ is the wetting phase dynamic viscosity, Φ $[-]$ is the porosity of the material, K $[L^2]$ is the intrinsic permeability, ρ_w $[ML^{-3}]$ is the wetting phase density and g $[LT^{-2}]$ is the gravitational acceleration constant. Both λ and p_d are the Brooks and Corey parameters ([4]) that can be experimentally estimated. Thus, the coefficient τ_S can be calculated for a given porous medium and wetting fluid.

The Stauffer model for the dynamic effect coefficient τ was obtained by correlating experimental data. The values of τ_S vary between $\tau_S = 2.7{\cdot}10^4\ Pa s$ and $\tau_S = 7.7{\cdot}10^4\ Pa s$, see [17, page 27]. However, other researchers suggest that the magnitude of τ should be in the order of $10^2 - 10^3\ Pa\,s$, [5], or, on the other hand, it should be also in the order of $10^4 - 10^8\ Pa\,s$ as estimated in [14].

Recently, a more general nonlinear dependence $\tau = \tau(S_w)$ is assumed to be more relevant in modelling of realistic two-phase flow displacement [21]. In this manuscript, both constant and linear model will be used in numerical simulations.

2 Mathematical model

A mathematical model describing the two-phase flow in a onedimensional domain is presented in this section. The aim is to investigate how the inclusion of the dynamic capillary pressure relationship (3) instead of the classical relationships in thermodynamic equilibrium influences solution of the two-phase flow system of equations (5).

2.1 Governing equations

The governing two-phase flow equations in one-dimensional domain $[0, L]$ are given by the $p_w - S_n$ formulation [1]:

$$\Phi \frac{\partial S_\alpha}{\partial t} = \frac{\partial}{\partial x}\left[\frac{K}{\mu_\alpha} k_{r\alpha}\left(\frac{\partial}{\partial x}(p_w + \delta_{\alpha n} p_c) - \rho_\alpha\, g\right)\right], \tag{5}$$

S_α	Saturation $[-]$,	p_α	Pressure $[ML^{-1}T^{-2}]$,
ρ_α	Density $[ML^{-3}]$,	μ_α	Dynamic viscosity $[ML^{-1}T^{-1}]$,
g	Gravitational acceleration $[LT^{-2}]$,	Φ	Porosity $[-]$,
K	Intrinsic conductivity $[L^2]$,	$k_{r\alpha}$	Relative permeability $[-]$,

where $S_w + S_n = 1$, $\delta_{\alpha n}$ is the Kronecker symbol, and $\alpha \in \{w, n\}$. The wetting fluid (water) and non-wetting (air, NAPL[1]) fluid are indexed by w and n, respectively. The initial and boundary conditions for (5) are given separatedly for each experimental problem.

2.2 Discrete problem

A standard finite difference discretization technique is used in order to determine approximate discrete solution $S^k_{n,i}$, $p^k_{w,i}$ of the problem (5), generally defined as $f^k_i = f(k\Delta t, i\Delta x)$, where $i = 0, 1, \ldots, m$, $m\Delta x = L$, and $k = 0, 1, \ldots$.

[1]Non-Aqueous Phase Liquid

Since the nonlinear problem (5) involves the dynamic capillary pressure function defined in (3) that includes time derivative of S_n, an implicit numerical scheme is proposed in the following form:

$$\Phi \frac{S_{\alpha,i}^{k+1} - S_{\alpha,i}^{k}}{\Delta t} = -\frac{u_{\alpha,i}^{k+1} - u_{\alpha,i-1}^{k+1}}{\Delta x}, \tag{6}$$

where $\alpha \in \{w, n\}$ and the discrete Darcy velocities u_α are given as follows

$$u_{\alpha,i}^{k+1} = -\frac{K}{\mu_\alpha} k_{r\alpha}(S_{\alpha,upw}^{k+1}) \Big(\underbrace{\frac{p_{w,i+1}^{k+1} - p_{w,i}^{k+1}}{\Delta x} + \delta_{\alpha n} \frac{p_{c,i+1}^{k+1} - p_{c,i}^{k+1}}{\Delta x} - \rho_\alpha g}_{\Delta \Phi_\alpha} \Big), \tag{7}$$

$$p_{c,i}^{k+1} = p_c \left(1 - S_{n,i}^{k+1}, -\frac{S_{n,i}^{k+1} - S_{n,i}^{k}}{\Delta t} \right).$$

and $S_{\alpha,upw}^{k+1}$ is the saturation taken from the upwind direction with respect to gradient of the phase potential Φ_α, i.e.

$$S_{\alpha,upw}^{k+1} = \begin{cases} S_{\alpha,i+1}^{k+1} & \text{if } \Delta\Phi_\alpha \geq 0. \\ \\ S_{\alpha,i}^{k+1} & \text{if } \Delta\Phi_\alpha < 0. \end{cases}$$

The numerical scheme is solved using the Newton-Raphson iteration method. The Jacobi matrix used in the Newton iteration method is block tridiagonal.

3 Numerical experiments

3.1 Validation of numerical scheme

The numerical scheme (6) is validated using the McWhorter and Sunada semi-analytical solution [18], [19], [24], [9], and [10]. A special configuration of the problem (5) is assumed in order to obtain such a semi-analytical solution. Neither gravity nor dynamic effect is considered and the inflow boundary condition at $x = 0$ consists of a time-dependent input flux $u_w(t, 0) = A/\sqrt{t}$, [7], [9], and [10]..

In Figure 1, the numerical solution is compared to the semi-analytical solution obtained for the same sand and fluid properties. As the numerical grid gets finer, the agreement of the numerical solution with respect to the semi-analytical solution is apparent. However, estimation of the error of convergence (EOC) cannot be used in this case because the semi-analytical solution cannot be obtained with a sufficient precison for finer grids Hence, only graphical representations are relevant.

3.2 Column experiment

The dynamic effect coefficient $\tau = \tau(S_w)$ was estimated as a result of a laboratory experiment held in CESEP, Colorado School of Mines for a given sand. In this section,

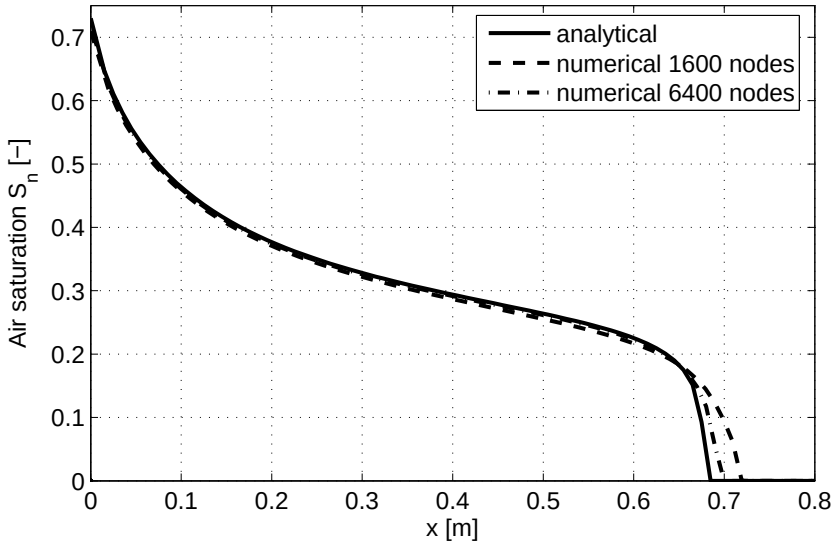


Figure 1: Numerical solution and semi-analytical solution at time $t = 1000\ s$.

the experimental setup is approximated by the one-dimensional problem (5) and the experimental dataset is used in numerical simulations.

The laboratory experiment consists of a single, vertically placed, 10 cm long tube filled with sand. Initially, the column is flushed with water such that no air phase is present inside. Then, a series of drainage and imbibition experiments is proceeded and values of the capillary pressure and the saturation of air are measured by probes in the middle of the column. As a result, two models of $\tau = \tau(S_w)$ were estimated : constant $\tau(S_w) = C$ and linear $\tau(S_w) = C(1 - S_w)$.

The numerical scheme is used to simulate drainage of the column with exponentially decreasing outflow of water at $x = 10\ cm$ for three different models for capillary pressure: static capillary pressure ($\tau = 0$), constant, and then linear model for dynamic capillary pressure. The numerical solutions plotted versus time in the middle of the column are shown in Figure (2).

As expected, the saturation profiles does not exhibit large differences between the models of dynamic effect term. On the contrary, the capillary pressure temporal profile for the constant model for dynamic effect term has completly different history than that of static capillary pressure or linear model for dynamic effect term. As a result of this observation, the constant model for dynamic effect gives significantly different results with respect to temporal monotonicity than the linear model or the static capillary pressure.

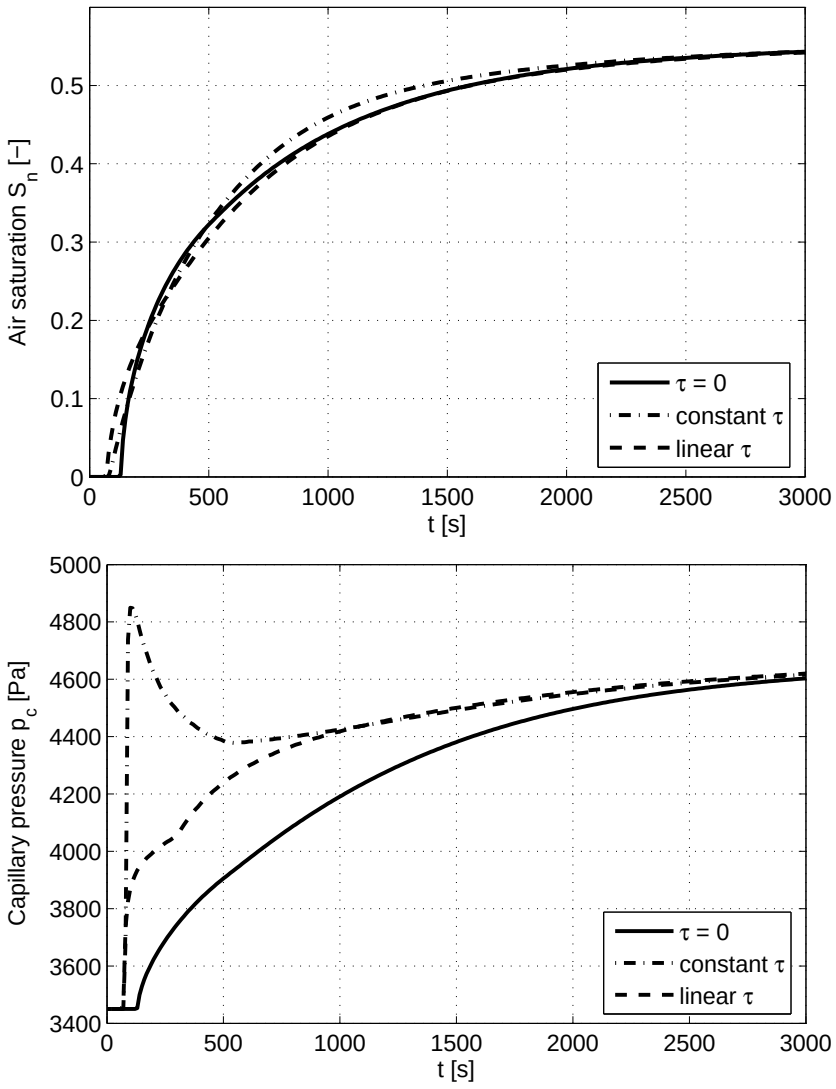


Figure 2: Numerical solutions of the column drainage simulation. Values of saturation and capillary pressure in the middle of the column are plotted versus time.

4 Conclusion and future work

This manuscript presents recently obtained numerical simulations using the non-classical dynamic capillary pressure in simulating two-phase incompressible flow in porous medium. Two main models for dynamic effect term $\tau = \tau(S_w)$ were used in order to determine their influence on a two-phase flow drainage problem.

As a result of the numerical simulation, the temporal profile of capillary pressure in the middle of the column is significantly different for the constant model of τ than for the linear model of τ or the static capillary pressure, but the differences between temporal profiles of air saturation were small.

These results indicate, that the dynamic effect may not be so important in drainage problems in homogeneous porous media. On the other hand, it may be of great importance in the highly heterogeneous media where the capillarity governs flow through material interfaces.

Acknowledgement

This work has been partly supported by:

- Project "Applied Mathematics in Technical and Physical Sciences" MSM 6840770010, Ministry of Education of the Czech Republic.
- Project "Mathematical Modelling of Multiphase Porous Media Flow" 201/08/P507 of the Czech Science Foundation (GA CR).

References

[1] P. Bastian. *Numerical Computation of Multiphase Flows in Porous Media*. Habilitation Dissertation, Kiel university (1999).

[2] J. Bear and A. Verruijt. *Modeling Groundwater Flow and Pollution*. D. Reidel, Holland, Dordrecht, (1990).

[3] A. Beliaev and S. Hassanizadeh. *A Theoretical Model of Hysteresis and Dynamic Effects in the Capillary Relation for Two-phase Flow in Porous Media*. Transport in Porous Media **43** (2001), 487–510.

[4] R. H. Brooks and A. T. Corey. *Hydraulic properties of porous media*. Hydrology Paper **3** (1964), 27.

[5] H. Dahle, M. Celia, and S. Majid Hassanizadeh. *Bundle-of-Tubes Model for Calculating Dynamic Effects in the Capillary-Pressure-Saturation Relationship*. Transport in Porous Media **58** (2005), 5–22.

[6] D. Das, S. Hassanizadeh, B. Rotter, and B. Ataie-Ashtiani. *A Numerical Study of Micro-Heterogeneity Effects on Upscaled Properties of Two-Phase Flow in Porous Media*. Transport in Porous Media **56** (2004), 329–350.

[7] R. Fučík. *Numerical Analysis of Multiphase Porous Media Flow in Groundwater Contamination Problems , Graduate Thesis.* FNSPE of Czech Technical University Prague, Prague, (2006).

[8] R. Fučík, M. Beneš, J. Mikyška, and T. H. Illangasekare. Generalization of the benchmark solution for the two-phase porous-media flow. In 'Finite Elements Models, MODFLOW, and More : Solving Groundwater problems', 181–184, (2004).

[9] R. Fučík, J. Mikyška, M. Beneš, and T. Illangasekare. *An Improved Semi-Analytical Solution for Verification of Numerical Models of Two-Phase Flow in Porous Media.* Vadose Zone Journal **6** (2007), 93–104.

[10] R. Fučík, J. Mikyška, M. Beneš, and T. Illangasekare. *Semianalytical Solution for Two-Phase Flow in Porous Media with a Discontinuity.* Vadose Zone Journal **7** (2008), 1001.

[11] R. Fučík, J. Mikyška, and T. H. Illangasekare. *Evaluation of saturation-dependent flux on two-phase flow using generalized semi-analytic solution.* Proceedings on the Czech Japanese Seminar in Applied Mathematics FNSPE CVUT Prague (2004), 25–37.

[12] W. Gray and S. Hassanizadeh. *Paradoxes and Realities in Unsaturated Flow Theory.* Water Resources Research **27** (1991), 1847–1854.

[13] W. Gray and S. Hassanizadeh. *Unsaturated Flow Theory Including Interfacial Phenomena.* Water Resources Research **27** (1991), 1855–1863.

[14] S. Hassanizadeh, M. Celia, and H. Dahle. *Dynamic Effect in the Capillary Pressure-Saturation Relationship and its Impacts on Unsaturated Flow.* Vadose Zone Journal **1** (2002), 38–57.

[15] S. Hassanizadeh and W. Gray. *Thermodynamic basis of capillary pressure in porous media.* Water Resources Research **29** (1993), 3389–3406.

[16] R. Helmig. *Multiphase Flow and Transport Processes in the Subsurface : A Contribution to the Modeling of Hydrosystems.* Springer Verlag, Berlin, (1997).

[17] S. Manthey. *Two-phase flow processes with dynamic effects in porous media - parameter estimation and simulation.* Institut fur Wasserbau der Universitat Stuttgart, Stutgart, (2006).

[18] D. McWhorter and D. Sunada. *Exact Integral Solutions for Two-Phase Flow.* Water Resources Research **26** (1990), 399–413.

[19] D. McWhorter and D. Sunada. *Reply to" Comment on'Exact integral solutions for two-phase flow'" by Z.-X. Chen, GS Bodvarsson, and PA Witherspoon.* Water Resources Research **28** (1992), 1479–1479.

[20] J. Mikyška. *Numerical Model for Simulation of Behaviour of Non-Aqueous Phase Liquids in Heterogeneous Porous Media Containing Sharp Texture Transitions, Ph.D. Thesis*. FNSPE of Czech Technical University, Prague, (2005).

[21] T. Sakaki, D. O'Carroll, and T. Illangasekare. *Direct laboratory quantification of dynamic coefficient of a field soil for drainage and wetting cycles*. American Geophysical Union, Fall Meeting 2007, abstract# H53F-1486 (2007).

[22] F. Stauffer. Time dependence of the relations between capillary pressure, water content and conductivity during drainage of porous media. In 'On scale effects in porous media, IAHR, Thessaloniki, Greece', (1978).

[23] M. T. van Genuchten. *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils.* Soil Science Society of America Journal **44** (1980), 892–898.

[24] G. S. B. Z.-X. Chen and P. A. Witherspoon. *Comment on 'exact integral solutions for two-phase flow'.* Water Resources Research **28** (1992), 1477–1478.

Semi-Regular Texture Modeling

Martin Hatka

3rd year of PGS, email: hadis@email.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Michal Haindl, Institute of Information Theory and Automation, AS CR

Abstract. This paper describes a method for seamless enlargement or editing of difficult colour textures containing both regular periodic and stochastic components. Such textures cannot be modeled using neither simple tiling nor using purely stochastic models. However these textures are often required for realistic appearance visualization. The principle of our near-regular texture synthesis and editing method is to automatically recognize and separate periodic and random components of the corresponding texture. The regular texture part is modeled using the roller method, while the random part is synthesized from its estimated exceptionally efficient Markov random field based representation. Both independently enlarged texture components from the original measured texture are combined in the resulting synthetic near-regular texture. In the editing application both enlarged texture components can be from different measurements. The presented texture synthesis method allows large texture compression and it is simultaneously extremely fast due to complete separation of the analytical step of the algorithm from the texture synthesis part. The method is universal and easily viable in a graphical hardware for purpose of real-time rendering of any type of near-regular static textures.

Abstrakt. Článek popisuje metodu pro zvětšování nebo editaci složitých barevných textur, které obsahují pravidelnou periodickou i stochastickou složku. Tyto textury nelze modelovat ani jednoduchým dlaždicováním, ani čistě stochastickými modely. Ovšem jsou často potřeba pro realistickou visualizaci. Princip popisované metody pro syntézu a editaci textur je založen na automatickém rozpoznání a oddělení periodické a náhodné složky textury. Pravidelná část je pak modelována metodou Roller, zatímco náhodná složka modelem využívajícím reprezentaci pomocí Markovských náhodných polí. Obě nezávisle zvětšené složky textury jsou pak zkombinovány ve výslednou syntetickou texturu. Při aplikacích editace textury lze zvětšené složky kombinovat z různých texturních měření. Prezentovaná metoda pro syntézu textur umožňuje vysokou kompresi a současně je extrémně rychlá, a to díky kompletní separaci analytické části algoritmu od části syntézy. Metoda je univerzální a je možné ji implementovat v grafickém hardware za účelem renderingu libovolných statických periodicko-stochastických textur v reálném čase.

1 Introduction

Physically correct virtual models require object surfaces covered with realistic nature-like colour textures to enhance realism in virtual scenes. To make virtual worlds realistic detailed scene models must be built. Satisfactory models require not only complex 3D shapes accorded with the captured scene, but also lifelike colour and texture. This will increase significantly the realism of the synthetic scene generated. Textures provide useful

cues to a subject navigating in such a VR environment, and they also aid in the accurate detailed reconstruction of the environment.

We define near-regular textures as textures containing global, possibly imperfect, regular structures as well as irregular stochastic structures simultaneously. This is more ambitious definition than to view [8] a near-regular textures as a statistical distortion of a regular texture. Our definition comprises types I and II from the near-regular texture categorization [8] while their type III is stochastic texture. Near regular textures are difficult to synthesize, however, these textures are ubiquitous in man-made environments such as buildings, wallpapers, floors, tiles, fabric but even some fully natural textures such as honeycomb, sand dunes or waves belong to this texture category. These textures can be either smooth or rough (also referred as the bidirectional texture function - BTF [3]). The rough textures which have rugged surfaces do not obey the Lambert law and their reflectance is illumination and view angle dependent. Both types of such near-regular textures occur in virtual scenes models.

The purpose of any synthetic texture is to reproduce a given digitized texture image so that ideally both natural and synthetic texture will be visually indiscernible. However modeling of an existing measured real texture is a very challenging and difficult task, due to unlimited variety of possible surfaces, illumination and viewing conditions simultaneously with the strong discriminative functionality of the human visual system. The related texture modeling approaches may be divided primarily into intelligent sampling and model-based-analysis and synthesis, but no ideal method for texture synthesis exists. Each of the existing approaches or texture models has its advantages and limitations.

Neither model-based or simple sampling algorithm alone can satisfactorily solve the difficult problem of near-regular texture modeling. Existing work [10, 7, 8, 12, 2, 9, 1, 11] usually tries to overcome this problem by user assisted modeling of the regular structures and then relies on regular tiling. However Lin et al. [2] experimentally observed that several of these general purpose sampling algorithms fail to preserve the structural regularity on more than 40% of their tested regular textures.

The presented fully automatic method proposes to combine advantages of both basic texture modeling approaches by factorizing a texture into factors that benefit best from each of these two basic different modeling concepts. The principle of the method is to separate regular and stochastic parts of the texture, to enlarge both parts separately and to combine these results (texture enlargement) or results from several different textures (texture editing) into the required resulting texture. The proposed solution is not only fully automatic, very fast due to strict separation of the analytical and very efficient synthesis steps, but it also allows significant data compression. Due to its stochastic modeling it completely eliminates visible repetitions (contrary to all mentioned tiling approaches) because there are never two identical tiles. Finally the method can be easily used to near-regular texture editing by either combining texture parts from different measurement or by changing stochastic model parameters.

The following section describes an automatic separation of the regular and stochastic texture parts. Section 3 is devoted to the regular part modeling using our simple sampling approach based on the repetition of a double toroidal tile carved from the original regular part texture measurement, while the subsequent section 4 defines our fast Markov random field model of the stochastic texture part. The overall algorithm results are reported in

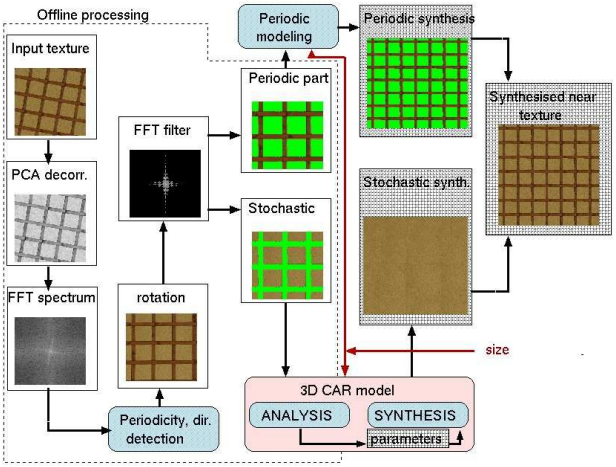


Figure 1: Presented method schema.

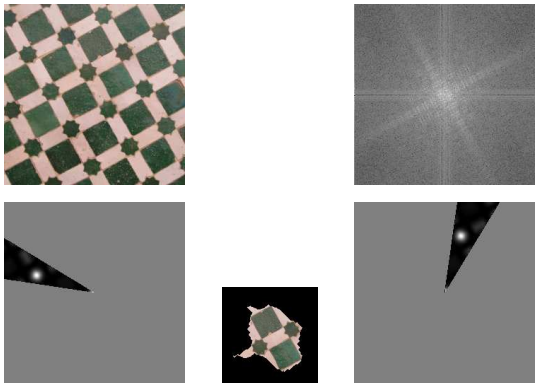


Figure 2: Original measured texture and its amplitude spectrum (upper row), detected spatial correlation sectors (bottom row) and the resulting toroidal tile.

the section 5, followed by conclusions in the last section.

2 Periodic and Non-Periodic Texture Separation

We can legitimately assume that the near-regular input textures have distinct amplitude spectrum parts for both periodic and random components. Otherwise the method would not be able to separate both texture parts. The overall schema of the method is illustrated in Fig.1 and detailed in the corresponding following sections. Periodic and non-periodic texture part are detected in the simplified monospectral texture space. The input colour texture is spectrally transformed using the principal component analysis (PCA). Let the digitized colour texture Y is indexed on a finite rectangular three-dimensional $M \times N \times d$ underlying lattice I, where $M \times N$ is the image size and d is the number of spectral bands (i.e., $d = 3$ for usual colour textures). PCA is performed on data vectors $Y_{r,\bullet}$, where the multiindex r has two components $r = [r_1, r_2]$, the first component is row and the second one column index, respectively, the notation $\bullet$ has the meaning of all possible values of the corresponding index. Then the periodic texture part is detected on the most informative transformed monospectral factor (first principal component), which corresponds to the largest eigenvalue.

2.1 Textural Periodicity Direction

Near-regular measured textures can have arbitrary periodicity directions (Fig.2), not necessarily simple axis aligned periodicity. The periodicity in two directions is detected from the spatial correlation field restricted with the help of Fourier amplitude spectrum (Fig.2-upper right). The method finds two largest Fourier amplitude spectrum coefficients provided that they do not represent parallel directions.

Tolerance sectors (Fig.2- bottom left, right), which accommodate for possible localization imprecision of local amplitude spectra maxima, are specified and for all their

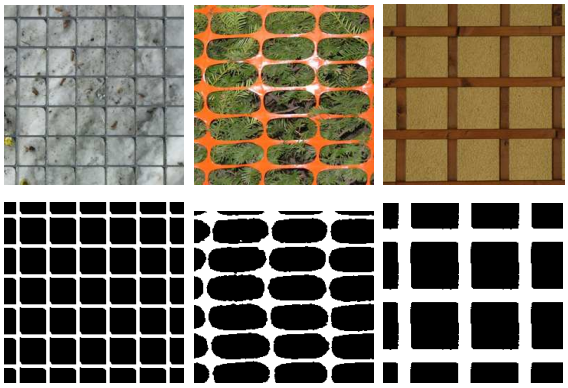


Figure 3: Near-regular measured textures and their detected periodic parts.

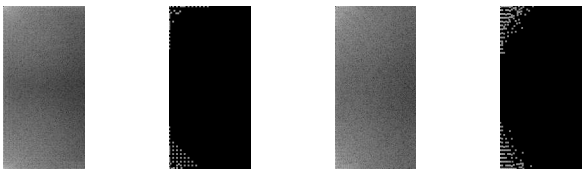


Figure 4: Near regular texture Fourier amplitude spectrum and its filtered version for the upper (two leftmost images) and the middle row textures in Fig.3.

indices the corresponding spatial correlations are evaluated. Local spatial correlation field maxima, larger than a threshold, are detected and the minimal periodicity maximum is selected. Detected periodicity $(\delta^{h^*}, \delta^{v^*})$ and its direction specify a rhomboid which contains the largest periodic part from the input texture. The rhomboid is further transformed to the rectangle to make periodic texture part detection as precisely as possible.

2.2 Amplitude Spectrum Filter

The texture cutout is re-sampled to the lattice size of power two required by the fast Fourier transformation based filter. Let $A_{\max}$ is the Fourier amplitude spectrum maximum coefficient detected from the Fourier amplitude spectrum (Fig.4-1.,3. leftmost images). The filter removes such coefficients, for which any of the following conditions (1), (2) holds (Fig.4-2.,4.).

$$A_r < k\, A_{\max} \ , \tag{1}$$

$$A_r \notin \mathcal{M} \ \wedge \ r \notin I_m \ , \tag{2}$$

where $\mathcal{M}$ is a set of amplitude spectrum local maxima, $k \in \langle 0; 1\rangle$ is a parameter and I_m is a contextual neighbourhood (we use the hierarchical neighbourhood of the first or the second order) of such a local maximum. Applying the inverse Fourier transformation and re-sampling the filtered tile back to the original size we get the filtered cutout.

2.3 Periodic Structure Separation

The filtered tile is binarized using a threshold $t_{bin} \in \langle 0; 1\rangle$. One label determines the periodic texture part and the remaining one the stochastic part. To find the labels correspondence to both periodical and non-periodical parts of the original texture Fig.3, the binary image is tested for periodicity δ^{h^*}, δ^{v^*}. The majority label complying to the periodicity test denotes the original texture periodic sites (Fig.3-right). When both periodic and stochastic parts are separated they can be independently modeled and enlarged

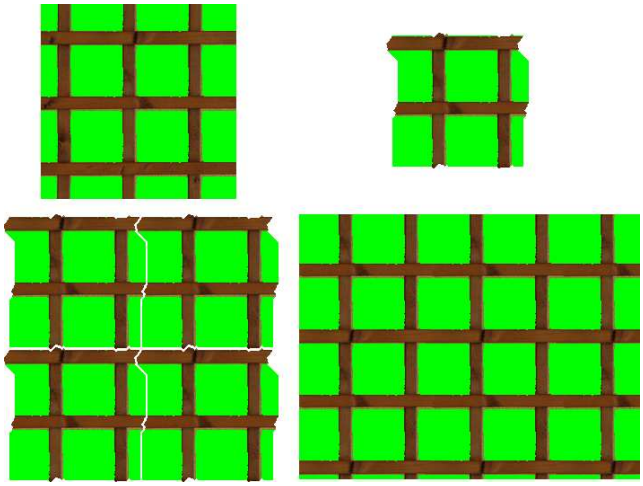


Figure 5: The double toroidal tile modeling principle - upper row input texture and toroidal tile, bottom row texture generation and the result, respectively.

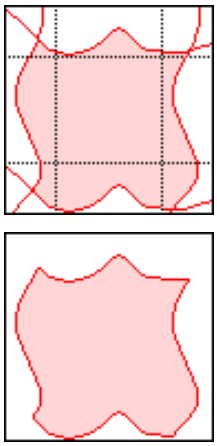


Figure 6: The optimal tile cuts in horizontal and vertical direction.

to any required size as it is detailed in two following sections. The required near-regular texture is simple composite of both synthetic parts.

3 Periodic Texture Modeling

The regular part of the texture is enlarged using a simplification of our previously published [5] method. The method selects double toroidal tiles as small as possible to compress the original measurements. The method starts with the minimal tile size detection which is limited by the size of texture measurements, the number of toroidal tiles we are looking for (n) and the sample spatial frequency content.

The optimal cuts for both the horizontal and vertical edge is searched using the dynamic programming method. Alternatively we can use some other suboptimal search such as the A^* algorithm if necessary to speed up also the analytical part of the method. The combination of both optimal vertical and horizontal cuts creates the toroidal tile as is demonstrated on the Fig.6.

Some textures with dominant irregular structures cannot be modeled by simple one tile repetition without clearly visible and visually disturbing regularly repeated effects. These textures are modeled by using multiple toroidal tiles which have the same border but differ in their interior.

Finally, the enhancement of any required periodic texture is simple repetition of either single double toroidal tile or randomly alternating repetition of several double toroidal tiles in both directions until the required texture is generated Fig.7.

4 Random Texture modeling

The random part of a texture is synthesized from the original input texture from where the detected periodic component was removed as described in section 2. If the stochastic

texture patches are too small (few hundred pixels area) to reliably learn the random field model statistics, we replace occluded stochastic texture areas by using a modification of the image quilting algorithm [4] (see Fig.8-left).

The random part of the texture is synthesized using an adaptive probabilistic spatial model, a multiresolution 3D causal autoregressive model (CAR) [6], which is an exceptionally efficient type from the Markov random field (MRF) family of models. The CAR model allows extreme compression (few tens of parameters to be stored only) and can be evaluated directly in procedural form to seamlessly fill an infinite texture space. An analyzed texture is decomposed into multiple resolutions factors using Laplacian pyramid and the intermediary Gaussian pyramid [6] which is a sequence of images in which each one is a low-pass down-sampled version of its predecessor. The Laplacian pyramid contains band-pass components and provides a good approximation to the Laplacian of the Gaussian kernel. It can be constructed by differentiate single Gaussian pyramid layers.

The CAR model synthesis is very simple and the CAR random field can be directly generated from the model equation using a multivariate Gaussian generator. The fine-resolution synthetic texture is obtained from the pyramid collapse procedure (Fig.8). The CAR model offers huge compression ration because only few parameters for each texture have to be stored or transmitted. The resulting near-regular texture is simple combination of both regular and stochastic synthesized factors.

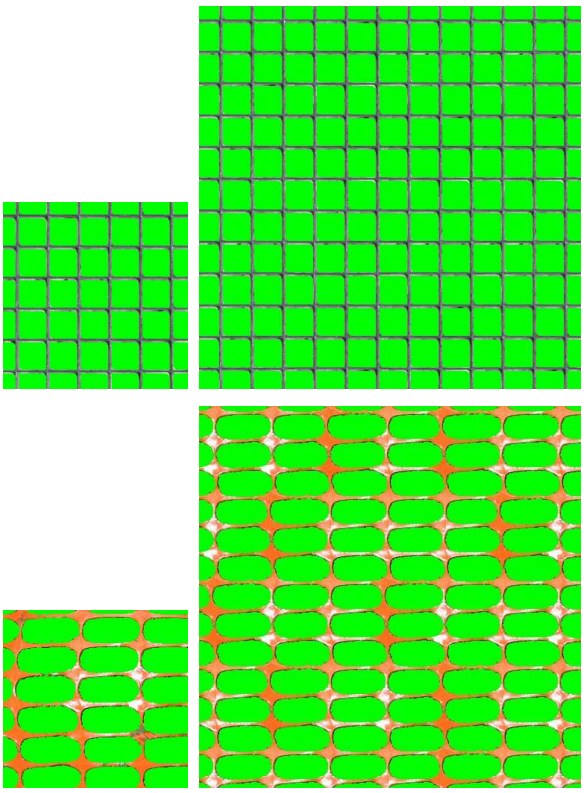


Figure 7: Periodic texture part synthesis (right).

Figure 8: Stochastic texture part synthesis (right).

5 Results

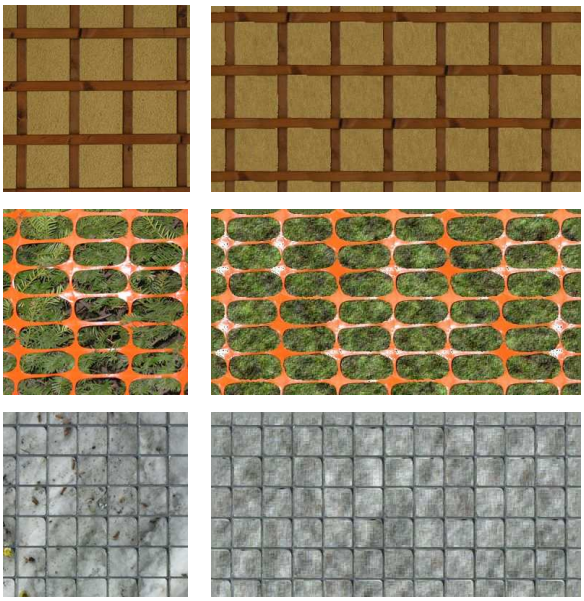


Figure 9: Near-regular textures and their synthesis (right).

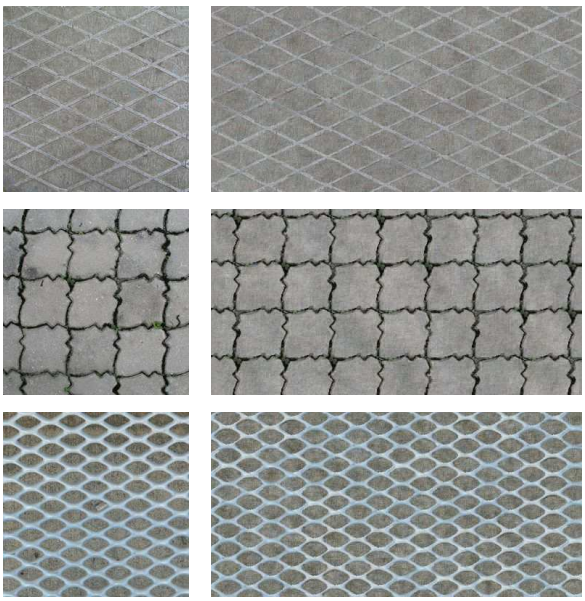


Figure 10: Near-regular textures and their synthesis (right).

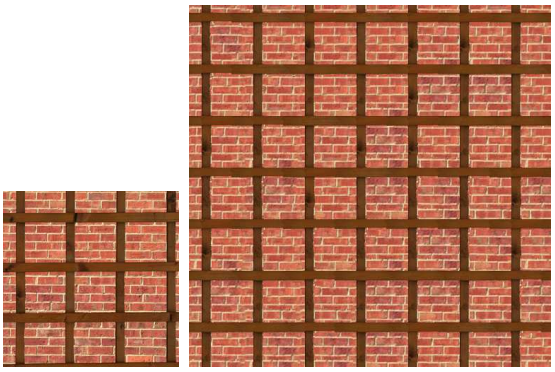


Figure 11: Near-regular texture with two types of regular structures (bricks and lattice - edited from two separate measurements) and its synthesis.

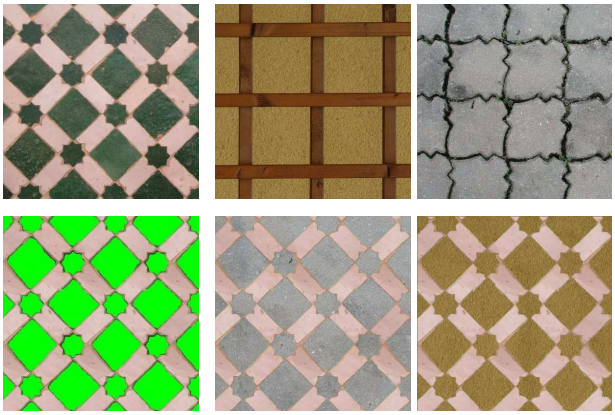


Figure 12: Near-regular texture editing. Input textures (upper row) and resulting lattice and edited textures.

We have tested the presented method on near-regular textures from our extensive texture database, which currently contains over 1000 colour textures. Tested near-regular textures were either man-made such as three textures on Fig.10 or combinations of man-made structures with natural background (Fig.9) such as grass, wood, plants, snow, sand, etc. Several of these results are demonstrated in the following images Figs.9,10. Both part of modeling were separately successfully tested on hundreds of colour or BTF textures with results reported elsewhere ([5]). Such unusually extensive testing was possible due

to simplicity and efficiency of both crucial parts of the algorithm and it allowed us to get insight into the algorithm properties. The method is even capable to synthesize some near-regular textures combined from two distinctive types of regular structures Figs.11,12 provided they can be adequately separated in the Fourier domain.

Resulting textures are mostly surprisingly good for such an automatic and fast algorithm. For example our results on the text texture ([5]) are indistinguishable (see [4]) from results on the same texture using much more complicated and slower image quilting algorithm [4]. Obviously there is no optimal texture modeling method and also the presented method fails on some textures. These are near-regular textures with similar amplitude spectrum parts for both periodic and random components, where our spectrum filter cannot separate both texture types without visible errors.

6 Conclusions

The test results of our method on available near-regular texture data are visually indiscernible from the measured textures for most of the tested colour textures. The test results of the method on our natural near-regular texture collection are encouraging. The presented method is extremely fast due to strict separation of the analytical and very efficient synthesis steps and fully automatic. The regular part modeling is easily implementable even in the graphical processing unit. The method offers larger compression ratio than alternative tiling methods for transmission or storing texture information due to the periodic part modeling approach. The MRF based random part model can reach itself a huge compression ratio, hence its storage requirements are negligible, and simultaneously eliminates visible repetitions typical for tiling approaches. The overall method is very fast - it has negligible computation complexity for the periodic model and exceptionally efficient computational model for the random part as well. The method's extension for alternative texture types, such as BTF textures or some other spatial data such as the reflectance models parametric spaces is straightforward. Finally the method can be easily used to near-regular texture editing by either combining texture parts from different measurement or by changing stochastic model parameters.

References

[1] W. chieh Lin, J. Hays, C. Wu, V. Kwatra, and Y. Liu. Quantitative evaluation of near regular texture synthesis algorithms. In 'In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', 427–434, (2006).

[2] W. chieh Lin, J. H. Hays, C. Wu, V. Kwatra, and Y. Liu. A comparison study of four texture synthesis algorithms on regular and near-regular textures. Technical report, (2004).

[3] K. J. Dana, S. K. Nayar, B. V. Ginneken, and J. J. Koenderink. Reflectance and texture of real-world surfaces authors. In 'CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)', 151, Washington, DC, USA, (1997). IEEE Computer Society.

[4] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In 'SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques', 341–346, New York, NY, USA, (2001). ACM Press.

[5] M. Haindl and M. Hatka. A roller - fast sampling-based texture synthesis algorithm. In 'The 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2005', V. Skala, (ed.), 80–83, Plzen, (February 2005). University of Western Bohemia.

[6] M. Haindl and V. Havlíček. A multiscale colour texture model. In 'Proceedings of the 16th International Conference on Pattern Recognition', R. Kasturi, D. Laurendeau, and C. Suen, (eds.), 255–258, Los Alamitos, (August 2002). IEEE Computer Society.

[7] Y. Liu, R. T. Collins, and Y. Tsin. *A computational model for periodic pattern perception based on frieze and wallpaper groups.* IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004), 354–371.

[8] Y. Liu, W.-C. Lin, and J. H. Hays. *Near regular texture analysis and manipulation.* ACM Transactions on Graphics (SIGGRAPH 2004) **23** (August 2004), 368 – 376.

[9] Y. Liu and Y. Tsin. *The promise and perils of near-regular texture.* International Journal of Computer Vision **62** (2005), 1–2.

[10] J. m. Dischler, K. Maritaud, B. Lévy, and D. Ghazanfarpour. *Texture particles.* Computer Graphics Forum **21** (2002), 401–410.

[11] A. Nicoll, J. Meseth, G. Müller, and R. Klein. *Fractional fourier texture masks: Guiding near-regular texture synthesis.* Computer Graphics Forum **24** (September 2005), 569–579.

[12] Y. Tsin, Y. Liu, and V. Ramesh. Texture replacement in real images. In 'In IEEE Computer Vision and Pattern Recognition Conference', 539–544. IEEE Computer Society Press, (2001).

Application of Marginalized Particle Filter to Linear–Gaussian Problems with Unknown Model Error Covariance Structure*

Radek Hofman

2nd year of PGS, email: hofman@utia.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Petr Pecha, Institute of Information Theory and Automation, AS CR

Abstract. The paper presents a scheme for estimation of spatio–temporal evolution of a quantity with unknown model error. Model error is estimated on basis of measured–minus–observed residuals evaluated upon measured and modeled values. Methods of Bayesian filtering are applied to the problem. The main contribution of this paper is application of general marginalized particle filter algorithm to the linear–Gaussian problem with unknown model error covariance structure. Methodology is demonstrated on the problem of modeling of spatio–temporal evolution of groundshine–dose from radionuclides deposited on terrain in long–time horizon.

Abstrakt. Příspěvek se zabývá asimilací časového vývoje prostorově rozložené veličiny s měřeními. Pokud je řešený problém chápán jako lineární s gaussovským rozdělením šumu, může být za předpokladu znalosti kovarianční struktury chyb modelu a měření řešen Kalmanovým filtrem. Pokud kovarianční strukturu chyb modelu neznáme, musí být nejprve odhadnuta. V příspěvku je popsána metodika aplikace marginalizovaného particle filtru na lineárně–Gaussovské problémy s neznámou kovarianční strukturou, která je odhadována pomocí sekvenčních M–C metod. Metodika je prezentována na odhadu vývoje dávky z depozice radionuklidů na terénu.

1 Introduction

The task of estimation of time evolution of a spatially distributed quantity is widely applied in different branches of "Earth sciences" such as meteorology and oceanography [12]. During the last years, there have arisen tendencies for application of an advanced data assimilation algorithms also in the field of radiation protection [16], [19], [20]. It is related to the renaissance of nuclear energy which can be observed. The application of advanced statistical methods can increase reliability of consequence predictions of possible releases from nuclear power–plants. Their reliability is in the field of radiation protection mission–critical as the problem deals with the population health.

There were developeded several models for modeling of evolution of living environment contamination for different release scenarios. The only connection with physical reality are measurements with errors (sparse both in time and in space). In our work, we attempt to make groundshine–dose model predictions more reliable in a way of adjusting them towards measurements incoming from terrain. This process is called data assimilation

*This work has been supported by the grant project GAČR No. 102/07/1596, which is funded by the Grant Agency of the Czech Republic.

[12]. Its principle consists in combining of the information provided by the model with the measured data. Exploiting information on sources of uncertainty, we are able to produce improved estimate of the true situation on terrain.

If the problem is treated as linear–Gaussian, it can be successfully solved via Kalman filter (KF) [11]. The unavoidable condition for utilization of Kalman filter is knowledge of model error covariance structure but in many cases it is unknown due to the problem background. In this paper is presented methodology for application of the Kalman filter to the problems where the model error covariance structure is unknown and has to be estimated upon actual data before application of the filter. This results in marginalized particle filter described in [22].

Model error covariance is represented by a covariance matrix. As the total number of its elements is much higher the number of measurements, we can't estimate all of them. Simplified model error covariance structure parametrization based on idealized assumptions is introduced. For finding the most plausible values of these parameters, the similar approach as proposed in [3] or [15] based on modeled–minus–observed residuals is used. Instead of maximum likelihood estimates, we use marginalized particle filter for estimation of both the model error covariance parameters and groundshine–dose distribution. The marginalized particle filter is a powerful combination of the particle filter and the Kalman filter, which can be used when the underlying model contains a linear substructure which is being subject to Gaussian noise.

The performance of this methodology is demonstrated on modeling of groundshine–dose evolution in long–time horizon of several months [6]. As the problem is complex, the groudshine–dose evolution model is an idealized approximation of the true physical process. Calculations are performed on a subset of polar network around the source of pollution. The model error covariance parametrization proposed here follows the physical background of the problem.

The outline of this paper is as follows. Section 2 briefly discusses Bayesian filtering. Kalman filter, particle filter and marginalized particle filter are successively presented there. In Section 3, the assimilation algorithm is proposed and the problem of model error covariance estimation is described. Application of the algorithm on modeling of long–term evolution of groundshine–dose from radionuclide deposition on terrain is presented in Section 4. Specific model error covariance parametrization suitable for the problem is developed there. In Section 5, experimental results with simulated measurements are demonstrated and the conclusion is given.

2 Bayesian filtering

Bayesian approach to filtering is applicable to all linear or nonlinear stochastic systems [7], [13]. Let the stochastic system be defined by discrete–time state–space transition equation (1) and observation equation (2)

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) + \boldsymbol{b}_t \tag{1}$$

$$\boldsymbol{y}_t = h(\boldsymbol{x}_t) + \boldsymbol{\epsilon}_t \tag{2}$$

Here, t is time index, $\boldsymbol{x}_t$ is unobservable system state vector, $\boldsymbol{b}_t$ is the additive dynamic noise vector. Vector $\boldsymbol{y}_t$ is the measurement vector which provides us indirect information

about the system state and $\boldsymbol{\epsilon}_t$ its noise. Both the densities of noise terms are assumed to be independent and known. Functions $f(\cdot)$ and $h(\cdot)$ are generally non–linear. State transition function $f(\cdot)$ propagates the prior state to the current one. Forward observation operator $h(\cdot)$ transforms vectors from state–space to the measurement space, thus constitutes relation of the actual measurements to the current state.

The goal is to acquire posterior density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ where $\boldsymbol{Y}_t = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t\}$ are available measurements. In the following text, the state process $\{\boldsymbol{x}_t\}$ is assumed to be Markovian of the first order. It means that given the present state, future states are independent of the past states:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots, \boldsymbol{x}_0) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{3}$$

Realization of the process at time t contains all information about the past, which is necessary in order to calculate the future behavior of the process.

Bayesian estimation procedure consists of two recursively repeated step. The first step transits the state estimate to the next time step according to the probability density function (PDF) $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. This step is called time update (4). In the second step called data update (5), the information provided by actual measurements $\boldsymbol{y}_t$ is included into the current estimate given by the PDF $p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1})$.

$$p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1}) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) p(\boldsymbol{x}_{t-1}|\boldsymbol{Y}_{t-1}) d\boldsymbol{x}_{t-1} \tag{4}$$

$$p(\boldsymbol{x}_t|\boldsymbol{Y}_t) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1})}{\int p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_{t-1}) p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1}) d\boldsymbol{x}_t} \tag{5}$$

The state evolution is initialized by a probability density function $p(\boldsymbol{x}_0|\boldsymbol{Y}_{-1}) = p(\boldsymbol{x}_0)$ which represents all the prior information on the problem and also our subjective judgments. This density is often called background–field or just the prior.

If both the measurement density $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ and the state transition density $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ are parametric, the problem can be solved analytically. Provided that the system is linear–Gaussian, the integrals (4, 5) lead to KF recursion.

2.1 Kalman filter

In the following text $N(\boldsymbol{\mu}, \boldsymbol{Q})$ is assumed to be a Gaussian PDF with mean value $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{Q}$. KF is simple implementation of the Bayesian filter and it provides the optimal Bayesian solution. Its usage is limited to the case of linear estimation with the Gaussian noise where

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = N(\boldsymbol{M}\boldsymbol{x}_{t-1}, \boldsymbol{Q}_t) \tag{6}$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = N(\boldsymbol{H}\boldsymbol{x}_t, \boldsymbol{R}_t) \tag{7}$$

Matrices $\boldsymbol{M}$ and $\boldsymbol{H}$ are matrices of linear (linearized) operators $f(\cdot)$ and $h(\cdot)$, respectively. Matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ are known error covariance matrices of model error and measurement error, respectively. Under these assumptions (4, 5) lead to KF equations for time update and data update steps [11]. The equations perform recursive update of the first two moments of estimated Gaussian distribution $p(\boldsymbol{x}|\boldsymbol{Y}) = N(\hat{\boldsymbol{x}}, \boldsymbol{P})$ – the mean value $\hat{\boldsymbol{x}}$ and its covariance matrix $\boldsymbol{P}$.

2.2 Particle filter

In more general cases where analytical solution of integrals (4, 5) is not known, there are their numerical approximations based on sequential Monte Carlo methods also known as particle filters.

Particle filter (PF) is more general implementation of Bayesian filter which can be used to approximate the posterior density function for the state in non–linear and non–Gaussian filtering problems [7]. It is based on recursive estimation of the PDF $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ which is represented as a set of M so called particles $\boldsymbol{x}_t^{(i)}$ and its associated normalized weights $\tilde{q}_t^{(i)}$ as $\{\tilde{q}_t^{(i)}, \boldsymbol{x}_t^{(i)}\}|_{i=1\ldots M}$. The posterior PDF $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ can be approximated with their help as $\hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t)$.

$$p(\boldsymbol{x}_t|\boldsymbol{Y}_t) \approx \hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t) = \sum_{i=1}^{M} \frac{1}{M}\delta(\boldsymbol{x}_t - \boldsymbol{x}_t^{(i)}), \tag{8}$$

where $\delta(\cdot)$ is the Dirac δ-function and $\boldsymbol{x}_t^{(i)}$ are samples from approximated PDF. In (8), all the weights $\tilde{q}_t^{(i)}$ are equal to $\frac{1}{M}$. Our goal is usually to estimate the mean value of a function defined on our approximated distribution $E_{\hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t)}[g(\boldsymbol{x}_t)]$. The approximation $\hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ satisfies condition

$$\lim_{M\to+\infty} E_{\hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t)}[g(\boldsymbol{x}_t)] \overset{a.s.}{\to} E_{p(\boldsymbol{x}_t|\boldsymbol{Y}_t)}[g(\boldsymbol{x}_t)], \tag{9}$$

where $\overset{a.s.}{\to}$ means almost sure convergence and $g(\boldsymbol{x}_t)$ is arbitrary function bounded for support $\Omega = \{\boldsymbol{x}_t|p(\boldsymbol{x}_t|\boldsymbol{Y}_t) > 0\}$.

In real cases we are not able to sample directly from $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ but we are able to evaluate it in discrete points (at least up to proportionality). We can draw independent samples $\boldsymbol{x}_t^{(i)}$ from a chosen known proposal distribution (importance function) $q(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ and use them for approximating of $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$. The estimated density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$, its approximation $\hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ and importance function $q(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ are related as follows

$$\begin{aligned} p(\boldsymbol{x}_t|\boldsymbol{Y}_t) &= \frac{p(\boldsymbol{x}_t|\boldsymbol{Y}_t)}{q(\boldsymbol{x}_t|\boldsymbol{Y}_t)} q(\boldsymbol{x}_t|\boldsymbol{Y}_t) \approx \\ &\approx \hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t) = \sum_{i=1}^{M} \frac{p(\boldsymbol{x}_t^{(i)}|\boldsymbol{Y}_t)}{q(\boldsymbol{x}_t^{(i)}|\boldsymbol{Y}_t)} \frac{1}{M}\delta(\boldsymbol{x}_t - \boldsymbol{x}_t^{(i)}), \end{aligned} \tag{10}$$

where $\frac{1}{M}\sum_{i=1}^{M}\delta(\boldsymbol{x}_t - \boldsymbol{x}_t^{(i)})$ is an approximation of $q(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ since $\boldsymbol{x}_t^{(i)}$ are sampled from this PDF. If we denote $q_t^{(i)} = \frac{p(\boldsymbol{x}_t^{(i)}|\boldsymbol{Y}_t)}{q(\boldsymbol{x}_t^{(i)}|\boldsymbol{Y}_t)}\frac{1}{M}$, the estimated PDF can be approximated as

$$\hat{p}(\boldsymbol{x}_t|\boldsymbol{Y}_t) = \sum_{i=1}^{M} \tilde{q}_t^{(i)}\delta(\boldsymbol{x}_t - \boldsymbol{x}_t^{(i)}), \tag{11}$$

where $\tilde{q}_t^{(i)} = q_t^{(i)}/\sum_{j=1}^{M} q_t^{(j)}$, $\sum_i^M \tilde{q}_t^{(i)} = 1$, $\tilde{q}_t^{(i)} \geq 0$ are normalized weights. This normalization will for finite M introduce a bias in the estimate. However, from the strong

law of large numbers the estimate is asymptotically unbiased. This algorithm is called sampling–importance–sampling (SIS).

If we choose the posterior PDF from the previous step as proposal distribution in the current, we can via recursive evaluation of normalized weights perform Bayesian filtering. In this case will weight update result in

$$q_t^{(i)} \propto \tilde{q}_{t-1}^{(i)} p(\boldsymbol{y}_t|\boldsymbol{x}_t^{(i)}) \tag{12}$$

This algorithm also suffers from degeneracy problem, so we have to implement a resampling algorithm, more in [4]. Resampling should eliminate particles with small weights and multiply particles with large weights. After resampling all the weights are set to $\frac{1}{M}$. If we perform resampling in each step, the weights can be computed as $q_t^{(i)} = p(\boldsymbol{y}_t|\boldsymbol{x}_t^{(i)})$. This modification of SIS algorithm with resampling in each step is also known as sampling–importance–resampling (SIR).

Disadvantage of this method is that we have to be able to generate random samples from complicated distributions and this is for high dimensional problems computationally prohibitive. The computational complexity rapidly increases along with the state–space dimension.

2.3 Marginalized particle filter

When structure of the model (1, 2) allows analytical marginalization over a subset of states, we can reduce the computational burden. Let's consider factorization of the state vector $\boldsymbol{x}_t = \begin{bmatrix}\boldsymbol{x}_t^l & \boldsymbol{x}_t^n\end{bmatrix}^T$ where $\boldsymbol{x}_t^l$ is a subset of analytically tractable states and $\boldsymbol{x}_t^n$ is the rest. Provided that the $\boldsymbol{x}_t^l$ and $\boldsymbol{x}_t^n$ are conditionally independent, substitution of the factorization into (8) and application of the chain rule gives

$$p(\boldsymbol{x}_t^l, \boldsymbol{x}_t^n|\boldsymbol{Y}_t) = p(\boldsymbol{x}_t^l|\boldsymbol{x}_t^n, \boldsymbol{Y}_t) p(\boldsymbol{x}_t^n|\boldsymbol{Y}_t), \tag{13}$$

where $p(\boldsymbol{x}_t^l|\boldsymbol{x}_t^n, \boldsymbol{Y}_t)$ is analytically tractable and $\boldsymbol{x}_t^n$ is given by the particle filter. Assuming that $\boldsymbol{x}_0^l \sim N(\hat{\boldsymbol{x}}_0, \boldsymbol{P}_0)$ and to be governed by a linear model implies that $p(\boldsymbol{x}_t^l|\boldsymbol{x}_t^n, Y_t)$ is conditionally linear–Gaussian and can be computed via Kalman filter [23]. Substitution of (8) into (13) for $\boldsymbol{x}_t^n$ leads to

$$p(\boldsymbol{x}_t|\boldsymbol{Y}_t) \approx \sum_{i=1}^{M} \tilde{q}_t^{(i)} \delta(\boldsymbol{x}_t^n - \boldsymbol{x}_t^{n,(i)}) N(\hat{\boldsymbol{x}}_t^{l,(i)}, \boldsymbol{P}_t^{(i)}) \tag{14}$$

The joint PDF is estimated as a mixture of a parametric distribution of Gaussian type and of a nonparametric one. The estimated PDF is represented by a weighted sum of Gaussians, where each particle has a Gaussian distribution attached to it. This modification of PF is called marginalized particle filter (MPF) and details on its implementation can be found in [22], [23].

3 Assimilation procedure based on MPF

The unavoidable condition for application of Kalman filter is knowledge of model error represented in (1) by the noise vetor $\boldsymbol{b}_t$. We assume $\{\boldsymbol{b}_t\}$ to be the white noise process

where $\boldsymbol{b}_t \sim N(\boldsymbol{0}, \boldsymbol{Q}_t)$. Matrix $\boldsymbol{Q}_t$ is corresponding covariance matrix. The value of $\boldsymbol{Q}$ should reflect total (unknown) model error, which is in each step contribution to the forecast error due to differences between the model and the true process. In KF [11], forecast error covariance matrix $\boldsymbol{P}$ evolves as

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{M}_{t|t-1}\boldsymbol{P}_{t-1|t-1}\boldsymbol{M}^T_{t|t-1} + \boldsymbol{Q}_t, \tag{15}$$

where $\boldsymbol{M}$ is matrix of linear (linearized) operator for the state transition from time $t-1$ to t. It is obvious that if $\boldsymbol{Q}$ is neglected, the predicted forecast error will be underestimated. This could cause divergence from the true state (its good estimate) because smaller model error will handicap the information provided by measurements.

We assume that the $\boldsymbol{Q}$ is unknown and attempt to estimate it in each assimilation step. As the total number of elements of $\boldsymbol{Q}$ to be estimated is much higher than the number of measurements, we can't estimate all of them. Simplified covariance model based on idealized assumptions has to be introduced.

Schematically, let the model error covariance matrix be approximated as a function $Q(\boldsymbol{\theta}) : \Re^{dim(\boldsymbol{\theta})} \rightarrow \Re^{[dim(\boldsymbol{x}),dim(\boldsymbol{x})]}$ of a parameter vector $\boldsymbol{\theta}$, where $\Re^{[m,n]}$ is a space of real matrices of dimension $m \times n$.

$$\boldsymbol{Q}_t = Q_t(\boldsymbol{\theta}_t) \tag{16}$$

Function Q has to be chosen properly in order to produce positive semi–definite symmetric matrices which can be covariance matrices.

For finding the most plausible values of $\boldsymbol{\theta}$ a similar approach as proposed in [3], [15] based on modeled–minus–observed residuals is used. Instead of maximum likelihood estimates proposed there we use MPF introduced in Section 2. When the measurements are available, we can evaluate residual vector $\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{H}\hat{\boldsymbol{x}}_t$ having the same dimension as the measurement vector. Covariance of $\boldsymbol{v}$ derived in [3] has the form

$$E[\boldsymbol{v}_t\boldsymbol{v}_t^T] = \boldsymbol{H}_t\boldsymbol{P}_{t|t-1}\boldsymbol{H}_t^T + \boldsymbol{R}_t = \boldsymbol{S}_t \tag{17}$$

We assume $\boldsymbol{v_t} \sim N(\boldsymbol{0}, \boldsymbol{S}_t)$. If we substitute (15) into (17) for $\boldsymbol{P}_{t|t-1}$ and use covariance parametrization (16) of $\boldsymbol{Q}_t$ we obtain

$$\boldsymbol{S}_t(\boldsymbol{\theta}) = \boldsymbol{H}_t[\boldsymbol{M}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{M}_t^T + Q_t(\boldsymbol{\theta})]\boldsymbol{H}_t + \boldsymbol{R}_t \tag{18}$$

From (15) can be seen that the parametrization of model error covariance leads to parametrization of forecast error covariance $\boldsymbol{P}$. The most plausible value of parameters are found in each time step via PF from likelihood $p(\boldsymbol{v}_t^{(i)}|\boldsymbol{\theta}_t^{(i)}) = N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{\theta}_t^{(i)}))$ for random parameter vectors $\boldsymbol{\theta}_t^{(1)}, \ldots, \boldsymbol{\theta}_t^{(M)}$ and corresponding residual vectors $\boldsymbol{v}_t^{(i)}$. The likelihood is the higher, the higher is the probability that difference between modeled and measured values is zero given covariance (18). These parameters are then used in (15, 16) for forecast error propagation. Incorporation of this algorithm into KF assimilation scheme results in MPF for estimation of joint PDF $p(\boldsymbol{x}_t, \boldsymbol{\theta}_t|\boldsymbol{Y}_t)$ which is the mixture of Gaussian and nonparametric distributions

$$\underbrace{p(\boldsymbol{x}_t, \boldsymbol{\theta}_t|\boldsymbol{Y}_t)}_{MPF} = \underbrace{p(\boldsymbol{x}_t|\boldsymbol{\theta}_t, \boldsymbol{Y}_t)}_{KF}\underbrace{p(\boldsymbol{\theta}_t|\boldsymbol{Y}_t)}_{PF}, \tag{19}$$

where $\boldsymbol{x}_t$ is the state vector and $\boldsymbol{\theta}_t$ is the vector of parameters used for estimation of current model error covariance structure.

4 Assimilation scenario

The algorithm described in Section 3 is demonstrated on assimilation scenario introduced in this section.

In case of an accidental aerial release of radioactive pollutants into the living environment, the radioactive plume is depleted during passing over the terrain. This phase is called the plume phase. Due to the deposition processes the plume leaves a radioactive trace on the ground.

After the plume phase (when the radioactive cloud exits the area of interest) post–emergency phase follows. It covers latter stages of accident consequence evolution. Post–emergency phase may extend over a prolonged period of several weeks or many years depending on the source of radiation and local conditions. It ends when environmental radiation levels resume to normal. The main exposure pathways in this phase are groundshine and ingestion. The deposited material cause irradiation and through the root system migrates to the edible parts of crops consumed by people and livestock. Among many radionuclides released during emergency situations we focus only on ^{137}Cs since it is one of the most important nuclides in long–time perspective. Its half–time of decay is long (30 years) and also analysis after the Chernobyl accident had shown that it is one of the most significant nuclides in these types of accidents having detrimental long–term effect on population health.

Our assimilation scenario covers the post–emergency phase. The source of pollution is placed into the centre of polar network. We perform our calculations on subset of this network in successive time steps $t \in \{0, 1, \ldots, t_{MAX}\}$. Groundshine–dose in ordered set of analyzed spatial points forms our state vector $\boldsymbol{x}$. We assume $\boldsymbol{x} \sim N(\hat{\boldsymbol{x}}, \boldsymbol{P})$. Let $\hat{\boldsymbol{x}}_0$ be an initial estimate of groundshine–dose and $\boldsymbol{P}_0$ its corresponding error covariance matrix. This background–field is given by probabilistic version of Atmospheric Dispersion Model (ADM) and constitutes the prior characterization of the problem. It is based on segmented Gaussian plume model and it is part of the HARP system, more in [16]. We assume sparse measurements $\boldsymbol{y}_t$ of actual gamma dose–rate to be available in each time step. These measurements are assumed to be conditionally independent with known error. Assimilation procedure consists of two iteratively repeated steps: In time update step (4) current state estimate together with its error covariance matrix are propagated forward in time. The model error is estimated and accounted for. Following data update step (5) produces so called analysis – adjusts the model prediction to be in accordance with actual measurements. Along with this two Kalman filter steps is in each time step estimated model error covariance structure.

4.1 Model error covariance parametrization

The idealized model of $\boldsymbol{Q}$ chosen for this example has three parameters $\boldsymbol{\theta} = (\alpha, \beta, L)|_{\ \alpha,\beta,L\geq 0}$

$$\boldsymbol{Q}_t = \alpha_t \left[\boldsymbol{Q}_t^{(1)} + \beta_t \boldsymbol{Q}_t^{(2)}(L_t) \right] \tag{20}$$

The model error is formally partitioned into two components representing different sources of uncertainty. The partitioning has physical background. Matrix $\boldsymbol{Q}^{(1)}$ concerns the uncertainty of forecast model parameters introduced in [10]. This component is found as

a covariance of sample obtained via Monte–Carlo simulation with many different settings of model parameters. Component $\boldsymbol{Q}^{(2)}$, scaled with β, is structured, homogeneous and isotropic error. All other sources of uncertainty are accommodated by introduction of $\boldsymbol{Q}^{(2)}$. This component is generated by means of second order autoregressive function $\rho_L(r)$ of length–scale parameter L and Euclidean distance between two spatial locations r [5].

$$\rho_L(r) = \left(1 + \frac{r}{L}\right) \exp\left(-\frac{r}{L}\right) \tag{21}$$

The value of length–scale parameter L controls how fast the correlation between two points decreases with their growing distance. The overall covariance is scaled with α. This parametrization allows for mutual scaling of unstructured noise component $\boldsymbol{Q}^{(1)}$ given upon numerical simulation and "additional" structured noise given by $\boldsymbol{Q}^{(2)}$. MPF algorithm according to [21] modified for this case is listed in the box ALGORITHM.

In Step 1), the particles are initialized with a prior distribution. In Step 2) are evaluated residuals upon measured and modeled values for purpose of normalized weights evaluation for different covariance parameter vectors $\boldsymbol{\theta}_t^{(i)}$. For each particle, the overall covariance given by (20) has to be evaluated. During Step 3) are particles resampled – those with small weights are replaced with particles "better" in terms of likelihood. Sometimes is also in this step introduced an artificial noise to prevent particle degeneracy problem – to maintain high diversity of particles. In Step 4) is performed data and time update of KF and time update of PF. If we omit Steps 4a) and 4c) we get the standard PF. In Step 4b) is set new importance function for the next time step.

5 Experimental Results and Conclusion

For experimental demonstration of the algorithm, an artificial scenario with local rain during the fifth hour of the plume phase was chosen. The rain increases depletion of the plume due the wet deposition. The area of interest is subset of polar network comprising of $N = 91$ analyzed points.

The measurements were simulated from the measurement equation (2) via linear forward observation operator $\boldsymbol{H}$ where the true initial deposition $\boldsymbol{x}_0$ was assumed to be two times higher than the prior estimate $\hat{\boldsymbol{x}}_0$ obtained from ADM. The background–field (initial distribution in time $t = 0$) was $N(\hat{\boldsymbol{x}}_0, \boldsymbol{P}_0)$ where forecast error covariance $\boldsymbol{P}_0$ was calculated according to

$$\boldsymbol{P}_0 = 2\bar{\boldsymbol{P}}_0 \circ \boldsymbol{\Omega}, \tag{22}$$

where $\boldsymbol{\Omega}$ is covariance matrix generated from (21) and the $\circ$ stands for element–wise matrix product (Schur product) [15]. This was done because the background–field error covariance matrix $\bar{\boldsymbol{P}}_0$ was modeled as sample covariance from multiple calls of ADM where the rain intensity was treated as a random variable. This accommodated the uncertainty in rain intensity into $\bar{\boldsymbol{P}}_0$ and provided us a valuable physical knowledge but this process also introduced strong covariances between states. In (22), these covariances were reduced, so the background–field became more conservative.

