DOKTORANDSKÉ DNY 2014

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

14. a 21. listopadu 2014

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2014
sborník workshopu doktorandů FJFI oboru Matematické inženýrství

P. Ambrož, Z. Masáková (editoři)
Kontakt petr.ambroz@fjfi.cvut.cz / 224 358 569

Vydalo České vysoké učení technické v Praze
Zpracovala Fakulta jaderná a fyzikálně inženýrská
Vytisklo Nakladatelství ČVUT-výroba, Zikova 4, Praha 6
Počet stran 274, Vydání 1.

ISBN 978-80-01-05605-9

Seznam příspěvků

Předmluva

Doktorandské dny jsou setkáním, na které se každoročně těší především doktorandi oboru Matematické inženýrství na FJFI, ale i jejich školitelé a další spřátelení členové akademické obce ČVUT. Letos se konají již podeváté a opět na nich studenti vystoupí s prezentacemi své výzkumné práce. Čekají nás příspěvky z oblasti diskrétní matematiky, teoretické i aplikované informatiky, numerické matematiky, stochastického modelování a v neposlední řadě také matematické fyziky. Tento sborník přináší texty k příspěvkům a abstrakty vystupujících studentů.

I tento ročník workshopu Doktorandské dny finančně podpořila Studentská grantová soutěž ČVUT, grant SVK 34/14/F4. Děkujeme.

Editoři

Limit Distribution and Three-State Quantum Walk

Iva Bezděková*

3rd year of PGS, email: bezdekova.iva@gmail.com
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Igor Jex, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Martin Štefaňák, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. If we begin with well-known two-state Hadamard walk, we see that the limit distribution depends on the choice of the initial state. In [1] the distribution is derived with respect to the initial state in standard basis. Nevertheless in [3] authors derived the distribution in different basis that is formed by eigevectors of the coin operator. This approach simplifies the resulting description of the walk. The aim of this work is to describe how the choice of the inicial state affects higher-dimensional quantum walks, namely we take two deformations of the three-state Grover walk called eigenvalue and eigenvector family. Both families exhibits the effect of localization, which means additional peak at the origin of the probability distribution. The construction of the eigenvalue and eigenvector family is based on continuous transfer of the Grover matrix to some trivial matrix [4]. If we take anti-diagonal permutation matrix as a coin, the walk will be localizing and the spectrum of this matrix and the Grover matrix differs in a sign of one eigenvalue, while the eigenvectors can be chosen the same. The addition of a phase factor into the spectral decomposition of the Grover matrix provides continuous transfer between these two matrices and give us so-called eigenvalue family. The eigenvector family is the opposite case. We have to parametrize eigenvectors of the Grover matrix in order to get trivial matrix with the same spectrum. Recently, quantum walk with eigenvector family as a coin operator were studied from the viewpoint of the limit distribution [2]. The author derived the distribution with respect to the initial state in standard basis. However, as in the case of Hadamard walk, one can also express the initial state in different and more convenient basis. The new basis formed by eigenvectors of the coin operator simplifies the result and reveals previously hidden features. Similar approach can be applied to the eigenvalue family. Both families lead to the localizing walk so the dependence of trapping probability at the origin on the eigenvector basis can be expressed.

The full paper: Martin Štefaňák, Iva Bezděková and Igor Jex, Phys. Rev. A 90, 012342 (2014) or arXiv:1405.7146.

Keywords: quantum walk, localization, limit distribution

Abstrakt. Již u Hadamardovy procházky s dvěma stavy chodce je vidět, že limitní rozdělení závisí na volbě počátečního stavu. V [1] je toto rozdělení odvozeno vzhledem k počátečnímu stavu vyjádřenému ve standardní bázi. Přechodem k jiné bázi se toto rozdělení dá zjednodušit,

*This work was supported from SGS13/217/OHK4/3T/14 and GAČR 13-33906S.

jako nejvhodnější se při tom chová báze tvořená vlastními vektory mince [3]. Zajímali jsme se o to, zda by šel podobný přístup použít i pro jiné typy procházek. Konkrétně jsme vybrali dvě zobecnění Groverovy procházky, které vykazují lokalizaci, tedy nenulový pík v počátku pravděpodobnostního rozdělení. Zobecnění budeme nazývat jako rodina vlastních hodnot a rodina vlastních vektorů. Jejich konstrukce probíhá jako parametrizace Groverovy mince společně s jistou triviální mincí [4]. V prvním případě, kdy jde o rodinu vlastních hodnot, bude naše triviální mince rovna antidiagonální permutační matici. Ta má až na znaménko u jedné vlastní hodnoty stejné spektrum jako Groverova matice, přičemž vlastní vektory obou matic můžeme volit shodné. Přidáme-li do spektrálního rozkladu Groverovy matice k oné vlastní hodnotě fázový faktor, můžeme spojitě přecházet mezi Groverovou a permutační maticí. Opačný případ je rodina vlastních vektorů, kdy je naopak potřeba parametrizovat vlastní vektory Groverovy matice tak, aby přešli ve vlastní vektory triviální matice se stejným spektrem. Nedávno byla rodina vlastních vektorů jako operátor mince studována z pohledu limitního rozdělení [2]. Autor zde odvodil limitní rozdělení vzhledem k počátečnímu stavu ve standardní bázi. Podobně jako v případě Hadamardovy procházky je vhodnější, pokud počáteční stav vyjádříme v jiné bázi. Ta bude opět tvořena vlastními vektory operátoru mince. Nová báze navíc odhalí dříve skryté zajímavé vlastnosti. Stejný postup se dá aplikovat i na procházku, kde jako minci volíme rodinu vlastních hodnot. Jelikož v obou případech jde o kvantovou procházku vykazující lokalizaci, spočteme rovněž pravděpodobnost záchytu částice v počátku v závislosti na vhodné volbě báze.

Plná verze článku: Martin Štefaňák, Iva Bezděková and Igor Jex, Phys. Rev. A 90, 012342 (2014) nebo arXiv:1405.7146.

Klíčová slova: kvantová procházka, lokalizace, limitní rozdělení

References

[1] N. Konno. *Limit theorems and absorption problems for quantum random walks in one dimension.* Quantum. Inf. Comput. **2** (2002), 578.

[2] T. Machida *Limit theorems of a 3-state quantum walk and its application for discrete uniform measures.* ArXiv:1401.1522.

[3] M. Štefaňák, S. M. Barnett, B. Kollár, T. Kiss and I. Jex. *Directional correlations in quantum walks with two particles.* New J. Phys. **13** (2011), 033029.

[4] M. Štefaňák, I. Bezděková and I. Jex. *Continuous deformations of the Grover walk preserving localization.* Eur. Phys. J. D **66** (2012), 142.

Phase Transition in Pedestrian Flow*

Marek Bukáček†

2nd year of PGS, email: marek.bukacek@seznam.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Miroslav Virius, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The transition between low and high density phases is a typical feature of systems with social interaction (e.g. traffic systems [2]). This contribution focuses on such characteristic in a simple evacuation design of one room with one entrance and one exit, which can be understood as one segment of larger network. The phase of the system is evaluated with respect to the inflow – controlled boundary condition. Critical values of inflow and outflow are described with respect to the transition from low density to congested state.

To handle this task, four passing-through experiments were organized [4], [7], [8]. By means of automatic image processing, pedestrians were detected, identified and significant qualities (e.g. travel time or occupation) were extracted. Moreover, due to the ability of specific participant identification, detailed microscopic analysis of travel time is provided. The data measured under experimental conditions are used to study general qualities of pedestrian flow as well, the fundamental diagram describing the dependency of flow or velocity to density is evaluated.

Simultaneously to this project, the same design was studied by means of Floor Field model [1], [3], a cellular automata tool frequently used to capture pedestrian flow. This model was enhanced by adaptive time span and principle of bounds [5], which improve the realistic behavior on the microscopic level. Considering the average travel time through the room and average room occupancy the settings incorporating the bounds and synchronous update seems to match the experimental data better [6].

Keywords: pedestrian behavior, egress experiments, image processing

Abstrakt. Fázový přechod mezi stavy nízkých a vysokých hustot je typickým prvkem pozorovaným v systémech se sociálními interakcemi (např. dopravní systémy [2]). Tento příspěvek se zaměřuje na výzkum tohoto jevu na jednoduchém scénáři – průchod skupiny chodců místností s jedním vchodem a jedním východem, který může reprezentovat jeden prvek rozsáhlejšího komplexu. Fázový přechod byl vyhodnocován z pohledu kontrolovaného příchodu osob do místnosti (okrajová podmínka). Kritické hodnoty toku do místnosti a ven byly stanoveny na základě přechodu systému ze stavu nízkých hustot do stavu kongesce.

V průběhu posledních let proběhly ve studovně FJFI čtyři evakuační experimenty s důrazem na automatické zpracování kamerových záznamů [4], [7], [8]. Jednotliví chodci byli s pomocí

*This work was supported by the grant SGS12/197/OHK4/3T/14 and the research program MSM 6840770039.

†This study has been provided in cooperation with Pavel Hrabák.

speciálních čepic rozpoznáni, identifikováni a na základě jejich drah byly vyhodnoceny významné veličiny (např. čas průchodu či obsazenost). Díky možnosti identifikace konkrétních účastníků jsou navíc představeny některé mikroskopické vlastnosti pohybu chodců. Měřená data byla použita i k ověření základních vlastností pohybu chodců, například fundamentální diagram popisující závislost rychlosti či toku na hustotě byl vyhodnocen.

Paralelně k tomuto projektu byl totožný design simulován pomocí Floor Field modelu [1], [3], celulárního automatonu běžně využívaného k modelování pohybu lidí. tento model byl doplněn o adaptivní časový krok a princip vazeb mezi chodci [5], což vylepšuje vlastnosti modelu na mikroskopické úrovni. Z pohledu průměrného času průchodu a průměrné obsazenosti se ukazuje, že nastavení modelu zahrnující vazby a synchronní update odpovídá experimentálně měřeným datům nejlépe [6].

Klíčová slova: chování chodců, evakuační experimenty, zpracování obrazu

References

[1] A. Kirchner and A. Schadschneider, *Simulation of Evacuation Processes Using a Bionics-Inspired Cellular Automaton Model for Pedestrian Dynamics*, Physica A **312** (2002), 260–276.

[2] A. Schadschneider, D. Chowdhury and K. Nishinari, *Stochastic Transport in Complex Systems*, Elsevier (2010), ISBN: 978-0080560526.

[3] P. Hrabák, M. Bukáček and M. Krbálek, *Cellular Model of Room Evacuation Based on Occupancy and Movement Prediction*, In 'ACRI 2012 Proceedings', LNCS **7495** (2012), 709–718.

[4] P. Hrabák, M. Bukáček and M. Krbálek, *Cellular Model of Room Evacuation Based on Occupancy and Movement Prediction, Comparison with Experimental Study*, Journal of Cellular Automata **8** (2013), 383 – 395.

[5] M. Bukáček, P. Hrabák and M. Krbálek, *Cellular Model of Pedestrian Dynamics with Adaptive Time Span*, In 'PPAM 2013 Proceedings', LNCS **8385** (2014), 669 – 678.

[6] M. Bukáček and P. Hrabák, *Case Study of Phase Transition in Cellular Models of Pedestrian Flow*, In 'ACRI 2014 Proceedings', LNCS **8751** (2014), 508 – 517.

[7] M. Bukáček, P. Hrabák and M. Krbálek, *Experimental Analysis of Two-Dimensional Pedestrian Flow in front of the Bottleneck*, In 'TGF 2013 Proceedings', in print. [available at http://arxiv.org/abs/1408.6107]

[8] M. Bukáček, P. Hrabák and M. Krbálek, *Experimental Study of Phase Transition in Pedestrian Flow*, In 'PED 2014 Proceedings', in print. [available at http://arxiv.org/abs/1408.6108]

Bethe Vectors for Heisenberg Spin Chains*

Jan Fuksa

3rd year of PGS, email: fuksajan@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Čestmír Burdík, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Aleksei P. Isaev, Bogoliubov Laboratory of Theoretical Physics, JINR Dubna

Abstract. Algebraic Bethe ansatz has turned out as remarkably sufficient tool in the theory of quantum integrable systems. Its origins come up to the 80's and are connected mainly with Leningrad shool. Since that time, it was used successfully to solve an amount of quantum models, cf. [3, 8]. This contribution deals with algebraic Bethe ansatz for XXX- and XXZ-spin chains which were used, e.g., to describe crystals with specific properties [9]. Our first aim is to describe Bethe vectors in fermionic representation [7]. At first, we prescribe a way how to write Bethe ansatz in terms of fermions. Then, generators of Yang-Baxter algebra are expressed in fermionic representation and used to calculate explicit form of Bethe vectors up to 3-magnons. We formulate a conjecture about general form of M-magnons, cf. [1, 2]. In the next part, we discuss N-component model and use it to find explicit form of Bethe vectors in terms of usual spin operators. This result turns out to be useful to prove our conjecture about Bethe vectors in fermionic representation [4, 1, 2]. We also discuss inhomogeneous XXX- and XXZ-spin chains and find explicit forms of their Bethe vectors [4].

This contribution is based on our texts [4, 1, 2].

Keywords: Yang–Baxter equation, algebraic Bethe ansatz, Bethe vectors, quantum integrable systems

Abstrakt. Algebraický Betheho ansatz se osvědčil jako velice úspěšná metoda kvantové integrability. Jeho počátky sahají do 80. let a jsou spjaty především s Leningradskou školou. Od té doby byl úspěšně použit k řešení množství kvantových modelů, viz. [3, 8]. Tento příspěvek se zabývá algebraickým Betheho ansatzem pro spinové XXX a XXZ řetízky, které byly použity např. k popisu krystalů se specifickými vlastnostmi [9]. Naším prvním cílem je popsat Betheho vektory pomocí fermionů [7]. Nejdříve popíšeme způsob, jak formulovat Betheho ansatz ve fermionové reprezentaci. Poté v této reprezentaci vyjádříme generátory Yang-Baxterovy algebry a použijeme je pro výpočet explicitní podoby Betheho vektorů až do 3-magnonů. Formulujeme hypotézu o obecné podobě M-magnonů, srov. [1, 2]. V další části se budeme zabývat N-komponentním modelem a použijeme ho k nalezení explicitní podoby Betheho vektorů v obvyklé reprezentaci pomocí spinových operátorů. Tento výsledek se ukáže být užitečný k důkazu naší domněnky o Betheho vektorech ve fermionové reprezentaci [4, 1, 2]. Rozebereme také případ nehomogenních XXX a XXZ řetízků a nalezneme explicitní tvar jejich Betheho vektorů [4].

Tento příspěvek je založen na našich textech [4, 1, 2].

Klíčová slova: Yang–Baxterova rovnice, algebraický Betheho ansatz, Betheho vektory, kvantové integrabilní systémy

*This work has been supported by the grant SGS12/198/OHK4/3T/14.

References

[1] Č. Burdík, J. Fuksa, A. P. Isaev, *Bethe vectors for XXX-spin chain*, submitted to J. Phys.: Conf. Ser.

[2] Č. Burdík, J. Fuksa, A. P. Isaev, S. O. Krivonos, O. Navrátil, *Remarks on the spectrum of the Heisenberg spin chain type models*, to be submitted to Particles & Nuclei.

[3] L. D. Faddeev, *How Algebraic Bethe Ansatz works for integrable model*, (1996), `arXiv:hep-th/9605187`.

[4] J. Fuksa, A. P. Isaev, N. A. Slavnov, *in preparation.*

[5] F. Göhmann, V. E. Korepin, *Solution of the quantum inverse problem*, J. Phys. A: Math. Gen. **33** (2000), pp.1199-1220.

[6] A. G. Izergin, V. E. Korepin, *The quantum inverse scattering method approach to correlation functions*, Comm. Math. Phys. **94**, No.1, (1984), pp.67-92.

[7] P. Jordan, E. Wigner, Z. Phys. **47** (1928), p.631.

[8] N. A. Slavnov, *Algebraic Bethe ansatz and quantum integrable models*, Uspekhi Fiz. Nauk **62** (2007), p.727.

[9] B. Lake, D. A. Tennant, J.-S. Caux, T. Barthel, U. Schollwöck, S. E. Nagler, C. D. Frost, *Multispinon continua at zero and finite temperature in a near-ideal Heisenberg chain*, Phys. Rev. Lett. 111, 137205 (2013).

Dynamic Texture Modelling and Editing Using Temporal Mixing Coefficients Reduction[*]

Michal Havlíček[†]

5th year of PGS, email: havlimi2@utia.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Haindl, Pattern Recognition Department
Institute of Information Theory and Automation, ASCR

Abstract. Real world materials often change their appearance over time. If these variations are spatially and temporally homogeneous then the visual appearance can be represented by a dynamic texture which is a natural extension of classic texture concept including the time as an extra dimension. In this article we present possible way to handle multispectral dynamic textures based on a combination of input data eigen analysis and subsequent processing of temporal mixing coefficients. The proposed method exhibits overall good performance, offers extremely fast synthesis which is not restricted in temporal dimension and simultaneously enables to compress significantly the original data and additionally perform texture editing.

Keywords: Dynamic texture, texture analysis, texture synthesis, texture editing, data compression, computer graphics.

Abstrakt. Skutečné materiály velmi často mění svůj vzhled v čase. Pokud jsou tyto změny prostorově a časově homogenní pak lze vizuální vlastnosti reprezentovat dynamickou texturou, která představuje přirozené rozšíření konceptu klasické textury o čas jako dodatečný rozměr. V tomto článku přestavujeme možný přístup k multispektrálním dynamickým texturám založený na kombinaci vlastní analýzy vstupních dat a následného zpracování časových směsových koeficientů. Navržená metoda vykazuje celkově dobré vlastnosti, nabízí extrémně rychlou syntézu, která není omezená v časovém rozměru a zároveň umožňuje výrazně komprimovat původní data a navíc textury editovat.

Klíčová slova: Dynamická textura, analýza textur, syntéza textur, editace textur, komprese dat, počítačová grafika.

1 Introduction

Dynamic textures (DT) can be defined as spatially repetitive motion patterns exhibiting homogeneous temporal properties. Good example might be smoke, fire or liquids, also waving trees, straws or some moving mechanical objects can be also sometimes considered as DT. A sequence of either monospectral or multispectral images which are called frames is the simplest representation of DT. Measured DT data are always represented by a finite length sequence, sometimes too short for an intended application. This property may

[*] This research was supported by the grants GACR 14-10911S, GACR 14-0265S.

[†] Pattern Recognition Department, Institute of Information Theory and Automation, ASCR.

limit possible use of DTs in virtual reality systems so temporally unconstrained synthesis of DT is an interesting and challenging research problem in several computer graphics, computer vision, and pattern recognition applications. On the other hand, synthesis can be considered when dynamics of the texture can be represented more efficiently.

The contribution of this paper is to propose straightforward multispectral DT modelling method with low computational demands enabling extremely fast synthesis of arbitrarily long DT sequence and in addition compression of original data and possible texture editing. The method is based on input data dimensionality reduction using eigen analysis and selective elimination of resulted temporal coefficients.

2 Related Work

Already published articles dealing with DTs can be divided according to the application to: recognition, representation and synthesis [2]. The DT synthesis is apparently the most difficult task and there are only few papers on this topic available [11, 10, 1, 3, 5]. Some methods [11, 10] are limited by time consuming synthesis algorithm (in addition method [11] requires some high level of temporal homogeneity of the input and is restricted to monospectral DTs), method [1] is limited to finite length sequence generation.

Another possibility is to utilize so called video editing techniques [12, 9, 6], developed for general video sequences originally, which can be used for DT synthesis as DT can be considered as a special case of general video sequence. For example video texture generation based on searching for transition points for looping with additional blending and morphing [9]. Evident drawback is using blend and morphing to achieve continuity of the synthesized sequence which may introduce blur and other unfavourable visual artifacts. This issue was solved in [6]. Another possibility is tree structured vector quantization published in [12]. This method sometimes fail to reproduce global structures which may appear in the original data [6]. Video editing techniques are also very often time demanding [3]. We compare most of the above mentioned methods in Section 8.

3 Method Overview

The whole modelling process can be divided into two phases: analysis and synthesis. The first step of the analysis is so called normalization of DT during that an average frame from all frames in the analyzed sequence is computed per-pixel and then subtracted per-pixel from each frame in the DT. Eigen analysis, described in detail in Section 4, follows. The results of that analysis are eigen images and temporal mixing coefficients which are further processed during temporal mixing coefficients reduction as explained in Section 5. Average frame, eigen images and reduced temporal mixing coefficients are saved for synthesis purposes. Synthesis procedure, described in detail in Section 6, consists of non-deterministic temporal mixing coefficients selection, sequence synthesis driven by chosen coefficients and denormalization which is an addition of the average frame to every single frame in the synthesized sequence i.e. inverse procedure to the normalization.

4 Dynamic Texture Eigen Analysis

Normalized DT sequence is analyzed and compressed by means of Principal Component Analysis (PCA) which is able to create a low-dimensional representation of the original data describing them as accurately as possible. Original data can be then reconstructed simply by linear combination of the low-dimensional basis. We used traditional PCA method for this task because of its optimal performance beside alternative choices as for example non-linear techniques [7].

Values corresponding to pixel intensities of individual frames from the normalized sequence are arranged into column vectors forming $(n \times t)$ matrix C where n is a number of values equals frame width × frame height × number of spectral planes in the DT and t is a number of DT frames. Then a covariance $(t \times t)$ matrix A is computed as $A = C^T C$. Thus the matrix A expresses how individual frames of the sequence depend on the others. The matrix A is decomposed using singular value decomposition so that $A = UDU^T$, where U is an orthogonal matrix of eigen vectors and D is a diagonal matrix of corresponding eigen numbers. Their values are proportional to their significance in data reconstruction so that some of them can be not used with simultaneous minimal impact on reconstruction error. Therefore only $k < t$ eigen vectors corresponding to eigen numbers which are expected to represent the most of the information are used. A threshold τ for selecting vectors which are used is computed from the values of the eigen numbers as:

$$\tau = \frac{1}{t} \sum_{i=1}^{t} D_{(i,i)} \ . \tag{1}$$

If $D_{(i,i)} > \tau, \ i \in \{1, \ldots, t\}$ then i-th column of D and i-th column of U are used. All used columns of D and U form new matrices D^* and U^* respectively. The $(n \times k)$ matrix I of eigen images can be computed as: $I = CT$, where T is a $(t \times k)$ matrix with elements:

$$T_{(i,j)} = \frac{U^*_{(i,j)}}{\sqrt{D^*_{(j,j)}}} \ .$$

Computed matrix I represents the reduced basis for the reconstruction of the original data therefore a matrix representing linear combination coefficient is needed. This role is played by a matrix of temporal mixing coefficients which is computed as: $M = I^T C$. The $(k \times t)$ matrix M, which in fact reflects the overall dynamics of the sequence, is a subject of further processing described in following Section.

5 Temporal Mixing Coefficients Reduction

The matrix of temporal mixing coefficients M described in previous Section can be further processed so that it provides additional compression and enables to apply non-deterministic synthesis algorithm guaranteeing potentially infinite DT sequence generation. For non-deterministic synthesis purposes analyzed DT is redefined in terms of graph theory so that the sequence is represented as a directed graph $\mathcal{G}$ where the individual frames play the role of vertices and the order of the frames plays the role of edges i.e. the graph structure defines for each frame the set of frames which may immediately follow. Thus before the reduction an adjacency matrix A of $\mathcal{G}$ is formed as:

$$A_{(i,j)} = \begin{cases} 1 & j = i+1 \\ 0 & \text{otherwise} \end{cases} ,$$

where $i \in \{1,\dots,t\}$, $j \in \{1,\dots,t\}$. The idea of the reduction is to keep only those columns of M for which there is no other sufficiently similar column in M, in terms of certain metric, or generally distance.

Let a $(t \times t)$ matrix Δ is composed of elements:

$$\Delta_{(i,j)} = \sum_{l=1}^{k} (M_{(l,i)} - M_{(l,j)})^2 \ , \tag{2}$$

i.e. Δ consists of all mutual distances of the columns of M. Apparently, Δ is symmetric, with zero diagonal, and therefore it is sufficient to take into account only those elements $\Delta_{(i,j)}$ for which $i < j$ holds and let $\Delta_{(i,j)} = 0$ otherwise. Distance (2) was chosen because of its proven reasonable properties for column comparison purpose and low computing demands.

An average distance δ is defined by the elements of Δ as:

$$\delta = \frac{1}{|Z|} \sum_{Z} \Delta_{(i,j)} \ ,$$

where Z is the set defined as $\{\Delta_{(i,j)} : i < j\}$. The average distance δ plays the role of the criterion determining the similarity of the columns of M, in the sense of the distance (2).

First adjacency matrix A is processed using Δ and δ as described in following algorithm: $\forall i \in< 1; t-1 >$: $\forall j \in< i+1; t >$ if $\Delta_{(i,j)} < \delta$ holds then update adjacency matrix as follows: $A_{(i,j)} = 0$, $A_{(i,i)} = 1$ and in addition if $j < t$ also holds then update adjacency matrix further like that: $A_{(j,j+1)} = 0$, $A_{(i,j+1)} = 1$. A brief demonstration of presented algorithm is shown in Figure 1.

Remaining r columns of M, i.e. every column $i : \exists j \in \{1,\dots,t\} : A_{(i,j)} = 1$, form new $(k \times r)$ matrix M^* which has to be stored and is later used for synthesis purposes described in following Section.

6 Dynamic Texture Synthesis

The goal of the synthesis is to create DT sequence of required length. Dimensions of its individual frames are identical to those of the original DT as the method is restricted to temporal synthesis and compression.

During the synthesis a $(k \times t^\dagger)$ matrix of temporal mixing coefficients $M^\dagger$, where $t^\dagger$ is a length of the synthesized sequence (in general different from t) is created column wise according to the following algorithm. The first column of $M^\dagger$ is randomly chosen column of M^*. Let the last chosen column of M^* has index $i \in \{1,\dots,r\}$ then the next column

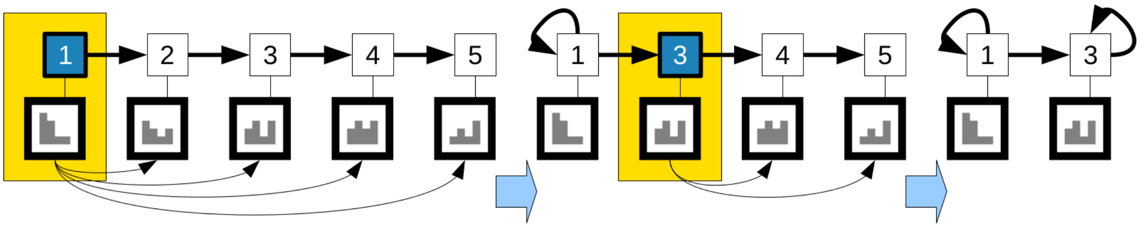


Figure 1: The reduction algorithm illustrated on the 5 frames long sample sequence. Numbering reflects the order of the frames in the original sequence. Vectors of the mixing coefficients corresponding to the individual frames are symbolically represented in the form of bar graphs (the height of each bar equals to the value of an individual coefficient). In this example reduction is performed in three steps: 1) vector of mixing coefficients of the first frame is compared with vectors of mixing coefficients of the frames of the rest of the sequence as indicated by thin dashed arrows, 2) second frame was evaluated as too similar to the first so it was replaced, graph structure was updated and vector of mixing coefficients of the third frame is compared with vectors of mixing coefficients of the remaining frames, 3) resulting graph structure describing reduced sequence.

is randomly chosen column of M^* from those which fulfill: $A_{(i,j)} = 1,\ j \in \{1, \ldots, r\}$. This provides required continuity of the synthesized sequence since the set from which the selection is performed consists of the frames such that there exists an edge between them and the last chosen one in $\mathcal{G}$ (see Section 5).