Initialization of particles in the very first step was following: $\alpha_1 \sim Gamma(1, 1)$, $\alpha_2 \sim N(10^2, 10^4)$ and $L \sim N(10^3, 10^6)$. The prediction was evaluated for the first eighth months of the post–emergency phase. Measurements were assumed to be available each

month. At each time step were simulated 10 irregularly spaced measurements. For clarity, all the measurements in this example are during computation located in the same positions, so the observation operator $\boldsymbol{H}_t = \boldsymbol{H}$ is constant.

ALGORITHM

1. Initialization:
   (a) *For* $i = 1, \ldots, M$ initialize $\boldsymbol{\theta}_0^{(i)} \sim p(\boldsymbol{\theta}_0)$
   (b) Set $\{\boldsymbol{x}_{0|-1}^{(i)}, \boldsymbol{P}_{0|-1}^{(i)}\} = \{\hat{\boldsymbol{x}}_0, \boldsymbol{P}_0\}$
2. Normalized weights evaluation:
   For $i = 1, \ldots, M$ evaluate:
   (a) Residuals $\boldsymbol{v}_t^{(i)} = \boldsymbol{y}_t - \boldsymbol{H}_t \hat{\boldsymbol{x}}_t^{(i)}$
   (b) Model error covariance matrix parametrization:
   $$\boldsymbol{Q}_t^{(i)} = Q\left(\boldsymbol{\theta}_t^{(i)} = \{\alpha_t^{(i)}, \beta_t^{(i)}, L_t^{(i)}\}\right)$$
      i. Evaluation of $\boldsymbol{Q}_t^{(i),(1)}$ via M-C simulation with multiple groudnshine model parameters setting
      ii. Evaluation of $\boldsymbol{Q}_t^{(i),(2)}(L_t^{(i)})$ via (21)
      iii. Evaluation of overall covariance via (20)
      $$\boldsymbol{Q}_t^{(i)} = \alpha_t^{(i)}\left[\boldsymbol{Q}_t^{(i),(1)} + \beta_t^{(i)}\boldsymbol{Q}_t^{(i),(2)}(L_t^{(i)})\right]$$
   (c) Residual covariance matrix $\boldsymbol{S}(\boldsymbol{\theta}_t^{(i)})$ via (18)
   (d) Importance weights $q_t^{(i)} = N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{\theta}_t^{(i)}))$
   (e) Normalize weights $\tilde{q}_t^{(i)} = \frac{q_t^{(i)}}{\sum_{j=1}^{M} q_t^{(j)}}$
3. PF measurement update – resampling:
   Resample M particles with replacement
   $$Pr(\boldsymbol{\theta}_{t|t}^{(i)} = \boldsymbol{\theta}_{t|t-1}^{(j)}) = \tilde{q}_t^{(j)}$$
4. KF data/time update and PF time update
   (a) KF data update:
   $$\hat{\boldsymbol{x}}_{t|t}^{(i)} = \hat{\boldsymbol{x}}_{t|t-1}^{(i)} + \boldsymbol{K}_t^{(i)}[\boldsymbol{y}_t - \boldsymbol{H}_t \hat{\boldsymbol{x}}_{t|t-1}^{(i)}]$$
   $$\boldsymbol{K}_t^{(i)} = \boldsymbol{P}_{t|t-1}^{(i)}\boldsymbol{H}^T(\boldsymbol{H}_t\boldsymbol{P}_{t|t-1}^{(i)}\boldsymbol{H}_t^T + \boldsymbol{R}_t)^{-1}$$
   $$\boldsymbol{P}_{t|t}^{(i)} = (\boldsymbol{I} - \boldsymbol{K}_t^{(i)}\boldsymbol{H}_t)\boldsymbol{P}_{t|t-1}^{(i)}$$
   (b) PF time update – prediction of new particles:
   $$\boldsymbol{\theta}_{t+1}^{(i)} \sim p(\boldsymbol{\theta}_{t+1}^{(i)}|\boldsymbol{\theta}_t^{(i)})$$
   (c) KF time update:
   $$\hat{\boldsymbol{x}}_{t+1|t}^{(i)} = \boldsymbol{M}\hat{\boldsymbol{x}}_{t|t}^{(i)}$$
   $$\boldsymbol{P}_{t+1|t}^{(i)} = \boldsymbol{M}\boldsymbol{P}_{t|t}^{(i)}\boldsymbol{M}^T + \boldsymbol{Q}_{t+1}(\boldsymbol{\theta}_{t+1}^{(i)})$$
5. Iterate from step 2) with $t := t + 1$

Multinomial resampling described in [4] was used as a resampling algorithm in MPF. Measurement error was set according to expert judgment based on the fact that the small measured values have higher relative error than high values due to the measurement methodology. In each step, first two moments of groundshine–dose distribution approximating the truth were predicted and updated.

The results are in compliance with our expectations for this special scenario. Model predictions were successfully adjusted in accordance with the measurements correcting the speed of dose mitigation. Even thought it seems that the methodology has a potential for improving of reliability of predictions in the late phase, the algorithm still has to be improved in terms of robustness and carefully tested.

References

[1] R. Daley. *Atmospheric data analysis.* Cambridge Univ. Press, Cambridge, (1991).

[2] D. P. Dee. A simple scheme for tuning forecast error covariance parameters. In 'ECMWF Workshop on Variational Assimilation', (1993).

[3] D. P. Dee. *On-line estimation of error covariance parameters for atmospheric data assimilation.* Monthly Weather Review **123** (1995).

[4] R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In 'Proceedings of the 4th International Symposium'. Image and Signal Processing and Analysis, (2005).

[5] G. Gaspari and S. E. Cohn. *Construction of correlation functions in two and three dimensions.* Quart. J. Roy. Meteor. Soc. **125** (1999), 723–757.

[6] F. Gering, W. Weiss, E. Wirth, R. Stapel, P. Jacob, H. Muller, and G. Prohl. *Assessment of evaluation of the radiological situation in the late phase of a nuclear accident.* Rad. Prot. Dosim **109** (2004).

[7] A. J. Haug. A tutorial on Bayesian estimation and tracking techniques applicable to nonlinear and non-Gaussian processes. Technical report, MITRE Corporation, (2005).

[8] R. Hofman. Assimilation scenario for long-term deposition of ^{137}Cs. In '8th International PhD Workshop on Systems and Control a Young Generation Viewpoint, Balatonfured, Hungary', (2007).

[9] R. Hofman and P. Pecha. Data assimilation of model predictions of long-time evolution of Cs-137 deposition on terrain. In '2008 IEEE International Geoscience & Remote Sensing Symposium', (2008). Boston, Massachusetts, U.S.A.

[10] T. Homma and T. Matsunaga. *OSCAAR Model - Description and Evaluation of Model Performance*, (2006).

[11] R. E. Kalman. *A new approach to linear filtering and prediction problems.* Trans. ASME J. Basic Eng. **82** (1960).

[12] E. Kalnay. *Atmospheric modeling, data assimilation and predictability.* Cambridge Univ. Press, Cambridge, (2003).

[13] M. Karny et al. *Optimized Bayesian Dynamic Advising: Theory and Algorithms.* Springer, London, (2005).

[14] A. C. Lorenc. *Analysis methods for numerical weather prediction.* Quarterly Journal of the Royal Meteorological Society **112** (1986), 1177–1194.

[15] H. L. Mitchell and P. L. Houtekamer. *An adaptive ensemble Kalman filter.* AMS Monthly Weathr Review **128** (1999), 416–433.

[16] P. Pecha and R. Hofman. Integration of data assimilation subsystem into environmental model of harmful substances propagation. In 'Harmo11 - 11th Internal Conf. Cambridge', (2007).

[17] P. Pecha, R. Hofman, and P. Kuča. Assimilation techniques in consequence assessment of accidental radioactivity releases. ECORAD 2008, Bergen, Norway, (2008).

[18] P. Pecha, R. Hofman, and E. Pechová. Training simulator for analysis of environmental consequences of accidental radioactivity releases. In '6th EUROSIM Congress on Modelling and Simulation, Ljubljana, Slovenia', (2007).

[19] C. Rojas-Palma. Data assimilation for off site nuclear emergency management. Technical report, SCK-CEN, DAONEM final report, RODOS(RA5)-RE(04)-01, (2005).

[20] C. Rojas-Palma et. al. *Data assimilation in the decision support system RODOS.* Rad. Prot. Dosim **104** (2003).

[21] T. B. Schön, F. Gustaffson, and N. P-J. *The marginalized particle filter - analysis, applications and generalizations.* IEEE Transaction on Signal Proceedings **53** (2005).

[22] T. B. Schön, R. Karlsson, and F. Gustaffson. *Marginalized particle filter for mixed linear/nonlinear state-space models.* ESAIM: PROCEEDINGS **19** (2007), 53–64.

[23] V. Šmídl and A. Quinn. *Variational Bayesian filtering.* IEEE Transactions on Signal Processing (2008). to be published.

Zpracování dat ze sčítání lidu pomocí statistického modelu*

Jan Hora

4. ročník PGS, email: hora@utia.cas.cz
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT
školitel: Jiří Grim, Ústav teorie informace a automatizace, AV ČR

Abstract. In last years we presented a method for interactive presentation of census results by means of the probabilistic expert system. The method is based on estimating a propabilistic model of the original microdata in form of a discrete distribution mixture of product components. The statistical information is derived from the estimated model without any risk of disclosure of individual respondents.

Now we managed to get the real microdata from census 2001 from the Czech Statistical Office and we present results of our first experiments made with these data.

Abstrakt. Využití statistického modelu pro prezentaci výsledků ze sčítání lidu je metoda, která novým způsobem umožňuje reprodukovat statistické vlastnosti populace při automatickém zachování bezpečnosti osobních údajů. Problematikou se zabýváme již velmi dlouho, teprve nedávno se však podařilo získat Český statistický úřad pro aktivní spolupráci, čímž bylo konečně umožněno aplikovat zkoumanou metodu na reálných datech.

V návaznosti na několikaletou snahu prezentujeme výsledky tříměsíční práce s reálnými daty ze sčítání lidu v České republice v roce 2001. Na práci je pohlíženo jako na pilotní studii ověřující možnosti aplikace této metody na reálné využití pro sčítání lidu v roce 2011 a zároveň jako na vytvoření datového zázemí pro následné zkoumání shlukovacích metod pro kategoriální data.

1 Úvod

Sčítání lidu je nákladné šetření, které produkuje obrovské množství dat. V důsledku nutnosti ochrany osobních údajů respondentů je však množství veřejně dostupných informací, které v pracně a nákladně získaných datech jsou, značně omezené.

Z jednotlivých dotazníků jsou sice odstraněny osobní údaje, je však obecně známo, že i takto anonymizovaný dotazník může být s využitím obecně dostupných informací jednoznačně identifikován. Proto nemohou být ani dotazníky zbavené osobních údajů volně přístupné veřejnosti.

Výsledky sčítání lidu se tedy obvykle zveřejňují souhrnně pro jednotlivé administrativní území části, např dle sčítacích okrsků. Takto agregované údaje pak představují velmi podrobnou a užitečnou informaci z hlediska geografického, avšak velká část obsažené informace se tím ztratí.

Jinou možností publikace výsledků jsou tištěné tabulky, tímto způsobem však lze zveřejnit jen velmi malou část zajímavých údajů, neboť počet tabulek velmi rychle narůstá, začneme-li uvažovat o subpopulacích podmíněných kombinací několika proměnných.

*This research was supported by the grant GACR 102/07/1594 of the Czech Grant Agency and by the projects of the Grant Agency of MŠMT 2C06019 ZIMOLEZ and 1M0572 DAR.

Metoda interaktivní reprodukce výsledků sčítání lidu pomocí statistického modelu (viz [2]) nabízí v této souvislosti nový, uživatelsky pohodlný přístup k výsledkům sčítání lidu při dokonalém zabezpečení ochrany dat. Jakákoli identifikace respondentů pomocí statistického modelu je znemožněna klesající spolehlivostí histogramů odvozených pro malé části populace (viz [2]).

Práce úzce navazuje na výzkumy prováděné na vzorku dat pražských domácností ze sčítání lidu v ČR v roce 1991 (viz např. [2], [10] a [11]) a rozvádí ji aplikací teoretických výsledků na reálná data ze sčítání lidu v roce 2001, která se podařilo získat až teprve v červnu tohoto roku. Oproti původním datům je zde třeba řešit ještě skutečnost, že ne vždy jsou všechny otázky vyplněné.

Cílem stávajícího výzkumu je ověřit možnosti metody na reálných datech a připravit tuto možnost pro plánované sčítání lidu v roce 2011. Druhým cílem je připravit podmínky pro zpracování dat pomocí metod informační a shlukové analýzy pro kategoriální data, které byly zkoumány v předchozích letech (viz např. [9]).

1.1 Stávající způsoby prezentace výsledků

Současné možnosti publikace statistických informací ze sčítání lidu lze zařadit do několika kategorií

- **Publikace výsledků v tištěné podobě** představuje nejtradičnější cestu zpřístupnění zjištěných statistických vlastností populace. Tištěné publikace se ovšem nutně omezují na nejzákladnější údaje a nejčastěji diskutované aspekty dat. Jak již bylo zmíněno v předchozím odstavci, tištěné materiály mohou pokrýt jen malou část reálně možných otázek, které mohou být ve specifických situacích formulovány různými uživateli.

- **Komerční služby statistických úřadů.** Jakýkoli dotaz týkající se sčítání lidu lze zodpovědět na základě specifického výpočtu s využitím původní databáze statistického úřadu. Bohužel, písemné zadání odpovídající zakázky příslušnému statistickému úřadu představuje těžkopádný a zdlouhavý způsob získávání informací, který není vhodný pro interaktivní výzkum, kdy formulaci otázky je třeba upřesňovat podle zjištěných výsledků.

- **Agregace dotazníků dle vybraných kritérií** Jednotlivé dotazníky jsou agregovány např. dle sčítacích okrsků. Tato metoda umožňuje přesné zobrazení rozložení různých vlastností populace dle geografického hlediska, ale již neumožňuje sledovat vlastnosti populací, které jdou napříč členěním použitým k agregaci.

- **Generování a publikace tabulek.** Obvykle mohou být uloženy a na různých paměťových médiích distribuovány pouze tabulky nízkého řádu (6 - 10 proměnných). Je zřejmé, že každá tabulka popisuje pouze statistické vztahy mezi tabelovanými proměnnými. Výběr subpopulace je tak omezen vždy jen na kombinace hodnot tabelovaných proměnných. Nabízené tabulky je navíc nutné ověřovat z hlediska spolehlivosti ochrany dat a vhodným způsobem anonymizovat identifikovatelné údaje [6]. Omezení identifikovatelnosti dat je ovšem nutně spojeno se ztrátou informace a vnášením nepřesností.

- **Poskytování podsouborů anonymizovaných mikrodat.** Z původního souboru individuálních dat jsou vybírány podsoubory a upravovány pomocí různých technik, jako je záměna údajů, pozměňování dat a pod., s cílem znemožnit jakoukoli identifikaci osobních údajů respondentů [6]. Soubor mikrodat představuje nejdokonalejší formu poskytování informací, která umožňuje analýzu dat v plné obecnosti bez jakýchkoli formálních omezení. Přesnost údajů, které lze odvodit z daného souboru mikrodat, bohužel klesá s jeho velikostí, závisí na kvalitě provedeného výběru a také na míře znehodnocení způsobené ochrannými anonymizačními postupy. Omezují se také možnosti analýzy malých subpopulací. Přístup k souborům mikrodat je umožněn ve většině zemí EU a je považován za doklad vysoké úrovně statistického servisu. Na druhé straně je tento postup značně citlivý z hlediska ochrany osobních údajů. Možnost pracovat s mikrodaty zpravidla podléhá schvalovací proceduře a není zaručena automaticky každému žadateli.

Ukazuje se, že ochrana osobních údajů, jakkoli nezbytná, je značně omezující z hlediska obvyklých požadavků ekonomických a sociálních výzkumů. V popředí zájmu je proto vytváření nových přístupů a metod, které mohou zkvalitnit a rozšířit informační nabídku statistických úřadů. Cílem je dosažení optimální rovnováhy mezi nutnou ochranou osobních údajů a dostupností užitečných informací.

2 Vstupní datový soubor

Datový soubor obsahuje vybrané odpovědi z dotazníků ze sčítání osob, bytů a domů České republiky z roku 2001. Jednotlivé vektory v souboru se skládají z vybraných údajů z dotazníku osob doplněné o údaje z odpovídajícího bytového dotazníku. Výsledný soubor obsahuje 10230060 záznamů s odpovědmi na 24 otázek, přičemž ne všechny odpovědi jsou vyplněné.

Formálně tedy uvažujeme konečný diskrétní N rozměrný prostor $\mathcal{X}$ ($N = 24$)

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_N, \qquad \mathcal{X}_n = \{\xi_{n,1}, ..., \xi_{n,K_n}\}, \tag{1}$$

kde $\mathcal{X}_n$ reprezentuje množinu možných odpovědí na otázku číslo n. Dále uvažujme datový soubor S

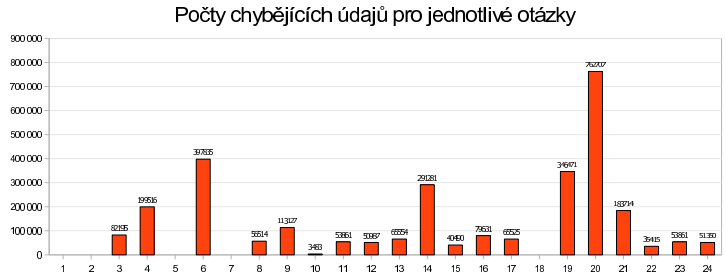
$$S = \left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(K)}\right\}, \qquad \boldsymbol{x}^{(i)} \in \hat{\mathcal{X}} = \hat{\mathcal{X}}_1 \times \hat{\mathcal{X}}_2 \times ... \times \hat{\mathcal{X}}_N, \qquad \hat{\mathcal{X}}_n = \mathcal{X}_n \cup \{\xi_{n,0}\} \tag{2}$$

kde hodnota $\xi_{n,0}$ reprezentuje skutečnost, že odpovídající otázka v dotazníku nebyla zodpovězena. Jedná se tedy o tzv. chybějící údaj.

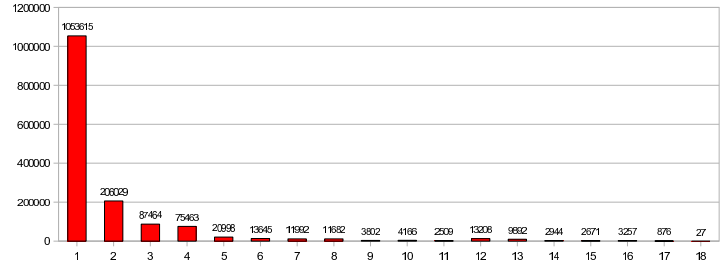
Graf 1 ukazuje nepravidelnost rozložení chybějících údajů dle jednotlivých otázek. Graf 2 zobrazuje počty vektorů dle počtu chybějících údajů ve vektoru. Z obrázku lze např. vyčíst, že u 32875 vektorů chybí více jak polovina odpovědí.

V celé problematice se zabýváme prakticky výlučně o takové podmožiny prostoru $\mathcal{X}$ resp. souboru S, které lze určit kombinací několika odpovědí. Mějme tedy vektor $\boldsymbol{x}_C$ s kombinací c odpovědí na c různých otázek

$$\boldsymbol{x}_C = (\xi_{n_1,k_1}, \xi_{n_2,k_2}, ..., \xi_{n_c,k_c}), \qquad \boldsymbol{x}_C \in \mathcal{X}_C = \mathcal{X}_{n_1} \times ... \times \mathcal{X}_{n_c}, \tag{3}$$



Obrázek 1: Počty chybějících údajů u jednotlivých otázek



Obrázek 2: množství vektorů dle počtu chybějících údajů ve vektoru

kde $1 \leq c \leq N$ a C představuje indexovou množinu odpovídající výběru otázek $C = \{n_1, n_2, ..., n_c\}$.

Potom subpopulací $A(\boldsymbol{x}_C)$ definovanou podmínkou $\boldsymbol{x}_C$ rozumíme takovou podmožinu prostoru $\mathcal{X}$, pro kterou platí

$$A(\boldsymbol{x}_C) = \left\{\boldsymbol{y} \in \mathcal{X} | \left(\boldsymbol{y}_{n_1}, \boldsymbol{y}_{n_2}, ..., \boldsymbol{y}_{n_c}\right) = \boldsymbol{x}_C\right\} \tag{4}$$

Typickým příkladem takové subpopulace je množina všech nezaměstnaných v praze apod. Skutečnou velikostí suboppulace $A(\boldsymbol{x}_C)$ pak rozumíme četnost výskytu kombinace $\boldsymbol{x}_C$ v souboru S. Tj.

$$sizeof(A(\boldsymbol{x}_C)) = \sum_{\boldsymbol{y} \in S} \delta(\boldsymbol{y}_C, \boldsymbol{x}_C), \qquad \boldsymbol{y}_C = \left(\boldsymbol{y}_{n_1}, \boldsymbol{y}_{n_2}, ..., \boldsymbol{y}_{n_c}\right) \tag{5}$$

kde $\delta(\boldsymbol{a}, \boldsymbol{b})$ značí standardní delta funkci, tj. $\delta(\boldsymbol{a}, \boldsymbol{b}) = 1$ pokud $\boldsymbol{a} = \boldsymbol{b}$, jinak $\delta(\boldsymbol{a}, \boldsymbol{b}) = 0$.

3 Reprodukce statistických vlastností souboru pomocí směsi

Je obecně známým faktem, že sčítání lidu představuje jednorázové šetření, které nelze opakovat jako náhodný experiment. Formálně však můžeme na vyplněný dotazník popsaný vektorem $\boldsymbol{x}$ pohlížet jako na realizaci nějakého neznámého náhodného vektoru v nabývající hodnot z $\mathcal{X}$ a na soubor S jako na posloupnost nezávislých realizací tohoto vektoru.

Veškeré statistické vlastnosti náhodného vektoru v jsou potom popsány jeho sdruženým rozložením pravděpodobnosti $P(\boldsymbol{x})$, které, zjednodušeně řečeno, popisuje chování náhodného respondenta. Pravděpodobnost výskytu vektoru $\boldsymbol{x}$ v souboru S pak aproximujeme pomocí diskrétní distribuční směsi součinových komponent

$$P(\boldsymbol{x}) = \sum_{m=1}^{M} w_m \prod_{n=1}^{N} p_n(x_n|m), \qquad \boldsymbol{x} \in \mathcal{X}, \tag{6}$$

kde M je počet komponent směsi, w_m jsou jednotlivé váhy komponent a $p_n(.|m)$ jsou jednorozměrné podmíněné distribuce v komponentě.

Protože prostor $\mathcal{X}$ je konečný, víme, že existuje směs s konečným počtem komponent $M \leq |\mathcal{X}|$, která popisuje rozložení pravděpodobnosti na prostoru $\mathcal{X}$ zcela přesně. Stejné tvrzení platí i za předpokladu, že je konečný soubor S (potom stačí $M \leq |S|$ komponent). Můžeme tedy tvrdit, že daný soubor S jsme schopni popsat konečnou směsí libovolně přesně. Abychom však zajistili bezpečnost osobních údajů, nemůže být model přesný příliš.

Velkou výhodou uvedeného modelu je velmi jednoduché vyjádření odhadu relativní velikosti subpopulace $A(\boldsymbol{x}_C)$ definované podmínkou $\boldsymbol{x}_C$ (viz (4)). Ten je roven pravděpodobnosti $P(\boldsymbol{x}_C)$, která lze vyjádřit prostým vynecháním členů v součinu ve výrazu (6)

$$P(\boldsymbol{x}_C) = \sum_{\boldsymbol{y} \in A(\boldsymbol{x}_y)} P(\boldsymbol{y}) = \sum_{m=1}^{M} w_m \prod_{i=1}^{c} p_{n_i}(x_{n_i}|m) \tag{7}$$

Tato vlastnost umožňuje velmi rychlé odvozování pravděpodobností, které nás převážně zajímají a které jsou potřeba jako vstupní informace pro interaktivní pravděpodobnostní expertní systém, který je součástí projektu.

4 Odhad parametrů modelu

Standardně se pro odhad parametrů směsi v podobných případech využívá iteračního EM algoritmu, který hledá maximálně věrohodný odhad tím, že monotónně zvyšuje hodnotu věrohodnostní funkce. Jako počáteční řešení volíme náhodně zašuměné uniformní rozložení. Použití tohoto algoritmu bylo popsáno např. v [3].

Pro odhad parametrů modelu na datech s chybějícími údaji je možné modifikovat schéma algoritmu prostým vynecháním odpovídajícch součinitelů v kroku E (viz [1]).

E-krok : $(m \in \mathcal{M},\ \boldsymbol{x} \in S)$

$$q^{(t)}(m|\boldsymbol{x}) \ = \ \frac{w_m^{(t)} \prod_{n=1, x_n \neq \xi_{(n,0)}}^{N} p_n^{(t)}(x_n|m)}{\sum_{j \in \mathcal{M}} w_j^{(t)} \prod_{n=1, x_n \neq \xi_{n,0}}^{N} p_n^{(t)}(x_n|j)} \tag{8}$$

M-krok : $(m \in \mathcal{M})$

$$w_m^{(t+1)} = \frac{1}{|S|} \sum_{\boldsymbol{x} \in \mathcal{S}} q^{(t)}(m|\boldsymbol{x}) \tag{9}$$

$$p_i^{(t+1)}(\xi|m) = \frac{1}{\sum_{\boldsymbol{x} \in \mathcal{S}} q^{(t)}(m|\boldsymbol{x})} \sum_{\boldsymbol{x} \in \mathcal{S}} \delta(x_i, \xi) q^{(t)}(m|\boldsymbol{x}) \tag{10}$$

Dalšího zpřesnění modelu lze dosáhnout doplněním chybějících údajů pomocí modelu optimalizovaného pomocní uvedenému schématu (8) a následným upřesněním modelu na souboru s doplněnými údaji. Vlastní výpočet se potom skládá ze tří fází - učení se na neúplných datech, doplnění dat a učení se na doplněných datech.

5 Ověřování přesnosti

Přirozeným kritériem pro měření přesnosti je průměrná absolutní chyba odhadu pravděpodobnosti všech možných podmnožin prostoru $\mathcal{X}$.

$$\epsilon = \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} |P(A) - \hat{P}(A)|, \tag{11}$$

kde $\mathcal{A}$ je třída všech podmnožin prostoru $\mathcal{X}$, $P(A)$ je pravděpodobnost množiny A odvozená ze statistického modelu

$$P(A) = \sum_{x \in A} P(\boldsymbol{x}), \tag{12}$$

a $\hat{P}(A)$ je (skutečná) relativní četnost výskytu dotazníků z množiny A v původním souboru $\mathcal{S}$

$$\hat{P}(A) = \sum_{x \in \mathcal{S}} \varphi_A(\boldsymbol{x}), \tag{13}$$

($\varphi_A(\boldsymbol{x})$ je charakteristická funkce množiny A rovná 1 pro $\boldsymbol{x} \in A$ a rovná 0 pro $\boldsymbol{x} \notin A$).

Je zřejmé, že toto kritérium je vhledem k rozsahu množiny $\mathcal{A}$ prakticky nepoužitelné. Vzhledem k tomu, že se v našem případě zabýváme zejména odvozováním pravděpodobností podmnožin, které lze učit kombinací hodnot, zjednodušíme výpočet kritéria tím, že se omezíme pouze na podmnožiny A, které lze určit pomocí kombinace několika hodnot (viz pojem subpopulace popisovaný v (4)).

Dále, uvážíme-li, že náším cílem je reprodukovat pouze dostatečně velké subpopulace, omezíme se pouze na tzv. relevantní podmnožiny, což jsou ty, jejichž skutečná velikost je větší než 1570. Hodnota tohoto prahu vychází ze statistické přesnosti odhadu, kdy požadujeme přesnost alespoň 5% na úrovni spolehlivosti 95%. Odvození tohoto prahu je možné nalézt např. v [11] či [8].

Označme tedy $\mathcal{A}_{rn}$ třídu všech relevantních podmožin, které lze určit pomocí kombinace maximálně n odpovědí.

$$\mathcal{A}_{rn} = \left\{ A(\boldsymbol{x}_C) \in \mathcal{A} \middle| \boldsymbol{x}_C \in \mathcal{X}_C, |C| < n \wedge \hat{P}(A) > 1570 \right\} \tag{14}$$

Průměrnou relativní ϵ_R resp. absolutní ϵ_A chybu pak počítáme následovně

$$\epsilon_A(\mathcal{A}_{rn}) = \frac{1}{|\mathcal{A}_{rn}|} \sum_{A \in \mathcal{A}_{rn}} |P(A) - \hat{P}(A)|, \qquad \epsilon_R(\mathcal{A}_{rn}) = \frac{1}{|\mathcal{A}_{rn}|} \sum_{A \in \mathcal{A}_{rn}} \frac{|P(A) - \hat{P}(A)|}{\hat{P}(A)} \quad (15)$$

Poznámka 1. U základní verze EM algoritmu platí, že jednorozměrné marginální pravděpodobnosti jsou reprodukovány zcela přesně hned po první iteraci agoritmu (tzn. $\epsilon_A(\mathcal{A}_{r1}) = 0$). V případě modifikace pro neúplná data tato skutečnost již obecně neplatí.

6 Experimentální část

6.1 Výpočty modelů a jejich přesnost

Na získaném datovém souboru ze sčítání lidu 2001 bylo provedeno několik různých výpočtů. Tabulka 1 obsahuje výpočty pro modely z různým počtem komponent m, hodnotu dosaženého věrohodnostního kritéria a průměrnou relativní chybu. Počáteční řešení bylo vždy voleno náhodně a výpočet byl zastaven v případě, že přírůstek věrohodnostního kritéria klesl pod stanovený práh, resp. dříve, pokud výpočet trval příliš dlouho.

Počet komponent	orientační čas výpočtu	kritérium L	relativní chyba $\epsilon_R(\mathcal{A}_{r3})$
10	1 min	-28.0078	0.2903
100	7,5 min	-21.7319	0.1357
1000	1 h	-21.1125	0.0677
10000	30 h	-20.9682	0.0521

Tabulka 1: Přesnost a dosažená hodnota věrohodnostního kritéria pro různě složité modely. Relativní chyba byla počítána na množině relevantních subpopulací, které lze určit až třemi podmínkami.

6.2 Přesnost souborů mikrodat

Reprodukce statistických vlastností datového souboru pomocí distribuční směsi je alternativou k dosud používaným souborům mikrodat. Soubor mikrodat je náhodný výběr vzorků z datového souboru, většinou 1 - 10% původního počtu. V praxi jsou soubory dále upravovány tak, aby byla zajištěna požadovaná ochrana osobních údajů, tj. aby byla vyloučena možnost identifikovat údaje o jednotlivcích.

Pro porovnání přesnosti statistického modelu a souboru mikrodat bylo vybráno několik náhodných podsouborů, u kterých byla měřena chyba odhadu na stejném souboru kontrolních subpopulací $\mathcal{A}_{r3}$. Anonymizační procedura nebyla vzhledem k její náročnosti aplikována, dá se však očekávat, že by vedla pouze k nepatrnému zhoršení přesnosti.

Při porovnání tabulek (1) a (2) vidíme, že směs s $m = 1000$ komponent je již z hlediska relativní chyby přesnější než soubor mikrodat obsahujícím cca 1% vektorů z původního souboru. Z hlediska přesnosti je tedy popisovaná metoda publikace výsledků sčítání lidu srovnatelná s využitím souborů mikrodat, které v současnosti patří k nejdokonalejším používaným způsobům publikace takových dat.

velikost	počet vektorů	relativní chyba ϵ_{r3}
1 %	102405	0.0793
5 %	511213	0.0348
10 %	1023442	0.0240

Tabulka 2: Přesnost různě velkých náhodně vybraných souborů mikrodat. Skutečná velikost a relativní chyba byla spočtena jako průměrná hodnota pro vždy tři náhodně vybrané podsoubory pro každou požadovanou velikost

6.3 Interaktivní prezentace výsledků

Navrhovaná metoda podstatně využívá faktu, že konečná směs součinových komponent je přímo použitelná jako báze znalostí pravděpodobnostního expertního systému PES (viz např. [7]).

Tento systém nabízí uživateli srovnatelné možnosti jako přímý kontakt s původním datovým souborem prostřednictvím databázového systému. Expertní systém odvozuje statistické informace přímo z odhadnutého modelu, bez nutnosti jakéhokoliv přístupu k původnímu datovému souboru. Ochrana osobních dat je tak dokonale zaručena, protože směsový model neumožňuje identifikaci jednotlivých dotazníků.

Informace expertního systému jsou uživateli nabídnuty ve formě podmíněných histogramů pro zadané subpopulace.

7 Závěr a další práce

Práce obsahuje první výsledky zpracování reálného datového souboru ze sčítání lidu v roce 2001, kdy se po dlouhé době podařilo získat Český statistiký úřad pro aktivní spolupráci. Na toto zpracování je pohlíženo jako na pilotní projekt testující použitelnost navrhované metody pro případné použití pro sčítání lidu v roce 2011.

Navrhovaná metoda umožňuje zpřístupnit statistické informace široké veřejnosti v daleko větší míře, než je tomu u stávajících forem zveřejňování výsledků sčítání lidu. Zároveň garantuje zachování bezpečnosti osobních údajů, neboť přesnost modelu klesá u malých subpopulací.

Výsledky uvedené v této práci již umožnují tvrdit, že z hlediska přesnosti měřené na množině relevantních subpopulací je navrhovaná metoda alespoň srovnatelná se soubory mikrodat, které také umožňují velmi obecně zkoumat statistické vlastnosti datového souboru.

Jako další navazující aktivitu plánujeme zkoumání možností informační a shlukové analýzy kategoriálních dat na zpracovávaném datovém souboru, který je typickým příkladem tohoto druhu dat.

Literatura

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, **39** (1977) l-38

[2] Grim J., Boček P., Pudil P. (2001): Safe dissemination of census results by means of interactive probabilistic models. In: *Proceedings of the ETK-NTTS 2001 Conference*, (Hersonissos (Crete), European Communities 2001, **2** (2001) 849-856

[3] J. Grim, "Probabilistic expert systems and distribution mixtures", *Computers and Artificial Intelligence*, Vol.9, No.3, pp. 241–256, 1990.

[4] Grim J.: EM cluster analysis for categorical data. In: *Structural, Syntactic and Statistical Pattern Recognition.* (Yeung D. Y., Kwok J. T., Fred A. eds.), (Lecture Notes in Computer Science. 4109). Springer, Berlin 2006, pp. 640-648.

[5] McLachlan G.J. and Peel D.: *Finite Mixture Models*, John Wiley & Sons, New York, Toronto, (2000)

[6] L.C.R.J. Willenborg, A.G. de Waal, *Elements of statistical disclosure control*, Springer Verlag, New York, 2001.

[7] J. Grim , P. Boček, "Statistical model of Prague households for interactive presentation of census data." In: *SoftStat'95. Advances in Statistical Software 5*, pp. 271 – 278, Lucius & Lucius: Stuttgart, March 1996.

[8] J. Hora "Interaktivní analýza výsledků sčítání lidu pomocí statistických modelů" - diplomová práce na FJFI, vedoucí práce Jiří Grim

[9] Grim J., Hora J.: Minimum Information Loss Cluster Analysis for Cathegorical Data. In: *Machine Learning and Data Mining in Pattern Recognition.* (Perner P. ed.), (Lecture Notes in Artificial Intelligence 4571). Springer, Berlin 2007, pp. 233 - 247.

[10] J. Grim , J. Hora, P. Boček, P. Somol, P. Pudil, "Information Analysis of Census Data by Using Statistical Models." *Statistics - Investment in the Future*, Praha, 6. - 7. září 2004.

[11] J. Grim, J. Hora, P. Pudil: Interaktivní reprodukce výsledků sčítání lidu pomocí statistického modelu se zaručenou ochranou anonymity dat , *Statistika vol.40*, 5 (2004), p. 400-414

Resonant Effect for Periodically Time-Dependent Singular Flux Tube and Homogeneous Magnetic Field

Tomáš Kalvoda*

2nd year of PGS, email: kalvotom@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Pavel Šťovíček, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. The dynamics of a classical charged particle confined to a plane, under the influence of a homogeneous magnetic field perpendicular to the plane and a time-periodic singular flux tube (so called Aharonov-Bohm flux) is investigated. For the description of the system we use the action-angle coordinates. The main tool of our analysis is von Zeipel's method, which is a classical perturbation method. We are interested especially in the resonant phenomena between the strength of the field and the frequency of the singular flux.

Abstrakt. Tento příspěvek se zabývá klasickou dynamikou nabité bodové částice pohyhující se v rovině pod vlivem homogenního magnetického pole, které je na tuto rovinu kolmé, a singulárního časově periodicky závislého magnetického toku (tzv. Aharonova-Bohmova toku). Výchozím bodem pro studium stability tohoto systému jsou souřadnice akce-úhel. Hlavním nástrojem pak klasická poruchová metoda pocházející od von Zeipela. Hlavní důraz je kladen na odhalení rezonančních efektů mezi sílou magnetického pole a frekvencí singulárního toku.

1 Introduction

In the present contribution we are interested in the qualitative behaviour of a classical charged particle which is under the influence of a homogeneous magnetic field and the time-dependent singular flux tube[1] piercing the origin of coordinate system. The basic description of the system is given in the following paragraphs. In the subsequent sections we will invoke standard perturbation technique due to von Zeipel. This method gives much better results than the Bogolyubov's averaging (see for example [5]) used in [4].

Let the Cartesian coordinates in the plane be denoted by $q = (q_1, q_2) \in \mathbb{R}^2$. The vector potential A consists of two parts. The homogeneous magnetic field of strength $b > 0$ (such choice can be made without loss of generality) perpendicular to the q-plane is generated by the potential

$$A_h(q) = \frac{-b}{2} q^\perp,$$

*I would like to express thanks to Dr. Joachim Asch from C.N.R.S. Marseille for many valuable discussions and help during my stay in Marseille.

[1] Sometimes also called Aharonov-Bohm flux tube.

where $q^\perp = (-q_2, q_1)$. The second part corresponds to a singular flux through the origin of the coordinate system and is given by

$$A_f(q,t) = \frac{\Phi(t)}{2\pi|q|^2} q^\perp,$$

where $|q| = \sqrt{q_1^2 + q_2^2}$ and $\Phi : \mathbb{R} \to \mathbb{R}$ is a periodic function. The total vector potential is given as a sum $A_h + A_f$. Passing to the polar coordinates and using the Legendre transform it is straightforward to arrive at the Hamiltonian

$$H(r,\theta,p_r,p_\theta,t) = \frac{1}{2m}\left(p_r^2 + \left(\frac{p_\theta - \frac{e\Phi(t)}{2\pi}}{r} + \frac{eb}{2}r\right)^2\right). \tag{1}$$

considered on phase space $(\mathbb{R}^+ \times S_1) \times \mathbb{R}^2$. The equations of motion are

$$\dot{p}_\theta = -\frac{\partial H}{\partial \theta} = 0, \quad \dot{\theta} = \frac{\partial H}{\partial p_\theta} = \frac{p_\theta - \frac{e\Phi(t)}{2\pi}}{mr^2} + \frac{eb}{2m}. \tag{2}$$

It is obvious that θ is a cyclic coordinate, therefore p_θ is an integral of motion. This fact enables us to treat p_θ as constant. And the question of stability is essentially contained in the single ordinary differential equation for radial distance r

$$\ddot{r} + \frac{e^2b^2}{4m^2}r = \frac{\left(p_\theta - \frac{e\Phi}{2\pi}\right)^2}{mr^3}. \tag{3}$$

From now on we set the charge and mass of the particle equal to one. In order to use classical perturbation techniques it is necessary to transform the system to the so called Action-Angle coordinates (for more details see [1]). In [4] it is shown that there is a canonical transformation from $r, p_r \in \mathbb{R}^+ \times \mathbb{R}$ to $(\varphi, I) \in S_1 \times \mathbb{R}^+$ coordinates which transforms the Hamiltonian (1) to new one

$$H_c(\varphi, I, t) = bI - \operatorname{sgn}(a(t))\dot{a}(t) \arctan\left(\frac{\sqrt{I}\cos\varphi}{\sqrt{I + |a(t)|} + \sqrt{I}\sin\varphi}\right) \tag{4}$$

And equations of motion are given by

$$\dot{\varphi} = b - \frac{a\dot{a}}{2}\frac{\cos\varphi}{\sqrt{I(I+|a|)}}\frac{1}{2I + |a| + 2\sqrt{I(I+|a|)}\sin\varphi}, \tag{5}$$

$$\dot{I} = -\frac{\operatorname{sgn} a}{2}\left(\dot{a} - \frac{|a|\dot{a}}{2I + |a| + 2\sqrt{I(I+|a|)}\sin\varphi}\right). \tag{6}$$

Moreover, in [4] it is also shown, that the question of stability is answered by the behaviour of the action coordinate I. More precisely, if (for certain initial conditions) the solution $I(t)$ of the above equation is bounded then also the radial distance of the particle is bounded. Or, in other words, the particle will not leave some bounded region of the plane. On the other hand, if it happens that $I(t) \to +\infty$ as $t \to \infty$, then the particle will get arbitrarily close to and far from[2] the flux tube during the time evolution.

[2]This means that 0 and $+\infty$ are accumulation points of the particular trajectory $\{r(t)\}_{t\geq 0}$.

2 Dynamics Generated by the Time-dependent Singular Flux Tube

In order to become acquainted with the dynamics of the system, it is appropriate to investigate the influence of magnetic field and time-dependent flux separately. If the flux is turned off, i.e. the particle is influenced only by the homogeneous magnetic field, then the classical trajectories are circles in the q-plane. This fact is a well known elementary result.

Let us investigate what happens if we turn off the magnetic field. To answer the stability question we look at the behaviour of solutions of equation (3) - of course we again put $e = m = 1$ and in addition also $b = 0$. In this subsection the flux need not to be periodic, but it must satisfy the conditions (7). The basic result is formulated in the following

Lemma 1. *Suppose that $a \in C(\mathbb{R})$ is such that $a(t) \neq 0$ for all $t \in \mathbb{R}$ and*

$$\int_T^\infty \frac{a(t)^2}{t^2}\mathrm{d}t < \infty,\ \int_{-\infty}^{-T} \frac{a(t)^2}{t^2}\mathrm{d}t < \infty,\ \int_T^\infty a(t)^2\mathrm{d}t = \int_{-\infty}^{-T} a(t)^2\mathrm{d}t = +\infty. \tag{7}$$

for certain $T > 0$. Then for any $(r_0, v_0) \in \mathbb{R}^+ \times \mathbb{R}$ there exists an unique solution $r(t)$, defined on $\mathbb{R}$, to the initial value problem

$$r''(t) = \frac{a(t)^2}{r(t)^3},\ r(0) = r_0,\ r'(0) = v_0. \tag{8}$$

Moreover, the solution satisfies the condition $r(t) \sim c_\pm t$ as $t \to \pm\infty$ where $c_+ > 0$ and $c_- < 0$.

Proof. The differential equation is equivalent to the dynamical system

$$x' = f(t,x) = \begin{pmatrix} x_2 \\ a(t)^2/x_1^3 \end{pmatrix},\ x \in U = \mathbb{R}^+ \times \mathbb{R}.$$

Since $f \in C(\mathbb{R} \times U, \mathbb{R}^2)$ is locally Lipschitz continuous in the second argument one can use Picard-Lindelöf theorem to establish existence and uniqueness of the local solutions. Thus, the solutions are either defined for all $t \in \mathbb{R}$, or they approach the boundary of $\mathbb{R}^+$ in finite time (more precisely they escape to infinity $r \to \infty$ or fall on the zero $r \to 0$). We will analyse the case $t > 0$, the rest is analogous.

First of all observe, that generally $r'(t)$ is increasing function. Therefore we will consider three situations according to the initial velocity.

Suppose that $v_0 > 0$ and that we have solution $r(t)$ of the IVP (8) defined on the interval (T_-, T_+). Hence $r(t)$ is increasing for all $t \in [0, T_+)$, so there always exists[3] $\lim_{t\uparrow T_+} r(t) > r_0 > 0$, therefore the solution can be prolonged to the infinite interval. We can assume that $T_+ = +\infty$. For any $t \geq 0$ it is true that

$$r(t) = r_0 + \int_0^t r'(s)\mathrm{d}s > r_0 + v_0 t, \quad r(t) = r_0 + v_0 t + \int_0^t (t-s)\frac{a(s)^2}{r(s)^3}\mathrm{d}s.$$

[3]The symbol $\lim_{t\uparrow a}$ denotes limit from the left.

The claim of the Lemma is equivalent to the existence of positive finite limit

$$\lim_{t\to\infty}\frac{r(t)}{t}=\lim_{t\to\infty}\left\{v_0+\int_0^t\frac{a(s)^2}{r(s)^3}\mathrm{d}s+\frac{1}{t}\left(r_0-\int_0^t s\frac{a(s)^2}{r(s)^3}\mathrm{d}s\right)\right\}.$$

But this is true since $1/r(t)<1/(r_0+v_0t)$ for any $t\geq 0$ and $\int_T^\infty a(s)^2/s^2\mathrm{d}s$ converges. The positivity of the limit is obvious.

Suppose now that $v_0=0$. Since a^2 is positive it is true that

$$r'(t)=\int_0^t\frac{a(s)^2}{r(s)^3}\mathrm{d}s>0$$

for t from domain of definition of r. Hence we can immediately pass to the preceding point.

Finally assume that $v_0<0$. Since $r'(t)$ is increasing our objective is to show that there exists $t>0$ such that $r'(t)=0$ and then again we can use the preceding considerations. Let us first show, that the solution can not approach the boundary $r=0$ in finite time. Let $r(t)$ be a solution of IVP (8) defined on (T_-,T_+), $T_+\in\mathbb{R}^+$, such that $\lim_{t\uparrow T_+}r(t)=0$. It must hold that for all $t\in[0,T_+)$ it is true that $r'(t)<0$. But

$$\begin{aligned}\lim_{t\uparrow T_+}r'(t)&=v_0+\lim_{t\uparrow T_+}\int_0^t\frac{a(s)^2}{r(s)^3}\mathrm{d}s=v_0+\lim_{t\uparrow T_+}\int_{1/r_0}^{1/r(t)}\frac{\rho a(r^{-1}(1/\rho))^2}{-r'(r^{-1}(1/\rho))}\mathrm{d}\rho=\\&=v_0+\int_{1/r_0}^{+\infty}\frac{\rho a(r^{-1}(1/\rho))^2}{-r'(r^{-1}(1/\rho))}\mathrm{d}\rho=+\infty,\end{aligned}$$

because the denominator in the integrand tends to zero or some constant and a is nonzero. This contradicts our hypothesis. Also in this case ($v_0<0$) the solution $r(t)$ of the IVP (8) can be prolonged to infinite interval. It remains to show that there is some $t^*>0$ such that $r'(t^*)=0$. So assume that we have solution obeying $\lim_{t\to\infty}r(t)=R\geq 0$ and $r'(t)<0$ for all $t>0$. Therefore $r(t)$ is decreasing. But now we have

$$r'(t)=v_0+\int_0^t\frac{a(s)^2}{r(s)^3}\mathrm{d}s>v_0+\frac{1}{r_0^3}\int_0^t a(s)^2\mathrm{d}s\to+\infty,\ \text{as } t\to+\infty.$$

This is impossible due to (7). □

With a little more effort we can treat also zeros of a:

Lemma 2. *Suppose that $a\in C^1(\mathbb{R})$ is such that if $a(t^*)=0$ for some $t^*\in\mathbb{R}$ then $a'(t^*)\neq 0$, and let the conditions (7) of Lemma 1 hold. Then all claims of Lemma 1 are true.*

Proof. The only part of proof of Lemma 1 which has to be changed is the proof of extensibility. In particular, suppose that $r(t)$ is a solution with domain (T_-,T_+) of the initial value problem (8) with $v_0<0$, $\lim_{t\uparrow T_+}r(t)=0$. If it happens so that $a(T_+)\neq 0$ one can use the same argument as in the proof of Lemma 1. However what if $a(T_+)=0$? Then we may write

$$r'(t)=v_0+\int_0^t\frac{a(s)^2}{r(s)^3}\mathrm{d}s=v_0+\int_0^t\frac{(a(T_+)+a'(\xi_s)(s-T_+))^2}{(r(T_+)+r'(\eta_s)(s-T_+))^3}\mathrm{d}s,\ t\in(0,T_+),$$

where $\eta_s, \xi_s \in (s, T)$. But the last integral diverges logarithmically as $t \uparrow T_+$ since

$$\int_0^t \frac{(a(T_+) + a'(\xi_s)(s - T_+))^2}{(r(T_+) + r'(\eta_s)(s - T_+))^3} \mathrm{d}s \geq \frac{1}{(-v_0)^3} \int_0^t \frac{a'(\xi_s)^2}{T_+ - s} \mathrm{d}s$$

and $a'(\xi_s) \to a'(T_+)$ as $s \uparrow T_+$. □

To complete our picture it is necessary to look at the equation for polar coordinate θ, (2). We immediately see that $\theta \to \text{const}$ as $t \to \pm\infty$. In other words the particle is "pushed from the origin". More precisely if the particle approaches from the infinity then it is deflected by the origin and asymptotically moves freely. All trajectories are unbounded, the particle escapes to infinity.

3 Von Zeipel's Method

Let us now look at the system with Hamilton's function (4) more closely. Suppose that $a(t)$ is a smooth periodic function with frequency Ω and that $a(t) = f(\Omega t) > 0$, where f is a 2π-periodic function. Therefore we have

$$H_c(\varphi, I, t) = bI - \Omega f'(\Omega t) \arctan \frac{\sqrt{I} \cos \varphi}{\sqrt{I + f(\Omega t)} + \sqrt{I} \sin \varphi}, \quad \varphi \in S_1,\ I \in \mathbb{R}^+.$$

In order to get rid of the time dependence let us introduce new phase $\varphi_2 = \Omega t$ and its conjugate variable I_2 (old variables φ, I are denoted by φ_1, I_1). We obtain so called extended Hamiltonian K which reads

$$K(\varphi, I) = (b, \Omega) \cdot I - \Omega f'(\varphi_2) \arctan \frac{\sqrt{I_1} \cos \varphi_1}{\sqrt{I_1 + f(\varphi_2)} + \sqrt{I_1} \sin \varphi_1}, \quad \varphi \in \mathbb{T}^2,\ I \in \mathbb{R}^+ \times \mathbb{R}. \tag{9}$$

Hamiltonians H_c and K are equivalent (in the sense that the corresponding solutions of Hamiltonian equations are the same up to parametrisation), provided that the initial conditions are properly matched (e.g. if $\varphi(0) = \varphi_0$ then $\varphi_2(0) = 0$ and $\varphi_1(0) = \varphi_0$). The extended Hamiltonian is in a form which is suitable for application of von Zeipel's method. This is a classical canonical perturbation method. The fundamental steps will be mentioned in course of the following computation. More details can be found in [2] or [3].

As a simple demonstrative example take $\Phi(t) = -2\pi\varepsilon \sin \Omega t$ (the procedure described below can be applied without any modification to more general fluxes, e.g. fluxes with finite number of nonzero Fourier coefficients). So $f(x) = p_\theta + \varepsilon \sin x$ and suppose that $0 < \varepsilon < p_\theta$. We will compute the approximate Hamiltonian up to order $\mathcal{O}(\varepsilon^3)$. For the sake of simplicity denote $(\omega_1, \omega_2) = (b, \Omega)$. Let us assume that (resonance condition)

$$\frac{\omega_2}{\omega_1} = \frac{p}{q},$$

where p, q are natural co-prime numbers. The extended Hamiltonian reads

$$K(\varphi, I, \varepsilon) = \omega \cdot I + \varepsilon\omega_2 \cos(\varphi_2) \arctan \frac{\sqrt{I_1} \cos \varphi_1}{\sqrt{I_1 + p_\theta - \varepsilon \sin \varphi_2} + \sqrt{I_1} \sin \varphi_1}.$$

Let us expand the Hamiltonian up to order $\mathcal{O}(\varepsilon^3)$

$$K(\varphi, I, \varepsilon) = \omega \cdot I + \varepsilon K_1(\varphi, I) + \varepsilon^2 K_2(\varphi, I) + \mathcal{O}(\varepsilon^3).$$

The first step of Von Zeipel's method consists of a near identity canonical transformation to new canonical coordinates ψ and J. The generating function of the transformation is sought in the form

$$S(\varphi, J, \varepsilon) = \varphi \cdot J + \varepsilon S_1(\varphi, J) + \varepsilon^2 S_2(\varphi, J) + \mathcal{O}(\varepsilon^3).$$

And one seeks new Hamiltonian in a similar form

$$\mathscr{K}(\psi, J, \varepsilon) = \omega \cdot J + \varepsilon \mathscr{K}_1(\psi, J) + \varepsilon^2 \mathscr{K}_2(\psi, J) + \mathcal{O}(\varepsilon^3).$$

Coefficients S_1, S_2 and $\mathscr{K}_1, \mathscr{K}_2$ are to be determined. Of course one can try to compute also higher order terms but the task is more and more difficult. The relation between the old and new Hamiltonian is given by the equality

$$K(\varphi, \partial_\varphi S, \varepsilon) = \mathscr{K}(\partial_J S, J, \varepsilon),$$

from which one finds that

$$\begin{aligned}
\mathscr{K}_1(\varphi, J) &= K_1(\varphi, J) + \omega \cdot \partial_\varphi S_1(\varphi, J), \\
\mathscr{K}_2(\varphi, J) &= K_2(\varphi, J) + \partial_J K_1(\varphi, J) \cdot \partial_\varphi S_1(\varphi, J) - \partial_\varphi \mathscr{K}_1(\varphi, J) \cdot \partial_J S_1(\varphi, J) \\
&\quad + \omega \cdot \partial_\varphi S_2(\varphi, J).
\end{aligned}$$

In the present situation the lattice of resonant frequencies is given by $\mathbb{K} = \mathbb{Z}(p, -q)$. Terms in the expansion of the new Hamiltonian $\mathscr{K}$ are chosen in the following way

$$\begin{aligned}
\mathscr{K}_1(\varphi, J) &= \langle K_1(\varphi, J) \rangle_{\mathbb{K}}, \\
\mathscr{K}_2(\varphi, J) &= \langle K_2(\varphi, J) + \partial_J K_1(\varphi, J) \cdot \partial_\varphi S_1(\varphi, J) - \partial_\varphi \mathscr{K}_1(\varphi, J) \cdot \partial_J S_1(\varphi, J) \rangle_{\mathbb{K}}.
\end{aligned}$$

The notation $\langle \cdot \rangle_{\mathbb{K}}$ means that we keep only resonant frequencies in the Fourier expansion. More precisely for a function $\eta : \mathbb{T}^n \to \mathbb{C}$ with Fourier expansion

$$\eta(\varphi) = \sum_{k \in \mathbb{Z}^n} \mathscr{F}[\eta(\varphi)]_k \exp(ik \cdot \varphi), \quad \mathscr{F}[\eta(\varphi)]_k = \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} f(\varphi) e^{ik \cdot \varphi} \mathrm{d}^n \varphi,$$

put

$$\langle \eta(\varphi) \rangle_{\mathbb{K}} = \sum_{k \in \mathbb{K}} \mathscr{F}[\eta(\varphi)]_k \exp(ik \cdot \varphi).$$

The bracket $\langle \cdot \rangle_{\mathbb{K}}$ is sometimes called the averaging operator. Functions $S_1(\varphi, J)$, $S_2(\varphi, J)$ are then obtained as solutions of the partial differential equations

$$\begin{aligned}
\omega \cdot \partial_\varphi S_1(\varphi, J) &= -\langle K_1(\varphi, J) \rangle_{\mathbb{K}^c}, \\
\omega \cdot \partial_\varphi S_1(\varphi, J) &= -\langle K_2(\varphi, J) + \partial_J K_1(\varphi, J) \cdot \partial_\varphi S_1(\varphi, J) - \partial_\varphi \mathscr{K}_1(\varphi, J) \cdot \partial_J S_1(\varphi, J) \rangle_{\mathbb{K}^c},
\end{aligned}$$

where $\mathbb{K}^c = \mathbb{Z}^2 \setminus \mathbb{K}$. It is possible to give solutions of these equations in form of formal series, but one has to be careful since it is here where the problem of small denominators

appears. However, if our flux function has finite number of nonzero Fourier coefficients then there is no problem.

For the computation below it is convenient to set

$$\beta \equiv \beta(J_1) = \sqrt{\frac{J_1}{J_1 + p_\theta}}.$$

Using crucial results from Appendix of [4] it is easy to derive that only the following Fourier coefficients, which we will need, are nonzero

$$\begin{aligned}
\mathscr{F}[K_1(\varphi, I)]_{(l,\pm 1)} &= \frac{\omega_2}{4il}\beta^{|l|}\exp\Big(il\frac{\pi}{2}\Big),\\
\mathscr{F}[\partial_{I_1}K_1(\varphi, I)]_{(l,\pm 1)} &= \frac{\omega_2 p_\theta \operatorname{sgn}(l)}{8i}\frac{\beta^{|l|+2}}{I_1^2}\exp\Big(il\frac{\pi}{2}\Big),\\
\mathscr{F}[K_2(\varphi, I)]_{(l,\pm 2)} &= \frac{\omega_2 \operatorname{sgn}(l)}{16ip_\theta}(1-\beta^2)\beta^{|l|}\exp\Big(il\frac{\pi}{2}\Big), \quad l \neq 0.
\end{aligned}$$

With these coefficients it is possible to compute terms $\mathscr{K}_1, \mathscr{K}_2, S_1, S_2$. There are three situations which need a separate treatment, in particular $q = 1$, $q = 2$, and $q \geq 3$. In the following subsections only results are presented, the computation itself is straightforward but tedious and space-consuming. The formulae were computed by hand and checked using the computer algebra system *Mathematica 6.0*.

3.1 The case of $q = 1$

In other words we have here

$$\frac{\omega_2}{\omega_1} = \frac{\Omega}{b} = p \in \mathbb{N}.$$

Under this assumption one can compute that

$$\begin{aligned}
\mathscr{K}_1(\psi, J) &= \frac{\omega_2}{2p}\beta^p \sin\Big(p\frac{\pi}{2} + p\psi_1 - \psi_2\Big),\\
\mathscr{K}_2(\psi, J) &= \frac{p\omega_2}{16p_\theta}\Big(\frac{1}{\beta} - \beta\Big)^2\left(\frac{1}{p} + (\beta^{2p} + \beta^{-2p})\ln(1-\beta^2) + \sum_{n=1}^{p-1}\frac{\beta^{2(p-n)} + \beta^{-2(p-n)}}{n}\right) +\\
&\quad + \frac{(-1)^p p\omega_2}{8p_\theta}\Big(\frac{1}{\beta} - \beta\Big)^2\left(\ln(1-\beta^2) + \sum_{l=1}^{p}\frac{\beta^{2l}}{l}\right)\cos(2p\psi_1 - 2\psi_2) +\\
&\quad + \frac{(-1)^p\omega_2}{8p_\theta}(1-\beta^2)\beta^{2p}\cos(2p\psi_1 - 2\psi_2).
\end{aligned}$$

Further, it is possible to give a closed form expression for S_1 but we will not need it here. It is important to observe that the new Hamiltonian system obtained by von Zeipel's method has additional integrals of motion, while in general the original system has only one integral of motion (namely the Hamiltonian K itself). The number of integrals produced depends on the number of independent resonance relations and the dimension of the phase space. Moreover, these integrals does not depend on the order of approximation, but only on the resonance relations.

To see this, let us denote

$$R = \begin{pmatrix} p & -1 \\ 1-p & 1 \end{pmatrix}.$$

The matrix R is chosen such that it has the basic resonant vector $(p, -1)$ in the first row and has unit determinant and integer entries. Next step is canonical transformation from ψ, J coordinates to χ, P coordinates generated by the function

$$W(\psi, P) = P \cdot R\psi.$$

Therefore

$$\chi = \partial_P W = R\psi, \quad J = \partial_\psi W = R^T P.$$

The point is that the Hamiltonian in these new coordinates $\mathscr{K}(\chi, P)$ does not depend on χ_2, therefore $P_2 = J_1 + pJ_2$ is an integral of motion of the approximate system. Moreover, in new coordinates one has

$$\mathscr{K}(\chi, P) = \omega_1 P_2 + \varepsilon \mathscr{K}_1(\chi, P) + \varepsilon^2 \mathscr{K}_2(\chi, P) + \mathcal{O}(\varepsilon^3). \tag{10}$$

Since P_2 is integral of motion and the Hamiltonian does not depend on χ_2 we can plot the level curves in the χ_1, P_1-plane. This is done in Figure 1. Let us first look at the level

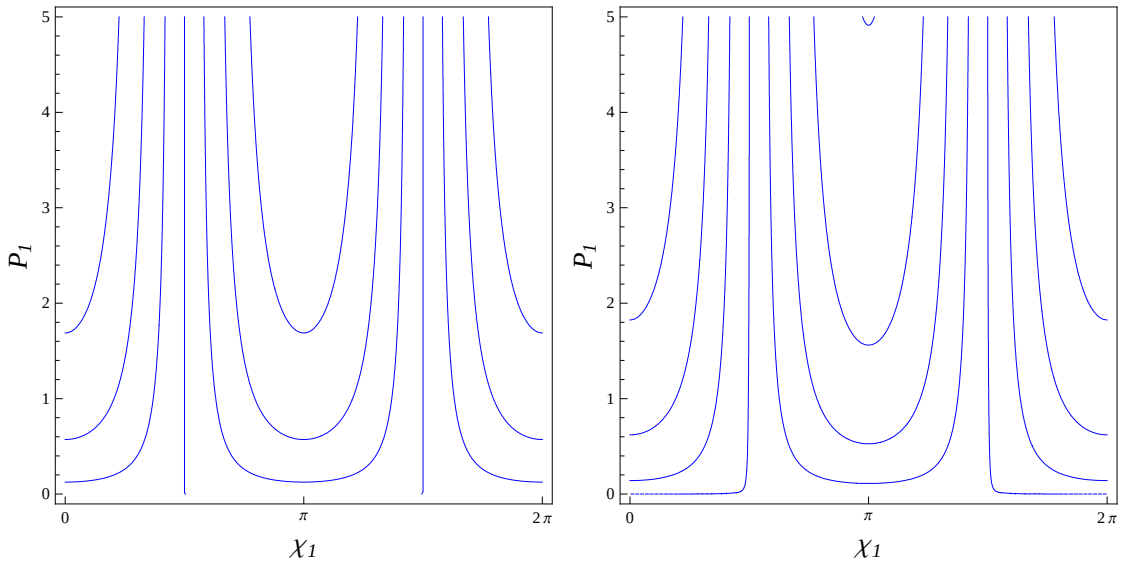


Figure 1: Typical level curves of Hamiltonian (10) without (left column) and with (right column) the $\mathcal{O}(\varepsilon^2)$ term in the case of $q = 1$. Note that in the left picture the curved lines does *not* approach asymptotes $\chi_1 = \pi/2, 3\pi/2$.

curves of the first order approximation. If we fix initial conditions $\chi(0) \in \mathbb{T}^2$, $P(0) \in \mathbb{R}^2$ then the equality

$$\mathscr{K}(\chi, P) = \mathscr{K}(\chi(0), P(0))$$

defines implicitly P_1 as a function of χ_1. Since P_2 is conserved one gets

$$\left(\frac{\beta}{\beta_0}\right)^p = \frac{\sin\left(p\frac{\pi}{2} + \chi_1(0)\right)}{\sin\left(p\frac{\pi}{2} + \chi_1\right)}$$

if $\sin\Big(p\pi/2+\chi_1(0)\Big)\neq 0$. Note that

$$0<\beta=\sqrt{\frac{pP_1+(1-p)P_2}{pP_1+(1-p)P_2+p_\theta}}<1,$$

therefore χ_1 varies between two roots of equation

$$\beta_0^{-p}=\frac{\sin\Big(p\frac{\pi}{2}+\chi_1(0)\Big)}{\sin\Big(p\frac{\pi}{2}+x\Big)}$$

which are nearest to $\chi_1(0)$ and P_1 approaches infinity as χ_1 tends to these roots. On the other hand, if $\sin\Big(p\pi/2+\chi_1(0)\Big)=0$ then $\chi_1=\chi_1(0)$ and P_1 can be arbitrary (more precisely bounded from below by $\frac{p-1}{p}P_2(0)$). This exactly corresponds to the left picture of Figure 1.

Since the second order correction $\mathscr{K}_2$ is complicated, it is impossible to carry out the procedure described above. However, since $\lim_{J_1\to\infty}\mathscr{K}_2(\psi,J)=0$ for any $\psi\in\mathbb{T}^2$ one can argue, that the picture described by the first order approximation will not be spoilt by the second order term. Also note that because $J_1=pP_1+(1-p)P_2$ and J_1 plays role of I_1 which was originally the action I, we just showed, that in the case $\Omega/b\in\mathbb{N}$ the resonant behaviour described at the end of Section 1 will occur.

3.2 The case of $q=2$

In general, for $q>1$ it is true that $\mathscr{K}_1(\psi,J)=0$. Also for any $q>1$ it is possible to compute[4] the function S_1:

$$S_1(\varphi,J)=\frac{1}{2}\Re\Bigg[-2i\arctan\frac{\beta\cos\varphi_1}{1+\beta\sin\varphi_1}+\frac{\omega_1}{\omega_1-\omega_2}i\beta e^{i\varphi_1}{}_2F_1\left(1,1-\frac{\omega_2}{\omega_1},2-\frac{\omega_2}{\omega_1},i\beta e^{i\varphi_1}\right)$$
$$+\frac{\omega_1}{\omega_1+\omega_2}i\beta e^{-i\varphi_1}{}_2F_1\left(1,1+\frac{\omega_2}{\omega_1},2+\frac{\omega_2}{\omega_1},-i\beta e^{-i\varphi_1}\right)\Bigg]\exp(-i\varphi_2)$$

The second term of the approximate Hamiltonian is nontrivial

$$\mathscr{K}_2(\varphi,J)=-\frac{\omega_2 p}{2^4p_\theta(1+p/2)}\beta^p\Big(1-\beta^2\Big)^2{}_2F_1\left(1,1+\frac{p}{2},2+\frac{p}{2},\beta^2\right)\cos(p\psi_1-2\psi_2+p\pi/2)$$
$$-\frac{p\omega_2}{2^5p_\theta}\Big(1-\beta^2\Big)^2\left(\frac{1}{1-p/2}{}_2F_1\left(1,1-\frac{p}{2},2-\frac{p}{2},\beta^2\right)+\frac{1}{1+p/2}{}_2F_1\left(1,1+\frac{p}{2},2+\frac{p}{2},\beta^2\right)\right)$$
$$+\frac{\omega_2}{8p_\theta}(1-\beta^2)\beta^p\sin(p\psi_1-2\psi_2+p\pi/2)$$

Again, we have one integral of motion. Following the same steps as in the end of the last subsection, but with the matrix

$$R=\begin{pmatrix}p & -2\\ \frac{1-p}{2} & 1\end{pmatrix},$$

[4]The symbol ${}_2F_1(a,b,c,z)$ stands for the Gauss hypergeometric function.

it follows that $P_2 = 2J_1 + pJ_2$ is conserved. For the level curves in the χ_1, P_1-plane see Figure 2. It appears that in this case it is not possible to have $P_1 \to \infty$. Let me just note

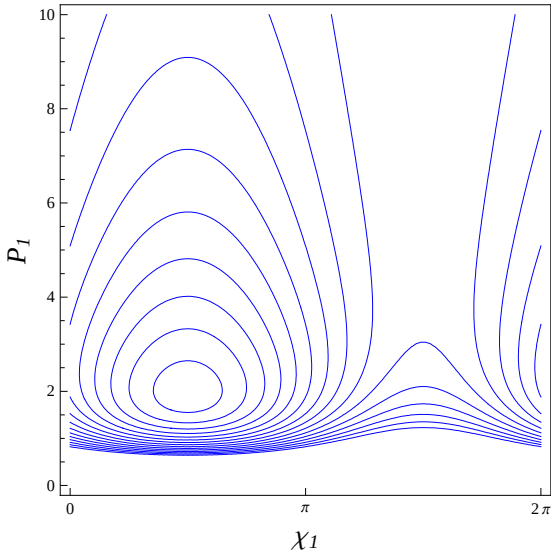


Figure 2: Typical level curves of the transformed Hamiltonian $\mathscr{K}(\chi, P) = \frac{\omega_1}{2}P_2 + \varepsilon^2\mathscr{K}_2(\chi, P)$ in the case of $q = 2$.

that it is possible to compare these level curves with numerical solution of the original system of equations.

3.3 The case of $q \geq 3$

As was said earlier, the first order term is trivial, $\mathscr{K}_1(\psi, J) = 0$. The second order term is independent[5] of ψ and reads

$$\mathscr{K}_2(\psi, J) = -\frac{\omega_2^2}{2^4 p_\theta}\left(1-\beta^2\right)^2$$
$$\times\left[\frac{1}{\omega_1-\omega_2}{}_2F_1\left(1, 1-\frac{\omega_2}{\omega_1}, 2-\frac{\omega_2}{\omega_1}, \beta^2\right) + \frac{1}{\omega_1+\omega_2}{}_2F_1\left(1, 1+\frac{\omega_2}{\omega_1}, 2+\frac{\omega_2}{\omega_1}, \beta^2\right)\right].$$

Therefore, the second order von Zeipel's Hamiltonian is given by

$$\mathscr{K}(\psi, J) = \omega \cdot J + \varepsilon^2\mathscr{K}_2(J).$$

The equations of motion for this Hamiltonian are trivial and can be easily integrated. Note that this was not possible in the preceding cases. The solution is simply

$$\psi(t) = \partial_J\mathscr{K}(\psi(0), J(0))t + \psi(0), \quad J(t) = J(0).$$

It follows that in this case the original action I is bounded and no resonance appear.

[5]It can be shown, that if $\omega_2/\omega_1 = p/q$, then the first term which depends on ψ is the q-th one. Therefore the slow evolution of action coordinates is negligible with increasing q.

4 Summary

Let us conclude that with the aid of von Zeipel's method it was shown that in case of simple sinusoidal flux the resonant behaviour can be observed only if the ratio of the flux frequency and the strength of the field is a natural number, i.e. $\Omega/b \in \mathbb{N}$. By resonance we mean here that the motion of the particle will be exactly as described in the end of the Section 1, in particular that for any initial conditions the set $\{r(t)\}_{t\geq 0} \subset \mathbb{R}$ has 0 and $+\infty$ as accumulation points.

5 Acknowledgement

The author greatly appreciates the support by the project of the Grant Agency of the Czech Republic No. 202/08/H072.

References

[1] V. Arnold. *Mathematical Methods of Classical Mechanics.* Springer-Verlag, (1989).

[2] V. Arnold, V. Kozlov, and A. Neishtadt. *Mathematical Aspects of Classical and Celestial Mechanics.* Encyclopaedia of Mathematical Sciences. Springer-Verlag, (2006).

[3] G. Giacaglia. *Perturbation Methods in Non-Linear Systems.* Applied Mathematical Science. Springer-Verlag, (1972).

[4] T. Kalvoda. Classical particle and time-periodic Aharonov-Bohm flux. In 'Doktorandské dny 2007', Prague, (2007).

[5] J. A. Sanders and F. Verhulst. *Averaging Methods in Nonlinear Dynamical Systems.* Springer-Verlag, (1985).

Towards Real Prediction of Bone Adaptation*

Václav Klika

2nd year of PGS, email: klika@it.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: František Maršík, Institute of Thermomechanics, AS CR

Abstract. Bone remodelling model that we formulated in previous years ago went through small modifications recently to better describe the bone renewal phenomenon. A rather large recherche was carried out to determine the model parameters to reach not only qualitative but also quantitative results. A great advantage of presented model is that all the parameters are real and measurable and thus by thorough search in literature we were able to set almost all of them. The remaining were obtained as a solution of nonlinear programming problem. As a consequence the model could be used for predictions on the tissue level.

Abstrakt. Model pro remodelaci kostí, který jsme formulovali v předchozích letech, prošel menšími úpravami, aby přesněji popisoval jev znovuobnovy kosti. Provedli jsme poměrně značnou rešerši, abychom určily parametry modelu, což umožní dosahovat nejen kvalitativních ale i kvantitativních výsledků. Velikou výhodou prezentovaného modelu je reálnost a měřitelnost jeho parametrů, a tedy pomocí důkladného prozkoumání dostupné literatury jsme jich byli schopni nastavit většinu. Zbylé byly získány jakožto řešení úlohy nelineárního programování. Nyní může být model používán pro predikci na tkáňové úrovni.

1 Introduction

Bone is a living tissue that is constantly being renewed. The cells that participate in the process are the osteoblasts(bone forming), osteoclasts(bone dissolving), and osteoclasts(bone cells). They form a temporary anatomical structure, called basic multicellular units, that carry out the remodelling process. A number of factors affect bone turnover, including hormones, cytokines, and mechanical stimuli. Mechanical loading is believed to be of very high significance as a stimulus for bone cells, which ensures proper bone strength and prevents high bone loss with age.

Bone remodelling also repairs an accumulated damage from everyday loading by renewing the tissue, plays an important role in metabolism since bone is used as a reservoir of many minerals (e.g calcium, potassium) and hormones (e.g. parathyroid hormone PTH) and remodelling process is a way to access these storages.

In our approach, we describe the mentioned phenomenon using the following stoichiometric equations:

$$RR + MCELL \overset{k_{\pm 1}}{\rightleftarrows} MNOC + N_4, \tag{1}$$

*This research has been supported by the Czech Science Foundation project no. 106/08/0557, by Research Plan No. AV0Z20760514 of the Institute of Thermomechanics AS CR, and by Research Plan MSM 6840770010 'Applied Mathematics in Technical and Physical Sciences' of the Ministry of Education, Youth and Sports of the Czech Republic.

where RR are ligand-receptor (RANKL-RANK) bounds between OB and $MCELL$ (precursor of osteoclast) that are needed to enable formation of multinucleated osteoclasts $MNOC$ from mononuclear cells ($MCELL$) [16]. N_4 is a remaining product from reaction (1). Bone decomposition can be characterised by following chemical reaction:

$$MNOC + Old\_B \overset{k_{\pm 2}}{\rightleftarrows} N_6 + LF, \tag{2}$$

where Old_B denotes old bone. During resorption, the osteoclasts release local factors LF (mainly growth factors) from bone, which play role in activation of osteoblasts OB [4].

$$LF + osteoprogenitor \overset{k_{\pm 3}}{\rightleftarrows} OB + N_{10} \tag{3}$$

$$OB + N_{11} \overset{k_{\pm 4}}{\rightleftarrows} Osteoid + N_{13}, \tag{4}$$

where $N_{10},\ N_{13}$ are remaining substratum. The longest period in bone remodelling process pertains to mineralisation (depositing calcium, etc. – N_{14} – into matrix) of osteoid. Ossification of *osteoid* (the primary ossification) into new bone tissue may be characterised by:

$$N_{14} + Osteoid \overset{k_{\pm 5}}{\rightleftarrows} New\_B + N_{16}, \tag{5}$$

where New_B denotes new bone formed by remodelling process and N_{16} is the residuum of bone formation reaction.

Kinetics of the above mentioned processes is governed by the following system of ordinary differential equations (obtained from *law of active mass*; for more details see some of our previous work - e.g. [8, 9, 10]):

$$\frac{\partial n_{MCELL}}{\partial \tau} = -\delta_1(\beta_1 + n_{MCELL})n_{MCELL} + \mathcal{J}_3 + \mathcal{J}_{New\_B} - \mathcal{D}_1 \tag{6}$$

$$\frac{\partial n_{Old\_B}}{\partial \tau} = -(\beta_3 - n_{MCELL} + n_{Old\_B})n_{Old\_B} - \mathcal{D}_2 + \mathcal{J}_{New\_B} \tag{7}$$

$$\begin{aligned}\frac{\partial n_{OB}}{\partial \tau} &= \delta_3(\beta_6 - n_{Old\_B} - (n_{OB} + n_{Osteoid} + n_{New\_B}))\cdot \\ &\quad \cdot(\beta_8 - (n_{OB} + n_{Osteoid} + n_{New\_B})) - \\ &\quad - \delta_4(\beta_{11} - (n_{Osteoid} + n_{New\_B}))n_{OB} + \mathcal{D}_3 - \mathcal{D}_4\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial n_{Osteoid}}{\partial \tau} &= \delta_4(\beta_{11} - (n_{Osteoid} + n_{New\_B}))n_{OB} - \\ &\quad - \delta_5(\beta_{14} - n_{New\_B})n_{Osteoid} + \mathcal{D}_4 - \mathcal{D}_5\end{aligned} \tag{9}$$

$$\frac{\partial n_{New\_B}}{\partial \tau} = \delta_5(\beta_{14} - n_{New\_B})n_{Osteoid} - \mathcal{J}_{New\_B} + \mathcal{D}_5, \tag{10}$$

where index i relates to substances and index α to reactions, β_i is a sum of normalised initial molar concentrations of relevant substances, δ_α relate to chemical reaction rate, parameter D_α describes the influence of dynamic loading on chemical reactions, and n_i is a normalised concentration of i-th substance.

2 Parameters setting

It is very important to know stationary solution of dynamic system (6)-(10) because (if stable) it gives us some idea about solution of ODEs and necessary conditions for

parameters may be derived. Because (6)-(10) describe evolution of normalised molar concentrations, it is needed to ensure that the solution is positive for all $t > 0$. Moreover, appropriate linear combinations of solution which represent all the other involved substances need to be positive too. There is just one positive stationary solution satisfying these necessary conditions.

All used parameters in this model are realistic and measurable. Unfortunately, we do not have nowadays enough knowledge for precise identification of all of them. However, we can perform reasonable estimation based on experiments and nowadays knowledge of the process found in literature. Firstly, the parameters δ_α will be determined. Since ODEs (6)-(10) are in dimensionless form, these parameters representing chemical reaction rate can be assigned just from ratio of reaction rates:

$$\delta_\alpha = \frac{k_{+\alpha}}{k_{+2}}\ [1], \quad k_{+2} \ldots \text{chemical reaction rate of } 2^{\text{nd}} \text{ reaction} \tag{11}$$

In literature, we may find that resorption carried out by $MNOC$ (second reaction (2)) lasts 20 days [16]. Further, the reversal phase (third reaction (3)) lasts approximately 9-10 days [4, 16]. The osteoid production by OB is the longest part of BR process and it lasts 90-140 days [16, 4]. Consecutive mineralisation is almost never ending but the primary ossification, which completes the formation of new bone from osteoid, has time span similar to osteoid formation [18]. Time needed for the creation of active resorbing osteoclasts ($MNOC$s) by merging osteoclast precursors ($MCELL$) was not found in literature but it can be assumed that it is much faster than the previous mentioned reactions. We postulate it to be one hour. In total, we have:

$$\begin{gathered}\delta_1 = \frac{k_{+1}}{k_{+2}} = \frac{T_2}{T_1} = \frac{20\text{days}}{1\text{h}} \doteq 480, \quad \delta_3 = \frac{k_{+3}}{k_{+2}} = \frac{T_2}{T_3} = \frac{20\text{days}}{9\text{days}} \doteq 2, \\ \delta_4 = \frac{k_{+4}}{k_{+2}} = \frac{T_2}{T_4} = \frac{20\text{days}}{140\text{days}} = \frac{1}{7} \doteq \delta_5,\end{gathered} \tag{12}$$

For further parameter setting we need to estimate resorption rate of bone (Old_B). Kanehisa [7] states that a single $MNOC$ resorbs $43\mu m^3$ to $1225\mu m^3$ of bone per hour with mean value $390\mu m^3/hr$, which will be used in following considerations. To obtain total resorption rate in bone, an estimation of total active $MNOC$ in body is needed. In typical BMU (basic multicellular unit - [2, 12]) there are 9 $MNOC$ 'at the front' of cutting cone and approximately 2000 OBs at the end [17]. We may verify this quantity of $MNOC$ present in BMU: Eriksen states that typical osteoclast ($MNOC$) diameter is 50 μm [1]. Thus really 9 or 10 $MNOC$s fill the front line of 'cutting cone' [12] with diameter of $200 - 250\ \mu m$. Further, Manolagas states that 1 milion BMU operates at any moment in body [11]. If we use these data, the resorption in human body per hour is

$$390 \cdot 9 \cdot 10^6 \mu m^3/h \doteq 3.5 mm^3/h. \tag{13}$$

In other words, the whole skeleton which has a volume of $1.75\ 10^6 mm^3$ [6] would be resorbed in $1.75\ 10^6/3.5 = 5\ 10^5 h \doteq 57$years. On the other hand, it is often stated that bone remodels once every 5-7 years. From here it is apparent that it is needed to modify the assertion of Manolagas and state that approximately 10^7 BMU operates at any moment in body instead of 10^6.

Now it is possible to determine concentration of $MNOC$, OB, and Old_B (=osteocytes - OCy) in human bone

$$[\mathrm{MNOC}] = \frac{9\ 10^7}{\mathrm{N_A}}\frac{1}{1.75}\frac{mol}{l} \doteq 5.14\frac{10^7}{\mathrm{N_A}}\frac{mol}{l}, \tag{14}$$

$$[\mathrm{OB}] = \frac{2000\ 10^7}{\mathrm{N_A}}\frac{1}{1.75}\frac{mol}{l} \doteq \frac{10^{10}}{\mathrm{N_A}}\frac{mol}{l}, \tag{15}$$

where $\mathrm{N_A}$ represents Avogadro's number. As was mentioned, ODEs (6)-(10) describing BR process are in dimensionless form which is very usefull for mathematical analysis. One consequence, of course, is that all concentration are normalised with respect to concentration of bone tissue - osteocytes. It is often stated that amount of $MNOC$ together with OB are around 5% in human bone. Robling mentions that the ratio of [OCy] to ([OB] + [MNOC]) is around 20 [16], which also supports our belief that [OCy] is determining for bone tissue concentration.

Correct estimation of chemical reaction rate k_{+2} is crucial for finding relation between real time t and computational time $\tau = t\ k_{+2}[\mathrm{Bo}]$, where [Bo] is initial concentration of bone tissue which is used for normalisation, i.e. $[\mathrm{Bo}] = [\mathrm{Old\_B_{ini}}] + [\mathrm{New\_B_{ini}}]$. From second reaction (2)

$$\frac{[\mathrm{Old\_B}]}{[\mathrm{Old\_B_{ini}}]} = \exp(-k_{+2}[\mathrm{MNOC}] \triangle\ t), \tag{16}$$

where $[\mathrm{Old\_B_{ini}}]$ is the initial concentration of old bone at time $t = 0$ and [Old_B] at time $t = \triangle\ t$. To set the k_{+2} parameter, it is needed to calculate the concentration change of old bone in time caused by $MNOC$. It was already mentioned that 1 $MNOC$ dissolves $390\mu m^3/h$ of bone tissue in average. Since

$$[\mathrm{Old\_B}] = [\mathrm{OCy}] \doteq 20[\mathrm{OB}], \tag{17}$$

we have

$$[\mathrm{Old\_B}] = 20\frac{10^{10}}{\mathrm{N_A}}\frac{mol}{l} = 2\ 10^{-4}\frac{1}{\mathrm{N_A}}\frac{mol}{\mu m^3}, \tag{18}$$

which means that there is approximately 1 OCy in every $5000\mu m^3$. To verify this number as well as previous estimates, we will calculate an average distance between OCy: $\mathrm{dist} = \sqrt[3]{5000} - \sqrt[3]{3}\sqrt[3]{5000} \doteq 17 - 29\mu m$, which is in very good correlation with Sugawara observation: $24.1 \pm 2.8\mu m$ [19]. In total, 1 $MNOC$ dissolves $3.9\ 10^2 \cdot 2\ 10^{-4} = 7.8\ 10^{-2}$ of OCy per hour, i.e. $\frac{7.8\ 10^{-2}}{3.6\ 10^3} = 2.17\ 10^{-5}\ OCy$ per 1s. We may (without loss of generality) further assume that this rate is independent on $MNOC$ concentration, i.e. noncompeting. Finally, the time change of [Old_B] decreases every second in following manner (using eq (14)):

$$\frac{\triangle\ [\mathrm{Old\_B}]}{\triangle\ t = 1s} = \frac{\triangle\ [\mathrm{OCy}]}{1s} = \frac{5.14\ 10^7 \cdot 2.17\ 10^{-5}}{\mathrm{N_A}} \doteq \frac{1.12\ 10^3}{\mathrm{N_A}}\frac{mol}{ls}, \tag{19}$$

and thus the value k_{+2} satisfies (relation (16) used)

$$\begin{aligned}\frac{[\mathrm{Old\_B}]}{[\mathrm{Old\_B_{ini}}]} = \frac{[\mathrm{Old\_B_{ini}}] - \triangle\ [\mathrm{Old\_B}]}{[\mathrm{Old\_B_{ini}}]} = \frac{20\ 10^{10} - 1.12\ 10^3}{20\ 10^{10}}\frac{\mathrm{N_A}}{\mathrm{N_A}} = 1 - \frac{5.6}{10^9} = \\ = \exp(-k_{+2} \cdot \frac{5.14\ 10^7}{\mathrm{N_A}}) = \sum_{n=0}^{\infty}\frac{\left(-k_{+2} \cdot \frac{5.14\ 10^7}{\mathrm{N_A}}\right)^n}{n!} \approx 1 - k_{+2} \cdot \frac{5.14\ 10^7}{\mathrm{N_A}}.\end{aligned} \tag{20}$$

By solving the last equation we obtain

$$k_{+2} = \frac{5.6}{5.14}\frac{\mathrm{N_A}}{10^{16}} \doteq 6.5\ 10^7 \frac{mol}{l}. \tag{21}$$

Interestingly, we may infer this value from a very different point of view: let us assume that OCy are located so that they are 'tuned' together to communicate. Speed of wave propagation in bone (fluid) is around 2900 m/s [14] and when realizing that typical distance between OCy is $20\mu m$ (which corresponds to $\lambda/2$) we have

$$f = \frac{2900m/s}{2 \cdot 20\ 10^{-6}m} \Rightarrow T = \frac{2\pi}{f} = \frac{8\pi}{29}10^{-7}s.$$

Since concentration changes with time proportionally to concentration (with coefficient $k_{+\alpha}$) we have

$$\frac{\mathrm{d}c}{\mathrm{d}t} = -k_{+\alpha}c \Rightarrow \frac{c(t)}{\mathrm{c_0}} = \exp(-k_{+\alpha}t) \tag{22}$$

¿From here it can be seen that $\frac{1}{k_{+\alpha}}$ equals to characteristic time, and it may be summarised from (22) that $k_{+\alpha} \sim \frac{1}{T} \doteq \frac{29}{8\pi}10^7 \doteq 1.15\ 10^7$. If we compare these two estimates, we see that they are closely related. It would be interesting to test the second hypothesis - to see whether the distance among OCy is so crucial for proper mechanosensing/function of bone adaptation.

Knowledge of the k_{+2} value enables us to find the relation between computational time τ and real time t

$$\tau = k_{+2}[\mathrm{Bo}]t = 6.5\ 10^7 \cdot \frac{20\ 10^{10}}{\mathrm{N_A}}t \doteq 2\ 10^{-5}t. \tag{23}$$

Useful information for further parameter setting is to know the time τ equivalent to 1 day and duration of BR cycle($1h + 20d + 9d + 140d + 140d \approx 310d$):

$$\tau_{day} = 2\ 10^{-5} \cdot 24 \cdot (60)^2 \doteq 1.7 \quad \tau_{BR} = \tau_{day} \cdot 310 = 527. \tag{24}$$

BR creates a new bone after 310 day by replacing old bone tissue. This new bone tissue, as it is called, is a regular bone tissue that has just been recently formed and has smaller mineral content since the secondary ossification has not started yet. Nevertheless it can be remodelled if needed. The model has the same features - it creates a new bone tissue which is transformed into Old_B. This transformation is realized by fluxes of particular substances - outflow of New_B ($\mathcal{J}_{New\_B}$) and inflow of Old_B ($\mathcal{J}_{Old\_B}$). In the model it holds:

$$\mathcal{J}_{Old\_B} = \mathcal{J}_{New\_B} \tag{25}$$

which guarantees us that 1 mol of New_B is changed into 1 mol of Old_B (actually, we may now rename Old_B simply into $Bone$ and New_B into $formationindex$ because new bone New_B after being formed is changed into Old_B which then represents total amount of bone). Now, we will calculate the value of $\mathcal{J}_{Old\_B}$. We know that 1 $MNOC$ resorbs $390\mu m^3/h$ and in whole skeleton there are 9 10^7 $MNOC$s which means that 3.5 $10^{10}\mu m^3/h = 35mm^3/h$ of bone tissue is removed. Because bone tissue is mostly

in equilibrium (resorption is balanced with formation), we may assume that the same amount of bone is produced and resorbed:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}}[\mathrm{Old\_B}] = \mathrm{resorbtion} + \mathrm{J_{Old\_B}} \overset{equilib}{=} 0 \Rightarrow \\ \Rightarrow \mathrm{J_{Old\_B}} = \mathrm{resorbtion} = 35mm^3/h \doteq 10^7 \mu m^3/s, \end{aligned} \tag{26}$$

and thus amount of bone resorbed per second in mols is

$$\frac{\triangle\ \#Old\_B}{\triangle\ t} = 10^7 \cdot 2\ 10^{-4}/\mathrm{N_A} = 2\ 10^3/\mathrm{N_A} mol \cdot s^{-1}, \tag{27}$$

where equation (18) was used and when realizing that skeleton has volume of $2l$ we may conclude

$$\frac{\triangle\ [\mathrm{Old\_B}]}{\triangle\ t} = \frac{2\ 10^3}{2\mathrm{N_A}} = \frac{10^3}{\mathrm{N_A}} \frac{mol}{l} = \mathrm{J_{Old\_B}} \tag{28}$$

$$\Rightarrow \mathcal{J}_{\mathrm{Old\_B}} = \mathrm{J_{Old\_B}} \cdot \frac{1}{k_{+2}[\mathrm{Bo}]^2} \doteq 3\ 10^{-4}. \tag{29}$$

Bone remodelling is a very long process. Cells taking part in it must be several times replaced. This fact is actually exploited by body itself as a control mechanism - e.g. estrogen promotes osteoclast apoptosis [17]. Apoptosis of $MNOC$ plays substantial role since its mean life in vivo is 3 days [20]. Using this knowledge we may determine $\mathrm{J_3}$(=negative flux of MNOC=MNOC apoptosis) analogically to (28):

$$\text{decrease of } [\mathrm{MNOC}]/s = \frac{\#MNOC}{volume \cdot time} = \frac{9\ 10^7}{2l \cdot 3 \cdot 24h} = \frac{1.7\ 10^2}{\mathrm{N_A}} \frac{mol}{l \cdot s} \tag{30}$$

$$\Rightarrow \frac{\mathcal{J}_3}{\mathcal{J}_{New\_B}} = \frac{1.7\ 10^2}{\mathrm{N_A}} \cdot \frac{\mathrm{N_A}}{10^3} = \frac{1}{6}. \tag{31}$$

Another family of parameters - β_i - are determined by sum of normalised initial concentrations of appropriate substances

$$\begin{aligned} \beta_6 = \frac{[\mathrm{Old\_B}_{ini}] + [\mathrm{New\_B}_{ini}] + [\mathrm{Osteoid}_{ini}] + [\mathrm{OB}_{ini}] + [\mathrm{N_{6}}_{ini}]}{[\mathrm{Bo}] = [\mathrm{Old\_B}_{ini}] + [\mathrm{New\_B}_{ini}]} \doteq \\ \doteq 1 + \frac{[\mathrm{Osteoid}_{ini}] + [\mathrm{OB}_{ini}] + [\mathrm{N_{6}}_{ini}]}{[\mathrm{Old\_B}_{ini}]} = 1 + \frac{0 + 1/20[\mathrm{Old\_B}_{ini}] + 0}{[\mathrm{Old\_B}_{ini}]} = 1.05, \end{aligned} \tag{32}$$

where relation between OB and OCy was used and a consideration that remaining product (N_6) and osteoid are not present in given volume when BR is initiated. Similarly

$$\beta_8 = \frac{[\mathrm{osteoprogenitor}_{ini}] + [\mathrm{New\_B}_{ini}] + [\mathrm{Osteoid}_{ini}] + [\mathrm{OB}_{ini}]}{[\mathrm{Bo}]} = \frac{1}{10}, \tag{33}$$

$$\beta_{14} = \frac{[\mathrm{New\_B}_{ini}] + [\mathrm{N_{14}}_{ini}]}{[\mathrm{Bo}]} = \frac{1}{20}, \tag{34}$$

$$\beta_{11} = \frac{[\mathrm{New\_B}_{ini}] + [\mathrm{Osteoid}_{ini}] + [\mathrm{N_{11}}_{ini}]}{[\mathrm{Bo}]} = 0.7, \tag{35}$$

$$\beta_1 = \frac{[\mathrm{RR}_{ini}] - [\mathrm{MCELL}_{ini}]}{[\mathrm{Bo}]} = \frac{1}{4}, \tag{36}$$

$$\beta_3 = \frac{[\mathrm{MNOC}_{ini}] - [\mathrm{Old\_B}_{ini}] + [\mathrm{MCELL}_{ini}]}{[\mathrm{Bo}]} = 0. \tag{37}$$

Last group of parameters, $\mathcal{D}_\alpha$, describes the effect of dynamic loading on rate of chemical reactions:

$$\mathcal{D}_\alpha = \frac{l_{\alpha v} d_{(1)}}{k_{+2}[\mathrm{Bo}]^2}[1], \tag{38}$$

$$r_\alpha = l_{\alpha v} d_{(1)} + l_{\alpha\alpha}\mathcal{A}_\alpha \tag{39}$$

where $d_{(1)} = divv = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3} = -\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}t}$ is the trace of the deformation rate tensor, ρ is concentration of material, r_α and $\mathcal{A}_\alpha$ is a chemical reaction rate and affinity of the $\alpha - th$ reaction, respectively. In case of thermodynamic equilibrium, all quantities depend on equilibrial values $(T, e_{eq}, [N_i]_{eq})$. When the system is deflected from equilibrium, they may be described using the following linear relations (Curie-Prigogine principle - linear nonequilibrium thermodynamics [13]):

$$r_\alpha = l_{\alpha\ v} d_{(1)} + l_{\alpha\ \alpha}\mathcal{A}_\alpha, \tag{40}$$

$$p_\alpha = l_{v\ \alpha} d_{(1)} + l_{v\ v}\mathcal{A}_\alpha, \tag{41}$$

where $l_{v\ v}, l_{\alpha\ v} = l_{v\ \alpha}, l_{\alpha\ \alpha}$ are functions of temperature T, and the invariants of a strain rate tensor d_{ij}, i.e. $d_{(1)}$ - volume rate, $d_{(2)}$ - shear rate. We assume that the process is isothermic (body temperature), further that the linear dependence in equation 40 is sufficient to describe the dependence on $d_{(1)}$, and that the influence of shear rate is constant, i.e. we assume that shape and size of canaliculi, lacunas, and osteocytes in bone does not change noticeably. Elastic deformations of canaliculi and lacunas induce both their volume deformation and shear bone fluid flow past osteocytes. The measure of these stimulations is summarised in the phenomenological coefficients $l_{v\ v}, l_{\alpha\ v}, l_{\alpha\ \alpha}$ which can be patient (genetically) dependent.

As can be seen from (38), we need to determine the influence of mechanical loading on each chemical reaction. The unknown parameters $l_{\alpha v}$ were calculated as a solution of constraint extremum problem (or minimisation problem of appropriate functional) for unknowns $l_{\alpha v}$:

$$[\mathrm{OB}]\% = 4.5\% \tag{42}$$

with constraints:

$$\frac{\rho_{\max}}{\rho_{\min}} = 20 \tag{43}$$

$$[\mathrm{MNOC}]\% = 0.023\% \Longleftrightarrow \frac{[\mathrm{OB}]}{[\mathrm{MNOC}]} = \frac{2000}{9} \doteq 200 \tag{44}$$

and conditions describing that concentrations of all substances in stationary solution are positive. We know that spongy bone is located in part of bone which experiences smaller deformations/strains, and conversely cortical bone creates weightbearing support on outer cortex. We used this fact for setting $l_{\alpha v}$ so that the maximal (found in cortical bone - the properly loaded case) and the minimal apparent density (found in spongy bone - unloaded case leads to minimal density) in stationary state are in the following relations:

$$\rho = \rho(d_{(1)}) \quad \Rightarrow \quad \frac{\rho_{\max}}{\rho_{\min}} = \frac{\rho_{\mathrm{cort,\ max}}}{\rho_{\mathrm{spongy,\ min}}} = \frac{2.0\ g/cm^3}{0.1\ g/cm^3} = 20 \tag{45}$$

which guarantees us correct range of apparent bone density [5, 15, 3]. The equation (42) ensures that the percentage of osteoblasts OB in a stationary state will be 4.5% and similarly the relation (44) ensures the correct $MNOC$ percentage.

Even if all stationary solutions are positive, it is still needed to check whether all the concentrations of substances are positive for all time $t > 0$. $\mathcal{D}_\alpha$ parameters that solve above mentioned constraint extremum problem and also satisfies conditions from previous sentence are listed in table 46.

$$\begin{gathered}
\delta_1 = 480, \quad \delta_3 = 2, \quad \delta_4 = \tfrac{1}{7} = \delta_5 \\
\beta_1 = -0.5, \quad \beta_3 = 0.4, \quad \beta_6 = 1.02, \quad \beta_8 = \tfrac{1}{10}, \quad \beta_{11} = \tfrac{1}{3}, \quad \beta_{14} = \tfrac{1}{20} \\
\mathcal{J}_{Old_B} = \mathcal{J}_{New_B} = 3\ 10^{-4}, \quad \mathcal{J}_3 = \tfrac{1}{6}\mathcal{J}_5, \\
l_{1v} = -8.96\ 10^{-13}\tfrac{mol}{l}, \quad l_{2v} = 4.1\ 10^{-19}\tfrac{mol}{l}, \quad l_{3v} = -2.82\ 10^{-17}\tfrac{mol}{l}, \\
l_{4v} = -7.91\ 10^{-18}\tfrac{mol}{l}, \quad l_{5v} = 5.91\ 10^{-19}\tfrac{mol}{l}.
\end{gathered} \tag{46}$$

3 Discussion and Conclusion

The bone remodelling process together with its control is still not fully understood even if there has been a great step forward in last decade, especially on the cellular level. It is very important to be able to predict response of bone to varying condition - both mechanical (e.g. joint implants) and biological (e.g. hormonal) changes. Models that are nowadays used for simulation of BR are still not sufficient.

The model here presented combines both the mechanical stimuli and biochemical control. With current settings of parameter the model has all the following features that already describe the bone remodelling process to reasonable extent:

- realistic and measurable model paramteres
- positiveness of all molar concentration of involved substances
- unique positive stationary solution
- rate of chemical reactions
- resorption rate of bone
- number of active BMU (active remodelling foci)
- molar concentrations: [MNOC], [OB], [OCy]
- relation between time scales(computational and real time)
- 1 mol of New_B transforms into 1 mol of Old_B (mass may differ)
- MNOC apoptosis (mean life in vivo is 3 days; compare to BR)
- initial concentration of involved substances
- the influence of mechanic stimuli on reaction rates-determined by solving the minimalisation of appropriate functional with constraints such as $\frac{\rho_{\max}}{\rho_{\min}} = 20$.

We are about to start using the presented model for predicting bone adaptation in humans and use the results for further verification.

Acknowledgement
I would like to thank to my supervisor Professor F. Maršík, Eng., D.Sc. for gratuitous passing of know-how, for advice, pleasant consultations, and help with writing this article.

This research has been supported by the Czech Science Foundation project number 106/08/0557, by Research Plan No. AV0Z20760514 of the Institute of Thermomechanics AS CR, and by Research Plan MSM 6840770010 'Applied Mathematics in Technical and Physical Sciences' of the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] E. F. Eriksen and M. Kassem. *The cellular basis of bone remodelling.* Triangle **31** (1992), 45–57.

[2] H. M. Frost. *Tetracycline-base histological analysis of bone remodelling.* Calcif Tissue Res (1969), 211–237.

[3] B. Helgason, E. Perilli, E. Schileo, and F. Taddei. *Mathematical relationships between bone density and mechanical properties: A literature review.* Clinical Biomechanics **23** (2008), 135–146. doi:10.1016/j.clinbiomech.2007.08.024.

[4] P. A. Hill. *Bone remodelling.* British Journal of Orthodontics **25** (1998), 101–107.

[5] R. Hodgskinson and J. D. Currey. *Youngś modulus, density and material properties in cancellous bone over a large density range.* Journal of Materials Science: Materials in Medicine **3** (1992), 377–381. doi:10.1007/BF00705371.

[6] W. S. S. Jee. *The skeletal tissues.* (1983). In: Weiss, L (ed) Histology: cell and tissue biology 5th ed.

[7] J. Kanehisa and J. N. Heersche. *Osteoclastic bone resorption: in vitro analysis of the rate of resorption and migration of individual osteoclasts.* Bone **9** (1988), 73–79.

[8] V. Klika. Mathematical and numerical analysis of differential equations of bone remodelling. Master thesis, Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Prague, (May 2006).

[9] V. Klika. Thermodynamic model of bone adaptation. In 'Doktorandské dny 2007', volume 2, 93–104. ČVUT, (2007). ISBN 978-80-01-03913-7.

[10] V. Klika, F. Maršík, and P. Barsa. *Remodelling of a living bone - numerical simulation.* Locomotor System **14** (2007), 112–117.

[11] S. C. Manolagas. *Editorial:cell number versus cell vigor-what really matters to a regenerating skeleton?* Endocrinology **140** (1999), 4377–4381.

[12] A. Parfitt. *Osteonal and hemi-osteonal remodeling: the spatial and temporal framework for signal traffic in adult human bone.* Journal of cellular biochemistry (1994), 273–286.

[13] I. Prigogine. *Étude Thermodynamique des Processes Irreversibles.* Desoer, Liege, (1947).

[14] J. Y. Rho, R. B. Ashman, and C. H. Turner. *Youngś modulus of trabecular and cortical bone material: ultrasonic and microtensile measurements.* Journal of Biomechanics **26** (1993), 111–119.

[15] J. C. Rice, S. C. Cowin, and J. A. Bowman. *On the dependence of the elasticity and strenght of cancellous bone on apparent density.* Journal of Biomechanics **21** (1988), 155–168.

[16] A. G. Robling, A. B. Castillo, and C. H. Turner. *Biomechanical and molecular regulation of bone remodeling.* Annual Review of Biomedical Engineering **8** (2006), 455–498. doi:10.1146/annurev.bioeng.8.061505.095721.

[17] D. Rucker, D. A. Hanley, and R. F. Zernicke. *Response of bone to exercise and aging.* Locomotor System **9** (2002), 6–22.

[18] D. Ruffoni, P. Fratzl, P. Roschger, K. Klaushofer, and R. Weinkamer. *The bone mineralization density distribution as a fingerprint of the mineralization process.* Bone **40** (2007), 1308–1319. doi:10.1016/j.bone.2007.01.012.

[19] Y. Sugawara, H. Kamioka, T. Honjo, K. Tezuka, and T. Takano-Yamamoto. *Three-dimensional reconstruction of chick calvarial osteocytes and their cell processes using confocal microscopy.* Bone **36** (2005), 877–883. doi:10.1016/j.bone.2004.10.008.

[20] R. S. Weinstein, J.-R. Chen, C. C. Powers, S. A. Stewart, R. D. Landes, T. Bellido, R. L. Jilka, A. M. Parfitt, and S. C. Manolagas. *Promotion of osteoclast survival and antagonism of bisphosphonate-induced osteoclast apoptosis by glucocorticoids.* J Clin Invest **109** (2002), 1041–1048. doi:10.1172/JCI200214538.

Complexity of Infinite Words Associated with Non-simple Parry Numbers

Karel Klouda

1st year of PGS, email: karel@kloudak.eu
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Edita Pelantová, Department of mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU,
Christiane Frougny, LIAFA, CNRS and Université Paris 8

Abstract. The aim of our work is to compute the (factor) *complexity function* $\mathcal{C}(n)$ of the infinite word $\mathbf{u}_\beta$ associated with *β-expansions*, where β is a *non-simple Parry number*. In general it is hard to find an explicit formula for the complexity function of an infinite word $\mathbf{u}$ and it seems it holds also for the case of $\mathbf{u}_\beta$. However, we are able to find all *left special factors* that, in a certain sense, completely determine the complexity. The notion of (right) special factor was introduced by Berstel in 1980 and considerably enhanced by Cassaigne in his paper from 1997. We introduce another slight enhancement, a tool that will help us to identify all *infinite left special branches* of a fixed point of substitutions satisfying some natural assumption. Further, the knowledge of the structure of left special factors will allow us to formulate a simple sufficient and necessary condition for the complexity of $\mathbf{u}_\beta$ to be affine.

Abstrakt. Cílem naší práce je najít (faktorovou) *komplexitu* $\mathcal{C}(n)$ nekonečného slova $\mathbf{u}_\beta$ příslušného *β-rozvojům*, kde β je *nejednoduché Parryho číslo*. Obecně je často nemožné najít explicitní formuli popisující komplexitu nekonečného slova a zdá se, že to platí i pro případ, kdy jako nekončné slovo bereme $\mathbf{u}_\beta$. Přesto se dá docílit alespoň nepřímého popisu faktorové komplexity a to za pomoci *levých speciálních faktorů*, které v jistém smyslu komplexitu zcela určují. Metoda výpočtu komplexity pomocí (pravých) speciálních faktorů byla poprvé uvedena v Berstelově článku v roce 1980 a významně rozvinuta v Cassaigneho článku z roku 1997. Výsledkem naší práce je pak další rozšíření, které umožňuje nalézt všechny *levé nekonečné speciální větve* pro pevné body substitucí splňujících poměrně obecné předpoklady. Dalším důležitým výsledkem je, mimo nalezení všech levých speciálních faktorů slova $\mathbf{u}_\beta$, také jednoduchá formulace nutné a postačující podmínky pro to, aby komplexita $\mathbf{u}_\beta$ byla afinní funkcí.

1 Introduction

Generally speaking, our aim is to understand the combinatorial structure of aperiodic infinite words over a finite alphabet. In particular, we are interested in the word $\mathbf{u}_\beta$ associated with β-numeration, where β is a non-simple Parry number. In order to be able to better explain the problem, we need some basic notation.

Definition 1. Let $\mathcal{A} = \{0, 1, \ldots, q-1\}, q \geq 1$ be a finite alphabet. An *infinite word* over the alphabet $\mathcal{A}$ is a sequence $\mathbf{u} = (u_i)_{i\geq 1}$ where $u_i \in \mathcal{A}$ for all $i \geq 1$. If $v = u_j u_{j+1} \cdots u_{j+n-1}$, $j, n \geq 1$, then v is said to be a *factor* of $\mathbf{u}$ of length n and the index j is an *occurrence* of v, ϵ is the factor of length 0.

By $\mathcal{L}_n(\mathbf{u})$ we denote the set of all factors of $\mathbf{u}$ of length $n \in \mathbb{N}$, the *language* of $\mathbf{u}$ is then the set $\mathcal{L}(\mathbf{u}) = \bigcup_{n\in\mathbb{N}} \mathcal{L}_n(\mathbf{u})$.

Definition 2. An infinite word $\mathbf{u}$ is said to be *eventually periodic* if $\mathbf{u} = v_1v_2v_2v_2\cdots = v_1(v_2)^\omega$, where v_1 and v_2 are finite words and v_2 is non-empty. If $\mathbf{u}$ is not eventually periodic, then it is *aperiodic*.

From our point of view, eventually periodic words are not interesting as their structure is completely described by the simple finite rule. No such a rule exists in the case of aperiodic words, therefor some tools how to measure their irregularity have been proposed. One of such basic tools is the (factor) complexity function $\mathcal{C} : \mathbb{N} \to \mathbb{N}$, which counts the number of factors of a given length, i.e.

$$\mathcal{C}(n) = \#\mathcal{L}_n(\mathbf{u}),$$

where $\#A$ is the number of elements of a set A. It is easy to realize, that the complexity of $\mathbf{u}$ is bounded if and only if $\mathbf{u}$ is eventually periodic. Other known results on the complexity functions are listed in the following proposition.

Definition 3. A mapping φ which maps each letter of a finite alphabet $\mathcal{A}$ to a finite word over the alphabet is a *substitution*.

A substitution φ is *primitive* if there exists $k \in \mathbb{N}$ such that for all $a, b \in \mathcal{A}$ the word $\varphi^k(a)$ contains b.

Proposition 4. *Let* $\mathbf{u}$ *be an infinite word over a finite alphabet* $\mathcal{A}$*.*

(i) $0 \le \mathcal{C}(n) \le (\#\mathcal{A})^n$*,*

(ii) $\mathbf{u}$ *is aperiodic if and only if the first difference of the complexity function is positive, i.e.* $\triangle\mathcal{C}(n) := \mathcal{C}(n+1) - \mathcal{C}(n) \ge 1$*, for all* $n \in \mathbb{N}$*,*

(iii) if $\mathbf{u}$ *is a fixed point of a primitive substitution then* $\mathcal{C}(n)$ *is a sublinear function, i.e.,* $\mathcal{C}(n) \le an + b$*, for some* $a, b \in \mathbb{N}$*,*

(iv) if $\mathbf{u}$ *is a fixed point of primitive substitution then* $\triangle\mathcal{C}(n)$ *is bounded.*

Items (i) and (ii) are obvious, (iii) is due to [15], (iv) was proved in [13] and in a more general context in [5]. For more properties see e.g. [1].

Infinite words appears in various fields of mathematics [4]. The word $\mathbf{u}_\beta$ we are interested in has origin in the theory of non-standard numeration, namely β-numeration. For more on this topic see [12].

β-numeration is a generalization of the classical numeration, when each number is represented as a sum of powers of an integer base $b > 1$. Humans use the representation in base $b = 10$, computers use binary representation $b = 2$. For each positive real number x one can found its representation in base b using a *greedy algorithm*:

1. Find $k \in \mathbb{N}$ such that $b^k \le x < b^{k+1}$.
2. Set $x_k := \lfloor x/b^k \rfloor$ and $r_k := \{x/b^k\}$.

3. For $0 \leq i < k$, let $x_i = \lfloor br_i \rfloor$ and $r_i := \{br_{i+1}\}$.

$\lfloor x \rfloor$ is the integer part and $\{x\}$ is the fractional part of a real number x. Obviously, digits x_i takes value in the set $\{0, 1, \ldots, \lfloor b \rfloor - 1\}$. If $b > 1$ is an integer, we obtain classical representation in an integer base. If we replace b by some non-integer number $\beta > 1$, we obtain the *β-expansion* of x.

For $x \in [0, 1)$, the β-expansion can be computed also by using the piecewise linear map $T_\beta : [0, 1) \to [0, 1)$ defined as $T_\beta(x) = \{\beta x\}$. The sequence $d_\beta(x) = x_1 x_2 x_3 \cdots$ is obtained by iterating T_β with $x_i = \lfloor \beta T_\beta^{i-1}(x) \rfloor$. The difference between β-expansion and $d_\beta(x)$ arises for $x = 1$ since the *Rényi expansion of unity* $d_\beta(1)$ is not a β-expansion. Parry [14] showed that $d_\beta(1)$ plays a very important role in the theory of β-numeration. Among other things, it allows us to define Parry numbers.

Definition 5. A real number $\beta > 1$ is said to be a Parry number if $d_\beta(1)$ is eventually periodic. In particular,

a) if $d_\beta(1) = t_1 \cdots t_m$ is finite, i.e. it ends in infinitely many zeros, then β is a *simple Parry number*,

b) if it is not finite, i.e. $d_\beta(1) = t_1 \cdots t_m (t_{m+1} \cdots t_{m+p})^\omega$, then β is called a *non-simple Parry number*.

Note, that the parameters $m, p > 0$ are taken the least possible. It implies that $t_m \neq t_{m+p}$ which will be a very important fact.

As the infinite word $\mathbf{u}_\beta$ is tightly connected with a geometrical interpretation of β-integers, we first introduce β-integers along with some of their properties.

Definition 6. The real number x is a *β-integer* if the β-expansion of $|x|$ is of the form $\sum_{i=0}^{k} a_i \beta^i$. The set of all β-integers is denoted by $\mathbb{Z}_\beta$.

The definition of β-integers coincides with the definition of classical integers in the case of β in $\mathbb{Z}$. But there are several new phenomena linked with the notion of β-integers when β is not an integer. For our purposes, the most interesting difference between classical integers and β-integers is the difference in their distribution on the real line. While the classical integers are distributed equidistantly, i.e. gaps between two consequent integers are always of the same length 1, the lengths of gaps between β-integers can take their values even in an infinite set. More precisely, Thurston [16] proved the following theorem.

Theorem 7. *Let $\beta > 1$ be a real number and $d_\beta(1) = (t_i)_{i \geq 1}$. Then the length of gaps between neighbors in $\mathbb{Z}_\beta$ takes values in the set $\{\triangle_0, \triangle_1, \ldots\}$, where*

$$\triangle_i = \sum_{k \geq 1} \frac{t_{k+i}}{\beta^k}, \quad \text{for } i \in \mathbb{N}.$$

Corollary 8. *The set of lengths of gaps between neighbors in $\mathbb{Z}_\beta$ is finite if and only if β is a Parry number. Moreover, if β is a simple Parry number, i.e. $d_\beta(1) = t_1 \cdots t_m$, the set reads $\{\triangle_0, \triangle_1, \ldots \triangle_{m-1}\}$, if β is a non-simple Parry number, i.e. $d_\beta(1) = t_1 \cdots t_m (t_{m+1} \cdots t_{m+p})^\omega$, we obtain $\{\triangle_0, \triangle_1, \ldots \triangle_{m+p-1}\}$.*

Now, let us suppose that we have drawn β-integers on the real line and assume that β is a Parry number. If we read the length of gaps from zero to the right, we obtain an infinite sequence, say $\{\triangle_{i_k}\}_{k\geq 0}$. Further, if we read only indices, we obtain an infinite word over the alphabet $\{0,\dots,m-1\}$ in the case of simple Parry numbers, and over the alphabet $\{0,\dots,m+p-1\}$ in the non-simple case. The obtained infinite word is just the word $\mathbf{u}_\beta$ we are interested in. However, there exists another way to define it. Fabre [9] proved that $\mathbf{u}_\beta$ can be defined as the unique fixed point of a substitution φ_β canonically associated with a Parry number β and defined as follows.

Definition 9. For a simple Parry number β the canonical substitution φ_β over the alphabet $\mathcal{A}=\{0,1,\dots,m-1\}$ is defined by

$$\varphi_\beta(k)=\begin{cases}0^{t_{k+1}}(k+1) & \text{if } k\in\mathcal{A}\setminus\{m-1\},\\ 0^{t_m} & \text{if } k=m.\end{cases}$$

Definition 10. For a non-simple Parry number β the canonical substitution φ_β over the alphabet $\mathcal{A}=\{0,1,\dots,m+p-1\}$ is defined by

$$\varphi_\beta(k)=\begin{cases}0^{t_{k+1}}(k+1) & \text{if } k\in\mathcal{A}\setminus\{m+p-1\},\\ 0^{t_{m+p}}m & \text{if } k=m+p-1.\end{cases}$$

We see that the definition of φ_β is given by $d_\beta(1)$ and that the only difference between simple and non-simple cases appears in the image of the last letters $m-1$ and $m+p-1$. While in the simple case the last letters of images $\varphi_\beta(k), k=0,1,\dots,m-1$, are all distinct and so the images form a suffix-free code, in the non-simple case either $\varphi_\beta(m)=0^{t_m}m$ is a prefix of $\varphi_\beta(m+p-1)=0^{t_{m+p}}m$ or vice versa. As we will see later on, this property is crucial from the point of view of computing the complexity of the infinite word $\mathbf{u}_\beta$.

Definition 11. Let $\beta>1$ be a Parry number. The unique fixed point of the canonical substitution φ_β is denoted by

$$\mathbf{u}_\beta=\lim_{n\to\infty}\varphi_\beta^n(0)=\varphi_\beta^\infty(0).$$

The uniqueness of $\mathbf{u}_\beta$ follows from the definitions of φ_β, the letter 0 is the only admissible starting letter of a fixed point.

2 Special factors and factor complexity

In this section, we will recall the notion of special factors of an arbitrary infinite word and we will explain how the structure of special factors of an infinite word determines its factor complexity.

In what follows, we shall restrict ourself to those infinite words which are fixed point of some substitution φ defined over a finite alphabet $\mathcal{A}$. We shall further assume that φ is injective and primitive.

It is well known that any fixed point of a primitive substitution is uniformly recurrent, i.e. if each factor occurs infinitely many times and the gaps between its two consecutive occurrences are bounded in length. It implies that each factor is extendable both to the right and to the left.

Definition 12. Let v be a factor of $\mathbf{u}$, the set of *left extensions* of v is defined as

$$\mathrm{Lext}(v) = \{a \in \mathcal{A} \mid av \in \mathcal{L}(\mathbf{u})\}.$$

If $\#\mathrm{Lext}(v) \geq 2$, then v is said to be a *left special (LS) factor* of $\mathbf{u}$.

In the analogous way we define the set of *right extensions* $\mathrm{Rext}(\mathbf{u})$ and a *right special (RS) factor*. If v is both left and right special, then it is called *bispecial*.

The connection between (left) special factors and the complexity follows from the following reasoning. Let us suppose that $\mathcal{L}_n(\mathbf{u}) = \{v_1, \ldots, v_k\}$ and let $\mathrm{Lext}(v_i) = \{a_1^{(i)}, \ldots, a_{l_i}^{(i)}\}, l_i \geq 1, i = 1, \ldots, k$. Now, it is not difficult to realize that

$$\mathcal{L}_{n+1}(\mathbf{u}) = \{a_1^{(1)}v_1, \ldots, a_{l_1}^{(1)}v_1, a_1^{(2)}v_2, \ldots, a_{l_{k-1}}^{(k-1)}v_{k-1}, a_1^{(k)}v_k, \ldots, a_{l_k}^{(k)}v_k\},$$

i.e. by concatenating all factors of length n and all their left extensions we obtain all factors of length $n+1$. It implies that

$$\#\mathcal{L}_{n+1}(\mathbf{u}) - \#\mathcal{L}_n(\mathbf{u}) = \triangle\mathcal{C}(n) = \sum_{\substack{v \in \mathcal{L}_n(\mathbf{u}) \\ v \text{ is LS}}} (\#\mathrm{Lext}(v) - 1). \tag{1}$$

Hence, if we know all LS factors along with the number of their left extensions, we are able to evaluate the complexity $\mathcal{C}(n)$ using this formula.

2.1 Classification of LS factors

Let $a, b \in \mathrm{Lext}(v)$ be left extensions of a factor v of $\mathbf{u}$, it means that both av and bv are factors of $\mathbf{u}$. If there exists a letter $c \in \mathrm{Rext}(av) \cap \mathrm{Rext}(bv)$, we say that v can be extended to the right such that it remains LS with left extensions a, b, indeed $a, b \in \mathrm{Lext}(vc)$.

Definition 13. Let $a, b \in \mathrm{Lext}(v)$ be distinct left extensions of a LS factor v of $\mathbf{u}$. v is an (a, b)-*maximal LS factor* if $\mathrm{Rext}(av) \cap \mathrm{Rext}(bv) = \emptyset$, in words, v can not be extended to the right such that it remains LS with left extensions a, b.

It can also happen that a LS factor v with left extensions a and b is extendable to the right infinitely many times remaining LS. In this way we obtain a so-called infinite LS branch.

Definition 14. An infinite word $\mathbf{w}$ called an *infinite LS branch* of $\mathbf{u}$ if each prefix of $\mathbf{w}$ is a LS factor of $\mathbf{u}$. We put

$$\mathrm{Lext}(\mathbf{w}) = \bigcap_{v \text{ prefix of } \mathbf{w}} \mathrm{Lext}(v).$$

Proposition 15.

(i) If $\mathbf{u}$ *is eventually periodic, then there is no infinite LS branch of* $\mathbf{u}$,

(ii) if $\mathbf{u}$ *is aperiodic, then there exists at least one infinite LS branch of* $\mathbf{u}$,

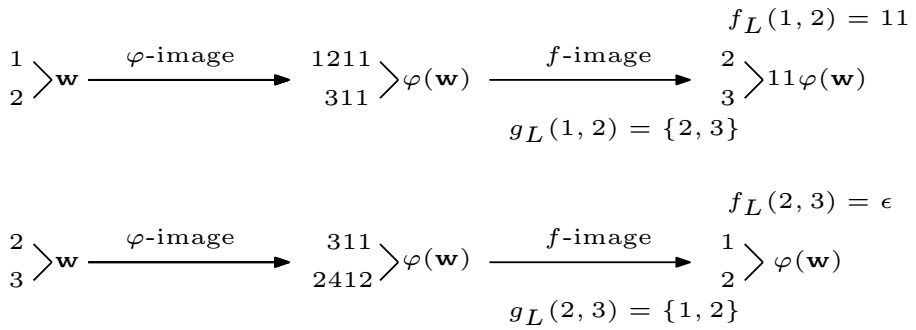


Figure 1: Images of LS factors.

(iii) if $\mathbf{u}$ *is a fixed point of a primitive substitution then the number of infinite LS branches is bounded.*

(i) is obvious, (iii) is a direct consequence of (1) and Proposition 4 (v). Item (ii) is a direct consequence of the famous König's infinity lemma [11].

Taking all together, our aim is to find all (a, b)-maximal LS factors and also all infinite LS branches of $\mathbf{u}$.

Remark 16. The term "special factor" (for us it was RS factor) was introduced in 1980 [2] and it has been used for computing the factor complexity since then (eg. [3], [8]). The notations introduced above are based on Cassaigne's article [6]. An (a, b)-maximal factor is a new term, actually it is a special case of a *weak bispecial factor* proposed there. It is also shown in the article that bispecial factors determine the second difference of the complexity in a similar way as LS factors determine the first difference of the complexity.

Remark 17. Everything what has been (and will be) defined or showed for LS factors can be defined or showed similarly for RS factors.

2.2 How to find infinite LS branches

Before introducing a new notion, let us consider for example the substitution

$$\varphi : 1 \mapsto 1211, 2 \mapsto 311, 3 \mapsto 2412, 4 \mapsto 435, 5 \mapsto 534 \tag{2}$$

with $\mathbf{u} = \varphi^\infty(1)$. Further, let w be a LS factor (or infinite LS branch) of $\mathbf{u}$ with left extensions 1 and 2. Is $\varphi(w)$ again LS factor? From Figure 1 (the first line) we see that it is not since the letter 1 is its only left extension. In order to obtain a LS factor, we have to prepend the factor 11 which is the longest common suffix of $\varphi(1) = 1211$ and $\varphi(2) = 311$, then $11\varphi(w)$ is a LS factor with left extensions 2 and 3. In the case when $\mathrm{Lext}(w) = \{2, 3\}$ (the second line in Figure 1), $\varphi(w)$ is a LS factor since the longest common suffix of $\varphi(2) = 311$ and $\varphi(3) = 2412$ is the empty word ϵ.

Definition 18. Let φ be a substitution defined over an alphabet $\mathcal{A}$. For each couple of distinct letters $a, b \in \mathcal{A}$ we define $f_L(a, b)$ as the longest common suffix of words $\varphi(a)$ and $\varphi(b)$.

Definition 19. Let φ be an injective substitution defined over an alphabet $\mathcal{A}$ having a fixed point $\mathbf{u}$. For each unordered couple of distinct letters $a, b \in \mathcal{A}$ such that $\mathrm{Rext}(a) \cap \mathrm{Rext}(b) \neq \emptyset$ we define the set $g_L(a, b)$ as follows.

(i) If $f_L(a, b)$ is a proper suffix of both $\varphi(a)$ and $\varphi(b)$, then $g_L(a, b)$ contains just the last letters of factors $\varphi(a)(f_L(a, b))^{-1}$ and $\varphi(b)(f_L(a, b))^{-1}$.

(ii) If $f_L(a, b) = \varphi(a)$ (i.e. W.L.O.G. $|\varphi(a)| < |\varphi(b)|$), then $g_L(a, b)$ contains the last letter of the factor $\varphi(b)(f_L(a, b))^{-1}$ and all the last letters of factors $\varphi(c)$, where $c \in \mathrm{Lext}(a)$ such that $\mathrm{Rext}(ca) \cap \mathrm{Rext}(b) \neq \emptyset$.

Assumption 20. A substitution φ defined over $\mathcal{A}$ is injective and it has a fixed point $\mathbf{u}$ such that for all $a, b \in \mathcal{A}$, for which $g_L(a, b)$ is defined, it holds that $\#g_L(a, b) = 2$.

Moreover, if $f_L(a, b) = \varphi(a)$ (i.e. W.L.O.G. $|\varphi(a)| < |\varphi(b)|$) and d is the last letter of the factor $\varphi(b)(f_L(a, b))^{-1}$, then for all $c \in \mathrm{Lext}(a)$ such that $\mathrm{Rext}(ca) \cap \mathrm{Rext}(b) \neq \emptyset$ it holds that d is not the last letter of $\varphi(c)$.

Assumption 20 is valid for all suffix-free substitutions since $g_L(a, b)$ from point (i) of Definition 19 contains always just two elements and the case when $f_L(a, b) = \varphi(a)$ never happens. If $f_L(a, b) = \varphi(a)$, then Assumption 20 says that if v is a LS factor with $\mathrm{Lext}(v) = \{a, b\}$, then the last letter of $\varphi(c)$ is the same for all $c \in \mathrm{Lext}(av)$ and, moreover, $d\varphi(a)$ is not a suffix of $\varphi(b)$ – in other words, for each LS factor v the factor $f_L(a, b)\varphi(v)$ is again LS. We will see that this complicated assumption is satisfied for the (not suffix-free) substitution φ_β, where β is a non-simple Parry number.

Definition 21. Let φ be a substitution satisfying Assumption 20. Then for each LS factor (or infinite LS branch) w having distinct left extensions a and b we define *f-image* of w as the factor $f_L(a, b)\varphi(w)$.

With respect to the preceding discussion, Assumption 20 says that f-image is always a LS factor and it has just two left extensions, namely two elements of $g_L(a, b)$, corresponding to two original left extensions a and b.

Assumption 20 along with the notation introduced above allow us to define the following graph.

Definition 22. Let φ be a substitution defined over an alphabet $\mathcal{A}$ satisfying Assumption 20. We define a directed labelled graph GL_φ as follows:

(i) vertices of GL_φ are all unordered couples of distinct letters a, b such that $\mathrm{Rext}(a) \cap \mathrm{Rext}(b) \neq \emptyset$,

(ii) there is an edge from a vertex (a, b) to a vertex (c, d) labelled by $f_L(a, b)$ if $g_L(a, b) = \{c, d\}$.

In fact, Assumption 20 states that out-degree of each vertex is exactly one. The graph GL_φ for our example substitution is drawn in Figure 2, this substitution satisfies Assumption 20 for it is suffix-free.

Now, let us consider the case when $\mathbf{w}$ is an infinite LS branch with $a, b \in \mathrm{Lext}(\mathbf{w}), a \neq b$. Obviously, f-image of $\mathbf{w}$ is uniquely given. For most substitutions even a "f-preimage" of each infinite LS branch exists.

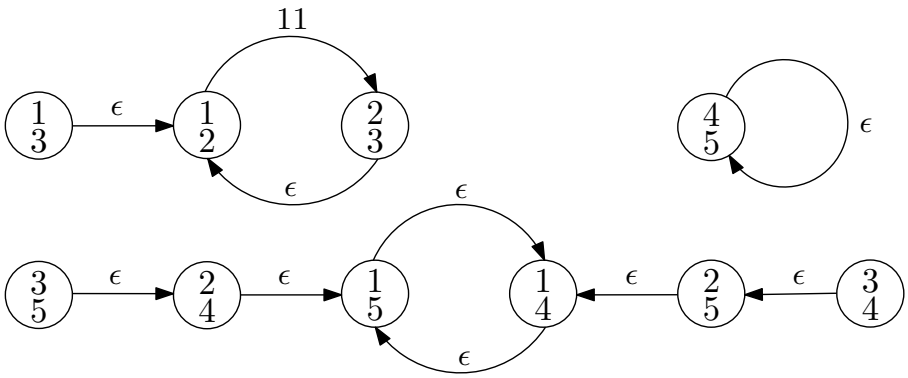


Figure 2: The graph GL_φ for the Substitution (2).

Assumption 23. An infinite word $\mathbf{u}$ is a fixed point of a substitution φ satisfying Assumption 20. For each infinite LS branch $\mathbf{w}$ of $\mathbf{u}$ with $a, b \in \mathrm{Lext}(\mathbf{w}), a \neq b$ there exists at least one infinite LS branch $\overline{\mathbf{w}}$ with left extensions c and d such that f-image of $\overline{\mathbf{w}}$ equals $\mathbf{w}$ and $g_L(c, d) = \{a, b\}$.

This assumption is very weak. Actually, we have not found any substitution not satisfying it.

Theorem 24. *Let* $\mathbf{u}$ *be a fixed point of a primitive injective substitution* φ *satisfying Assumption 23 and let* $\mathbf{w}$ *be an infinite LS branch with* $a, b \in \mathrm{Lext}(\mathbf{w}), a \neq b$. *Then either* $\mathbf{w}$ *is a* periodic point *of* φ, *i.e*

$$\mathbf{w} = \varphi^l(\mathbf{w}) \quad \textit{for some } l \geq 1, \tag{3}$$

and (a, b) *is a vertex of a cycle in* GL_φ *labelled by* ϵ *only or* $\mathbf{w} = s\varphi^l(s)\varphi^{2l}(s)\cdots$ *is the unique solution of the equation*

$$\mathbf{w} = s\varphi^l(\mathbf{w}), \tag{4}$$

where (a, b) *is a vertex of a cycle in* GL_φ *containing at least one edge with non-empty label,* l *is the length of this cycle and*

$$s = f_L(g_L^{l-1}(a, b)) \cdots \varphi^{l-2}(f_L(g_L(a, b))\varphi^{l-1}(f_L(a, b)). \tag{5}$$

3 Results for $\mathbf{u}_\beta$

Definition 25. Let $\beta > 1$ be a non-simple Parry number. The set $\mathcal{S}$ is defined as follows: β belongs to $\mathcal{S}$ if and only if one of the following conditions is satisfied

a) $d_\beta(1) = t_1 \cdots t_m (0 \cdots 0 t_{m+p})^\omega$ and $t_m > t_{m+p}$,
b) $d_\beta(1) = t_1 \cdots \underbrace{t_{m-qp}}_{\neq 0} \underbrace{0 \cdots 0}_{qp-1} t_m (t_m + 1 \cdots t_{m+p})^\omega$, $q \geq 1$, $t_m < t_{m+p}$.

As an outstanding subset of $\mathcal{S}$, we define a set $\mathcal{S}_0 = \{\beta > 1 \mid d_\beta(1) = t_1(0 \cdots 0(t_1 - 1))^\omega\}$.

Due to the previous lemma, $\beta \in \mathcal{S}$ if and only if $z = sp, s \in \mathbb{N}$.

Proposition 26. *Let $\beta > 1$ be a non-simple Parry number and let $\mathbf{u}_\beta$ be the fixed point of the canonical substitution φ_β. Then*

(i) if $p > 1$, then $\mathbf{u}_\beta$ is an infinite LS branch with left extensions $\{m, m+1, \ldots, m+p-1\}$,

(ii) if $\beta \notin \mathcal{S}$, then $\mathbf{u}_\beta$ is the only one infinite LS branch,

(iii) if $\beta \in \mathcal{S}$, then there are m infinite LS branches

$$\begin{gathered} 0^t m\varphi^m(0^t m)\varphi^{2m}(0^t m)\cdots \\ \vdots \\ \varphi^{m-1}(0^t m)\varphi^{2m-1}(0^t m)\varphi^{3m-1}(0^t m)\cdots . \end{gathered}$$

There are no other infinite LS branches of $\mathbf{u}_\beta$.

We have found all infinite LS branches. To obtain complete knowledge of the structure of LS factors we need to find all (a, b)-maximal LS factors as well. It is possible to do so but it requires introducing a lot of notations. Therefore, we present only the most important result formulated as the following lemma.

Lemma 27. *$\beta \in \mathcal{S}_0$ if and only if $\mathbf{u}_\beta$ contains a finite number of (a, b)-maximal LS factors for any $a, b \in \mathcal{A}$.*

It is important since one can prove the following.

Lemma 28. *The complexity of $\mathbf{u}_\beta$ is affine if and only if $\mathbf{u}_\beta$ contains a finite number of (a, b)-maximal LS factors for any $a, b \in \mathcal{A}$.*

This equivalence is not valid in general, for a counter example see [7]. These two lemmas give us our main result.

Theorem 29. *Let $\beta > 1$ be a non-simple Parry number and let $\mathbf{u}_\beta$ be the fixed point of the canonical substitution φ_β. The factor complexity of $\mathbf{u}_\beta$ is affine if and only if $\beta \in \mathcal{S}_0$. Then, $\mathcal{C}(n) = (m + p - 1)n + 1$. Moreover,*

(i) if $p > 1$ and $\beta \in \mathcal{S}_0$, then $\mathbf{u}_\beta$ and $0^{-1}\mathbf{u}_\beta$ are the only infinite LS branches,

(ii) $\mathbf{u}_\beta$ is Sturmian if and only if $p = 1$ and $\beta \in \mathcal{S}_0$, i.e. $d_\beta(1) = t_1(t_1 - 1)^\omega$.

The characterization of the Sturmian case is given in [10]. Remark that numbers from $\mathcal{S}_0$ are all Pisot numbers (Frougny).

References

[1] P. Ambrož. *Algebraic and combinatorial properties of non-standard numeration systems.* PhD thesis, Université Paris VII and Czech Technical University, (2006).

[2] J. Berstel. Mots de Fibonacci. Séminaire d'informatique théorique, LITP, Paris, Année 1980/81, 57-78.

[3] J. Berstel. Properties of infinite words: recent results. In 'Proceedings of the 6th Annual Symposium on Theoretical Aspects of Computer Science on STACS 89', 36–46, New York, NY, USA, (1989). Springer-Verlag New York, Inc.

[4] J. Berstel and D. Perrin. *The origins of combinatorics on words.* Eur. J. Comb. **28** (2007), 996–1022.

[5] J. Cassaigne. Special factors of sequences with linear subword complexity. In 'Developments in Language Theory II', 25–34. World Scientific, (1996).

[6] J. Cassaigne. *Complexité et facteurs spéciaux.* Bull. Belg. Math. Soc. Simon Stevin **4** (1997), 67–88.

[7] R. V. Chacon. *Weakly mixing transformations which are not strongly mixing.* Proceedings of the American Mathematical Society **22** (1969), 559–562.

[8] A. de Luca and S. Varricchio. *On the factors of the Thue-Morse word on three symbols.* Inf. Process. Lett. **27** (1988), 281–285.

[9] S. Fabre. *Substitutions et beta-systèmes de numération.* Theoretical Computer Science **137** (1995), 219–236.

[10] C. Frougny, Z. Masáková, and E. Pelantová. *Infinite special branches in words associated with beta-expansions.* Discrete Math. Theor. Comput. Sci. **9** (2007), 125–144.

[11] D. König. *Theorie der endlichen und unendlichen Graphen.* Akademische Verlagsgesellschaft, Leipzig, (1936).

[12] M. Lothaire. *Algebraic Combinatorics on Words.* Cambridge University Press, (2002).

[13] B. Mossé. *Notions de reconnaissabilité pour les substitutions et complexité des suites automatiques.* Bull. Soc. Math. France **124** (1996), 101–108.

[14] W. Parry. *On the β-expansions of real numbers.* Acta Math. Acad. Sci. Hunger. **11** (1960), 401–416.