If there does not exist any such column in M^* then the column closest to the i-th column of M^*, in sense of (2), is chosen. This can occur if the graph representing the DT after reduction step described in Section 5 is not connected. Above mentioned rule provides continuous sequence with no need to utilize any additional technique such as morphing which could introduce some unfavourable artifacts to the visual information.

It is important to not allow choose the same columns of M^* several times in succession to avoid texture dynamics turn to static but it was observed that up to three consecutive frames of the same has almost no observable impact on the result. Synthesized normalized DT sequence $C^\dagger$ which is a $(n \times t^\dagger)$ matrix can be then computed simply as: $C^\dagger = IM^\dagger$ as explained in Section 4. The last step is an addition of the average frame to each synthesized frame in the sequence that is inverse procedure to the normalization mentioned in Section 3.

7 Dynamic Texture Editing

There are several possibilities how to edit DT using this approach i.e. finely adjust overall appearance so that edited DT still looks realistic. In general texture editing is complex, very often user directed task (image editing software) and although some attempts have been made to automate this process it still remains an open problem. It is possible to replace average frame and eigen images of the analyzed DT with average frame and eigen images of another DT or even edit original average frame as a normal multispectral image. In this case characteristic features of DT such as colours and lower

clouds		23.0534			15.8105		4.4309
flame	9.6099						3.1155
fountain	29.7381		6.756				2.8560
grass	15.2763		18.4564				6.0214
ocean	34.1274	31.3391	4.5209	20.3366			6.3397
pond	14.2968		13.7776				7.4999
river		40.3987					11.9223
smoke	36.9984	15.4730		23.8452			1.7904
sparkle	9.5897						1.5501
waterfall	18.7805		21.4468			55.8353	5.0756
waterfall2		10.9446			17.6794		1.9639

Table 1: Comparison of MAD quality criterion values on the Graphcut texture database for six alternative DT synthesis methods, from the left: Method [6] (temporal cut only), [6] (spatio temporal cut), [9], [12], [1], [10] and our proposed method.

frequencies are affected by new average frame, higher frequencies by eigen images and overall dynamics is still driven by temporal mixing coefficients. Similarly it is possible to replace either matrix of eigen images or mixing coefficients (or both) with similar effect (or even only average frame) but in this case both DTs should be of a similar nature (i.g. similar movement). There are another limitations of this approach. Both textures should be rather dynamic textures than dynamic scenes (i.e. general video containing several different and in general moving objects or dynamic textures) and it is necessary to both DTs have the same dimensions. Some of the achieved results is showed in Fig.2. Apart from the above mentioned options it is also possible to apply frequency swap strategy [4] to the individual frames of the DTs as they can be considered as multispectral textures.

8 Results

For a testing purposes we used dynamic texture data sets from DynTex texture database [8] as a source of data. Each dynamic texture from this database is typically represented by 250 frames, which equals 10 seconds, long video sequence. We extracted its frames, converted, saved and used as 400×300 RGB colour images, so that $(n = 360000, t = 250)$. As test DTs were chosen: smoke, steam, streaming water, sea waves, river, candle light, detail of running escalator, sheet, waving flag, leaves, straws and branches. For a comparison with alternative synthesis methods we also tested our method on video textures from Graphcut texture database [1] [6]. This database consists of several very different textures including corresponding textures synthesized by alternative approaches. In case of this database textures very often differ from each other in dimensions. Some achieved results are showed in Fig.8. It is really hard to compare visual quality of the results of those methods exactly as robust and reliable similarity comparison even between two static textures is still unsolved problem up to now. We decided to compare differences

[1] http://www.cc.gatech.edu/cpl/projects/graphcuttextures/

between original DT and synthesized DTs of individual methods. Let the original DT O is a L_O long sequence of $W_O \times H_O$ images with S_O spectral planes and the synthesized DT S is a L_S long sequence of $W_S \times H_S$ images with S_S spectral planes then Mean Absolute Difference of the original DT and the synthesized DT (MAD) is defined as:

$$MAD = \frac{1}{LSHW} \sum_{l=1}^{L} \sum_{k=1}^{S} \sum_{j=1}^{H} \sum_{i=1}^{W} |O_{(i,j,k,l)} - S_{(i,j,k,l)}| \ , \tag{3}$$

where $\alpha = \min\{\alpha_O; \alpha_S\}$ and $\alpha \in \{L, S, H, W\}$. From the results listed in Tab.1 it is apparent that our method outperforms, with one exception, the others, in this concept. Although there is not exist any sample in used database which would offer results obtained by all alternative methods, Tab.1 clearly demonstrates certain quality of our method. Another comparison is further discussed in following Section.

Criterion (1) allows adjust the level of compression for each type of texture. Loss of the information can be expressed in the amount of energy (sum of used eigen values divided by the sum of all eigen values) which was preserved. In case of tested DTs we achieved these results: fire: 73%, clouds: 92%, flame: 78%, fountain: 82%, grass: 65%, ocean: 83%, pond: 77%, river: 80%, smoke: 93%, sparkle: 87%, waterfall: 62%, waterfall2: 90%. To demonstrate computing time requirements several examples are available in following table (in this case synthesized sequence was as long as original one).

Texture	Number of frames	Frame resolution	Analysis time	Synthesis time
clouds	61	128 × 128	10s	2s
flame	89	320 × 240	112s	4s
grass	100	224 × 144	66s	3s
smoke	32	160 × 112	4s	1s

9 Discussion

The main advantage of this method is its simplicity, efficiency and performance, using optimal methods for compression. Extremely fast synthesis can be even more efficiently performed by contemporary graphical hardware since only elementary instructions and matrix operations have to be realized. Our synthesis algorithm is less demanding than in case of most other methods. The synthesis is not restricted on number of frames to be generated, unlike [1], and it is not necessary to verify synthesized frames to prevent extremely long sequence to turn static as for example like in case [3]. However eigen analysis may cause observable loss of information (high frequencies which may occur in the original more precisely). In contrast to the sampling based DT modelling method called dynamic roller [5], our method cannot simultaneously enlarge the frames of the DT. On the other hand, this avoids possible spatial repetition of patterns which may appear in the synthesized DT produced by the dynamic roller.

10 Conclusion

We presented a novel method for fast synthesis of multispectral dynamic textures (DT). The main part of the approach is based on reduction of temporal coefficients resulted

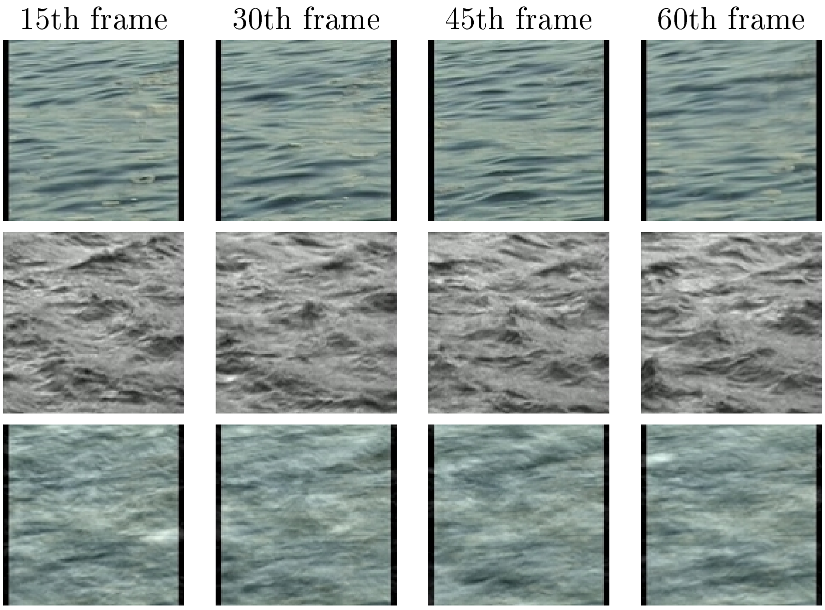


Figure 2: Several frames of original ocean DT (top row), original river DT (middle row) and edited DT (using average frame and eigen images of ocean DT and temporal mixing coefficients of river DT) (bottom row).

from DT dimensionality analysis step using the singular value decomposition which enables compress significantly the original data. This solution also enables extremely fast synthesis of arbitrary number of required multispectral DT frames, which can be even more efficiently performed by contemporary graphical hardware. We also presented several possibilities how to edit dynamic texture utilizing this approach. From many showed results it is apparent that the visual properties of the original DTs stayed preserved in the synthesized ones. We also compared our method with several existing DT synthesis and video texture generation approaches. This method avoids some problems of the alternative methods. On the other hand, proposed synthesis algorithm does not extend DT in spatial domain. Overall, this method represents interesting alternative to the existing approaches.

References

[1] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman. *Texture Mixing and Texture Movie Synthesis using Statistical Learning.* IEEE Transactions on Visualization and Computer Graphics, volume 7, (2001), 120–135.

[2] D. Chetverikov and R. Péteri. *A Brief Survey of Dynamic Texture Description and Recognition.* In Proceedings 4th Int. Conf. on Computer Recognition Systems (CORES05), Springer Advances in Soft Computing, (2005), 17–26.

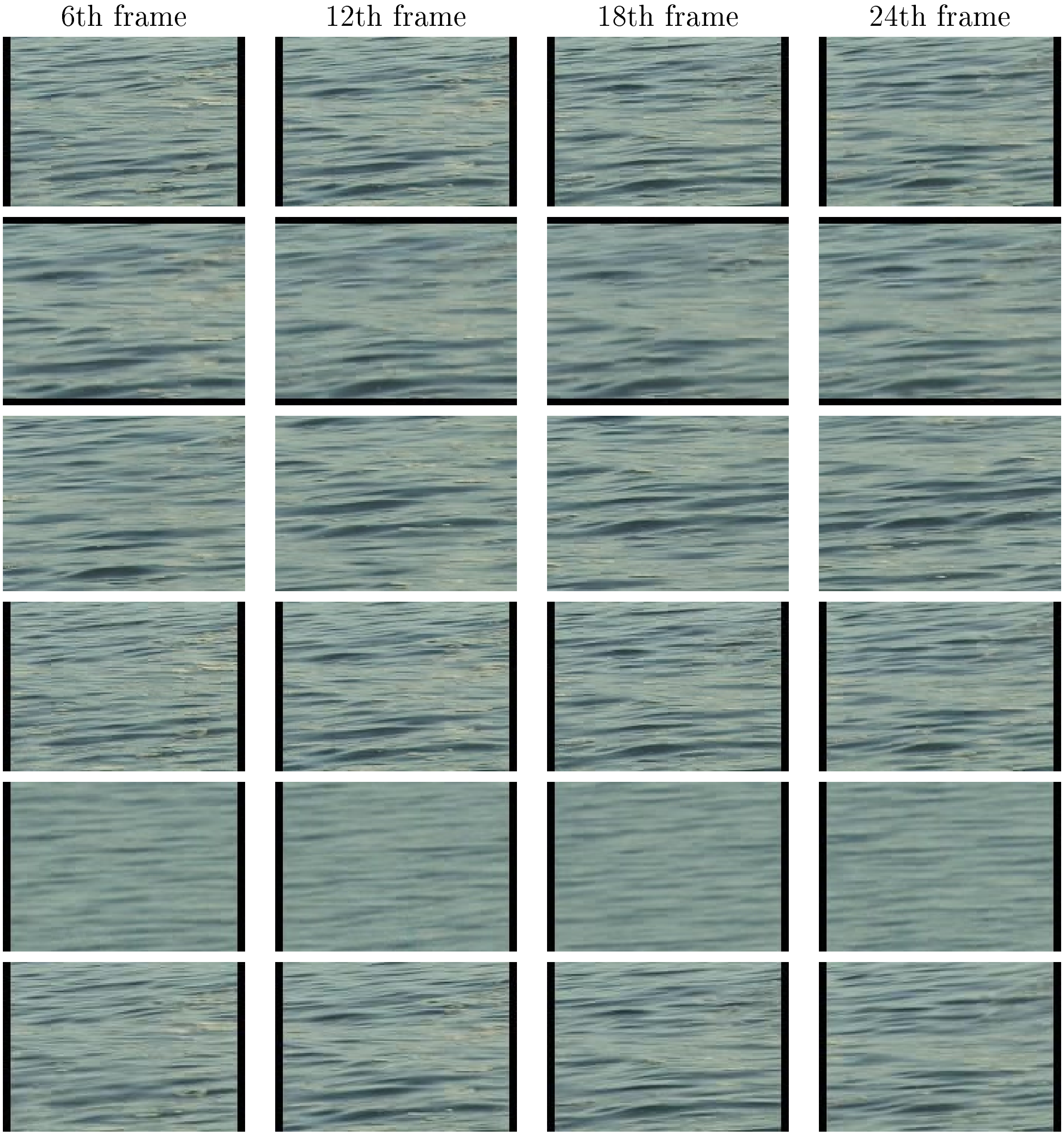


Figure 3: Several original ocean DT frames (first row) and their synthesis using the methods (topdown): [6] (temporal cut), [6] (spatio temporal cut), [9], [12], and our method.

[3] J. Filip, M. Haindl, and D. Chetverikov. *Fast Synthesis of Dynamic Colour Textures.* Proceedings of 18th International Conference on Pattern Recognition, IEEE Computer Society Press, volume 4, (2006), 25–28.

[4] M. Haindl, V. Havlíček. *Texture Editing Using Frequency Swap Strategy.* Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, volume 5702, Springer Berlin Heidelberg, (2009), 1146–1153.

[5] M. Haindl and R. Richtr. *Dynamic Texture Enlargement.* Proceedings of the 29th Spring conference on Computer Graphics (SCCG), ACM, Comenius University, Bratislava, (2013), 13–20.

[6] V. Kwatra, A. Schödl, I. Essa, and A. Bobick. *Graphcut Textures: Image and Video Synthesis Using Graph Cuts.* ACM Trans. Graph., volume 22, number 3, ACM, (2003), 277–286.

[7] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. *Dimensionality Reduction: A Comparative Review.* Tilburg University Technical Report, TiCC-TR 2009-005, (2009).

[8] R. Péteri, S Fazekas, and M. J. Huiskes. *DynTex: A Comprehensive Database of Dynamic Textures.* Pattern Recognition Letters, volume 31, number 12, (2010), 1627–1632.

[9] A. Schödl, R. Szeliski, D. Salesin, and I. Essa. *Video Textures.* Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH 2000, ACM Press/Addison-Wesley Publishing Co., (2000), 489–498.

[10] S. Soatto, G. Doretto, and Y. Wu. *Dynamic Textures.* ICCV, (2001), 439–446.

[11] M. Szummer and R. Picard. *Temporal Texture Modeling.* IEEE International Conference on Image Processing, (1996), 823–826.

[12] L. Wei and M. Levoy. *Fast Texture Synthesis using Tree-structured Vector Quantization.* Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH 2000, ACM Press/Addison-Wesley Publishing Co., (2000), 479–488.

Beta-expansions of Rational Numbers in the Quadratic Pisot Bases*

Tomáš Hejda

3rd year of PGS, email: tohecz@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Edita Pelantová, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Wolfgang Steiner, LIAFA, CNRS UMR 8079, Uniersité Paris Diderot, France

Abstract. We study the purely periodic β-expansions of rational numbers. We give an algorithm for determining the value of the function $\gamma(\beta)$ for quadratic Pisot numbers β. For numbers satisfying $\beta^2 = a\beta + b$ with b dividing a, we show a necessary and sufficient condition for $\gamma(\beta) = 1$, i.e., that all rational numbers $p/q \in [0, 1)$ with $\gcd(q, b) = 1$ have a purely periodic β-expansion.

Keywords: beta-expansions, natural extension, periodic point, purely periodic expansions, algorithm

Abstrakt. V tomto příspěvku se zabýváme čistě periodickými β-rozvoji racionálních čísel. Umíme určit hodnotu $\gamma(\beta)$ algoritmicky pro kvadratická Pisotova čísla β. Pokud je β kořen rovnice $\beta^2 = a\beta + b$, kde a je násobkem b, dokazujeme nutnou a postačující podmínku pro to, aby $\gamma(\beta) = 1$, tedy, aby všechna racionální čísla $p/q \in [0, 1)$ taková, že q je nesoudělné s b, měla čistě periodický β-rozvoj.

Klíčová slova: beta-rozvoje, přirozené rozšíření, periodické body, čistě periodické rozvoje, algoritmus

This work was presented at 15$^{\mathrm{e}}$ Journées Montoises d'Informatique Théorique held in Nancy on September 23–26, 2014 [1].

References

[1] Tomáš Hejda, Wolfgang Steiner. *Beta-expansions of rational numbers in the quadratic Pisot bases.* In 'Journées Montoises (Nancy, 2014)', 6 pp. http://jm2014.sciencesconf.org/44328

*Partially funded by Grant Agency of the Czech Technical University in Prague grant SGS11/162/OHK4/3T/14 and ANR/FWF project "FAN – Fractals and Numeration" (ANR-12-IS01-0002, FWF grant I1136).

Pseudo-Random Number Generators in Statistical Thermodynamics

Ivan Horňák[*]

1st year of PGS, email: hornaiva@fjfi.cvut.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Miroslav Virius, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. First, George Marsaglia's Diehard battery of tests completed by autocorrelation tests is used for standard testing of pseudo-random number generators. Autocorrelation and bit correlation of generators are tested by means of Random Walk Test and Autocorrelation Matrix Test. Suitability of generators to evaluating integrals by means of Monte Carlo method is also studied. The virial coefficients of the gases are calculated analytically and evaluated by means of Monte Carlo method. In this paper, it requires integration in dimensions up to three. Suitability of generators for Monte Carlo simulation in statistical physics is also tested. From an initial random disequilibrium configuration a fluid reaches equilibrium. Once the equilibrium is reached, compressibility factor is measured, and the results of the simulations are compared with the result of analytical calculation. Pseudo-random number generators are compared via statistical test.

Keywords: Mersenne-Twister, KISS, Xorshift, DIEHARD, Metropolis Algorithm, Lennard-Jones Potential

Abstrakt. Sada testů Diehard Geroge Marsaglii doplněná autokorelačními testy je použita ke standardnímu testování generátorů pseudonáhodných čísel. Korelace a bitová korelace jsou testovány Testem náhodné procházky a Testem autokorelační matice. Dále je studován vliv generátorů v Monte Carlo metodách. Virialní koeficienty plynů jsou spočítány analyticky a vyhodnoceny Monte Carlo integrací. V této práci to vyžaduje integraci od jedné do tří dimenzí. Dále jsou provedeny testy generátorů prostředky Monte Carlo simulace ve statistické fyzice. Z původní náhodné nerovnovážné konfigurace plyn dosahuje rovnováhy. Po dosažení rovnováhy je změřen kompresibilní faktor a výsledky simulace jsou porovnány s analytickým vyhodnocením. Výsledky generátorů pseduonáhodných čísel jsou porovnány ttestem.

Klíčová slova: Mersenne-Twister, KISS, Xorshift, DIEHARD, Metropolisův algoritmus, Lennard-Jonesův potenciál

1 Introduction

The question of the choice of a right pseudo-random generator is more important every day. Many possible ways how to generate pseudo-random numbers can be found, and many of them are very easy to use. However, there are some pitfalls that user has to be aware of. Many applications tolerate even bad generators, because good statistics can

[*] Special thanks go to Jaromír Kukal for his assistance

still be obtained. On the other hand, the bad choice of generator can be fatal in more complex applications.

Some standard generators have poor quality. It was mentioned in [2] that these standard generators have serious defect. Some of them are:

- Standard Perl rand
- Python random() (versions before V2.3; V2.3 and above are OK)
- Java.util.Random
- C-library rand(), random() and drand48()
- Matlab's rand
- Mathematica's SWB generator

From the example of flawed standard generators it can be seen that the testing is absolutely necessary and generators have to be used carefully.

The testing of the generator is very complex subject. There is not one test, that verifies the quality of the generator. Sets of statistical tests, which can provide complex view of the generator's qualities, were created. The perfect example is George Marsaglia's Diehard battery of tests, which can serve as good initial tool for judging generator's qualities. Diehard does not contain an autocorrelation test. Thus further testing should be performed, because correlation can strongly influence statistics. Can we choose generators, which passed Diehard, to a specific application without further doubts? We want to employ aditional tests aimed to discover suitability for application in statistical physics.

The aim of this short text is to provide testing of popular generators known for its qualities beyond Diehard battery of test and to verify if they can be used for Monte Carlo evaluation of integrals and simulation of statistical fluids. This paper is organized as follows. In the next section, fluid modeling preliminaries are briefly reviewed. Section 3 deals with various potentials, and in section (4) various approaches to generator testing are considered, and finally, we focus on tested generators and after that results of our tests are provided.

2 Fluid modeling preliminaries

The basic frame of fluid modeling preliminaries enables to study the role of pseudo-random generators by validity of results. We consider a traditional design of N particles in an abstract box of volume $V = a^3$ with periodic extension, where a is the dimensionless side of the box. Particle interactions are performed via dimensionless spherically symmetric potential function $u(r)$ satisfying $u(1) = 0$. Reduced virial coefficients can be expressed as:

$$B_2^* = 3\int_0^\infty f(r_{12})r_{12}^2 \mathrm{d}r_{12}, \tag{1}$$

$$B_3^* = 6\int_0^\infty \int_0^\infty \int_{r_{12}-r_{13}}^{r_{12}+r_{13}} f(r_{12})f(r_{13})f(r_{23})r_{12}r_{13}r_{23}\mathrm{d}r_{12}\mathrm{d}r_{13}\mathrm{d}r_{23}. \tag{2}$$

using dimensionless temperature $T^* > 0$ and dimensionless Mayer function

$$\mathrm{f}(r) = \exp\Big(-\frac{u(r)}{T}\Big) - 1. \tag{3}$$

Supposing low density $\rho = N/V$, we can approximate the virial expansion of the compressibility factor

$$z = 1 + \sum_{i=2}^{\infty} B^*_{i-1}\rho^{i-1} \tag{4}$$

by its first three terms as

$$z = 1 + B^*_2\rho + B^*_3\rho^2. \tag{5}$$

Formulas (1), (2) will be used for testing of the suitability of generators for Monte Carlo evaluation of integrals. Formula (5) presents theoretical value of compressibility factor for evaluation of Monte Carlo simulation.

The Metropolis method of importance sampling is used to reach equilibrium from initial random non-equilibrium configuration. Dimensionless potential energy is expressed as

$$U(\mathbf{r}_1, ..., \mathbf{r}_N) = \sum_{i<j} U(r_{ij}) \tag{6}$$

where $r_{ij} = ||\mathbf{r}_i - \mathbf{r}_j||$, and $\mathbf{r}_i \in R^3$ is the position vector of i^{th} particle. A random perturbation of a particle position is performed in every step of the Metropolis algorithm. The new configuration is accepted with probability

$$p = \min(1, \exp(-\Delta U/T^*)) \tag{7}$$

where $\Delta U = U_{\text{new}} - U$ is the change of dimensionless potential energy. Once the equilibrium is reached, compressibility factor is measured by means of Metropolis algorithm using

$$z = 1 - \frac{2W_{\mathrm{f}}}{3NT^*} \tag{8}$$

where

$$W_{\mathrm{f}} = \frac{1}{2}\sum_{i=1}^{N} \mathbf{r}_i \frac{\partial U}{\partial \mathbf{r}_i} \tag{9}$$

is a force virial.

The resulting values of z form a statistical sample which is easy to compare with its theoretical value (5).

3 Included models

Three models of particle interaction are used for evaluation of virial coefficients by integration and for fluid modeling testing. Lennard-Jones model (LJ) is selected as a representative of a traditional physically motivated potential expressed as:

$$u(r) = 4(r^{-12} - r^{-6}). \tag{10}$$

Table 1: Virial coefficients of Lennard-Jones fluid

T^*	B_2^*	B_3^*
1	-2.538081	0.42968
3	-0.115234	0.35230
5	0.243344	0.31506

Table 2: Virial coefficients of Constrained Potential Fluid

T^*	B_2^*	B_3^*
1	0.207277	0.010017
3	0.078072	0.000567
5	0.048065	0.000134

Its virials can not be evaluated analytically, so numerical integration is necessary. The obtained values of B_2^*, B_3^* in agreement with [1] are listed in Tab. 1. Considering the real models often do not permit analytical evaluation of virial coefficients, we establish trivial models of particle repulsion permitting an analytical evaluation of virial coefficients. The first novel model called Constrained Potential Fluid (CP) with dimensionless potential

$$u(r) = \max(1-r,0) \tag{11}$$

is introduced here. We can directly calculate

$$B_2^* = 1 - 3T^* + 6(T^*)^2 + 6(T^*)^3(Q-1) \tag{12}$$

and

$$\begin{aligned} B_3^* = \frac{5}{8} - \frac{15}{4}T^* + \frac{3(T^*)^2}{4}(13-2Q) + \frac{3(T^*)^3}{2}(25Q-6) - \frac{243(T^*)^4}{4}Q + \\ \frac{9(T^*)^5}{4}(51Q-16) + \frac{3(T^*)^6}{8}(160-537Q+384Q^2-7Q^3) \end{aligned} \tag{13}$$

where $Q = \exp(-\frac{1}{T^*})$. The obtained values of B_2^*, B_3^* are listed in Tab. 2.

The last model included is Logarithmic Potential Fluid (LOG) with dimensionless potential

$$u(r) = \max(-\text{ln}r, 0). \tag{14}$$

Using (1), we can express the second reduced virial coefficient of LOG as:

$$B_2^* = (1+3T^*)^{-1}. \tag{15}$$

LOG permits analytical evaluation of the third virial coefficient only for dimensionless temperatures $T^* = 1/n$, where $n \in N$. Values of virial coefficients for three dimensionless temperatures are listed in Tab. 3.

Main advance of developed models (CP , LOG) is its similarity to hard sphere model for $r > 1$, but they can be used for simulation via standard Monte Carlo techniques.

Table 3: Virial coefficients of Logarithmic Potential Fluid

T^*	B_2^*	B_3^*
1/5	0.6200	686005/3139136
1/3	0.5000	1313/10500
1	0.2500	97/5040

4 Generator testing

4.1 Tests of pseudo random number generators

The goal of this text is to make generator testing very thorough. For this reason, Diehard battery (DIEHARD) [3] battery of tests is used. DIEHARD was presented by George Marsaglia in 1995. In order to provide tests that are more stringent than usual easy-to-pass tests, DIEHARD combines various challenging tests. The battery is very popular and it is often considered to be a standard in good quality generator testing.

DIEHARD lacks a test aimed to detect correlation. We think this is the flaw of the battery. Correlation can be a serious defect of a generator. Recall that even some of good quality generators are suspected to have a problem with correlation.

Considering the fact DIEHARD discovered no difference between the generators, we focused on autocorrelation testing which is very close to movement simulation in pseudo-experiment. Random Walk Test (RWT) was used for correlation testing.

The movement was simulated by m-step random walk in R^n according

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{e}_k, \tag{16}$$

where $\mathbf{x}_0 = 0$, $\mathbf{e}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$

Under the hypothesis H_0: Independence of $\mathbf{e}_k$ realization, the criterion $A = ||\mathbf{x}_m||_2^2/m$ has χ_n^2 distribution. The simulation and testing was performed for multiple values of m and n at significance level 0.05.

Digit correlation were studied by means of Autocorrelation Matrix Test (AMT) using representation of depth as

$$x_k = \sum_{j=1}^{N} b_{kj} 2^j. \tag{17}$$

Using statistical sample x_1, x_2, ..., x_N, we tested bit autocorrelation as follows: Every individual test operates with i_{th} and j_{th} bits with time delay $t \in N_0$. We study relationship $b_{k,i}$ and $b_{k+t,j}$ for $k = 1, ..., m - t$ in the term of χ^2-test for adequate 2×2 contingency table with one degree of freedom on critical level 0.05. Previous tests are useful in general case. Additional tests concern with suitability of generators for application in statistical thermodynamics.

4.2 Testing on virial integrals

Virial coefficients are calculated analytically and numerically. For the case of LJ, where analytical solution is not possible, values are compared with other sources [1]. As can be seen from (1) and (2), the calculations of the second and the third reduced virial coefficients are evaluations of integrals up to three dimensions. These integrals are evaluated by means of Monte Carlo integration. Stratified Sampling method of integration is used for $n = 1$, $n = 2$ or $n = 3$ with the number of segments N^n for $N = 10$. Results of multiple simulations ($M = 100$) are compared with theoretical value. Supposing normal distribution, we test hypothesis H_0: $\mu = \mu_0$ on a significance level 0.05 using one sampled two sided t-test. Here, μ_0 is value of virial coefficient calculated using formulas (1) and (2), respectively.

4.3 Fluid modeling testing

Model is realized for isochoric NVT ensemble with periodic boundary conditions, which reaches equilibrium from initial random non-equilibrium configuration. Once the equilibrium is reached, a sample of size 10^4 of z values is collected. The equilibrium is studied by 20 samples from multiple runs of various seeds connected into one. This sample is then used to test random number generator.

Supposing normal distribution, we test hypothesis H_0: $\mu = \mu_0$ on a significance level 0.05 using one sampled two sided t-test. Here, μ_0 is theoretical value of compressibility factor. We performed test 100 times and counted the number of successful runs.

5 Generators

5.1 Mersenne-Twister

Generator Mersenne-Twister was first introduced in publication [5] Up to date, it is one of the most used pseudo-random number generators. For a particular choice of parameters it has extremely long period of $2^{19937} - 1$.

A downside is its complexity of generation - it is often considered to be too elaborate. The period, which is one of the advantages of the generator, is sometimes considered to be unnecessarily long.

Mersenne-Twister generator is very well described. Therefore, we will not concern with it in this work.

5.2 Xorshift generator

Mersenne-Twister has a very good reputation, but it is maybe too complicated for every day use. There is an ambition to create a more simple, yet good quality generator. George Marsaglia described Xorshift generator in [4]. Its generation is based repeated use of a simple computer construction: exclusive-or of a computer word with a shifted version of itself.

Combining such Xorshift operations for various shifts and arguments provides extremely fast and simple RNGs that seem to perform very well in tests of randomness.

Xorshift generators are based on three Xorshift operations. Marsaglia stated, that there are over hundred of possible combinations of the numbers of shifted bits. Combination $[17, 13, 5]$ is used.

5.3 KISS generator

KISS is primarily intended for scientific applications such as Monte Carlo simulations. As mentioned in [6] advantages of KISS are: It is proposed by well known and respected authors, it has a reasonably long but not excessive period (about 2^{123}), and it does not need to warm up.

KISS (32-bit version) consist of combination of four subgenerators of three kinds:

- one linear congruential generator modulo 2^{32}, $a = 69069$, $b = 1234567$
- one Xorshift generator (combination $[17, 13, 5]$),
- two multiply-with-carry generators modulo 2^{16}

With its simplicity which does not deteriorate quality of pseudo random numbers, KISS generators become very popular and they find new applications very often.

Note that Xorshift generator is one of the four subgenerators of KISS. Combining different RNGs is now considered to be a sound practice in designing good RNGs by many experts in the field. The flaws of one generator are likely to be compensated by others generators. For further information about KISS and his subgenerators, [6] is recommended.

6 Results

Selected generators are tested as follows: first, standard tests are used, then testing by Monte Carlo integration is performed, and, finally, testing by means of simulation in statistical physics is studied. The main idea behind testing is to study relation between standard and advance testing related to integration and simulations.

6.1 Standard testing

Tab. 4 consists of information, whether the generator passes the test (1/0 - pass/fail). Here, AMT and RWT still stand for Autocorrelation Matrix Test and Random Walk Test, respectively. BRT represents Binary Rank Test and DIEHARD-BRT means Diehard battery of tests excluding Binary Rank Test.

Mersenne-Twister passes several stringent statistical tests, including Diehard. This result is in a compliance with [5]. When dealing with KISS generator, the same results are obtained. These two generators pass RWT and AMT. Testing by means of BRT and AMT reveals a serious flaw of Xorshift. From Tab. 4 it can be seen that DIEHARD excluding BRT makes no difference between generators. BRT and AMT reveals insufficiency of Xorshift generator. The fail of Xorshift generators in BRT was previously described in [4].

The goal of successive testing is to perform advance testing of generators beyond standard tests.

Table 4: Standard generator testing

model	DIEHARD - BRT	BRT	RWT	AMT
Mersenne-Twister	1	1	1	1
KISS	1	1	1	1
Xorshift	1	0	1	0

6.2 Monte Carlo integration

Considering just one dimensional Monte Carlo integration, no difference among generators appears, ie all p-values are above 0.05. The different situation arises when dealing with 3D integration. From Tab. 5 we can see that Xorshift can not be used for 3D integration.

Except for LJ Xorshift failed in evaluating of B_3^*; therefore, we conclude that Xorshift can not be used for multiple dimensions evaluation in general.

The fail in case of CP and LOG can be explained by the fact that the potentials are not smooth and by effect of zero potential like in the case of hard spheres fluid.

Table 5: Testing by 3D integration (p-values)

model	T^*	Mersenne-Twister	KISS	Xorshift
LJ	1	0.86	0.72	0.62
	3	0.91	0.86	0.91
	5	0.89	0.93	0.92
CP	1	0.95	0.91	0.04
	3	0.92	0.89	0.01
	5	0.90	0.84	0.03
LOG	1/5	0.95	0.91	0.04
	1/3	0.92	0.89	0.01
	1	0.90	0.84	0.03

6.3 Fluid modeling testing

When dealing with simulation, the situation is different from the case of Monte Carlo integration. Simulation were performed 100 times. We get p-value of every run performed. Significance level is 0.05. In the pursuit of objectivity, we perform 100 runs, and we study the number of successful runs with no significant difference between theory and experiment. Tab. 6 contains frequency of successful results. From Tab. 6 it can be seen that results are comparable; Although, KISS showed the best results. On the the hand, Xorshift have the worst results. Pair t-test is used to identify significant difference in efficiency between pseudo-random number generators. KISS outperformes Mersenne-Twister. No significance difference between Xorshift and Mersenne-Twister is discovered. The same situation arises when dealing with KISS and Xorshift.

Table 6: Testing via Monte Carlo simulation (passing frequency [%])

model	T^*	Mersenne-Twister	KISS	Xorshift
LJ	1	60	69	50
	3	56	66	55
	5	52	65	51
CP	1	57	68	52
	3	28	68	55
	5	29	70	54
LOG	1/5	57	65	54
	1/3	53	68	51
	1	51	69	49

7 Conclusion

Tests of popular pseudo-random number generators have been made. Diehard battery of tests has been used to determine the quality of generator. Flaws of some generators have been discovered. Xorshift generator do not pass Binary Rank test in Diehard and it fails in autocorrelation testing performed by Autocorrelation Matrix test. The suitability of pseudo-random number generators for integral evaluation in three or less dimension was tested Generators have been tested via fluid modeling. It has been shown, that some generators are not suitable for physical simulation even if they pass Diehard. The importance of additional testing beyond DIHARD was confirmed.

References

[1] R. Caligaris and A. Rodriguez. *Second and third virial coefficients for the lennard-jones potential.* Molecular Physics: An International Journal at the Interface Between Chemistry and Physics **17** (1971), 1131–1132.

[2] D. Jones. *Good Practice in (Pseudo) Random Number Generation for Bioinformatics Applications.* University College London, URL: `http://www0.cs.ucl.ac.uk/staff/d.jones/GoodPracticeRNG.pdf`, (2010).

[3] G. Marsaglia. The marsaglia random number cdrom including the diehard battery of tests of randomness. `http://www.stat.fsu.edu/pub/diehard/`. Accessed: 2014-08-30.

[4] G. Marsaglia. *Xorshift rngs.* Journal of Statistical Software **8** (2003), 1–6.

[5] M. Matsumoto and T. Nishimura. *Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator.* ACM Transactions on Modeling and Computer Simulations **8** , 3–30.

[6] G. Rose. *Kiss: A bit too simple.* IACR Cryptology ePrint Archive **2011** (2011), 7.

Headway Distribution for TASEP with Parallel Updates*

Pavel Hrabák

5th year of PGS, email: `pavel.hrabak@fjfi.cvut.cz`
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Milan Krbálek, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Contribution focuses on four different updating procedures of the totally asymmetric simple exclusion process (abbr. TASEP), namely the fully-parallel, forward-sequential, backward-sequential updates [5], and generalized backward update [2]. The goal is to introduce the distance- and time-headway distribution for those updates to extend the random-sequential update studies in [4] and [3]. All systems are considered on the circle of L sites. Corresponding Markov processes are investigated in the large L approximation. Using the car-oriented mean-field approximation [6] or mapping to the mass transport process [7], the stationary distribution is derived, from which the time-headway distribution can be easily obtained. Furthermore, the derivation of time-headway distribution of fully-parallel update [1] has been extended to other parallel updates mentioned in this contribution.

Results presented within the conference Doktorandské Dny are being prepared for the submission to *Journal of Physics A: Mathematical and Theoretical.*

Keywords: TASEP, parallel updates, headway distribution

Abstrakt. Příspěvek se zaměřuje na čtyři různá updatovací schémata totálně asymetrického jednoduchého vylučovacího procesu (TASEP), jmenovitě plně-paralelné, dopředný, zpětný update [5] a zobecněný update [2]. Cílem je představit rozdělení vzdálenostních a časových rozestupů těchto updatů a rozšířit tak výsledky pro náhodný update z [4] a [3]. Všechny systémy jsou uvážovány na kruhové mřížce tvořené L pozicemi. Příslušné markovské procesy jsou zkoumány ve stacionárním stavu. Pomocí car-oriented mean-field aproximace [6] nebo převedením na mass transport process [7] je odvozeno stacionární rozdělení, pomocí něhož lze snadno získat rozdělení vzdálenostních rozestupů. Dále, odvození časových rozestupů pro plně-parallelní update [1] je rozšířeno pro ostatní uvažované updaty.

Výsledky prezentované na konferenci Doktorandské Dny jsou připravovány pro odeslání do *Journal of Physics A: Mathematical and Theoretical.*

Klíčová slova: TASEP, paralelní update, rozdělení rozestupů

References

[1] D. Chowdhury, A. Pasupathy, and S. Sinha. *Distributions of time- and distance-headways in the nagel-schreckenberg model of vehicular traffic: Effects of hindrances.* European Physical Journal B **5** (1998), 781–786.

*This work has been supported by the CTU grant SGS12/197/OHK4/3T/14

[2] A. E. Derbyshev, S. S. Poghosyan, A. M. Povolotsky, and V. B. Priezzhev. *The totally asymmetric exclusion process with generalized update.* Journal of Statistical Mechanics: Theory and Experiment **2012** (2012), P05014.

[3] P. Hrabák and M. Krbálek. *Distance- and time-headway distribution for totally asymmetric simple exclusion process.* In '14th Euro Working Group on Transportation', J. Zak, (ed.), volume 20 of *Procedia - Social and Behavioral Sciences*, Elsevier Science B.V. (2011), 406–416.

[4] M. Krbálek and P. Hrabák. *Inter-particle gap distribution and spectral rigidity of totally asymmetric simple exclusion process with open boundaries.* Journal of Physics A: Mathematical and Theoretical **44** (2011), 175203–175224.

[5] N. Rajewsky, L. Santen, A. Schadschneider, and M. Schreckenberg. *The asymmetric exclusion process: Comparison of update procedures.* Journal of Statistical Physics **92** (1998), 151–194.

[6] A. Schadschneider and M. Schreckenberg. *Car-oriented mean-field theory for traffic flow models.* Journal of Physics A: Mathematical and General **30** (1997), L69–L75.

[7] R. K. P. Zia, M. R. Evans, and S. N. Majumdar. *Construction of the factorized steady state distribution in models of mass transport.* Journal of Statistical Mechanics: Theory and Experiment **2004** (2004), L10001.

Mathematical Modeling of the Swelling Behavior of Articular Cartilage*†

Jana Hradilová

3rd year of PGS, email: hradilova.jana@email.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Zdeněk Převorovský, Department of Impact and Waves in Solids
Institute of Thermomechanics, ASCR

Kay Raum, Julius Wolff Institut, Charité – Universitätsmedizin Berlin

Abstract. Articular cartilage changes its dimensions and volume when bathed in saline solutions of different concentrations. The change of dimensions, swelling or shrinkage, depends on its fixed-charge density, stiffness of its collagen-proteoglycan matrix, and the ion concentrations in the interstitium. These parameters are fundamental of the electro-mechano-chemical behavior of the cartilage. The investigation of degenerative state of the articular cartilage was performed on a small-animal model. The mathematical model has beed developed and results of experiments were compared with simulations to explain the observed phenomena.

Keywords: Articular cartilage, Osmotic swelling, Ultrasound, COMSOL

Abstrakt. Kloubní chrupavka mění své rozměry a objem při ponoření do externích solných lázní o různých koncentracích soli. Změna rozměrů, otékání a splaskávání, závisí na hustotě pevných nábojů v chrupavce, na tuhosti pevné matice, tvořené především kolagenovými vlákny a proteoglykany, a na koncentracích iontů v mezibuněčném prostoru. Tyto parametry jsou základem elektro-mechano-chemackého chování kloubní chrupavky. Výzkum degenerativních stavů chrupavky byl proveden na modelu malých zvířat. Byl vyvinut matematický model a výsledky experimentů byly porovnány se simulacemi pro vysvětlení pozorovaných jevů.

Klíčová slova: Kloubní chrupavka, Osmotické otékaní, Ultrazvuk, COMSOL

1 Introduction

Articular cartilage (AC) is a hydrated soft tissue covering the bone ends and aiding the joint in absorbing mechanical shock. It also provides joints with lubrication allowing smooth motion and maintains efficient bearing system for the body. Structurally, articular cartilage comprises three main structural components: water, collagen fibrils and proteoglycan macromolecules. A proteoglycan monomer consists of a protein core and glycosaminoglycan chains that carry negative charges. This is a fundament of the electro-mechano-chemical behavior of the cartilage. The negatively charged groups of proteoglycans attract cations and water into the tissue to generate a substantial Donnan osmotic

[*] This work has been supported by grants SGS12/197/OHK4/3T/14of the Czech Ministry of Education, IT ASCR No. 904150 and with institutional support RVO: 61388998

[†] Publication in a peer-reviewed periodical is expected.

pressure [1] and cause cartilage swelling. However, the swelling is balanced by the elastic force of collagen fibril network.

Osteoarthrosis (OA) is one of the most common musculoskeletal diseases which affects articular cartilage and joints, typically knees. OA causes significant constraints in the quality of patient's life because of pain and limitations of mobility, which results in considerable loss of working ability and economic hardship. Possible causes of OA are aging, injury, extensive loading and obesity. OA causes progressive degenerative changes to cartilage surface, matrix and subchondral bone. In addition, the ability of cartilage to repair itself is limited. One of the first symptoms of OA is cartilage tissue softening, followed by cartilage fibrillation and, in later stages, disruption of the collagen network [2]. While early changes in the cartilage might still be reversible, treatment in later stages is possible only with intra-articular injection or surgically, often only with short-term effect. However, the first subjective signs (e.g. pain) occur in advanced, irreversible, stages of cartilage tissue damage. To avoid advanced stages of OA, the need of early changes (e.g. cartilage surface fibrillation) detecting methods arises. To date, early signs of cartilage degeneration are commonly detected by histologic evaluation; this is not possible to do noninvasively with current imaging methods (radiography, magnetic resonance imaging) because of their insufficient resolution [4].

High-frequency ultrasound (US) has the potential to provide sufficient information about early degenerative cartilage changes noninvasively. Similar to the histologic analysis, high-frequency US backscatter signals allow distinct evaluation of signals originating from the cartilage surface, tissue matrix and subchondral bone boundary [2].

It has been been suggested that the quantification of the swelling effects in articular cartilage can be used to characterize the degenerative changes associated with OA. This is due to increased water content and swelling of the tissue induced by degeneration of collagen fibrils. Our aim is to use ultrasound to characterize AC (normal and degenerated) in a nondestructive way by measuring the transient swelling behavior induced by changing the concentration of bathing solution. To achieve this, we investigate Dunkin-Hartley guinea pigs (one control and several treated groups of animals) which develop OA naturally by aging during first year of their life.

To better understand the underlying mechano-electro-chemical processes we developed a one-dimensional mathematical description of the cartilage swelling problem based on triphasic theory [1] and utilize the commercial finite-element code COMSOL Multiphysics to simulate the steady state and transient behavior of the AC in response to changing salt concentration of the surrounding solution. The simulation results are compared with available experimental data to show the accuracy of the model.

2 Materials and methods

Normal saline, or physiological saline, is a solution isotonic to body fluids. It is solution of $9.0g$ of NaCl dissolved in one liter of sterile water, it means $0.15mol/L$ NaCl (mole NaCl per liter). At the physiological state of $0.15mol/L$ NaCl, cartilage is in a swollen state, with a swelling pressure resisted by the elastic stress in the collagen-proteoglycan silod matrix. Articular cartilage changes its dimensions, volume, and weight when the ion concentration in the bathing solution is changed. For an unloaded specimen, the tissue dimensions decrease with increasing NaCl concentration; this decrease approaches an asymptote at the concentrations as high as $2.5mol/L$ NaCl [1].

Articular cartilage specimens: Experiments were performed on fresh medial thibial plateau (left and right) of three-months-old guinea pigs. Samples were visually classified as healthy.

Ultrasound system: Commercial high-frequency ultrasound imaging system Ultrasonix was used to monitor transient swelling behavior of small animal AC noninvasively. Swelling was induced by changing the concentration of bathing saline. The system included a $12MHz$ focused ultrasound transducer $L40 - 8/12Linear$, with focal range $0.2 - 3cm$, which allows to scan a field of $16mm$.

Ultrasound measurements: Fresh cartilage specimens were placed rigidly on the bottom of the container and then submerged in a $0.15mol/L$ saline solution. The ultrasound transducer was fixed at a position over the central part of the medial thibia plateau with the focal zone of its beam located inside the AC tissue. The AC specimen was tested at the temperature of 36°C ± 0.5°C for all of the following procedures. After $30min$ the AC specimen was supposed to reach the equilibrium, the $0.15mol/L$ saline was replaced with $0.3mol/L$ saline within $30s$ and the AC was monitored with ultrasound for next $30min$. The AC was supposed to shrink based on the Donnan theory of osmotic pressure [1]. The echo signals reflected from the surface of the AC and from the AC/bone interface were continuously recorded with a sampling period of $3s$. The bathing solution was further changed from $0.3mol/L$ saline back to $0.15mol/L$ physiological saline to see the backward effect. Collected data were then postprocessed using the methods of signal analysis.

3 Mathematical model and simulations

Swelling of AC depends on its fixed charge density (i.e. density of the fixed charges attached to the extracellular matrix), the stiffness of its collagen-proteoglycan matrix, and the ion concentrations in the interstitium. Simulation of AC swelling/shrinkage characteristics require the mathematical formulation of coupled chemo-electro-mechanical mechanisms between the AC and a surrounding bath solution. Following mathematical description couples the Nernst-Planck equation, Poisson's equation and mechanical deformation equation for one dimensional case. The Nernst–Planck Equation describes the concentrations of chemical species in a fluid medium. It extends the Fick's law of diffusion for the case where the particles diffuse under the influence of both ionic concentration gradient and electrostatic forces:

$$\frac{\partial c_i(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D_i(x)\frac{\partial c_i(x,t)}{\partial x}\right) - \frac{\partial}{\partial x}\left(D_i(x)\frac{z_i e}{k_B T}c_i(x,t)E(x,t)\right), \tag{1}$$

where t, D_i, c_i, z_i, e, k_B, T and E are, respectively, time, the diffusion coefficient of the i-th ion, the concentration of the i-th ion, valence of the i-th ion, elementary charge, the Boltzmann constant, temperature and electric field. In our case $i \in \{Na, Cl\}$. In this equation, first term represents the diffusive flux due to concentration gradient, the second term describes the migration flux due to electric potential gradient. Second term

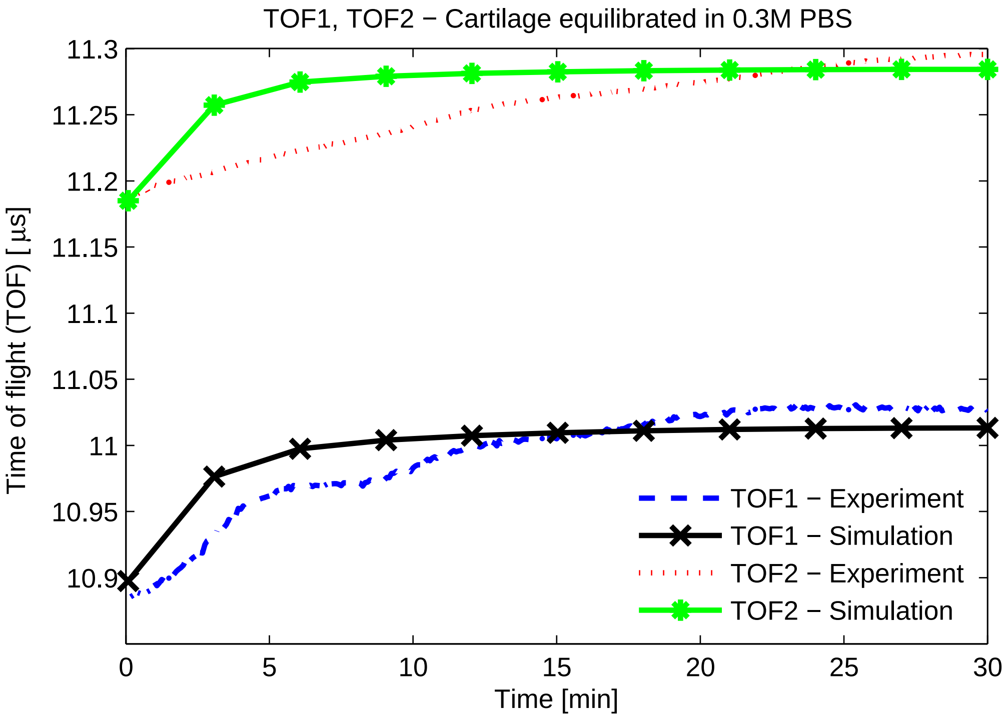


Figure 1: Times of flight for cartilage surface (TOF1), cartilage/bone interface (TOF2) and corresponding numerical simulations for cartilage equilibrated in $0.30 mol/L$ solution.

is coupled with the equation for electric field:

$$\frac{dE(x,t)}{dx} = \frac{\rho(x,t)}{\varepsilon_0 \varepsilon_r} = \frac{F}{\varepsilon_0 \varepsilon_r} \left(c_{Na}(x,t) - c_{Cl}(x,t) - c_F \right), \tag{2}$$

where ρ, ε_0, ε_r, F and c_F are the charge density in the cartilage, the dielectric constant of the vacuum, the relative dielectric constant of the solvent, the Faraday constant and the fixed-charge concentration in the cartilage, respectively. The cartilage swelling or shrinking is described by the following equation which we derived from the triphasic theory of [1]:

$$de_x(x,t) = -\frac{RT}{(\lambda_s + 2\mu_s)} \sum_{i \in \{Na, Cl, H_2O\}} dc_i(x,t), \tag{3}$$

where e_x is the swelling strain describing relative deformation, R is universal gas constant, T is temperature, and λ_s and μ_s are Lamé parameters for solid extracellular matrix. Calculation of H_2O concentration has been done using the relation for the density of saline water which depends on temperature, pressure and salinity, [3].

4 Results and discussion

Figures 1 and 2 show the progress of the swelling experiment. Figure 1 refers to the period when the saline bath was changed from the physiological concentration of $0.15 mol/L$

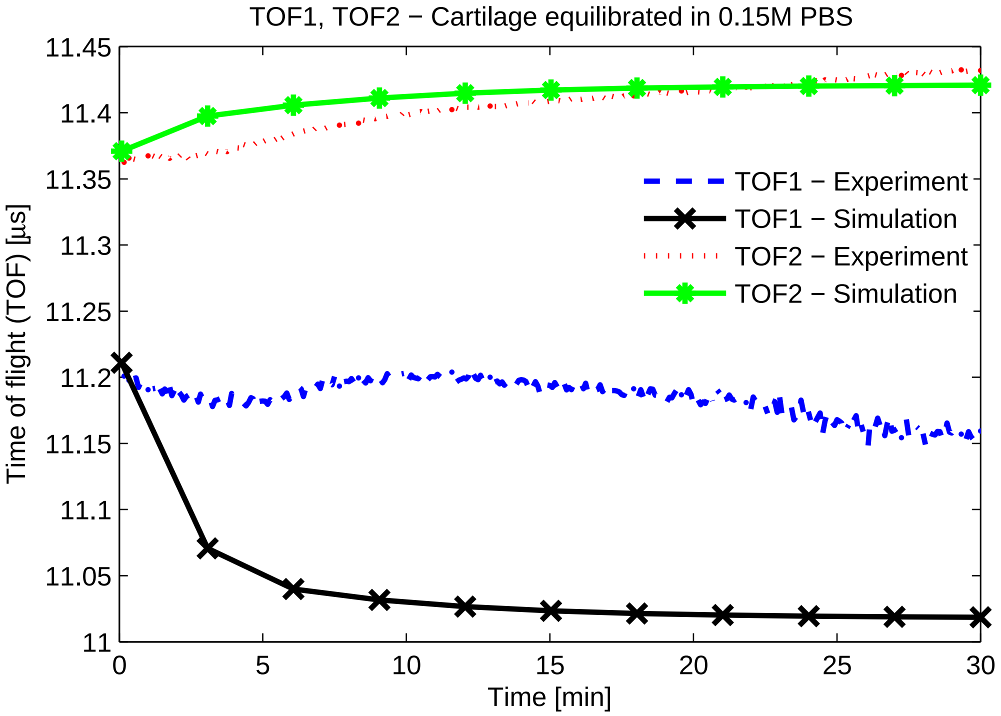


Figure 2: Times of flight for cartilage surface (TOF1), cartilage/bone interface (TOF2) and corresponding numerical simulations for cartilage equilibrated in $0.15 mol/L$ solution.

to $0.30 mol/L$. Figure 2 refers to the period when the saline bath was changed from the higher concentration of $0.30 mol/L$ back to $0.15 mol/L$.