[15] M. Queffélec. *Substitution dynamical systems–spectral analysis*, volume 1284 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, (1987).

[16] W. Thurston. Groups, tilings and finite state automata. AMS Colloquium Lecture Notes, (1989).

Propagators Associated to Periodic Hamiltonians

Petra Košťáková

1st year of PGS, email: petra.kostakova@gmail.com
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Pavel Šťovíček, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. We consider an invariant quantum Hamiltonian $H = -\Delta_{LB} + V$ in the L^2 space based on a Riemannian manifold $\tilde{M}$ with a discrete symmetry group Γ. Typically, $\tilde{M}$ is the universal covering space of a multiply connected manifold M and Γ is the fundamental group of M. To any unitary representation Λ of Γ one can relate another operator on $M = \tilde{M}/\Gamma$, called H_Λ, which formally corresponds to the same differential operator as H but which is determined by quasi-periodic boundary conditions. We give a brief review of the Bloch decomposition of H and of a formula relating the propagators associated to the Hamiltonians H_Λ and H. Then we concentrate on the example of the Aharonov-Bohm effect with two vortices.

Abstrakt. Mějme invariantní Hamiltonián $H = -\Delta_{LB} + V$ na $L^2(\tilde{M})$, kde $\tilde{M}$ je Riemanovská varieta se spočetně konečnou grupou symetrií Γ. $\tilde{M}$ je nejčastěji univerzální nakrývací prostor variety M a Γ je její fundamentální grupa. Ke každé unitární reprezentaci Λ grupy Γ lze přiřadit operátor H_Λ na $M = \tilde{M}/\Gamma$. Ten je formálně stejný jako operátor H, navíc je určen kvazi-periodickými okrajovými podmínkami. V následujícím textu stručně nastíníme konstrukci Blochova rozkladu operátoru H a rozklad propagátoru náležející operátorům H_Λ a H. Tento postup je následně aplikován na Aharono-Bohmův jev se dvěma cívkami.

1 Introduction

Suppose that there is given a connected Riemannian manifold $\tilde{M}$ with a discrete symmetry group Γ. Let us consider a Γ-periodic Hamilton operator in $L^2(\tilde{M})$ of the form $H = -\Delta_{LB} + V$ where Δ_{LB} is the Laplace-Beltrami operator and V is a Γ-invariant bounded real function on $\tilde{M}$. To any unitary representation Λ of Γ one can relate another operator on $M = \tilde{M}/\Gamma$, called H_Λ, which formally corresponds to the same differential operator as H but which is determined by quasi-periodic boundary conditions. In the framework of the Feynman path integral there was derived a remarkable formula relating the propagators $\mathcal{K}^\Lambda_t(x, x_0)$ and $\mathcal{K}_t(x, x_0)$ associated respectively to the Hamiltonians H_Λ and H [11, 12].There exists also an opposite point of view when one decomposes the operator H into a direct integral with components H_Λ where Λ runs over all irreducible unitary representations of Γ [14, 1, 4]. The evolution operator then decomposes correspondingly. This type of decomposition is an essential step in the Bloch analysis. Let us also note that an alternative approach to the Bloch analysis, based on a more algebraic point of view, has been proposed recently in [5].

The both relations, the propagator formula on the one hand and the generalized Bloch decomposition on the other hand, are in a sense mutually inverse [8]. In the current paper we give a brief review of basic results concerning this relationship and further consider

the example of the Aharonov-Bohm effect with two vortices. In this case $\tilde{M}$ is identified with the universal covering space of the plane with two excluded points and Γ is the fundamental group of the same manifold.

The paper is organized as follows. In Section 2 we give a brief review of basic results concerning the relationship between the generalized Bloch analysis and the formula for propagators associated to periodic Hamiltonians. In Section 3 we explain the construction of the propagator on the universal covering space in the case of the Aharonov-Bohm effect with two vortices and discuss the application of the propagator formula in this particular case.

2 Propagators associated to periodic Hamiltonians

2.1 Periodic Hamiltonians

Let $\tilde{M}$ be a connected Riemannian manifold with a discrete and at most countable symmetry group Γ. The action of Γ on $\tilde{M}$ is assumed to be smooth, free and proper (also called properly discontinuous). Denote by $\tilde{\mu}$ the measure on $\tilde{M}$ induced by the Riemannian metric. The quotient $M = \tilde{M}/\Gamma$ is a connected Riemannian manifold with an induced measure μ. This way one gets a principal fiber bundle $\pi : \tilde{M} \to M$ with the structure group Γ. The L^2 spaces on the manifolds M and $\tilde{M}$ are everywhere tacitly understood with the measures μ and $\tilde{\mu}$, respectively.

Typically, $\tilde{M}$ is the universal covering space of M and $\Gamma = \pi_1(M)$ is the fundamental group of M. For example, this is the case when one is considering the Aharonov-Bohm effect.

To a unitary representation Λ of Γ in a separable Hilbert space $\mathscr{L}_\Lambda$ one relates the Hilbert space $\mathscr{H}_\Lambda$ formed by Λ-equivariant vector-valued functions on $\tilde{M}$. This means that any function $\psi \in \mathscr{H}_\Lambda$ is measurable with values in $\mathscr{L}_\Lambda$ and satisfies

$$\forall s \in \Gamma,\ \psi(s \cdot y) = \Lambda(s)\psi(y) \ \text{ almost everywhere on } \tilde{M}.$$

Moreover, the norm of ψ induced by the scalar product is finite. If $\psi_1, \psi_2 \in \mathscr{H}_\Lambda$ then the function $y \mapsto \langle \psi_1(y), \psi_2(y) \rangle$ defined on $\tilde{M}$ is Γ-invariant and so it projects to a function s_{ψ_1,ψ_2} defined on M, and the scalar product is defined by

$$\langle \psi_1, \psi_2 \rangle = \int_M s_{\psi_1,\psi_2}(x)\, d\mu(x).$$

As already announced, our discussion concerns Γ-periodic Hamiltonians on $\tilde{M}$ of the form $H = -\Delta_{LB} + V$ where Δ_{LB} is the Laplace-Beltrami operator and $V(y)$ is a Γ-invariant measurable bounded real function on $\tilde{M}$. Here we accept the Friedrichs extension as the preferred self-adjoint extension of semibounded symmetric operators defined on test functions.

To the same differential operator, $-\Delta_{LB} + V$, one can relate a selfadjoint operator H_Λ in the space $\mathscr{H}_\Lambda$ for any unitary representation Λ of Γ. Let us define $\Phi_\Lambda : C_0^\infty(\tilde{M}) \otimes \mathscr{L}_\Lambda \to \mathscr{H}_\Lambda$ by

$$\forall \varphi \in C_0^\infty(\tilde{M}), \forall v \in \mathscr{L}_\Lambda,\ \ (\Phi_\Lambda \varphi \otimes v)(y) = \sum_{s \in \Gamma} \varphi(s \cdot y)\, \Lambda(s^{-1}) v.$$

Since the action of Γ is proper, the vector-valued function $\Phi_\Lambda\,\varphi\otimes v$ is smooth. Moreover, $\Phi_\Lambda\,\varphi\otimes v$ is Λ-equivariant and the norm of $\Phi_\Lambda\,\varphi\otimes v$ in $\mathscr{H}_\Lambda$ is finite. Furthermore, the range of Φ_Λ is dense in $\mathscr{H}_\Lambda$. The Laplace-Beltrami operator is well defined on $\mathrm{Ran}(\Phi_\Lambda)$ and it holds

$$\Delta_{LB}\Phi_\Lambda[\varphi\otimes v]=\Phi_\Lambda[\Delta_{LB}\varphi\otimes v].$$

One can also verify that the differential operator $-\Delta_{LB}$ is positive on the domain $\mathrm{Ran}(\Phi_\Lambda)\subset\mathscr{H}_\Lambda$. Since the function $V(y)$ is Γ-invariant, the multiplication operator by V is well defined in the Hilbert space $\mathscr{H}_\Lambda$. The Hamiltonian H_Λ is defined as the Friedrichs extension of the differential operator $-\Delta_{LB}+V$ considered on the domain $\mathrm{Ran}\,\Phi_\Lambda$.

2.2 A generalization of the Bloch analysis

Let $\hat\Gamma$ be the dual space to Γ (the quotient space of the space of irreducible unitary representations of Γ). In the first step of the generalized Bloch analysis one decomposes H into a direct integral over $\hat\Gamma$ with the components being equal to H_Λ. As a corollary one obtains a similar relationship for the evolution operators $U(t)=\exp(-itH)$ and $U_\Lambda(t)=\exp(-itH_\Lambda)$, $t\in\mathbb{R}$. To achieve this goal a well defined harmonic analysis on the group Γ is necessary.

It is known that the harmonic analysis is well established for locally compact groups of type I [13]. So all formulas presented bellow are perfectly well defined provided Γ is a type I group. A countable discrete group is type I, however, if and only if it has an Abelian normal subgroup of finite index [17, Satz 6]. This means that there exist multiply connected configuration spaces of interest whose fundamental groups are not of type I. For example, the fundamental group in the case of the Aharonov-Bohm effect with two vortices is the free group with two generators and it is not of type I. Fortunately, in this case, too, there exists a well defined harmonic analysis [16].

Let us recall the basic properties of the harmonic analysis on discrete type I groups [13]. In that case the Haar measure on Γ is chosen as the counting measure. Let $\mathrm{d}\hat m$ be the Plancherel measure on $\hat\Gamma$. Denote by $\mathscr{I}_2(\mathscr{L}_\Lambda)\equiv\mathscr{L}_\Lambda^*\otimes\mathscr{L}_\Lambda$ the Hilbert space formed by Hilbert-Schmidt operators on $\mathscr{L}_\Lambda$ ($\mathscr{L}_\Lambda^*$ is the dual space to $\mathscr{L}_\Lambda$). The Fourier transformation is defined as a unitary mapping

$$\mathscr{F}:L^2(\Gamma)\to\int_{\hat\Gamma}^{\oplus}\mathscr{I}_2(\mathscr{L}_\Lambda)\,\mathrm{d}\hat m(\Lambda).$$

For $f\in L^1(\Gamma)\subset L^2(\Gamma)$ one has

$$\mathscr{F}[f](\Lambda)=\sum_{s\in\Gamma}f(s)\Lambda(s).$$

Conversely, if f is of the form $f=g*h$ (the convolution) where $g,h\in L^1(\Gamma)$, and $\hat f=\mathscr{F}[f]$ then

$$f(s)=\int_{\hat\Gamma}\mathrm{Tr}[\Lambda(s)^*\hat f(\Lambda)]\,\mathrm{d}\hat m(\Lambda).$$

It is known that if Γ is a countable discrete group of type I then $\dim\mathscr{L}_\Lambda$ is a bounded function of Λ on the dual space $\hat\Gamma$ [17, Korollar I]. Using the unitarity of the Fourier

transform one finds that

$$\hat{m}(\hat{\Gamma}) \leq \int_{\hat{\Gamma}} \dim \mathscr{L}_\Lambda \, \mathrm{d}\hat{m}(\Lambda) = 1.$$

The following rule satisfied by the Fourier transformation is also of crucial importance:

$$\forall s \in \Gamma, \forall f \in L^2(\Gamma),\ \mathscr{F}[f(s \cdot g)](\Lambda) = \Lambda(s^{-1}) \mathscr{F}[f(g)](\Lambda).$$

Now we are going to construct a unitary mapping

$$\Phi : L^2(\tilde{M}) \to \int_{\hat{\Gamma}}^{\oplus} \mathscr{L}_\Lambda^* \otimes \mathscr{H}_\Lambda \, \mathrm{d}\hat{m}(\Lambda)$$

which makes it possible to decompose the Hamiltonian H. Observe that the tensor product $\mathscr{L}_\Lambda^* \otimes \mathscr{H}_\Lambda$ can be naturally identified with the Hilbert space of $1 \otimes \Lambda$-equivariant operator-valued functions on $\tilde{M}$ with values in $\mathscr{L}_\Lambda^* \otimes \mathscr{L}_\Lambda \equiv \mathscr{I}_2(\mathscr{L}_\Lambda)$. For $f \in L^2(\tilde{M})$ and $y \in \tilde{M}$ set

$$\forall s \in \Gamma,\ f_y(s) = f(s^{-1} \cdot y).$$

The norm $\|f_y\|$ in $L^2(\Gamma)$ is a Γ-invariant function of $y \in \tilde{M}$, and the projection of this function onto M can be checked to be square integrable. Hence for almost all $x \in M$ and all $y \in \pi^{-1}(\{x\})$ it holds $f_y \in L^2(\Gamma)$. We define the component $\Phi[f](\Lambda)$, $\Lambda \in \hat{\Gamma}$, by the prescription

$$\Phi[f](\Lambda)\,(y) := \mathscr{F}[f_y](\Lambda) \in \mathscr{I}_2(\mathscr{L}_\Lambda).$$

In particular, if $f \in L^1(\tilde{M}) \cap L^2(\tilde{M})$ then

$$\Phi[f](\Lambda)\,(y) = \sum_{s \in \Gamma} f(s^{-1} \cdot y) \Lambda(s).$$

Equivalently one can define Φ in the following way. For $\varphi \in C_0^\infty(\tilde{M})$, $v \in \mathscr{L}_\Lambda$ and $y \in \tilde{M}$ set

$$\Phi[\varphi](\Lambda)(y)v = (\Phi_\Lambda \, \varphi \otimes v)(y). \tag{1}$$

Then Φ introduced in (1) is an isometry and extends unambiguously to a unitary mapping.

Finally one can verify the formula

$$\Phi H \Phi^{-1} = \int_{\hat{\Gamma}}^{\oplus} 1 \otimes H_\Lambda \, \mathrm{d}\hat{m}(\Lambda)$$

which represents the sought Bloch decomposition. As a corollary we have

$$\Phi U(t) \Phi^{-1} = \int_{\hat{\Gamma}}^{\oplus} 1 \otimes U_\Lambda(t) \, \mathrm{d}\hat{m}(\Lambda). \tag{2}$$

2.3 A construction for propagators associated to periodic Hamiltonians

In equality (2), the evolution operator $U(t)$ is expressed in terms of $U_\Lambda(t)$, $\Lambda \in \hat{\Gamma}$. It is possible to invert this relationship and to derive a formula for the propagator associated to H_Λ which is expressed in terms of the propagator associated to H.

The propagators are regarded as distributions which are introduced as kernels of the corresponding evolution operators. Recall that by the Schwartz kernel theorem (see, for example, [7, Theorem 5.2.1]), to every $B \in \mathscr{B}(L^2(\tilde{M}))$ there exists one and only one $\beta \in \mathscr{D}'(\tilde{M} \times \tilde{M})$ such that

$$\forall \varphi_1, \varphi_2 \in C_0^\infty(\tilde{M}), \quad \beta(\overline{\varphi_1} \otimes \varphi_2) = \langle \varphi_1, B\varphi_2 \rangle.$$

Moreover, the map $B \mapsto \beta$ is injective. One calls β the kernel of B.

The kernel theorem can be extended to Hilbert spaces formed by Λ-equivariant vector-valued functions. In this case the kernels are operator-valued distributions. To every $B \in \mathscr{B}(\mathscr{H}_\Lambda)$ there exists one and only one $\beta \in \mathscr{D}'(\tilde{M} \times \tilde{M}) \otimes \mathscr{B}(\mathscr{L}_\Lambda)$ such that

$$\begin{aligned} &\forall \varphi_1, \varphi_2 \in C_0^\infty(\tilde{M}), \forall v_1, v_2 \in \mathscr{L}_\Lambda, \\ &\langle v_1, \beta(\overline{\varphi_1} \otimes \varphi_2) v_2 \rangle = \langle \Phi_\Lambda\, \varphi_1 \otimes v_1, B\, \Phi_\Lambda\, \varphi_2 \otimes v_2 \rangle. \end{aligned}$$

The distribution β is Λ-equivariant:

$$\forall s \in \Gamma, \quad \beta(s \cdot y_1, y_2) = \Lambda(s)\beta(y_1, y_2),\ \beta(y_1, s \cdot y_2) = \beta(y_1, y_2)\Lambda(s^{-1})$$

In this case, too, the map $B \mapsto \beta$ is injective.

Denote by $\mathcal{K}_t \in \mathscr{D}'(\tilde{M} \times \tilde{M})$ the kernel of $U(t) \in \mathscr{B}(L^2(\tilde{M}))$, and by $\mathcal{K}_t^\Lambda \in \mathscr{D}'(\tilde{M} \times \tilde{M}) \otimes \mathscr{B}(\mathscr{L}_\Lambda)$ the kernel of $U_\Lambda(t) \in \mathscr{B}(\mathscr{H}_\Lambda)$. Here and everywhere in this section, t is a real parameter. The kernel $\mathcal{K}_t^\Lambda$ is Λ-equivariant:

$$\forall s \in \Gamma, \quad \mathcal{K}_t^\Lambda(s \cdot y_1, y_2) = \Lambda(s)\mathcal{K}_t^\Lambda(y_1, y_2),\ \mathcal{K}_t^\Lambda(y_1, s \cdot y_2) = \mathcal{K}_t^\Lambda(y_1, y_2)\Lambda(s^{-1}).$$

First we rewrite the Bloch decomposition of the propagator (2) in terms of kernels. It is possible to prove that, for all $\varphi_1, \varphi_2 \in C_0^\infty(\tilde{M})$, the function $\Lambda \mapsto \mathrm{Tr}[\mathcal{K}_t^\Lambda(\overline{\varphi_1} \otimes \varphi_2)]$ is integrable on $\hat{\Gamma}$ and

$$\mathcal{K}_t(\varphi_1 \otimes \varphi_2) = \int_{\hat{\Gamma}} \mathrm{Tr}[\mathcal{K}_t^\Lambda(\varphi_1 \otimes \varphi_2)]\, \mathrm{d}\hat{m}(\Lambda).$$

An inverse relation was derived by Schulman in the framework of path integration [11, 12] and reads

$$\mathcal{K}_t^\Lambda(x, y) = \sum_{s \in \Gamma} \Lambda(s)\, \mathcal{K}_t(s^{-1} \cdot x, y). \tag{3}$$

It is possible to give (3) the following rigorous interpretation. Suppose that $\varphi_1, \varphi_2 \in C_0^\infty(\tilde{M})$ are fixed but otherwise arbitrary. Set

$$F_t(s) = \mathcal{K}_t\big(\varphi_1(s^{-1} \cdot y_1) \otimes \varphi_2(y_2)\big) \text{ for } s \in \Gamma,$$

and

$$G_t(\Lambda) = \mathcal{K}_t^{\Lambda}(\varphi_1 \otimes \varphi_2) \in \mathscr{I}_2(\mathscr{L}_\Lambda) \text{ for } \Lambda \in \hat{\Gamma}.$$

One can show that $F_t \in L^2(\Gamma)$ and G_t is bounded on $\hat{\Gamma}$ in the Hilbert-Schmidt norm. Recalling that $\hat{m}(\hat{\Gamma}) \leq 1$ we have $\|G_t(\cdot)\| \in L^1(\hat{\Gamma}) \cap L^2(\hat{\Gamma})$. In [8] it is verified that

$$F_t = \mathscr{F}^{-1}[G_t].$$

and, consequently,

$$G_t = \mathscr{F}[F_t]. \tag{4}$$

Rewriting (4) formally gives equality (3).

3 The Aharonov-Bohm effect with two vortices

3.1 The propagator on the universal covering space

The configuration space for the Aharonov-Bohm effect with two vortices is the plane with two excluded points, $M = \mathbb{R}^2 \setminus \{a, b\}$. This is a flat Riemannian manifold and the same is true for the universal covering space $\tilde{M}$. Let $\pi : \tilde{M} \to M$ be the projection. It is convenient to complete the manifold $\tilde{M}$ by a countable set of points $\mathcal{A} \cup \mathcal{B}$ which lie on the border of $\tilde{M}$ and project onto the excluded points, $\pi(\mathcal{A}) = \{a\}$ and $\pi(\mathcal{B}) = \{b\}$.

$\tilde{M}$ looks locally like $\mathbb{R}^2$ but differs from the Euclidean space by some global features. First of all, not every two points from $\tilde{M}$ can be connected by a geodesic segment. Fix a point

$y \in \tilde{M}$ which can be connected with x by a geodesic segment. The domain $D(x)$ is one sheet of the covering $\tilde{M} \to M$. It can be identified with $\mathbb{R}^2$ cut along two halflines with the limit points a and b, respectively. Thus the border $\partial D(x)$ is formed by four halflines. The universal covering space $\tilde{M}$ can be imagined as a result of an infinite process of glueing together countably many copies of $D(x)$ with each copy having four neighbors.

The fundamental group of M, called Γ, is known to be the free group with two generators g_a and g_b. For the generator g_a one can choose the homotopy class of a simple positively oriented loop winding once around the point a and leaving the point b in the exterior. Analogously one can choose the generator g_b by interchanging the role of a and b. One-dimensional unitary representations Λ of Γ are determined by two numbers α, β, $0 \leq \alpha, \beta < 1$, such that

$$\Lambda(g_a) = e^{2\pi i\alpha},\ \Lambda(g_b) = e^{2\pi i\beta}.$$

The standard way to define the Aharonov-Bohm Hamiltonian with two vortices is to choose a vector potential $\overrightarrow{A}$ for which rot $\overrightarrow{A} = 0$ on M and such that the nonintegrable phase factor [18] for a closed path from the homotopy class g_a or g_b equals $e^{2\pi i\alpha}$ or $e^{2\pi i\beta}$, respectively (assuming that $0 < \alpha, \beta < 1$). The Hamiltonian then acts as the differential operator $(-i\nabla - \overrightarrow{A})^2$ in $L^2(M)$. A unitarily equivalent and for our purposes more convenient possibility is to work with the Hamiltonian $H_\Lambda = -\Delta$ in the Hilbert space $\mathscr{H}_\Lambda$ of Λ-equivariant functions on $\tilde{M}$, as introduced in Section 2.1. Parallelly one considers the free Hamiltonian $H = -\Delta$ in $L^2(\tilde{M})$. H is Γ-periodic. In order to compute, according to prescription (3), the propagator $\mathcal{K}^{\Lambda}(t, x, y)$ associated to H_Λ one needs to

derive a formula for the free propagator $\mathcal{K}(t,x,y)$ on $\tilde{M}$. Such a formula is recalled below following [15].

Let ϑ be the Heaviside step function. For $x,y \in \tilde{M}\cup\mathcal{A}\cup\mathcal{B}$ set $\chi(x,y)=1$ if the points x, y can be connected by a geodesic segment, and $\chi(x,y)=0$ otherwise. Given in addition $t\in\mathbb{R}$ we define

$$Z(t,x,y)=\vartheta(t)\chi(x,y)\,\frac{1}{4\pi i t}\,\exp\!\left(\frac{i}{4t}\,\mathrm{dist}^2(x,y)\right),$$

Furthermore, for $x_1,x_2,x_3\in\tilde{M}\cup\mathcal{A}\cup\mathcal{B}$ such that $\chi(x_1,x_2)=\chi(x_2,x_3)=1$, and for $t_1,t_2>0$ we set

$$V\begin{pmatrix} x_3,x_2,x_1\\ t_2,t_1\end{pmatrix}=2i\left(\left(\theta-\pi+i\log\!\left(\frac{t_2r_1}{t_1r_2}\right)\right)^{-1}-\left(\theta+\pi+i\log\!\left(\frac{t_2r_1}{t_1r_2}\right)\right)^{-1}\right)$$

where $\theta=\angle x_1,x_2,x_3\in\mathbb{R}$ is the oriented angle and $r_1=\mathrm{dist}(x_1,x_2)$, $r_2=\mathrm{dist}(x_2,x_3)$. Note that if the inner vertex x_2 belongs to the set of extreme points $\mathcal{A}\cup\mathcal{B}$ then the angle θ can take any real value.

We claim that the free propagator on $\tilde{M}$ equals

$$\mathcal{K}(t,x,x_0)=\sum_{\gamma}\mathcal{K}_\gamma(t,x,x_0) \tag{5}$$

where the sum runs over all piecewise geodesic curves $\gamma: x_0\to C_1\to\ldots\to C_n\to x$ with the inner vertices C_j, $1\le j\le n$, belonging to the set of extreme points $\mathcal{A}\cup\mathcal{B}$. This means that it should hold $\chi(x_0,C_1)=\chi(C_1,C_2)=\ldots=\chi(C_n,x)=1$. Let us denote by $|\gamma|=n$ the length of the sequence $(C_1,C_2,\ldots,C_n)$. In particular, if $|\gamma|=0$ then γ designates the geodesic segment $x_0\to x$. To simplify notation we set everywhere where convenient $C_0=x_0$ and $C_{n+1}=x$. With this convention, the summands in (5) equal

$$\begin{aligned}&\mathcal{K}_\gamma(t,x,x_0)\\ &=\int_{\mathbb{R}^{n+1}}\mathrm{d}t_n\ldots\mathrm{d}t_0\,\delta(t_n+\ldots+t_0-t)\prod_{j=0}^{n-1}V\begin{pmatrix}C_{j+2},C_{j+1},C_j\\ t_{j+1},t_j\end{pmatrix}\prod_{j=0}^{n}Z(t_j,C_{j+1},C_j).\end{aligned} \tag{6}$$

In particular, if $|\gamma|=0$ then $\mathcal{K}_\gamma(t,x,x_0)=Z(t,x,x_0)$, and if $|\gamma|=1$ then γ designates a path composed of two geodesic segments $x_0\to C\to x$, with $C\in\mathcal{A}\cup\mathcal{B}$, and

$$\mathcal{K}_\gamma(t,x,x_0)=\vartheta(t)\int_0^t V\begin{pmatrix}x,C,x_0\\ t-s,s\end{pmatrix}Z(t-s,x,C)Z(s,C,x_0)\,\mathrm{d}s\,.$$

For detailed derivation of this formula see [9].

3.2 The propagator associated to H_Λ

Without loss of generality we can suppose that the vortices are located in the points $a=(0,0)$ and $b=(\varrho,0)$. Let (r_a,θ_a) be the polar coordinates centered at the point a

and (r_b, θ_b) be the polar coordinates centered at the point b. To express the propagator for H_Λ it is convenient to pass to a unitarily equivalent formulation. Let us cut the plane along two half-lines,

$$L_a =]-\infty, 0[\times \{0\} \text{ and } L_b =]\varrho, +\infty[\times \{0\}.$$

The values $\theta_a = \pm\pi$ correspond to the two sides of the cut L_a, and similarly for θ_b and L_b. The unitarily equivalent Hamiltonian H'_Λ is formally equal to $-\Delta$ in $L^2(\mathbb{R}^2, \mathrm{d}^2x)$ and is determined by the boundary conditions along the cut,

$$\begin{aligned}
&\psi(r_a, \theta_a = \pi) = e^{2\pi\, i\, \alpha}\psi(r_a, \theta_a = -\pi),\ \partial_{r_a}\psi(r_a, \theta_a = \pi) = e^{2\pi\, i\, \alpha}\partial_{r_a}\psi(r_a, \theta_a = -\pi)\,,\\
&\psi(r_b, \theta_b = \pi) = e^{2\pi\, i\, \beta}\psi(r_b, \theta_b = -\pi),\ \partial_{r_b}\psi(r_b, \theta_b = \pi) = e^{2\pi\, i\, \beta}\partial_{r_b}\psi(r_b, \theta_b = -\pi)\,.
\end{aligned}$$

In addition, one should impose a boundary condition at the vortices, namely $\psi(a) = \psi(b) = 0$.

Let us denote $D = \mathbb{R}^2 \setminus (L_a \cup L_b)$. Then one can embed $D \subset \tilde{M}$ as a fundamental domain. We wish to find a formula for the propagator $\mathcal{K}'^{\Lambda}(t, x, x_0)$ associated to the Hamiltonian H'_Λ. It can be simply obtained as the restriction to D of the propagator $\mathcal{K}^\Lambda(t, x, x_0)$ associated to the Hamiltonian H_Λ. On the other hand, to construct $\mathcal{K}^\Lambda(t, x, x_0)$ one can apply formula (3) and the knowledge of the free propagator on $\tilde{M}$, see (5), (6). Thus we get

$$\mathcal{K}^\Lambda(t, x, x_0) = \sum_{g\in\Gamma}\sum_{\gamma} \Lambda(g^{-1})\, \mathcal{K}_\gamma(t, g\cdot x, x_0). \tag{7}$$

Fix $t > 0$ and $x_0, x \in D$. One can classify piecewise geodesic paths in $\tilde{M}$,

$$\gamma : x_0 \to C_1 \to \ldots \to C_n \to g\cdot x, \tag{8}$$

with $C_j \in \mathcal{A} \cup \mathcal{B}$ and $g \in \Gamma$, according to their projections to M. Let $\overline{\gamma}$ be a finite alternating sequence of points a and b, i.e., $\overline{\gamma} = (c_1, \ldots, c_n)$, $c_j \in \{a, b\}$ and $c_j \neq c_{j+1}$. The empty sequence $\overline{\gamma} = ()$ is admissible. Relate to $\overline{\gamma}$ a piecewise geodesic path in M, namely $x_0 \to c_1 \to \ldots \to c_n \to x$. Suppose that this path is covered by a path γ in $\tilde{M}$, as given in (8). Then $C_j \in \mathcal{A}$ iff $c_j = a$ and $C_j \in \mathcal{B}$ iff $c_j = b$. Denote the angles $\angle\, x_0, c_1, c_2 = \theta_0$ and $\angle\, c_{n-1}, c_n, x = \theta$. Then the angles in the path γ in (8) take the values $\angle\, x_0, C_1, C_2 = \theta_0 + 2\pi k_1$, $\angle\, C_{n-1}, C_n, g\cdot x = \theta + 2\pi k_n$ and $\angle\, C_j, C_{j+1}, C_{j+2} = 2\pi k_{j+1}$ for $1 \le j \le n-2$ (if $n \ge 3$), where $k_1, \ldots, k_n$ are integers. Any values $k_1, \ldots, k_n \in \mathbb{Z}$ are possible. In that case the representation Λ applied to the group element g occurring in (8) takes the value

$$\Lambda(g) = \exp(2\pi i(k_1\sigma_1 + \ldots + k_n\sigma_n))$$

where $\sigma_j \in \{\alpha, \beta\}$ and $\sigma_j = \alpha$ if $c_j = a$, and $\sigma_j = \beta$ if $c_j = b$.

Using the equalities

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}} \exp(2\pi \mathrm{i}\alpha k)\left(\frac{1}{\theta + 2k\pi - \pi + \mathrm{i}s} - \frac{1}{\theta + 2\pi k + \pi + \mathrm{i}s}\right)\\
&= -\sin(\pi\alpha)\int_{-\infty}^{+\infty} \frac{\exp((\theta + \mathrm{i}s)\tau)}{\sin(\pi(\alpha + \mathrm{i}\tau))}\,\mathrm{d}\tau
\end{aligned}$$

and

$$\int_{-\infty}^{\infty} \frac{\exp((\theta+\mathrm{i}s)\tau)}{\sin(\pi(\alpha+\mathrm{i}\tau))}\,\mathrm{d}\tau = 2\,\frac{\exp(-\alpha(s-\mathrm{i}\theta))}{1+\exp(-s+\mathrm{i}\theta)},$$

that are valid for $0<\alpha<1$, $|\theta|<\pi$, one can carry out a partial summation in (7) over the integers $k_1,\ldots,k_n$. This way the double sum in (7) reduces to a sum over finite alternating sequences $\overline{\gamma}$.

Let us conclude our contribution by giving the resulting formula for $\mathcal{K}'^{\Lambda}(t,x,x_0)$. We set

$$\zeta_a = 1 \text{ or } \zeta_a = e^{2\,\pi\,i\,\alpha} \text{ or } \zeta_a = e^{-2\,\pi\,i\,\alpha}$$

depending on whether the segment $\overline{x_0x}$ does not intersect L_a, or $\overline{x_0x}$ intersects L_a and x_0 lies in the lower half-plane, or $\overline{x_0x}$ intersects L_a and x_0 lies in the upper half-plane. Analogously,

$$\zeta_b = 1 \text{ or } \zeta_b = e^{2\,\pi\,i\,\beta} \text{ or } \zeta_b = e^{-2\,\pi\,i\,\beta}$$

depending on whether the segment $\overline{x_0x}$ does not intersect L_b, or $\overline{x_0x}$ intersects L_b and x_0 lies in the upper half-plane, or $\overline{x_0x}$ intersects L_b and x_0 lies in the lower half-plane. Furthermore, let us set

$$\zeta_a = e^{i\,\alpha\,\eta_a},\ \zeta_b = e^{i\,\beta\,\eta_b}, \quad \text{where } \eta_a,\eta_b \in \{0,2\,\pi,-2\,\pi\}.$$

Then one has

$$\begin{aligned}
&\mathcal{K}'^{\Lambda}(t,x,x_0)\\
&\quad= \zeta_a\zeta_b\,\frac{1}{4\pi i t}\,\exp\!\left(i\,\frac{|x-x_0|^2}{4t}\right)\\
&\qquad-\zeta_a\,\frac{\sin(\pi\alpha)}{4\pi^2 i}\int_0^\infty\frac{\mathrm{d}t_1}{t_1}\int_0^\infty\frac{\mathrm{d}t_0}{t_0}\,\delta(t_1+t_0-t)\\
&\qquad\qquad\times\exp\!\left(i\left(\frac{r_a^2}{4t_1}+\frac{r_{0a}^2}{4t_0}\right)\right)\frac{\exp[-\alpha(s_a-i(\theta_a-\theta_{0a}-\eta_a)]}{1+\exp(-s_a+i\theta_a-i\theta_{0a})}\\
&\qquad-\zeta_b\,\frac{\sin(\pi\beta)}{4\pi^2 i}\int_0^\infty\frac{\mathrm{d}t_1}{t_1}\int_0^\infty\frac{\mathrm{d}t_0}{t_0}\,\delta(t_1+t_0-t)\\
&\qquad\qquad\times\exp\!\left(i\left(\frac{r_b^2}{4t_1}+\frac{r_{0b}^2}{4t_0}\right)\right)\frac{\exp[-\beta(s_b-i(\theta_b-\theta_{0b}-\eta_b)]}{1+\exp(-s_b+i\theta_b-i\theta_{0b})}\\
&\qquad+\frac{1}{4\pi i}\sum_{\overline{\gamma},n\geq 2}(-1)^n\int_0^\infty\frac{\mathrm{d}t_n}{t_n}\ldots\int_0^\infty\frac{\mathrm{d}t_0}{t_0}\,\delta(t_n+\ldots+t_0-t)\\
&\qquad\qquad\times\exp\!\left(\frac{i}{4}\left(\frac{r^2}{t_n}+\frac{\varrho^2}{t_{n-1}}+\ldots+\frac{\varrho^2}{t_1}+\frac{r_0^2}{t_0}\right)\right)S_{\overline{\gamma}}(s,\theta,\theta_0),
\end{aligned}$$

where

$$\begin{aligned}
S_{\overline{\gamma}}(s,\theta,\theta_0) \;=\;& \frac{\sin(\pi\sigma_n)}{\pi}\frac{\exp[-\sigma_n(s_n-i\theta)]}{1+\exp(-s_n+i\theta)}\frac{\sin\pi\sigma_{n-1}}{\pi}\frac{\exp(-\sigma_{n-1}s_{n-1})}{1+\exp(-s_n)}\\
&\times\ldots\times\frac{\sin(\pi\sigma_1)}{\pi}\frac{\exp[-\sigma_1(s_1-i\theta_0)]}{1+\exp(-s_1+i\theta_0)},
\end{aligned}$$

and

$$s_a = \log\left(\frac{t_1 r_{0a}}{t_0 r_a}\right),\ s_b = \log\left(\frac{t_1 r_{0b}}{t_0 r_b}\right),\ s_j = \log\left(\frac{t_j r_{j-1}}{t_{j-1} r_j}\right) \text{ for } 1 \leq j \leq n.$$

In addition, (r, θ) are the polar coordinates of the point x with respect to the center c_n, (r_0, θ_0) are the polar coordinates of the point x_0 with respect to the center c_1. The sum $\sum_{\overline{\gamma}, n \geq 2}$ runs over all finite alternating sequences of length at least two, $\overline{\gamma} = (c_1, \ldots, c_n)$, such that for all j, $c_j \in \{a, b\}$, $c_j \neq c_{j+1}$, and $\sigma_j = \alpha$ (resp. β) depending on whether $c_j = a$ (resp. b).

Acknowledgments

This work was supported by the project of the Grant Agency of the Czech Republic No. 202/08/H072.

References

[1] J. Asch, H. Over and R. Seiler. *Magnetic Bloch analysis and Bochner Laplacians.* J. Geom. Phys. **13**, (1994), 275-288.

[2] M. F. Atiyah. *Elliptic operators, discrete groups and von Neumann algebras.* Astérisque **32-33**, (1976), 43-72.

[3] O. Giraud, A. Thain and J. H. Hannay. *Shrunk loop theorem for the topology probabilities of closed Brownian (or Feynman) paths on the twice punctured plane* J. Phys. A: Math. Gen. **37**, (2004), 2913-2935.

[4] M. J. Gruber. *Bloch theory and quantization of magnetic systems.* J. Geom. Phys. **34**, (2000), 137-154.

[5] M. J. Gruber. *Noncommutative Bloch theory* J. Math. Phys. **42**, (2001), 2438-2465.

[6] J. H. Hannay and A. Thain. *Exact scattering theory for any straight reflectors in two dimensions* J. Phys. A: Math. Gen. **36**, (2003), 4063-4080.

[7] L. Hörmander. *The Analysis of Linear Partial Differential Operators I.* (Berlin: Springer), (2003).

[8] P. Kocábová and P. Šťovíček. *Generalized Bloch analysis and propagators on Riemannian manifolds with a discrete symmetry.* J. Math. Phys. **49**, (2008)

[9] P. Kocábová, P. Šťovíček. *Propagators associated to periodic Hamiltonians: an example of the Aharonov-Bohm Hamiltonian with two vortices.* arXiv:0802.0755, (2008).

[10] S. Mashkevich, J. Myrheim and S. Ouvry. *Quantum mechanics of a particle with two magnetic impurities.* Phys. Lett. A **330**, (2004), 41-47.

[11] L. S. Schulman. *Approximate topologies.* J. Math. Phys. **12**, (1971), 304-308.

[12] L. S. Schulman. *Techniques and Applications of Path Integration.* New York: Wiley, (1981).

[13] A. I. Shtern. *Unitary representation of a topological group.* The Online Encyclopaedia of Mathematics (Berlin: Springer), Online: http://eom.springer.de/, (2001).

[14] T. Sunada. *Fundamental groups and Laplacians.* Geometry and analysis on manifolds, Lect. Notes Math. **1339** (Berlin: Springer), (1988), 248-277.

[15] P. Šťovíček. *The Green function for the two-solenoid Aharonov-Bohm effect* Phys. Lett. A **142**, (1989), 5-10.

[16] A. Figà-Talamanca and M. A. Picardello. *Spherical functions and harmonic analysis on free groups.* J. Func. Anal. **47**, (1982), 281-304.

[17] E. Thoma. *Über unitäre Darstellungen abzälbarer, diskreter Gruppen.* Math. Annalen **153**, (1964), 111-138.

[18] T. T.Wu and C. N. Yang. *Concept of nonintegrable phase factors and global formulation of gauge fields.* Phys. Rev. D **12**, (1978) 3845-3857

Evolutionary Algorithms for Constrained Optimization Problems

David Kozub

1st year of PGS, email: zub@linux.fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Martin Holeňa, Institute of Computer Science, AS CR

Abstract. This paper presents an overview of the techniques used to solve constrained optimization problems using evolutionary algorithms. The construction of the fitness function together with the handling of feasible and infeasible individuals is discussed. Approaches using penalty functions, special representations, repair algorithms, methods based on separation of objective and constraints and multiobjective techniques are mentioned.

Abstrakt. Tento příspěvek podává přehled metod pro řešení optimalizačních úloh s omezeními pomocí evolučních algoritmů. Zmíněny jsou některé způsoby vytváření fitness funkce spolu se zpracováním přípustných a nepřípustných jedinců. Zahrnuty jsou přístupy využívající penalizační funkce, speciální reprezentace, opravné algoritmy, metody založené na oddělení účelové funkce a omezení a vícekriteriální metody.

1 Introduction

Evolutionary algorithms have been successfully used in a range of applications. [1] Majority of the papers presented pertain to unconstrained optimization problems. As [2] argues, virtually all real problems are constrained. Thus, the study of constraint-handling methods that can be used with evolutionary algorithms is an important subject.

Evolutionary algorithms are based on a analogy with the evolution process occurring in nature: The individuals have genes that encode the solution. The individuals are compared with others and those that perform better (have higher fitness) get higher probability of propagating their genes into the next generation. The genes of the offspring population are the product of applying genetic operators to the genes of their parent individuals.

For an evolutionary algorithm, the following is needed:

- A representation of the potential solution (an individual).
- A way of initializing the population of the individuals.
- Genetic operators that act on the (parent) population – typically recombination and mutation.
- Selection operator that chooses which individuals propagate to the next generation.

Evolutionary algorithm can be formally defined as follows (based on [1]):

Definition 1. (Evolutionary algorithm) The following algorithm is called an Evolutionary Algorithm:

1. $t \leftarrow 0$
2. initialize:
$$P_0 = \left\{a_0, \ldots, a_{\mu^{(0)}}\right\} \subseteq \mathcal{I}$$
3. while ($\iota\left((P_0, \ldots, P_t)\right) \neq 1$) do
   (a) recombine:
   $$P'_t \leftarrow r^{(t)}_{\phi_r^{(t)}}(P_t)$$
   (b) mutate:
   $$P''_t \leftarrow m^{(t)}_{\phi_m^{(t)}}(P'_t)$$
   (c) select: if $\chi = 1$:
   $$P_{t+1} \leftarrow s^{(t)}_{(\phi_s^{(t)})}(P''_t)$$
   else:
   $$P_{t+1} \leftarrow s^{(t)}_{(\phi_s^{(t)})}(P''_t \cup P_t)$$
   (d) $t \leftarrow t + 1$

where:

- $\mathcal{I} \neq \varnothing$ is the individual space
- $a_0, \ldots, a_{\mu^{(0)}}$ is the initial population
- $\left(\mu^{(i)}\right)_{i \in \mathbb{N}_0}$ is a sequence of the parent population sizes
- $\left(\mu'^{(i)}\right)_{i \in \mathbb{N}_0}$ is a sequence of the offspring population sizes
- $\iota : \left\{ \left(\mathcal{I}^{\mu^{(i)}}\right)_{i=0}^{t} \middle| t \in \mathbb{N}_0 \right\} \rightarrow \{0, 1\}$ is the terminating criterion
- $\chi \in \{0, 1\}$ chooses between (μ, λ) and $(\mu + \lambda)$ selection method
- $\left(r^{(i)}\right)_{i \in \mathbb{N}_0}$ is a sequence of recombination operators:
  $$r^{(i)} : \Xi_r^{(i)} \rightarrow \left[\mathcal{I}^{\mu^{(i)}} \rightarrow \mathcal{I}^{\mu'^{(i)}}\right]$$
  where $\Xi_r^{(i)}$ is the set of recombination parameters and $\theta_r^{(i)} \in \Xi_r^{(i)}$
- $\left(m^{(i)}\right)_{i \in \mathbb{N}_0}$ is a sequence of mutation operators:
  $$m^{(i)} : \Xi_m^{(i)} \rightarrow \left[\mathcal{I}^{\mu'^{(i)}} \rightarrow \mathcal{I}^{\mu'^{(i)}}\right]$$
  where $\Xi_m^{(i)}$ is the set of mutation parameters and $\theta_m^{(i)} \in \Xi_m^{(i)}$

- $\left(s^{(i)}\right)_{i\in\mathbb{N}_0}$ is a sequence of selection operators:

$$s^{(i)} : \Xi_s^{(i)} \rightarrow \left[\mathcal{I}^{\mu'^{(i)}+\chi\mu'^{(i)}} \rightarrow \mathcal{I}^{\mu^{(i+1)}}\right]$$

  where $\Xi_s^{(i)}$ is the set of mutation parameters and $\theta_s^{(i)} \in \Xi_s^{(i)}$

In this paper we focus on applying evolutionary algorithms to constrained optimization problems. By this we mean the following:

$$\min_{x\in\Omega} f(x) \tag{1}$$

subject to:

$$g_i(x) \leq 0 \quad \forall i \in \{1,\ldots,n_g\} \tag{2}$$
$$h_j(x) = 0 \quad \forall j \in \{1,\ldots,n_h\} \tag{3}$$

where the set Ω is the *search space*. Let n denote the total number of constraints:

$$n = n_g + n_h$$

The constraints (3) and (2) implicitly define the feasible set Φ:

$$\Phi = \left\{x \in \Omega | g_i(x) \leq 0 \wedge h_j(x) = 0 \right.$$
$$\left. \forall i \in \{1,\ldots,n_g\}, \forall j \in \{1,\ldots,n_h\} \right\}$$

We make no additional assumptions about the feasible set. In general it can be a non-convex, even a disconnected set.

Defining $\Upsilon = \Omega - \Phi$, it can be stated that the search space Ω is partitioned into two disjoint sets: the feasible set Φ and the infeasible set Υ.

The level of violation of the constraints (2) and (3) by a point $x \in \Omega$ can be measured as follows:

$$G_i(x) = \max\{0, g_i(x)\} \tag{4}$$
$$H_j(x) = |h_j(x)| \tag{5}$$

Note that for all $x \in \Phi$

$$G_i(x) = 0$$
$$H_j(x) = 0$$

for all $i \in \{1,\ldots,n_g\}$, $j \in \{1,\ldots,n_h\}$.

An equality constraint $h_j(x) = 0$ can be transformed into inequality constraints in the following way:

$$|h_j(x)| \leq \varepsilon$$

where ε is a small constant specifying the tolerance.

This approach allows the equality constraints to be treated as inequalities, which can be useful for methods that do not treat equality constraints separately.

2 Fitness function

The fitness function is a function $F : \mathcal{I} \to \mathbb{R}$ that evaluates the individuals according to how well they solve given problem.

The design of the fitness function can be a non-trivial task even for an unconstrained problem. In case of constrained problems, the design of a good fitness function is even more difficult. In [2] the following points guiding the design of the fitness function are listed:

1. How should two feasible points be compared?
2. How should two infeasible points be compared?
3. How are the functions for feasible and infeasible points related? Should feasible points be always "better" than infeasible ones?
4. Should infeasible points be considered harmful and removed from the population?
5. Should infeasible points be "repaired"?
6. If individuals are repaired, should this repaired individual be used only for evaluating its fitness (*Baldwin effect*) or should the individual be replaced (*Lamarckian evolution*)?
7. Should infeasible individuals be penalized?
8. Should the algorithm start with a feasible population and keep the feasibility throughout the run of the algorithm?

During the run of the algorithm, the population can generally contain both feasible and infeasible individuals. In the end though, the answer must be a feasible solution, as the infeasible individual, no matter its fitness from the point of view of the evolutionary algorithm, is not a solution to the original problem.

An obvious method of ensuring this works by removing all the infeasible solutions, so that the population never contains an infeasible individual. While this method has been used, in many problems it does not work. (See section 3 for more information on this approach.)

This leads to the conclusion that the evolutionary algorithm should allow the infeasible individuals in the population. Because of this, a decision has to be made on how to compare the feasible and the infeasible individuals.

One way to tackle this task is to define the fitness function as follows:

$$F(x) = \begin{cases} F_\Phi(x) & x \in \Phi \\ F_\Upsilon(x) & x \in \Upsilon \end{cases} \tag{6}$$

When evaluating F_Φ, the actual value of the constraints should not be important, as the point is in the feasible set. When evaluating F_Υ, the question is if the value of the objective function f should be taken into account. F_Υ should react to the fact that the solution is not feasible and direct the search into the feasible set. Yet, should it be

based on the amount of the violation, or should it only reflect the number of violated constraints?

While the inclusion of the objective f in F_Υ might help guide the search, sometimes (in case the objective is not defined outside of the feasible region Φ) this is not possible.

It should be noted that in some evolutionary algorithms the fitness function is not explicitly needed. For example, if the evolutionary algorithm uses the tournament selection, all that is needed is an ordering relation defined over the individual space $\mathcal{I}$. Still, this does not relieve us of the burden of satisfactorily answering the aforementioned questions.

An overview of some of the methods that were used to solve constrained optimization problems follows. The methods differ by how they answer the aforementioned questions.

3 Penalty functions

The oldest and most common approach to solving constrained optimization problems using evolutionary algorithms is the use of a penalty function. The method is based in the idea of adding to the objective function f a function that penalizes solutions laying in the infeasible set, thus decreasing their fitness.

There are two basic options: *interior penalty functions* – this approach starts from a feasible solution and the penalty function is defined so that its value approaches to infinity as the solution moves towards the boundary of the feasible set, and *exterior penalty functions* – this approach starts from any (generally infeasible) point in the search space and the penalty is used to guide the search into the feasible set.

An advantage of the exterior approach is that it does not require an initial feasible population.

The generic formula for the fitness function with an exterior penalty is:

$$F(x) = f(x) + P^{(t)}(x) \tag{7}$$

where $P^{(t)} : \mathcal{I} \to \langle 0, +\infty)$ is the penalty function satisfying for all $x \in \Phi$ and for all $t \in \mathbb{N}_0$:

$$P^{(t)}(x) = 0$$

A problem with this approach is the choice of the value of the penalty: Too small penalty value does not discourage the algorithm from the infeasible set, possibly resulting in an infeasible optimum. On the other hand, too high penalty value might prohibit the algorithm from crossing the feasible set boundary (which might be useful or even necessary in case the feasible set is non-convex or disconnected) and from exploring the boundary of the feasible set.

In [3] author suggests the relation between an infeasible individual and the feasible set plays an important role in the penalization. There are several ways how this relationship could be reflected in the penalty function:

1. the penalty is constant – the individual is being penalized for being infeasible

2. the penalty reflects the amount of constraint violation

3. the penalty reflects the effort needed to make the individual feasible

This method was advanced in several directions in order to tackle this issue:

static penalties In this approach, the value of the penalties is independent of the generation number. Typical choice for $P^{(t)}$ is:

$$P^{(t)}(x) = \sum_{i=1}^{n_g} a_i G_i(x)^\beta + \sum_{j=1}^{n_h} b_j H_j(x)^\gamma$$

with $\beta, \gamma \in \{1, 2\}$, a_i, b_i positive constants called *penalty factors* and G_i, H_j as defined in (4) and (5).

dynamic penalties In this approach, the value of the penalties is dependent on the generation number. Typically, the penalties rise over time. This enables the population to explore the search space (low penalties) and eventually move into the feasible set. An example of this approach is:

$$P^{(t)}(x) = (ct)^\alpha \left(\sum_{i=1}^{n_g} a_i G_i(x)^\beta + \sum_{j=1}^{n_h} b_i H_j(x)^\gamma \right)$$

annealing penalties This method was inspired by simulated annealing: The penalties change when the algorithm gets stuck in a local optimum. The penalty rises over time to penalize infeasible solutions in the end of the run of the algorithm.

adaptive penalties Within this approach, the penalty uses the previous states of the algorithm: The penalty with respect to a constraint is increased if all the individuals in the previous generation were infeasible. The penalty is decreased if all the individuals in the previous generation were feasible.

co-evolutionary penalties In this approach, there are more populations, for example a population for the evolution of solutions and a population for the evolution of the penalty factors. A co-evolution scheme is then used.

death penalty This is a simple method that works by eliminating all the non-feasible individuals form the population. While it can be easily implemented, it tends to work only if the feasible set is a reasonably large subset of the search space and when the feasible set is convex. [2]

Another approach in this category works by focusing the search on the boundary of the feasible set Φ. According to [1], many real-world tasks have optimum for which at least some constraints are active, so the focus on the boundary of the feasible set seems reasonable. The way the border is explored is by varying a penalty and thus forcing the individuals to cross between the feasible and the infeasible set.

The main disadvantage of the penalty methods is their dependency on multiple parameters. While some guidance has been provided, often the parameters have to be empirically determined. [1] Also, penalty methods often do not perform well when the problem is highly-constrained or when the feasible set is disconnected. [2]

4 Special representations

This approach tackles the optimization problem by designing a special, problem-dependent, representation of the individuals. This in turn calls for special operators to be used on those individuals. The operators used typically preserve the feasibility of the population. The motivation behind this approach is to simplify the feasible set Ω.

The representation is problem-specific. While the approach was successfully used on specific problems, it is difficult to generalize this approach.

5 Repair algorithms

This approach works by repairing infeasible individuals. Two ways are possible: The repaired individual is used only to evaluate the fitness of the original, or the infeasible individual is replaced with the repaired one.

The resulting individual is not necessarily feasible, but the amount of constraint violation is reduced.

This method was generalized into the area of constrained multiobjective evolutionary optimization in [4] and [5].

The repair approach often has problems with keeping the diversity of the population. Also, the repair operator can sometimes introduce a strong bias into the search process. [3]

6 Separation of constraints and objectives

The following approaches do not mix the objective and the constraints together. There are several different methods reported in [2] and [3].

6.1 Superiority of feasible points

In this approach feasible individuals are always considered superior to infeasible ones.

One way to ensure this is to map the objective function onto a bounded-above interval, e. g. $(-\infty, 1)$ and specify the fitness function like:

$$F(x) = \begin{cases} f(x) & x \in \Phi \\ L(x) & x \in \Upsilon \end{cases} \tag{8}$$

where $L : \Upsilon \to (1, +\infty)$ is a function measuring the level of constraint violation.

An interesting adaptation that does not require the objective to be bounded-above is:

$$F(x) = \begin{cases} f(x) & x \in \Phi \\ f_{max}^{(t)} + L(x) & x \in \Upsilon \end{cases} \tag{9}$$

where $f_{max}^{(t)} = \max_{x \in P_{(t)} \cap \Phi} f(x)$ and $L : \Upsilon \to \mathbb{R}^+$ is a function measuring the level of constraint violation.

A different way to ensure the feasible points are always superior is to use tournament selection with the rules (x and y denotes the individuals being compared) from table 1.

Table 1: Tournament selection for the superiority of feasible points method

$x \in \Phi$	$y \in \Upsilon$	x is preferred over y
$x \in \Upsilon$	$y \in \Phi$	y is preferred over x
$x \in \Phi$	$y \in \Phi$	decide based on $f(x)$ and $f(y)$
$x \in \Upsilon$	$y \in \Upsilon$	decide based on constraint violation

6.2 Behavioral memory

This method requires a linear ordering of the constraints. Then it proceeds as follows:

1. initialize the population randomly
2. evolve the individuals to minimize the violation of the first constraint; stop when the percentage of individuals feasible with respect to the first constraint surpasses given percentage
3. $j \to 2$
4. while $j \leq n$ do:
   (a) evolve the individuals to minimize the violation of the j-th constraint while removing individuals which do not satisfy any of the constraints $1 \ldots j$; stop when the percentage of individuals feasible with respect to the j-th constraint surpasses given percentage
   (b) $j \to j + 1$
5. evolve the individuals to minimize the objective f while removing infeasible individuals from the population (*death penalty* – see section 3)

This approach is similar to the lexicographic ordering approach mentioned in subsection 7. A drawback is that the initial ordering of the constraints influences the results obtained.

Those methods do not work well when the size of the feasible set is relatively small (when the constraints are difficult to satisfy). Another problem mentioned in [3] is the difficulty of maintaining the diversity of the population.

An interesting point to make is that those approaches never evaluate the objective on infeasible points, making it interesting for problems with hard constraints.

7 Multiobjective techniques

The technique works by transforming the original constrained optimization problem into an unconstrained multiobjective problem, turning the original constraints into additional objectives. The problem (1) – (3) turns into:

$$\min_{x \in \Omega}(f, G_1(x), \ldots, G_{n_g}(x), H_1(x), \ldots, H_{n_h}(x)) \tag{10}$$

Table 2: Tournament selection for the min-max approach in [6]

$x \in \Phi$	$y \in \Upsilon$	x is preferred over y
$x \in \Upsilon$	$y \in \Phi$	y is preferred over x
$x \in \Phi$	$y \in \Phi$	decide based on $f(x)$ and $f(y)$
$x \in \Upsilon$	$y \in \Upsilon$	select the individual having the smallest maximal constraint violation.

The ideal solution of (10) is an $x^{ideal} \in \Phi$ such that:

$$\begin{aligned} f(x^{ideal}) &= min_{x \in \Phi} f(x) \\ G_i(x^{ideal}) &= 0 \ \forall i \in \{1, \ldots, n_g\} \\ H_j(x^{ideal}) &= 0 \ \forall j \in \{1, \ldots, n_h\} \end{aligned}$$

Unlike in actual multiobjective optimization, here we are not interested in finding good trade-offs between the objectives (the original objective (1) and the constraints): Any feasible point might be acceptable, no matter the actual value of the constraint violation values. On the other hand, a global minimum that lies in the infeasible set is no solution to the original problem, even if it means a good trade-off in the multiobjective problem.

In [6] a min-max-like approach was described: The evolutionary algorithm uses the tournament selection with the rules (x and y denotes the individuals that are compared) according to table 2.

8 Conclusion

This paper presents several ways of handling constrains together with evolutionary optimization. Majority of the approaches does need to evaluate the objective outside the feasible set, which renders the methods unusable for constraints that cannot be relaxed. Handling such problems with evolutionary algorithms seems therefore like an interesting option for further research.

References

[1] C. A. Coello Coello, D. A. Van Veldhuisen, G. B. Lamont. *Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers*, 2002.

[2] Z. Michalewicz, M. Schmidt *Evolutionary Algorithms and Constrained Optimization. Evolutionary Optimization*, New York, Kluwer Academic Publishers, pp. 57–86, 2003.

[3] C. A. Coello Coello, *Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. Computer Methods in Applied Mechanics and Engineering*, vol. 191, pp. 1245–1287, 2002.

[4] K. Harada, J. Sakuma, K. Ikeda, I. Ono, S. Kobayashi, *Local search for multi-objective function optimization: Pareto descent method*. In: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, New York, NY, ACM Press, pp. 659–666, 2007.

[5] K. Harada, J. Sakuma, I. Ono, S. Kobayashi *Constraint-Handling Method for Multi-objective Function Optimization: Pareto Descent Repair Operator*. In: *Proceedings of the Evolutionary Multi-Criterion Optimization (EMO 2007)*, Springer, Berlin, 156–170, 2007.

[6] F. Jiménez, J. L. Verdegay *Evolutionary Techniques for Constrained Optimization Problems*. In: *Seventh European Congress on Intelligent Techniques and Soft Computing*, Springer, Aachen, 1999.

Equivalence Problem in Compositional Models*

Václav Kratochvíl

3rd year of PGS, email: velorex@utia.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Radim Jiroušek, Institute of Information Theory and Automation, AS CR

Abstract. Compositional models theory (originally developed by Radim Jiroušek) represents an alternative approach to Bayesian networks. This text should familiarize the reader with new results in this theory, namely with partial solution of the equivalence problem (in the sense independence equivalence). Four different operations on persegram which preserve independence model are introduced. By help of these operations we may generate the class of persegrams equivalent to a given one.

Abstrakt. Teorie kompozicionálních modelů (zformulovaná Radimem Jirouškem) představuje určitou alternativu k Bayesovským sítím. V tomto článku jsou uvedeny nejnovější poznatky v této oblasti, konkrátně částečné řešení problému ekvivalence (ve smyslu nezávislostní ekvivalence). Zavádíme čtyři různé operace na persegramu (dvě jsou zveřejněny poprvé) které zachovávají nezávislostní model a umožňují generovat různé ekvivalentní persegramy.

1 Introduction

The ability to represent and process multidimensional probability distributions is a necessary condition for application of probabilistic methods in Artificial Intelligence. Among the most popular approaches are the methods based on Graphical Markov Models, e.g., Bayesian Networks. An alternative approach to Graphical Markov Models are the so-called Compositional models, which try to be more efficient than Bayesian networks (more efficient in computations, etc.). Nevertheless, the theory has not been finished yet and many substantial problems remain to be solved.

2 Compositional Models

Bayesian networks may be defined as a multidimensional distribution factorizing with respect to an acyclic directed graph. Alternatively, the Bayesian network may be uniquely defined by its graph and an appropriate system of low-dimensional (oligodimensional) *conditional distributions*. Similarly, Compositional models are defined as a multidimensional distribution assembled from a sequence of oligodimensional *unconditional distributions*, with the help of operators of composition. The main advantage of both approaches lies in the fact that oligodimensional distributions could be easily stored in computer

*The research was partially supported by Ministry of Education of the Czech Republic under grants no 1M0572 and 2C06019.

memory. However, computing with a multidimensional distribution that is split into many pieces is exceptionally complicated. The advantage in comparison with Bayesian networks consists in the fact that compositional models explicitly express some marginals, whose computation in the Bayesian network may be demanding.

2.1 Notation and Basic Properties

In this paper we consider a system of finite-valued random variables with indices from a non-empty finite set N. All probability distributions discussed in the paper will be denoted by Greek letters. For $K \subset N$, $\pi(x_K)$ denotes a distribution of variables $\{X_j\}_{j\in K}$.

Consider a distribution $\pi(x_K)$ and three disjoint subsets $A, B, C \subset K$ such that $A \neq \emptyset \neq B$. $A \perp\!\!\!\perp B|C[\pi]$ denotes that two groups of variables $\{X_j\}_{j\in A}$ and $\{X_j\}_{j\in B}$ are conditionally independent given $\{X_j\}_{j\in Z}$. Suppose, that $L \subset K$, we denote its corresponding marginal distribution either $\pi(x_L)$, or $\pi^{\downarrow L}$. These symbols are used to highlight the variables for which the marginal distribution is defined.

To describe how to compose low-dimensional distributions to get a distribution of a higher dimension we use the following operator of composition.

Definition 1. For arbitrary two distributions $\pi(x_K)$ and $\kappa(x_L)$ their *composition* is given by the formula

$$\pi(x_K) \triangleright \kappa(x_L) = \begin{cases} \frac{\pi(x_K)\kappa(x_L)}{\kappa(x_{K\cap L})} & \text{when } \pi^{\downarrow K\cap L} \ll \kappa^{\downarrow K\cap L}, \\ \text{undefined} & \text{otherwise,} \end{cases}$$

where the symbol $\pi(x_M) \ll \kappa(x_M)$ denotes that $\pi(x_M)$ is *dominated* by $\kappa(x_M)$, which means (in the considered finite setting)

$$\forall x_M \in \times_{i\in M}\mathbf{X}_i; (\kappa(x_M) = 0 \Longrightarrow \pi(x_M) = 0).$$

The result of the composition (if defined) is a new distribution. We can iteratively repeat the process of composition to obtain a multidimensional model. This is why these multidimensional distributions are called *compositional models*. To describe such a model it is enough to introduce an ordered system of low-dimensional distributions $\pi_1, \pi_2, \ldots, \pi_n$. If all compositions are defined, we view this ordered system as a *generating sequence*, in which the composition operator is applied from left to right:

$$\pi_1 \triangleright \pi_2 \triangleright \pi_3 \triangleright \ldots \triangleright \pi_{n-1} \triangleright \pi_n = (\ldots((\pi_1 \triangleright \pi_2) \triangleright \pi_3) \triangleright \ldots \triangleright \pi_{n-1}) \triangleright \pi_n.$$

In that case we say that a generating sequence defines (or represents) a multidimensional compositional model. From now on, we consider generating sequences $\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})$ which define a distribution

$$\pi_1(x_{K_1}) \triangleright \pi_2(x_{K_2}) \triangleright \ldots \triangleright \pi_n(x_{K_n}).$$

Therefore, whenever distribution π_i is used, we assume it is defined for variables $\{X_j\}_{j\in K_i}$. In addition, each set K_i can be divided into two disjoint parts. We denote them R_i and S_i with the following sense. R_i denotes variables from K_i emerging in the sequence (meaning from left to right) the first time. R_i denotes the already used.

$$R_i = K_i \backslash (K_1 \cup \ldots \cup K_{i-1}), S_i = K_i \cap (K_1 \cup \ldots \cup K_{i-1}) .$$

In the proofs of the upcoming lemmata will be used the following assertion, which is proved e.g. in [1].

Lemma 2. *Let $M \subseteq K_1 \cup K_2$. If $M \supseteq K_1 \cap K_2$ then for any probability distributions $\pi_1(x_{K_1})$ and $\pi_2(x_{K_2})$*

$$(\pi_1 \triangleright \pi_2)^{\downarrow M} = \pi_1^{\downarrow K_1 \cap M} \triangleright \pi_2^{\downarrow K_2 \cap M}.$$

2.2 Perfect Sequence Models

Not all generating sequences are equally efficient in their representations of multidimensional distribution. Among them, so-called perfect sequences hold an important position. From the original definition (e.g. in [1]) one can hardly see the importance of this generating sequences class. Instead, for the purpose of this text let us define it by another equivalence property, which is more suitable for our needs.

Definition 3. A generating sequence $\pi_1, \pi_2, \ldots, \pi_n$ is perfect *iff* all the distributions π_i are marginal to the represented distribution, i.e., for all $i = 1, 2, \ldots, n$

$$(\pi_1 \triangleright \ldots \triangleright \pi_n)^{\downarrow K_i} = \pi_i.$$

Perfect sequences have many pleasant properties which are advantageous for multidimensional distributions representation. One of them says that, for a perfect sequence model, all distributions in model are pair-wise consistent. This feature is in other parts of this paper highly used.

2.3 Conditional Independencies

It is well-known that one can read conditional independence relations of a Bayesian network from its graph. A similar technique is used in compositional models. An appropriate tool for this is a **persegram**. Persegram is used to visualize the structure of a compositional model and is defined bellow.

Definition 4. Persegram of a generating sequence is a table in which rows correspond to variables (in an arbitrary order) and columns to low-dimensional distributions; ordering of the columns corresponds to the generating sequence ordering. A position in the table is marked if the respective distribution is defined for the corresponding variable. Markers for the first occurrence of each variable (i.e., the leftmost markers in rows) are squares (we call them **box-markers**) and for other occurrences there are **bullets**.

All persegrams discussed in the paper are denoted by P, modified by P'. Since i-th column corresponds to π_i, we denote the markers in i-th column K_i. In accordance with the other marking of variables in the i-th distribution $\pi_i(x_{K_i})$, box-markers in i-th column are denoted like R_i and bullets like S_i. $K_i = R_i \cup S_i$.

Example 5. Let $\pi_1(x_{K_1}), \ldots, \pi_6(x_{K_6})$ be a generating sequence. $K_1 = \{1, 4\}, K_2 = \{4, 2\}, K_3 = \{2, 5\}, K_4 = \{5, 3\}, K_5 = \{5, 7\}, K_6 = \{4, 5, 6\}$. Then the corresponding persegram P is in Figure 1.

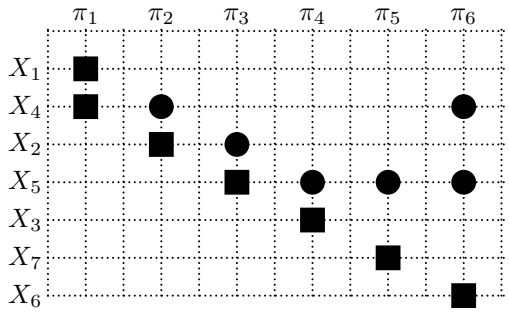


Figure 1: Persegram corresponding to the model in Example 5

Like the Bayesian networks, conditional independence of groups of variables is indicated by the absence of a *trail connecting relevant markers and avoiding the respective subset* which is defined below.

Definition 6. Consider a generating sequence $\pi_1(x_{K_1}), \ldots, \pi_n(x_{K_n})$, its corresponding persegram and a subset $Z \subset K_1 \cup \ldots \cup K_n$. A sequence of markers $m_0, \ldots, m_t$ is called a **Z-avoiding trail** that connects m_0 and m_t if it meets the following 4 conditions:

1. for each $s = 1, \ldots, t$ a couple (m_{s-1}, m_s) is in the same row (i.e., horizontal connection) or in the same column (vertical connection);
2. each vertical connection must be adjacent to a box-marker (one of the markers is a box-marker);
3. no horizontal connection corresponds to a variable from X_Z;
4. vertical and horizontal connections regularly alternate with the following possible exception: two vertical connections may be in direct succession if their common adjacent marker is a box-marker of a variable from X_Z;

If a Z-avoiding trail connects two-box markers corresponding to variables X_j and X_k, we also say that these **variables are connected by a Z-avoiding trail**. Such situations will be denoted $X_j \rightsquigarrow_Z X_k$.