Figure 1 shows different times of flight (TOF): TOF1 corresponds to the echoes reflected from the articular cartilage surface during the equilibration in the saline bath. The shift of TOF1 represents the thickness change of cartilage tissue. We can conclude that the articular cartilage shrink over the time. TOF2 describes behavior of echoes from the cartilage/bone interface. The shift of TOF2 represents combination of, first, increasing salt concentration inside the cartilage and, second, shrinking of the cartilage. These phenomena lead to the change of the speed of sound in the cartilage tissue. We can see from the Figure 1 that although the speed of sound is increasing over the time period, the cartilage shrinking causes TOF2 to increase. Both variables, TOF1 and TOF2, agree reasonably well with numerical simulations.

The same meaning of TOF1 and TOF2 holds for Figure 2. From TOF1 we can conclude that the cartilage does not swell as much as mathematical model suggests. This could be attributed to the change in effective values of the Lamé coefficients in Equation (3). The shift of TOF2 represents a combination of, first, decreasing salt concentration inside the cartilage and, second, swelling of the cartilage tissue. From the slightly increasing TOF2 curve in Figure 2 we can see that decreasing speed of sound has a greater influence than swelling of the cartilage tissue. TOF2 corresponds well with the mathematical model.

References

[1] W.M. Lai, J.S. Hou, V.C. Mow. *A triphasic theory for the swelling and deformation behaviors of articular cartilage.* J. Biomech. Eng. **113** (1991), 245–258.

[2] M. Schöne, N. Männicke, M. Gottwald, F. Göbel, K. Raum. *3-D high-frequency ultrasound improves the estimation of surface properties in degenerated cartilage.* Ultrasound Med Biol. **39(5)** (2013), 834–844.

[3] W. Wilson, D. Bradley. *Specific volume of sea water as a function of temperature, pressure and salinitz.*Deep-Sea Research. **15** (1968), 355–363.

[4] Y.P. Zheng, J. Shi, L. Qin, S.G. Patil, V.C. Mow, K.Y. Zhou. *Dynamic depth–dependent osmotic swelling and solute diffusion in articular cartilage monitored using real–time ultrasound.* Ultrasound Med Biol. **30(6)** (2004), 841–849.

Dynamic Model and Requirements Engineering

Radek Hřebík

3rd year of PGS, email: Radek.Hrebik@seznam.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Vojtěch Merunka, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This contribution deals with the initial phase of software developing process and suggests some improvement in this phase to make the whole developing process more effective. The improvements are based on dynamic model of requirements. The model representing the application requirements is created with the Business Object Relation Modeling tool. The contribution deals with the representation of this model as a finite state machine. The aim is to use such model based on object-oriented paradigm and finite state machines for improvements in task verification and find a way to improve software development process.

Keywords: Requirements, dynamic model, finite state machine, BORM, verification, database

Abstrakt. Tento příspěvek se zabývá počátečné fází procesu vývoje softwaru a navrhujě některá vylepšení v této oblasti. Navrhovaná vylepšení při vývojů softwaru vychází z dynamického modelu, který reprezentuje požadavky na aplikaci. K vytvoření dynamického modelu požadavků je využito nástroje podporujícího metodu BORM. Příspěvek zkoumá možnost reprezentace modelu požadavků v podobě konečného automatu. Cílem je využití modelu požadavků založeného na prinicpech objektově orientovaného programování a konečných automatů pro zlepšní v oblasti verifikace zadání a nalezení cesty pro další zlepšení v procesu vývoje softwaru.

Klíčová slova: Požadavky, BORM, konečný automat, dynamický model, verifikace

1 Introduction

Research is based on connection of requirements engineering, dynamic model representation by finite state machines. The key role in the research plays requirements engineering. This implies the applying of scientific knowledge to ensure that requirements are fully correct. The obvious definition of requirements engineering is discipline based on understanding software requirements. Another possible definition comes from Laplante [11] and defines requirements engineering as process of eliciting, analyzing, documenting, validating and managing requirements. Requirements affect the whole software development process and can be changed. Software life-cycle in connection with requirements is detail discussed in [2].

Naturally it has to be confirmed with Laplante [11] that requirements have to be documented. Documented requirements can lately serve as an evidence. Requirements document has to be processed and readable. Requirements should be clear, precise and unambiguous. If they are not well understood the system will not meet expectations and the final version of program will be probably not delivered on time and the costs will be much higher than originally expected. The costs of fixing errors in later development phases may be very high as discussed for example in [18]. All these problems can be prevented by using knowledge of requirements engineering. The research is based on representing software requirements as dynamic model.

2 Tools for requirement engineering

The need of requirement engineering is indisputable. Tools are available and commonly helping requirements engineers in their work. Often they determined for a widely range of use in the whole developing process. The mainstream tools are used in a standard way. Next are briefly mentioned some commercial tools as Jama Software, DOORS and CaliberRM. As the research is focused not only to market tools, working with standard ways of representing requirements, there is also mentioned ModellicaML representing very interesting project of one Swedish PhD student. Project represents using requirements to improve the software developing process.

The first selected tool Jama Contour represents web application helping users to manage the entire requirements management life-cycle. It enables users to collaborate with the requirements, reuse and trace them. The changing of requirements and testing management is also available. Application is adapted to any process whether waterfall, iterative or agile. The Contour enable user to publish requirements documents as a software requirements specification or as a product requirements document. The natural languages requirements are not the main kind of expressing requirements. [5]

The IBM Rational Dynamic Object Oriented Requirements System (DOORS) represents requirements management tool for systems and advanced information technology applications. Tool is accessible from web and to prevent the danger of conflicting changes gives users the possibility to lock sections of documents for editing. As big advantage of this tool can be seen the requirements interface bringing comprehensive roundtrip traceability. Users are able to manage, track and report implanted processes. [4]

The next chosen example of requirements management tool comes from Micro Focus and it is CaliberRM. As declared by producer the CaliberRM enable users the powerful facilities to capture, analyse and validate requirements. As expected the tool also provides central and secure repository for deposited requirements. [8]

ModelicaML represented new language developed by PhD student Schamai in Sweden and enables requirement formalization and integrates UML and Modelica. The language is implemented in a prototype based on Eclipse Papyrus UML, Acceleo, and Xtext for modeling, and OpenModelica tools for simulation. The simulation results produced are then used to draw conclusions on requirement violations. This approach supports the development and dynamic verification of cyber-physical systems. ModelicaML facilitates a holistic view of the system by enabling engineers to model and verify multi-domain system behavior using mathematical models and state-of-the-art simulation capabilities. Using this approach, requirement inconsistencies, incorrectness, or infeasibilities, as well as design errors, can be detected and avoided early on in system development. [14]

As because there still exists communication gap between analyst and domain experts, the aim of this contribution is suggest an approach avoiding this. The research aims to represent the dynamic requirements model as a finite state machine (FSM). In such case it can be used for quantifying the real processing of the application. As the dynamic model was selected representation using Business Object Relation Modeling (BORM) which was not used in any mentioned commercial tool. The reason is simple. Any modeling and simulation tool and any diagramming technique should be comprehensible to the stakeholders, many of whom are not software engineering literate. Moreover, these diagrams must not deform or inadequately simplify requirement information. The correct mapping of the problem into the model and subsequent visualization and possible simulation is very hard task with standard diagramming techniques used in major commercial tools. The business community needs a simple yet expressive tool for process modeling.

3 BORM

Business Object Relation Modeling (BORM) represents an approach to both process modeling and the subsequent development of information systems. [6, 7] It provides the description of how real business systems evolve, change and behave. BORM was originally developed in 1993 and was intended to provide seamless support for the building of object oriented software systems based on pure object-oriented languages, databases and distributed environments. Subsequently, it has been realized that this method has significant potential in business process modeling and other related business issues. Process approach and object orientation are the pillars of the BORM method. It is the application of principles that are successful in the field of modeling and software.

The basis of the object approach is the notion that each action must have an object that executes it. Each object must have some activity in a conceptual model. Every action means there has to be an object, object means there has to be some action. This is BORM interpretation of the MDA approach. [10]

3.1 Combination of the OOP and FSM

There was not a standard solution to the problem of gathering and representing knowledge. That is the reason why own UML-based BORM process diagramming technique [20] was developed and successfully used. It represents the way of starting object-oriented business system analysis recommended by Taylor [19] and together with Scheldbauer [15] it prefers this approach before the semantically different Business Process Model and Notation (BPMN) [3, 17]. BORM innovation is based on the reuse of old thoughts from the beginning of 1990s regarding the description of object properties and behavior using finite state machines (FSM). The first work expressing the possible merge of Object-Oriented Paradigm (OOP) and FSM was the book by Shlaer and Mellor ([16]). One of the best books about the applicability of OOP to the business modeling was written by Taylor ([19]). These works together with practical experience is why to believe that the business requirements modeling and software modeling could be unified on the platform of OOP and FSM.

The object-oriented approach has its origins in the 1970s in the researching of operating systems and graphic user interfaces. The object describing data structures and their behavior is the basic element. This is the difference from other modeling approaches where data and behavior are described separately and independently. OOP has been and still is explained in many books, for example Schach in [13]. To one of the best publications belongs Rubin and Goldberg [12].

In the field of theoretical informatics, the theory of automata is a study of abstract automatons and the problems they can save. An automaton is a mathematical model for a device that reacts to its surroundings, gets input, and provides output. Automatons can be configured in a way that the output from one of them becomes input for another. An automaton's behavior is defined by a combination of its inner structure and its newly - accepted input. The automata theory is a basis for language and translation theory, and for system behavior descriptions. The usage for modeling and simulation in software engineering activities has been described for example by Shlaer and Mellor in [16]. The idea of automata also inspired behavioral aspects of the UML.

3.2 Modeling cards

The BORM development methodology starts from an informal problem specification and the advantage is that it provides both, methods and techniques, to enable this informal specification to be transformed into an initial set of interacting objects. The main technique used here are modified modeling cards from the Object Behavior Analysis (OBA) being firstly published in [12]. Original OBA is only text-based method and used a large set of form sheets, textual lists and tables for storing and manipulating the information being processed. Modeling cards are structured texts, various lists and tables and so-called modeling cards (textual forms).

In BORM it is not started directly by drawing the process diagrams. Process diagrams are the subsequent refined visual representation of the information collected by the modeling cards. Modeling card of a scenario clarifies the entire process contours, process participants, necessary legislation, documents etc. Modeling card of a participant is a textual description of some role in a process. It has similar structure as scenario card, but seen from the different perspective of particular participant. Participant modeling cards are subsequently refined into several FSM.

Business process diagrams in BORM, or Object-Relationship Diagrams (ORD), are visual representation of processes and objects inside of processes obtained by modeling cards technique. Process diagram consists of participants, their states and transitions and their mutual communications. Each participant is composed of a set of states, activities and transitions (communications). Formally, it is a Mealy-type FSM. [16] Conceptual link within one participant can be considered as a transition between states, it contains no data, because it is only behavioral concept. On the other side communication between more participants may contain the data and therefore can be considered as data flows between activities of these participants making together some concrete process. Therefore a whole process diagram can be seen as a set of several finite state machines where each FSM represents just one participant.

In the basic concept of the FSM, each participant is represented as a unique entity, defined as 5-tuple $P_i(S_i, I_i, \delta_i, s_i^0, s_i^e)$, where:

- S_i is a finite non-empty set of states which the participant may be in it,
- I_i is a finite non-empty set of all possible inputs

- δ_i represents the activities carried out, i.e. transitions between states,
$$\delta_i : I_i \times S_i \rightarrow S_i,$$
- s_i^0 is the initial state of the process,
- s_i^e is the final state of the process.

The participant starts from the state s_i^0 and according user input I_i and actual state transfers itself into a next state. In case the participant ends in the state s_i^e we say that P_i accepts user word from input I_i^*. We allow reading an empty symbol X so the participant can continue to the next state even without any user input.

The model is possible to extend by N not communicating participants $P_1 ... P_N$ simulated together. User input symbols differs for each participant.

$$\bigcap_{i=1}^{N} I_i \subseteq \{\varepsilon\} \tag{1}$$

it can be defined a finite state machine $P_{\sum}$, which is composition of the partial automata representing individual participants. That machine simulates all participants together.

$$S = S_1 \times \cdots S_N, s_0 = (s_1^0, s_2^0, \cdots, s_N^0), s_\varepsilon = (s_1^\varepsilon, s_2^\varepsilon, \cdots, s_N^\varepsilon) \tag{2}$$

$$I = \bigcup_{i=1}^{N} I_i \tag{3}$$

Therefore inner states of $P_{\sum}$ are tuples of length N composed of individual participant's inner states. Same can be said about start state and end state. User input symbols are different for each automata to easily define action function for compound FSM with the help of individual action function X.

$$\delta : I \times S \rightarrow S \tag{4}$$

$$\delta(i, s_1, \cdots, s_N) = (s_1, \cdots, s_{j-1}, s_j', s_{j+1}, \cdots, s_N) | \exists j \in \hat{N}, \delta_j(i, s_j) = s_j' \tag{5}$$

The following process description is based on the basic concept of a finite automaton, but it enhances the part of the model of communicating finite state machines that are necessary to capture the mutual communication participants.

Business-process diagram representing a particular process can be defined as a finite set of participants.

$$\mathbf{BP} = \{\mathbf{P^i}\} \tag{6}$$

Each participant then can be described as an ordered 6-tuple

$$\mathbb{P}^i = (\mathbb{S}^i, -\mathbb{M}^i, +\mathbb{M}^i, f^i, g^i, s_1^i)$$

where:

- S^i is a finite set of all possible states which the participant may be in it
- $-M^i$ is a finite set of all outgoing messages
- $+M^i$ is a finite set of all received messages
- f^i represents the activities carried out, i.e. transition between states. The transition function can be defined as $f^i : S^i \times +M^i \rightarrow S^i$
- g^i is output function that can be defined as $g^i : S^i \times +M^i \rightarrow -M^i$
- s_1^i is the initial state of participant P^i, if $s_1^i \in S^i$

Without loss of generality, it can be assumed that the participant will not send the message to itself

$$-\mathbf{M^i} \cap +\mathbf{M^i} = \emptyset \tag{7}$$

The set of all messages participant will be the union of the set of all outgoing and incoming messages

$$M^i = -M^i \cup +M^i \tag{8}$$

Without loss of generality, it can be assumed that within the entire communications of a system is just one identical mi. Each message has only one recipient and one sender.

$$\forall m_1 \in \mathbf{M^1}, m_2 \in \mathbf{M^1} : m_1 \neq m_2 \tag{9}$$

The set of all messages $\mathbf{M^i}$ consists of an ordered triple $\left\langle \sigma^i, \mathrm{in}^i, \mathrm{out}^i \right\rangle$

$$M^i = \left\{ \left\langle \sigma^i, \mathrm{in}^i, \mathrm{out}^i \right\rangle \right\} = m^i \tag{10}$$

where σ^i is symbol, $textin^i$ and $textout^i$ are data.
Each message has its sender and recipient

$$\forall \mathbf{P^i} : \bigcup_i -M^i = \bigcup_i +M^i \tag{11}$$

The functions data(P^i) and in(m^i) can be defined for recipient where

$$data(\mathbf{P^i}; in(m^i) = in^i; m^i = \left\langle \sigma^i, \mathrm{in}^i, \mathrm{out}^i \right\rangle \tag{12}$$

$$in(m^i) = in\left(\left\langle \sigma^i, \mathrm{in}^i, \mathrm{out}^i \right\rangle \right) = m^i \tag{13}$$

Analogously functions data(P^j) and out(m^i) for sender:

$$data(P^j); \mathrm{out}(m^i) = \mathrm{out}^i; m^i = \left\langle \sigma^i, \mathrm{in}^i, \mathrm{out}^i \right\rangle \tag{14}$$

$$out(m^i) = \mathrm{out}\left(\left\langle \sigma^i, \mathrm{in}^i, \mathrm{out}^i \right\rangle \right) = m^i \tag{15}$$

Although the exchange of messages carried out from the perspective of BORM semantics at the same time, if we apply the theory of finite automata, we must distinguish the state before the transition and the new state after transition. Sent or received message at time $t+1$ will depend on the state at time t.

Data interchange between participants we can define for recipient as:

$$data^{t+1}(\mathrm{P}^i) = data^t(\mathrm{P}^i) \cup in(m^{ij}) \wedge in(m^{ij}) \subseteq data^t(\mathrm{P}^j) \tag{16}$$

And for sender:

$$data^{t+1}(\mathrm{P}^j) = data^t(\mathrm{P}^j) \cup out(m^{ij}) \wedge out(m^{ij}) \subseteq data^t(\mathrm{P}^i) \tag{17}$$

Using formal description based on the theory of finite state machines was defined process participant which can communicate with other participants by sending messages. Process diagram consists of a set of participants who obtained partial composition of finite state machines into a single comprehensive machine. Such a complex machine will represent a whole process diagram. Chance composing machines, thereby reducing the complexity of the resulting process has been known for decades. In detail, the issue of composition, minimization and generalization of finite state machines deals such as [1], which describes the specific algorithms for their composition.

Based on the derived definition of communicating participants, it can be described any process diagram in BORM as a finite state machine. This composed machine will consist of a set of finite state machines each of which will represent just one participant. If we apply the algorithm for the composition of finite state machines described in [16] on a set of participants, we get

$$\mathrm{FSM}^{BP} = (AE, \hat{E}, \varphi, \gamma, \sigma_1) = \{P^i\} = (\mathbb{S}^i, -\mathbb{M}^i, +\mathbb{M}^i, f^i, g^i, s_1^i)$$

where:

The transition function φ for the composed finite state machine obtained using the transition functions of each automaton represents participants.

The output function γ for the composed finite state machine obtained using the transition functions of each automaton represents participants.

As it has been shown above it is possible to describe any process model in BORM using FSM theory. The practical impact is the ability to use all the theoretical assumptions and practices that are known from the theory of FSM for process models in BORM. Therefore it can verify that all states of participants of the process are reachable and that all activities performed. It is also possible to automatically identify the state in which could lead to deadlock the process and evaluate the consistency of the model. Formal description also opens up opportunities for better implementation of the method BORM in CASE tools, especially in the construction process simulators.

4 Dynamic Model and Quantitative Methods

The main about BORM is that there is detail plan about data flows in the program. The data flows are the main thing to be used in the research. There is no need to wait for first code, first prototype to test the application and find out for example bad defined requirements in so late phase of software development process. The aim is to simulate the real function of software in the possible earliest phase of developing process. This is possible thanks to representation of dynamic request model as finite state machine. There is a possibility to simulate and quantify the data flows in the application. It is not only problem of function, there is a big potential of using the results of such testing based on BORM diagrams to design the application data structure.

The reason is quite simple. Save the time for developing the whole application and then starting the test phase just for the plain code. Never verify requirements after coding. Realisation of requirements based testing in the early phase of developing process means big cost savings because fixing errors in later developing phase is very expensive. This represents relatively small time investment in early phase of developing process that can means a big costs savings in later phases.

As already said, the main advantage is seen in using the BORM model represented by finite state machine to design the data structure of application. The research is based on modelling and quantifying the possible application data flows. at this time there is seen big potential in task verification and next in creating diagrams just from the dynamic requirements model based on FSM.

4.1 Task Verification

As the BORM diagram is fully dynamic, it can easily serve as something like testing scenario of real application. There is a big chance to reach big costs savings. Because as it well known, the testing phase of application development process is most expensive to fix errors. The testing phase means in this case not the testing of meeting application with ist requirements, but meeting requirements with practice use of application. There is a big problem, when the application is well designed, but the practical use shows it do not match the requirements. In the future research there is a plan to use the data from testing process to suggest also the programming language. But this needs to have the data set about the programming languages and then search the best language for the application represented by the finite state machine representing the BORM diagram.

4.2 Database diagram

The simple example to present our aim of using BORM method to create something like pattern for the data structure is database model. There are objects with its own states communicating each other. Each state is also telling the detailed information about itself. As illustrative can be mentioned the project of library and the database of books. Each book in the library has ist own state, is for example free to lend, lent or reserved. Each state needs to be described more in detail, for example the time for lending, the time of reservation, the number of lends in last period and so on.

There are some probabilities of being in some state. Then there are some statistics of lending time. All these data has steakholder or will be collected in coordination with him. Based on such data can be designed for example entity relation of the database. The key role is in collecting users requirements. The data flows collected from testing phase based at BORM diagram are the key to plan the correct database model. The simulations of flows can serve as the best input for new database. There is a big potential with big potential cost savings in software developing process. The savings are based mainly on testing the applications and also requirements in early phase of developing process.

5 Summary

Firstly was mentioned the basement of requirements engineering. As there already exist a lot of tools working with requirements the briefly overview was given. The market tools give similar functions and work quite well. At this tools was seen that there is nothing like the theme of presented research. The main problem is still seen in the existing gap between analyst and domain exerts. In this contribution was presented the possible use of dynamic requirements model represented by finite state machines. The model is created in early phase of developing process and that its big advantage. This means that the results from working with it, can be easily use in later software developing phases. It also means to prevent fixing errors in later phases of software developing process and that means the possibility to make big costs savings. As the possible way of using the BORM diagram is seen the task verification and creating database models. The concrete way of task verification, creating database diagram, estimation of application robustness is task for following research.

References

[1] BARJIS, J. (2007) *Developing Executable Models of Business Systems*, in: Proceedings of the ICEIS - International Conference on Enterprise Information Systems, pp. 5-13. INSTICC Press.

[2] HULL, E., JACKSON K., DICK J. (2010) *Requirements Engineering.* Springer, London.

[3] GROSSKOPF, A., DECKER, G., WESKE, M. (2006) *Business Process Modeling Using BPMN*, Meghan Kiffer Press, ISBN 978-0-929652-26-9.

[4] IBM: *Integrate requirements and change management with IBM Rational software.* In: IBM. Available at: `http://public.dhe.ibm.com/common/ssi/ecm/en/rad14034usen/RAD14034USEN.PDF`

[5] Jama Software: *The agile way to communicate requirements and manage complex projects.* In: Jama Software. Available at: http://www.jamasoftware.com/contour/

[6] KNOTT, R. P., MERUNKA, V., POLAK, J. (2003) *The BORM methodology: a third-generation fully object-oriented methodology*, Knowledge-Based Systems Elsevier Science International New York, ISSN 0950-7051

[7] KNOTT, R.P., MERUNKA, V., and POLAK, J. (2000) *Process Modeling for Object Oriented Analysis Using BORM Object Behavioral Analysis.* 4th International Conference on Requirements Engineering, Proceedings. N.p., 2000. 7 –16. IEEE Xplore.

[8] Micro Focus: CaliberRM. In: Micro Focus. Available at: http://www.microfocus.com/products/caliber/caliberrm/index.aspx

[9] Micro Focus: Document Factory. In: Micro Focus Documentation. Available at: `http://documentation.microfocus.com/help/index.jsp?topic=\%2Fcom.borland.caliberrm.doc\%2Fhtml\%2Fcreatedocumentdocfactory.htm`

[10] MDA – The Model Driven Architecture, OMG – The Object Management Group, http://www.omg.org.

[11] LAPLANTE, P. (2007) *What Every Engineer Should Know about Software Engineering.* CRC Press.

[12] RUBIN, K., GOLDBERG, A. (1992) *Object Behavioral Analysis.* Communications of the ACM - Special issue on analysis and modeling in software development CACM, Vol. 35 Issue 9.

[13] SCHACH, S. (2008) *Object-Oriented Software Engineering*, McGraw Hill, Singapore, ISBN 978-007-125941-5.

[14] SCHAMAI, W. (2013) *Model-Based Verification of Dynamic System Behavior aganist Requirements: Method, Language, and Tool.* Linköping. Dissertation. Linköping University.

[15] SCHELDBAUER, M. (2010) The Art of Business Process Modeling - The business Analyst Guide to Process Modeling with UML and BPMN, Cartris Group, Sudbury MA, ISBN 1-450-54166-6.

[16] SHLAER, S. MELLOR, S. (1992) Object Lifecycles: Modeling the World in States, Yourdon Press, ISBN 0136299407.

[17] SILVER, B. (2011) BPMN Method and Style, 2nd Edition, with BPMN Implementer's Guide: A Structured Approach for Business Process Modeling and Implementation Using BPMN 2.0, Cody-Cassidy Press.

[18] STECKLEIN, S., et al. *Error Cost Escalation Through the Project Life Cycle.* In: NASA Technical Reports Server. [online]. `http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100036670_2010039922.pdf`.

[19] TAYLOR, D. A. (1995) *Business Engineering with Object Technology*, John Wiley ISBN 0-471-04521-7.

[20] The UML standard, OMG – The Object Management Group, http://www.omg.org, ISO/IEC 19501.

Application of Edge Detectors to Alzheimer's Disease Diagnosis from 3D SPECT Image

Martina Jandová

2nd year of PGS, email: jandoma1@fjfi.cvut.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Aleš Procházka, Department of Computing and Control Engineering
Faculty of Chemical Engineering, ICT in Prague

Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. 3D SPECT Images of human brain are frequently used in diagnostics of Alzheimer's disease. Our approach is based on digital morphology of 3D gray image. We compare traditional edge detectors and morphological filters together with various approaches to data preprocessing in this study. Space fluctuations of detected signals are studied via statistical testing of skewness or kurtosis. Pilot study on small patient's group brings statistical significant characteristics which are suitable for diagnosis of neural degenerative disease.

Keywords: high frequency filter, edge detection, mathematical morphology, digital 3D image, MATLAB, Alzheimer's disease

Abstrakt. 3D obrazy lidského mozku získaného pomocí metody SPECT se často využívají při diagnostice Alzheimerovy choroby. V této práci jsou porovnávány tradiční hranové detektory s morfologickými filtry za použití různých metod předzpracování obrazu. Výkyvy detekovaných signálů jsou podrobeny statistickému testování pomocí šikmosti a špičatosti. Získané výsledky s malou skupinou pacientů přináší statisticky významné charakteristiky, které jsou vhodné pro diagnostiku neurodegenerativní onemocnění.

Klíčová slova: vysokofrekvenční filtr, hranový detektor, matematická morfologie, digitální 3D obraz, MATLAB, Alzheimerova choroba

1 Introduction

Alzheimer's disease [15] is a disabling and distressing disorder that affects 5 % of the population older than 65 and 20 % of those over 80. For last three decades a great progress were made in medical engineering and computing technologies. One of non-invasive diagnostic tools that provide clinical information regarding biochemical and physiologic processes in patients is called Single-Photon Emission Computed Tomography (SPECT) [5]. This methods provides a good inside into brain neuron activities. Using various radio markers and measuring their activities during scan. Neural degenerative diseases [13] are about structural changes of gray mater. The paper is oriented to structure morphology of 3D image because changes of brain structure will change the structure of image and therefor can be indicated and then used to disease classification. The paper is organized as follows. Basic notations related to 3D image and its local processing are remembered first. Novel method of Alzheimer's disease classification is based on three step processing. Linear and median filters with various masks are used as low-pass filters in the first step. Traditional edge detectors and morphological detectors as high-pass filters magnitudes are applied in second step. Final third step consists of overall skewness and kurtosis evaluations. The efficiency of recognition process will be investigated by statistical testing of sample median differences between Alzheimer's and control groups.

2 Preliminaries of 3D Image

The paper is oriented to digital processing of 3D image. It is necessary to remember basic terms and notations as will be used in next chapters.

Let $m, n, h \in \mathbb{N}$ be number of rows, columns, and levels of 3D image. The image can by represented as matrix $\mathbf{X} \in \left(\mathbb{R}_0^+\right)^{m \times n \times h}$. Any local image processing is driven by mask, which is also a matrix. We prefer integer mask as $\mathbf{M} \in \mathbb{Z}^{3\times3\times3}$ of small size in this study. The mask slides over the data and we can easily perform local image processing as follows. Let $\mathbf{x} \in (i, j, k)$ be position of mask center, where $2 \le i \le m-1$, $2 \le j \le n-1$, $2 \le k \le h-1$. Using mask $\mathbf{M}$ around this point we collect the values from original image into matrix $\mathbf{B} \in \left(\mathbb{R}_0^+\right)^{3\times3\times3}$ according to formula $b_{u,v,w} = x_{i+u-2,j+v-2,k+w-2}$ for $u, v, w \in \{1, 2, 3\}$.

Any local characteristic [3] is only a function $\mathrm{ch}(\mathbf{x}) = \mathrm{g}\left(\mathbf{B}(\mathbf{x}), \mathbf{M}_1, \dots \mathbf{M}_k\right)$. Individual characteristics differ in number of masks, their elements, and type of processing function g.

3 Local Image Processing

Detection of structures and their variabilities is based here on local image processing of three kinds. Original 3D image can be preprocessed using image filtering. Edge detection can be performed using traditional approaches [3] or grey morphological operators [2], [10].

3.1 Image filtering

Low-pass digital filters of two kinds: weighted arithmetic mean [4] and weighted median [3], are optionally used for image smoothing with single mask. Various masks were used as follows

$$\mathbf{F}_1 = \left[\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right],$$

$$\mathbf{F}_2 = \left[\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 6 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right],$$

$$\mathbf{F}_3 = \left[\begin{array}{ccc|ccc|ccc} 1 & 2 & 1 & 2 & 4 & 2 & 1 & 2 & 1 \\ 2 & 4 & 2 & 4 & 8 & 4 & 2 & 4 & 2 \\ 1 & 2 & 1 & 2 & 4 & 2 & 1 & 2 & 1 \end{array}\right],$$

$$\mathbf{F}_4 = \left[\begin{array}{ccc|ccc|ccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right].$$

The mask $\mathbf{F}_1$ is also useful in greyscale morphology, mask $\mathbf{F}_3$ is a good approximation of gaussian smoothing, and mask $\mathbf{F}_4$ represents box filter.

3.2 Traditional Edge Detectors

Digital edge detectors try to approximate norm of image gradient. General [10] formula of traditional edge detector is

$$\mathrm{ch}(\mathbf{x}) = |\nabla \mathbf{f}(\mathbf{x})| \approx \left(\sum_{k=1}^{N} G_k^2 \right)^{1/2}$$

where G_k is a gradient approximation using k^{th} mask. When $\mathbf{m}_k$ is a vector formed from $\mathbf{M}_k$ and $\mathbf{b}$ is a vector formed from $\mathbf{B}$, we direct calculate scalar product $G_k = \mathbf{m}_k \cdot \mathbf{b}$.

Roberts detector

Roberts edge detector [8], [2] applies four masks. First of them is

$$\mathbf{T}_1 = \left[\begin{array}{ccc|ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

The other masks are also diagonals in $2 \times 2 \times 2$ cube.

Prewitt operator

Prewitt edge detector [7], [11] applies nine masks. Two of them are

$$\mathbf{P}_1 = \left[\begin{array}{ccc|ccc|ccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{array}\right],$$

$$\mathbf{P}_2 = \left[\begin{array}{ccc|ccc|ccc} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 0 & -1 & -1 & 0 & -1 & -1 & 0 & -1 & -1 \end{array}\right].$$

The other masks are obtained by rotation.

Sobel operator

Sobel edge detector [3] applies nine masks. Two of them are

$$\mathbf{S}_1 = \left[\begin{array}{ccc|ccc|ccc} 1 & 2 & 1 & 2 & 4 & 2 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -1 & -2 & -4 & -2 & -1 & -2 & -1 \end{array}\right],$$

$$\mathbf{S}_2 = \left[\begin{array}{ccc|ccc|ccc} 2 & 1 & 0 & 4 & 2 & 2 & 2 & 1 & 0 \\ 1 & 0 & -1 & 2 & 0 & 0 & 1 & 0 & -1 \\ 0 & -1 & -2 & 0 & -2 & -4 & 0 & -1 & -2 \end{array}\right].$$

The other masks are obtained by rotation.

Robinson operator

Robinson edge detector [9], [4] applies only single mask

$$\mathbf{R}_1 = \left[\begin{array}{ccc|ccc|ccc} 1 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 \\ 1 & -2 & 1 & 2 & -4 & 2 & 1 & -2 & 1 \\ -1 & -1 & -1 & -2 & -2 & -2 & -1 & -1 & -1 \end{array}\right].$$

Laplacian operator

Laplacian edge detector [2] applies only single mask

$$\mathbf{L}_1 = \left[\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & -6 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right].$$

Laplacian of Gaussian

Laplacian of Gaussian detector (LoG) [6], [3] is based on convolution of two masks: $\mathbf{F}_3$ and $\mathbf{L}_1$. It is possible to perform it as a sequence of linear filtering based on gaussian smoothing and then Laplacian detector.

3.3 Morphological Detectors

Using $\mathbf{m}$ and $\mathbf{b}$ as vector representations of mask and local image, we can define elementary morphological operations using mask

$$\mathbf{M} = \left[\begin{array}{ccc|ccc|ccc} 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{array}\right].$$

Dilation

Elementary dilation [10] of single size has local characteristics

$$\mathrm{ch}(\mathbf{x}) = \max_{j:m_j=1} b_j.$$

Dilation of higher size can be performed as repeated sequence of elementary dilation.

Erosion

Elementary erosion [10] of single size has local characteristics

$$\mathrm{ch}(\mathbf{x}) = \min_{j:m_j=1} b_j.$$

Erosion of higher size can be performed as repeated sequence of elementary erosion.

Opening

Opening operator [2] is performed as dilation followed by erosion of the same mask size.

Closing

Closing operator [12] is performed as erosion followed by dilation of the same mask size.

High-pass morphological detectors

Let $\mathbf{X}, \mathbf{D}, \mathbf{E}, \mathbf{O}, \mathbf{C}$ be original image, its dilation, erosion, opening, and closing with fixed mask size. Morphological high-pass filters [3] can be designed as voxel by voxel calculations using formulas

$$\begin{aligned}
\mathbf{H}_1 &= \mathbf{D} - \mathbf{E}, \\
\mathbf{H}_2 &= \mathbf{D} - \mathbf{X}, \\
\mathbf{H}_3 &= \mathbf{X} - \mathbf{E}, \\
\mathbf{H}_4 &= \mathbf{O} - \mathbf{X}, \\
\mathbf{H}_5 &= \mathbf{X} - \mathbf{C}, \\
\mathbf{H}_6 &= \min(\mathbf{H}_2, \mathbf{H}_3), \\
\mathbf{H}_7 &= \min(\mathbf{H}_4, \mathbf{H}_5).
\end{aligned}$$

4 Fluctuation Measures

Local image processing operations from previous section will produce new image. Its intensity is not constant in general and vary voxel by voxel. Standard statistical characteristic are used as fluctuation measures. Employing arithmetic mean as

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

We remember basic statistical characteristics [14] as variance, skewness, and kurtosis according to formulas

$$s^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2,$$

$$skew = \frac{1}{n}\sum_{k=1}^{n}\frac{(x_k - \bar{x})^3}{s^3}, \text{and}$$

$$kurt = \frac{1}{n}\sum_{k=1}^{n}\frac{(x_k - \bar{x})^4}{s^4}$$

where n is total number of voxels. We suppose the patients will differ mainly in skewness and kurtosis after local image transforms.

5 Case Study: Alzheimer's Disease Diagnosis

Previous principals of image transformation and analysis were applied to actual problem of Alzheimer's diseases diagnosis. Original data were preprocessed via edge detectors of various kinds and statistical analysis of differences between Alzheimer's disease and control group was performed.

5.1 Data Description

Two groups of patients were studied. First group consist of 10 patients of Alzheimer's disease (AD). Second group consist of 10 patients with Amyotrophic Lateral Sclerosis (ALS) as control one with no changes in structure and functionality of brain. Patient brains were scanned using 3D SPECT technic. Resulting 3D images of size $128\times128\times128$ were subjects of data processing and statistical analysis. Therefore $m = n = h = 128$. Nonnegative intensities were individually divided into patient's maximum intensity to obtain normalized 3D image of every patient.

5.2 Selected Characteristic

Patient's characteristics are results of three step process: optional low-pass preprocessing, edge detection, and calculation of global statistical measures. Four masks were used for mean and median filtering in the first step. But the first step can be skipped. Therefore, there are nine possibilities of low pass image preprocessing. We used 6 traditional and 7 morphological edge detectors in the second step. Finally, both skewness and kurtosis of whole 3D edge scans are calculated in the third step. We obtained 234 various patient's characteristics in this way.

5.3 Hypotheses Testing

Individual patient's characteristics were tested using hypotheses H_0 of median equity and standard Wilcokson-Mann-Whitney test on significance level 0,05. Results of testing are included in tabs 1 - 4 as corresponding p-values. But it is a kind of multiple hypothesis testing. Therefore, we use False Discovery Rate [1] supposing independent or positive dependent testing. Corrected significant level is $\alpha_{\text{FDR}} = 0,0312$. Statistically significant differences after this Bonferroni like correction are highlighted.

The results can be interpret as follows. Independently on preprocessing algorithm and its masks, and independently on skewness or kurtosis evaluation, traditional filters as Roberts, Laplacian, and LoG are suitable for diagnosis of Alzheimer's disease. Similar observation related to preprocessing and statistical analysis recommends to use first morphological detectors, which are based on pair differences between original and its dilation, erosion, and opening. Therefore, Minkowski sausage, dilation minus original filter, original minus erosion filter, and white top hat detectors are comparable with the best traditional ones.

Table 1: Skewness after Application of traditional Edge Detectors

Pre-processing	**Mask**	**Roberts**	**Prewitt**	**Sobel**	**Robinson**	**Laplacian**	**LoG**
Mean	$\mathbf{F}_1$	**0,007285**	0,212294	0,273036	0,212294	**0,021134**	**0,025748**
	$\mathbf{F}_2$	**0,011330**	0,212294	0,241322	0,241322	**0,011330**	0,037635
	$\mathbf{F}_3$	**0,014019**	0,344704	0,384673	0,307489	0,031209	**0,009108**
	$\mathbf{F}_4$	0,045155	0,472676	0,570750	0,472676	**0,004586**	**0,002202**
Median	$\mathbf{F}_1$	**0,014019**	**0,021134**	0,088973	0,031209	**0,011330**	**0,007285**
	$\mathbf{F}_2$	**0,014019**	0,053903	0,140465	0,045155	**0,011330**	0,053903
	$\mathbf{F}_3$	**0,009108**	0,031209	0,088973	0,037635	**0,009108**	**0,007285**
	$\mathbf{F}_4$	**0,009108**	0,053903	0,121225	0,053903	**0,007285**	**0,009108**
None	None	**0,014019**	0,064022	0,161972	0,037635	**0,011330**	0,053903

Table 2: Kurtosis after Application of traditional Edge Detectors

Pre-processing	**Mask**	**Roberts**	**Prewitt**	**Sobel**	**Robinson**	**Laplacian**	**LoG**
Mean	$\mathbf{F}_1$	**0,005795**	**0,021134**	0,037635	**0,021134**	0,031209	**0,004586**
	$\mathbf{F}_2$	**0,011330**	**0,014019**	**0,025748**	**0,011330**	**0,011330**	**0,011330**
	$\mathbf{F}_3$	**0,007285**	0,053903	0,053903	0,053903	**0,009108**	**0,002202**
	$\mathbf{F}_4$	**0,017257**	0,140465	0,161972	0,140465	**0,002827**	**0,001706**
Median	$\mathbf{F}_1$	**0,011330**	**0,011330**	**0,017257**	**0,014019**	**0,009108**	**0,007285**
	$\mathbf{F}_2$	**0,014019**	**0,009108**	**0,007285**	**0,007285**	**0,009108**	**0,021140**
	$\mathbf{F}_3$	**0,011330**	**0,017257**	**0,025748**	**0,017257**	**0,007285**	**0,009108**
	$\mathbf{F}_4$	**0,005795**	0,031209	0,031209	0,037635	**0,005795**	**0,009108**
None	None	**0,014019**	**0,009108**	**0,007285**	**0,009108**	**0,009108**	**0,021134**

Table 3: Skewness after Application of Morphological Detectors

Pre-processing	**Mask**	$\mathbf{H}_1$	$\mathbf{H}_2$	$\mathbf{H}_3$	$\mathbf{H}_4$	$\mathbf{H}_5$	$\mathbf{H}_6$	$\mathbf{H}_7$
Mean	$\mathbf{F}_1$	**0,009108**	**0,005795**	**0,007285**	0,031209	0,064022	0,570750	**0,011330**
	$\mathbf{F}_2$	**0,007285**	**0,009108**	**0,009108**	0,031209	0,733730	0,427355	**0,014019**
	$\mathbf{F}_3$	**0,011330**	**0,005795**	**0,007285**	**0,007285**	**0,014019**	0,520523	**0,014019**
	$\mathbf{F}_4$	**0,021134**	**0,005795**	**0,004586**	**0,005795**	**0,011330**	0,791337	**0,014019**
Median	$\mathbf{F}_1$	**0,011330**	**0,007285**	**0,009108**	**0,014019**	0,909722	0,520523	0,273036
	$\mathbf{F}_2$	**0,011330**	**0,011330**	**0,011330**	**0,004586**	0,969850	0,140465	0,088973
	$\mathbf{F}_3$	**0,017257**	**0,005795**	**0,007285**	**0,011330**	0,969850	0,677585	0,161972
	$\mathbf{F}_4$	**0,009108**	**0,005795**	**0,005795**	**0,017257**	0,088973	0,677585	0,045155
None	None	**0,005795**	**0,011330**	**0,014019**	0,088973	0,623176	**0,017257**	0,064022

Table 4: Kurtosis after Application of Morphological Detectors

Pre-processing	**Mask**	$\mathbf{H}_1$	$\mathbf{H}_2$	$\mathbf{H}_3$	$\mathbf{H}_4$	$\mathbf{H}_5$	$\mathbf{H}_6$	$\mathbf{H}_7$
Mean	$\mathbf{F}_1$	**0,007285**	**0,009108**	**0,011330**	**0,017257**	0,037635	0,037635	**0,009108**
	$\mathbf{F}_2$	**0,004586**	**0,009108**	**0,011330**	**0,017257**	0,677585	**0,014019**	**0,011330**
	$\mathbf{F}_3$	**0,014019**	**0,005795**	**0,011330**	**0,014019**	**0,017257**	0,140465	**0,011330**
	$\mathbf{F}_4$	**0,011330**	**0,004586**	**0,009108**	**0,007285**	**0,011330**	0,273036	**0,017257**
Median	$\mathbf{F}_1$	**0,003611**	**0,009108**	**0,009108**	**0,014019**	0,969850	0,031209	0,140465
	$\mathbf{F}_2$	**0,003611**	**0,009108**	**0,009108**	**0,004586**	1,000000	**0,003611**	0,241322
	$\mathbf{F}_3$	**0,002827**	**0,009108**	**0,009108**	**0,009108**	0,969850	0,088973	0,075662
	$\mathbf{F}_4$	**0,002827**	**0,005795**	**0,009108**	**0,021134**	0,088973	0,273036	0,075662
None	None	**0,011330**	**0,017257**	**0,011330**	0,045155	0,472676	**0,002202**	0,045155

6 Conclusion

In this paper, traditional edge detectors and morphological filters was tested for diagnosis of Alzheimer's disease. Independency these filters on preprocessing algorithm and its masks is proven by results of tests. The same applies for skewness or kurtosis evaluation. Roberts, Laplacian, and LoG filters and morphological detectors, which are based on pair differences between original and its dilation, erosion, and opening, have produced statistically significant results. Therefore, these filters can be used for diagnosis of Alzheimer's disease.

References

[1] Y. Benjamini, Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* In 'Journal of the Royal Statistical Society', Series B (Methodological), 1995, Vol.57, No. 1, 289-300.

[2] G. Dougherty, *Digital Image Processing for Medical Applications*. Cambridge: Cambridge University Press, 2009. ISBN 978-0-521-86085-7.

[3] R. C. Gonzales, R. E. Woods, EDDINS, Steven L., *Digital Image Processing Using MATLAB*, 2nd edition. United States of America: Gatesmark Publishing, 2009. ISBN 978-0-9820854-0-0.

[4] V. Hlavac, M. Sedlacek, *Zpracovani signalu a obrazu*, 3rd edition. Praha: CVUT, 2009. ISBN 978-80-01-04442-1.

[5] T. A. Holly, B. G. Abbott, D. A. Calnon, *Single photon-emission computed tomography*, 2010, vol. 17, issue 5, 941-973.

[6] D. Marr, E. C. Hildreth, *Theory of Edge Detection* In 'Proceedings of the Royal Society of London', Series B, Biological Sciences 207, 1980, 187-217.

[7] J. M. S. Prewitt, *Object Enhancement and Extraction*, Picture Processing and Psychopictorics (B. Lipkin and A. Rosenfeld, Ed.) New York, Academic Press, 1970.

[8] L. G. Roberts, *Machine perception of three-dimensional solids* In 'Optical and Electro-Optical Information Processing' (J. Tippett, Ed.) Cambridge, Massachussets: MIT Press, 1965.

[9] G. S. Robinson, *Edge detection by compass gradient masks* In 'Computer Graphics and Image Processing', Elsevier Inc., 1977.

[10] F. Y. Shih, *Image Analysis and Mathematical Morphology - Fundamentals and Application*, 1st edition. Boca Raton: CRC Press, 2009. ISBN 978-1-4200-8943-1.

[11] M. Sonka, V. Hlavac, R. Boyle, *Image Processing, Analysis, and Machine Vision.* Stamford: Cengage Learning, 2014. ISBN 978-1-133-59360-7.

[12] J. Starck, F. Murtagh, J. M. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity.* New York: Cambridge University Press, 2010. ISBN 978-0-521-11913-9.

[13] R. E. Tanzi, *The Genetics of Alzheimer Disease* in Cold Spring Harb Perspect Med, 2012, vol. 2(10), 1-11.

[14] S. P. Washington, M. G. Karlaftis, F. L. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis.* 2nd edition. Boca Raton, Florida: CRC Press, 2011. ISBN 978-1-4200-8286-9.

[15] Alzheimer's Association: Alzheimer's Disease Facts and Figures, *Alzheimer's & Dementia*, 2013, vol. 9, issue 2, 1-71.

Spectral Asymptotics of a Strong δ' Interaction Supported by a Surface*

Michal Jex

3rd year of PGS, email: jexmicha@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Pavel Exner, Department of Theoretical Physics
Nuclear Physics Institute, AS CR

Abstract. We derive asymptotic expansion for the spectrum of Hamiltonians with a strong attractive δ' interaction supported by a smooth surface in $\mathbb{R}^3$, either infinite and asymptotically planar, or compact and closed. Its second term is found to be determined by a Schrödinger type operator with an effective potential expressed in terms of the interaction support curvatures.

This paper was published in Physical Letters A, A378 (2014), 2091–2095 and it was presented at the conference Kent Spectral Theory in Canterbury (April 14th to 17th, 2014).

Keywords: δ' surface interaction, strong coupling expansion

Abstrakt. Odvodíme asymptotický rozvoj bodového spektra Hamiltoniánu, který popisuje silnou δ' interakci lokalizovanou na hladké ploše v $\mathbb{R}^3$, která je nekonečná a asymptoticky plochá nebo uzavřená a kompaktní. Druhý člen rozvoje je určen Schrödingerovským operátorem s efektivním potenciálem závislým na křivostech plochy.

Tento příspěvek byl publikován v časopise Physical Letters A, A378 (2014), 2091–2095 a byl přednesen na konferenci Kent Spectral Theory v Canterbury (14.–17.4.2014).

Klíčová slova: δ' povrchová interakce, asymptotický rozvoj silné vazby

References

[Ex08] P. Exner: *Leaky quantum graphs: a review*, Proceedings of the Isaac Newton Institute programme “Analysis on Graphs and Applications”, AMS “Proceedings of Symposia in Pure Mathematics” Series, vol. 77, Providence, R.I., 2008; pp. 523–564.

[BLL13] J. Behrndt, M. Langer, V. Lotoreichik: Schrödinger operators with δ and δ'-potentials supported on hypersurfaces, *Ann. Henri Poincaré* **14** (2013), 385–423.

[AGHH05] S. Albeverio, F. Gesztesy, R. Høegh-Krohn, H. Holden: *Solvable Models in Quantum Mechanics*, 2nd edition with an appendix by P. Exner, AMS Chelsea Publishing, Providence, R.I., 2005.

[CS98] T. Cheon, T. Shigehara: Realizing discontinuous wave functions with renormalized short-range potentials, *Phys. Lett.* **A243** (1998), 111–116.

*The research was supported by the Czech Science Foundation within the project 14-06818S and by Grant Agency of the Czech Technical University in Prague, grant No. SGS13/217/OHK4/3T/14.

[AN00] S. Albeverio, L. Nizhnik: Approximation of general zero-range potentials, *Ukrainian Math. J.* **52** (2000), 582–589.

[ENZ01] P. Exner, H. Neidhardt, V.A. Zagrebnov: Potential approximations to δ': an inverse Klauder phenomenon with norm-resolvent convergence, *Commun. Math. Phys.* **224** (2001), 593–612.

[EK03] P. Exner, S. Kondej: Bound state due to a strong δ interaction supported by a curved surface, *J. Phys. A: Math. Gen.* **36** (2003), 443–457.

[BEL13] J. Behrndt, P. Exner, V. Lotoreichik: Schrödinger operators with δ- and δ'-interactions on Lipschitz surfaces and chromatic numbers of associated partitions, `arXiv:1307.0074 [math-ph]`

[CEK04] G. Carron, P. Exner, D. Krejčiřík: Topologically non-trivial quantum layers, *J. Math. Phys.* **45** (2004), 774-784.

[F06] M. Fecko: *Differential Geometry and Lie Groups for Physicists*, Cambridge University Press, 2006.

[BK] G. Berkolaiko, P. Kuchment: *Introduction to Quantum Graphs*, Amer. Math. Soc., Providence, R.I., 2013.

[DEK01] P. Duclos, P. Exner, D. Krejčiřík: Bound states in curved quantum layers, *Commun. Math. Phys. 223* **223** (2001), 13–28.

[EJ13] P. Exner, M. Jex: Spectral asymptotics of a strong $\delta\prime$ interaction on a planar loop, *J. Phys. A: Math. Theor.* **46** (2013), 345201.

Výpočet fázové rovnováhy systému $CO_2 - H_2O$ při daném objemu, teplotě a složení s aplikací v CO_2 sekvestraci *

Tereza Jindrová

2. ročník PGS, email: jindrter@fjfi.cvut.cz
Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Jiří Mikyška, Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. Studying $CO_2 - H_2O$ system phase behaviour is motivated by CO_2 sequestration, which is from an ecology point of view a possibility of protection against the greenhouse effect by capturing emissions of CO_2 at the source and storing them into deep geological repositories or salt-water reservoirs. For such operations, it is essential to fully understand the thermodynamics of the processes in the subsurface and to have a model which describes the behaviour of CO_2 correctly under wide range of natural geological conditions.

Injecting CO_2 into a reservoir, it may dissolve in water or it can mix with it and the $CO_2 - H_2O$ mixture can split into two or more phases. Let us consider a closed system of total volume V containing a $CO_2 - H_2O$ mixture with mole numbers N_w, N_c at temperature T. First, we are interested to find out whether the system is under given conditions in single-phase or splits into two or more phases; this is the problem of single-phase stability at constant volume, temperature and moles (the so-called VT-stability). In case of phase-splitting we want to establish volumes of both phases, mole numbers of each component in both phases, and consequently the equilibrium pressure of the system from the equation of state; this is the problem of two-phase split calculation at constant volume, temperature and moles (the so-called VT-flash). In the previous work [3, 4, 6], these problems were formulated and the algorithms were proposed and tested on many examples.

The contribution deals with the investigation of multi-phase equilibrium of $CO_2 - H_2O$ system at constant volume, temperature and moles. Studying $CO_2 - H_2O$ mixture under natural geological conditions (pressures typically below 50 MPa and temperatures typically $298 - 383$ K) for different system composition, two-phase and three-phase states were observed. Recently, we have developed and successfully tested a fast and robust algorithm for constant-volume two-phase split calculation, which is based on the direct minimization of the total Helmholtz free energy of the mixture with respect to the mole- and volume-balance constraints [6]. The algorithm uses modified Newton-Raphson minimization method with line-search and modified Cholesky decomposition of the Hessian matrix to produce a sequence of states with decreasing values of the total Helmholtz free energy. Using of the Newton-Raphson method ensures fast convergence towards the exact solution. To initialize the algorithm, an initial guess is constructed using the

[*] Práce vznikla v rámci projektů Výpočetní metody v termodynamice vícesložkových směsí LH12064 MŠMT ČR, Vývoj počítačových modelů pro simulaci ukládání CO_2 do podzemí P105/11/1507 Grantové agentury České republiky, Aplikovaná matematika v technických a fyzikálních vědách MSM6840770010 MŠMT ČR, a Pokročilé superpočítačové metody implementace matematických modelů SGS11/161/OHK4/3T/14 Studentské grantové soutěže ČVUT v Praze.

results of constant-volume stability testing which has been developed in [4]. Now we extend the results for CO_2-H_2O system and propose a fast and robust algorithm for three-phase equilibrium computation at constant volume, temperature and moles. The performance of the proposed algorithm has been tested on many examples of two- and three-phase equilibrium calculations of $CO_2 - H_2O$ mixture, which is described using the Cubic-Plus-Association equation of state. In the contribution, we present several examples which are important for CO_2 sequestration.

This work was presented at Interpore Conference 2014 in Milwaukee, Wisconsin (27.-30.5.2014) and the full article [7] has been accepted to the IAENG Journal of Applied Mathematics.

Keywords: phase equilibrium, constant-volume flash, CO_2 sequestration, Helmholtz free energy minimization, Newton-Raphson method, modified Cholesky decomposition.

Abstrakt. Zkoumání fázového chování směsi $CO_2 - H_2O$ je motivováno CO_2 sekvestrací; z ekologiského hlediska se jedná o jednu z možností ochrany před skleníkovým jevem zachycováním emisí CO_2 přímo u zdroje a ukládáním těchto emisí do hlubinných geologických úložišť nebo slaných akviferů. Pro takové operace je nezbytné jednak rozumět termodynamice procesů probíhajících v podzemí a zároveň mít takový model, který dokáže správně popsat chování CO_2 v širokém rozsahu běžných geologických podmínek.