Theorem 7. *Consider a generating sequence $\pi_1(x_{K_1}), \ldots, \pi_n(x_{K_n})$, and three disjoint subset $J_1, J_2, Z \subset K_1 \cup \ldots \cup K_n$ such that $J_1 \neq \emptyset \neq J_2$. If there does not exist a trail $X_j \rightsquigarrow_Z X_k$ in the corresponding persegram with $j \in J_1$ and $k \in J_2$ then:*

$$X_{J_1} \perp\!\!\!\perp X_{J_2} \,|\, X_Z[\pi_1 \rhd \ldots \rhd \pi_n].$$

Definition 8. Let P be a persegram over N. The formal **independence model** $\mathcal{M}_P = \{\langle A, B|C\rangle \in \mathcal{T}(N); A \perp\!\!\!\perp B|C[P]\}$ is a model induced by persegram P, where $\mathcal{T}(N)$ is a system of all triples of disjoint subsets of N where $A \neq \emptyset \neq B$.

3 Equivalence problem

By the equivalence problem we understand the problem how to recognize whether two given persegrams P, Q over N induce the same independence model. It is of special

importance to have an easy rule to recognize that two persegrams are equivalent in this sense and an easy way to convert P into Q in terms of some elementary operations on persegrams. Another very important aspect is the ability to generate all persegrams which are equivalent to a given persegram. For all these problems, the last one is partially solved in this paper.

Definition 9. Persegrams P, Q (over the same variable set N) are called **independently equivalent**, if they induce the same independence model $\mathcal{M}_P = \mathcal{M}_Q$.

Like in Bayesian networks, it may happen that different persegrams induce the same independence model.

Example 10. 1. The following example is simple: $N = \{a, b\}$ and the following two persegrams P, Q:

P: Q:

$\mathcal{M}_P = \mathcal{M}_Q = \emptyset$ in this case.

2. On the other hand, the persegrams which have the same variable sets in columns in different order do not have to be equivalent. Let $N = \{a, b, c\}$ and consider the following persegrams:

P: Q:

$a \perp\!\!\!\perp b|\emptyset[P]$ but $a \not\perp\!\!\!\perp b|\emptyset[Q]$. On the contrary, $a \not\perp\!\!\!\perp b|c[P]$ but $a \perp\!\!\!\perp b|c[Q]$. The order of the columns in persegram is important.

Four different simple operations on persegram preserving independence model were discovered. We call them IE operations (Independence Equivalent). These operations can be divided into two groups according to the behavior of columns in a persegram: Either changing their order (this group is called *permutations*) or adding/removing them (*extensions/reductions*). Let us note, that when these operations are applied on compositional model (its persegram), its generating sequence is accordingly modified.

To facilitate the reader survey, basic overview of the mentioned operations is presented in the form of definition.

Definition 11. Let P be a persegram over N and two adjacent columns $i, i+1$ with K_i, K_{i+1} markers. The so called **IE operations** are the following set of operations with columns.

- **Independent permutation** is swapping of columns $i, i+1$ when no box-marker turns into bullet and vice-versa. ($\cup_{j=1}^{i-1} Kj \supseteq K_i \cap K_j$.)
- **Intersection permutation** is swapping of columns $i, i+1$ if all their bullets belong to their intersection ($S_i \cup S_{i+1} \subseteq K_i \cap K_{i+1}$).
- Removing of a column containing bullets only is called **Bullets extension/reduction** ($K_i = S_i$).
- Removing of a column i, which is a subset of the column $i+1$ that has box-markers elsewhere only, is called **Subset extension/reduction**. ($K_i = S_{i+1}$.)

Graphic representation of a compositional model - persegram is inherently connected to real data - distributions. If one applies the above defined IE operations to persegram (assume that we have already proved, these operations preserve independence model), we know that modified persegram has the same "power" as the original one - It expresses the same (un)conditional independencies (or dependencies). However, imagine that we change the order of distributions (or add/remove some) in the generating sequence as well as in the corresponding persegram. Will the resulting multidimensional distribution be the same?

In other words, first, identity of independence models has to be proven. Then, we have to show that multidimensional distributions represented by the original and the modified generating sequence are equal. Denote by $\pi_1, \pi_2, \ldots \pi_n$ the original generating sequence. We can iteratively repeat IE operations to obtain a new multidimensional model represented by sequence $\pi'_1, \pi'_2, \ldots \pi'_m$. We need to prove $\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n = \pi'_1 \triangleright \pi'_2 \triangleright \ldots \triangleright \pi'_m$. Or, if that is not valid in general, under what conditions.

In order to simplify the following lemmata we will work with the model where generating sequence consists of three distributions π_1, π_2, π_3. This simplification is not in any way at the expense of universality. (π_1 can be internally composed from several distributions and π_1, π_2, π_3 can be a beginning of much longer sequence.)

Lemma 12. *(Independent permutation) If* $K_1 \supseteq (K_2 \cap K_3)$ *then* $\pi_1 \triangleright \pi_2 \triangleright \pi_3 = \pi_1 \triangleright \pi_3 \triangleright \pi_2$.

Proof of this assertion can be found for example in [1]. The declaration of this lemma can be translated into the language of persegrams as following: *"Two columns in persegram can be swapped, if no bullet turns into box-marker and vice-versa."*

The proof of the assertion that this operation preserves the independence model is obvious: If no box-marker turns into a bullet and vice-versa, then all Z-avoiding trails from definition 6 are maintained. (The vertical connections are moved with swaped columns and the horizontal ones shortened/extended.)

Lemma 13. *(Intersection permutation) If* π_2 *and* π_3 *are consistent then*

$$S_2 \cup S_3 \subseteq K_2 \cap K_3 \Longrightarrow \pi_1 \triangleright \pi_2 \triangleright \pi_3 = \pi_1 \triangleright \pi_3 \triangleright \pi_2. \tag{1}$$

Remark 14. The condition of lemma 13 is given in the form $S_2 \cup S_3 \subseteq K_2 \cap K_3$ since it seems to be closer to the verbal designation of condition: "*All swapped distributions bullets must be included in their intersection.*" However, for the purposes of the proof, we will rewrite it into its equivalent form. The idea is outlined in the following form.

$$S_2 \cup S_3 \subseteq K_2 \cap K_3 \Leftrightarrow \left\{ \begin{array}{l} S_2 \subseteq K_2 \cap K_3 \\ S_3 \subseteq K_2 \cap K_3 \Leftrightarrow S_3 = K_2 \cap K_3 \end{array} \right\} \Leftrightarrow S_2 \subseteq S_3 \subseteq K_2.$$

Proof. (*lemma 13*) First, let us show, that under given assumptions, $\pi_1 \triangleright \pi_2 \triangleright \pi_3$ is undefined *iff* $\pi_1 \triangleright \pi_3 \triangleright \pi_2$ is undefined. From the definition of the operator we know that $\pi_1 \triangleright \pi_2 \triangleright \pi_3$ is not defined *iff*

$$\pi_1^{\downarrow K_1 \cap K_2} \not\ll \pi_2^{\downarrow K_1 \cap K_2} \tag{2}$$

or

$$(\pi_1 \triangleright \pi_2)^{\downarrow (K_1 \cup K_2) \cap K_3} \not\ll \pi_3^{\downarrow (K_1 \cup K_2) \cap K_3} \tag{3}$$

Analogously, $\pi_1 \triangleright \pi_3 \triangleright \pi_2$ is not defined *iff*

$$\pi_1^{\downarrow K_1 \cap K_3} \not\ll \pi_3^{\downarrow K_1 \cap K_3} \tag{4}$$

or

$$(\pi_1 \triangleright \pi_3)^{\downarrow (K_1 \cup K_3) \cap K_2} \not\ll \pi_2^{\downarrow (K_1 \cup K_3) \cap K_2} \tag{5}$$

Because of the remark 14:

$$\begin{aligned} K_1 \cap K_3 &= K_1 \cap S_3 \\ &= (K_1 \cap S_2) \cup (K_1 \cap (R_2 \cap S_3)) \\ &= (K_1 \cap S_2) \cup (K_1 \cap R_2') \\ &= K_1 \cap S_2 = K_1 \cap (K_1 \cap K_2) = K_1 \cap K_2 \end{aligned} \tag{6}$$

and

$$\begin{aligned} (K_1 \cup K_3) \cap K_2 &= (K_1 \cap K_2) \cup (K_3 \cap K_2) \\ &= (K_1 \cap K_3) \cup (K_2 \cap K_3) \\ &= (K_1 \cup K_2) \cap K_3 = S_3. \end{aligned} \tag{7}$$

Regarding the fact that in our case π_2 and π_3 are consistent and the fact that $K_1 \cap K_2 = K_1 \cap K_3$, (2) is equivalent to (4). Since $K_1 \cup K_2 \supseteq S_3 \supseteq K_1 \cap K_2$ and $K_1 \cup K_3 \supseteq S_3 \supseteq K_1 \cap K_3$ we can apply lemma 2 getting

$$(\pi_1 \triangleright \pi_2)^{\downarrow S_3} = \pi_1^{\downarrow S_3} \triangleright \pi_2^{\downarrow S_3} = \pi_1^{\downarrow S_3} \triangleright \pi_3^{\downarrow S_3} = (\pi_1 \triangleright \pi_3)^{\downarrow S_3},$$

where the second equality follows from the consistency of π_2 and π_3. Thus we got that (3) is equivalent to (5) and both conditions coincide.

Now, assume that both expressions in formula (1) are defined. Because of (6), (7) and the fact that π_2 and π_3 are consistent, the expressions

$$\pi_1 \triangleright \pi_2 \triangleright \pi_3 = \frac{\pi_1 \pi_2 \pi_3}{\pi_2^{\downarrow K_1 \cap K_2} \pi_3^{\downarrow K_3 \cap (K_1 \cup K_2)}},$$

$$\pi_1 \triangleright \pi_3 \triangleright \pi_2 = \frac{\pi_1 \pi_2 \pi_3}{\pi_3^{\downarrow K_1 \cap K_3} \pi_2^{\downarrow K_2 \cap (K_1 \cup K_3)}}$$

are mutually equivalent, which finishes the proof. □

Lemma 15. *Let P be a persegram. If P' arises from P by applying of Intersection permutation then $\mathcal{M}_P = \mathcal{M}_{P'}$.*

Proof. Let P is a persegram over variable set N. Suppose, that two adjacent columns i, $i+1$ meet the condition $K_i \cap K_{i+1} = S_i \cup S_{i+1}$.

We have to consider the following two situations:

$$(a)\, S_i = S_{i+1},$$
$$(b)\, S_i \subset S_{i+1}.$$

It is needless to consider

$$(c)\, S_i \supset S_{i+1}$$

because it is in dispute with assumptions.

Consider the situation (a) where $S_i = S_{i+1}$ (i.e the intersection contains bullets only). By swapping the corresponding columns, no bullet will change into a box-marker and vice-versa. It passes into proof of *Independent permutation*, which is evident.

Now consider the situation (b). Regarding the fact, that (un)conditional independencies in persegram are indicated by absence of corresponding Z-avoiding trails, we have to prove, that the sequence of markers remain Z-avoiding trail after *Intersection permutation*.

Suppose, that there is a Z-avoiding trail which passes through swapped columns. Horizontal parts remain the same. Vertical parts have to be connected with a box-marker. Assume, that the original trail fulfilled all the conditions imposed. After reordering, the corresponding vertical connection may contain:

- two box-markers $\rightarrow$ In this case everything is all right.
- one box-marker $\rightarrow$ In this case everything is all right.
- no box-marker $\rightarrow$ In this case vertical connection contains two bullets. According to the assumptions they belong into both columns. Hence, vertical connection can be transferred into the other column. Then the vertical connection will contain, at least, one box-marker there, which corresponds to the box-marker from the original vertical connection.

□

In accordance with the definition 11, lemmata about *Bullets extension/reduction* should follow now. The first of them can be found e.g. in [1].

Suppose that we remove/add column of bullets. It is easy to prove that this operation preserves the Independence model. According to the definition 6 of Z-avoiding trail, no vertical connection of that trail can pass through column without any box-marker. Therefore, the removal/addition of such column will bring no change in its independence model. Now proceed with the last of IE operations - *Subset extension/reduction*.

Lemma 16. *(Subset extension/reduction) If π_2 and π_3 are consistent then*

$$K_2 = S_3 \Longrightarrow \pi_1 \triangleright \pi_2 \triangleright \pi_3 = \pi_1 \triangleright \pi_3. \tag{8}$$

Proof. Let us start, again, by showing that, under given assumptions $\pi_1 \triangleright \pi_2 \triangleright \pi_3$ is undefined *iff* $\pi_1 \triangleright \pi_3$ is undefined. From the definition of operator $\triangleright$ follows that $\pi_1 \triangleright \pi_2 \triangleright \pi_3$ is undefined if

$$\pi_1^{\downarrow K_1 \cap K_2} \not\ll \pi_2^{\downarrow K_1 \cap K_2}$$

or

$$(\pi_1 \triangleright \pi_2)^{\downarrow (K_1 \cup K_2) \cap K_3} \not\ll \pi_3^{\downarrow (K_1 \cup K_2) \cap K_3} \tag{9}$$

Because of $(K_1 \cup K_2) \cap K_3 = S_3 = K_2$ and the consistence of π_2 and π_3 , the expression (9) can be rewritten into the following form: $\pi_1^{\downarrow K_2} \triangleright \pi_2 \not\ll \pi_2$. This condition is invalid under any circumstances. Therefore the condition (9) is invalid and under given assumptions, $\pi_1 \triangleright \pi_2 \triangleright \pi_3$ is not defined *iff*

$$\pi_1^{\downarrow K_1 \cap K_2} \not\ll \pi_2^{\downarrow K_1 \cap K_2} \tag{10}$$

Analogously, $\pi_1 \triangleright \pi_3$ is not defined *iff*

$$\pi_1^{\downarrow K_1 \cap K_3} \not\ll \pi_3^{\downarrow K_1 \cap K_3} \tag{11}$$

Under the given assumption $K_2 = S_3$, these two conditions (10), (11) coincide because

$$S_2 = K_1 \cap K_2 = K_1 \cap S_3 = K_1 \cap K_3$$

and π_2, π_3 are consistent.

Now, supposing that both expressions in (8) are defined,

$$\pi_1 \triangleright \pi_2 \triangleright \pi_3 = \frac{\pi_1 \pi_2 \pi_3}{\pi_2^{\downarrow S_2} \pi_3^{\downarrow S_3}} = \frac{\pi_1 \pi_3}{\pi_2^{\downarrow S_2}} = \frac{\pi_1 \pi_3}{\pi_3^{\downarrow K_1 \cap K_3}}$$

which finishes the proof. □

Lemma 17. *Let P be a persegram. If P' arises from P by applying of Subset extension/reduction then $\mathcal{M}_P = \mathcal{M}_{P'}$.*

Proof. This lemma can by proved the same way as lemma 15, or one can realize that Subset extension/reduction can be spread out into Subset permutation and Bullets extension/reduction, where both of them preserve Independence model $\mathcal{M}_P$. □

The previous proof narrows the set of IE operations in three of them (Subset extension/reduction can be omitted since it can be substituted by the sequence of the others).

Lemma 18. *Let P be a persegram. If P' become from P by iterative application of the IE operations then $\mathcal{M}_P = \mathcal{M}_{P'}$.*

Proof. Since all of IE operations preserve the $\mathcal{M}_P$, the proof is clear. □

Lemma 19. *Let $\pi_1, \pi_2, \ldots, \pi_n$ be a perfect sequence. If $\pi'_1, \pi'_2, \ldots, \pi'_m$ is obtained by iterative application of IE operations, then*

$$\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n = \pi'_1 \triangleright \pi'_2 \triangleright \ldots \triangleright \pi'_m$$

Proof. Because $\pi_1, \pi_2, \ldots, \pi_n$ is a perfect sequence, then $\pi_1, \pi_2, \ldots, \pi_n$ are pairwise consistent. □

Example 20. An example of four different persegrams with the same independence model is on Figure 2. They were produced by iterative application of IE operations from the most left persegram.

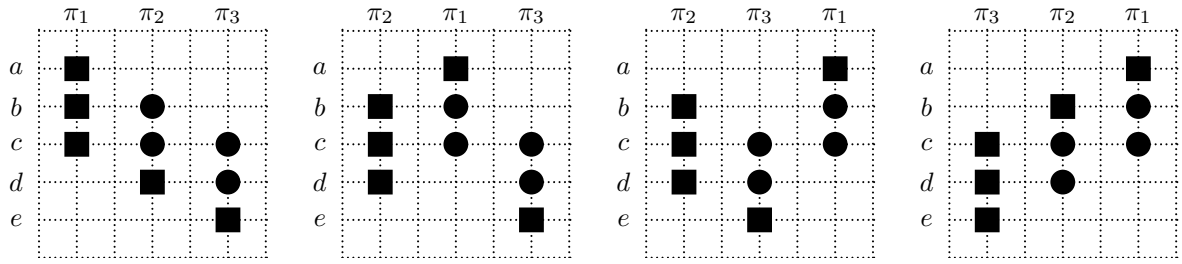


Figure 2: Process of application IE operations in persegram

4 Conclusion

The main achievements of this report are various operations in persegram, denoted as IE operations, which preserve the (un)conditional independencies expressed by persegram. By iterative application of the IE operations we can obtain big amount of various persegrams. However, may we obtain all of them? All persegrams with the same independent model? May two persegram with the same independent model be converted each other by application of IE operations only ?

According to our preliminary studies, the answer is YES. Nevertheless, the corresponding proof has not been finished yet. To do it, a number of different assertions has to be bringed out, but it goes beyond the scope of this paper.

References

[1] R. Jiroušek. *Multidimensional Compositional Models*. Preprint DAR - ÚTIA 2006/4, ÚTIA AV ČR, Prague, (2006).

[2] T. Kočka, R. R. Bouckaert, M. Studený. *On the Inclusion Problem*. Research report 2010, ÚTIA AV ČR, Prague (2001).

[3] M. Studený. *O strukturách podmíněné nezávislosti*. Rukopis série přednášek. Prague (2008).

[4] R. Merris: Graph Theory. Wiley Interscience, New York 2001.

Multiagent Exploation from Renewable Resources

Karel Macek

1st year of PGS, email: macek@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Jaromír Kukal, Department of Software Engineering in Economics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. Real complex dynamic systems are subject of advanced modelling yet. Knowledge, decision and action in such systems are often distributed into a set of individuals called agents. Present paper introduces a simple model of an environment with renewable resources. In this environment, agents are operating, namely they exploit from the resources. The system depends on various parameters and global rules which are assessed and optimized in the second part of the paper.

Abstrakt. Reálné složité dynamické systémy jsou předmětem pokročilého modelování. Znalosti, rozhodování i akce v takových systémech jsou často rozloženy mezi jedinci - agenti. Předložený příspěvek představuje jednoduchý model prostředí s obnovitelnými zdroji. V tomto prostředí působí agenti tak, že čerpají z těchto zdrojů. Systém závisí na několika parametrech, které jsou zkoumány a optimalizovány v druhé části článku.

1 Intoduction

Modeling of dynamic systems has a long tradition and stochastic dynamic programing and control theory has introduced many useful concepts how to act on a system and so influence its behavior [2].

In present, new works occurs dealing with operation of multiagent society on the system. There are two main tendencies: first group of researchers are experts[3] in AI and they intend to introduce lot of logic, communication and so on[4]. On the other hand there are experts who the agents only admit into their sophisticated physical models.

This paper is different. The model is constructed step by step from both points of view and the objective to assess the operation of agents in a particular task is achieved. Section 2 introduces the model and section 3 shows results and conclusion on this model. Section 4 summarizes most interesting results.

2 Model Desription

For a multi-agent system, it is important to specify the environment, its state, flow of time and particular components. The model applied for this work simulates agents' behavior in the environment with renewable resources. The state of the system can represented as

$\mathbb{x} \in \mathbb{R}^{n+m}$ where n is number of resources and m is number of agents; x_j is the amount of material in agent or resource j. The time is discrete and has a finite horizon t_0.

Basic system dynamic cosnsist in discrete steps. In each steps two affairs happen:

1. Natural changes - from states $\mathbb{x}(t)$ evolves temporary new states $\mathbb{x}^{(*)}(t+1)$.
2. Agents' actions - from temporary new states $\mathbb{x}^{(*)}(t+1)$ evolves final new states $\mathbb{x}(t+1)$.

2.1 System Components and their Properties

The system is composed from two kinds of entities, viz. of resources and of agents. Abstraction of both is the class natural object. An natural object contains an amount of material. The capacity have upper and lower limits. The amount of material in natural object j varies in time as a difference equation with limits according following formula:

$$x_j^*(t+1) = \min\left(x_j^{max}, \max(x_j^{min}, x_j(t) \cdot g_j + a_j)\right) \tag{1}$$

where a_j and g_j are parameters specific for each natural object and corresponds to linear (arithmetic) or exponential (geometric) trends.

The system contains a set of resources. A resource is - in fact - a natural objects. A renewable resource have $a_j > 0$ or $g_j > 1$. Possitive a_j stands for resources with regular feed, e.g. for water source. Situation with $g_j > 1$, is typical for living natural objects, e.g. for growing wood in a forest. Resources considered in our simulation are considered to have $g_j > 1$, $a_j = 0$, and $x_j^{min} = 0$. Therefore, if a resource is exploited totaly, it is not able to recover its state.

Second set contained in the system are agents. From the natural object's point of view, it holds $a_j < 0$ and $g_j = 0$. It means that the agents consumes regulary a portion of the material thay contain. However, the agents differ from other natural objects in a more important aspect: they act. In each time step, they choose an action. There are following options what an agent may do

- load the material from the actual resource,
- move to another resource,
- or wait doing nothing.

Nevertheless, there is a condition to act, namely $x_j(t) > 0$ because if $x_j(t) \leq 0$, the agent is not able to consume usual amount a_j and dies.

Waiting is very simple to be implemented and movement as well. All resources are considered to be conected by equaly long way. Hence, the only parameter for movement is which resources is to be visited next.

If agent i loads from resource j, is the situation more complicated. The agent strives to load its maximum, but there are two limits: agent's free capacity and the speed of loading. Nevertheless, the resource may not contain enough material. Therefore the loading is given by following formula:

$$l_i(t) = \min\left(x_j^*(t) - x_j^{min}, \quad \max(x_i^{max} - x_j^*(i), l_i^{max})\right) \tag{2}$$

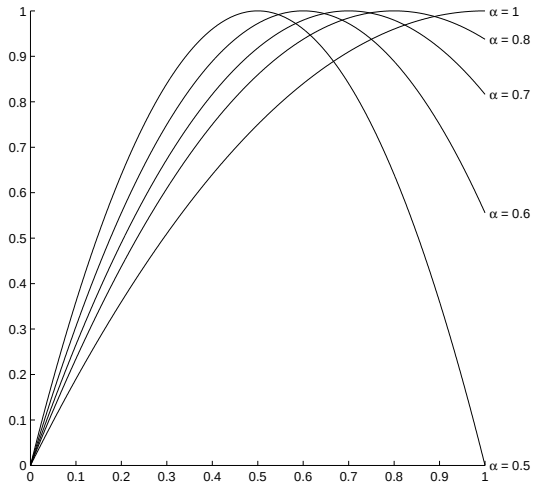


Figure 1: Pugnancity depending on relative state $\frac{x_j}{x_j^{max}}$ for different α

Hence after actions the states of loading agents will be $x_i(t) = x_i^*(t) + l_i(t)$ and the state of resources as follows:

$$x_j(t) = x_j^*(t) - \sum_{i \in L_j(t)} l_i(t) \tag{3}$$

where $L_j(t)$ are agents loading from the resource at time t.

The problem occurs if more than one agents are at the same resource. In this case, agents are in a queue and load one after other. The point is how to sort this queue, who will load first. Let $p_i(t)$ is actual pugnancity of agent i. The ordering can have e.g. such form:

- The strongest agents load first - selfish apporach.
- The order is random - random approach.
- The weakest agents load first - altruistic approach.

A sorting algorithm called groggy sort has been developed which can parametrize this. The system parameter $\gamma \in [-1, 1]$ passes from the altruistic approach through the random to the selfish one smoothly.

The pugnancity depends on relative filling of the agent. If an agent is almost empty, it means it is hungry and is weak. Nevertheless, the agent may lose the power if it is nearly completely full. The relation is given by following formula:

$$p_j = -\alpha^{-2} \left(\frac{x_j}{x_j^{max}} - \alpha \right)^2 + 1 \tag{4}$$

The pugnancity express how the agent's vigour depends on the relative filling and is always $\in [0, 1]$. Parameter $\alpha \in [0.5, 1]$ sets the relative filling by maximal pugnancity, as shown in Figure 1.

The last remaining aspect of the system is the mechamism how the agents decide. The selection of an action is random. Each action have a score that is updated. Agents

have memory about all visited resources and amounts of material that they loaded there. The score of movement to a resource is calculated as follows:

$$s_{i,j}(t) = c\frac{v_{i,j}^{successful}(t)}{v_{i,j}^{total}(t)}\frac{l_{(t_{last})}}{l_i^{max}}\frac{1}{n}\tanh(\tau(t-t_{last})) \tag{5}$$

where c is a constant > 0 and represents the traveling preference. The first fraction is the ratio of visits when the load was not zero. The second fraction express the ratio of successfulness of last visit. Third fraction is important to ensure the comparability of the action travel and action load for all possible amount of resources n. The latter term express the time influence: longer absence, bigger curisousity to visit the resource. The parameter τ can be denoted as nostalgia. The function tanh is applied in order to keep also this term in [01].

2.2 Objectives

The system is running with given parameters finite period. During this time natural changes happen and the agents act. There are some thinks that can be considered as objectives:

Total production - the sum of loaded material during the simulation is maximal.

Humanism - the amount of not empty agents after the simulation is maximal. In other words, this approach maximizes the amount of living agents in the system.

Ecology - the amount of not empty resources after the simulation is maximal. In other words, this approach attempts to keep resources able to be renewed.

Egoism - the sum of loaded material during a period is maximal. If the system has m agents, there are m egoist criteria. This model would suppose heterogenous agents. As a simplification, this objective has been skipped.

It is obvious some criteria are conflicting. E.g. if the agents would not load the material, they will die soon and no resource will be used. Nevertheless, in a long term horizon, it can be supposed that these conflicting criteria are not conflicting in fact. If there is no production, no agent can survive. If there are no resources able to be renewed, the agents can not survive as well.

The conflict of several criteria is solvable by multicriterial methods. One of them is presented bellow. For a proper formulation of a multicriterial problem it is necessary to state also the input space, i.e. parameters and values from which can be the parameters of the system selected.

2.3 Parameters

The system has following parameters The fourth column states which parameters are fixed for performed simulation. In fact, the parameters that are considered not to be

Parameter	Stands for	Range	Default
n	Number of resources	$\mathbb{N}$	20
m	Number of agents	$\mathbb{N}$	20
a	$a = -a_i$ for all agents i	$\mathbb{R}+$	1
a_{max}	maximal load limit for each agent	$\mathbb{R}+$	1
g	$g = g_j$ for all resources j	$1 + \mathbb{R}+$	1.3
α	pugnancity parameter	$[0.5, 1]$	-
γ	groggy sort parameter	$[-1, 1]$	-
c	traveling preference	$\mathbb{R}+$	-
τ	time parameter	$\mathbb{R}+$	-
$x_{ag}^{max}, x_{res}^{max}$	maximal states for agents and for resources	$\mathbb{R}+$	10, 20
$x_{ag}^{init}, x_{res}^{init}$	initial states for agent and resources	$\mathbb{R}+$	4, 9
$x_{ag}^{min}, x_{res}^{min}$	minimal states for agent and resources	$\mathbb{R}+$	0, 0

Table 1: Considered system parameters

subject of decision making have been fixed. Parameters g, α influence the ordering of agents if more than one come to the same resource. Parameters c, τ are part of the decision making procedure of each agent.

3 Multicriterial Genetic Optimization

The system described was described completely. The question is which values shall the system parameters have with respect to proposed objectives.

Multicriterial decision making has been a subject of research for a long time. Nevertheless, usual methods consider convex decision space and special (linear, quadratic, convex, etc.) objective functions.

However, if the objective function corresponds to a result of a simulation, the assumptions are not possible. Therefore, another approach is necessary. Following text presents a modification of algorithm described [1].

First, let the multicriterial optimization problem be formulated properly. The input space is an interval $S \subset \mathbb{R}^4$ where particular attributes corresponds to parameters α, g, c, τ. The objective function has 3 components: total production, number of non empty agents, and number non empty resources. The objectives will be denoted f_1, f_2, and f_3. Because the system is stochastic, these values are obtained as average from s simulation. The parameter $s = 20$ was used.

The basic principle of multicriterial analysis consist in elimination of dominated variants. A variant is dominated if there is another one that is at least same in all criteria and in at least one better. The aim of presented method is to find nondominated variants. The dominanace will be denoted

$$\mathbb{x}^{(i)} \succ \mathbb{x}^{(j)} \tag{6}$$

and a non dominated variant within a set

$$\mathbb{x}^{(i)} \succ S \tag{7}$$

finally a subset dominanting particular solution will be denoted S_i. Of course, S_i will be empty for all non dominated variants. The method is a modification of an usual genetic algorithm with selection, crossover, and mutation.

3.1 Selection

This operatdion respects the levels of nondominance. I.e. non dominated individuals have the probability highest. They have rank 1. After omitting them, other individuals are non dominated. They have rank 2 etc. Formaly

$$r_i = 1 \quad \text{if} \quad \mathbb{x}^{(i)} \succ S \tag{8}$$

$$r_i = 1 + \max_{j \in S_i} r_j \tag{9}$$

The ranking is afterwards used for caclulation of two distribution functions:

$$\rho_i = k_1(\exp r_j + 1) \tag{10}$$

$$\theta_i = \frac{k_2}{r_j} \tag{11}$$

where k_1, k_2 are constants do $\sum \rho_i = \sum \theta_i = 1$. In each time step given number of crossovers and mutations are performed. ρ is used for sampling parents of these operations, while θ is used for locating the place.

3.2 Crossover

The crossover is perfomed per components, i.e. the component of the first child are selected with 0.5 probablility form the first parent and with 0.5 from the other one. The not selected component is put into the second child.

3.3 Mutation

The mutation adds random noise form normal distribution with the mean $\mu = 0$ and standard deviation that may differ for each attribute is defined as follws:

$$\sigma_j = \frac{1}{\log(\log(t))} \tag{12}$$

$$\sigma_j = \frac{1}{s_j \log(\log(t))} \tag{13}$$

where s_j is standard deviation of the $j-$th component of the population.

4 Results

Practical part of the work consist of two parts. First, the model was implemented in Java. Afterwards an multicriterial algorithm was coded in Matlab. This algorithm uses the Java classes for the objective function calculation, of course.

Number of iterations	10000
Size of population	100
Number of crossovers in an iteration	1000
Number of mutations in an iteration	1000
Simulations required for fitness calculation	20

Table 2: Optimization parameters

	α	γ	τ	c
Agents alive	−0.12366	−0.036367	−0.19365	−0.36289
Non empty resources	−0.014428	−0.00093429	−0.013190	−0.0072702
Total production	−0.12346	−0.033927	−0.18654	−0.33625

Table 3: Correlation matrix for all generated results

Simulation and optimization parameters are to be distinguished. Some simulation parameters have been fixed, some of them were variable. The objective was to find such values of these variable parameters, leading to non dominated solutions with respect to above mentioned objectives. Table 1 shows in column Default which parameters are fixed and their values. Only such parameters were selected to be subject of optimization that influece the behavior of agents. Table 2 shows the optimization parameters.

The optimization algorithm discovered 5 non dominated solutions. For basic orientation in dependencies between parameters of non dominated solutions and corresponding values of objective functions, correlation analysis was performed. Table 3 provides part of correlation matrix for the entire population, Table 4 for non dominated solutions. The first one represents all solutions, but we are interested only in the non dominated ones. GFurthermore, the values of Table 3 are not s

The correlations for non dominated solutions are more signifficant. At level 0.1 for t-test, four correlations are signifficant:

- Agents alive - τ
- Total production -τ
- Agents alive - c
- Total production -c

Both parameters τ and c influence the agent's will to travel. Agents obtain better results in exploiting if they load instead of travel.

Other correlations between variable parameters and objectives are not so significant. Parameters α and γ have the opposite effect. It seems that the exploaition is more effective if more pugnant agents are prefered and the pugnancy grows with agent's material.

5 Discusion

Present work have introduced simple multi-agent model of an environment with renewable resources. Parameters influencing the problem were mentioned. Afterwards an multicri-

	α	γ	τ	c
Agents alive	0.10876	0.22804	−0.75014	−0.82158
Non empty resources	−0.29206	−0.17625	0.47344	0.47121
Total production	0.10872	0.22842	−0.75076	−0.82220

Table 4: Correlation matrix for all non dominated solutions

terial optimization algorithm was introduced and applied. Results were presented and discussed.

Main benefits are these: design and implementation of a multi-agent system, general framework for conflict modelling and resolution via pugnancity and groggy sort, basic modelling of agents' memory and decision making, and finally modification and implementation of an multicriterial optimization evolution algorithm.

The work opens also some challenges for further research. The model could deal also with placement of the resources in a plane or in a graph so the movement of agents is not so easy, but more realistic. The behavior of resources could depend on other resources (phreatic water) or external conditions (wheather). The model of agent could be improoved as well, especially with respect to communuication, knowledge sharing, reasoning, coolation formation etc.

Regarding the optimization of the system, other criteria and parameters can be involved, more testing can be performed. The optimization method can employ other selection, mutation or crossover. The relationships between variable simulation parameters and optimization objectives could be examined by advanced methods than applied correlation analysis.

There is lot of open work and I intent to deal with it within next phases of my PhD course.

References

[1] P. Davidsen H. Qudrat-Ullah, M. Spector, editor. *Complex Decision Making: Theory and Practice*. Springer, February 2007.

[2] David M. Nicol, Corrado Priami, Hanne Riis Nielson, and Adelinde M. Uhrmacher, editors. *Simulation and Verification of Dynamic Systems, 17.04. - 22.04.2006*, volume 06161 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2006.

[3] V. Šmídl. *Software analysis of Bayesian distributed dynamic decision making*. PhD thesis, University of West Bohemia, Faculty of Applied Sciences, Pilsen, Czech Republic, Plzeň, 2005.

[4] František Zbořil. Simulation languages for agent systems. In *Proceedings of the Fifth International Scientific Conference ECI 2002*, pages 73–77. Faculty of Electrical Engineering and Informatics, University of Technology Košice, 2002.

Qualitative Study of the Gray-Scott Model

Jan Mach

2nd year of PGS, email: jan.mach@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Michal Beneš, Department of mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. This contribution deals with numerical solution of the Gray-Scott model. We introduce two numerical schemes for the 2D GS model based on the method of lines. To perform spatial discretization we use FDM in first case and FEM in the second case. Resulting systems of ODEs are solved using the Runge-Kutta-Merson method. We present our numerical simulations.

Abstrakt. V tomto příspěvku se věnujeme numerickému řešení Grayova-Scottova modelu. Představujeme dvě numerická schémata pro 2D GS model založená na metodě přímek. K prostorové diskretizaci používáme v prvním případě FDM, ve druhém FEM. Vzniklé systémy ODEs řešíme metodou Runge-Kutta-Merson. Uvádíme výsledky numerických simulací.

1 Introduction

Reaction-diffusion systems are a class of systems of partial differential equations of parabolic type. It includes mathematical models describing various phenomena in the field of physics, biology and chemistry. They describe how the concentration of one or more substances distributed in space changes under the influence of two processes: local chemical reactions in which the substances are converted into each other, and diffusion which causes the substances to spread out in space. Reaction and diffusion of chemical species can produce a variety of patterns.

Gray-Scott model is one of these models. It was first introduced in 1984 in an article by P. Gray and S. K. Scott [1] as a mathematical model of autocatalytic chemical reaction

$$\begin{aligned} U + 2V &\longrightarrow 3V \\ V &\longrightarrow P, \end{aligned} \tag{1}$$

where U, V are input reactants and P is inert product. Gray-Scott model can be written as the following system two PDEs of parabolic type (see [3, 4])

$$\begin{aligned} \frac{\partial u}{\partial t} &= a\nabla^2 u - uv^2 + F(1-u), \\ \frac{\partial v}{\partial t} &= b\nabla^2 v + uv^2 - (F+k)v. \end{aligned} \tag{2}$$

Here u, v are unknown functions representing concentrations of chemical substances U, V. Parameter F denotes the rate at which the chemical substance U is being added

during the chemical reaction, $F + k$ is the rate of $V \to P$ transformation and a, b are constants characterizing the environment where the chemical reaction takes place.

We solve (2) on a finite domain Ω, which is a square or line depending on whether we are solving the system in 2D or 1D. We use zero Neumann boundary conditions. Our choice of initial data is such that $v(x,0)) = v_{ini}$, $u(x,0)) = 1 - v_{ini}$. We usually take $F > 0$, $k > 0$. For $a = 0$, $b = 0$ the system (2) is a model of the reaction (1) in continuously fed well stirred tank reactor (CSTR), the CSTR model. If $a > 0$, $b > 0$ then the system (2) models the reaction in continuously fed unstirred reactor (CFUR), the CFUR model (see [7]). Dynamics of the CSTR model is rich and covers standing pulses, traveling pulses, traveling fronts, self-replicating patterns, spatio-temporal chaos and others (see [3]). Most of these pattern have been observed also in the CFUR model (see i.e. [9]).

For other dimensionless forms of the Gray-Scott model and their application see i.e. [4, 5, 6].

2 Numerical schemes

We use two numerical schemes to solve initial-boundary-value problem for the Gray-Scott model (2). Both of them are based on the method of lines. For spatial discretization we used finite difference method (FDM) in the first case and finite elements method (FEM) in the second case. We use structured numerical grids (see Fig. 1). To solve resulting systems of ordinary differential equations Runge-Kutta-Merson method is used.

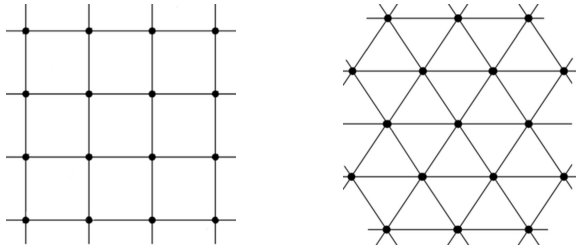


Figure 1: Numerical grids we used for our numerical simulations.

2.1 FDM based numerical scheme

Let h be mesh size such that $h = \frac{L}{N-1}$ for some $N \in \mathbb{N}^+$. We define numerical grid as a set

$$\begin{aligned} \omega_h &= \{(ih, jh) \mid i = 1, \ldots, N-2, j = 1, \ldots, N-2\}, \\ \overline{\omega}_h &= \{(ih, jh) \mid i = 0, \ldots, N-1, j = 0, \ldots, N-1\}. \end{aligned}$$

For function $u : \mathbb{R}^2 \to \mathbb{R}$ we define a projection on $\overline{\omega}_h$ as $u_{ij} = u\,(ih, jh)$. We introduce finite differences

$$u_{x_1,ij} = \frac{u_{i+1,j} - u_{i,j}}{h}, u_{\overline{x}_1,ij} = \frac{u_{i,j} - u_{i-1,j}}{h}$$
$$u_{x_2,ij} = \frac{u_{i,j+1} - u_{i,j}}{h}, u_{\overline{x}_2,ij} = \frac{u_{i,j} - u_{i,j-1}}{h},$$

and define approximation Δ_h of the Laplace operator Δ as follows

$$\Delta_h u_{ij} = u_{\overline{x}_1 x_1,ij} + u_{\overline{x}_2 x_2,ij}.$$

Then semi-discrete scheme has the following form

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u_{ij}(t) &= \frac{a}{h^2}\Delta_h u_{ij} + F(1 - u_{ij}) - u_{ij}v_{ij}^2, \\ \frac{\mathrm{d}}{\mathrm{d}t}v_{ij}(t) &= \frac{b}{h^2}\Delta_h v_{ij} - (F + k)v_{ij} + u_{ij}v_{ij}^2, \end{aligned} \tag{3}$$

plus corresponding initial and boundary conditions.

2.2 FEM based numerical scheme

To induce the semi-discrete scheme we begin with variation formulation of the initial-boundary-value problem for the Gray-Scott model (2). Let

$$\begin{aligned} \varphi_1(x), \varphi_2(x) &\in C_0^\infty(\Omega), \\ \psi_1(t), \psi_2(t) &\in C_0^\infty(0, T) \end{aligned}$$

are test functions and

$$\begin{aligned} f_1(u, v) &= F(1 - u) - uv^2, \\ f_2(u, v) &= -(F + k)v + uv^2 \end{aligned}$$

denote right-hand sides of differential equations (2). Using standard approach (see [8]) we induce weak formulation of the problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(u, \varphi_1) + a(\nabla u, \nabla\varphi_1) &= (f_1, \varphi_1), \\ \frac{\mathrm{d}}{\mathrm{d}t}(v, \varphi_2) + b(\nabla v, \nabla\varphi_2) &= (f_2, \varphi_2), \\ u(\cdot, 0) &= u_{ini}, \\ v(\cdot, 0) &= v_{ini}, \end{aligned} \tag{4}$$

with solution u, v from the Sobolev space $W_2^{(1)}(\Omega)$. We are looking for Galerkin approximation

$$\begin{aligned} u_h(t) &= \sum_{i=1}^{N} \alpha_i(t)\Phi_i, \\ v_h(t) &= \sum_{i=1}^{N} \beta_i(t)\Phi_i \end{aligned}$$

of this weak solution in the finite dimensional space $S_h \subset W_2^{(1)}(\Omega)$, where $\Phi_1, \ldots, \Phi_N$ are its basis functions. Functions α_i, β_i are real functions which we get using common technique as solutions of initial value problems. Choosing basis functions Φ_i in the form of pyramidal functions

$$\Phi_i(P_j) = \delta_{ij} \quad \text{for all grid nodes} \quad P_j,$$

and using mass-lumping we can rewrite the problem for finding functions α_i, β_i in the following form

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u_{ij}(t) &= \frac{2a}{3h^2}[u_{i+1,j} + u_{i+1,j+1} + u_{i,j-1} + u_{i,j+1} + u_{i-1,j} + \\ &\quad + u_{i-1,j+1} - 6u_{ij}] + F(1 - u_{ij}) - u_{ij}v_{ij}^2 \\ \frac{\mathrm{d}}{\mathrm{d}t}v_{ij}(t) &= \frac{2b}{3h^2}[v_{i+1,j} + v_{i+1,j+1} + v_{i,j-1} + v_{i,j+1} + v_{i-1,j} + \\ &\quad + v_{i-1,j+1} - 6v_{ij}] - (F + k)v_{ij} + u_{ij}v_{ij}^2 \end{aligned} \tag{5}$$

plus corresponding initial and boundary conditions. For details on induction of presented semi-discrete schemes we refer reader to [9].

3 Numerical experiments

3.1 EOC measurements

To determine the order of convergence of our numerical algorithm based on the FDM based semi-discrete scheme (3) we use experimental order of convergence (EOC). For our measurements we used formula

$$\frac{\| v - v_{h2} \|}{\| v - v_{h1} \|} = \left(\frac{h2}{h1}\right)^{\alpha}, \tag{6}$$

where v is numerical solution computed on the grid of size 2000×2000 and substitutes the analytical solution, v_{h2}, v_{h1} are numerical solutions computed on courser grids with mesh sizes $h2$, $h1$ and α is the EOC coefficient. We present some of our measurements for different Gray-Scott model parameter values and initial conditions (see Table 1, Table 2, Table 3). According to the presented results the EOC coefficient depends notably on initial concentration data and model parameter values. Our results vary between the values of 1 and 2. More research into this problem is needed including EOC measurement for the FEM based numerical algorithm.

3.2 Diversity of solutions

In this section we present some of our numerical results. In the figures Figure 3 and Figure 2 we can see spatial distribution over the domain Ω of chemical substance V concentration for given Gray-Scott model parameter values and time. These results demonstrate the diversity of GS model solutions. We can see that patterns vary between geometrically simple ones and those which are more complex. In the Figure 2 we can see growing-line like patterns which we were able to observe for parameter values $a = 2{\cdot}10^{-5}$, $b = 1{\cdot}10^{-5}$, $F = 0.0737$, $k = 0.061882$, $L = 0.5$ and different initial conditions.

$N_x \times N_y$	h	EOC L_2	EOC L_∞
100x100	0.0050505	-	-
150x150	0.0033557	1.6479179	1.6364127
200x200	0.0025125	1.8042298	1.5663398
250x250	0.0020080	1.9112146	1.7531840
300x300	0.0016722	1.9725610	1.8660718
350x350	0.0014326	2.0089377	1.8995297
400x400	0.0012531	2.0336490	1.9882238

Table 1: Table of EOC coefficients.

$N_x \times N_y$	h	EOC L_2	EOC L_∞
100x100	0.0101010	-	-
150x150	0.0067114	0.8225371	0.5550153
200x200	0.0050251	0.9222231	0.7584173
250x250	0.0040160	0.9995422	0.9052681
300x300	0.0033444	1.0667171	1.0124643
350x350	0.0028653	1.1237827	1.0727512
400x400	0.0025062	1.1754085	1.1689477

Table 2: Table of EOC coefficients.

$N_x \times N_y$	h	EOC L_2	EOC L_∞
100x100	0.0050505	-	-
150x150	0.0033557	2.0466270	1.0203486
200x200	0.0025125	2.0460521	0.9659226
250x250	0.0020080	2.0512043	1.1006299
300x300	0.0016722	1.9143909	0.9491632
350x350	0.0014326	1.5423185	1.0946135
400x400	0.0012531	1.5552072	0.9893100

Table 3: Table of EOC coefficients.

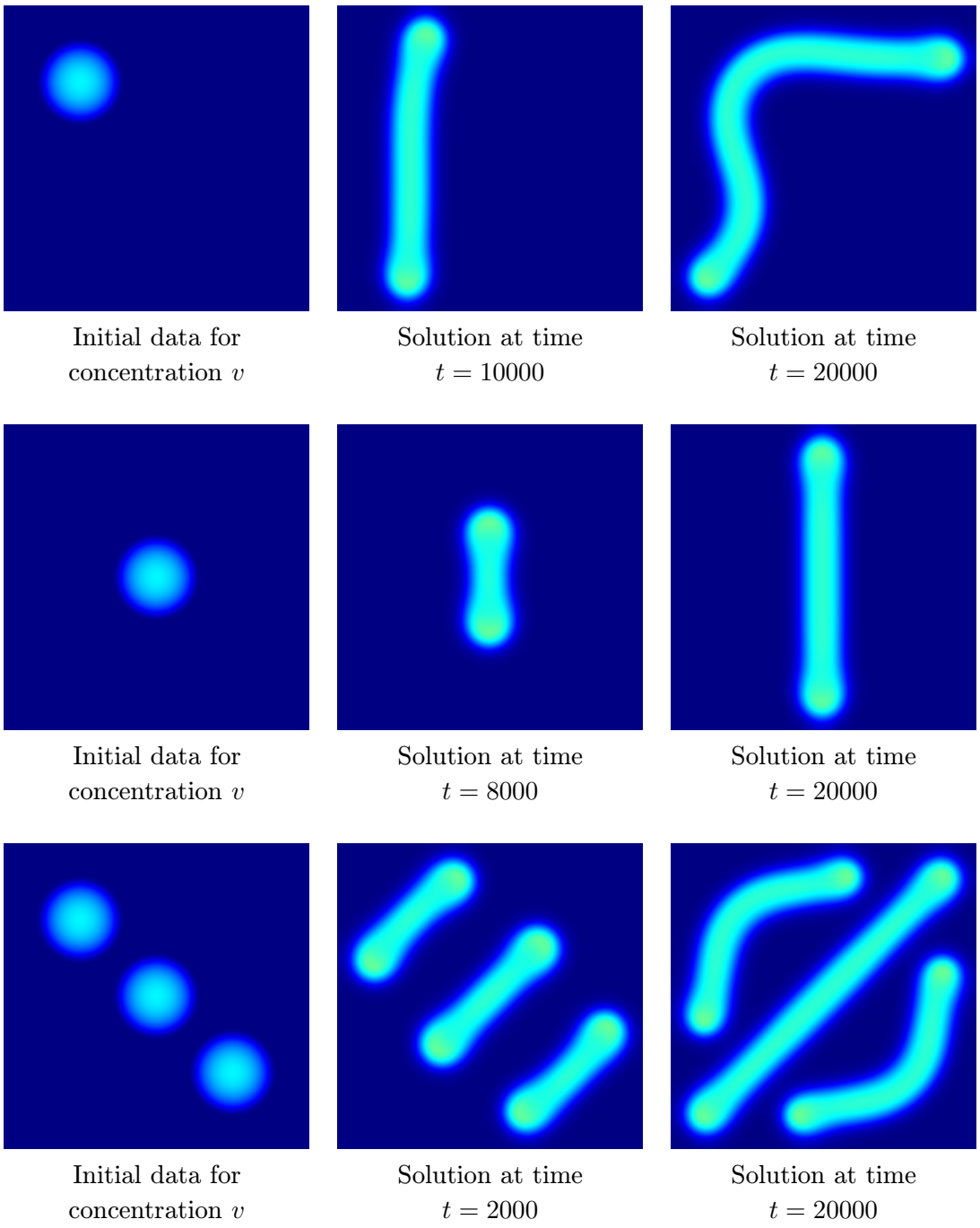
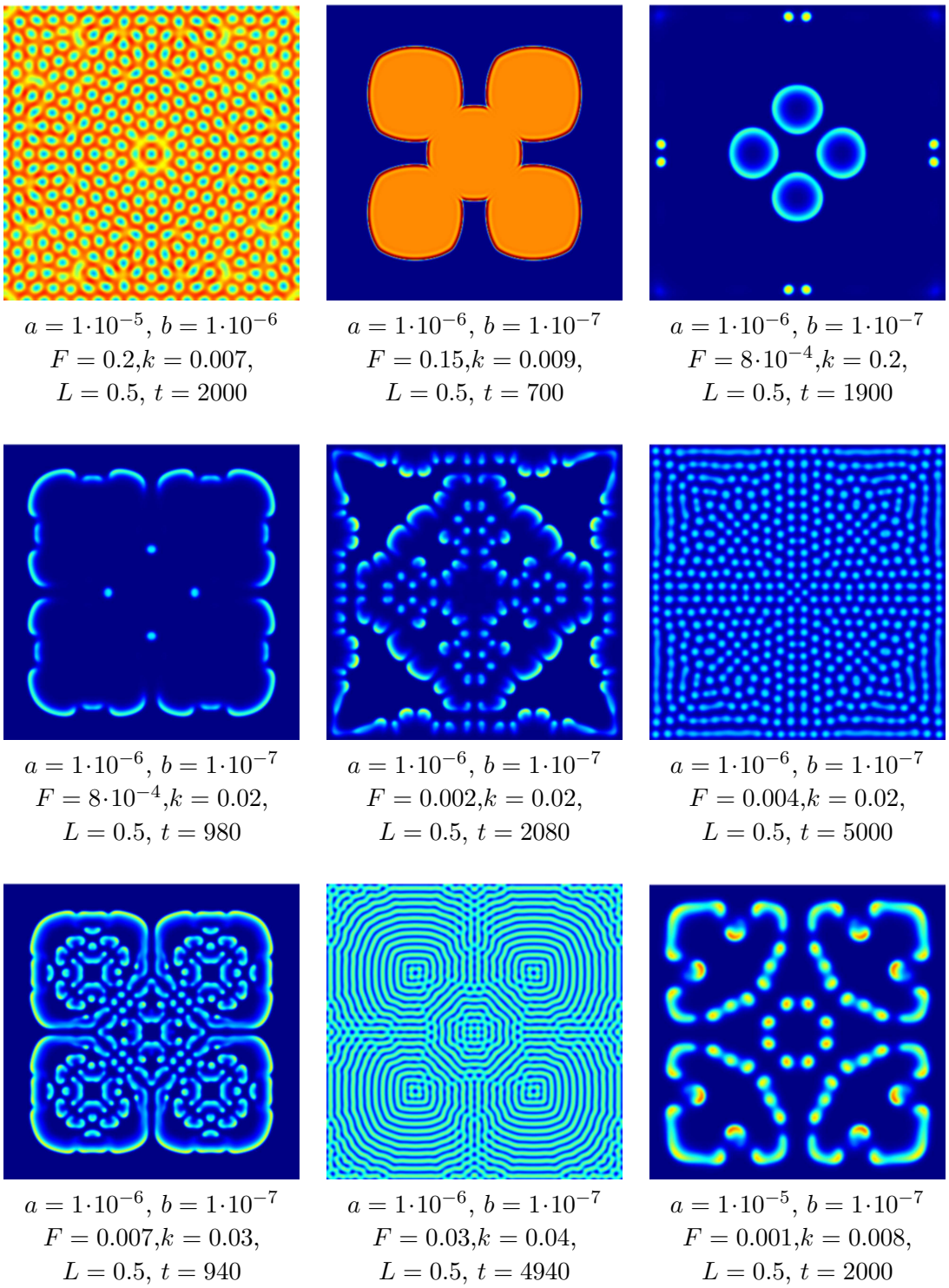


Figure 2: Growing line-like patterns. Parameter values: $a = 2 \cdot 10^{-5}$, $b = 1 \cdot 10^{-5}$, $F = 0.0737$, $k = 0.061882$, $L = 0.5$. Grid size: 1000×1000. Numerical method: FDM. Time evolution of concentration v is shown for different initial data.



$a = 1{\cdot}10^{-5}$, $b = 1{\cdot}10^{-6}$
$F = 0.2$,$k = 0.007$,
$L = 0.5$, $t = 2000$

$a = 1{\cdot}10^{-6}$, $b = 1{\cdot}10^{-7}$
$F = 0.15$,$k = 0.009$,
$L = 0.5$, $t = 700$

$a = 1{\cdot}10^{-6}$, $b = 1{\cdot}10^{-7}$
$F = 8{\cdot}10^{-4}$,$k = 0.2$,
$L = 0.5$, $t = 1900$

$a = 1{\cdot}10^{-6}$, $b = 1{\cdot}10^{-7}$
$F = 8{\cdot}10^{-4}$,$k = 0.02$,
$L = 0.5$, $t = 980$

$a = 1{\cdot}10^{-6}$, $b = 1{\cdot}10^{-7}$
$F = 0.002$,$k = 0.02$,
$L = 0.5$, $t = 2080$

$a = 1{\cdot}10^{-6}$, $b = 1{\cdot}10^{-7}$
$F = 0.004$,$k = 0.02$,
$L = 0.5$, $t = 5000$

$a = 1{\cdot}10^{-6}$, $b = 1{\cdot}10^{-7}$
$F = 0.007$,$k = 0.03$,
$L = 0.5$, $t = 940$

$a = 1{\cdot}10^{-6}$, $b = 1{\cdot}10^{-7}$
$F = 0.03$,$k = 0.04$,
$L = 0.5$, $t = 4940$

$a = 1{\cdot}10^{-5}$, $b = 1{\cdot}10^{-7}$
$F = 0.001$,$k = 0.008$,
$L = 0.5$, $t = 2000$

Figure 3: Results demonstrating diversity of solutions of the GS model computed using FDM based numerical scheme (3) and grid size 400×400 for different parameter value combinations. Spatial distribution of concentration v is presented.

Acknowledgement

The work has been performed under the Project HPC-EUROPA (RII3-CT-2003-506079), with the support of the European Community - Research Infrastructure Action under the FP6 "Structuring the European Research Area" Programme.

Partial support of the project of "Necas Center for Mathematical modeling", No. LC06052 and of the project "Applied Mathematics in Physics and Technical Sciences", No. MSN6840770010 of the Ministry of Education, Youth and Sports of the Czech Republic is acknowledged.

References

[1] P. Gray and S. K. Scott. *Autocatalytic reactions in the isothermal, continuous stirred tank reactor: oscillations and instabilities in the system* $A + 2B \to 3B$, $B \to C$. Chem. Eng. Sci. 39:1087-1097 (1984).

[2] P. Gray and S. K. Scott. *Chemical Oscillations and Instabilities: Non-linear Chemical Kinetics*. Oxford University Press, Oxford, 1990.

[3] Y. Nishiura, D. Ueyama. *Spation-Temporal Chaos for the Gray-Scott model*. Physica D 150 (2001), 137-162.

[4] J. Wei. *Pattern formation in two-dimensional Gray-Scott model: existence of single-spot solutions and their stability*. Physica D 148 (2001), 20-48.

[5] J. Wei and M. Winter. *Asymmetric spotty patterns for the Gray-Scott model in* $\mathbb{R}^2$. Stud. Appl. Math. 110 (2003), no. 1, 63-102.

[6] T. Kolokolnikov, M. J. Ward, J. Wei. *The existence and stability of spike equlibria in the one-dimensional Gray-Scott model on a finite domain*. Appl. Math. Letters 18 (2005), 951-956.

[7] J. S. McGough, K. Riley. *Pattern formation in the Gray-Scott model*. Nonlinear Analysis: Real World Application 5 (2004), 105-121.

[8] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer-Verlag Berlin Heidelberg, 1997.

[9] J. Kodovský. *Dynamics of reaction-diffusion equations, mathematical and numerical analysis*. Master thesis, FNSPE CTU, Prague, 2006.

Numerical Simulation of Dislocation Dynamics*

Petr Pauš

2nd year of PGS, email: pauspetr@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Michal Beneš, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. This paper deals with the numerical simulation of dislocation dynamics. Dislocations are described by means of the evolution of a family of closed and open smooth curves $\Gamma(t) : S^1 \to \mathbb{R}^2$, $t \geqq 0$. The curves are driven by the normal velocity v which is the function of curvature κ and the position vector $x \in \Gamma(t)$. In this case the equation is defined this way: $v = -\kappa + F$. The equation is solved using direct approach by two numerical schemes, ie. semi-implicit and semi-discrete. Results of the dislocation dynamics simulation are presented.

Abstrakt. Tento článek se zabývá numerickou simulací dislokační dynamiky. Dislokace jsou popsány pomocí časového vývoje množiny uzavřených a otevřených hladkých křivek $\Gamma(t) : S^1 \to \mathbb{R}^2$, $t \geqq 0$. Vývoj křivek je ovlivňován normálovou rychlostí v, jenž je funkcí křivosti κ a polohového vektoru $x \in \Gamma(t)$. V tomto případě má rovnice tvar $v = -\kappa + F$. Rovnice je řešena přímou metodou pomocí dvou různých numerických schémat, semi-implicitním a semi-diskrétním. Výsledky simulace dislokační dynamiky jsou také uvedeny.

1 Introduction

In the field of material science, the dislocations are defined as an irregularity or error in crystal structure of the material. The presence of dislocations strongly influences many of the material properties, that is why it is important to develop suitable physical and mathematical model. The physical model already exists but there still is a lot of to do concerning mathematical model. From the mathematical point of view, the dislocations are defined as smooth closed or open plain curves which evolve in time. The example of dislocation in the material is shown in Figure 1.

2 Mathematical model

The evolving curves can be mathematically described in several ways. One possibility is to use the *level-set method* [1, 2, 3], where the curve is defined by the zero level of some surface function. One can also use the *phase-field method* [4]. Finally, it is possible to use the *direct (parametric) method* [5, 6] where the curve is parametrized in usual way. This article discusses this direct approach.

*This work is supported by grant no. MSM 6840770010, project no. LC06052 of Nečas center for mathematical modeling.

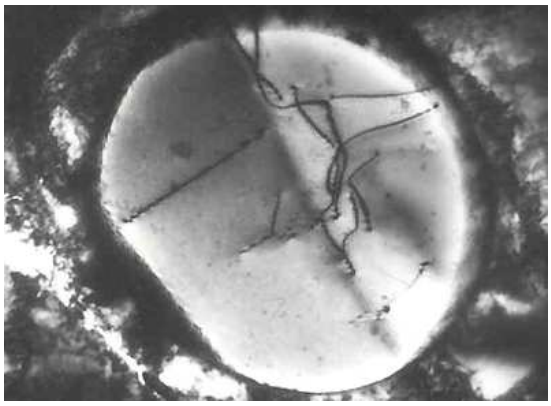


Figure 1: Dislocation in steel. http://en.wikipedia.org/wiki/Dislocations

3 Parametric description

When using the parametric approach, the dislocation curve $\Gamma(t)$ is described by a smooth time-dependent vector function

$$X : S \times I \to \mathbb{R}^2,$$

where $S = (0, 1)$ is a fixed interval for curve parametrization and $I = \langle 0, T\rangle$ is the time interval. Dislocation curve $\Gamma(t)$ is then given as

$$\Gamma(t) = \{X(u, t) = (X^1(u, t), X^2(u, t)), u \in S\}.$$

The family of curves satisfies the equation for time evolution

$$v = -\kappa + F \tag{1}$$

where v is the normal velocity of the curve evolution. The normal velocity v is the function of the curvature κ and the position vector $\mathbf{x}$. κ is the mean curvature and F is the forcing term.

The evolution law (1) is transformed into the parametric form. The unit tangential vector $\vec{T}$ is defined as $\vec{T} = \partial_u X/|\partial_u X|$. The unit normal vector $\vec{N}$ is perpendicular to the tangential vector and $\vec{N} \cdot \vec{T} = 0$ holds. The curvature κ is defined as

$$-\kappa = \frac{\partial_u X^\perp}{|\partial_u X|} \cdot \frac{\partial^2_{uu} X}{|\partial_u X|^2} = \vec{N} \cdot \frac{\partial^2_{uu} X}{|\partial_u X|^2},$$

where $X^\perp$ is a vector perpendicular to X. The normal velocity v is defined as a time derivative of X projected into the normal direction,

$$v = \partial_t X \cdot \frac{\partial_u X^\perp}{|\partial_u X|}.$$

The equation (1) can now be written as

$$\partial_t X \cdot \frac{\partial_u X^\perp}{|\partial_u X|} = \frac{\partial^2_{uu} X}{|\partial_u X|^2} \cdot \frac{\partial_u X^\perp}{|\partial_u X|} + F$$

which holds provided

$$\partial_t X = \frac{\partial^2_{uu} X}{|\partial_u X|^2} + F(u,t)\frac{\partial_u X^\perp}{|\partial_u X|}. \tag{2}$$

The term $\partial^2_{uu} X/|\partial_u X|^2$ in (2) contains some tangential force which makes curve points to move along curve. To neglect this tangential force, some term α in the tangential direction must be subtracted, so the equation changes to

$$\partial_t X = \frac{\partial_{uu} X}{|\partial_u X|^2} - \alpha\frac{\partial_u X}{|\partial_u X|} + F(u,t)\frac{\partial_u X^\perp}{|\partial_u X|}. \tag{3}$$

One can derive that the tangential force contained in the equation has the form

$$\alpha = \frac{\partial_{uu} X \cdot \partial_u X}{|\partial_u X|}. \tag{4}$$

We obtain the equation where there is no tangential force at all. The equation has the following form:

$$\partial_t X = \frac{\partial_{uu} X}{|\partial_u X|^2} - \frac{\partial_{uu} X \cdot \partial_u X}{|\partial_u X|^2}\partial_u x + F(u,t)\frac{\partial_u X^\perp}{|\partial_u X|}. \tag{5}$$

This equation is not suitable for numerical simulations because points cannot move along curve and often create areas with high density of points and areas where points are very sparse causing very slow computation. The equation (2) is better for numerical simulations but still for long time simulations similar grouping of points usually happens. One of the solutions is to use some algorithm for tangetial redistribution of points.

For long time computations with time and space variable external force $F(u,t)$, the algorithm for curvature adjusted tangential velocity is used. This algorithm moves points along the curve according to the curvature, i.e., areas with higher curvature contain more points than areas with lower curvature. This improves numerical stability and also precision of computation. Unlike the case with no tangential force (5), the term α is not given by a simple formula but it is based on relative local length between points. Details are described in [12].

4 Numerical scheme

For numerical approximation we consider a regularized form of (3) which reads as

$$\partial_t X = \frac{\partial^2_{uu} X}{Q(\partial_u X)^2} - \alpha\frac{\partial_u X}{Q(\partial_u X)} + F(u,t)\frac{\partial_u X^\perp}{Q(\partial_u X)}, \tag{6}$$

where $Q(x_1, x_2) = \sqrt{x_1^2 + x_2^2 + \varepsilon^2}$. Two numerical schemes are used for the numerical solution of the differential equation (3), semi-implicit and semi-discrete. With two numerical schemes it is possible to compare the solution and error of computation.

In the semi-discrete scheme spatial derivatives are approximated by fourth-order central differences. The first derivative is approximated as

$$\partial_u X \approx \left[\frac{X^1_{j-2} - 8X^1_{j-1} + 8X^1_{j+1} - X^1_{j+2}}{12h}, \frac{X^2_{j-2} - 8X^2_{j-1} + 8X^2_{j+1} - X^2_{j+2}}{12h}\right],$$

and the second one as

$$\partial^2_{uu}X \approx \big[\frac{-X^1_{j-2}+16X^1_{j-1}-30X^1_j+16X^1_{j+1}-X^1_{j+2}}{12h^2}, \frac{-X^2_{j-2}+16X^2_{j-1}-30X^2_j+16X^2_{j+1}-X^2_{j+2}}{12h^2}\big],$$

where X^i_j denotes an approximation of $X^i(jh,\cdot), i \in \{1,2\}$, $h = 1/m$. Here m is a number of points on the curve. Differences are denoted as X_u for the first derivative and X_{uu} for the second derivative.

The equation (6) in semi-discrete scheme has the following form:

$$\frac{dX_j}{dt} = \frac{X_{uu,j}}{Q^2(X_{u,j})} - \alpha_j\frac{X_{u,j}}{Q(X_{u,j})} + F(u,t)\frac{X^\perp_{u,j}}{Q(X_{u,j})}, \quad j = 1,\cdots,m-1, t \in (0,T), \quad (7)$$

where again $Q(x_1,x_2) = \sqrt{x_1^2+x_2^2+\varepsilon^2}$, $X^\perp_{u,j}$ is a vector perpendicular to $X_{u,j}$, and α_j is redistribution coeficient. The term ε serves as a regularization to avoid singularities when the curvature tends to infinity. This scheme is solved by the fourth order Runge-Kutta method.

Second approach uses the semi-implicit scheme. In this case lower order differences are used. The first derivative is discretized by backward difference as follows

$$\partial_u X \approx \left[\frac{X^1_j - X^1_{j-1}}{h}, \frac{X^2_j - X^2_{j-1}}{h}\right],$$

and the second derivative as

$$\partial^2_{uu}X \approx \left[\frac{X^1_{j+1} - 2X^1_j + X^1_{j-1}}{h^2}, \frac{X^2_{j+1} - 2X^2_j + X^2_{j-1}}{h^2}\right].$$

The approximation of the first derivative is denoted as $X_{\bar{u},j}$ and the second derivative as $X_{\bar{u}u,j}$.