Při injektování CO_2 do rezervoáru může docházet buď k jeho rozpuštění ve vodě nebo k jeho smíchání s vodou a následnému rozdělení směsi CO_2-H_2O do dvou nebo více fází. Uvažujeme-li uzavřený systém obsahující směs CO_2-H_2O s látkovými množstvími N_w, N_c o celkovém objemu V při teplotě T, nachází se tento systém za daných podmínek buď stabilně v jedné fázi nebo je nestabilní a dojde k jeho rozdělení do dvou nebo více fází; jedná se o problém jednofázové stability při konstantním objemu, teplotě a složení (tzv. VT-stabilita). V případě rozdělení do fází určíme objemy a složení obou fází, a následně vypočítáme rovnovážný tlak systému ze stavové rovnice; jedná se o problém výpočtu dvoufázové rovnováhy při konstantním objemu, teplotě a složení (tzv. VT-flash). V předchozí práci [3, 4, 6] byly tyto problémy formulovány a příslušné výpočetní algoritmy byly navrženy a testovány na řadě příkladů.

Článek pojednává o vyšetřování vícefázové rovnováhy směsi $CO_2 - H_2O$ při konstantním objemu, teplotě a složení. Při vyšetřování směsi $CO_2 - H_2O$ za běžných geologických podmínek (typicky tlaky pod 50 MPa a teploty v rozmezí 298 – 383 K) a při různém složení směsi byly pozorovány dvou- a třífázové stavy. Nedávno byl navržen a úspěšně testován rychlý a robustní algoritmus pro výpočet dvoufázové rovnováhy za konstantního objemu, teploty a složení založený na přímé minimalizaci celkové Helmholtzovy volné energie směsi při zachování podmínek na bilanci hmoty a objemu [6]. Numerický algoritmus je zde založen na modifikované Newtonově-Raphsonově minimalizační metodě s použitím metody line-search a modifikovaného Choleskyho rozkladu matice Hessiánu, čímž je vytvořena posloupnost stavů s klesajícími hodnotami celkové Helmholtzovy volné energie. Použití Newtonovy-Raphsonovy metody navíc zajišťuje rychlou konvergenci k přesnému řešení. Algoritmus pro VT-flash je inicializován počátečním odhadem z testování stability [4]. Nyní rozšíříme výsledky pro směs $CO_2 - H_2O$ a navrhneme rychlý a robustní algoritmus pro výpočet třífázové rovnováhy při konstantním objemu, teplotě a složení. Navržený algoritmus byl testován na mnoha příkladech výpočtu dvou- a třífázové rovnováhy směsi CO_2-H_2O, která je popsána pomocí kubické stavové rovnice s asociačním členem. V článku prezentujeme několik příkladů, které hrají významnou roli v CO_2 sekvestraci.

Tato práce byla prezentována na konferenci Interpore 2014 v Milwaukee, Wisconsin (27.-30.5.2014) a celý článek [7] byl přijat do časopisu IAENG Journal of Applied Mathematics.

Klíčová slova: fázová rovnováha, flash při konstantním objemu, CO_2 sekvestrace, minimalizace Helmholtzovy volné energie, Newtonova-Raphsonova metoda, modifikovaná Choleskyho dekompozice.

Literatura

[1] A. Firoozabadi. *Thermodynamics of Hydrocarbon Reservoirs*, McGraw-Hill, New York, 1999.

[2] A. Firoozabadi, Z. Li. *Cubic-Plus-Association Equation of State for Water-Containing Mixtures: Is "Cross Association" Necessary?.* AIChE Journal **55(7)** (2009), 1803–1813.

[3] A. Firoozabadi, J. Mikyška. *A New Thermodynamic Function for Phase-Splitting at Constant Temperature, Moles, and Volume.* AIChE Journal **57(7)** (2011), 1897–1904.

[4] A. Firoozabadi, J. Mikyška. *Investigation of Mixture Stability at Given Volume, Temperature, and Number of Moles.* Fluid Phase Equilibria **321** (2012), 1-9.

[5] P.E. Gill, W. Murray, M.H. Wright. *Practical Optimization*, Academic Press, London and New York, 1981.

[6] T. Jindrová, J. Mikyška. *Fast and Robust Algorithm for Calculation of Two-Phase Equilibria at Given Volume, Temperature, and Moles.* Fluid Phase Equilibria **353** (2013), 101-114.

[7] T. Jindrová, J. Mikyška. *Phase Equilibria Calculation of* $CO_2 - H_2O$ *System at Given Volume, Temperature, and Moles in* CO_2 *Sequestration.* přijato do IAENG International Journal of Applied Mathematics (2014).

[8] M.L. Michelsen. *The Isothermal Flash Problem. Part I. Stability.* Fluid Phase Equilibria **9** (1982), 1–19.

[9] M.L. Michelsen. *The Isothermal Flash Problem. Part II. Phase-split computation.* Fluid Phase Equilibria **9** (1982), 21-40.

[10] M.L. Michelsen. *State Function Based Flash Specifications.* Fluid Phase Equilibria **158** (1999), 617-626.

[11] M.L. Michelsen, J.M. Mollerup. *Thermodynamic Models: Fundamentals & Computational Aspects.* Tie-Line Publications (2004).

[12] D.Y. Peng, D.B. Robinson. *A New Two-Constant Equation of State.* Industrial & Engineering Chemistry Fundamentals **15(1)** (1976), 59-64.

Weyl Group Orbit Functions in Image Processing*

Ondřej Kajínek

2nd year of PGS, email: kajinond@fjfi.cvut.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Aleš Procházka, Department of Computing and Control Engineering
Faculty of Chemical Engineering, ICT in Prague

Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This article is concerned with the Fourier-like analysis of functions on discrete grids in two-dimensional simplexes using $C-$ and $E-$ Weyl group orbit functions. The convolution theorem is presented for these cases. An example of application of image processing using the $C-$ functions and the convolutions for spatial filtering of the treated image.

The full version of the article was published as G. Chadzitaskos, L. Háková and O. Kajínek, *Weyl Group Orbit Functions in Image Processing*, Applied Mathematics **5** (2014), 501–511.

Keywords: orbit functions, convolution, image processing

Abstrakt. Objektem zájmu tohoto článku je analýza funkce na diskrétních mřížkách ve dvourozměrných simplexech pomocí $C-$ a $E-$ funkcí na orbitách Weylových grup. Pro tyto případy je zobecněn konvoluční teorém. Jako příklad aplikace ve zpracování obrazu je použita $C-$ funkce pro provedení konvoluce jakožto nástroje filtrace obrazu.

Plná verze tohoto příspěvku byla publikována v článku as G. Chadzitaskos, L. Háková and O. Kajínek, *Weyl Group Orbit Functions in Image Processing*, Applied Mathematics **5** (2014), 501–511.

Klíčová slova: funkce na orbitách, konvoluce, zpracování obrazu

References

[1] R. C. Gonzalez, R. E. Woods. *Digital Image Processing*, Addison Wesley Longman, Inc., (1992).

[2] H.G. Granlund a H. Knutsson. *Signal processing for computer vision*, Kluwer Academic Publisher, Dordrecht, The Netherlands, (1995).

[3] N. X. Thao, V.A. Kakichev, V. A., V. K. Tuan, *On the generalized convolutions for Fourier cosine and sine transforms*, East-West J. Math. **1**, (1998).

*Vznik článku byl podpořen Evropskou Unií v rámci programu pro zvýšení kvality výzkumných skupin v České republice

[4] J. Hrivnák, J. Patera, *On discretization of tori of compact simple Lie groups*, J. Phys. A: Math. Theor. **42** (2009).

[5] J. Hrivnák, J. Patera, *On E−discretization of tori of compact simple Lie groups*, J. Phys. A: Math. Theor. **43** (2010).

[6] A. Klimyk, J. Patera, *Orbit functions*, SIGMA (Symmetry, Integrability and Geometry: Methods and Applications) **2** (2006), 006.

[7] A. Klimyk, J. Patera, *Antisymmetric orbit functions*, SIGMA (Symmetry, Integrability and Geometry: Methods and Applications) **3** (2007), paper 023, 83 pages.

[8] A. Klimyk, J. Patera, *E−orbit functions*, SIGMA (Symmetry, Integrability and Geometry: Methods and Applications) **4** (2008), 002, 57 pages.

[9] L. Háková, J. Hrivnák, J. Patera, *Six types of E−functions of the Lie groups* $O(5)$ *and* $G(2)$, J. Phys. A: Math. Theor. **45** (2012).

[10] R. V. Moody, L. Motlochová, and J. Patera, *New families of Weyl group orbit functions*, arXiv:1202.4

Cartanovy podalgebry a řešitelná rozšíření Lieových algeber[*]

Dalibor Karásek

4. ročník PGS, email: dalibor.karasek@fjfi.cvut.cz
Katedra fyziky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Libor Šnobl, Katedra fyziky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. Cartan subalgebras have proven their value when the semisimple Lie algebras were classified. However, they seem to have their utilization also in the classification of solvable Lie algebras. The aim of this text is to present one example how they can be harnessed. At the same time, we try to thoroughly demonstrate the connection of cohomologies of Lie algebras and complete reducibility of representations.

Keywords: Lie algebra, solvable, extensions, cohomology, Cartan subalgebras

Abstrakt. Cartanovy podalgebry mají své velké opodstatnění při klasifikaci poloprostých Lieových algeber. Ukazuje se ovšem, že by mohly jít využít i při klasifikaci řešitelných algeber. Prezentovat jedno takové využití je cílem tohoto příspěvku. Současně se snaží zevrubně ukázat, jak souvisí úplná reducibilita reprezentací a invariantní doplňky s kohomologiemi.

Klíčová slova: Lieova algebra, řešitelná, rozšíření, kohomologie, Cartanova podalgebra

1 Úvod

Klasifikace Lieových algeber je jedním z doposud nedokončených velkých úkolů matematické fyziky. Velký kus práce byl přitom paradoxně odveden už chvíli po jejich objevení. Cartan klasifikoval poloprosté Lieovy algebry a Levi dokázal, že každá Lieova algebra lze rozložit na polopřímý součet poloprosté a řešitelné. V tom okamžiku se vývoj zastavil a už přes sto let se spíše pomalu prodíráme kupředu. Ukazuje se totiž, že klasifikovat řešitelné algebry je dosti obtížný úkol. Množství neekvivalentních tříd totiž s dimenzí velmi rychle narůstá.

Částečný postup byl zaznamenán použitím odlišné metody. Jedná se o řešitelná rozšíření, kdy se ze série nilpotentních algeber podobného typu ale libovolně velké dimenze vytvoří množina algeber řešitelných. Tyto třídy můžeme po jejich klasifikaci využít k pozorování a hledání zajímavých vlastností, které by mohli přispět ke klasifikaci a v dalších oblastech.

Jednou takovou pozorovanou vlastností se zabývá právě tento text. Dříve než se k ní na konci 4. sekce dostaneme, je potřeba definovat základní pojmy, které se týkají modulů, kohomologie Lieových algeber a Cartanových podalgeber. Tomu se věnujeme v 2. sekci.

[*] Tato práce byla podpořena grantem SGS13/217/OHK4/3T/14 ze Studentské grantové soutěže ČVUT.

Velký prostor věnujeme v sekci 3 aplikaci teorie kohomologií na kritérium existence invariantních doplňků k invariantním podprostorům pro reprezentace Lieových algeber. V neposlední řadě je nutno vysvětlit, co myslíme řešitelným rozšířením a jak se takové rozšíření hledá a klasifikuje. Toho se týká začátek sekce 4.

2 Definice

2.1 Moduly

Před samotnou definicí kohomologie se nám bude hodit zadefinovat si moduly, jelikož reprezentace Lieových algeber není nic jiného než teorie jejich modulů. Více podrobností a příkladů týkajících se modulů lze nalézt v monografii [2].

Definice 2.1. O uspořádané trojici $(R, U, \cdot)$ řekneme, že jde o R-modul U, když platí, že

1. Množina U je abelovská grupa.
2. Množina R je okruh.
3. Zobrazení $\cdot : R \times U \to U$ je levou akcí.

Axiomy pro násobení ve vektorovém prostoru V nad tělesem $\mathbb{T}$ nám říkají, že násobením číslem je levou akcí okruhu (s jednotkou). Každý vektorový prostor je tedy $\mathbb{T}$-modul. Máme-li Lieovu algebru L, lze se na ní dívat jako na okruh a její reprezentace na vektorovém prostoru V jsou v 1-1 vztahu s L-moduly na V. Přechod mezi reprezentací ρ a modulem vyjadřuje vztah $x \cdot v = \rho(x)v$ a mezi těmito dvěma popisy budeme často přecházet.

Definice 2.2. Mějme R-moduly U a V. Zobrazení $\varphi : U \to V$ je (R-)homomorfismus modulů, když je aditivní a zároveň platí $\varphi(x \cdot u) = x \cdot \varphi(u)$ pro všechny $x \in R$ a $u \in U$.

Množinu všech R-homomorfismů budeme značit $\mathrm{Hom}_R(U, V)$.

Vrátíme-li se k interpretaci vektorového prosturu nad $\mathbb{T}$ jako $\mathbb{T}$-modulu, tak prvky množiny $\mathrm{Hom}_{\mathbb{T}}(U, V)$ nejsou nic jiného než lineární zobrazení mezi U a V. Podobně když chceme vyšetřovat reprezentace Lieovy algebry L, jsou pro nás zajímavá zobrazení z $\mathrm{Hom}_L(U, V)$.

Věta 2.3. Mějme Lieovu algebru L a dva L-moduly U a V. Množina $\mathrm{Hom}_{\mathbb{T}}(U, V)$ s operací $\cdot_2$ definovanou pobodově $(x \cdot_2 \varphi)(u) := x \cdot_V \varphi(u) - \varphi(x \cdot_U u)$ je také L-modul.

2.2 Kohomologie

Nyní postupme k samotné definici kohomologie Lieovy algebery. Ta vychází z takzvaných Cartanových vztahů pro diferenciální formy na varietě. Od této chvíle uvažujeme pouze *komplexní* Lieovy algebry.

Definice 2.4. Mějme Lieovu algebru L a L-modul V. Prvky množiny všech totálně antisymetrických q-lineární zobrazení s hodnotami ve V ($q \in \mathbb{N}_0$) se nazývají kořetězce stupně q, nebo také q-kořetězce.

$$C^q(L,V) := \mathrm{Hom}_{\mathbb{C}}\left(\bigwedge^q L, V\right), \tag{1}$$

$$C^0(L,V) := V. \tag{2}$$

Dále definujme gradovaný operátor $\mathrm{d}^{(q)}$.

$$\begin{aligned}
\mathrm{d}^{(q)} &: C^q(L,V) \to C^{q+1}(L,V),\\
(\mathrm{d}^{(0)}\,\omega^0)(x) &:= x\cdot\omega^0, \qquad (3)\\
(\mathrm{d}^{(q)}\,\omega^q)(x_0,\dots,x_q) &:= \sum_{i=0}^{q}(-1)^i x_i\cdot\omega^q(x_0,\dots,\hat{x}_i,\dots,x_q)+\\
&+\sum_{i<j}(-1)^{i+j}\omega^q([x_i,x_j],x_0,\dots,\hat{x}_i,\hat{x}_j,\dots,x_q), \qquad (4)
\end{aligned}$$

kde q značí stupeň kořetězce a stříška znamená, že je nutno daný vektor vynechat.

Zobrazení d se označuje jako operátor kohranice a je nilpotentní

$$\mathrm{d}^{(q+1)}\circ\mathrm{d}^{(q)} = 0. \tag{5}$$

To nám umožňuje definovat kocykly Z^q, kohranice B^q a kohomologické grupy $H^q(L,V)$.

$$Z^q(L,V) := \ker\mathrm{d}^{(q)}, \tag{6}$$

$$B^q(L,V) := \mathrm{d}^{(q-1)}(C^{q-1}(L,V)), \tag{7}$$

$$H^q(L,V) := {Z^q(L,V)}\Big/{B^q(L,V)}. \tag{8}$$

Pro různé volby akcí L na V dostáváme různé dimenze kohomologických grup. Ty často zachycují důležité invarianty Lieovy algebry, případně jejího modulu. Jedná se například o dimenzi jádra, množství vnějších derivací, jestli je algebra perfektní a podobně.

2.3 Cartanovy podalgebry

Poslední struktura, která je zapotřebí zadefinovat, je Cartanova podalgebra. Cartanovy podalgebry jsou nesmírně důležité zejména při klasifikaci poloprostých Lieových algeber, ale my je využijeme i při studiu těch řešitelných a nilpotentních. Důkazy vět v této podkapitole a mnohá další tvrzení obsahují například knihy [4, 5].

Definice 2.5. Mějme Lieovu algebru L. Podalgebru C označíme jako Cartanovu, pokud splňuje dvě podmínky:

1. Je nilpotentní.

2. Sama sebe normalizuje, tedy nelze nalézt větší podalgebru W ($C \subset\subset W \subset\subset L$), pro níž by C bylo ideálem.

Tato definice je sice jednoduchá a elegantní, ale nedává nám návod, jak takovou Cartanovu podalgebru najít a zda vůbec existuje. To ovšem řeší následující věta.

Věta 2.6. Buď množina $\ker_\infty A := \{x \in L | \exists n, A^n x = 0\}$ zobecněné jádro operátoru $A \in \mathrm{Hom}_\mathbb{C}(L, L)$. Označme $\mathcal{C} = \{\ker_\infty \mathrm{ad}_x | x \in L\}$ množinu zobecněných jader vnitřních derivací.

Pak
$$C \text{ je Cartanova podalgebra} \iff C \in \underset{B \in \mathcal{C}}{\arg\min}\,(\dim B)\,. \tag{9}$$

Tato věta tvrdí, že Cartanovy podalgebry jsou právě zobecněná jádra vnitřních derivací ad_x (pro většinu vektorů x), což jsou objekty, se kterými se pracuje mnohem lépe, a navíc můžeme volit vektor x tak, aby nám to vyhovovalo. Také je vidět, že Lieova algebra má většinou vícero Cartanových podalgeber, ale všechny mají stejnou dimenzi.

3 Kohomologie a reducibilita

Pro zobecnění výsledků se nám bude hodit způsob, jak kvantifikovat úplnou reducibilitu pomocí kohomologie. Při zpracování tohoto tématu jsme vycházeli z [3].

Definice 3.1. Mějme Lieovu algebru L. Její reprezentace ρ na V (L-modul) je *reducibilní*, když existuje netriviální společný invariantní podprostor pro $\rho(L)$. Pokud pro každý společný invariantní podprostor existuje invariantní doplněk, říkáme, že je tato reprezentace *úplně reducibilní.*

Mějme nyní Lieovu algebru L a její reducibilní reprezentaci ρ na V. Nechť dále W je invariantní podprostor reprezentace ρ, neboli $\rho(L)W \subset W$. Zvolme doplněk U. Vektorový prostor jde rozložit $V = W \dotplus U$ a reprezentace jde rozepsat blokově jako

$$\rho(x) = \begin{pmatrix} \rho_W(x) & \lambda_U(x) \\ 0 & \rho_U(x) \end{pmatrix}, \tag{10}$$

kde bloky jdou chápat jako $\rho_W \in \mathrm{Hom}_\mathbb{C}(W, W)$, $\rho_U \in \mathrm{Hom}_\mathbb{C}(U, U)$ a $\lambda_U \in \mathrm{Hom}_\mathbb{C}(U, W)$.

Tvrzení 3.2. Zobrazení ρ definované blokově vztahem (10) je reprezentace s invariantním podprostorem W, právě když blokové komponenty mají tyto tři vlastnosti:

1. Zobrazení ρ_W je reprezentace na W.

2. Zobrazení ρ_U je reprezentace na U.

3. Pro všechna $x, y \in L$ platí vztah

$$\rho_W(x) \circ \lambda_U(y) - \lambda_U(y) \circ \rho_U(x) - (\rho_W(y) \circ \lambda_U(x) - \lambda_U(x) \circ \rho_U(y)) - \lambda_U([x, y]) = 0. \tag{11}$$

Jelikož W i U jsou L-moduly, lze využít větu 2.3 a přepsat vztah (11) jako

$$x \cdot_2 \lambda_U(y) - y \cdot_2 \lambda_U(x) - \lambda_U([x,y]) = 0, \tag{12}$$

což není nic jiného než požadavek, aby λ_U byl kocyklus z $Z^1(L, \mathrm{Hom}_{\mathbb{C}}(U,W))$.

Doplněk U je možno samozřejmě zvolit jinak a naším cílem je zjistit, jestli ho lze zvolit tak, aby bylo λ_U nulové zobrazení.

Různé volby doplňku (a báze v něm) můžeme popsat také pomocí zobrazení z prostoru $\mathrm{Hom}_{\mathbb{C}}(U,W)$.

Věta 3.3. Doplňky U a $\tilde{U}$ jsou isomorfní pomocí vztahu

$$\tilde{U} = \mathbb{Y}U = (\mathbb{1} - \tilde{\mathbb{A}})U, \tag{13}$$

kde $\tilde{\mathbb{A}} \in \mathrm{Hom}_{\mathbb{C}}(U,W)$. A každé dva doplňky lze takto popsat.

Důkaz. Neprve ukážeme, že takto definované $\tilde{U}$ je opět doplněk. Zobrazení $\mathbb{Y}$ je operátor na V a jeho zúžením na podprostor U bychom měli dostat prosté zobrazení.

Buď $u \in \ker \mathbb{Y}$. Potom

$$0 = \mathbb{Y}u = (1 - \tilde{\mathbb{A}})u = u - \tilde{\mathbb{A}}u. \tag{14}$$

Máme rozklad nulového vektoru do direktního součtu $U \dot{+} W$. Obě složky musí být nulovými vektory. Tedy jak u, tak $\tilde{\mathbb{A}}u$. Vektor u tedy je nulový vektor, zobrazení $\mathbb{Y}$ je prosté a z definice $\tilde{U}$ ve vztahu (13) plyne, že $\dim U = \dim \tilde{U}$.

Nyní stačí dokázat, že $W \cap \tilde{U} = \{0\}$ a z dimenzionální analýzy (první věta o dimenzi) plyne, že $\tilde{U}$ je doplněk.

Mějme $x \in W \cap \tilde{U}$. Jelikož $x \in \tilde{U}$, jde najít jeho $\mathbb{Y}$-vzor u a zapsat ho jako $x = u + \tilde{\mathbb{A}}u$. Tím jsme ovšem nalezli zároveň rozklad vektoru x do direktního součtu $W \dot{+} U$, a jelikož předpokládáme, že $x \in W$, musí být jeho část patřící do U nulová, a tedy $u = 0$, což okamžitě implikuje $x = u + \tilde{\mathbb{A}}u = 0 + 0 = 0$.

Pro opačný směr musíme najít $\tilde{\mathbb{A}} \in \mathrm{Hom}_{\mathbb{C}}(U,W)$ pro zadaný doplněk $\tilde{U}$. Ukazuje se, že tím správným $\tilde{\mathbb{A}}$ je projektor $\mathbb{X}$ na W podle $\tilde{U}$, který zúžíme na U.

Zaprvé $\mathbb{Y} := \mathbb{1} - \mathbb{X}$ je projektor na $\tilde{U}$ podle W a platí tedy $\tilde{U} \subset \mathbb{Y}(U)$ a pokud $\tilde{u} \in \tilde{U}$, pak lze vzít jeho rozklad $\tilde{u} = w + u$ do $W \dot{+} U$ a hledaným vzorem pro důkaz opačné inkluze bude u.

$$\mathbb{Y}(u) = \mathbb{Y}(\tilde{u} - w) = \mathbb{Y}\tilde{u} - \mathbb{Y}w = \tilde{u} - 0 = \tilde{u}, \tag{15}$$

kde jsme využili vlastností projektoru. □

Nyní si vezmeme reprezentaci ρ, zadanou pomocí $(\rho_W, \rho_U, \lambda_U)$, přejdeme k jinému doplňku $\tilde{U}$ pomocí zobrazení $\tilde{\mathbb{A}}$ a ztotožníme ho s U pomocí vztahu (13).

Ukazuje se, že reprezentace se změní a je popsaná

$$(\rho_W, \rho_{\tilde{U}}, \lambda_{\tilde{U}}) = (\rho_W, \rho_U, \lambda_U - \mathrm{d}^{(1)}\,\tilde{\mathbb{A}}), \tag{16}$$

takže změna nastala pouze v poslední komponentě, která se liší o kohranici z prostoru $B^1(L, \mathrm{Hom}_{\mathbb{C}}(U,W))$.

Dospěli jsme tedy k tomu, že λ_U musí být kocyklus a při změně doplňku se toto zobrazení změní o kohranici. S trochou další práce, která je víceméně technickou záležitostí, z toho plyne, že každé reprezentaci L na V s invariantním podprostorem W odpovídá jedna třída ekvivalence z $H^1(L, \mathrm{Hom}_{\mathbb{C}}(V \,/\, W, W))$. Pokud je tato třída ekvivalence nulová, lze nalézt invariantní doplněk.

Důsledek 3.4. Pokud $H^1(L, \mathrm{Hom}_{\mathbb{C}}(V \,/\, W, W))$=0, lze nalézt invariantní doplněk pro libovolnou reprezentaci se zadaným ρ_W a $\rho_{V\,/\,W}$.

Z tohoto důsledku mimojiné plyne známý fakt, že poloprosté algebry jsou úplně reducibilní, neboť Whiteheadovo lemma tvrdí, že $H^1(L, V)$ pro poloprostou algebru L je vždy triviální.

4 Řešitelná rozšíření

4.1 Definice řešitelného rozšíření

Tato sekce se zabývá řešitelnými rozšířeními nilpotentních algeber. Další podrobnosti včetně velkého množství příkladů lze nalézt v monografii [11]. Mějme nilpotentní Lieovu algebru N dimenze n. Zajímají nás všechny řešitelné algebry S, jejichž nilradikál je s N isomorfní. Tyto algebry nazýváme *řešitelná rozšíření algebry N*.

Pokud v N zvolíme bázi $\mathcal{E} := (e_1, \ldots, e_n)$, můžeme definovat strukturní koeficienty ${c^i}_{jm}$ vztahem $[e_j, e_m] = {c^i}_{jm} e_i$. Naším cílem při řešitelném rozšíření o k prvků je definovat násobení na vektorovém prostoru $N \dotplus \mathbb{C}^k$. Nechceme ho však definovat libovolně, ale tak, aby se na N shodovalo s původní Lieovou závorkou. Zároveň bychom byli rádi, aby N byl ideál a výsledná algebra byla řešitelná, díky čemuž nám stačí přidat dodatečné strukturní konstanty $(D_a)^i_k$ a γ^i_{ab}, kde $a, b = 1, \ldots, k$, a pomocí nich definovat násobení na bázi $(e_1, \ldots, e_n, s_1, \ldots, s_k)$, vzniknuvší sloučením báze $\mathcal{E}$ a báze v $\mathbb{C}^k$:

$$[e_j, e_m] := {c^i}_{jm} e_i, \tag{17}$$

$$[s_a, e_j] := {(D_a)^i}_j e_i, \tag{18}$$

$$[s_a, s_b] := {\gamma^i}_{ab} e_i. \tag{19}$$

Strukturní konstanty samozřejmě nelze volit libovolně. Koneckonců chceme, abychom dostali Lieovu algebru. Násobení musí být antisymetrické a splňovat Jacobiho identitu. Antisymetrie implikuje ${c^i}_{jm} = -{c^i}_{mj}$, což je splněno díky tomu, že N Lieova algebra už byla, a ${\gamma^i}_{ab} = -{\gamma^i}_{ba}$. Všimněme si, že D_a lze interpretovat jako lineární operátor na N a γ jako lineární zobrazení $\mathbb{C}^k \wedge \mathbb{C}^k \xrightarrow{\gamma} N$.