The semi-implicit scheme for equation (3) has the form of

$$X^{k+1}_j - \tau\frac{X^{k+1}_{\bar{u}u,j}}{Q^2(X^k_{\bar{u},j})} + \tau\alpha_j\frac{X^{k+1}_{\bar{u},j}}{Q(X^k_{\bar{u},j})} = X^k_j + \tau F(u,t)\frac{X^{\perp k}_{\bar{u},j}}{Q(X^k_{\bar{u},j})}, \quad j = 1,\cdots,m-1, k = 0,\cdots,N_T-1, \quad (8)$$

where $Q(x_1,x_2)$, $X^\perp_{\bar{u},j}$, m, and α_j have the same meaning as for semi-discrete scheme. $X^k_j \approx X(jh,k\tau)$, τ is a time step and N_T is the number of time steps. The matrix structure of one component X^{k+1} looks like

$$\begin{pmatrix} 1+\frac{2t}{h^2Q^2}-\frac{t\alpha}{hQ} & \frac{-t}{h^2Q^2} & 0 & \cdots \\ \frac{-t}{h^2Q^2}+\frac{t\alpha}{hQ} & \ddots & \ddots & \ddots \\ 0 & \ddots & & \\ \vdots & \ddots & & \end{pmatrix}.$$

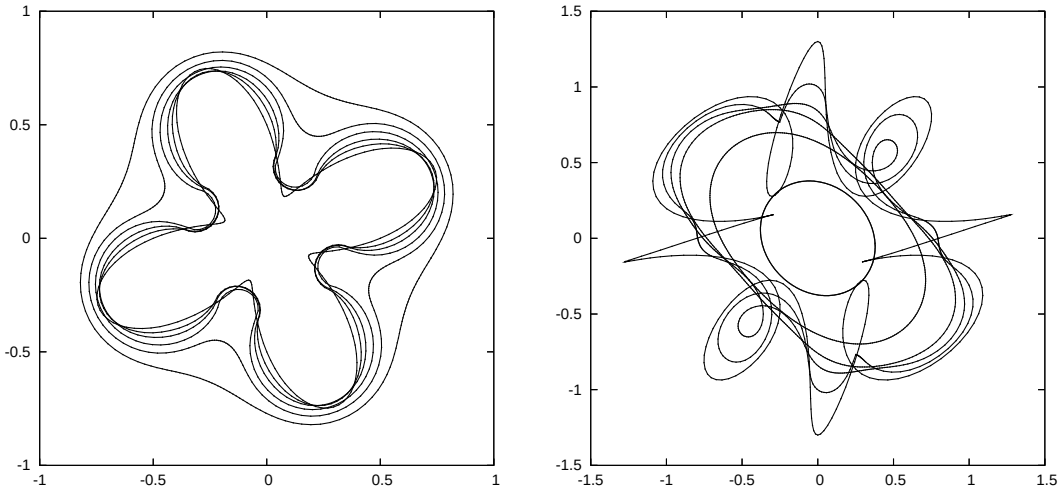
The scheme (8) is solved for each k by means of a factorization method.

5 Results of numerical simulation

In this section, the results of numerical simulation by previous schemes will be presented. Schemes were tested on open and closed curves with and without tangential redistribution of points. At first, we simulated evolution of a circle and compared with analytical solution. Experimental order of convergence and absolute error were measured. See [13].

Figure 2(a) illustrates the evolution of a closed curve with external force variable in space. Values are as follows: $F = 10$ for $|X| < 0.35$, $F = -5$ for $|X| > 0.35$. The initial curve is a four-leaf clover curve. The positive force moves the curve to the center but the negative force move the rest of the curve from the center. In a short time, high curvature appears and neglects the positive external force $F = 10$. It causes the whole curve to expand.

Figure 2(b) shows the evolution of the curve which intersects itself. Intersections cause singularities and it is not possible to continue evolution because curvature goes to infinity. That is why we added regularization term $Q(x_1, x_2) = \sqrt{x_1^2 + x_2^2 + \varepsilon^2}$ to the scheme. This allows the curve to evolve beyond singularities. One can see that the curve evolves to the circle.



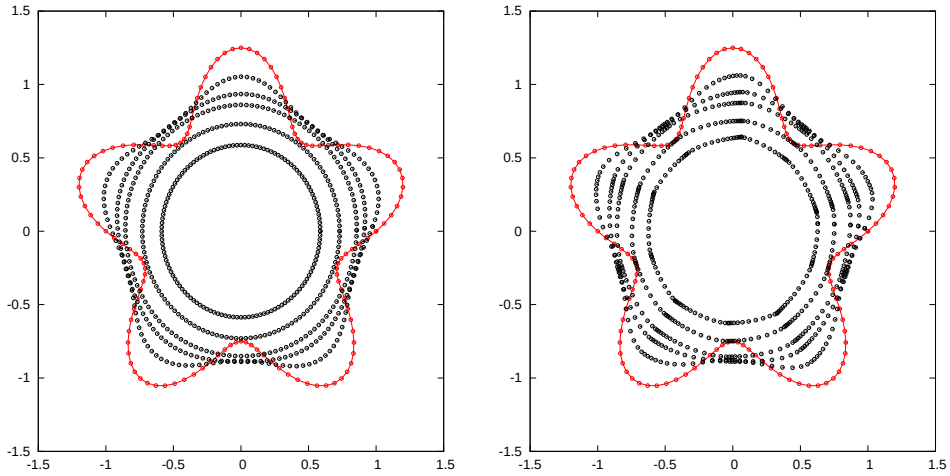
(a) $t \in (0, 0.184)$, $h = 1/200$, $F = 10$ for $|X| < 0.35$, $F = -5$ for $|X| > 0.35$

(b) $F = 0$, $t \in (0, 0.495)$, $h = 1/400$

Figure 2: Time evolution of closed curves, scheme (7)

Figures 3(a) and 3(b) show the evolution of star shaped curve using the scheme (8) for $\alpha = 0$ and α computed by (4). It was already said that the equation (2) contains some tangential force which helps to move points along curve and improve the stability of the computation. In Figure 3(a), one can see that the points are equally distributed at the end of simulation. On the other hand, when the tangential force is completely removed, points stay in groups causing long computation times and worse precision (see Figure 3(b)).

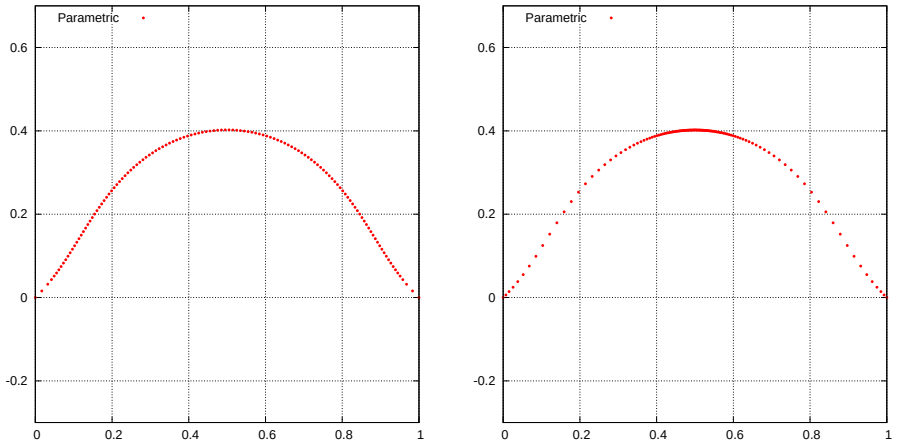
For the simulation of dislocation dynamics, long time computations with periodical



(a) With tangential force. Eq. (8), $\alpha = 0$ (b) Without tangential force. Eq. (8), α according to (4)

Figure 3: Comparison of evolution with and without tangential force.

change of the external force are needed. In this case, grouping of points happens for both equations (2) and (5) and one has to use for example the algorithm mentioned at the end of section 3 (see also [12]). Figures 4(a) and 4(b) present the position of an open curve at $t = 1.38$. There is an external force $F = 3$ which periodically changes the sign (i.e., $F = 3$ or $F = -3$). This force causes the curve to move up and down. Why we need this periodic force is described in the next chapter. Figure 4(b) shows the evolution by equation (2). One can see that the middle part of the curve contains many points while ends are very sparse. If tangential redistribution is used (Figure 4(a)), all points are equally redistributed along the curve.

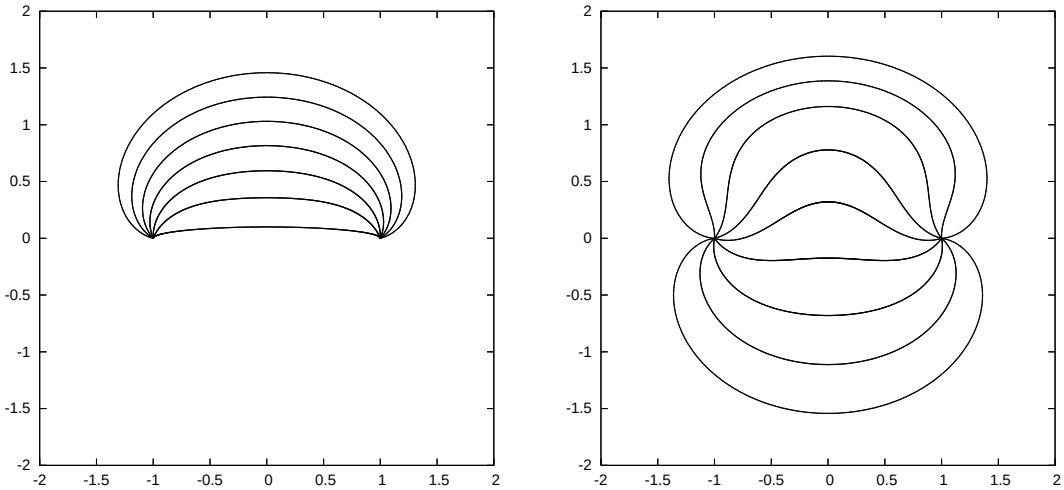


(a) With tangential redistribution (b) Without tangential redistribution

Figure 4: Comparison of evolution with and without tangential redistribution.

6 Dislocation dynamics

The main purpose of this work is to simulate dislocation dynamics. Dislocation curves in the material evolves in time. It means they change the shape, the topology, etc. The following simulations should tell us whether this way can be used for this purpose.



(a) Dislocation curve expands ($F = -3$) for $t \in (0, 0.54)$.

(b) Curve goes back and expands to the other side ($F = 3$) for $t > 0.54$.

Figure 5: The evolution of the dislocation curves with variable external force F.

Figure 5 illustrates the evolution of dislocation curves in time. The external force $F = -3$ is applied to the curve which causes the expansion in the up direction. At time $t = 0.54$, the direction of the force is changed. In the real material, one can observe similar behavior.

During the curve evolution, a barrier which blocks the curve evolution can appear. According to the value of external force in the barrier, the curve can be either locked or can pass through it. Figure 6 shows the case with weak force. Dislocation curve expands by means of $F = -3$ until it reaches the barrier made by the spatially variable force $F = 9$ at $x_2 = 1.7$. This barrier is not strong enough to lock the curve because at the ends of the barrier there is a very high curvature. High curvature causes strong force against the external force. The curve can leave the barrier and continues to expand. The simulation in Figure 7 was computed for $t \in (0, 2.1)$.

In the case of strong external force, the curve is locked in the barrier and cannot continue in evolution. The curve can only expand to sides. The barrier is again at $x_2 = 1.7$ and the value of barrier force is $|F| = 35$. Figure 7(a) illustrates the curve expansion by $F = -3$ and the case when it is locked at the barrier ($t \in (0, 1.5)$). Figure 7(b) shows the curve shrinking by $F = 3$ for $t \in (1.5, 3)$. The curve is locked at the barrier and cannot go back to a straight line. This example should simulate the real dislocation curve expansion when the curve is locked at so called *channel*.

The evolution of the curve at the endless channel is shown in Figure 8. Again, the

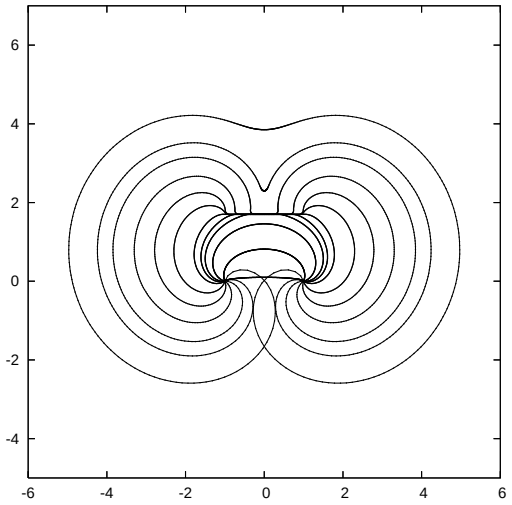
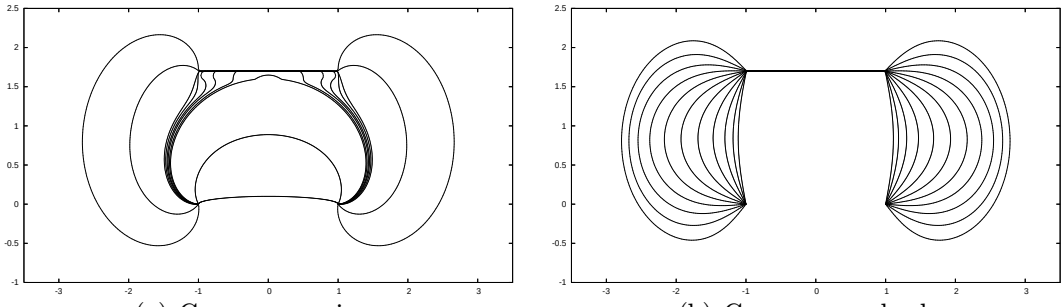


Figure 6: The dislocation curve expands over a barrier created by spatially variable external force.



(a) Curve expansion (b) Curve moves back

Figure 7: Spatially variable external force F with high value, $t \in (0, 1.5)$.

endless channel is created by spatially variable external force. The curve cannot cross these barriers (at $x_2 = 1.2$ and $x_2 = 0$).

7 Conclusion

The dislocation dynamics simulation is important in practice because dislocations affect many material properties. Dislocation dynamics can be mathematically simulated by mean curvature flow. We presented a method based on a parametric approach and two numerical schemes. We applied the model to situations similar to the real context. The scheme had to be improved by an algorithm for tangential redistribution of points.

Acknowledgement. This work was partly supported by the project MSM No. 6840770100 "Applied Mathematics in Technical and Physical Sciences" and by the project No. LC06052 "Nečas Center for Mathematical Modelling" of the Ministry of Education, Youth and

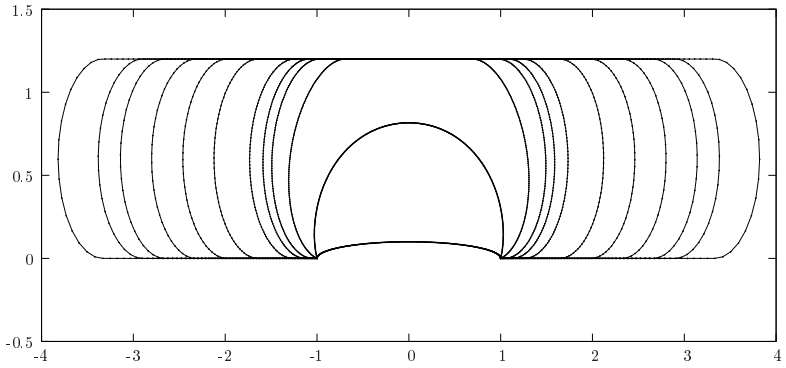


Figure 8: Curve evolution at the channel.

Sport of the Czech Republic.

References

[1] J. A. Sethian. *Level set methods*. Cambridge University Press, 1996.

[2] J. A. Sethian. *Level set methods and fast marching methods*. Cambridge University Press, 1999.

[3] G. Dziuk, A. Schmidt, A. Brillard, and C. Bandle. *Course on mean curvature flow*. Manuscript 75p., Freiburg, 1994.

[4] M. Beneš. *Phase field model of microstructure growth in solidification of pure substances*. Acta Math. Univ. Comenian, 2001.

[5] K. Deckelnick, G. Dziuk. *Mean curvature flow and related topics*. Freiburg, 2002.

[6] K. Mikula, D. Ševčovič. *Computational and qualitative aspects of evolution of curves driven by curvature and external force*. Computing and Visualization in Science, Vol. 6, No. 4, pp. 211–225, 2004.

[7] K. Mikula, D. Ševčovič. *Evolution of plane curves driven by a nonlinear function of curvature and anisotropy*. SIAM Vol. 61, No. 5, pp. 1473–1501, 2001.

[8] K. Mikula, D. Ševčovič. *A direct method for solving an anisotropic mean curvature flow of plane curves with and external force*. Mathematical methods in the applied sciences, Vol. 27, No. 13, pp. 1545–1565, 2004.

[9] V. Minárik, J. Kratochvíl. *Dislocation Dynamics – Analytical Description of the Interaction Force between Dipolar Loops*. Proceedings of the Czech Japanese Seminar in Applied Mathematics. Prague: Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, 2006.

[10] V. Minárik, J. Kratochvíl, K. Mikula. *Numerical Simulation of Dislocation Dynamics by Means of Parametric Approach*. Proceedings of the Czech Japanese Seminar in Applied Mathematics. Prague: Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague. ISBN 80-01-03181-0, pp. 128–138, 2005.

[11] V. Minárik, J. Kratochvíl, K. Mikula, M. Beneš. *Numerical simulation of dislocation dynamics*. Numerical Mathematics and Advanced Applications, ENUMATH 2003, pp. 631–641, Springer Verlag, ISBN 3-540-21460-7, 2004.

[12] D. Ševčovič, S. Yazaki. *On a motion of plane curves with a curvature adjusted tangential velocity*. Preprint, 2007.

[13] P. Pauš. *Numerical simulation of dislocation dynamics*. Proceedings of Slovak-Austrian Congress, Podbanské, 2007.

Parallel Algorithms for Diffusion-based Tensor Field Visualization in Mathematics and Medicine

Pavel Strachota

1st year of PGS, email: pavel.strachota@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Michal Beneš, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. We propose a vector/tensor field visualization technique based on solving an initial boundary value problem for the Allen-Cahn equation with diffusion anisotropy controlled by a tensor field. Focus is put on the details of the numerical solution of the given problem by means of the method of lines, presenting the results of both theoretical and experimental convergence analysis. Afterwards, the aspects of the parallel implementation of the numerical algorithm are dealt with, concentrating on the efficiency benchmarks. Finally, vector field visualization results are presented and the possibilities of applying the method in MR tractography are outlined.

Abstrakt. Vyvinuli jsme metodu pro zobrazování vektorových a tenzorových polí založenou na řešení smíšené úlohy pro Allenovu-Cahnovu rovnici s anizotropní difuzí, která je řízena tenzorovým polem. Tento článek se soustředí na detaily numerického řešení daného problému metodou přímek a uvádí výsledky teoretické i experimentální konvergenční analýzy. Dále se zabývá aspekty paralelní implementace numerického algoritmu s důrazem na testování efektivity. Nakonec jsou prezentovány výsledky vizualizace a jsou nastíněny možnosti uplatnění této metody v MR traktografii.

1 Introduction

Vector fields or tensor fields are a common output of simulations in computational fluid dynamics and are also produced as an intermediate result of the Diffusion Tensor Imaging (DTI) medical examination technique [3]. DTI represents one of the applications of a magnetic resonance (MR) scanner and is capable of tracking the diffusion of H_2O molecules in human brain (as well as some other tissues of an animal). This motion is directly related to the neural fiber structures in the brain. In order to interpret the described kind of data, an appropriate visualization technique needs to be chosen. In this paper, we propose an approach based on solving a problem for the Allen-Cahn partial differential equation [7, 2], introduce a numerical method for its solution and investigate properties of the method itself as well as the properties of its parallel implementation.

The main idea of the method is as follows. Suppose a static vector field $\boldsymbol{v}$ is defined in a rectangular domain $\Omega = (0, L^1) \times (0, L^2)$. Generating a noisy texture in Ω and making it undergo an anisotropic diffusion process with the diffusion focused in the direction $\boldsymbol{v}(\boldsymbol{x})$ at each point $\boldsymbol{x}$, the streamlines of the vector field emerge as "smudges". In addition to smearing, one may impose advection on the texture in order to interpret the flow of the fluid along the vector field.

2 Formulation

Let $p : \mathcal{J} \times \Omega \mapsto \mathbb{R}$, $p = p(t, \boldsymbol{x})$ be the function of texture intensity at each point $\boldsymbol{x} \in \Omega$ and at the time $t \in \bar{\mathcal{J}}$, where $\mathcal{J} = (0, T)$ is the time interval. The initial boundary value problem for the Allen-Cahn equation with advection (see [8]) reads

$$\xi \frac{\partial p}{\partial t} + \xi \boldsymbol{v} \cdot \nabla p = \xi \nabla \cdot T^0 (\nabla p) + \frac{1}{\xi} f_0(p) + c_0 F \qquad \text{in } \mathcal{J} \times \Omega, \tag{1}$$

$$p|_{\partial\Omega} = 0 \qquad \text{on } \mathcal{J} \times \partial\Omega, \tag{2}$$

$$p|_{t=0} = I \qquad \text{in } \Omega, \tag{3}$$

where

$$f_0(p) = p(1-p)\left(p - \frac{1}{2}\right).$$

In (1), the term $\nabla \cdot T^0(\nabla p)$ is responsible for anisotropic diffusion of p focused into the direction of the vector field. Consider a vector $\boldsymbol{\eta} = (\eta^1, \eta^2)^{\mathrm{T}} \in \mathbb{R}^2$ and denote the coordinates of $\boldsymbol{\eta}$ in the orthonormal basis $(\frac{1}{v}\boldsymbol{v}, \frac{1}{v}\boldsymbol{v}^\perp)$ by $\tilde{\eta}^1, \tilde{\eta}^2$. The anisotropic operator T^0 is defined as

$$T^0(\boldsymbol{\eta}) = \Phi^0(\boldsymbol{\eta}) \Phi^0_\eta(\boldsymbol{\eta}),$$

where

$$\Phi^0(\boldsymbol{\eta}) = \sqrt{\alpha \cdot (\tilde{\eta}^1)^2 + \beta \cdot (\tilde{\eta}^2)^2}\,, \quad \Phi^0_\eta(\boldsymbol{\eta}) = \begin{pmatrix} \partial_{\eta^1} \Phi^0(\boldsymbol{\eta}) \\ \partial_{\eta^2} \Phi^0(\boldsymbol{\eta}) \end{pmatrix}. \tag{4}$$

The coefficients α, β depend on the vector field and should be chosen such that the absolute value of T^0 is largest in the case when the directions of $\boldsymbol{v}$ and ∇p coincide. Our choice is

$$\alpha = \kappa\left(1 + \sigma\left|\boldsymbol{v}\right|\right),\ \beta = \kappa, \quad \kappa, \sigma > 0.$$

The term $\boldsymbol{v} \cdot \nabla p$ in (1) causes texture advection [7, 2]. The polynomial f_0 makes *nucleation* occur during the time. In this context, nucleation is a formation of areas where the value of p is near 0 or 1. As described for example in [7, 1], the parameter ξ is proportional to the diffuse interface layer between such areas. ξ is chosen such that it is small in comparison with the dimensions of Ω. The sense of the parameter F is related to the problem of mean curvature flow and is explained e.g. in [7, 2].

In the context of visualization, if $I : \Omega \mapsto \mathbb{R}$ represents the intensity of a noisy texture at each point, the solution p will reflect the gradual diffusion of the initial image I with increasing time. Both the state of p at some final time T and the entire solution evolution can be regarded as the result.

2.1 Tensor field visualization

The anisotropy introduced with the T^0 operator is a generalization of the *diffusion tensor* model [9], based on replacing $T^0(\nabla p)$ in (1) by the term

$$\boldsymbol{D} \nabla p,$$

where $\boldsymbol{D}$ is a symmetric positive definite matrix. Indeed, it is easy to verify that defining

$$\Phi^0(\boldsymbol{\eta}) = \sqrt{\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\eta}},$$

we obtain

$$T^0(\nabla p) = \boldsymbol{D}\nabla p.$$

On the other hand, our special choice (4) can be expressed in terms of the diffusion tensor model. The corresponding tensor is such that it has the form

$$\boldsymbol{D} = \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix}$$

expressed in the basis $(\frac{1}{v}\boldsymbol{v}, \frac{1}{v}\boldsymbol{v}^\perp)$.

3 Numerical solution

For numerical solution, we use the *method of lines,* which converts the problem (1-3) to the solution of the system of ODEs in the form

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \boldsymbol{f}(t, \boldsymbol{p}). \tag{5}$$

The spatial discretization is carried out by the finite difference method; for the temporal discretization, we employ the 4th-order Runge-Kutta-Merson solver with adaptive time stepping. First, let us introduce the notations

$$\begin{gathered}
\boldsymbol{h} = (h^1, h^2),\ h^k := \frac{L^k}{m^k},\ k \in \{1, 2\}, \\
\boldsymbol{x}_{i,j} = (x_i^1, x_j^2) = (i \cdot h^1, j \cdot h^2), \\
\omega_h = \left\{ \boldsymbol{x}_{i,j} \middle| i = 1, ..., m^1 - 1,\ j = 1, ..., m^2 - 1 \right\}, \\
\bar{\omega}_h = \left\{ \boldsymbol{x}_{i,j} \middle| i = 0, ..., m^1,\ j = 0, ..., m^2 \right\},\ \gamma_h = \bar{\omega}_h - \omega_h, \\
\mathcal{H}_h = \left\{ u \middle| u : \bar{\omega}_h \to \mathbb{R} \right\},\ u_{i,j} = u(\boldsymbol{x}_{i,j}), \\
\mathcal{P}_h w = w\big|_{\bar{\omega}_h} \in \mathcal{H}_h \text{ defined for any } w : \Omega \mapsto \mathbb{R}.
\end{gathered} \tag{6}$$

In the sense of (6), we introduce the following difference quotients approximating the derivatives, gradient and divergence:

$$\begin{gathered}
u_{\bar{x}^1,i,j} = \frac{u_{i,j} - u_{i-1,j}}{h^1} \quad , \quad u_{x^1,i,j} = \frac{u_{i+1,j} - u_{i,j}}{h^1}, \\
u_{\bar{x}^2,i,j} = \frac{u_{i,j} - u_{i,j-1}}{h^2} \quad , \quad u_{x^2,i,j} = \frac{u_{i,j+1} - u_{i,j}}{h^2}, \\
\bar{\nabla}_h u = (u_{\bar{x}^1}, u_{\bar{x}^2}) \quad , \quad \nabla_h u = (u_{x^1}, u_{x^2}), \\
\nabla_h \cdot \boldsymbol{V} = V^1_{x^1} + V^2_{x^2},\ \bar{\nabla}_h \cdot \boldsymbol{V} = V^1_{\bar{x}^1} + V^2_{\bar{x}^2},\ \boldsymbol{V} = \left(V^1, V^2\right)^{\mathrm{T}}.
\end{gathered}$$

Using the above definitions, we assemble the semi-discrete scheme of the problem (1-3) for the unknown grid function $p^h : \mathcal{J} \to \mathcal{H}_h$ which represents the vector of functions of time $\boldsymbol{p}$ in (5):

$$\xi\frac{\mathrm{d}p^h}{\mathrm{d}t} + \xi\mathcal{P}_h(\boldsymbol{v}) \cdot \bar{\nabla}_h p^h = \xi\nabla_h \cdot T^0(\bar{\nabla}_h p^h) + \frac{1}{\xi} f_0(p^h) + c_0 F \qquad \text{in } \mathcal{J} \times \omega_h, \tag{7}$$

$$p^h|_{\gamma_h} = 0 \qquad \text{on } \mathcal{J} \times \gamma_h, \tag{8}$$

$$p^h(0) = \mathcal{P}_h I \qquad \text{in } \omega_h. \tag{9}$$

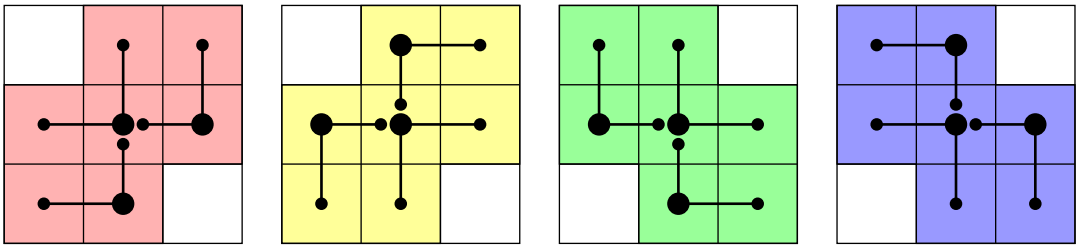


Figure 1: Versions of $\nabla \cdot T^0(\nabla p)$ discretization used for assembling the weighted scheme.

4 Numerical scheme weighting

The original numerical scheme (7-9) suffers from artificial numerical isotropic diffusion, which in accordance with the spectral error analysis theory [4] affects structures in p^h containing high frequencies. As a result, the formation of streamlines is degraded. However, due to the asymmetry of the scheme, the amount of additional isotropic diffusion depends on the direction of the vector field $\boldsymbol{v}$. This property of the scheme has been exploited to design mixed forward/backward difference quotients approximating the gradient by

$$\begin{aligned} \widetilde{\nabla}_h u &= (u_{\bar{x}^1}, u_{x^2})^{\mathrm{T}}, \\ \breve{\nabla}_h u &= (u_{x^1}, u_{\bar{x}^2})^{\mathrm{T}} \end{aligned}$$

and the divergence by

$$\begin{aligned} \widetilde{\nabla}_h \cdot \boldsymbol{V} &= V^1_{\bar{x}^1} + V^2_{x^2}, \\ \breve{\nabla}_h \cdot \boldsymbol{V} &= V^1_{x^1} + V^2_{\bar{x}^2}. \end{aligned}$$

These expressions allow four versions of discretization of the term $\nabla \cdot T^0(\nabla p)$ in (1), as listed in Figure 1. Two complementary scheme asymmetries are obtained, corresponding to two perpendicular directions of the strongest numerical diffusion. Finally, all discretization versions are combined into a single scheme, weighting them with respect to the direction of the vector field. As a result, the weighted scheme always prefers the discretization version with a weaker numerical diffusion. The improvement can be observed in Figure 2.

5 Convergence analysis

The work [8] contains a detailed convergence analysis, proving the following theorem:

Theorem 1. *Let* $I \in \mathrm{H}^1_0(\Omega) \cap \mathrm{C}(\bar{\Omega})$, $\boldsymbol{v} \in \mathrm{C}(\bar{\Omega})^2$. *Then the solution* p^h *of the semidiscrete scheme (7-9) converges in* $\mathrm{L}_2(\mathcal{J}; \mathrm{L}_2(\Omega))$ *to the unique weak solution* p *of the anisotropic*

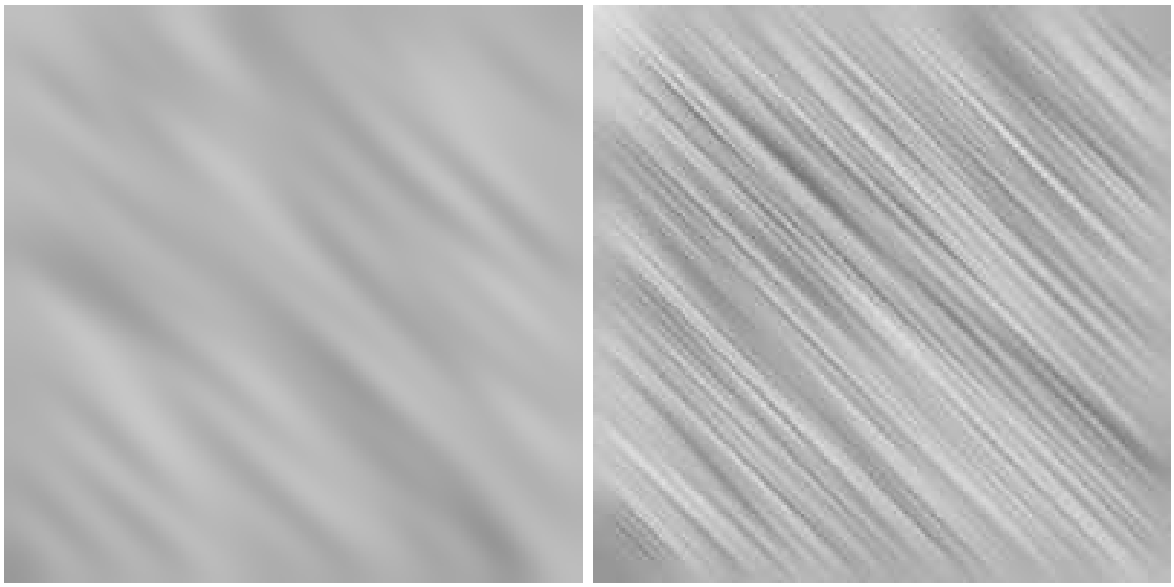


Figure 2: Visualization of the straight vector field in the direction corresponding to the strongest artificial diffusion in the original scheme. Result obtained by the original scheme (left) and the weighted scheme (right). All parameters were identical for both computations.

diffusion problem (1-3), where p satisfies

$$\begin{aligned} p &\in \mathrm{L}_2(\mathcal{J}; \mathrm{H}_0^1(\Omega)), \\ \frac{\partial p}{\partial t} &\in \mathrm{L}_2(\mathcal{J}; \mathrm{L}_2(\Omega)). \end{aligned}$$

The proof is based on interpolation theory, suitable a priori estimates and the method of compactness.

5.1 Experimental proof of convergence

In addition to the theoretical results, the measurement of the experimental order of convergence (EOC) has been performed for both the original and the weighted schemes. EOC is obtained by computing the solution on a sequence of gradually refining grids and is defined as

$$\mathrm{EOC} = \frac{\log\left(\frac{\mathrm{Error}_i}{\mathrm{Error}_{i-1}}\right)}{\log\left(\frac{\|\boldsymbol{h}_i\|}{\|\boldsymbol{h}_{i-1}\|}\right)},$$

where $\|\boldsymbol{h}\| = \max_j h^j$ and Error_i is the difference of the i-th solution from the precise solution measured in an appropriate norm. The results indicating the convergence are summarized in Table 1 and Table 2.

Grid size	$\boldsymbol{h}$	$L_\infty(\mathcal{J}; L_2(\Omega))$ error	$L_\infty(\mathcal{J}; L_\infty(\Omega))$ error	EOC in $L_\infty(\mathcal{J}; L_2(\Omega))$	EOC in $L_\infty(\mathcal{J}; L_\infty(\Omega))$
100×100	0.01	0.0257814	0.2909448	-	-
200×200	0.005	0.0082178	0.1145193	1.6495124	1.3451547
400×400	0.0025	0.0027553	0.0465855	1.5765111	1.2976364
800×800	0.00125	0.0007728	0.0133288	1.8339980	1.8053344

Table 1: Experimental order of convergence of the original scheme (7-9).

Grid size	$\boldsymbol{h}$	$L_\infty(\mathcal{J}; L_2(\Omega))$ error	$L_\infty(\mathcal{J}; L_\infty(\Omega))$ error	EOC in $L_\infty(\mathcal{J}; L_2(\Omega))$	EOC in $L_\infty(\mathcal{J}; L_\infty(\Omega))$
100×100	0.01	0.0249912	0.2056547	-	-
200×200	0.005	0.0073023	0.0633514	1.7750009	1.6987763
400×400	0.0025	0.0022840	0.0196849	1.6768129	1.6862861
800×800	0.00125	0.0006455	0.0060901	1.8230700	1.6925603

Table 2: Experimental order of convergence of the weighted scheme.

6 Parallelization

In order to allow reasonably fast calculations on large grids, a parallel implementation of the numerical algorithm has been developed by means of the MPI library (see [6]). Very fine grids are necessary e.g. for the convergence verification of several numerical scheme modifications.

The idea of parallelization of the finite difference algorithm is to decompose the grid ω_h into blocks, each of those being handled by a different process. Our choice was to compose a block of several rows of the grid. The processes belonging to the adjacent blocks need to interchange (synchronize) data in order to complete each step of the Runge-Kutta method.

Since the method of lines is extremely demanding on the amount of synchronization, much attention has been paid to benchmarking and scalability improvement of the code. Using the nonblocking communication operations, we are able to optimize the flow of the calculation by requesting the operations as soon as possible and completing them as late as possible. Since the synchronized data is used for calculation of the border nodes of the blocks only, we can calculate the value of the right hand side of (5) in the interior of the block before the communication is complete.

6.1 Dynamic load balancing

In addition to message passing optimization, an interesting method of dynamic load balancing during the calculation has been developed, making it possible to utilize non-homogeneous clusters for efficient computation. The technique is based on the changes of the block sizes, corresponding to the particular processes. For a given period, each

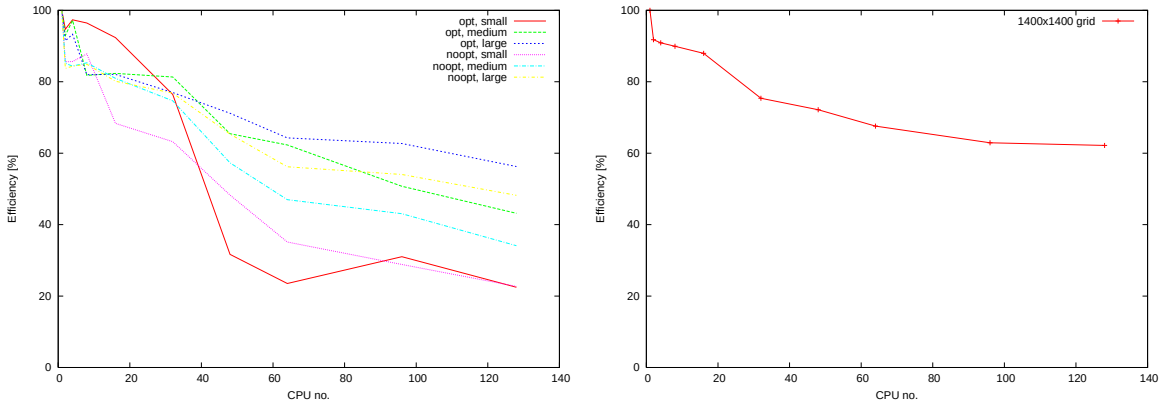


Figure 3: Efficiency testing on CLX, different grid sizes (small=200×200, medium=400×400, large=800×800) and MPI communication order (opt=nonblocking "request soon, complete late", noopt=nonblocking "all at once").

process accumulates the wall time of its autonomous calculations (between synchronizations). The acquired time values are then converted to relative speeds of the processes. Afterwards, the master process calculates the new block sizes, proportional to the process speeds. We assume that with such block sizes, the idle times of the processes (waiting for synchronization) should be eliminated. Rearrangement of the blocks requires data to be redistributed among the blocks. The algorithm implementation tries to minimize the amount of data being sent and provides mechanisms to avoid meaningless rearrangements (when the changes to be made are negligible).

Of course, the proposed load balancing system is not suitable for advanced homogeneous cluster solutions controlled by load sharing managers such as LSF or PBS. On such a system, all nodes utilized by the user application have the same performance and they are fully at its disposal for the whole program run time. No load balancing is therefore necessary.

Extensive efficiency benchmarks have been performed on the CLX Linux cluster at CINECA, Italy. Some efficiency results are shown in Figure 3.

7 Visualization results

The results of the numerical algorithm based on the weighted scheme and applied to some sample vector fields are displayed in Figure 4. Color visualization has been achieved by separately solving the above problem for the R, G, B components of the image. The advection term in (1) together with a suitable choice of the boundary condition may be useful for flow visualization, as depicted in Figure 5.

8 Application in MR Tractography

As already suggested in the introduction, each DTI examination generates a tensor field describing the directional distribution of water diffusion in human brain [10, 3, 5]. As

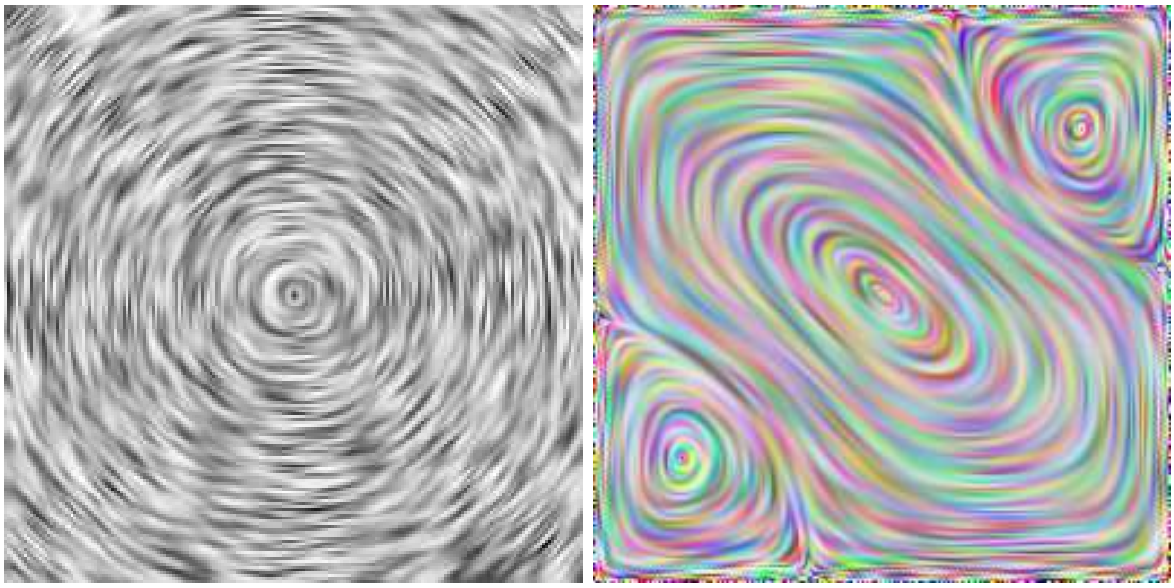


Figure 4: Sample vector field visualizations.

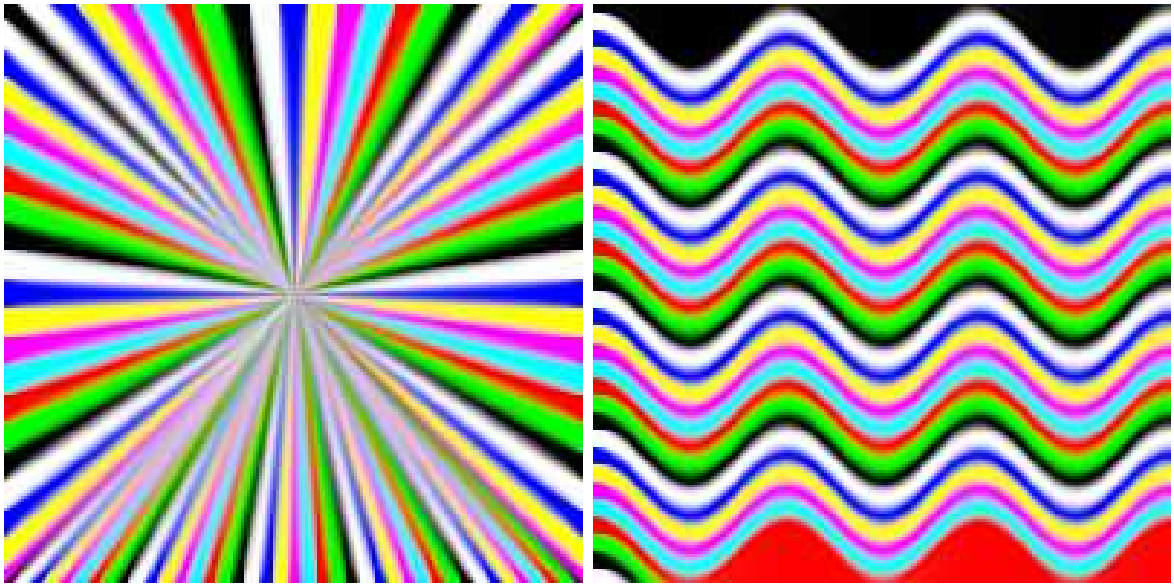


Figure 5: Flow visualization by means of advection together with a stripe-like Dirichlet boundary condition.

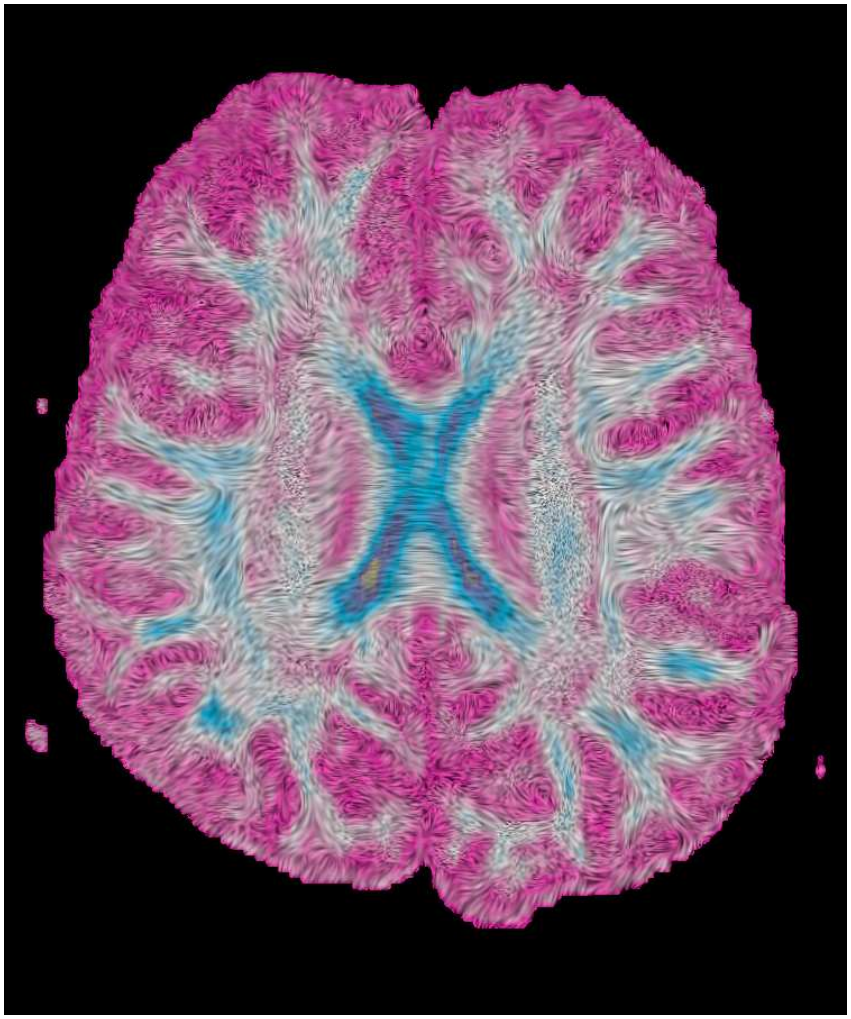


Figure 6: MR tractography using anisotropic diffusion. Colorized by fractional anisotropy [9].

neural fibers act as tubes for the H_2O molecules, tracking the direction of the strongest diffusion may help discover the pathways of the neural tracts. This process is called *tractography*. Using the choice

$$T^0(\boldsymbol{\eta}) = \boldsymbol{D}\boldsymbol{\eta}$$

in (1), we are able to employ our visualization approach to reveal the streamlines of the tensor field, interpretable as neural fiber bundles. A sample result of neural tract visualization in a transverse plane is displayed in Figure 6.

9 Conclusions

We have developed an optimized parallel algorithm for the numerical solution of the anisotropic diffusion problem (1-3). The solution is suitable for use as a vector or tensor field visualization technique, as demonstrated on several examples. The convergence analysis justifies the suitability of both the original and the weighted schemes. Thorough

tests of algorithm efficiency prove the possibility to create a well scalable parallel implementation of the method of lines despite the huge amount of necessary communication.

Acknowledgements: This work was carried out under the HPC-EUROPA project (RII3-CT-2003-506079), with the support of the European Community - Research Infrastructure Action under the FP6 Structuring the European Research Area Programme. Partial support of the project "Jindřich Nečas Center for Mathematical Modeling", No. LC06052.

References

[1] M. Beneš. *Mathematical analysis of phase-field equations with numerically efficient coupling terms.* Interfaces and Free Boundaries **3** (2001), 201–221.

[2] M. Beneš. *Diffuse-interface treatment of the anisotropic mean-curvature flow.* Applications of Mathematics **48** (2003), 437–453.

[3] D. L. Bihan et al. *Diffusion tensor imaging: Concepts and applications.* Journal of Magnetic Resonance Imaging **13** (2001), 534–546.

[4] H. Lomax, T. H. Pulliam, and D. W. Zingg. *Fundamentals of Computational Fluid Dynamics.* Springer, (2001).

[5] S. Mori and J. Zhang. *Principles of diffusion tensor imaging and its applications to basic neuroscience research.* Neuron **51** (2006), 527–539.

[6] M. Snir, S. Otto, S. Huss-Ledermann, D. Walker, and J. Dongarra. *The Complete MPI Reference.* The MIT Press, (1995).

[7] P. Strachota. Degenerate diffusion in mathematical visualization. Research work, Czech Technical University in Prague, (2005). (in Czech).

[8] P. Strachota. Anisotropic diffusion in mathematical visualization. Master's thesis, Czech Technical University in Prague, (2007). (in Czech).

[9] D. Tschumperlé and R. Deriche. Tensor field visualization with PDE's and application to DT-MRI fiber visualization. INRIA Sophia-Antipolis, Odyssée Lab, France, (2004).

[10] C. F. Westin et al. *Processing and visualization for diffusion tensor MRI.* Medical Image Analysis **6** (2002), 93–108.

Static vs. Dynamic Classifier Systems in Classifier Aggregation*

David Štefka

3rd year of PGS, email: stefka@cs.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Martin Holeňa, Institute of Computer Science, AS CR

Abstract. Classifier aggregation is a method for improving quality of classification – instead of using just one classifier, a team of classifiers is created, and the outputs of the individual classifiers are aggregated into the final prediction. Common methods for classifier aggregation are *static*, i.e., they do not adapt to the currently classified pattern. In this paper, we introduce a formalism of *dynamic* classifier systems, which use the concept of dynamic classification confidence to dynamically adapt to the currently classified pattern. Results of experiments with quadratic discriminant classifiers on four artificial and four real-world benchmark datasets show that dynamic classifier systems can significantly outperform static classifier systems.

Abstrakt. Spojování klasifikátorů je metoda pro zlepšení kvality klasifikace – místo používání jednoho klasifikátoru je vytvořen tým klasifikátorů a výstupy jednotlivých klasifikátorů jsou poté agregovány pro získání finální predikce. Většina metod pro agregaci klasifikátorů je *statická*, tj. agregace se nepřizpůsobuje konkrétním klasifikovaným vzorům. V tomto článku popíšeme *dynamické* systémy klasifikátorů, které používají koncept dynamické konfidence klasifikace, aby se přizpůsobily konkrétnímu vzoru. Výsledky experimentů na 4 umělých a 4 reálných datových množinách ukazují, že dynamické systémy mohou dosahovat signifikantně lepších výsledků než statické systémy.

1 Introduction

Classification is a process of dividing objects (called *patterns*) into disjoint sets called *classes* [7]. Many machine learning algorithms for classification have been developed – for example naive Bayes classifiers, linear and quadratic discriminant classifiers, k-nearest neighbor classifiers, support vector machines, neural networks, or decision trees. If the quality of classification (i.e., the classifier's predictive power) is low, there are several methods we can use to improve it.

One comonly used technique for improving classification quality is called *classifier combining* [11] – instead of using just one classifier, we create and train a team of classifiers, let each of them predict independently, and then combine (aggregate) their results. It can be shown that a team of classifiers can perform better in the classification task than any of the individual classifiers.

*The research presented in this paper was partially supported by the Program "Information Society" under project 1ET100300517 and by the grant ME949 of the Ministry of Education, Youth and Sports of the Czech Republic.

There are two main approaches to classifier combining: *classifier selection* [1, 17] and *classifier aggregation* [12, 10]. If a pattern is submitted for classification, the former technique uses some rule to select one particular classifier, and only this classifier is used to obtain the final prediction. The latter technique uses some aggregation rule to aggregate the results of all the classifiers in a team to get the final prediction.

A common drawback of classifier aggregation methods is that they are static, i.e., they are not adapted to the particular patterns that are currently classified. In other words, the aggregation is specified during a training phase, prior to classifying a test pattern. However, if we use the concept of dynamic classification confidence (i.e., the extent to which we can "trust" the output of the particular classifier for the currently classified pattern), the aggregation algorithms can take into account the fact that "this classifier is not good *for this particular pattern*".

Surprisingly, such dynamic classifier systems are not used very often in classifier combining. However, there has already been some work done in the field of dynamic classifier systems – Robnik-Šikonja and Tsymbal et al. [13, 14] study dynamic aggregation of random forests [4], i.e., dynamic classifier systems of decision trees. The authors report significant improvements in classification quality when using dynamic voting compared to simple voting. However, they study dynamic classifier systems only in the context of random forests, and they use only confidence measures based on the so-called margin.

In this paper, we provide a general formalism of dynamic classification confidence measures and dynamic classifier systems, and we experimentally study the performance of confidence-free classifier systems (i.e., systems that do not utilize classification confidence at all), static classifier systems (i.e., systems that use only "global" confidence of a classifier), and dynamic classifier systems (i.e., systems that adapt to the particular pattern submitted for classification).

The paper is structured as follows. In Section 2, we introduce the formalism of classifier combining, namely in Section 2.1, we define basic concepts of classification, in Section 2.2 we introduce the concept of classification confidence, and we introduce three dynamic confidence measures, in Section 2.3 we deal with classifier teams and ensembles, and in Section 2.4, we finally define classifier systems and show several examples of dynamic classifier systems. In Section 3, we experimentally compare performance of the proposed dynamic classifier systems. Section 4 then concludes the paper.

2 Formalism of Classifier Combining with Classification Confidence

2.1 Classification

Throughout the rest of the paper, we use the following notation. Let $\mathcal{X} \subseteq \mathbf{R}^n$ be a n-dimensional *feature space*, an element $\vec{x} \in \mathcal{X}$ of this space is called a *pattern*, and let $C_1, \ldots, C_N \subseteq \mathcal{X}$, $N \geq 2$, be disjoint sets called *classes*. The index of the class a pattern $\vec{x}$ belongs to will be denoted as $c(\vec{x})$ (i.e., $c(\vec{x}) = i$ iff $\vec{x} \in C_i$). The goal of classification is to determine to which class a given pattern belongs, i.e., to predict $c(\vec{x})$ for unknown patterns.

Definition 1. We call a *classifier* every mapping $\phi : \mathcal{X} \to [0,1]^N$, where $[0,1]$ is the unit interval, and $\phi(\vec{x}) = (\mu_1(\vec{x}), \ldots, \mu_N(\vec{x}))$ are *degrees of classification (d.o.c.)* to each class.

The d.o.c. to class C_j expresses the extent to which the pattern belongs to class C_j (if $\mu_i(\vec{x}) > \mu_j(\vec{x})$, it means that the pattern $(\vec{x})$ belongs to class C_i rather than to C_j). Depending on the classifier type, it can be modelled by probability, fuzzy membership, etc.

Remark 2. This definition is of course not the only way how a classifier can be defined, but in the theory of classifier combining, this one is used most often [11].

Definition 3. Classifier ϕ is called *crisp*, iff $\forall \vec{x} \in \mathcal{X}\ \exists i$, such that:

$$\mu_i(\vec{x}) = 1, \text{ and } \forall j \neq i\ \mu_j(\vec{x}) = 0.$$

Definition 4. Let ϕ be a classifier, $\vec{x} \in \mathcal{X}$, $\phi(\vec{x}) = (\mu_1(\vec{x}), \ldots, \mu_N(\vec{x}))$. *Crisp output* of ϕ on $\vec{x}$ is defined as $\phi_{cr}(\vec{x}) = \arg\max_{i=1,\ldots,N} \mu_i(\vec{x})$.

2.2 Classification Confidence

Classification confidence expresses the degree of trust we can give to a classifier ϕ when classifying a pattern $\vec{x}$. It is modelled by a mapping κ_ϕ.

Definition 5. Let ϕ be a classifier. We call a *confidence measure* of classifier ϕ every mapping $\kappa_\phi : \mathcal{X} \to [0,1]$.

The higher the confidence, the higher the probability of correct classification. $\kappa_\phi(\vec{x}) = 0$ means that the classification may not be correct, while $\kappa_\phi(\vec{x}) = 1$ means the classification is probably correct. However, κ_ϕ does not need to be modelled by a probability measure.

A confidence measure can be either *static*, i.e., it is a constant of the classifier, or *dynamic*, i.e., it adjusts itself to the currently classified pattern.

Definition 6. Let ϕ be a classifier and κ_ϕ its confidence measure. We call κ_ϕ *static*, iff it is constant in $\vec{x}$, we call κ_ϕ *dynamic* otherwise.

Remark 7. Since static confidence measures are constant, independent on the currently classified pattern, we will omit the pattern $(\vec{x})$ in the notation, i.e., we will denote them just κ_ϕ.

Remark 8. In the rest of the paper, we will use the indicator operator I, defined as $I(\text{true}) = 1$, $I(\text{false}) = 0$.

2.2.1 Static confidence measures

After the classifier has been trained, we can use a validation set to assess its predictive power as a whole (from a global point of view). These methods include accuracy, precision, sensitivity, resemblance, etc. [7, 9], and we can use these measures as static confidence measures. In this paper, we will use the Global Accuracy measure.

Global Accuracy (GA) of a classifier ϕ is defined as the proportion of correctly classified patterns from the validation set:

$$\kappa_{\phi}^{(GA)} = \frac{\sum_{\vec{y}\in\mathcal{M}} I(\phi_{cr}(\vec{y}) \stackrel{?}{=} c(\vec{y}))}{|\mathcal{M}|}, \tag{1}$$

where $\mathcal{M}$ is the validation set of ϕ (i.e., a set of patterns ϕ was not trained on, intended for parameter fine-tuning), and $\phi_{cr}(\vec{y})$ is the crisp output of ϕ on $\vec{y}$.

2.2.2 Dynamic confidence measures

An easy way how a dynamic confidence measure can be defined is to compute some property on patterns neighboring with $\vec{x}$. Let $N(\vec{x})$ denote a set of neighboring validation patterns. In this paper, we define $N(\vec{x})$ as the set of k patterns nearest to $\vec{x}$ under Euclidean metric. Now we will define three dynamic confidence measures which use $N(\vec{x})$:

Euclidean Local Accuracy (ELA) measures the local accuracy of ϕ in $N(\vec{x})$:

$$\kappa_{\phi}^{(ELA)}(\vec{x}) = \frac{\sum_{\vec{y}\in N(\vec{x})} I(\phi_{cr}(\vec{y}) \stackrel{?}{=} c(\vec{y}))}{|N(\vec{x})|}, \tag{2}$$

where $\phi_{cr}(\vec{y})$ is the crisp output of ϕ on $\vec{y}$.

Euclidean Local Match (ELM) is based on the ideas from [5], and measures the proportion of patterns in $N(\vec{x})$ from the same class as ϕ is predicting for $\vec{x}$:

$$\kappa_{\phi}^{(ELM)}(\vec{x}) = \frac{\sum_{\vec{y}\in N(\vec{x})} I(\phi_{cr}(\vec{x}) \stackrel{?}{=} c(\vec{y}))}{|N(\vec{x})|}, \tag{3}$$

where $\phi_{cr}(\vec{x})$ is the crisp output of ϕ on $\vec{x}$.

Euclidean Average Margin (EAM) is defined as mean value of the margin [4, 13, 14] in $N(\vec{x})$:

$$\kappa_{\phi}^{(EAM)}(\vec{x}) = \frac{\sum_{\vec{y}\in N(\vec{x})} mg(\phi(\vec{y}))}{|N(\vec{x})|}, \tag{4}$$

where the margin is defined as $mg(\phi(\vec{y})) =$

$$\begin{cases} \mu_{c(\vec{y})}(\vec{y}) - \max\limits_{\substack{i=1,\ldots,N \\ i\neq c(\vec{y})}} \mu_i(\vec{y}) & \text{if } \phi_{cr}(\vec{y}) = c(\vec{y}), \\ 0 & \text{otherwise.} \end{cases}, \tag{5}$$

where $\phi(\vec{y}) = (\mu_1(\vec{y}), \ldots, \mu_N(\vec{y}))$, and $\phi_{cr}(\vec{y})$ is the crisp output of ϕ on $\vec{y}$.

The dynamic confidence measures defined in this section have one drawback – they need to compute $N(\vec{x})$, which can be time-consuming, and sensitive to the similarity measure used. There are also dynamic confidence measures, which compute the classification confidence directly from $\phi(\vec{x})$, e.g., the ratio of the highest degree of classification to the sum of all degrees of classification. However, our preliminary experiments with such measures with quadratic discriminant classifiers and random forests show that such confidence measures give very poor results.

2.3 Classifier Teams

In classifier combining, instead of using just one classifier, a team of classifiers is created, and the team is then aggregated into one final classifier. If we want to utilize classification confidence in the aggregation process, each classifier must have its confidence measure defined.

Definition 9. *Classifier team* is a tuple $(\mathcal{T}, \mathcal{K})$, where $\mathcal{T} = (\phi_1, \ldots, \phi_r)$, $r \in \mathbf{N}$, $r \geq 2$ is a set of classifiers, and $\mathcal{K} = (\kappa_{\phi_1}, \ldots, \kappa_{\phi_r})$ is a set of corresponding confidence measures.

If a classifier team consists only of classifiers of the same type, which differ only in their parameters, dimensionality, or training sets, the team is usually called an *ensemble of classifiers*. For this reason the methods which create a team of classifiers are sometimes called *ensemble methods*. The restriction to classifiers of the same type is not essential, but it ensures that the outputs of the classifiers are consistent. Well-known methods for ensemble creation are *bagging* [3], *boosting* [8], *error correction codes* [11], or *multiple feature subset* methods [2].

If a pattern is submitted for classification, the team of classifiers gives us two different informations – outputs of the individual classifiers (a *decision profile*), and values of classification confidences of the classifiers (a *confidence vector*).

Definition 10. Let $(\mathcal{T} = (\phi_1, \ldots, \phi_r), \mathcal{K} = (\kappa_{\phi_1}, \ldots, \kappa_{\phi_r}))$ be a classifier team, and let $\vec{x} \in \mathcal{X}$. Then we define *decision profile* $\mathcal{T}(\vec{x}) \in [0,1]^{r,N}$ and *confidence vector* $\mathcal{K}(\vec{x}) \in [0,1]^r$ as

$$\mathcal{T}(\vec{x}) = \begin{pmatrix} \phi_1(\vec{x}) \\ \phi_2(\vec{x}) \\ \vdots \\ \phi_r(\vec{x}) \end{pmatrix} = \begin{pmatrix} \mu_{1,1}(\vec{x}) & \mu_{1,2}(\vec{x}) & \ldots & \mu_{1,N}(\vec{x}) \\ \mu_{2,1}(\vec{x}) & \mu_{2,2}(\vec{x}) & \ldots & \mu_{2,N}(\vec{x}) \\ & & \ddots & \\ \mu_{r,1}(\vec{x}) & \mu_{r,2}(\vec{x}) & \ldots & \mu_{r,N}(\vec{x}) \end{pmatrix}, \; \mathcal{K}(\vec{x}) = \begin{pmatrix} \kappa_{\phi_1}(\vec{x}) \\ \kappa_{\phi_2}(\vec{x}) \\ \vdots \\ \kappa_{\phi_r}(\vec{x}) \end{pmatrix} \tag{6}$$

Remark 11. Here we use the notation $\mathcal{T}$ for both the set of classifiers, and for the decision profile, and similarly for $\mathcal{K}$. To avoid any confusion, the decision profile and confidence vector will be always followed by $(\vec{x})$.

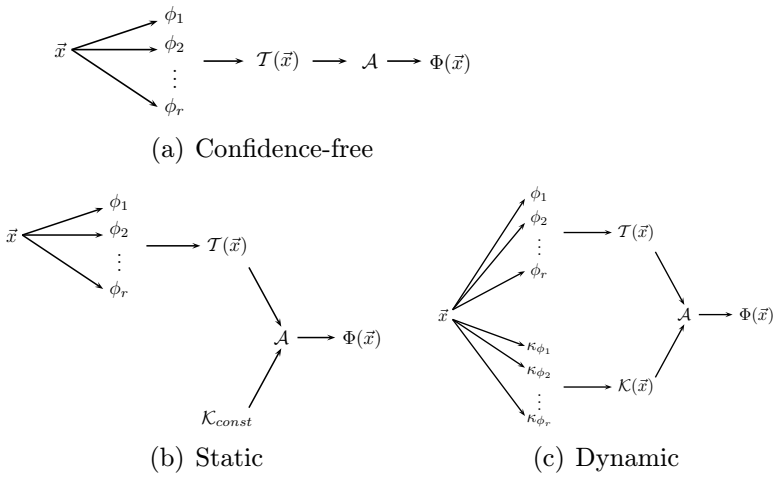
2.4 Classifier Systems

After the pattern $\vec{x}$ has been classified by all the classifiers in the team, and the confidences were computed, these outputs have to be aggregated using a *team aggregator*, which takes the decision profile as its first argument, the confidence vector as its second argument, and returns the aggregated degrees of classification to all the classes.

Definition 12. Let $r, N \in \mathbf{N}$, $r, N \geq 2$. A *team aggregator* of dimension (r, N) is any mapping $\mathcal{A} : [0,1]^{r,N} \times [0,1]^r \to [0,1]^N$.

A classifier team with an aggregator will be called a *classifier system*. Such system can be also viewed as a single classifier.

Definition 13. Let $(\mathcal{T}, \mathcal{K})$ be a classifier team, and let $\mathcal{A}$ be a team aggregator of dimension (r, N), where r is the number of classifiers in the team, and N is the number of classes.



(a) Confidence-free

(b) Static

(c) Dynamic

Figure 1: Schematic comparison of confidence-free, static, and dynamic classifier systems.

The triple $\mathcal{S} = (\mathcal{T}, \mathcal{K}, \mathcal{A})$ is called a *classifier system*. We define an *induced classifier* of $\mathcal{S}$ as a classifier Φ, defined as

$$\Phi(\vec{x}) = \mathcal{A}(\mathcal{T}(\vec{x}), \mathcal{K}(\vec{x})).$$

Depending on the way how a classifier system utilizes the classification confidence, we can distinguish several kinds of classifier systems.

Definition 14. Let $(\mathcal{T}, \mathcal{K})$ be a classifier team. $(\mathcal{T}, \mathcal{K})$ is called *static*, iff $\forall \kappa \in \mathcal{K} : \kappa$ is a static confidence measure. $(\mathcal{T}, \mathcal{K})$ is called *dynamic*, iff $\forall \kappa \in \mathcal{K} : \kappa$ is a dynamic confidence measure.

Definition 15. Let $\mathcal{A}$ be a team aggregator of dimension (r, N). We call $\mathcal{A}$ *confidence-free*, iff it is constant in the second argument.

Definition 16. Let $\mathcal{S} = (\mathcal{T}, \mathcal{K}, \mathcal{A})$ be a classifier system. We call $\mathcal{S}$ *confidence-free*, iff $\mathcal{A}$ is confidence-free. We call $\mathcal{S}$ *static*, iff $(\mathcal{T}, \mathcal{K})$ is static, and $\mathcal{A}$ is not confidence-free. We call $\mathcal{S}$ *dynamic*, iff $(\mathcal{T}, \mathcal{K})$ is dynamic, and $\mathcal{A}$ is not confidence-free.

Confidence-free systems do not utilize the classification confidence at all (for example a team of classifiers aggregated by simple voting). Static systems utilize classification confidence, but only as a global property (for example a team of classifiers aggregated by weighted voting with constant classifier weights). Dynamic systems utilize classification confidence in a dynamic way, i.e. the aggregation is adapted to the particular pattern submitted for classification (for example a team of classifiers aggregated by weighted voting with classifier weights computed for every pattern). The different approaches are schematically shown in Fig. 1.

Many methods for aggregating the team of classifiers into one final classifier have been proposed in the literature [11, 12]. These methods comprise simple arithmetic rules (voting, sum, product, maximum, minimum, average, weighted average, etc.), fuzzy integral, Dempster-Shafer fusion, second-level classifiers, decision templates, and many others.

In the following text, we define several team aggregators. We will use the notation from Def. 10 and Def. 13. Let $\Phi(\vec{x}) = \mathcal{A}(\mathcal{T}(\vec{x}), \mathcal{K}(\vec{x})) = (\mu_1(\vec{x}), \ldots, \mu_N(\vec{x}))$.

Mean value aggregation (MV) is the most common aggregation technique. Its aggregator is defined as

$$\mu_j(\vec{x}) = \frac{\sum_{i=1,\ldots,r} \mu_{i,j}(\vec{x})}{r}. \tag{7}$$

If the classifiers in the team are crisp, MV coincides with voting.

Static weighted mean aggregation (SWM) computes aggregated d.o.c. as weighted mean of d.o.c. given by the individual classifiers, where the weights are static classification confidences:

$$\mu_j(\vec{x}) = \frac{\sum_{i=1,\ldots,r} \kappa_{\phi_i} \mu_{i,j}(\vec{x})}{\sum_{i=1,\ldots,r} \kappa_{\phi_i}}. \tag{8}$$

Dynamic weighted mean aggregation (DWM) has the same aggregator as SWM, but the weights are dynamic classification confidences:

$$\mu_j(\vec{x}) = \frac{\sum_{i=1,\ldots,r} \kappa_{\phi_i}(\vec{x}) \mu_{i,j}(\vec{x})}{\sum_{i=1,\ldots,r} \kappa_{\phi_i}(\vec{x})}. \tag{9}$$

Filtered mean aggregation (FM) has the same aggregator as MV, but prior to computing the aggregated values, the classifiers which have (dynamic) classification confidence lower than $T \in [0, 1]$ are discarded:

$$\mu_j(\vec{x}) = \frac{\sum\limits_{\substack{i=1,\ldots,r \\ \kappa_{\phi_i}(\vec{x}) > T}} \mu_{i,j}(\vec{x})}{|\{\phi \in \mathcal{T} | \kappa_{\phi_i}(\vec{x}) > T\}|}. \tag{10}$$

3 Experiments

To compare confidence-free, static, and dynamic classifier systems, we implemented the algorithms described in Sec. 2.4, and we tested their performance on four artificial (Clouds, Concentric, Gauss_3D, Waveform) and four real-world (Breast, Phoneme, Pima, Satimage) datasets from the Elena database [15] and from the UCI repository [6].

For all the classifier systems we used, the classifier team $(\mathcal{T}, \mathcal{K})$ was an ensemble of quadratic discriminant classifiers [7], created either by the bagging algorithm [3] (which creates classifiers trained on random samples drawn from the original training set with replacement), or by the multiple feature subset method [2] (which creates classifiers using different combinations of features), depending on which method was more suitable for the particular dataset.

For the comparison, we designed the following classifier systems (refer to Section 2.2 and Section 2.4 for the description of the algorithms):

MV confidence-free system aggregated by mean value aggregation

SWM cl. system aggregated by static weighted mean aggregation; as a confidence measure, we used GA

DWM cl. system aggregated by dynamic weighted mean; as a confidence measure, we used ELA, ELM, and EAM

FM cl. system aggregated by filtered mean; as a confidence measure, we used ELA, ELM, and EAM

We also compared the systems' performance with the so-called *non-combined classifier* (NC), i.e., a common quadratic discriminant classifier (the NC classifier represents an approach which we had to use if we could use only one classifier).

All the methods were implemented in Java programming language, and a 10-fold crossvalidation was performed to obtain the results. For the dynamic confidence measures, we used $k = 20$. The threshold T for FM aggregators was set to $T = 0.8$ or $T = 0.9$, depending on the particular dataset. The parameters were set based on some preliminary testing; no fine-tuning or optimization was done.

The results of the testing are shown in Table 1. Mean error rate and standard deviation of the error rate of the induced classifiers from a 10-fold crossvalidation was measured. We also measured statistical significance of the results – at 5% confidence level by the analysis of variance using the Tukey-Kramer method (by the 'multcomp' function from the Matlab statistics toolbox).

The results show that for most datasets, the dynamic classifier systems outperform both confidence-free and static classifier systems. For three datasets, these results were statistically significant. FM usually gives better results than DWM, and if we compare the three dynamic confidence measures, we can say that ELM gives usually the best results, ELA and ELM being slightly worse. However, the performance of the individual confidence measures depends on the particular dataset [16]. Generally speaking, the FM-ELM was the most successfull algorithm in this experiment.

It should be noted that the experimental results from this paper are relevant only to quadratic discriminant classifiers, because for any other classifier types (k-NN, SVM, decision trees, etc.), the dynamic confidence measures could give quite different results.

4 Summary

In this paper, we have studied dynamic classifier aggregation. We have introduced the formalism of classifier systems which can be used with (dynamic) classification confidence, and we have defined confidence-free, static, and dynamic classifier systems. We have introduced three dynamic classification confidence measures (ELA, ELM, EAM), and we have shown a way how these measures can be used in dynamic classifier systems – we have introduced two algorithms for dynamic classifier aggregation.

In our experiments, we have compared the performance of confidence-free, static, and dynamic classifier systems of quadratic discriminant classifiers. The results show that dynamic classifier systems can significantly outperform both confidence-free and static classifier systems.

Table 1: Comparison of the aggregation methods – non-combined classifier (NC), mean value (MV), static weighted mean (SWM) using GA confidence measure, dynamic weighted mean (DWM) using confidence measures ELA, ELM, EAM, and filtered mean (FM) using confidence measures ELA, ELM, EAM. Mean error rate (in %) ± standard deviation of error rate from a 10-fold crossvalidation was measured. The best result is displayed in boldface, statistically significant (at 5% level) improvements to NC, MV, and SWM are marked by footnote signs. The (B/M) after dataset name means whether the ensemble was created by Bagging or Multiple feature subset algorithm.

	Non-combined	Conf.-free	Static		Dynamic		
Dataset	NC	MV	κ	SWM	κ	DWM	FM
Clouds (M)	25.0 ± 1.7	25.0 ± 2.1	GA	24.7 ± 1.6	ELA	23.4 ± 1.5	22.3 ± 1.5 *†‡
					ELM	23.2 ± 1.2	$\mathbf{22.0 \pm 2.1}$ *†‡
					EAM	23.5 ± 1.5	23.3 ± 1.4
Concentric (B)	3.5 ± 1.0	3.8 ± 0.6	GA	4.0 ± 0.8	ELA	3.2 ± 1.1	2.1 ± 1.3 †‡
					ELM	2.9 ± 1.6	$\mathbf{1.8 \pm 0.8}$ *†‡
					EAM	3.8 ± 1.3	4.3 ± 1.5
Gauss_3D (B)	21.4 ± 1.7	21.6 ± 1.1	GA	21.5 ± 2.1	ELA	21.5 ± 1.4	21.7 ± 1.3
					ELM	$\mathbf{21.3 \pm 2.0}$	22.0 ± 1.3
					EAM	21.5 ± 2.0	21.7 ± 1.3
Waveform (B)	14.9 ± 2.5	15.0 ± 1.4	GA	14.8 ± 0.9	ELA	14.7 ± 1.9	15.0 ± 1.2
					ELM	14.8 ± 2.5	$\mathbf{14.5 \pm 1.2}$
					EAM	14.6 ± 2.0	15.5 ± 1.0
Breast (M)	4.8 ± 2.9	4.7 ± 2.5	GA	4.2 ± 2.4	ELA	3.0 ± 2.1	2.9 ± 1.8
					ELM	3.0 ± 1.9	3.1 ± 2.1
					EAM	3.2 ± 2.0	$\mathbf{2.9 \pm 1.7}$
Phoneme (M)	24.7 ± 1.1	23.5 ± 1.6	GA	24.0 ± 1.4	ELA	21.5 ± 1.9 *‡	17.2 ± 1.4 *†‡
					ELM	21.2 ± 1.8 *‡	$\mathbf{16.9 \pm 2.0}$ *†‡
					EAM	21.9 ± 0.9 *	20.7 ± 1.7 *†‡
Pima (M)	27.1 ± 4.4	25.4 ± 3.6	GA	25.0 ± 5.6	ELA	25.8 ± 6.5	24.0 ± 2.7
					ELM	24.0 ± 4.1	25.0 ± 7.4
					EAM	24.8 ± 6.3	$\mathbf{23.5 \pm 5.4}$
Satimage (B)	15.6 ± 1.7	15.5 ± 1.2	GA	15.5 ± 1.7	ELA	15.3 ± 1.6	15.2 ± 2.4
					ELM	15.3 ± 1.3	$\mathbf{14.4 \pm 1.0}$
					EAM	15.5 ± 1.2	15.0 ± 1.5

*Significant improvement to NC
†Significant improvement to MV
‡Significant improvement to SWM

The main contribution of this paper is the verification that the concept of dynamic classification confidence can significantly improve the classification quality, and that it is a general concept, which can be incorporated into the theory of classifier aggregation in a systematic way.

In our future work, we plan to study dynamic classification confidence measures for other classifiers than quadratic discriminant classifier, mainly decision trees and support vector machines, and to study model-specific confidence measures for these classifier types. We will also incorporate local classification confidence into more sophisticated classifier aggregation methods.

References

[1] M. Aksela. Comparison of classifier selection methods for improving committee performance. In 'Multiple Classifier Systems', 84–93, (2003).

[2] S. D. Bay. *Nearest neighbor classification from multiple feature subsets.* Intelligent Data Analysis **3** (1999), 191–209.

[3] L. Breiman. *Bagging predictors.* Machine Learning **24** (1996), 123–140.

[4] L. Breiman. *Random forests.* Machine Learning **45** (2001), 5–32.

[5] S. J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh. Generating estimates of classification confidence for a case-based spam filter. In 'Case-Based Reasoning, Research and Development, 6th Int. Conf., ICCBR 2005, Chicago, USA', H. Muñoz-Avila and F. Ricci, (eds.), volume 3620 of *LNCS*, 177–190. Springer, (2005).

[6] C. B. D.J. Newman, S. Hettich and C. Merz. UCI repository of machine learning databases, (1998). http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition).* Wiley-Interscience, (2000).

[8] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In 'International Conference on Machine Learning', 148–156, (1996).

[9] D. J. Hand. *Construction and Assessment of Classification Rules.* Wiley, (1997).

[10] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. *On combining classifiers.* IEEE Trans. Pattern Anal. Mach. Intell. **20** (1998), 226–239.

[11] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms.* Wiley-Interscience, (2004).

[12] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin. *Decision templates for multiple classifier fusion: an experimental comparison.* Pattern Recognition **34** (2001), 299–314.

[13] M. Robnik-Šikonja. Improving random forests. In 'ECML', J. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, (eds.), volume 3201 of *Lecture Notes in Computer Science*, 359–370. Springer, (2004).

[14] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Dynamic integration with random forests. In 'ECML', J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, (eds.), volume 4212 of *Lecture Notes in Computer Science*, 801–808. Springer, (2006).

[15] UCL MLG. Elena database, (1995). http://www.dice.ucl.ac.be/mlg/?page=Elena.

[16] D. Štefka. Dynamic classifier systems for classifier aggregation. In 'Proccedings of PhD. Conference 2008'. Institute of Computer Science/MatfyzPress, Prague.

[17] K. Woods, J. W. Philip Kegelmeyer, and K. Bowyer. *Combination of multiple classifiers using local accuracy estimates.* IEEE Trans. Pattern Anal. Mach. Intell. **19** (1997), 405–410.

High-energy Asymptotics of the Spectrum of a Rectangular Periodic Network

Ondřej Turek

2nd year of PGS, email: turekond@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Pavel Exner, Nuclear Physics Institute, AS CR,
Pierre Duclos, Centre de Physique Théorique, C.N.R.S

Abstract. The aim of this paper is to study asymptotical behaviour of an infinite two-dimensional periodic rectangular network. A general singular coupling in the vertices is supposed. We will show that for certain vertex couplings the system behaves significantly differently from the spectrum of the corresponding one-dimensional system, i.e. that the classification of spectral asymptotics of one-dimensional periodic networks is not applicable here.

Abstrakt. Předmětem práce je studium asymptotického chování spektra nekonečné dvoudimenzionální periodické obdélníkové mřížky. Ve vrcholech předpokládáme obecnou singulární vazbu. Ukážeme, že pro určité vazby může tento systém vykazovat chování, které se silně odlišuje od chování spektra příslušných jednodimenzionálních systémů, jinak řečeno, že zde nelze využít klasifikaci spektra známou pro jednodimenzionální sítě.

1 Introduction

The term *quantum graph* denotes an ordered pair (Γ, H), where Γ is a metric graph (an undirected graph with a metric) and H is a Hamiltonian on Γ, i.e. self-adjoint differential operator of the second order acting on the graph edges as a minus second derivative (see [4]). These mathematical objects serve as natural models of graph-like structures of nanometer sizes, which may be made of various materials, usually of semiconductors. The technological progress in last decades of the twentieth century has enabled a mass production of such microscopic structures and, consequently, their practical utilization. As a result, the theory of quantum graphs gained a wide application potential, which is hitherto growing. This fact attracted the attention of mathematical physicist, and at the end of the eighties an intensive study in this field has begun, which continues till this time. However, it is a relatively new theory with many open problems remaining.

One of the open problems concerns spectra of infinite periodic systems. It is well known from a more general theory that a periodic system has a band spectrum. The interesting and important question is, how the asymptotics of the spectral bands looks like.

The easiest situation is a line with periodically located point interactions of the same type. One can consider either δ-interaction, which is a classical, very well examined Kronnig-Penney model, or a general point interaction. The case of infinite one-dimensional periodic network with a general point interaction in each vertex has been already described in the work [1]. The authors studied high-energy asymptotics of the

spectrum and have derived the following result: The system has a purely absolutely continuous spectrum and the structure of its spectral bands can conform only to one of the following three situations:

- band widths are asymptotically constant, gap widths grow asymptotically linearly,
- widths of both bands and gaps are growing,
- band widths grow asymptotically linearly, gap widths are asymptotically constant.

In this paper we will deal with a natural generalization of a one-dimensional network, namely with a planar rectangular network (see Fig. 1). We will ask if it is true that the assymptotics of the spectral bands is described by the three situations enumerated above, or if there is an interaction for which a new type of assympotics arises.

2 Vertex coupling

Let $v \in V$ be a vertex with n outgoing edges. Let us denote the wavefunctions on these edges by $\psi_1 \dots, \psi_n$. The limits of these functions and their first derivatives (in the outgoing sense) in the vertex v form two vectors:

$$\Psi(0) = \begin{pmatrix} \psi_1(0) \\ \vdots \\ \psi_n(0) \end{pmatrix}, \quad \Psi'(0) = \begin{pmatrix} \psi_1'(0) \\ \vdots \\ \psi_n'(0) \end{pmatrix}.$$

All physically admissible boundary conditions can be described by the group of unitary matrices in the following sense: Boundary conditions in a vertex are admissible if and only if there is a unitary matrix U such that

$$(U - I)\Psi(0) + \mathrm{i}(U + I)\Psi'(0) = 0\,. \tag{1}$$

As a result, a family of admissible boundary conditions can be parametrized by n^2 real parameters.

The most common type is the δ-coupling, already mentioned in the introduction. It corresponds to a unitary matrix U given as a sum $a \cdot I + b \cdot J$, where $a = -1$, $b = \frac{2}{n+\mathrm{i}\alpha}$ ($\alpha \in \mathbb{R}$), I is an identity matrix and J is a matrix, whose all elements are equal to 1.

3 Spectral condition

Consider an infinite rectangular network with the cell parameters a and b (see Fig. 1). Let a coupling corresponding to a given unitary matrix U be imposed on all vertices - on every vertex the same coupling. Our aim is to describe the high-energy assymptotics of the spectrum and to find if there is a matrix U for which the spectral properties show a significantly different behaviour with respect to infinite one-dimensional networks.

The considered graph is obviously a periodic system, thus it is natural to analyse it using the Floquet decomposition. Let us consider an elementary cell according to the Fig. 1, for the wavefunction we use the notation marked in the figure.

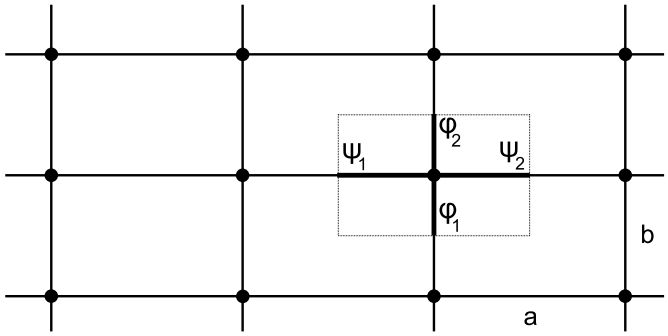


Figure 1: A periodic two-dimensional network

Let there be a particle with the energy E confined to this graph. First of all, we realize that for any matrix U the corresponding Hamiltonian is bounded below, thus its spectrum is bounded below as well. Therefore, to study its high-energy asymptotics it suffices to consider $E > 0$. For notation purposes, let us denote $E = k^2$, $k > 0$. Since the Hamiltonian acts as a minus second derivative, the wavefunction on each edge has to be a linear combination of the functions $\mathrm{e}^{\mathrm{i}kx}$ and $\mathrm{e}^{-\mathrm{i}kx}$, i.e.

$$\begin{aligned}
\psi_1(x) &= C_1^+\mathrm{e}^{\mathrm{i}kx} + C_1^-\mathrm{e}^{-\mathrm{i}kx}, \quad x \in [-a/2, 0] \\
\psi_2(x) &= C_2^+\mathrm{e}^{\mathrm{i}kx} + C_2^-\mathrm{e}^{-\mathrm{i}kx}, \quad x \in [0, a/2] \\
\varphi_1(x) &= D_1^+\mathrm{e}^{\mathrm{i}kx} + D_1^-\mathrm{e}^{-\mathrm{i}kx}, \quad x \in [-b/2, 0] \\
\varphi_2(x) &= D_2^+\mathrm{e}^{\mathrm{i}kx} + D_2^-\mathrm{e}^{-\mathrm{i}kx}, \quad x \in [0, b/2]
\end{aligned} \tag{2}$$

Moreover, the wavefunctions have to satisfy the boundary conditions in the vertex, i.e.

$$(U - I)\begin{pmatrix} \psi_1(0) \\ \psi_2(0) \\ \varphi_1(0) \\ \varphi_2(0) \end{pmatrix} + \mathrm{i}(U + I)\begin{pmatrix} -\psi_1'(0) \\ \psi_2'(0) \\ -\varphi_1'(0) \\ \varphi_2'(0) \end{pmatrix} = 0\,. \tag{3}$$

For the Floquet decomposition we suppose that the wavefunctions satisfy the conditions

$$\begin{aligned}
\psi_2(a/2) &= \mathrm{e}^{\mathrm{i}\theta_1}\psi_1(-a/2) \qquad & \psi_2'(a/2) &= \mathrm{e}^{\mathrm{i}\theta_1}\psi_1'(-a/2) \\
\varphi_2(b/2) &= \mathrm{e}^{\mathrm{i}\theta_2}\varphi_1(-b/2) \qquad & \varphi_2'(b/2) &= \mathrm{e}^{\mathrm{i}\theta_2}\varphi_1'(-b/2)
\end{aligned} \tag{4}$$

for some $\theta_1, \theta_2 \in [-\pi, \pi)$.

Substituting (2) into (4) enables one to rewrite (3) in the form

$$[U(M - kN) - (M + kN)]\begin{pmatrix} C_1^+ \\ C_1^- \\ D_1^+ \\ D_1^- \end{pmatrix} = 0\,, \tag{5}$$

where the matrices M and N are given by

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \mathrm{e}^{\mathrm{i}(\theta_1 - ak)} & \mathrm{e}^{\mathrm{i}(\theta_1 + ak)} & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \mathrm{e}^{\mathrm{i}(\theta_2 - bk)} & \mathrm{e}^{\mathrm{i}(\theta_2 + bk)} & 0 & 0 \end{pmatrix}, \qquad N = \begin{pmatrix} -1 & 1 & 0 & 0 \\ \mathrm{e}^{\mathrm{i}(\theta_1 - ak)} & -\mathrm{e}^{\mathrm{i}(\theta_1 + ak)} & 0 & 0 \\ 0 & 0 & -1 & 1 \\ \mathrm{e}^{\mathrm{i}(\theta_2 - bk)} & -\mathrm{e}^{\mathrm{i}(\theta_2 + bk)} & 0 & 0 \end{pmatrix}.$$

The functions (2) correspond to a nonzero solution iff the vector $\left(C_1^+, C_1^-, D_1^+, D_1^-\right)$ is nonzero. Therefore, a number k^2 belongs to the spectrum of the Hamiltonian if and only if (5) has a non-trivial solution for certain pair (θ_1, θ_2), in other words, if there is a pair (θ_1, θ_2) such that

$$\det\left(U(M - kN) - (M + kN)\right) = 0\,. \tag{6}$$

It can be easily shown that the determinant on the LHS is equal to the term

$$\begin{aligned}\left[C_{22}\left(\mathrm{e}^{\mathrm{i}\theta_2}\right)^2 + C_{21}\mathrm{e}^{\mathrm{i}\theta_2} + C_{20}\right]\left(\mathrm{e}^{\mathrm{i}\theta_1}\right)^2 + \\ + \left[C_{12}\left(\mathrm{e}^{\mathrm{i}\theta_2}\right)^2 + C_{11}\mathrm{e}^{\mathrm{i}\theta_2} + C_{10}\right]\mathrm{e}^{\mathrm{i}\theta_1} + C_{02}\left(\mathrm{e}^{\mathrm{i}\theta_2}\right)^2 + C_{01}\mathrm{e}^{\mathrm{i}\theta_2} + C_{00}\,,\end{aligned}$$

where C_{ij} are expression not containing the Floquet parameters θ_1 and θ_2.

It is convenient to divide equation (6) by $\mathrm{e}^{\mathrm{i}\theta_1}\mathrm{e}^{\mathrm{i}\theta_2}$ and then rearrange the terms:

$$\begin{aligned}C_{11} + \left(C_{21}\mathrm{e}^{\mathrm{i}\theta_1} + C_{01}\mathrm{e}^{-\mathrm{i}\theta_1}\right) + \left(C_{12}\mathrm{e}^{\mathrm{i}\theta_2} + C_{10}\mathrm{e}^{-\mathrm{i}\theta_2}+\right) + \\ + \left(C_{22}\mathrm{e}^{\mathrm{i}\theta_1}\mathrm{e}^{\mathrm{i}\theta_2} + C_{00}\mathrm{e}^{-\mathrm{i}\theta_1}\mathrm{e}^{-\mathrm{i}\theta_2}\right) + \left(C_{20}\mathrm{e}^{\mathrm{i}\theta_1}\mathrm{e}^{-\mathrm{i}\theta_2} + C_{02}\mathrm{e}^{-\mathrm{i}\theta_1}\mathrm{e}^{\mathrm{i}\theta_2}\right) = 0\,,\end{aligned} \tag{7}$$

where

$$\begin{aligned}C_{22} &= 16k^2(u_{12}u_{34} - u_{32}u_{14})\\ C_{00} &= 16k^2(u_{21}u_{43} - u_{41}u_{23})\\ C_{20} &= 16k^2(u_{12}u_{43} - u_{42}u_{13})\\ C_{02} &= 16k^2(u_{21}u_{34} - u_{31}u_{24})\end{aligned}$$

$$\begin{aligned}C_{21} =& k^3 \cdot 8\mathrm{i}\sin bk \cdot (u_{12} + \det U(2,1) - u_{32}u_{13} - u_{42}u_{14} + u_{12}u_{33} + u_{12}u_{44}) +\\ &+ k^2 \cdot 16\cos bk \cdot (-u_{12} + \det U(2,1)) +\\ &+ k \cdot 8\mathrm{i}\sin bk \cdot (u_{12} + \det U(2,1) + u_{32}u_{13} + u_{42}u_{14} - u_{12}u_{33} - u_{12}u_{44})\\ C_{01} =& k^3 \cdot 8\mathrm{i}\sin bk \cdot (u_{21} + \det U(1,2) - u_{31}u_{23} - u_{41}u_{24} + u_{21}u_{33} + u_{21}u_{44}) +\\ &+ k^2 \cdot 16\cos bk \cdot (-u_{21} + \det U(1,2)) +\\ &+ k \cdot 8\mathrm{i}\sin bk \cdot (u_{21} + \det U(1,2) + u_{31}u_{23} + u_{41}u_{24} - u_{21}u_{33} - u_{21}u_{44})\end{aligned}$$

$$\begin{aligned}C_{12} =& k^3 \cdot 8\mathrm{i}\sin ak \cdot (u_{34} + \det U(4,3) - u_{14}u_{31} - u_{24}u_{32} + u_{11}u_{34} + u_{22}u_{34}) +\\ &+ k^2 \cdot 16\cos ak \cdot (-u_{34} + \det U(4,3)) +\\ &+ k \cdot 8\mathrm{i}\sin ak \cdot (u_{34} + \det U(4,3) + u_{14}u_{31} + u_{24}u_{32} - u_{11}u_{34} - u_{22}u_{34})\\ C_{10} =& k^3 \cdot 8\mathrm{i}\sin ak \cdot (u_{43} + \det U(3,4) - u_{13}u_{41} - u_{23}u_{42} + u_{11}u_{43} + u_{22}u_{43}) +\\ &+ k^2 \cdot 16\cos ak \cdot (-u_{43} + \det U(3,4)) +\\ &+ k \cdot 8\mathrm{i}\sin ak \cdot (u_{43} + \det U(3,4) + u_{13}u_{41} + u_{23}u_{42} - u_{11}u_{43} - u_{22}u_{43})\end{aligned}$$

$$\begin{aligned}
C_{11} = &- k^4 \cdot 4 \sin ak \sin bk \cdot (1 + \det U + u_{11} + \det U(1,1) + u_{22} + \det U(2,2) + \\
&\quad + u_{33} + \det U(3,3) + u_{44} + \det U(4,4) + \\
&\quad + u_{11}u_{22} + u_{11}u_{33} + u_{11}u_{44} + u_{22}u_{33} + u_{22}u_{44} + u_{33}u_{44} - \\
&\quad - u_{12}u_{21} - u_{13}u_{31} - u_{14}u_{41} - u_{23}u_{32} - u_{24}u_{42} - u_{34}u_{43}) + \\
&+ k^3 \cdot 8\mathrm{i} \cos ak \sin bk \cdot (-1 + \det U - u_{33} + \det U(3,3) - u_{44} + \det U(4,4) + \\
&\quad + u_{11}u_{22} - u_{12}u_{21} - u_{33}u_{44} + u_{34}u_{43}) + \\
&+ k^3 \cdot 8\mathrm{i} \cos bk \sin ak \cdot (-1 + \det U - u_{11} + \det U(1,1) - u_{22} + \det U(2,2) + \\
&\quad + u_{33}u_{44} - u_{34}u_{43} - u_{11}u_{22} + u_{12}u_{21}) + \\
&+ k^2 \cdot 16 \cos ak \cos bk \cdot (1 + \det U + u_{12}u_{21} - u_{11}u_{22} + u_{34}u_{43} - u_{33}u_{44}) + \\
&+ k^2 \cdot 8 \sin ak \sin bk \cdot (-1 - \det U + u_{11}u_{44} + u_{22}u_{33} + u_{11}u_{33} + u_{22}u_{44} - \\
&\quad - u_{14}u_{41} - u_{23}u_{32} - u_{13}u_{31} - u_{24}u_{42} + \\
&\quad + u_{12}u_{21} - u_{11}u_{22} + u_{34}u_{43} - u_{33}u_{44}) + \\
&+ k \cdot 8\mathrm{i} \cos ak \sin bk \cdot (-1 + \det U + u_{33} - \det U(3,3) + u_{44} - \det U(4,4) + \\
&\quad + u_{11}u_{22} - u_{12}u_{21} - u_{33}u_{44} + u_{34}u_{43}) + \\
&+ k \cdot 8\mathrm{i} \cos bk \sin ak \cdot (-1 + \det U + u_{11} - \det U(1,1) + u_{22} - \det U(2,2) + \\
&\quad + u_{33}u_{44} - u_{34}u_{43} - u_{11}u_{22} + u_{12}u_{21}) + \\
&- 4 \sin ak \sin bk \cdot (1 + \det U - u_{11} - \det U(1,1) - u_{22} - \\
&\quad - \det U(2,2) - u_{33} - \det U(3,3) - u_{44} - \det U(4,4) + \\
&\quad + u_{11}u_{22} + u_{11}u_{33} + u_{11}u_{44} + u_{22}u_{33} + u_{22}u_{44} + u_{33}u_{44} - \\
&\quad - u_{12}u_{21} - u_{13}u_{31} - u_{14}u_{41} - u_{23}u_{32} - u_{24}u_{42} - u_{34}u_{43})
\end{aligned}$$

Lemma 1. *Let $U \in \mathbb{C}^{n,n}$ be a unitary matrix, let us denote $\det U = \mathrm{e}^{i\varphi}$. Then:*

$$\begin{aligned}
(1 + \det U) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &\in \mathbb{R} \\
\mathrm{i}(1 - \det U) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &\in \mathbb{R}
\end{aligned} \tag{8}$$

$$\begin{aligned}
(u_{jj} + \det U(j,j)) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &\in \mathbb{R} \qquad \textit{for all} \quad j \in \hat{n} \\
\mathrm{i}(-u_{jj} + \det U(j,j)) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &\in \mathbb{R} \qquad \textit{for all} \quad j \in \hat{n}
\end{aligned} \tag{9}$$

$$\begin{aligned}
(-u_{jk} + \det U(k,j)) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &= \overline{(-u_{kj} + \det U(j,k))\, \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}}} \qquad \textit{for all} \quad j,k \in \hat{n}, j \neq k \\
(u_{jk} + \det U(k,j)) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &= -\overline{(u_{kj} + \det U(j,k))\, \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}}} \qquad \textit{for all} \quad j,k \in \hat{n}, j \neq k
\end{aligned} \tag{10}$$

If moreover $n = 4$ and $\{j,k,\ell,m\} = \{1,2,3,4\}$, then

$$\begin{aligned}
\left[(u_{jj}u_{kk} - u_{jk}u_{kj}) + (u_{\ell\ell}u_{mm} - u_{\ell m}u_{m\ell})\right] \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &\in \mathbb{R} \\
\mathrm{i}\left[(u_{jj}u_{kk} - u_{jk}u_{kj}) - (u_{\ell\ell}u_{mm} - u_{\ell m}u_{m\ell})\right] \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} &\in \mathbb{R}
\end{aligned} \tag{11}$$

$$(u_{jk}u_{\ell\ell} + u_{jk}u_{mm} - u_{j\ell}u_{\ell k} - u_{jm}u_{mk})\, \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} = -\overline{(u_{kj}u_{\ell\ell} + u_{kj}u_{mm} - u_{k\ell}u_{\ell j} - u_{km}u_{mj})\, \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}}} \tag{12}$$

$$(u_{jk}u_{\ell m} - u_{jm}u_{\ell k}) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} = \overline{(u_{kj}u_{m\ell} - u_{j\ell}u_{mk}) \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}}} \tag{13}$$

We would like to stress that double indices wherever in this lemma *do not* mean the Einstein summation.

Proof. The validity of (8) is obvious, to prove other equalities, the following well known formula is useful:

$$[U^{-1}]_{jk} = \frac{(-1)^{j+k} \cdot \det U(k,j)}{\det U}\,.$$

Since U is unitary, it holds $U^{-1} = U^*$, i.e. $[U^{-1}]_{jk} = \overline{u_{kj}}$. Together we have

$$\overline{u_{kj}} = \frac{(-1)^{j+k} \cdot \det U(k,j)}{\det U}\,. \tag{14}$$

□

This lemma implies the following proposition.

Proposition 2. *There are real numbers*

- $V_4, V_3, V_3', V_2, V_2', V_1, V_1', V_0$,
- $W_3, W_3', W_2, W_2', \tilde{W}_2, \tilde{W}_2', W_1, W_1'$,
- $\alpha_3, \beta_3, \alpha_2, \beta_2, \tilde{\alpha}_2, \tilde{\beta}_2, \alpha_1, \beta_1$

that depend only on U, such that equation (7) *can be written as*

$$\begin{aligned}
&- k^4 \cdot \sin ak \sin bk \cdot V_4 + \\
&+ k^3 \cdot [\cos ak \sin bk \cdot V_3 + \cos bk \sin ak \cdot V_3' + \\
&\qquad + \sin bk \cdot W_3 \sin(\theta_1 + \alpha_3) + \sin ak \cdot W_3' \sin(\theta_2 + \beta_3)] + \\
&+ k^2 \cdot [\cos ak \cos bk \cdot V_2 + \sin ak \sin bk \cdot V_2' + \cos bk \cdot W_2 \cos(\theta_1 + \alpha_2) + \\
&\qquad + \cos ak \cdot W_2' \cos(\theta_2 + \beta_2) + \tilde{W}_2 \cdot \cos(\theta_1 + \theta_2 + \tilde{\alpha}_2) + \tilde{W}_2' \cdot \cos(\theta_1 - \theta_2 + \tilde{\beta}_2)\Big] + \\
&+ k \cdot [\cos ak \sin bk \cdot V_1 + \cos bk \sin ak \cdot V_1' + \\
&\qquad + \sin bk \cdot W_1 \sin(\theta_1 + \alpha_1) + \sin ak \cdot W_1' \sin(\theta_2 + \beta_1)] + \\
&+ \sin ak \sin bk \cdot V_0 = 0\,.
\end{aligned} \tag{15}$$

Proof. It suffices to multiply equation (7) by $\mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}}$, then the statement follows almost immediately from Lemma 1. Equalities (8), (8) and (8) imply that the term $C_{11} \cdot \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}}$ can be written as

$$\begin{aligned}
&- k^4 \cdot \sin ak \sin bk \cdot V_4 + k^3 \cdot (\cos ak \sin bk \cdot V_3 + \cos bk \sin ak \cdot V_3') + \\
&+ k^2 \cdot (\cos ak \cos bk \cdot V_2 + \sin ak \sin bk \cdot V_2') + k \cdot (\cos ak \sin bk \cdot V_1 + \cos bk \sin ak \cdot V_1') + \\
&\qquad + \sin ak \sin bk \cdot V_0
\end{aligned}$$

for certain $V_4, V_3, V_3', V_2, V_2', V_1, V_1', V_0 \in \mathbb{R}$. Using Lemma 1, all pairs of terms of (7), that are coupled in parentheses, can be decomposed into several expressions according to the power of k as well. □

4 Spectral behaviour

The positive spectrum of the considered periodic network contains such numbers k^2, for which there exist parameters $\theta_1, \theta_2 \in [-\pi, \pi)$ such that equation (15) is satisfied. Such numbers k^2 form bands, which can have various structure. In the introduction we have reffered to a result concerning one-dimensional network saying that the band structure can be classified into three groups. The common property of all of them is the following: the bands are either assymptotically growing, or assymptotically constant. The aim of this section is to show that for the two-dimensional case such classification is not sufficient. We will study the structure of (15) in order to find a concrete example of a coupling, for which the spectral behaviour is different.

The examination of (15) can be divided into two essentially different situations according to the value of V_4. If $V_4 \neq 0$, then the higer order of k contained in (15) is equal to 4, otherwise it it less or equal to 3. For our purposes it suffices to consider the first case, i.e. $V_4 \neq 0$. In such situation one may divide the whole equation by k^4 and separate the term $\sin ak \sin bk \cdot V_4$ as follows:

$$\begin{aligned}
\sin ak \sin bk \cdot V_4 = {} & \frac{1}{k} \cdot [\cos ak \sin bk \cdot V_3 + \cos bk \sin ak \cdot V_3' + \\
& + \sin bk \cdot W_3 \sin(\theta_1 + \alpha_3) + \sin ak \cdot W_3' \sin(\theta_2 + \beta_3)] + \\
& + \frac{1}{k^2} \cdot [\cos ak \cos bk \cdot V_2 + \sin ak \sin bk \cdot V_2' + \cos bk \cdot W_2 \cos(\theta_1 + \alpha_2) + \\
& + \cos ak \cdot W_2' \cos(\theta_2 + \beta_2) + \tilde{W}_2 \cos(\theta_1 + \theta_2 + \tilde{\alpha}_2) + \tilde{W}_2' \cos(\theta_1 - \theta_2 + \tilde{\beta}_2)\Big] + \\
& + \frac{1}{k^3} \cdot [\cos ak \sin bk \cdot V_1 + \cos bk \sin ak \cdot V_1' + \\
& + \sin bk \cdot W_1 \sin(\theta_1 + \alpha_1) + \sin ak \cdot W_1' \sin(\theta_2 + \beta_1)] + \\
& + \frac{1}{k^4} \sin ak \sin bk \cdot V_0 \,.
\end{aligned} \tag{16}$$

Since all the terms V_j, W_j etc. are constant with respect to k, the RHS is of the order $\mathcal{O}(\frac{1}{k})$, and the same has to hold for the product $\sin ak \sin bk$ at the LHS. Therefore, the bands correspond to either $\sin ak$ small or to $\sin bk$ small. Let us suppose $a > b$. To find a coupling that do not fall within the 1-D classification, we will focus on bands corresponding only to $\sin ak$ small. Let us denote $J_a := \left\{ n \in \mathbb{N} \;\middle|\; |\sin b\tfrac{n\pi}{a}| \geq \tfrac{1}{3} \right\}$ and $A := \bigcup_{n \in J_a} \left(\frac{n\pi}{a} - \frac{\pi}{12b}, \frac{n\pi}{a} + \frac{\pi}{12b} \right)$. Since $a > b$ and $\frac{1}{3} < \frac{\sqrt{2}}{2}$, it holds $|J_a| = \infty$, therefore A is a countable set of equally long intervals. The following inequality will be usefull:
Let $k \in A$, $k \in \left(\frac{n\pi}{a} - \frac{\pi}{12b}, \frac{n\pi}{a} + \frac{\pi}{12b} \right)$, let us put $k = \frac{n\pi}{a} + \frac{\delta}{b}$, $|\delta| \leq \frac{\pi}{12}$. Then

$$|\sin bk| = \left| \sin b\frac{n\pi}{a} \cdot \cos\delta + \cos b\frac{n\pi}{a} \cdot \sin\delta \right| \geq \frac{1}{3} \cdot \cos\frac{\pi}{12} - \sin\frac{\pi}{12} = \frac{2 - \sqrt{3}}{3\sqrt{2}} > 0 \,, \tag{17}$$

i.e. the set $\{ |\sin bk| \mid k \in A \}$ is bounded below by some positive constant.

We will study asymptotical behaviour of solutions of (16) that are contained in the

set A. Let us assemble all the terms of (16) containing $\sin ak$ on the LHS:

$$\begin{aligned}
&\sin ak \left(\sin bk \cdot V_4 - \frac{1}{k} \cos bk \cdot V_3' - \frac{1}{k} \cdot W_3' \sin(\theta_2 + \beta_3) - \frac{1}{k^2} \cdot \sin bk \cdot V_2' - \right. \\
&\qquad \left. - \frac{1}{k^3} \cdot \cos bk \cdot V_1' - \frac{1}{k^3} W_1' \sin(\theta_2 + \beta_1) - \frac{1}{k^4} \sin bk \cdot V_0 \right) = \\
&= \frac{1}{k} \cdot [\cos ak \sin bk \cdot V_3 + \sin bk \cdot W_3 \sin(\theta_1 + \alpha_3)] + \\
&+ \frac{1}{k^2} \cdot [\cos ak \cos bk \cdot V_2 + \cos bk \cdot W_2 \cos(\theta_1 + \alpha_2) + \cos ak \cdot W_2' \cos(\theta_2 + \beta_2) + \\
&\qquad + \tilde{W}_2 \cdot \cos(\theta_1 + \theta_2 + \tilde{\alpha}_2) + \tilde{W}_2' \cdot \cos(\theta_1 - \theta_2 + \tilde{\beta}_2) \Big] + \\
&+ \frac{1}{k^3} \cdot [\cos ak \sin bk \cdot V_1 + \sin bk \cdot W_1 \sin(\theta_1 + \alpha_1)] \, .
\end{aligned} \tag{18}$$

We distinguish two situations:

(a) $W_3 \neq 0$,

(b) $W_3 = 0$.

In the case (a) one can write (18) in the following way:

$$\sin ak \left(\sin bk \cdot V_4 + \mathcal{O}(\frac{1}{k}) \right) = \frac{1}{k} \cdot [\cos ak \sin bk \cdot V_3 + \sin bk \cdot W_3 \sin(\theta_1 + \alpha_3)] + \mathcal{O}(\frac{1}{k^2}) \, . \tag{19}$$

If k is sufficiently big, inequality (17) enables us to divide the whole equation by the expression in the parentheses:

$$\sin ak = \frac{1}{k} \cdot \left[\cos ak \cdot \frac{V_3}{V_4} + \frac{W_3}{V_4} \sin(\theta_1 + \alpha_3) \right] + \mathcal{O}(\frac{1}{k^2}) \, . \tag{20}$$

The expression in the brackets is uniformly bounded with respect to k, therefore the RHS is of the order $\frac{1}{k}$, i.e. $|\sin ak| = \mathcal{O}\left(\frac{1}{k}\right)$. Let k be a solution; for the notation purposes we put

$$ak = n\pi + \delta \quad |\delta| \leq \pi \, ; \tag{21}$$

obviously $\mathcal{O}(k) = \mathcal{O}(n)$ and $\frac{1}{k} = \frac{a}{n\pi + \delta} = \frac{a}{n\pi} + \delta \cdot \mathcal{O}(\frac{1}{n^2})$. Since $\sin ak = (-1)^n \sin \delta$ and $|\sin \delta| \geq \frac{2}{\pi} |\delta|$, we have $\delta \leq \frac{\pi}{2} |\sin ak| = \mathcal{O}\left(\frac{1}{k}\right) = \mathcal{O}\left(\frac{1}{n}\right)$. This allows us to write

$$\begin{aligned}
\sin ak &= (-1)^n \cdot \delta \left(1 + \mathcal{O}(\delta^2)\right) = (-1)^n \cdot \delta \left(1 + \mathcal{O}(\frac{1}{n^2}) \right) \, , \\
\cos ak &= (-1)^n \cdot \left(1 + \mathcal{O}(\delta^2)\right) = (-1)^n \cdot \left(1 + \mathcal{O}(\frac{1}{n^2}) \right) \, , \\
\frac{1}{k} &= \frac{a}{n\pi} + \mathcal{O}(\frac{1}{n^3}) \, .
\end{aligned} \tag{22}$$

Putting all together, we may transform (20) into

$$\delta = \frac{a}{n\pi} \cdot \left(\frac{V_3}{V_4} + \frac{W_3}{V_4} \sin(\theta_1 + \alpha_3) \cdot (-1)^n \right) + \mathcal{O}(\frac{1}{n^2})$$

(we have divided both sides by $(-1)^n$). Finally, when the parameter θ_1 runs through $[-\pi, \pi)$, the term $\sin(\theta_1 + \alpha_3)$ takes all values from $(-1, 1)$. Therefore, δ runs through

$$\left(\frac{a}{n\pi}\cdot\left(\frac{V_3}{V_4}-\left|\frac{W_3}{V_4}\right|\right)+\mathcal{O}(\frac{1}{n^2}), \frac{a}{n\pi}\cdot\left(\frac{V_3}{V_4}+\left|\frac{W_3}{V_4}\right|\right)+\mathcal{O}(\frac{1}{n^2})\right).$$

With respect to (21), the value of k belongs to the set

$$\left(\frac{n\pi}{a}+\frac{1}{n\pi}\cdot\frac{V_3}{V_4}-\frac{1}{n\pi}\cdot\left|\frac{W_3}{V_4}\right|+\mathcal{O}(\frac{1}{n^2}), \frac{n\pi}{a}+\frac{1}{n\pi}\cdot\frac{V_3}{V_4}+\frac{1}{n\pi}\cdot\left|\frac{W_3}{V_4}\right|+\mathcal{O}(\frac{1}{n^2})\right).$$

We immediately see that for sufficiently big n, this interval lies whole in the set A. The easy computation of the the length of the corresponding interval for k^2 gives the result

$$\frac{4}{a}\cdot\left|\frac{W_3}{V_4}\right|+\mathcal{O}(\frac{1}{n^2}),$$

i.e. the we have found an infinite set of assymptotically constant bands.

Consider now the situation (b), i.e. $W_3 = 0$. We will proceed in a similar way, but this time we take into account more terms of (18):

$$\begin{aligned}
&\sin ak\left(\sin bk\cdot V_4-\frac{1}{k}\cos bk\cdot V_3'-\frac{1}{k}\cdot W_3'\sin(\theta_2+\beta_3)+\mathcal{O}(\frac{1}{k^2})\right)=\\
&=\frac{1}{k}\cdot\cos ak\sin bk\cdot V_3+\\
&+\frac{1}{k^2}\cdot[\cos ak\cos bk\cdot V_2+\cos bk\cdot W_2\cos(\theta_1+\alpha_2)+\cos ak\cdot W_2'\cos(\theta_2+\beta_2)+\\
&\qquad+\tilde{W}_2\cdot\cos(\theta_1+\theta_2+\tilde{\alpha}_2)+\tilde{W}_2'\cdot\cos(\theta_1-\theta_2+\tilde{\beta}_2)\Big]+\mathcal{O}(\frac{1}{k^3}).
\end{aligned}$$

For all $k \in A$, $\sin bk$ is uniformly bounded below by a positive constant, thus for sufficiently big k one can divide both sides of the equation by the term standing in the parentheses on the LHS and obtain

$$\begin{aligned}
&\sin ak=\frac{1}{k}\cdot\cos ak\cdot\frac{V_3}{V_4}+\frac{1}{k^2}\cdot\frac{V_3\cos ak}{V_4^2\sin bk}\cdot[\cos bkV_3'+W_3'\sin(\theta_2+\beta_3)]+\\
&+\frac{1}{k^2}\cdot\frac{1}{V_4\sin bk}\cdot[\cos ak\cos bk\cdot V_2+\cos bk\cdot W_2\cos(\theta_1+\alpha_2)+\cos ak\cdot W_2'\cos(\theta_2+\beta_2)+\\
&\qquad+\tilde{W}_2\cdot\cos(\theta_1+\theta_2+\tilde{\alpha}_2)+\tilde{W}_2'\cdot\cos(\theta_1-\theta_2+\tilde{\beta}_2)\Big]+\mathcal{O}(\frac{1}{k^3}).
\end{aligned}$$

Now we use (22) similarly as in the case (a), subsequently we again divide the whole equation by the expression $(-1)^n$, arriving at

$$\begin{aligned}
\delta=&\frac{a}{n\pi}\cdot\frac{V_3}{V_4}+\frac{a^2}{n^2\pi^2}\cdot\frac{1}{V_4\sin bk}\cdot\left[\cos bk\cdot\left(\frac{V_3V_3'}{V_4}+V_2\right)+\right.\\
&+\frac{V_3W_3'}{V_4}\sin(\theta_2+\beta_3)+\cos bk\cdot(-1)^n\cdot W_2\cos(\theta_1+\alpha_2)+W_2'\cos(\theta_2+\beta_2)+\\
&\qquad\left.+\tilde{W}_2\cdot(-1)^n\cos(\theta_1+\theta_2+\tilde{\alpha}_2)+\tilde{W}_2'\cdot(-1)^n\cos(\theta_1-\theta_2+\tilde{\beta}_2)\right]+\mathcal{O}(\frac{1}{n^3}).
\end{aligned}$$

If the parameters θ_1, θ_2 run through the interval $[-\pi, \pi)$, the values of the expression in the bracket form an interval. Since the values are uniformly bounded, as well as $\frac{1}{V_4 \sin bk}$ by virtue of $k \in A$ and (17), there is a $\gamma > 0$ independent of k (and n) such that

$$\delta \in \left(\frac{a}{n\pi} \cdot \frac{V_3}{V_4} - \frac{a^2}{n^2\pi^2}\gamma, \frac{a}{n\pi} \cdot \frac{V_3}{V_4} + \frac{a^2}{n^2\pi^2}\gamma \right) ,$$

and therefore the solutions of (16) form an interval that is contained in

$$\left(\frac{n\pi}{a} + \frac{1}{n\pi} \cdot \frac{V_3}{V_4} - \frac{a}{n^2\pi^2}\gamma, \frac{n\pi}{a} + \frac{1}{n\pi} \cdot \frac{V_3}{V_4} + \frac{a}{n^2\pi^2}\gamma \right) .$$

We observe that if k is sufficiently big, the whole interval lies in the set A. The length of the band, i.e. of the corresponding interval for k^2, is bounded above by the term

$$\left(\frac{n\pi}{a} + \frac{1}{n\pi} \cdot \frac{V_3}{V_4} + \frac{a}{n^2\pi^2}\gamma \right)^2 - \left(\frac{n\pi}{a} + \frac{1}{n\pi} \cdot \frac{V_3}{V_4} - \frac{a}{n^2\pi^2}\gamma \right)^2 = \frac{4\gamma}{n\pi} , \tag{23}$$

thus the bands are neither assymptotically growing, nor assymptotically constant. It is a situation that does not occur in the case of one-dimensional network.

However, to prove that such situation really exists, it is necessary to find an example of a unitary matrix $U \in \mathbb{C}^{n,n}$ such that $V_4 \neq 0$, $W_3 = 0$, and moreover to show that infinitely many of the intervals (23) do not collapse to single points.

Let us consider the following matrix:

$$U = \frac{1}{2} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix} . \tag{24}$$

The unitarity is obvious. A simple calculation gives $V_4 = 32 \neq 0$, $W_3 = 0$. To exclude the collapsing case, we will show that the expression

$$\begin{aligned} &\frac{V_3 W_3'}{V_4} \sin(\theta_2 + \beta_3) + \cos bk \cdot (-1)^n \cdot W_2 \cos(\theta_1 + \alpha_2) + W_2' \cos(\theta_2 + \beta_2) + \\ &+ \tilde{W}_2 \cdot (-1)^n \cos(\theta_1 + \theta_2 + \tilde{\alpha}_2) + \tilde{W}_2' \cdot (-1)^n \cos(\theta_1 - \theta_2 + \tilde{\beta}_2) \end{aligned}$$

is not constant with respect to (θ_1, θ_2). It can be shown that for U given by (24) this expression is equal to

$$-\cos bk \cdot (-1)^n \cdot 32 \cos\theta_1 + 32 \cos\theta_2 - 16 \cdot (-1)^n \cos(\theta_1 + \theta_2) - 16 \cdot (-1)^n \cos(\theta_1 - \theta_2) ,$$

i.e. obviously not constant.

5 Conclusions

We have studied the spectral properties of an infinite periodic network with a rectangular cell. Our results demonstrate that the structure of its spectrum may strongly differ from the case of one-dimensional network, namely that the spectral bands may asymptotically shorten. It would be interesting and useful to find a complete classification of high-energy asymptotics. We aim to study the problem in more detail, believing that the results will be much more complex that in the one-dimensional case.

References

[1] P. Exner, H. Grosse: Some properties of the one-dimensional generalized point interactions (a torso).

[2] P. Exner: Lattice Kronig-Penney models, *Phys. Rev. Lett.* **74** (1995), 3503–3506.

[3] P. Exner: Contact interactions on graph superlattices, *J. Phys. A: Math. Gen.* **29** (1996), 87–102.

[4] P. Kuchment. *Quantum graphs: I. Some basic structures.* Waves Random Media **14**, (2004), S107-S128.

On the Spectrum of a Quantum Dot with Impurity in the Lobachevsky Plane

Matěj Tušek

2nd year of PGS, email: tusekmat@fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Pavel Šťovíček, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. A model of a quantum dot with impurity in the Lobachevsky plane is considered. Relying on explicit formulae for the Green function and the Krein $Q-$function which have been derived in a previous work we focus on the numerical analysis of the spectrum. The analysis is complicated by the fact that the basic formulae are expressed in terms of spheroidal functions with general characteristic exponents. The effect of the curvature on eigenvalues and eigenfunctions is investigated. Moreover, there is given an asymptotic expansion of eigenvalues as the curvature radius tends to infinity (the flat case limit).

Abstrakt. Příspěvěk pojednává o modelu kvantové tečky v Lobačevského rovině. Numerická analýza energetického spektra se opírá o znalost explicitních předpisů pro Greenovu funkci a Kreinovu $Q-$funkci, které byly odvozeny v předchozí práci. Analýza je ztížena výskytem sferoidálních funkcí s obecným charakteristickým exponentem právě v těchto předpisech. Vliv křivosti na vlastní hodnoty a vlastní funkce je podroben zkoumání. Navíc předkládáme asymptotické rozvoje vlastních hodnot pro poloměr křivosti jdoucí k nekonečnu (plochá limita).

1 Introduction

The influence of the hyperbolic geometry on the properties of quantum mechanical systems is a subject of continual theoretical interest for at least two decades. Numerous models have been studied so far, let us mention just few of them [7, 1, 10, 11]. Naturally, the quantum harmonic oscillator is one of the analyzed examples [5, 6]. It should be stressed, however, that the choice of an appropriate potential on the hyperbolic plane is ambiguous in this case, and several possibilities have been proposed in the literature. In [9], we have modeled a quantum dot in the Lobachevsky plane by an unbounded potential which can be interpreted, too, as a harmonic oscillator potential for this nontrivial geometry. The studied examples also comprise point interactions [3] which are frequently used to model impurities.

A Hamiltonian describing a quantum dot with impurity has been introduced in [9]. The main result of this paper is derivation of explicit formulae for the Green function and the Krein $Q-$function. The formulae are expressed in terms of spheroidal functions which are used rather rarely in the framework of mathematical physics. Further analysis is complicated by the complexity of spheroidal functions. In particular, the Green function depends on the characteristic exponent of the spheroidal functions in question rather than directly on the spectral parameter. In fact, it seems to be possible to obtain a

more detailed information on eigenvalues and eigenfunctions only by means of numerical methods. The particular case, when the Hamiltonian is restricted to the eigenspace of the angular momentum with eigenvalue 0, is worked out in [13]. In the current contribution we aim to extend the numerical analysis to the general case and to complete it with additional details.

The Hamiltonian describing a quantum dot with impurity in the Lobachevsky plane, as introduced in [9], is a selfadjoint extension of the following symmetric operator:

$$H = -\left(\frac{\partial^2}{\partial\varrho^2} + \frac{1}{a}\coth\Big(\frac{\varrho}{a}\Big)\frac{\partial}{\partial\varrho} + \frac{1}{a^2}\sinh^{-2}\Big(\frac{\varrho}{a}\Big)\frac{\partial^2}{\partial\phi^2} + \frac{1}{4a^2}\right) + \frac{1}{4}a^2\omega^2\sinh^2\Big(\frac{\varrho}{a}\Big),$$
$$\mathrm{Dom}(H) = C_0^\infty((0,\infty)\times S^1) \subset L^2\big((0,\infty)\times S^1, a\,\sinh(\varrho/a)\mathrm{d}\varrho\,\mathrm{d}\phi\big),$$

where (ϱ,ϕ) are the geodesic polar coordinates on the Lobachevsky plane and a stands for the so called curvature radius which is related to the scalar curvature by the formula $R = -2/a^2$. The deficiency indices of H are known to be $(1,1)$ and we denote each selfadjoint extension by $H(\chi)$ where the real parameter χ appears in the boundary conditions for the domain of definition: $f(\varrho,\phi)$ belongs to $\mathrm{Dom}(H(\chi))$ if there exist $f_0, f_1 \in \mathbb{C}$ so that $f_1 : f_0 = \chi : 1$ and

$$f(\varrho,\phi) = -\frac{1}{2\pi}f_0\log(\varrho) + f_1 + o(1) \quad \text{as } \varrho\to 0+$$

(the case $\chi = \infty$ means that $f_0 = 0$ and f_1 is arbitrary), see [9] for details. $H(\infty)$ is nothing but the Friedrichs extension of H. The Hamiltonian $H(\infty)$ is interpreted as corresponding to the unperturbed case and describing a quantum dot with no impurity.

After the substitution $\xi = \cosh(\varrho/a)$ and the scaling $H = a^{-2}\tilde{H}$, we make use of the rotational symmetry (which amounts to a Fourier transform in the variable ϕ) to decompose $\tilde{H}$ into a direct sum as follows

$$\tilde{H} = \bigoplus_{m=-\infty}^{\infty} \tilde{H}_m,$$
$$\tilde{H}_m = -\frac{\partial}{\partial\xi}(\xi^2-1)\frac{\partial}{\partial\xi} + \frac{m^2}{\xi^2-1} + \frac{a^4\omega^2}{4}(\xi^2-1) - \frac{1}{4},$$
$$\mathrm{Dom}(\tilde{H}_m) = C_0^\infty(1,\infty) \subset L^2((1,\infty),\mathrm{d}\xi).$$

Let us denote by H_m, $m\in\mathbb{Z}$, the restriction of $H(\infty)$ to the eigenspace of the angular momentum with eigenvalue m. This means that H_m is a self-adjoint extension of $a^{-2}\tilde{H}_m$. It is known (Proposition 2.1 in [9]) that $\tilde{H}_m$ is essentially selfadjoint for $m\neq 0$. Thus, in this case, H_m is the closure of $a^{-2}\tilde{H}_m$. Concerning the case $m=0$, H_0 is the Friedrichs extension of $a^{-2}\tilde{H}_0$. For quite general reasons, the spectrum of H_m, for any m, is semibounded below, discrete and simple [14]. We denote the eigenvalues of H_m in ascending order by $E_{n,m}(a^2)$, $n\in\mathbb{N}_0$.

The spectrum of the total Hamiltonian $H(\chi)$, $\chi\neq\infty$, consists of two parts (in a full analogy with the Euclidean case [4]):

1. The first part is formed by those eigenvalues of $H(\chi)$ which belong, at the same time, to the spectrum of $H(\infty)$. More precisely, this part is exactly the union of eigenvalues of H_m for m running over $\mathbb{Z}\setminus\{0\}$. Their multiplicities are discussed below in Section 5.

2. The second part is formed by solutions to the equation

$$Q^H(z) = \chi \tag{1}$$

with respect to the variable z where Q^H stands for the Krein Q-function of $H(\infty)$. Let us denote the solutions in ascending order by $\epsilon_n(a^2, \chi)$, $n \in \mathbb{N}_0$. These eigenvalues are sometimes called the point levels and their multiplicities are at least one. In more detail, $\epsilon_n(a^2, \chi)$ is a simple eigenvalue of $H(\chi)$ if it does not lie in the spectrum of $H(\infty)$, and this happens if and only if $\epsilon_n(a^2, \chi)$ does not coincide with any eigenvalue $E_{\ell,m}(a^2)$ for $\ell \in \mathbb{N}_0$ and $m \in \mathbb{Z}$, $m \neq 0$.

Remark. The lowest point level, $\epsilon_0(a^2, \chi)$, lies below the lowest eigenvalue of $H(\infty)$ which is $E_{0,0}(a^2)$, and the point levels with higher indices satisfy the inequalities $E_{n-1,0}(a^2) < \epsilon_n(a^2, \chi) < E_{n,0}(a^2)$, $n = 1, 2, 3, \ldots$.

2 Spectrum of the unperturbed Hamiltonian $H(\infty)$

Our goal is to find the eigenvalues of the mth partial Hamiltonian H_m, i.e., to find square integrable solutions of the equation

$$H_m\psi(\xi) = z\psi(\xi),$$

or, equivalently,

$$\tilde{H}_m\psi(\xi) = a^2 z\psi(\xi).$$

This equation coincides with the equation of the spheroidal functions (A.1) provided we set $\mu = |m|$, $\theta = -a^4\omega^2/16$, and the characteristic exponent ν is chosen so that

$$\lambda_\nu^m\left(-\frac{a^4\omega^2}{16}\right) = -a^2 z - \frac{1}{4}\,.$$

The only solution (up to a multiplicative constant) that is square integrable near infinity is $S_\nu^{|m|(3)}(\xi, -a^4\omega^2/16)$.

Proposition 3 describes the asymptotic expansion of this function at $\xi = 1$ for $m \in \mathbb{N}$. It follows that the condition on the square integrability is equivalent to the equality

$$\mathrm{e}^{i(3\nu+1/2)\pi} K_{-\nu-1}^m\left(-\frac{a^4\omega^2}{16}\right) + K_\nu^m\left(-\frac{a^4\omega^2}{16}\right) = 0. \tag{2}$$

Furthermore, in [9] we have derived that

$$S_\nu^{0(3)}(\xi, \theta) = \alpha \log(\xi - 1) + \beta + O((\xi - 1)\log(\xi - 1)) \quad \text{as } \xi \to 1+,$$

where

$$\alpha = \frac{i\tan(\nu\pi)\,\mathrm{e}^{-i(2\nu+1/2)\pi}}{2\pi s_\nu^0(\theta)}\left(\mathrm{e}^{i(3\nu+1/2)\pi} K_{-\nu-1}^0(\theta) + K_\nu^0(\theta)\right).$$

Taking into account that the Friedrichs extension has continuous eigenfunctions we conclude that equation (2) guarantees square integrability in the case $m = 0$, too.

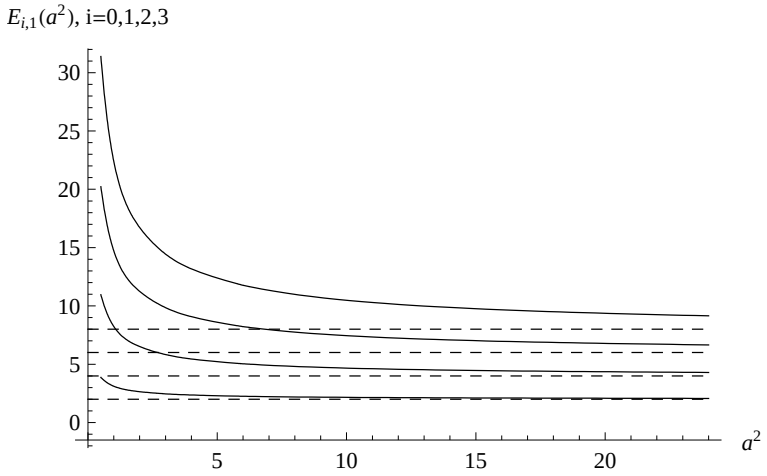


Figure 1: Eigenvalues of the partial Hamiltonian H_1

As far as we see it, equation (2) can be solved only by means of numerical methods. For this purpose we made use of the computer algebra system *Mathematica 6.0*. For the numerical computations we set $\omega = 1$. The particular case $m = 0$ has been examined in [13]. It turns out that an analogous procedure can be also applied for nonzero values of the angular momentum. As an illustration, Figure 1 depicts several first eigenvalues of the Hamiltonian H_1 as functions of the curvature radius a. The dashed asymptotic lines correspond to the flat limit ($a \to \infty$).

Denote the nth normalized eigenfunction of the mth partial Hamiltonian $\tilde{H}_m$ by $\tilde{\psi}_{n,m}(\xi)$. Obviously, the eigenfunctions for the values of the angular momentum m and $-m$ are the same and are proportional to $S_\nu^{|m|(3)}(\xi, -a^4\omega^2/16)$, with ν satisfying equation (2). Let us return to the original radial variable ϱ and, moreover, regard $\tilde{H}_m$ as an operator acting on $L^2(\mathbb{R}^+, \mathrm{d}\varrho)$. This amounts to an obvious isometry

$$L^2(\mathbb{R}^+, a^{-1}\sinh(\varrho/a)\mathrm{d}\varrho) \to L^2(\mathbb{R}^+, \mathrm{d}\varrho) : \ f(\varrho) \mapsto a^{-1/2}\sinh^{1/2}(\varrho/a) f(\varrho).$$

The corresponding normalized eigenfunction of $\tilde{H}_m$, with an eigenvalue $a^2 z$, equals

$$\psi_{n,m}(\varrho) = \left(\frac{1}{a}\sinh\left(\frac{\varrho}{a}\right)\right)^{1/2} \tilde{\psi}_{n,m}\left(\cosh\left(\frac{\varrho}{a}\right)\right). \tag{3}$$

At the same time, relation (3) gives the normalized eigenfunction of H_m (considered on $L^2(\mathbb{R}^+, \mathrm{d}\varrho)$) with the eigenvalue z. The same Hilbert space may be used also in the limit Euclidean case ($a = \infty$). The eigenfunctions $\Phi_{n,m}$ in the flat case are well known and satisfy

$$\Phi_{n,m} \propto \varrho^{|m|+1/2} \mathrm{e}^{-\omega\varrho^2/4}\, {}_1F_1\left(-n, |m|+1, \frac{\omega\varrho^2}{2}\right). \tag{4}$$

The fact that we stick to the same Hilbert space in all cases facilitates the comparison of eigenfunctions for various values of the curvature radius a. We present plots of several first eigenfunctions of H_1 (Figures 2, 3, 4) for the values of the curvature radius $a = 1$ (the

solid line), 10 (the dashed line), and ∞ (the dotted line). Again, see [13] for analogous plots in the case of the Hamiltonian H_0. Note that, in general, the smaller is the curvature radius a the more localized is the particle in the region near the origin.

3 The point levels

As has been stated, the point levels are solutions to equation (1) with respect to the spectral parameter z. Since, in general, $Q(\bar{z}) = \overline{Q(z)}$ the function $Q(z)$ takes real values on the real axis. Let $\tilde{H}(\infty) = a^2 H(\infty)$ be the Friedrichs extension of $\tilde{H}$. An explicit formula for the Krein Q-function $Q^{\tilde{H}}(z)$ of $\tilde{H}(\infty)$ has been derived in [9]:

$$Q^{\tilde{H}}(z) = -\frac{1}{4\pi a^2}\left(-\log(2) - 2\Psi(1) + 2\,\Psi s_\nu\left(-\frac{a^4\omega^2}{16}\right) s^0_\nu\left(-\frac{a^4\omega^2}{16}\right)\right) + \frac{1}{2a^2\tan(\nu\pi)}\left(\mathrm{e}^{i\pi(3\nu+3/2)}\,\frac{K^0_{-\nu-1}(-\frac{a^4\omega^2}{16})}{K^0_\nu(-\frac{a^4\omega^2}{16})} - 1\right)^{-1} + \frac{\log(2a^2)}{4\pi a^2},$$

where ν is chosen so that

$$\lambda^0_\nu\left(-\frac{a^4\omega^2}{16}\right) = -z - \frac{1}{4}\,.$$

The symbol $K^0_\nu(\theta)$ stands for the so called spheroidal joining factor,

$$\Psi s_\nu(\theta) := \sum_{r=-\infty}^{\infty} (-1)^r a^0_{\nu,r}(\theta)\,\Psi(\nu+1+2r),$$

where the coefficients $a^0_{\nu,r}(\theta)$, $r \in \mathbb{Z}$, come from the expansion of spheroidal functions in terms of Bessel functions (for details see [9, the Appendix])), and $s^0_\nu(\theta)$ is defined by the formula

$$(s^m_\nu(\theta))^{-1} := \sum_{r=-\infty}^{\infty} (-1)^r a^m_{\nu,r}.$$

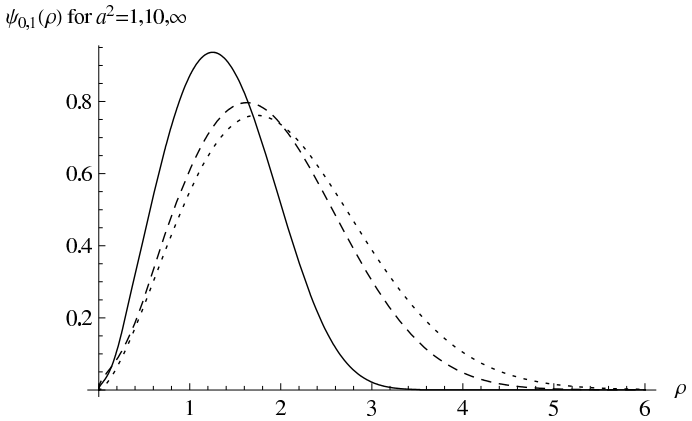


Figure 2: The first eigenfunction of the partial Hamiltonian H_1

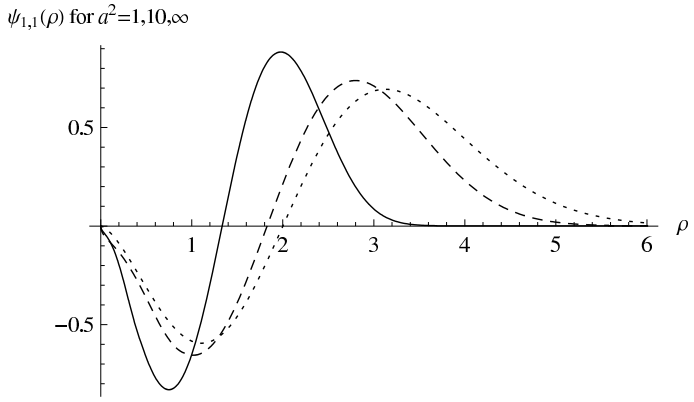


Figure 3: The second eigenfunction of the partial Hamiltonian H_1

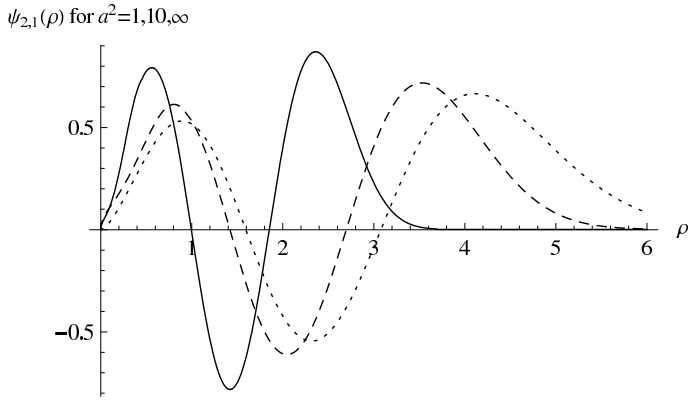


Figure 4: The third eigenfunction of the partial Hamiltonian H_1

One can obtain the Krein Q-function of $H(\infty)$ simply by scaling $Q^H(z) = a^2\, Q^{\tilde{H}}(a^2 z)$.