Jacobiho identita

$$[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0 \tag{20}$$

musí být splněna pro všechny vektory x, y, z. Zjevně stačí ověřit platnost na bázi. Podobně jako v případě antisymetrie, pokud za x, y, z zvolíme vektory z N, nedá nám Jacobiho identita žádná nová omezení, protože N už Lieova algebra byla.

Na druhou stranu ostatní Jacobiho identity (JI) nám omezí volbu D_a a γ. Pro trojice typu s_a, e_j, e_m jsou Jacobiho identity ekvivalentní s požadavkem, aby D_a jako lineární operátor byl derivace, a JI typu s_a, s_b, e_i nám svážou D_a a γ, neboť je lze ekvivalentně zapsat jako

$$[D_a, D_b] = \mathrm{ad}_{\gamma(s_a,s_b)}, \tag{21}$$

kde závorkami na levé straně rovnice je myšlen komutátor operátorů. Tato rovnost v sobě také skrývá podmínku, že komutátor $[D_a, D_b]$ musí být vnitřní derivace pro všechny a, b.

Poslední Jacobiho identity (pro vektory čistě z $\mathbb{C}^k$) nám ještě víc provážou derivace D_a a zobrazení γ. S využitím antisymetrie γ je můžeme ekvivalentně zapsat jako

$$D_a\gamma(s_b, s_c) - D_b\gamma(s_a, s_c) + D_c\gamma(s_a, s_b) = 0. \tag{22}$$

Dospěli jsme tedy k poznatku, že libovolná množina $(D_1, \ldots, D_k; \gamma)$, pokud splňuje Jacobiho identity, definuje rozšíření.

Zbývá vyřešit poslední věc. Po našem řešitelném rozšíření vyžadujeme, aby N bylo nilradikálem, jinak bychom totiž neměli shora omezenou dimenzi S, nehledě k tomu, že by se nám při klasifikaci vytvořily duplicitní třídy. Je tedy potřeba, abychom přidáním vektorů s_a nezvětšili nilradikál. Toho se vyhneme tím, že naložíme ještě jednu dodatečnou podmínku na derivace $D_1, \ldots, D_k$: Musí být lineárně nilnezávislé (linearly nilindependent), t.j. jediná lineární kombinace, která z nich vytvoří nilpotentní zobrazení, je ta triviální.

4.2 Klasifikace řešitelných rozšíření

Když teď víme, jak rozšířit nilpotentní algebru, věnujme se chvíli klasifikaci těchto rozšíření. Dvě množiny dat $(D_1, \ldots, D_k; \gamma)$ a $(\tilde{D}_1, \ldots, \tilde{D}_k; \tilde{\gamma})$ popisující řešitelná rozšíření jsou pro nás (slabě) ekvivalentní, právě když vzniklá rozšíření S a $\tilde{S}$ jsou isomorfní jako Lieovy algebry. Tato definice nám generuje tři operace s daty, které nám nezmění třídu ekvivalence.

1. Můžeme vybrat v lineárním obalu $\mathrm{span}\{D_1, \ldots, D_k\}$ jinou bázi.

2. Můžeme přejít pomocí Φ, automorfismu na N, k datům $(\Phi D_1 \Phi^{-1}, \ldots, \Phi D_k \Phi^{-1}; \Phi\gamma)$.

3. Můžeme k libovolnému D_a přičíst libovolnou vnitřní derivaci na N a současně upravit γ tak, aby stále splňovalo (21). Například $\tilde{D}_1 := D_1 + \mathrm{ad}_{e_3}$ a $\tilde{\gamma}(s_1, s_a) := \gamma(s_1, s_a) - D_a(e_3)$.

V praxi většinou probíhá klasifikace tím způsobem, že nejprve klasifikujeme řešitelná rozšíření o jeden vektor. Tam stačí zklasifikovat vnější derivace pomocí druhé a třetí operace—v podstatě klasifikujeme třídy vnějších derivací pomocí automorfismů modulo vnitřní derivace. Potom zkoumáme dvoudimenzionální podprostory lineárně nilnezávislých derivací a využijeme již získaných výsledků k výběru vhodného reprezentanta D_1, kterého pomocí automorfismů upravíme do jednoduchého tvaru. Následně zjistíme, jak nám podmínka (21) a především její důsledek o tom, že komutátor $[D_1, D_2]$ musí být vnitřní derivace, omezí tvary D_2. Prozkoumáme různé možnosti výběru γ a využijeme zbylé ekvivalentní transformace. Dál postup opakujeme s trojdimenzionálními podprostory…

V průběhu klasifikace se dá s výhodou využít několika faktů. Prvním je, že algebra všech derivací je Lieova algebra grupy všech automorfismů. Druhým je, že řešitelné algebry kanonicky obsahují svaz ideálů, které jsou zároveň invariantními podprostory pro všechny automorfismy (a derivace). Zároveň jsou koeficienty automorfismů a derivací určeny malým množstvím parametrů, které popisují chování těchto zobrazení na vektorech, které patří do doplňku k takzvané derivované algebře $[N, N]$. Tyto jevy jsou detailněji zpracovány v [7] a [1].

4.3 Předpovídané vlastnosti řešitelných rozšíření

Metoda hledání řešitelných rozšíření byla použita profesorem Winternitzem a jeho spolupracovníky k vytvoření velké množiny klasifikovaných řešitelných Lieových algeber libovolně velké dimenze, vycházejících ze speciálních tříd nilpotentních algeber. Jedná se například o Heisenbergovy algebry [6], Borelovy algebry [10], nebo algebry filiformní [9] a jim podobné [8]. Tento seznam samozřejmě není vyčerpávající, ale umožňuje nám pozorovat společné vlastnosti a vyslovit hypotézy o vlastnostech těchto rozšíření.

Jedna z těchto hypotéz vychází z pozorování, že při klasifikaci šla vždy zvolit taková báze, že přidané vektory spolu téměř komutovaly.

Hypotéza 4.1. V každé řešitelné algebře lze zvolit bázi $(e_1, \ldots, e_n, s_1, \ldots, s_k)$ tak, že prvních n vektorů patří do nilradikálu a výsledek Lieovy závorky zbylých k bazických vektorů leží v centru nilradikálu.

Pro velkou třídu řešitelných rozšíření se nám povedlo najít příčinu, proč mají tuto vlastnost. Ukazuje se, že platí tvrzení

Věta 4.2. Nechť S je řešitelné rozšíření algebry N, a C jeho Cartanova podalgebra. Pak C působí na S ad-reprezentací a platí, že

$$H^0\left(C, \mathrm{Hom}_{\mathbb{C}}\left({}^{S}\big/_{N}, N\right)\right) = 0 \Longrightarrow S = N \dotplus A, \tag{23}$$

kde A je abelovská podalgebra. To znamená, že lze zvolit bázi tak, že přidané vektory komutují.

Důkaz. Začněme tím, že si ujasníme, jak vypadá ad-reprezentace na $S\,/\,N$. Jelikož S je řešitelná algebra a N je její nilradikál, musí být reprezentace na faktorprostoru $S\,/\,N$ triviální (plyne to z faktu, že $[S, S] \subset N$).

Dále využijeme známý výsledek pro nilpotentní algebry, který lze nalézt například v [4]. Ten tvrdí, že pro nilpotentní algebry—a Cartanova algebra je nilpotentní z definice—je nultá kohomologická grupa triviální právě tehdy, je-li triviální první kohomologická grupa.

Víme tedy, že $H^1(C, \mathrm{Hom}_{\mathbb{C}}(S\,/\,N, N)) = 0$. Tato kohomologická grupa nám ale popisuje, zda existuje invariantní doplněk k N. Použijeme tedy důsledek 3.4 a vidíme, že lze invariantní doplněk U nalézt. Reprezentace je poté v blokově diagonálním tvaru

$$\rho(x) = \begin{pmatrix} \rho_N(x) & 0 \\ 0 & 0 \end{pmatrix}. \tag{24}$$

Z něho plyne, že pro všechna $x \in C$ a všechna $u \in U$, platí

$$\rho(x)u = [x, u] = \mathrm{ad}_x\, u = 0 \tag{25}$$

a u leží v jádru vnitřní derivace ad_x. Speciálně tam leží i pro $c \in C$, pomocí kterého je popsaná Cartanova podalgebra jako $C = \ker_\infty \mathrm{ad}_c$. To ovšem implikuje, že $U \subset C$.

V dalším kroku znovu využijeme vztah (25), tentokrát za vektor x z Cartanovy podalgebry zvolíme také vektor z U, a $\rho(u_1)u_2 = 0$ nám dokazuje, že U je ona hledaná abelovská podalgebra A. □

Tuto větu nyní využijeme na případ, kdy řešitelné rozšíření lze popsat daty $(D_1, \ldots, D_k; \gamma)$, kde první derivace D_1 je regulární zobrazení. Ač se může zdát, že je to dosti omezující předpoklad, ukazuje se, že ve skutečnosti popisuje velké množství případů. Už jenom proto, že můžeme permutovat derivace D_i, případně dělat jejich lineární kombinace.

Nyní můžeme najít Cartanovu podalgebru jako zobecněné jádro ad_{s_1}. To nám zajistí, že vektor s_1 bude v naší Cartanově podalgebře. Uvažovaná ad-reprezentace algebry C, zúžená na ideál N tak obsahuje regulární derivaci $\rho(s_1)\restriction_N = D_1$.

Dokažme nyní, že $H^0\left(C, \mathrm{Hom}_{\mathbb{C}}\left(S \,/\, N, N\right)\right) = 0$. Jelikož se jedná o nultou kohomologickou grupu, neobsahuje žádné kohranice a stačí zjistit, co jsou kocykly.

$$\begin{aligned} \mathbb{A} \in Z^0\left(C, \mathrm{Hom}_{\mathbb{C}}\left({}^{S}\!\big/_{N}, N\right)\right) &\Leftrightarrow \forall x \in C, (\mathrm{d}\,\mathbb{A})(x) = 0 \\ 0 = x \cdot_2 \mathbb{A} = \rho_N(x) \circ \mathbb{A} - \mathbb{A} \circ \rho_{S\,/\,N}(x) &= \rho_N(x) \circ \mathbb{A} - \mathbb{A} \circ 0 = \rho_N(x) \circ \mathbb{A}. \end{aligned} \tag{26}$$

Tato rovnost musí platit pro všechny vektory $x \in C$, musí tedy platit i pro s_1, ale $\rho_N(s_1) = D_1$ je regulární operátor, takže z toho plyne, že $\mathbb{A} = 0$ a jediným kocyklem je nulové zobrazení. Z toho plyne trivialita nulté kohomologické grupy.

5 Závěr

V tomto příspěvku jsme definovali základní pojmy týkající se modulů, kohomologií, Cartanových podalgeber a řešitelných rozšíření. Poté jsme použili kohomologické metody, abychom odhalili příčiny jednoduchého tvaru velké třídy řešitelných rozšíření. Ve větě 4.2 jsme zformulovali postačující podmínku a poté jsme ji použili pro speciální, ale velmi často se vyskytující případ.

Pro úplnost dodejme, že pro tento konkrétní případ není třeba budovat celou mašinerii kohomologií, a lze si v zásadě vystačit s Jordanovým tvarem matice. Pokud ovšem hodláme využít tento jednoduchý příklad jako výchozí bod pro další pr8ci, je vhodné formulovat tvrzení tak, aby šly snáze zobecnit. A tvrzení o „současném převedení vícero matic do Jordanova tvaru“ je zachyceno právě pomocí doplňků a kohomologií.

Vypadá to, že Cartanovy podalgebry se hodí nejenom ke klasifikaci poloprostých Lieových algeber, ale lze je s výhodou využít i pro ty řešitelné.

Literatura

[1] P. Benito and D. de-la-Concepción. *On Levi extensions of nilpotent Lie algebras.* Linear Algebra Appl. **439** (2013), 1441–1457.

[2] D. S. Dummit and R. M. Foote. *Abstract algebra.* John Wiley & sons, Hoboken, NJ, (2004).

[3] M. Goto and F. D. Grosshans. *Semisimple Lie algebras.* Marcel Dekker Inc, New York, (1978). Lecture Notes in Pure and Applied Mathematics, Vol. 38.

[4] J. Hilgert and K.-H. Neeb. *Structure and geometry of Lie groups.* Springer, New York; London, (2012).

[5] J. E. Humphreys. *Introduction to Lie Algebras and Representation Theory.* Springer-Verlag, Berlin, (1994).

[6] J. L. Rubin and P. Winternitz. *Solvable Lie algebras with Heisenberg ideals.* J. Phys. A **26** (1993), 1123–1138.

[7] L. Šnobl. *On the structure of maximal solvable extensions and of Levi extensions of nilpotent Lie algebras.* J. Phys. A **43** (2010), 505202, 17 pp.

[8] L. Šnobl and D. Karásek. *Classification of solvable Lie algebras with a given nilradical by means of solvable extensions of its subalgebras.* Linear Algebra Appl. **432** (2010), 1836–1850.

[9] L. Šnobl and P. Winternitz. *A class of solvable Lie algebras and their Casimir invariants.* J. Phys. A **38** (2005), 2687–2700.

[10] L. Šnobl and P. Winternitz. *Solvable Lie algebras with Borel nilradicals.* J. Phys. A **45** (2012), 095202, 18 pp.

[11] L. Šnobl and P. Winternitz. *Classification and Identification of Lie Algebras.* AMS and Centre de Recherches Mathématiques, Providence, (2014).

Effective Implementation of Grid Coarsening Algorithm for Algebraic Multigrid on GPU

Vladimír Klement

3rd year of PGS, email: wlada@post.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Tomáš Oberhuber, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This article deals with coarsening algorithm used for algebraic multigrid method in open source library Paralution and its effective implementation on GPU. One of the most important problems when creating algebraic multigrid is the way of obtaining translation operators and coarser version of the system matrix, which is basically a graph clustering problem. Even though there exist a number of ways how to solve this task, most of them are inherently sequential and therefore cannot be directly used on graphics cards. In this article we tried to find effective way how to convert one such method on the GPU, despite its sequential nature. This implementation was successful and we were able to obtain performance comparable to the CPU version.

Keywords: GPU, Algebraic Multigrid, Coarsening

Abstrakt. Tento článek se zabývá algoritmem zhrubování sítě pro potřeby metody algebraického multigridu a jeho implementací na GPU. Jeden ze základních problémů při implementaci algebraického multigridu je způsob, jakým získat operátory přenosu mezi hrubší a jemnější sítí respektive hrubší matici systému. V základu se jedná o problem rozdělení grafu do několika komponent. Ačkoliv existuje množství algoritmů, které takovouto úlohu řeší, většina z nich je sekvenční a nejde proto přímo použít na grafické kartě. Tento článek se bude zabývat právě efektivní implementací jedné takové metody na GPU. Danou metodu se podařilo úspěšně implementovat a dosáhnout výkonu srovnatelného s verzí na procesoru.

Klíčová slova: GPU, Algebraický Multigrid, Zhrubování

1 Introduction

Algebraic multigrid methods are a group of algorithms for solving linear systems from differential equations using a hierarchy of matrices. Their main advantage compared to geometric multigrid is that they are able to create coarser matrices on their own just from the original system matrix. This coarsening is however quite complex problem, with unknown optimal solution. Therefore there exist number of methods how to solve it. Moreover most of them are inherently sequential and therefore cannot be directly used on parallel architectures. This is not much of an issue for standard multicore processors because normally only small amount of time is spent in the creation of coarser system during the computation, so it does not hinder the performance even though it isn't parallelized. However it can be quite a bottleneck on architectures with dedicated memory such as graphics card, because if coarsening must run on the processor although rest of the program runs on the GPU, data must be constantly shifted between CPU and GPU memory, which can significantly reduce performance.

We are currently working with open source library *Paralution*[4], where this is exactly the case. The library contains numerous methods for solving linear systems, which are mostly implemented also on GPU and can be launched there to accelerate the computation. One of the methods is even algebraic multigrid, where however the coarsening algorithm is implemented only on CPU. Therefore if many matrices has to be solved during the computation, e.g. if evolutionary problems are solved, data need to be often copied between memories, which spoils the benefits of GPU acceleration.

To improve this situation we tried to convert the coarsening part on the GPU. The algorithm is quite simple, but it is inherently sequential which makes its implementation complicated. On the other hand it is memory bandwidth limited, so it should be alright if GPU computation performance is not fully utilized as long as we can use its bigger memory bandwidth. Moreover we do not need to speed up the coarsening computation, performance comparable to CPU should suffice as the main contribution will be that the data transfers can be eliminated.

2 Algebraic multigrid

We will be interested mainly in the setup phase of AMG, where the matrix hierarchy and transitions operators are created. It consist of the following steps:

- Variables clustering
- Defining transition operators
- Creating coarse problem matrix

which will be described in more detail.

Once the problem hierarchy is created the main iteration is same as in the case of geometric multigrid and so any standard multigrid cycle can be used to obtain the final solution.

2.1 Variables clustering

First and most important part is to divide the variables into clusters (groups), so that all variables in the same group are joined to one variable on the coarser level. To achieve this we will need to define *strong dependence*:

Definition 1: Given a threshold value $0 < \theta \leq 1$, the variable (unknown) u_i *strongly depends* on the variable u_j if

$$abs(a_{ij}) \geq \theta abs(a_{ii}). \tag{1}$$

This means that variable u_i strongly depends on the variable u_j if the coefficient a_{ij} is comparable in magnitude to the diagonal coefficient in the ith equation.

Now we can proceed to the variable clustering itself. It is basically graph clustering problem where the graph is created in following way:

1. All variables $u_i, i \in 0...N$ become nodes
2. There is oriented edge from u_i to u_j if u_i strongly depends on u_{ij}.

To divide resulting graph into components Paralution uses simple approach that works as follows:

1. Create array *groupId* of group ids for each variable, group id for variable u_i will be denoted as $groupId_i$
2. Set all group ids $groupId_i$ to $undefined$
3. Set group counter *last_id* to -1.
4. For each node u_i do:
   (a) If $groupId_i$ is not $undefined$ continue with next u_i
   (b) Increase group counter *last_id* by one
   (c) Set $groupId_i$ to *last_id*
   (d) For each neighbor(u_j) of u_i, if $groupId_j$ is not *removed* set it to *last_id*
   (e) For each neighbor(u_k) of each neighbor u_j, if $groupId_k$ is $undefined$ set it to *last_id*

2.2 Defining transition operators and coarser system matrix

When the clusters have been selected, the next goal is to define transition operators. Each cluster will form one variable on the coarser grid. In this case prolongation operator (I^h_{2h}) will simply distribute value from coarser variable to all its finer descendants i.e. the ith component of $I^h_{2h}e$ is

$$(I^h_{2h}e)_i = e_{gi}, \tag{2}$$

where e is the vector that should be prolonged to the finer grid and gi is group number for given variable u_i so $gi = groupId_i$.

Restriction operator can be then constructed from the interpolation one by simple transpose:

$$I^{2h}_h = (I^h_{2h})^T, \tag{3}$$

and restricted matrix is produced by

$$A^{2h} = I^{2h}_h A^h I^h_{2h}. \tag{4}$$

3 GPU programming

GPU is shared memory parallel architecture so all threads that run on it use the same memory. Unlike multi-core programming where there are typically 2-32 computational cores running at once, GPU can spawn hundreds of concurrently running threads. These threads are, however, not completely independent and all run the same function (called *kernel*) so it is the SIMD (simple instruction multiple data) type of architecture.

There are some key principles which must be taken into account when creating program for GPU, which come from the type of calculations graphics cards were designed for. The most important are:

Limited communication Computational threads form a two layer hierarchy. On first one threads are grouped to blocks, and on second all blocks create the so called grid. Number of blocks in the grid is completely up to the programmer and it should match the size of the solved problem. Size of the block can be also chosen, however it must be less than 513. The reason for this two level hierarchy is that only threads that are in the same block can communicate between each other. This means that blocks have to be completely independent.

Branching Threads on the GPU aren't completely independent, groups of 32 threads in the same block forms the so called *warp*. Threads in the warp has to always execute same instruction at the same time or wait, so if the kernel contains divergent branches and not all threads in the warp take the same one, complete computational time for each thread will be equal to the sum of all taken branches.

Coalescing Very important feature for numerical computation on GPU is the *coalescing*. Graphics card have much bigger bandwidth than standard RAM when reading blocks of data. More precisely when half warp (16 consecutive threads) try to read or write continuous block of data it can be coalesced into single operation and so whole block can be loaded more than ten times faster. Since most numerical applications are limited by memory accesses, utilizing this feature is absolutely crucial when implementing numerical problems on GPU. There are several ways how coalescing can be achieved even when data aren't naturally read in right order:

- Best solution, if it is possible, is to reorder data so that access to them will be coalesced. One classic example is to use structure of arrays instead of array of structures (i.e. group data by type, not by the thread they belong to).

- Threads in the same block can pre-fetch data to shared memory (shared within block), even random accesses to this memory are very cheap. This is especially useful when needed data form a continuous region, but are accessed randomly.

- If data are needed to be ordered differently in different kernels they can be duplicated (unless memory is a strong concern) this can be especially useful in the case of constant data (for example data describing mesh on which problem is solved).

Transports between GPU and CPU memory GPU don't use same memory as CPU, it has its own video RAM (VRAM). This isn't issue when problem is completely solved on GPU, but in case of converting only most computational demanding parts on GPU and doing rest of the work on processor, constant copying can cause a significant overhead.

4 GPU implementation

The GPU version of the coarsening algorithm was implemented in CUDA, which is a technology from NVidia company designed for general purpose programming on GPU, same as rest of the Paralution GPU code. The coarsening algorithm itself consist of several parts that were parallelized individually. Namely they were

- Deciding which neighbors are strongly connected
- Removing nodes without neighbors
- Creating the groups
- Creating prolongation operator
- Creating restriction operator and coarser system matrix

which will be now subsequently described.

4.1 Finding strongly connected neighbors

This means for each connection between two variables (therefore for each $A_{i,j}, i \neq j$) decide whether it is strong connection or not. This is done by comparison of value $A_{i,j}$ to diagonal element $A_{i,i}$. In CPU version the diagonal is firstly obtained by Paralution's function *ExtractDiagonal*(which works also on GPU) and then simple for cycle is used. The for cycle is easily parallelizable, so same approach was used on GPU in kernel

```
template <typename ValueType>
__global__ void kernel_csr_amg_connect(const int nrow, const int *row_offset,
    const int * cols, const ValueType * vals, const ValueType * vec_diag,
    int *cast_conn, ValueType theta2)
{
  int i = blockIdx.x*blockDim.x+threadIdx.x;
  if (i >= nrow) return;

  ValueType theta_dia_i = theta2 * vec_diag[i];
  for (int j=row_offset[i]; j<row_offset[i+1]; ++j) {
    int       c = cols[j];
    ValueType v = vals[j];
    //Strong connection if not diagonal element and bigger then theta*diagonal
    cast_conn[j] = (c != i) && (v * v > theta_dia_i * vec_diag[c]);
  }
}
```

Result is stored as a mask for all $A_{i,j}$ elements in variable *cast_con*

4.2 Removing nodes without neighbors

Also this part was quite easy it consist in loop over each unknown and setting its *group_id* either to *undefined* if it has some strongly connected neighbors or to *removed* if it has not. Such kernel can look as follows:

```
__global__ void kernel_csr_amgaggregate_remove_empty(const int nrow,
    const int *row_offset, int *cast_conn, int* cast_agg)
{
  int ai = blockIdx.x*blockDim.x+threadIdx.x;
  int aj;

  if (ai <nrow) {
    int state = removed;
    for (aj=row_offset[ai]; aj<row_offset[ai+1]; ++aj) {
      if (cast_conn[aj]>0) state = undefined;
    }
    cast_agg[ai]=state;
  }
}
```

Resulting group ids are stored in variable *cast_agg*.

4.3 Creating the groups

This part was hardest to implement because it is the sequential one. The algorithm works as described in section 2.1 and therefore new group cannot be created unless the previous one was already fully processed. To overcome this issue we use only one block of threads which can be manually synchronized. The *last_id* index is managed by only one thread (the one with $id == 0$) and others are used only to spread the current value to all neighbors and neighbors of neighbors. To employ GPU memory throughput, we tried to use coalescing for most of the memory accesses:

```
__global__ void kernel_csr_amgaggregate_plain(const int nrow,
    const int *row_offset, const int*cols,
    int *cast_conn, int* cast_agg)
{
  const int BLOCK_SIZE=512;
  int id = threadIdx.x;
  __shared__ int last_g;
  __shared__ int s_ca[BLOCK_SIZE];
  __shared__ int s_ro[BLOCK_SIZE+1];

  if (id==0) last_id = -1;
  for (int actOff=0; actOff < nrow; actOff+=BLOCK_SIZE)
  {
    int i = actOff+id;
    s_ca[id] = cast_agg[i];
    s_ro[id] = row_offset[i];

    if (id==0) s_ro[BLOCK_SIZE] = row_offset[actOff + BLOCK_SIZE];
    __syncthreads();

    for (int j=0; j < BLOCK_SIZE; j++)
    {
      __syncthreads();
      if (s_ca[j] != undefined) continue;
      if (id==0) s_ca[j] = ++last_id; //New seed
      __syncthreads();

      // Include its neighbors as well.
      for (int i1 = s_ro[j]+id, end1=s_ro[j+1]; i1 < end1; i1+=BLOCK_SIZE)
      {
        int c = cols[i1];
        int * ca = (c>=actOff && c<actOff+BLOCK_SIZE)?
         &s_ca[c-actOff] : &cast_agg[c];
        if (cast_conn[i1] && *ca != removed)
        {
          *ca=last_id;

          //Neigbours of neigbours
          for (int i2=row_offset[c], end2=row_offset[c+1]; i2<end2; i2++){
            if (cast_conn[i2]==0) continue; //Not valid link
            int c2 = cols[i2];
            int * ca2 = (c2>=actOff && c2<actOff+BLOCK_SIZE)?
              &s_ca[c2-actOff] : &cast_agg[c2];
            if ( *ca2 == undefined)
              *ca2=last_id;
          }
        }
      }
      __syncthreads();
    }
    int i = actOff+id;
    cast_agg[i] = s_ca[id];
    __syncthreads();
  }
}
```

4.4 Creating interpolation operator

The main part of the algorithm for creation of prolongation operator consist from the following code

```
for (int i=0, j=0; i < nrow; ++i)
{
    if (cast_agg[i] >= 0) {
      col[j] = cast_agg->vec_[i];
      val[j] = 1.0;
      ++j;
    }
}
```

which for each fine unknown u_i with set group number gi creates one line with only one 1 value on the G_i position. Therefore prolongation is done so that value of coarse point given by gi is simply copied to the u_i. This part was not yet implemented on the GPU but should be quite straightforward. One block of threads will be employed, all threads will be used to save and write data to/from global memory, but the computation itself will be done only by one thread and will use shared memory.

4.5 Creating restriction operator and coarser system matrix

Last part was quite easy because it consist only from matrix transposition and matrix matrix multiplication for which Paralution already have functions usable also on GPU. Therefore no work was needed.