Since we know the explicit expression for the Krein Q-function as a function of the characteristic exponent ν rather than of the spectral parameter z itself it is of importance to know for which values of ν the spectral parameter z is real. Propositions 1 and 2 give the answer. For $\nu \in \mathbb{R}$ and for ν of the form $\nu = -1/2 + it$ where t is real, the spheroidal eigenvalue $\lambda_\nu^m(-a^4\omega^2/16)$ is real, and so the same is true for z. Moreover, these values of ν reproduce the whole real z axis. With this knowledge, one can plot the Krein Q-function $Q^H = Q^H(z)$ for an arbitrary value of the curvature radius a. Note that for $a = \infty$, the Krein Q-function is well known as a function of the spectral parameter z [8] and equals (setting $\omega = 1$, Ψ is the logarithmic derivative of the gamma function)

$$Q(z) = \frac{1}{4\pi}\left(-\Psi\left(\frac{1-z}{2}\right) + \log(2) + 2\Psi(1)\right).$$

Again, equation (1) can be solved only numerically. Fixing the parameter χ one may be interested in the behavior of the point levels as functions of the curvature radius a. See Figure 5 for the corresponding plots, with $\chi = 0$, where the dashed asymptotic lines again correspond to the flat case limit ($a = \infty$). Note that for the curvature radius a large

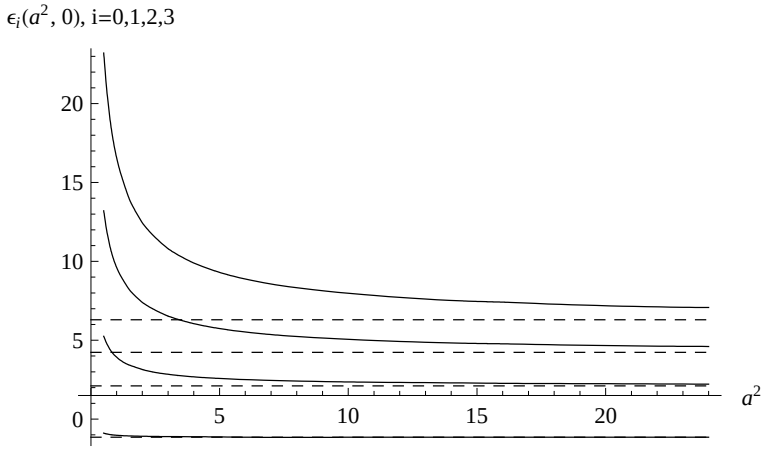


Figure 5: Point levels for H(0)

enough, the lowest eigenvalue is negative provided χ is chosen smaller than $Q(0) \simeq 0.1195$.

4 Asymptotic behavior for large values of a

The mth partial Hamiltonian H_m, if considered on $L^2(\mathbb{R}^+, \mathrm{d}\varrho)$, acts like

$$H_m = -\frac{\partial^2}{\partial\varrho^2} + \frac{m^2 - \frac{1}{4}}{a^2 \sinh^2(\frac{\varrho}{a})} + \frac{1}{4}\, a^2\omega^2 \sinh^2\Big(\frac{\varrho}{a}\Big) =: -\frac{\partial^2}{\partial\varrho^2} + V_m(a, \varrho).$$

For a fixed $\varrho \neq 0$, one can easily derive that

$$V_m(a, \varrho) = \frac{m^2 - \frac{1}{4}}{\varrho^2} + \frac{1}{4}\omega^2\varrho^2 + \frac{\frac{1}{4} - m^2}{3a^2} + \frac{\omega^2\varrho^4}{12a^2} + O\Big(\frac{1}{a^4}\Big) \quad \text{as } a \to \infty.$$

Recall that the mth partial Hamiltonian of the isotropic harmonic oscillator on the Euclidean plane, H_m^E, if considered on $L^2(\mathbb{R}^+, \mathrm{d}\varrho)$, has the form

$$H_m^E := -\frac{\partial^2}{\partial\varrho^2} + \frac{m^2 - \frac{1}{4}}{\varrho^2} + \frac{1}{4}\omega^2\varrho^2.$$

This suggests that it may be useful to view the Hamiltonian H_m, for large values of the curvature radius a, as a perturbation of H_m^E,

$$H_m \sim H_m^E + \frac{1}{12a^2}(1 - 4m^2 + \omega^2\varrho^4) =: H_m^E + \frac{1}{12a^2}U_m(\varrho).$$

The eigenvalues of the compared Hamiltonians have the same asymptotic expansions up to the order $1/a^2$ as $a \to \infty$.

Let us denote the nth eigenvalue of the Hamiltonian H_m^E by $E_{n,m}^E$, $n \in \mathbb{N}_0$. It is well known that

$$E_{n,m}^E = (2n + |m| + 1)\,\omega$$

Table 1: Comparison of numerical and asymptotic results for the eigenvalues, $a^2 = 24$

	$E_{0,0}$	$E_{1,0}$	$E_{2,0}$	$E_{0,1}$	$E_{1,1}$	$E_{2,1}$
numerical	1.0265	3.162	5.42	2.060	4.259	6.58
asymptotic	1.0268	3.169	5.46	2.058	4.258	6.59
error (%)	-0.03	-0.22	-0.74	0.10	0.02	-0.15

and that the multiplicity of $E^E_{n,m}$ in the spectrum of H^E equals $2n + |m| + 1$. The asymptotic behavior of $E_{n,m}(a^2)$ may be deduced from the standard perturbation theory and is given by the formula

$$E_{n,m}(a^2) = E^E_{n,m} + \frac{1}{12a^2}\frac{\langle \Phi_{n,m}, U_m \Phi_{n,m}\rangle}{\langle \Phi_{n,m}, \Phi_{n,m}\rangle} + O\left(\frac{1}{a^4}\right) \quad \text{as } a \to \infty, \tag{5}$$

where $\Phi_{n,m}$ denotes a (not necessarily normalized) eigenfunction of H^E_m associated with the eigenvalue $E^E_{n,m}$ (see (4)). The scalar products occurring in formula (5) can be readily evaluated in $L^2(\mathbb{R}^+, \mathrm{d}\varrho)$ with the help of Proposition 4. The resulting formula takes the form

$$E_{n,m}(a^2) = (2n + |m| + 1)\,\omega + \left(2n(n + |m| + 1) + |m| + \frac{3}{4}\right)\frac{1}{a^2} + O\left(\frac{1}{a^4}\right) \tag{6}$$

as $a \to \infty$. This asymptotic approximation of eigenvalues has been tested numerically for large values of the curvature radius a. The asymptotic eigenvalues for $a^2 = 24$ are compared with the precise numerical results in Table 1. It is of interest to note that the asymptotic coefficient in front of the a^{-2} term does not depend on the frequency ω.

5 The multiplicities

Since $H_{-m} = H_m$ the eigenvalues $E_{n,m}(a^2)$ of the total Hamiltonian $H(\infty)$ are at least twice degenerated if $m \neq 0$. From the asymptotic expansion (6) it follows, after some straightforward algebra, that no additional degeneracy occurs and thus theses eigenvalues are exactly twice degenerated at least for sufficiently large values of a.

Applying the methods developed in [4] one may complete the analysis of the spectrum of the total Hamiltonian $H(\chi)$ for $\chi \neq \infty$. Namely, the spectrum of $H(\chi)$ contains eigenvalues $E_{n,m}(a^2)$, $m > 0$, with multiplicity 2 if $Q^H(E_{n,m}(a^2)) \neq \chi$, and with multiplicity 3 if $Q^H(E_{n,m}(a^2)) = \chi$. The rest of the spectrum of $H(\chi)$ is formed by those solutions to equation (1) which do not belong to the spectrum of $H(\infty)$. The multiplicity of all these eigenvalues in the spectrum of $H(\chi)$ equals 1.

Appendix: Auxiliary results

In this appendix we summarize several auxiliary results. For the page limit we omit the proofs. Firstly, for our purposes we need the following observations concerning spheroidal

functions. The spheroidal functions are solutions to the equation

$$(1-\xi^2)\frac{\partial^2\psi}{\partial\xi^2} - 2\xi\frac{\partial\psi}{\partial\xi} + \left[\lambda_\nu^\mu(\theta) + 4\theta(1-\xi^2) - \mu^2(1-\xi^2)^{-1}\right]\psi = 0. \tag{A.1}$$

For the notation and properties of spheroidal functions see [2]. A detailed information on this subject can be found in [12], but be aware of somewhat different notation. A very brief overview of spheroidal functions is also given in the Appendix of [9].

In the last named source, the following proposition has been proved in the particular case $m = 0$. But, as one can verify by a direct inspection, the proof applies to the general case $m \in \mathbb{Z}$ as well.

Proposition 1. *Let $\nu, \theta \in \mathbb{R}$, $m \in \mathbb{Z}$. Then $\lambda_\nu^m(\theta) \in \mathbb{R}$.*

The following claim is also of interest.

Proposition 2. *Let $\nu = -1/2 + it$ where $t \in \mathbb{R}$, and $\theta \in \mathbb{R}$, $m \in \mathbb{Z}$. Then $\lambda_\nu^m(\theta) \in \mathbb{R}$.*

Another auxiliary result concerns the asymptotic expansion of the radial spheroidal function of the third kind.

Proposition 3. *Let $\nu \notin \{-1/2 + k |\ k \in \mathbb{Z}\}$, $m \in \mathbb{N}$. Then*

$$S_\nu^{m(3)}(\xi,\theta) \sim \frac{(-1)^m 2^{m/2-1}\Gamma(m)\tan(\nu\pi)}{\pi s_\nu^m(\theta)\,\mathrm{e}^{-i(\nu+3/2)\pi}}\left(K_{-\nu-1}^m(\theta) + \frac{K_\nu^m(\theta)}{\mathrm{e}^{i(3\nu+1/2)\pi}}\right)(\xi-1)^{-m/2}$$
$$\text{as } \xi \to 1+. \tag{A.2}$$

Further some auxiliary computations follow that we need for evaluation of scalar products of eigenfunctions (see (5)).

Proposition 4. *Let ${}_1F_1(a,b,t)$ stand for the Kummer confluent hypergeometric function, and $n, m, l \in \mathbb{N}_0$. Then*

$$\begin{aligned}&\int_0^\infty t^{m+l}\mathrm{e}^{-t}\,{}_1F_1(-n,1+m,t)^2\,\mathrm{d}t \\ &= (m!)^2 \sum_{k=\max\{0,n-l\}}^{n} (-1)^{n+k}\binom{n}{k}\frac{(k+l)!}{(k+m)!}\binom{k+m+l}{n+m}.\end{aligned} \tag{A.3}$$

Corollary 5. *In the case $l = 0$, (A.3) takes a particularly simple form:*

$$\int_0^\infty t^m\mathrm{e}^{-t}\,{}_1F_1(-n,1+m,t)^2\,\mathrm{d}t = \frac{n!}{(m+n)!}\,.$$

Acknowledgments

The authors wish to acknowledge gratefully partial support of the Ministry of Education of Czech Republic under the research plan MSM6840770039 (P.Š.) and from the grant No. LC06002 (M.T.).

References

[1] M. Antoine, A. Comtet, and S. Ouvry. *Scattering on a Hyperbolic Torus in a Constant Magnetic Field*. J. Phys. A: Math. Gen. **23** (1990), 3699-3710.

[2] H. Bateman and A. Erdélyi. *Higher Transcendental Functions III*. McGraw-Hill Book Company, 1955.

[3] J. Brüning and V. Geyler. *Gauge-Periodic Point Perturbations on the Lobachevsky Plane*. Theor. Math. Phys. **119** (1999), 687-697.

[4] J. Brüning, V. Geyler, and I. Lobanov. *Spectral Properties of a Short-Range Impurity in a Quantum Dot*. J. Math. Phys. **46** (2004), 1267-1290.

[5] D. V. Bulaev, V. A. Geyler, and V. A. Margulis. *Effect of Surface Curvature on Magnetic Moment and Persistent Currents in the Two-Dimensional Quantum Ring and Dots*. Phys. Rev. B **69** (2004), 195313.

[6] J. F. Cariñena, M. F. Rañada, and M. Santander. *The Quantum Harmonic Oscillator on the Sphere and the Hyperbolic Plane*. Ann. Physics **322** (2007), 2249-2278.

[7] A. Comtet. *On the Landau Levels on the Hyperbolic Plane*. Ann. Physics **173** (1987), 185-209.

[8] V. Geyler and I. Popov. *Eigenvalues Imbedded in the Band Spectrum for a Periodic Array of Quantum Dots*. Rep. Math. Phys. **39** (1997), 275-281.

[9] V. Geyler, P. Šťovíček, and M. Tušek. *A Quantum Dot with Impurity in the Lobachevsky Plane*. Operator Theory: Advances and Applications **188** (2008), 143-156.

[10] Yu. A. Kuperin, R. V. Romanov, and H. E. Rudin. *Scattering on the Hyperbolic Plane in the Aharonov-Bohm Gauge Field*. Lett. Math. Phys. **31** (1994), 271-278.

[11] O. Lisovyy. *Aharonov-Bohm Effect on the Poincaré Disk*. J. Math. Phys. **48** (2007), 052112.

[12] J. Meixner and F.V. Schäfke. *Mathieusche Funktionen und Sphäroidfunktionen*. Springer-Verlag, 1954.

[13] P. Šťovíček and M. Tušek. *On the Harmonic Oscillator on the Lobachevsky Plane*. Russian J. Math. Phys. **14** (2007), 493-497.

[14] J. Weidmann. *Linear Operators in Hilbert Spaces*. Springer, 1980.

Strategy Design for Futures Trading*

Jan Zeman

2nd year of PGS, email: janzeman3@seznam.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Tatiana Valentine Guy, Institute of Information Theory and Automation, AS CR

Abstract. The paper addresses the design of trading strategy for futures markets. The problem is formulated as dynamic decision making task and as such is solved. Iterations-spread-in-time and Monte Carlo methods are employed to the solution. The results of off-line real-data experiments are presented.

Abstrakt. Text popisuje návrh obchodní strategie určené pro trhy s futures kontrakty. Návrh se sestává z definice úlohy jako problému dynamického rozhodování. Následně je úloha řešena pomocí iterací rozložených v čase a metody Monte Carlo. Text obsahuje výsledky experimentů prováděných na reálných datech.

1 Introduction

The paper describes a part of research aiming to design automatic trading system for futures markets. The trading on exchanges is based on knowledge and prediction of the price of given commodity, which represents a very complex task.

The futures trading problem is formulated as a particular decision making (DM) task. DM reformulates the task as mathematical problem, which leads to integral equations. We need to solve the equations, but to find the analytical solution is almost impossible and the numerical calculation leads to bad conditioned or long calculated solutions. DM task is necessary to solve in given time, e.g. when the trader on exchange needs the solution each day, the calculation cannot take 3 days and is restricted by 24 hours. Although the reformulation like a DM task is good, we need feasible solution, which calls for an approximation. This paper considers by task redefinition and introduces the approximations.

The paper's outline is as follows. Section 2 introduces terminology of futures exchange, recalls main terms of DM theory and reformulates futures trading problem as dynamic DM task. Section 3 contains approximation of DM. Section 4 presents the experimental results obtained on real data. Section 5 addresses open questions as well as possible directions to approach's improvement.

*This work has been supported by the grants MŠMT 2C06001 and GA ČR 102/08/0567.

2 Preliminaries

2.1 Trading futures

The following definition by of the futures exchange is proposed by [2]. A *futures exchange* is a central financial exchange where people can trade standardized futures contracts; that is, a contract to buy specific quantities of a commodity (basic resources and agricultural products such as iron ore, coal, sugar, coffee beans, wheat, gold, etc) or financial instrument (cash, evidence of an ownership interest in an entity) at a specified price with delivery set at a specified time in the future. A futures contract gives the holder the obligation to buy or sell.

The term position means a commitment to buy or sell a given amount of commodities. The basic types of position are distinguished: *short*, *long* and *flat*.

A *long position* yields a trader's benefit when the price increases, and trader's loss otherwise. This position refers to the situation when

- a trader buys an option contract that he has not already written (i.e. sold), he is said to be opening a long position.
- a trader sells an option contract that he already owns, he is said to be closing a long position.

A *short position* yields a trader's profit when the price decreases, and trader's loss otherwise. This position refers to the situation when

- a trader sells an option contract that he does not already own, he is said to be opening a short position.
- a trader buys an option contract that he has written (i.e. sold), he is said to be closing a short position.

A *flat position* denotes the state when no other type of position is active. Flat position means neither trader's profit nor trader's lose with any price change.

The aim of trader is design such a strategy of positions selecting, which ensures trader's profit with minimal risk. The strategy design is based on prediction of price behavior and is very sensitive i.e. the small impreciseness in strategy make big change of profit.

2.2 Decision making under uncertainty

Decision maker is either human being or device aiming to influence a part of the World he is interested in (so called *System*) The influence desired is expressed by DM aim. To reach this DM aim a decision maker designs and applies a *DM strategy*, R_t. This strategy maps observations of the system's behavior $y_1, \ldots, y_t$ available to decision maker and past decisions $x_1, \ldots x_{t-1}$ to *decisions* x_t:

$$R_t : [y_1, \ldots, y_t, x_1, \ldots x_{t-1}] \rightarrow x_t.$$

The available knowledge grows with time, because it is extended each time step by new system output y_t and also by new decision x_t. The decision typically influences the system, therefore decision maker works with respect to closed loop 'decision maker - system'.

All knowledge about system available to decision maker to design decision x_t is called *experience* $\mathcal{P}_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1}, y_t)$. *Ignorance* $\mathcal{F}_t$ is knowledge about system unavailable to decision maker. *System behavior* consists of experience, decision and ignorance $\mathcal{Q} = (\mathcal{P}_t, x_t, \mathcal{F}_t)$.

Gain is mapping of system behavior to real non-negative number $G : \mathcal{Q} \to [0, \infty]$. Gain express the success of reaching the decision maker aims with given decision making strategy. The gain is not causal and it is necessary to measure the potential strategy success. Therefore the expected value is defined. Conditioned *expected value* $\mathcal{E}(.|.)$ is functional which returns the value of the gain independent on ignorance for the given strategy and conditioned by experience.

The expected gain conditioned by experience is chosen as following integral:

$$\mathcal{E}\big[G(\mathcal{P}_t, x_t, \mathcal{F}_t)\big|\mathcal{P}_t, x_t\big] = \int_{\mathcal{F}_t} G(\mathcal{P}_t, x_t, \mathcal{F}_t) f(\mathcal{F}_t|\mathcal{P}_t, x_t) d\mathcal{F}_t, \tag{1}$$

where $f(\mathcal{F}_t|\mathcal{P}_t, x_t)$ is probability density function of the ignorance conditioned by experience, this terms stands for the decision makers imagination of the ignorance based on experience. See [3] for general derivation of this equation.

The decision maker chose the decision $x_t \in \mathcal{X}$ to maximize of expected value in each time t:

$$x_t = \arg\max_{x_t \in \mathcal{X}} \mathcal{E}\big[G(\mathcal{Q})\big|\mathcal{P}_t, x_t\big], \tag{2}$$

which is the idea based on principle of optimality - see [1].

2.3 Futures trades as DM task

This subsection reformulates futures trading task as a decision making problem.

The system is exchange with one kind of futures contract. The system output y_t is a price of the contract. We design the strategy for discrete time starting from 1, finishing by horizon T. The strategy starts and finishes with the flat position.

The decision maker designs in each time t an integer number $x_t \in \mathcal{Z}$ as decision. The decision x_t characterizes traders position, i.e. $|x_T|$ characterizes count of contracts and $\text{sign}(x_T)$ characterizes the type of position 1 long, -1 short and 0 flat. The flat position at the beginning and at the horizon is expresses as: $x_0 = x_T = 0$.

The profit influenced only by the decision x_t is expressed via:

$$g(x_t, x_{t-1}, y_{t+1}, y_t) = (y_{t+1} - y_t)x_t - C|x_{t-1} - x_t|, \tag{3}$$

where $(y_{t+1} - y_t)x_t$ is profit caused by the change of price and C is normalized transaction costs for position change and $|x_{t-1} - x_t|$ is change of position. The gain from the whole trading can be expressed as a sum of partial gain (3) over time $t \in \{1, 2, \ldots, T\}$. The gain function $G_t(.)$ expresses the profit caused by decisions $x_t, \ldots, x_T$:

$$G_t(x_{t-1}, \ldots, x_T, y_t, \ldots, y_T) = -C|x_{T-1} - x_T| + \sum_{k=t}^{T-1} (y_{k+1} - y_k)x_k - C|x_{k-1} - x_k|, \tag{4}$$

Easy to see, that the function $G_t(.)$ is additive and backward recursive

$$G_t(x_{t-1}, \ldots, x_T, y_t, \ldots, y_T) = g(x_t, x_{t-1}, y_{t+1}, y_t) + G_{t+1}(x_t, \ldots, x_T, y_{t+1}, \ldots, y_T) \tag{5}$$

with initial condition

$$G_T(x_{T-1}, x_T, y_T) = -C|x_{T-1} - x_T|. \tag{6}$$

2.4 Solution of dynamic DM problem

To maximize the profit, the gain over the decisions $x_1, \ldots, x_T$ should be maximized:

$$\max_{\{x_1,\ldots,x_T\}} G_1(x_0, \ldots, x_T, y_1, \ldots, y_T). \tag{7}$$

Using the optimality principle (see [1] for details) and additivity of the gain function the optimal gain in time t can be expressed:

$$B_t(x_{t-1}, \ldots, x_T, y_t, \ldots, y_T) = \max_{x_t} \Big[g(x_{t-1}, x_t, y_t, y_{t+1}) + \max_{\{x_{t+1},\ldots,x_t\}} G_{t+1}(x_t, \ldots, x_T, y_t, \ldots, y_T) \Big].$$

Function $B_t(.)$ is called Bellman's function and hold the following recursive shape:

$$B_t(x_{t-1}, \ldots, x_T, y_t, \ldots, y_T) = \max_{x_t} \Big[g(x_{t-1}, x_t, y_t, y_{t+1}) + B_{t+1}(x_t, \ldots, x_T, y_t, \ldots, y_T) \Big],$$

where the maximal argument is the optimal decision at time t. But to find this argument, the knowledge of future decisions and prices is needed, i.e. $x_{t+1}, \ldots, x_T, y_t, \ldots, y_T$. These variables are the part of ignorance, therefore the expected value must be used:

$$\mathcal{V}_t(x_{t-1}, y_t) = \max_{x_t} \mathcal{E}\Big[g(x_{t-1}, x_t, y_t, y_{t+1}) + \mathcal{V}_{t+1}(x_t, y_{t+1}) \Big| x_0, \ldots, x_t, y_1, \ldots, y_t \Big], \tag{8}$$

where $\mathcal{V}_t(.)$ is called admissible Bellman's function.

3 Approximation of decision making

The substitution (3) into the equation (8) results in more suitable form:

$$\begin{aligned} \mathcal{V}_t(x_{t-1}, y_t) \;=\; & \max_{x_k} \Big[-y_t x_t - C|x_{t-1} - x_t| + x_t \underbrace{\mathcal{E}(y_{t+1}|x_0, \ldots, x_t, y_1, \ldots, y_t)}_{(*)} \\ & + \underbrace{\mathcal{E}\Big(\mathcal{V}_{t+1}(x_t, y_{t+1}) \Big| x_0, \ldots, x_t, y_1, \ldots, y_t \Big)}_{(**)} \Big]. \end{aligned} \tag{9}$$

This paragraph concerns expressing the term $(*)$, which characterizes expected value of future price y_{k+1} conditioned by the experience.

The probability density function $f(y_{k+1}|x_0, \ldots, x_t, y_1, \ldots, y_t)$ is required to express the expected value $(*)$. The probability density function can be written in the parameterized form:

$$f(y_{t+1}|x_0, \ldots, x_t, y_1, \ldots, y_t) = \int_\theta f(y_{t+1}|\theta, x_0, \ldots, x_t, y_1, \ldots, y_t) f(\theta|x_0, \ldots, x_t, y_1, \ldots, y_t) d\theta \tag{10}$$

The last expression consists of two density functions: $f(\theta|x_0,\ldots,x_t,y_1,\ldots,y_t)$ is the density of model parameters conditioned by experience, where θ is vector of the parameters. $f(y_{t+1}|\theta,x_0,\ldots,x_t,y_1,\ldots,y_t)$ is density of price y_{t+1} conditioned by model parameters and experience.

The assumed model is autoregressive and has following shape:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_N y_{t-N} + b + e_t, \tag{11}$$

where $\theta = (a_1,\ldots,a_N,b)$ are model parameters, N denotes model's order and e_t is white noise with distribution $N(0,\sigma^2)$, therefore the model prediction is normally distributed:

$$f(y_{t+1}|\theta,x_0,\ldots,x_t,y_1,\ldots,y_t) = N(a_1 y_t + a_2 y_{t-1} + \ldots + a_N y_{t-N+1} + b, \sigma^2). \tag{12}$$

The density function of model parameters $f(\theta|x_0,\ldots,x_t,y_1,\ldots,y_t)$ is estimated using software MIXTOOLS [4], which works with the distribution $f(\theta|x_0,\ldots,x_t,y_1,\ldots,y_t)$ and generates samples of model parameters.

This scheme corresponds with principles of Monte Carlo method and the expected value of the future price can be calculated using the following formula:

$$\hat{y}_{k+1} = \sum_{i\in S}(a_{1,i}y_k + a_{2,i}y_{k-1} + \ldots + a_{N,i}y_{k-N+1} + b_i)p_i, \tag{13}$$

where S is a set of samples, i is an index of sample, $(a_{1,i},\ldots,a_{N,i},b_i)$ is a sample vector and p_i is probability of the sample i.

Let approximate the term $(**)$ of the equation (9). The main problem of calculating the term is backward character of equation (8), where the future value of Bellman's function $\mathcal{V}_{t+1}(.)$ is needed to calculation the $\mathcal{V}_t(.)$. This problem is solvable two ways: expressing the generalized shape of Bellman's function or approximation by suitable shape.

We need to find formal solution of equation (9) to express the generalized shape of Bellman's function. The desired solution must be valid for all sequences $y_1,\ldots,y_T$. However this task is very complex and it seems impossible to find the formal solution.

The approximation of Bellman's function is more promising way. The approximation must be suitable for further computing, but at the same time contains the parameters of Bellman's function, therefore the following shape has been chosen:

$$\mathcal{V}_t(x_{t-1},y_t) \approx V_t(x_{t-1},y_t) \equiv p(x_{t-1})y_t + q(x_{t-1}), \tag{14}$$

where $p(.)$ and $q(.)$ are real functions. The approximation does not depend on ignorance, therefore the expected value in term $(**)$ is expressed as follows:

$$\mathcal{E}\Big(\mathcal{V}_{k+1}(x_k,y_{k+1})\Big|x_0,\ldots,x_t,y_1,\ldots,y_t\Big) \approx V_{k+1}(x_k,y_{k+1}). \tag{15}$$

The tasks is to design algorithm how to find functions $p(.)$ and $q(.)$ in definition (14). The approximation generates a non-preciseness in equation (8):

$$V_k(x_{k-1},y_k) + e_k = \max_{x_k}\mathcal{E}\Big[g(x_k,x_{k-1},y_{k+1},y_k) + V_{k+1}(x_k,y_{k+1})\Big|x_0,\ldots,x_t,y_1,\ldots,y_t Big], \tag{16}$$

where e_k is introduced non-preciseness, which is restricted by constant.

All terms in equation (16) are known or calculable. The design assumes, that Bellman's function shape does not vary. Therefore if the tth approximation of Bellman's function is $\hat{V}_t(x_{t-1}, y_t)$, the non-preciseness of approximation in time t can be expressed via:

$$e_t = \max_{x_t} \mathcal{E}\left[g(x_t, x_{t-1}, y_{t+1}, y_t) + \hat{V}_{t+1}(x_{t-1}, y_t)\middle| x_0, \ldots, x_t, y_1, \ldots, y_t\right] - \hat{V}_t(_{t-1}, y_t). \quad (17)$$

Then we minimize the sum of squares $\min_{\hat{V}_t} \sum_{k=1}^{t} e_k^2$ and arguments of minimum are the best approximation of the function $\hat{V}_t(.)$. The minimization leads to least squares method.

4 Experimental part

This section describes the experimental setup, data and results obtained. The designed trading strategy is defined at discrete time $t \in \{1, 2, \ldots, T\}$. The time step corresponds with interval of 24 hours. The trading period is given by available data.

The data used for design of the strategy are so-called close prices, which are collected once a day. It is the last price, when the exchange closes trading. The economic specialists grant that close price is the most stable price. The close price y_t is assumed to be known in time t, i.e. y_t is available to design the decision x_t.

The part of data sets is transaction costs c_t. Moreover the price changes during the day and the close price represent the best approximation, but the risk constant is demanded. Therefore the slippage constant c_s, which characterizes typically change of the price in delay between decision and real trading is employed. This constant is used as penalization for each action in design. And the whole transaction costs C (firstly used in the equation (3)) is defined as $C = c_t + c_s$.

The general equations used in this paper do not specify the restriction to decision x_t. The restrictions depend on the trader's account, because traders must own money to buy or sell contract at an exchange and the range of contracts to position is limited by owned money. We use following values of decision $x_t \in \{-1, 0, 1\}$. This three values are enough for experiments, because the wider range of actions leads only to use the extremal values of decision. This phenomenon is caused by the shape of gain function (3), which is partially linear function of decision x_t. The strategy starts and ends with flat position, therefore $x_0 = x_T = 0$.

The order of model (see equation (11)) is set to $N = 2$, because this value gives the best profit of strategies in the previous research. Predictions are generated by Monte Carlo method. The count of Monte Carlo samples is chosen dynamically: The decision is final, when it is not influenced by new Monte Carlo samples.

4.1 Used data

There are 35 available price sequences for the experiments. The sequences contain prices for more than 15 year, i.e. about 3900 trading days in each sequence. The experiment set is too wide to present all results here, therefore the following five futures contracts were chosen as reference markets.

Ticker	Description
CC	Cocoa - CSCE
CL	Petroleum-Crude Oil Light
FV2	5-Year U.S. Treasury Note
JY	Japanese Yen - FOREX
W	Wheat - CBT

The reference markets were chosen by economic specialist to include all typical kind of markets - i.e. cocoa and wheat are typical agriculture product, petroleum-crude oil is mined material, Japanese Yen is typical foreign currency and treasury note stands for bond markets.

4.2 Results

There are many ways, how evaluate the quality of designed strategy. The net profit calculated by (4) is the main criterion, secondary criteria are gross loss (sum of loss trades profit), gross profit (sum of win trades profit), count of winning and losing trades. By using these criteria it is possible to calculate sum of the transaction cost and sum of slippages.

The main non-quantitative pointer is the plot of cumulative gain depending on time. It is difficult to analyze it but it gives important information about the strategy. In ideal case, the plot increases.

	CC	**CL**	**FV2**	**JY**	**W**
Net profit	-40530.00	29390.00	-26368.75	-76992.50	-13210.00
Gross profit	23020.00	120360.00	52692.50	180000.00	54707.50
Gross loss	-63550.00	-90970.00	-79061.25	-256992.50	-67917.50
Transaction cost	-1780.00	-1580.00	-1900.00	-3080.00	-2060.00
Slippages	-8900.00	-6320.00	-17812.50	-38500.00	-15450.00
Trades	89	79	95	154	103
Wining trades	24	42	31	50	39
Losing trades	65	37	64	104	64

Table 1: Result overview

The results overview is in Table 1. The system designed good strategy for exchange with oil futures (CL), where the net profit is positive and the profit grows almost all the time (see Figure 1). Worst results were at cocoa future market, where the profit decreases in time. Other markets finished with negative profit, but the curve of cumulative gain shows only local decreasing, e.g. the FV2 curve decreases only at interval [1000,2500] and the other parts stagnate (see Figure 2).

The practical approach of presented design is good, because the algorithm works at one of reference markets. And three reference markets seems that the better settings or small algorithm changes can improve them to positive results.

Although the results do not suffice the requirements to usage at real trading, the theoretical results brought improvements. The methods of Monte Carlo and iterations-spread-in-time were applied and tested to new task, where the properties of both methods can be explored.

5 Future work

The main directions of the further research are:

Bellman's function - the used approximation is oversimplified. A more complex approximation is typically used to reach better results. The analytical properties of Bellman function should be explored to find the better approximation, which should lead to higher profit.

High dimensional model - present model uses only the close price to prediction, but other data channels are available too. The usage of the high dimensional model is traditional way, how obtain better results. Additional channels contain new prices, information about traders positions etc., which brings the new important information for decision maker.

Prediction quality influenced indirectly the trading system quality. Testing of prediction quality is related with model and settings of Monte Carlo method, which can be innovated by knowledge about prediction behavior.

The listed open problems should lead to improve the results and better knowledge about the approximate dynamic programming. The further approach should support the usage of this design to trading in markets as fully automatic system.

References

[1] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Nashua, US, (2001). 2nd edition.

[2] Wikipedia, the free encyklopedia.

[3] I. Nagy, L. Pavelková, E. Suzdaleva, J. Homolová, and M. Kárný. *Bayesian decision making*. CAS, Prague, (2005).

[4] P. Nedoma, M. Kárný, I. Nagy, and M. Valečková. Mixtools. matlab toolbox for mixtures. Technical Report 1995, ÚTIA AV ČR, Prague, (2000).

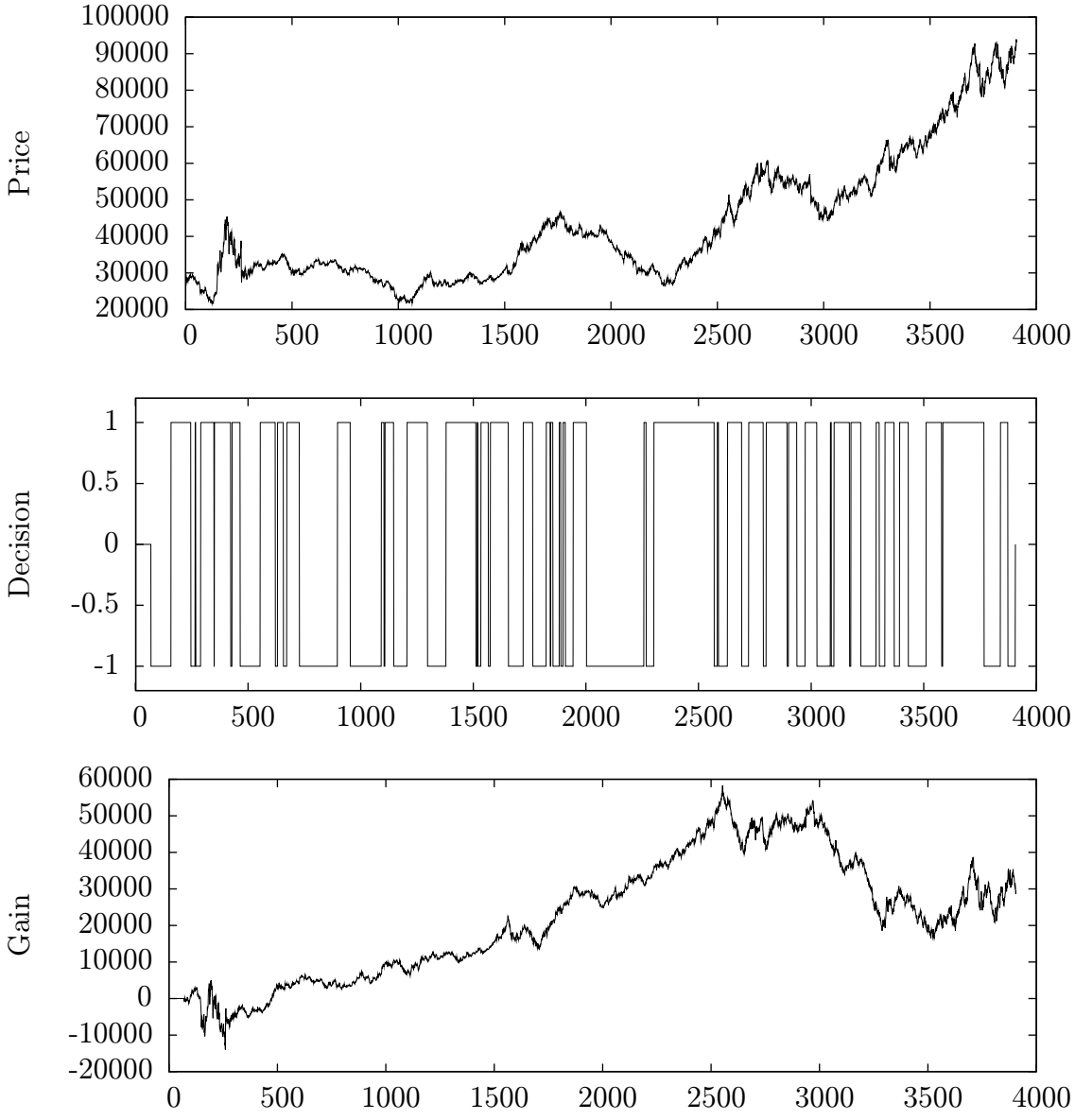


Figure 1: Results on CL

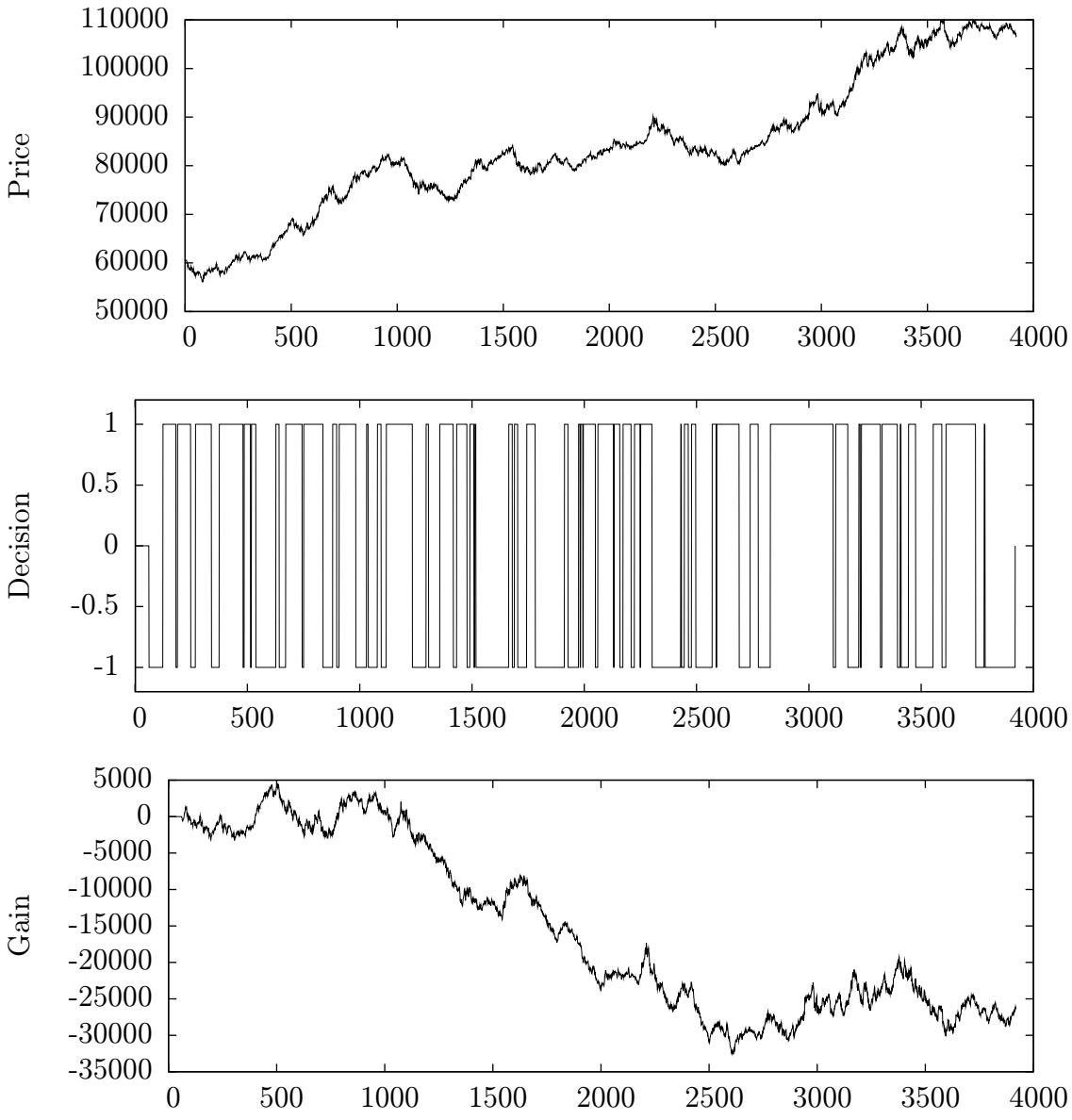


Figure 2: Results on FV2

Efficient Scheduling of Data Transfers and Job Allocations*

Michal Zerola

1st year of PGS, email: michal.zerola@ujf.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Michal Šumbera, Nuclear Physics Institute, AS CR
Jérôme Lauret, Brookhaven National Laboratory, USA
Roman Barták, Faculty of Mathematics and Physics, Charles University

Abstract. For the past decade, High Energy and Nuclear Physics experiments have been heading towards a distributed computing model in an effort to concurrently process tasks over enormous data sets, that have been increasing in size as a function of time. In order to optimize all available resources, it is necessary to face also the question of efficient data transfers and placements. In this work a model tackling this issue, based on Constraint Programming technique (CP) is introduced, as well as the representation of most important aspects of a real life scenario such as the sharing of infrastructure both when it comes to networking or storage. Methods for reducing a search tree and their side by side comparison are shown. Performance of scheduler based on Choco library is compared also with a Peer-2-Peer network simulator. Based on the preliminary results, using the CP model seems to be promising and gives good expectations for ongoing extensions.

Abstrakt. Fyzikálne experimenty vysokých energií v posledných rokoch napredujú smerom distribuovaného výpočtového modelu v snahe paralelizovať výpočty nad enormným množstvom dát, ktoré sa zvyšuje z roka na rok. Za účelom optimalizácie využitia všetkých dostupných zdrojov je potrebné čeliť i otázke efektívneho presunu a rozmiestnenia dát v distribuovanom prostredí. V tejto práci predstavíme model založený na programovaní s obmedzujúcimi podmienkami a reprezentáciu najdôležitejších vlastností reálneho prostredia. Budeme sa zaoberať i metódami zúženia prehľadávaného priestoru a predvedieme ich vzájomné porovnanie. Plánovač implementovaný za využitia knižnice Choco porovnáme tiež so simulátorom Peer-2-Peer siete. Na základe výsledkov sa použitie modelu zdá byť sľubné a dáva predpoklady na ďalšie rozšírenia.

1 Introduction

1.1 Problem area

Computationally challenging experiments such as the one from the High Energy and Nuclear Physics community (HENP) have developed a distributed computing approach (a.k.a. Grid computing model) to face the massive needs of their Peta-scale experiments. The era of data intensive computing has surely opened a vast arena for computer scientists to resolve practical and exciting problems. One of such HENP experiments is the

*The investigations have been partially supported by the IRP AVOZ10480505, by the Grant Agency of the Czech Republic under Contract No. 202/07/0079, by the grant LC07048 of the Ministry of Education of the Czech Republic and by the U.S. Department Of Energy

STAR[1] (Solenoidal Tracker at Relativistic Heavy Ion Collider) experiment located at the Brookhaven National Laboratory (USA).

In addition to a typical Peta-scale challenge and large computational needs, this experiment, as a running experiment acquires a new set of valuable real data every year, introducing other dimension of safe data transfer to the problem. From the yearly data sets, the experiment may produce many physics ready derived data sets which differ in accuracy as the problem is better understood as time passes. Thus, demands for a large-scaled storage management and efficient scheme to distribute data grows as a function of time, while on the other hand, end-users may need to access data sets from previous years and consequently at any point in time. Coordination is needed to avoid random access destroying efficiency.

The user's task is typically embarrassingly parallel; that is, a single program can run N times on fraction of the whole data set split into N sub-parts without any impact on science reliability, accuracy, or reproducibility. For a computer scientist, the issue then becomes how to split the embarrassingly parallel task into N jobs in the most efficient manner while knowing the data set is spread over the world and/or how to spread 'a' dataset and best place the data for maximal efficiency and fastest processing of the task.

The purpose of this work is to design and develop an automated system that would efficiently use all available computational and storage resources. It will relieve end users of making decisions among possible ways of their task execution (which includes locating and transferring data to desired sites that appear optimal to user) while preserving fairness. Users' knowledge of the whole system and data transfer tools will be reduced just to the communication with the future planner that will guarantee its decision to spread the task and data sets over chosen sites was, under current circumstances, the most efficient and optimal.

1.2 Milestones

Rather than trying to solve the problem directly from a task scheduling perspective within a grid environment, we split the problem into several stages. By isolating data transfer/placement and computational challenges from each other we get an opportunity to study the behavior of both sets of constraints separately.

Since individual tasks depend on a dataset with a non-trivial size, the time required for its staging and transfers is also inconsiderable. Therefore, the **first milestone** is to design and develop the data transfer planner/scheduler. For a given dataset needed at some site, its aim is to create a plan with an objective to prepare files from the dataset at a given site within the shortest time. The next requirement is to define and achieve fair share transfers within a multiuser environment. This means that if one user asked for a huge amount of data at some site, then another user who asked just for one file shouldn't wait until the first user's plan is finished.

The **next milestone** generalizes data transfer planning between sites. The goal for this stage is not to transfer files to one particular site, but do the transfer to several destinations. More precisely, the planner's goal is to achieve presence of each file (from user's input task) at one out of all possible destinations, while still having the objective

[1] http://www.star.bnl.gov

in mind, to minimize the finish time of the last file transfer the user waits for.

The second milestone is highly corellated with the **final milestone** - scheduling the data transfers together with particular tasks (jobs) on a grid. The subtask is not finished after a file is transfered at some destination site, but when the user's job executed at the same site (and dependent on this file) is finished. Thus, the planner still has the freedom of choosing a destination site for each file, but it has to consider that each site has a specific characteristic of its computational performance. These attributes include, for example, the number of available CPUs at current site or the actual load, so it can be more effective to transfer some files over the slower link to the computationally high performance site (or vice versa). The final objective is to minimize the finish time of the last user's job. In this article we focus on the first milestone.

2 Problem formalization

In the following part we will present a formal description of the problem and an approach based on Constraint Programming technique, used in artificial intelligence and operations research, where we search for assignment of given variables from their domains, in such a way that all constraints are satisfied and value of an objective function is optimal [3].

We will introduce the transfer network consisting of sites holding information which files are available at the site. For each file we will search for a path leading to the destination and time slots for each link on transfer path, when a particular file transfer should occur.

The network consists of a set of nodes $\mathbf{N}$ and a set of directed edges $\mathbf{E}$. The set $\mathbf{OUT}(n)$ consists of all edges leaving node n, the set $\mathbf{IN}(n)$ of all edges leading to node n. Input received from a user is a set of file names needed at a destination site **dest**. We will refer to this set of file names as to demands, represented by $\mathbf{D}$. For every demand d we have a set of sources ($\mathbf{orig}(d)$), sites where the file (d) is already available. We will use one decision variable for every demand and link of the network (edge in graph). The $\{0, 1\}$ variable X_{de} denotes whether demand d is routed over edge e of the network. The second variable $start_{de}$ denotes start time of transfer corresponding to the demand d over edge e. More approaches can be found in [5].

$$\min_{X_{de}, start_{de}} \max_{e \in \mathbf{E}} \underbrace{\left(start_{de} + \frac{size(d)}{speed(e)} \right)}_{end_{de}} \cdot X_{de} \tag{1}$$

$$\forall d \in \mathbf{D} : \sum_{e \in \cup \mathbf{OUT}(n | n \in \mathbf{orig}(d))} X_{de} = 1, \sum_{e \in \cup \mathbf{IN}(n | n \in \mathbf{orig}(d))} X_{de} = 0 \tag{2}$$

$$\forall d \in \mathbf{D} : \sum_{e \in \mathbf{OUT}(dest(d))} X_{de} = 0, \sum_{e \in \mathbf{IN}(dest(d))} X_{de} = 1 \tag{3}$$

$$\forall d \in \mathbf{D}, \forall n \notin \{orig(d) \cup dest(d)\} :$$
$$\sum_{e \in \mathbf{OUT}(n)} X_{de} \leq 1, \sum_{e \in \mathbf{IN}(n)} X_{de} \leq 1, \sum_{e \in \mathbf{OUT}(n)} X_{de} = \sum_{e \in \mathbf{IN}(n)} X_{de} \tag{4}$$

$$\forall e \in \mathbf{E}, \forall d \in \mathbf{D} : X_{de} = 1 : \bigcap [start_{de}, \underbrace{start_{de} + \frac{size(d)}{speed(e)}}_{end_{de}}] = \emptyset \tag{5}$$

$$\forall n \in \mathbf{N}, \forall d \in \mathbf{D} : \sum_{e \in \mathbf{IN}(n)} \underbrace{\left(start_{de} + \frac{size(d)}{speed(e)} \right)}_{end_{de}} \cdot X_{de} \leq \sum_{e \in \mathbf{OUT}(n)} start_{de} \cdot X_{de} \tag{6}$$

$$X_{de} \in \{0, 1\}$$
$$start_{de} \in \mathcal{N}^{+}$$

The *path constraints* (2, 3, 4) state that there is a single path for each demand (path starting right in one of origin sites, leading to the destination). Equation (5) ensures there is only one active file transfer over every edge in time. The last equation states that a transfer of the file at any site can start only if the file is already available at the site (Eq. 6)(i.e., a transfer of the file to this site has finished). The objective (Eq. 1) is to minimize the latest finish time of transfer over the whole files.

2.1 Constraint model

For implementation of the solver we use Choco [2], a Java based library for constraint satisfaction problems (CSP), constraint programming (CP). Among 70 available constraints Choco provides also a set for scheduling and resource allocation, we require most. Closer illustration of several Choco uses can be found in [1], [2], and [6]. In addition, Java based platform allows us an easier scheduler integration with currently used tools in the STAR environment.

Constraints introduced in the previous section were used directly via appropriate Choco structures, except the equation 5, that ensures at most single file transfer in any time on any link. For this, we used the **cumulative** scheduling constraint and notation of tasks and resources. Tasks are represented by their duration, by ranges for starting and ending times, and by resource consumption respectively. They are allocated to the resource(s) in such a way that in any time resource capacity can not be exceeded.

In our case, each link acts as a separate resource with capacity 1 (unary resource) and each file demand creates a single task on every resource, which duration depends on the current link speed (resource characteristic) and consumption of the resource corresponds to the value of variable X, i.e. no consumption if the transfer path for demand does not include current link (resource), or consumption 1 otherwise. In the Figure (1) is shown one possible schedule for transferring one file (F) with an origin at $Site_1$ and $Site_2$ to a destination $Dest$. Values of the X variables define the path, while the resource profile for each link is on the right side.

The search strategy, following Choco notation, is split into two *goals*. First one is to find assignment for X variables, i.e. paths for each transfer, while the second is to allocate time slot, assign *start* variables, for each transfer at chosen links. For both goals

[2] http://choco.sourceforge.net

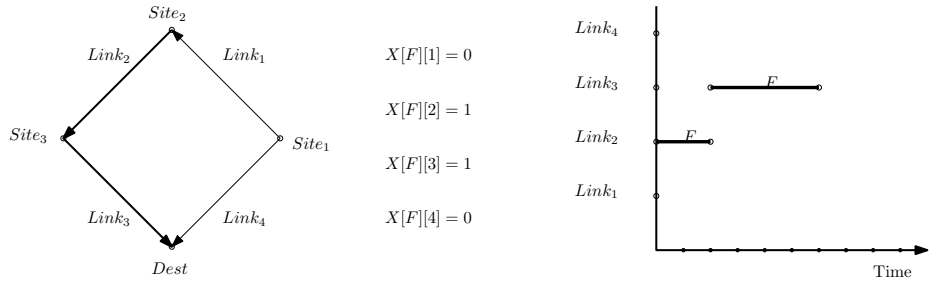


Figure 1: Example of a schedule solution with file F and its origin at $Site_1$ and $Site_2$.

the default *'minimum domain'* variable selection and *'increasing value'* value selection heuristic were used.

3 Direct connections

In order to closely analyze the problem, its scale, and behavior of used techniques, we started with several restrictions that simplify the case. We started to explore the network, where only **direct connections** for data movement are allowed. In other words, file can not be transferred from its origin to the destination by a path longer than one.

One can think that such a restriction shrinks the search space enormously, but closer look reveals that the number of possible combinations is still large:

Let's suppose that we have a network of 5 sites, all connected to the destination and 100 files available at each site ($|orig(f)| = 5$). The number of decision variables X is therefore 500 ($= |D| * |E|$). Even if an upper bound for all possible combinations (2^{500}) is reduced by a propagation to 5^{100} (solver has a freedom of 5 choices of an origin for each file), brute-force methods can run 'forever'.

With an intend to stay close to a reality, we fixed the number of sites to 5, which approximately represents the number of sites currently available in the STAR experiment. For each link we introduced a *slowdown factor* that influences the transfer time needed to move the data over this link. Slowdown factor 1 means that file of size 1 unit can be transferred in 1 unit of time, but with a slowdown factor 4 only in 4 units of time, etc.

Considering the second part of the input, the file demands, we studied the following cases: a) every file is available only at one particular site [**distinct**]; b) file is available at sites given by a probability function, that represents the reality [**weighted**]; c) file is available at all sites [**shared**]. For all cases we fixed the file size to a 1 unit, i.e. all files have the same size.

3.1 Shared links

So far we have assumed that all links incoming or outgoing from any site have their own bandwidth (slowdown factor) that is not affected by other ones. Nevertheless, in reality this is not always feasible, since several links leading to a site usually share the same router and/or physical fiber which bandwidth (capacity) is less than the sum of their

own values. Hence, simultaneously one can't use all links at their maximum bandwidths. We express this constraint by adding an additional resource per each group of shared links. Capacity of the resource will be the bandwidth of a shared link or a router, while tasks correspond to the scheduled transfers using any link belonging to this group with consumptions equal to its slowdown factor.

3.2 Reducing a search time

We studied also several techniques for reducing the time spent during a search.

3.2.1 Symmetry breaking

One of the common techniques for reducing the search tree is detecting and breaking variable symmetries. This is usually done by adding variable symmetry breaking constraints that can be expressed easily and propagated efficiently using lexicographic ordering. One idea that can be applied in the studied case (direct connections and fixed file size) is following: if two files have the same origin sets, links selected for the first one and for the second one respectively must be ordered. The reason behind is that both files must be transferred to the destination and their size is equal, it is not necessary to check also 'swapped' case, since the transfer time can not be shorter.

3.2.2 Decomposition and search limits

Another approach is based on the idea, where instead of searching for a global optimal solution that can be very computing time consuming, we try to find an optimal solution for smaller parts of the input, where sum of the time spent will be just a portion of time needed otherwise. This principle is even more suitable for our needs, since network link speeds vary in time, some sites can be down after the schedule is produced, generally, transferring all data files takes significant amount of time and during this time a lot of factors can be different to the ones the scheduler considered at the beginning. Thus the computed optimal schedule for the full input doesn't have to be valid anymore.

One of the approximations is splitting the input files into chunks and producing an optimal schedule for each chunk separately, while propagating the results from the previous ones. More precisely, result of the scheduler for a given chunk of files is information of computed starting/ending times for each file at particular links. In other words, current solver receives times for each link, by which the link will be busy, thus further scheduling for current chunk cannot place file transfer in these time-slots. We achieve this by allocating a fake task, with fixed starting and ending times, that were propagated from previous schedules (Figure: 2).

Also limits can be imposed on the search algorithm to avoid spending too much time in the exploration. One of them is fixing the time limit on a search tree. When the execution time is equal to the time limit, the search stops whether an optimal solution is found or not. One of the algorithms we studied was based on this, with a time-limit linearly dependent on the number of files in a request.

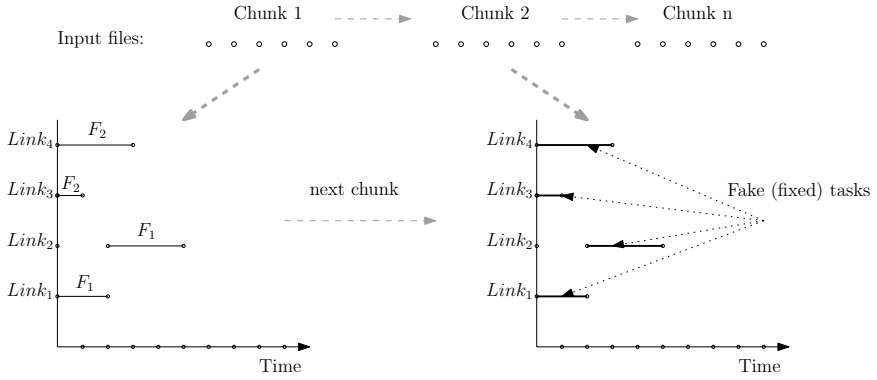


Figure 2: Allocating fake tasks according to the previous schedule.

4 Directed (simple) paths

Considering the model, no changes are necessary to perform in order to allow solver search for transfer paths longer than one. However, since data set transit takes some storage space, one must be sure that during file transfer from site A to C, using site B, there is enough space at intermediate site B.

4.1 Storage capacity

In order to respect storage restrictions we introduce the next attribute for each site, the available (free) space, or the storage capacity. All the time during the execution of a schedule, the storage capacity constraint for each site must be respected.

For each site we consider all possible ways (pairs of *inLink* and *outLink* how a file can be transferred trough it. Whether or not a pair is really used for the demand d is expressed by $channelingVariable$, using which we define also consumption of the task (Figure: 3).

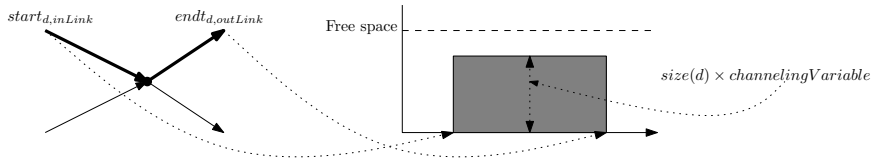


Figure 3: Storage resource.

If the pair is not used, the consumption is set to zero and storage resource is invariable to this task, otherwise the consumption is set to the file size.

5 Comparative studies

In this section we present the performance comparison of several methods of the CSP solver introduced in previous sections as well as of the Peer-2-Peer simulator. We will also

show an effect of one constraint (storage based) for a simple paths case and an example of the optimal schedule produced by the solver.

5.1 Peer-2-Peer simulator

To provide a base comparison with the results of our CSP based solver we chose to implement a Peer-2-Peer (P2P) model as well. This model is well known and successfully used in similar fields like file sharing, telecommunication, or media streaming. We implemented a P2P simulator by creating the following work-flow: **a)** put an observer for each link leading to the destination; **b)** if an observer detects the link is free, it picks up the file at his site (link starting node), initiate the transfer, and waits until the transfer is done. We introduced a heuristic for picking up a file as typically done for P2P. Link observer picks up a file with a smallest cardinality in the sense of its $|origin|$, i.e. the file that is available at the smallest number of sites and if there are more files available with the same cardinality, it randomly picks any of them. After each transfer, the file record is removed from the list of possibilities over all sites. This process is typically resolved using distributed hash table (DHT) [4], however in our simulator only simple structures were used. Finally an algorithm finishes when all files reach the destination, thus no observer has any more work to do.

5.2 Results

In Figure 4, we show a comparison of times needed to produce the schedules and divergence of the results (makespan) to the optimal solution between several algorithms. We present results only for **weighted** case with direct connections and will only describe the qualitative features for the other cases. Weights (probabilities) that were used for sites considering file's origins were $1.0, 0.6, 0.01$, and 0.01.

The X axes denote the number of files in a request while Y is the time (in units) needed to generate the schedule and percentage loss on optimal solution. We can see that time to find an optimal schedule without any additions grows exponentially and is usable only for a limited number of files, 50 in the weighted case and 20 in the shared case. This difference is induced by a higher number of possible configurations as long as any site can be selected as an origin. By introducing symmetry breaking, the solving time is improved, but still not usable for more than 200 files. Using the time-limit on the other hand we are moving apart from an optimal solution with increasing files in request, which is even more visible in shared case. Thus setting the time-limit as a linear function to the number of files, while using a default search strategy based on minimal domains, is not sufficient.

In contrast, splitting the input into chunks is giving the best performance results both in the running time and also in the quality of the makespan. Even scheduling by chunk of size 1, i.e. file by file, doesn't produce worse result than using larger chunks due to previous conditions propagation. We note as well the efficacious performance of a simple P2P algorithm, but it is worth to mention that this model is usable only in a direct connection case, while our intent is to study more complex networks with much more restrictions.

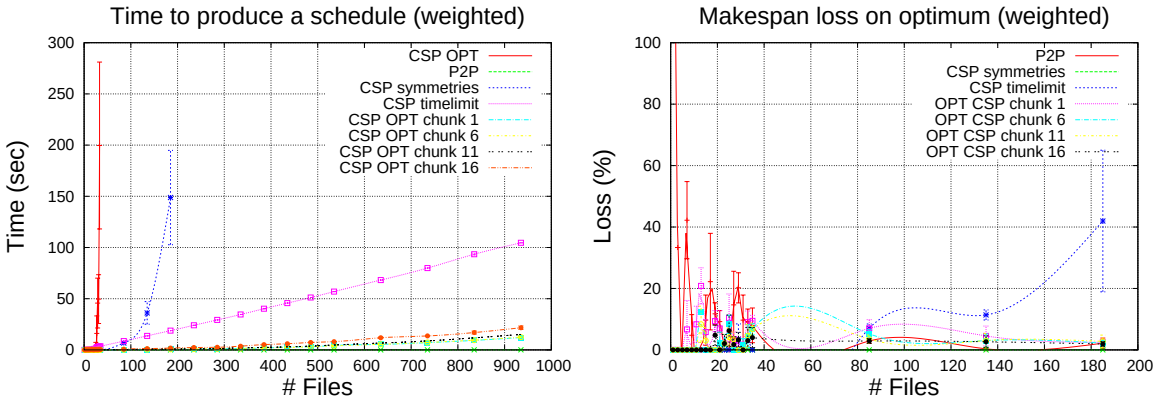


Figure 4: Approximation of the runtime (left) and makespan loss on optimal schedule (right) for weighted case.

To see the real effect of the storage constraint, in Gantt charts (Figure 5) are shown two schedules (without and with enabled constraint) for the same dataset, considering the funnel network displayed in the upper part of the figure with a limited available space at $Site_3$ only for one file size unit. This extreme example permits only a single transfer via site $Site_3$, that fills available space until a file is fully transfered to the destination $Site_4$. After that, the space at $Site_3$ is again released and another file can go trough.

6 Conclusion

We presented an approach using a Constraint Programming model to tackle the efficient data transfers/placements and job allocations problem within a distributed environment. Usage of constraints and declarative type of programming offers straightforward and more error prone way of representing many real life restrictions. On the other hand, since a search space is usually extensive, methods like symmetry breaking or approximations and understanding the scale of the problem are fundamental. We showed that using the scheduling of data transfers by sequence of smaller chunks gives results close to the optimal solution and provides very acceptable running time performance. We have implemented also several constraints for dealing with shared network links or limited storage capacities at sites and actual results indicate that it is worth to continue research with this technique.

References

[1] D. Benavides, S. Segura, P. T. Martín-Arroyo, and A. R. Cortés. Using java CSP solvers in the automated analyses of feature models. In 'GTTSE', R. Lämmel, J. Saraiva, and J. Visser, (eds.), volume 4143 of *Lecture Notes in Computer Science*, 399–408. Springer, (2006).

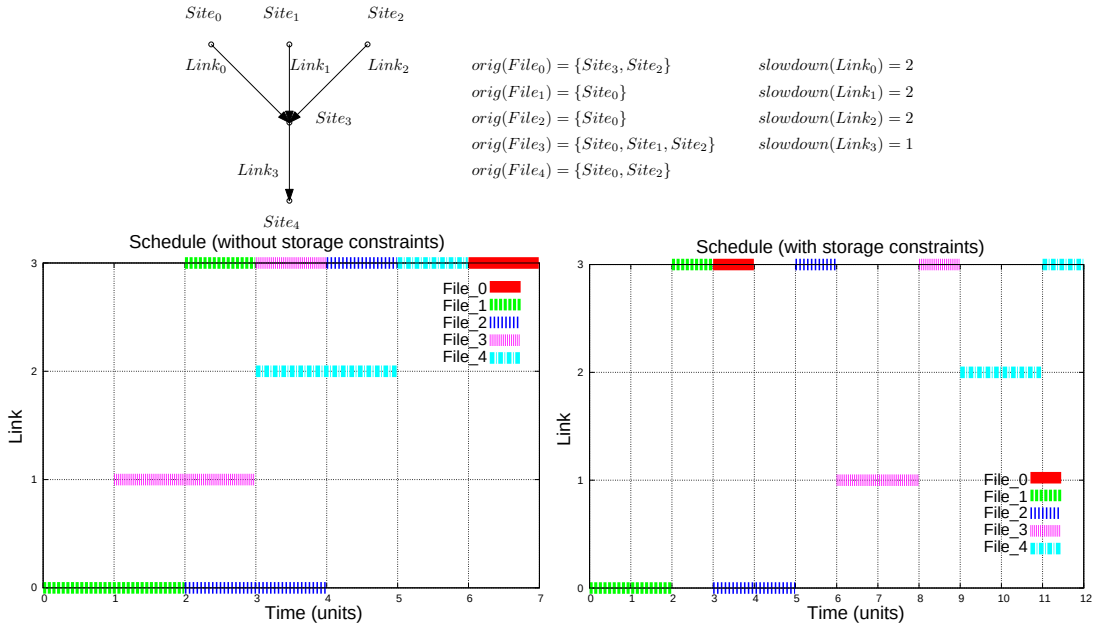


Figure 5: Gantt chart of a schedule without storage constraints (left) and a schedule with storage constraints (right) generated on the funnel network with limited storage capacity (up).

[2] A. Lazovik, M. Aiello, and R. Gennari. Choreographies: using constraints to satisfy service requests. In 'AICT/ICIW', 150. IEEE Computer Society, (2006).

[3] K. Marriott and P. Stuckey. *Programming with Constraints*. MIT Press, Cambridge, Massachusetts, (1998).

[4] Naor and Wieder. A simple fault tolerant distributed hash table. In 'International Workshop on Peer-to-Peer Systems (IPTPS), LNCS', volume 2, (2003).

[5] H. Simonis. Challenges for constraint programming in networking. In 'CP', M. Wallace, (ed.), volume 3258 of *Lecture Notes in Computer Science*, 13–16. Springer, (2004).

[6] J. White, D. C. Schmidt, K. Czarnecki, C. Wienands, G. Lenz, E. Wuchner, and L. Fiege. Automated model-based configuration of enterprise java applications. In 'EDOC', 301–312. IEEE Computer Society, (2007).