5 Results

The computations were done on the system equipped by Intel Core 2 Duo 2.6Ghz CPU and Nvidia Geforce GTX480 GPU. All simulations were computed in double precision.

First table (Tab 1) compares speed of the first part of the coarsening algorithm, where each connection is evaluated whether it is strong or not. This part was entirely parallel so there can be seen nice speedup.

Second table (Tab 2) shows results for the main part of coarsening algorithm, the clustering. From the result it is obvious that GPU version is quite slower than the CPU one.

	Time	Relative time
CPU edge detection	10.3 s	1
GPU edge detection	2.9 s	0.28

Table 1: Time of finding strongly connected edges

	Time	Relative time
CPU clustering	9.8 s	1
GPU clustering	23.8 s	2.43

Table 2: Time of the groups creating part

Most important and interesting is the final table (Tab 3), which compares total execution time of both versions of coarsening algorithms. Here one can see that GPU is still slower but now only by a small margin. This difference should be however quite negligible in real application where most of the time is not spent in the coarsening part but in the iterative one. In this case it should still be profitable to use GPU version to avoid copying data between memories.

However this could not be tested because, as was stated in the previous chapter, GPU version isn't yet complete, the part for creating interpolation operator is, due to the lack of time, still missing. Therefore in future we would like to finish the GPU version and try to compare the performance on larger spectrum of different matrices.

	Time	Relative time
CPU complete	18.9 s	1
GPU complete	26.4 s	1.4

Table 3: Complete time of coarsening algorithm

6 Summary

This article presented key principles of coarsening algorithm for algebraic multigrid implemented in Paralution software library. This algorithm was then, despite its serial nature, converted to semi parallel version feasible for implementation on GPU and this implementation was thoroughly described. From the obtained results it is obvious that the GPU is not perfectly suited for this task, but the performance is not worse by a large margin. Therefore it should be profitable to use GPU coarsening if this prevents the need to copy data between GPU and CPU memory which was our main goal. Unfortunately this was not yet tested and therefore will be addressed in future research.

References

[1] W. L. Briggs, V. E. Henson and S.F. McCormick. *A Multigrid Tutorial.* Society for6 Industrial and Applied Mathematics, 2000.

[2] D. Göddeke. *Dissertation thesis: Fast and Accurate Finite-Element Multigrid Solvers for PDE Simulations on GPU Clusters.* Technischen Universität Dortmund, 2010

[3] N. Klimanis *Generic Programming and Algebraic Multigrid.* VDM Verlag Dr. Mueller e.K., 2008

[4] Home page of *Paralution* open source library, http://www.paralution.com/

[5] H. Nguyen, *GPU Gems 3*, Adison-Wesley, 2007

[6] Nvidia company, *Nvidia CUDA Programing Guide version 2.2*, Nvidia, 2009

[7] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, 2003

Radon, Carbon Dioxide and Fault Displacements in Central Europe Related to the Tōhoku Earthquake*

Zuzana Knejflová

1st year of PGS, email: knoflice@gmail.com
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Vojtěch Merunka, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Tectonic instability may be measured directly using extensometers installed across active faults or it may be indicated by anomalous natural gas concentrations in the vicinity of active faults. This paper presents the results of fault displacement monitoring at two sites in the Bohemian Massif and Western Carpathians. These data have been supplemented by radon monitoring in the Mladeč Caves and by carbon dioxide monitoring in the Zbrašov Aragonite Caves. A significant period of tectonic instability is indicated by changes in the fault displacement trends and by anomalous radon and carbon dioxide concentrations. This was recorded around the time of the catastrophic $M_W = 9.0$ Tōhoku Earthquake, which hit eastern Japan on 11 March 2011. It is tentatively suggested that the Tōhoku Earthquake in the Pacific Ocean and the unusual geodynamic activity recorded in the Bohemian Massif and Western Carpathians both reflect contemporaneous global tectonic changes.

Full version of this paper: M. Briestenský, L. Thinová, R. Praksová, J. Stemberk, M.D. Rowberry and Z. Knejflová, *Radon, carbon dioxide and fault displacements in central Europe related to the Tōhoku earthquake*, Radiat Prot Dosimetry **160** (2014) 78–82.

Keywords: Fault displacement, Tōhoku earthquake, Radon, Carbon dioxide, Data Analysis

Abstrakt. Tektonické nestability mohou být měřeny přímo s využitím extenzometru v místě aktivního zlomu. Další možností jak zjišťovat tyto nestability je detekce nezvyklé koncentrace přírodních plynů v okolí aktivních zlomů. Tento článek se zabývá výsledky naměřenými na tektonické poruše na dvou místech – České vysočině a v západních Karpatech. Data byla doplněna o sledování radonu v Mladečských jeskyních a oxidu uhličitého v Zbrašovských argonitových jeskyních. Významné období tektonické nestability je zaznamenáno změnou velikosti trhlin v podloží a podle nezvyklých koncentrací výše zmíněných přírodních plynů. Data byla sledována v průběhu katastrofického zemětřesení Tōhoku ve východním Japonsku 11. března 2011, které dosáhlo až 9 stupňů Momentové škály. Předběžně lze říci, že zemětřesení Tōhoku v Tichém Oceánu a neobvyklé geodynamické aktivity, které byly zaznamenány v České vysočině a západních Karpatech, odráží současné globální tektonické změny.

Plná verze příspěvku: M. Briestenský, L. Thinová, R. Praksová, J. Stemberk, M.D. Rowberry and Z. Knejflová, *Radon, carbon dioxide and fault displacements in central Europe related to the Tōhoku earthquake*, Radiat Prot Dosimetry **160** (2014) 78–82.

*This work was supported by the Czech Science Foundation [P210/12/0573] and the Czech Ministry of Education, Youth, and Sports [LM2010008]. It contributes to the research plan of the Institute of Rock Structure and Mechanics [AVOZ30460519].

Klíčová slova: Poruchy podloží, zemětřesení Tohoku, radon, oxid uhličitý, Analýza dat

References

[1] Heinicke, J., Koch, U. and Martinelli, G. *CO2 and radon measurements in the Vogtland Area (Germany)—a contribution to earthquake prediction research.* Geophys. Res. Lett. **22**, 771–774 (1995).

[2] Sugisaki, R., Ido, M., Takeda, H., Isobe, Y., Hayashi, Y., Nakamura, N., Satake, H. and Mizutani, Y. *Origin of hydrogen and carbon dioxide in fault gases and its relation to fault activity.* J. Geol. **91**, 239–258 (1983)

[3] Briestenský, M., Thinová, L., Stemberk, J. and Rowberry, M. D. *The use of caves as observatories for recent geodynamic activity and radon gas concentrations in the Western Carpathians and Bohemian Massif.* Radiat. Prot. Dosim. **145**, 166–172 (2011).

[4] Heinicke, J., Koch, U. and Martinelli, G. *CO2 and radon measurements in the Vogtland Area (Germany)—a contribution to earthquake prediction research.* Geophys. Res. Lett. **22**, 771–774 (1995).

[5] Briestenský, M., Stemberk, J., Michalik, J., Bella, P. and Rowberry, M. D. *The use of a karstic cave system in a study of active tectonics: fault movements recorded at Driny Cave, Malé Karpaty Mts (Slovakia).* J. Cave Karst Stud. **73**, 114–123 (2011).

Geometrical Applications of Mean Curvature Flow: from Analytical Description to Numerical Solution[*]

Miroslav Kolář

2nd year of PGS, email: kolarmir@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Beneš, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This contribution deals with the consrained mean curvature flow for closed planar curves and open planar curves with fixed endpoints. We particularly focus on the area preserving mean curvature flow, which conserves area enclosed by the closed nonselfintersecting curve or area enclosed by the open curve and the lines connecting the fixed endpoints with origin of the coordinates. We deal with such geometrical equation by means of the parametric approach and discuss the effect of tangential redistribution. Resulting system of PDEs is numerically solved and results of particular numerical experiment are presented. We also summarize some results recently published and submitted.

Keywords: mean curvature flow, tangential redistribution, parametric method

Abstrakt. Tento příspěvek se zabývá speciálním případem pohybu křivek (uzavřených nebo otevřených s pevnými konci) v závislosti na střední křivosti, který zachovává jistou geometrickou veličinu. Zaměřujeme se především na případ, kdy se zachovává plocha - v případě uzavřené křivky plocha pod křivkou, v případě otevřené křivky plocha pod křivkou a spojnicemi pevných konců s počátkem souřadnic. Je diskutován parametrický popis problému včetně vlivu tangenciální redistribuce. Výsledný systém PDR je numericky řešen a výsledky numerického experimentu jsou prezentovány. Také shrnujeme nedávné publikované výsledky a výsledky zaslané k recenzi.

Klíčová slova: pohyb křivek řízený střední křivostí, tangenciální redistribuce, parametrizace

1 Introduction

This contribution deals with the applications of the mean curvature flow, i.e., the motion of curves, boundaries or interfaces in dependence on their mean curvature and under effect of external forces. The most general dimensionless form reads as the following geometric evolution equation

$$\text{normal velocity} = \text{mean curvature} + \text{force}. \tag{1}$$

[*] This work has been supported by the grant Two scales discrete-continuum approach to dislocation dynamics, project No. P108/12/1463 and by the grant Multidisciplinary research cntre for advanced materials, project No. 14-36566G of the Grant Agency of the Czech Republic.

In this contribution, our investigation is restricted to the the motion of open and closed planar curves in geometrical context and intended application in discrete dislocation dynamics as in [1].

More precisely, the flow (1) of a planar curve Γ is mathematically described by the following equation

$$\begin{aligned} v_\Gamma &= \kappa_\Gamma + F, & (2)\\ \Gamma|_{t=0} &= \Gamma_{ini}, & (3) \end{aligned}$$

where Γ is either a C^1 smooth closed curve or an C^1 smooth open curve with fixed endpoints in $\mathbb{R}^2$. The quantity v_Γ is the velocity in the direction of the outer normal, κ_Γ is the mean curvature of Γ and F is the force term.

In dependence on the force term F, equation (2) exhibits either global or local character. The global character of the forcing term F often occurs in the so called constrained mean curvature flow, where F depends on global geometrical quantities of the curve Γ, such as its length $L(\Gamma)$ or enclosed area $A(\Gamma)$. Here we compile several known particular constrained motions, and appropriate choices of the corresponding force terms F:

- Area preserving mean curvature flow (see [5, 6]):

  $$F = \frac{1}{L(\Gamma)} \int_\Gamma \kappa_\Gamma \mathrm{d}s.$$

  Choice of this forcing term causes that during time evolution according to (2), the area enclosed by initial curve Γ_{ini} is conserved.

- Length preserving mean curvature flow (see [14]):

  $$F = \frac{\int_\Gamma \kappa_\Gamma^2 \mathrm{d}s}{\int_\Gamma \kappa_\Gamma \mathrm{d}s}.$$

  Time evolution of Γ is constrained in such a way that the lenth of the initial curve Γ_{ini} is preserved.

- Isoperimetric gradient flow

  $$F = \frac{L(\Gamma)}{2A(\Gamma)}.$$

  This motion is constrained in such a way that it minimizes the isoperimetric ratio in relative geometry. More details are discussed by Ševčovič and Yazaki in [14].

Another well-known non-local character of the geometric governing equation (2) concerns the recrystallization phenomena, where a fixed previously melted volume of the liquid phase solidifies – see [8].

The local character of the force F typically occurs in applications of (2) in the field of digital image processing (image segmentation usually; the force here locally depents on the intensity of the processed image – see [9]). Many other particular forms of the force term F are investigated in problems with physical context, especially in the field of discrete dislocation dynamics (see [1, 11]).

2 Parametric Approach

There are several approaches how to treat equation (2). Popular and widely chosen methods come from the family of interface capturing approaches, such as the phase-field method ([10]) or the level set method ([4, 7]). One of their most discussed advantage is their ability to deal with topological change, such as merging of multiple curves into a single one or splitting a single curve into several others. On the other hand, in the case of evolution of planar curves, it is in fact required to solve a 2D problem, and consequently extract the wanted curve (which is one-dimensional object) from a 2D solution. This typically causes computational complexity and slowness of these methods.

A fast and straightforward approach for evolving planar curves is provided by the parametric method (also called direct or Lagrange method [1, 9, 11, 12]). In the direct method, when treating (2), one can describe the family of smooth time-dependent planar curves as

$$\Gamma_t = \{\mathrm{Image}(\vec{X}(t,u)) : u \in [0,1]\}, \quad t \geq 0$$

by means of the parametrization

$$\vec{X} = \vec{X}(t,u),$$

where the spatial parameter u belongs to the fixed interval $[0,1]$. The parametrization $\vec{X}$ is chosen to be oriented conterclockwise. For closed curves, we impose periodic boundary conditions at $u = 0$ and $u = 1$, i.e., $\vec{X}(t,0) = \vec{X}(t,1)$ and $\partial_u\vec{X}(t,0) = \partial_u\vec{X}(t,1)$. For open curves with fixed endpoints, the Dirichlet boundary conditions at $u = 0$ and $u = 1$, i.e., $\vec{X}(t,0) = \vec{X}_0$ and $\vec{X}(t,1) = \vec{X}_1$ are prescribed. Consequently, we can describe the geometric quantities of interest by means of the parametrization $\vec{X}$. The unit tangential vector $\vec{t}$ and the unit normal vector $\vec{n}$ are given by the following formulae:

$$\vec{t} = \frac{\partial_u\vec{X}}{|\partial_u\vec{X}|}, \quad \text{and} \quad \vec{n} = \frac{\partial_u\vec{X}^{\perp}}{|\partial_u\vec{X}|},$$

where $\perp$ is the symbol of perpendicularity. The vector $\vec{n}$ is chosen in such a way that $\det(\vec{n},\vec{t}) = 1$ holds, i.e., we consider the outer unit normal vector. The mean curvature is expressed as

$$\kappa_\Gamma = \frac{1}{|\partial_u\vec{X}|}\partial_u\left(\frac{\partial_u\vec{X}}{\partial_u\vec{X}}\right)\cdot\vec{n}.$$

The normal velocity is defined straightforwardly as

$$v = \partial_t\vec{X}\cdot\vec{n}.$$

Then equation (2) is valid provided the parametrization $\vec{X}$ satisfies the following parametric equation

$$\partial_t\vec{X} = \frac{\partial_{uu}\vec{X}}{|\partial_u\vec{X}|^2} + F\frac{\partial_u\vec{X}^{\perp}}{|\partial_u\vec{X}|}. \tag{4}$$

3 Constrained motion

Let us denote $A = A_t$ the following quantity parametrized by time

$$A = \frac{1}{2}\int_{\Gamma_t} \mathrm{det}(\vec{X}, \vec{t})\mathrm{d}s. \tag{5}$$

Then the following mean curvature flow

$$v_{\Gamma_t} = \kappa_{\Gamma_t} + \frac{1}{L(\Gamma_t)}\int_{\Gamma_t} \kappa_{\Gamma_t}\mathrm{d}s \tag{6}$$

preserves the quantity A, i.e., $A = A_t = A_0$ for all $t \geq 0$. Considering Γ_t a Jordan curve, the quantity A represents the enclosed area, i.e., $A = \int_{\mathrm{int}(\Gamma_t)} \mathrm{d}x$. In the case where Γ_t is an open curve with fixed endpoints, the quantity A_t represents the area, which is enclosed by the curve Γ_t and by the lines connecting the fixed endpoints with the origin of the coordinates.

In [3], we proposed the following proposition
Proposition: *Suppose Γ_t is a family of C^1 smooth curves in the plane for $t \geq 0$, either closed curves or open curves with fixed endpoints evolving according to the mean curvature flow (2). Then*

$$\frac{dA}{dt} = -\int_{\Gamma_t} v_{\Gamma_t} ds.$$

Particularly, if the time evolution is given by constrained mean curvature flow (6), then

$$\frac{dA}{dt} = 0. \tag{7}$$

In general, area-preserving relation (7) is valid for each geometric flow in the form

$$v_{\Gamma_t} = f - \frac{1}{L(\Gamma_t)}\int_{\Gamma_t} f ds.$$

4 The Effect of Tangential Redistribution

It is known when tracking a curve motion, the tangential terms do not affect its shape and hence when analyzing, it is sufficient to take into the account only the terms in the normal direction to the curve. Hovewer, numerical experiments show that the parametric equations (4) are not always apropriate for the numerical computation and instabilities can occur. Since the curve is discretized by a certain number of grid points, in certain cases, we can observe that during the evolution, the grid (discretized) points are accumulated somewhere and, on the other hand, very sparse somewhere else. One possible way to overcome this problem is to complement the equation (4) with a tangential term responsible for redistribution of discretization points

$$\partial_t\vec{X} = \frac{\partial_{uu}\vec{X}}{|\partial_u\vec{X}|^2} + \alpha\frac{\partial_u\vec{X}}{|\partial_u\vec{X}|} + F\frac{\partial_u\vec{X}^\perp}{|\partial_u\vec{X}|}. \tag{8}$$

The term α, is a (possibly nonlocal) function of the position vector $\vec{X}$, its first and second derivatives and the time. Generally, the tangential terms affect the discretization points and move them along the curve without affecting its shape. If correctly chosen, the numerical algorithm is more stable and has higher accuracy. On the other hand, wrong choice of tangential terms can lead to the errors and in the worst case, to the failure of the algorithm.

The problem of tangential redistribution has been extensivelly studied by many authors. We use the curvature adjusted tangential redistribution, which was originally proposed by D. Ševčovič and S. Yazaki in [12] for closed curves. Here which one can also find a brief overview and a critical discussion of redistribution methods. In paper [1], we adapted their original algorithm and developed a modification suitable for open curves with fixed endpoints.

The impact of the tangential redistribution is shown on the Figure 1.

According to the [12], the tangential component has been proposed as the soluton of the following problem

$$\begin{aligned} \partial_s(\varphi(\kappa_{\Gamma_t})\alpha) &= H, \\ H &= f - \frac{\varphi(\kappa_{\Gamma_t})}{\langle\varphi(\kappa_{\Gamma_t})\rangle}\langle f\rangle + \omega\left(\frac{L(\Gamma_t)}{|\partial_u\vec{X}|}\langle\varphi(\kappa_{\Gamma_t})\rangle - \varphi(\kappa_{\Gamma_t})\right), \end{aligned} \tag{9}$$

where ∂_s denotes the derivative with respect to the arc-length, i.e. $\partial_s\vec{X} = \partial_u\vec{X}/|\partial_u\vec{X}|$ and $\mathrm{d}s = |\partial_u\vec{X}|\mathrm{d}u$. The parameter ω is a given positive constant. The other factors in the problem (9) are as follows

$$\begin{aligned} \varphi(\kappa_{\Gamma_t}) &= 1 - \varepsilon + \varepsilon\sqrt{1 - \varepsilon + \varepsilon^2}, \\ f &= \varphi(\kappa_{\Gamma_t})\kappa_{\Gamma_t}(\kappa_{\Gamma_t} + F) - \varphi'(\kappa_{\Gamma_t})(\partial_s^2\kappa_{\Gamma_t} + \partial_s^2 F + \kappa_{\Gamma_t}^2(\kappa_{\Gamma_t} + F)), \\ \langle F(\cdot,t)\rangle &= \frac{1}{L(\Gamma_t)}\int_{\Gamma^t} F(s,t)\mathrm{d}s. \end{aligned}$$

The function $\varphi(\kappa_{\Gamma_t})$ plays an important role because it controls the redistribution of the grid points. The special choice $\varphi(\kappa_{\Gamma_t}) = 1$ produces the uniform redistribution for $\omega = 0$ and asymptotically uniform redistribution for $\varepsilon > 0$. The function $\varphi = |\kappa_\Gamma|$ was proposed for the crystalline curvature flow (see [13]). Choosing $\varepsilon \in (0,1)$, we obtain curvature adjusted redistribution [12].

The redistribution coefficient α is (up to an additive constant) uniquely determined from (9). For closed curves, Ševčovič and Yazaki used renormalization constraint

$$\langle\alpha(\cdot,t)\rangle = 0.$$

For open curves with fixed endpoints (see [1, 3]), we have to ensure

$$\alpha(0,t) = \alpha(L(\Gamma_t),t) = 0$$

for all $t \geq 0$. As $\varphi(\kappa(L(\Gamma_t))) > 0$, setting $\alpha(0,t) = 0$ and integrating (9) over the curve Γ_t with respect to the arc-length yields

$$\varphi(\kappa_{\Gamma_t})\alpha(s,t)|_{s=L(\Gamma_t)} = \varphi(\kappa_{\Gamma_t})\alpha(s,t)|_{s=0} + \int_{\Gamma_t} H(s)\mathrm{d}s,$$

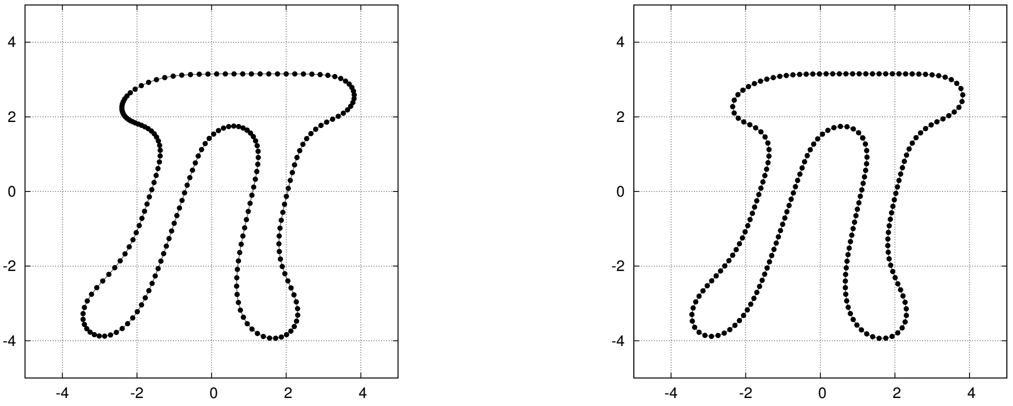


Figure 1: The impact of the tangential redistribution. On the left figure there is a case without the tangential velocity, the curve on the right figure was computed with the usage of the uniform redistribution.

because one can easily see that

$$\int_{\Gamma_t} H \mathrm{d}s = \omega(L(\Gamma_t)\langle\varphi\rangle - L(\Gamma_t)\langle\varphi\rangle) = 0$$

and the uniqueness condition

$$\alpha(0,t) = \alpha(L(\Gamma_t),t) = 0$$

holds.

5 Numerical Solution

For the numerical computations we can use either the scheme based on flowing finite volume method proposed by D. Ševčovič and S. Yazaki [12], which is also discussed in cite. In cite, it is also proposed, the flowing finite volume scheme has order of approximation $\mathcal{O}(h^2)$, where $h = 1/M$ for the number of finite volumes M. Another possibility, which we propose in this paper, is the fully discrete semi-implicit scheme with spatial discretization based on finite differences, such as is cite

$$\vec{X}_j^{k+1} - \tau \frac{\vec{X}_{uu,j}^{k+1}}{\mathcal{Q}^2(\vec{X}_{u,j}^k)} - \tau\alpha_j^{k+1}\frac{\vec{X}_{u,j}^{k+1}}{\mathcal{Q}(\vec{X}_{u,j}^k)} = \vec{X}_j^k + \tau F \frac{\vec{X}_{u,j}^{\perp,k}}{\mathcal{Q}(\vec{X}_{u,j}^k)},$$

where $\vec{X}_j^k \approx \vec{X}(jh, k\tau)$ for the spatial step h and the time step τ, $\mathcal{Q}(\vec{X}) = \sqrt{X_1^2 + X_2^2 + \varepsilon^2}$ serving as the regularization term since it is necessary to avoid dividing by zero. The symbols $\vec{X}_{u,j}^k$ and $\vec{X}_{uu,j}^k$ denote the first and the second central differences.

6 Computational Results

We presented the results of the numerical experiment of the area preserving mean curvature flow (6) for a particular closed curve, where the initial condition has shape of a snowflake. As expected, the solution approaches steady state in circular shape in finite time (see e.g., [2, 3]). This result is in agreement with the hypothesis, which is also supported by our results for closed curves in [2]. In [3], where we focused on area preserving mean curvature flow for open curves with fixed endpoints, one can find another computational study and numerical results of convergence analysis.

In this numerical experiment, the enclosed area of initial condition is ≈ 3.258 and the area enclosed by the steady state solution is ≈ 3.263. Time evolution is depicted in Figure 2.

7 Conclusion

We presented geometrical equation describing area preserving mean curvature flow for closed curves and open curves with fixed endpoints in the plane. We discussed the parametric description of the problem and the enhancement of the parametric equation by employing the tangental redistribution, including its modification for open curves with fixed endpoints. We presented our results of one particular numerical experiment, which is in a good agreement with the theory. We also summarized our published and submitted results.

References

[1] M. Kolář, M. Beneš, D. Ševčovič and J. Kratochvíl. *Mathematical Model and Computational Studies of Discrete Dislocation Dynamics.* IAENG International Journal of Applied Mathematics, to appear (2001).

[2] M. Kolář, M. Beneš and D. Ševčovič. *Computational Studies of Conserved Mean-Curvature Flow.* Mathematica Bohemica, to appear (2014).

[3] M. Kolář, M. Beneš and D. Ševčovič. *Computational Analysis of the Conserved Mean-Curvature Flow for Open and Closed Curves in the Plane.* Computational and Applied Mathematics, submitted (2014).

[4] S. A. Sethian. *Level set method and fast marching methods.* Cambridge University Press (Cambridge, 1999).

[5] M. Gage. *On area-preserving evolution equation for plane curves.* Contempt Math 51 (1986), 51-62.

[6] I. C. Dolcetta, S. F. Vita and R. March. *Area preserving curve shortening flows: from phase separation to image processing.* Interfaces and Free Boundaries 4 (2003), 325-353.

[7] S. Osher and R. P. Fedkiw. *Level set method and dynamic implicit surfaces.* Springer (New York, 2003).

[8] I. V. Markov. *Cystal Growth for Beginners: Fundamentals of Nucleation, Crystal Growth, and Epitaxy, 2nd edn.* World Scientific Publishing Company (2004).

[9] M. Beneš, M. Kimura, P. Pauš, T. Tsujikawa and S. Yazaki. *Application of a curvature adjusted method in image segmentation.* Bulletin of the Institute of Mathematics, Academia Sinica (New Series) 3 (2008), 509-523.

[10] T. Ohta, M. Mimura and R. Kobayashi. *Higher-dimensional localized patterns in excitable media.* Physica D: Nonlinear Phenomena, 34 (1989), 115-144.

[11] M. Beneš, J. Kratochvíl, J. Křišťan, V. Minárik and P. Pauš. *A parametric simulation method for discrete dislocation dynamics.* The European Physical Journal ST, 177 (2009), 177-192.

[12] D. Ševčovič and S. Yazaki. *Evolution of plane curves with a curvature adjusted tangential velocity.* Journal of Industrial and Applied Mathematics, Vol. 28, Issue 3 (Japan 2011), 413-442.

[13] S. Yazaki. *On the tangential velocity arising in a crystalline approximation of evolving plane curves.* Kybernetika, Vol. 43, No. 6 (2007), 913-918.

[14] D. Ševčovič and S. Yazaki. *On a gradient flow of plane curves minimizing the anisoperimetric ratio.* IAENG International Journal of Applied Mathematics Vol. 43, Issue 3 (2013), 160-171.

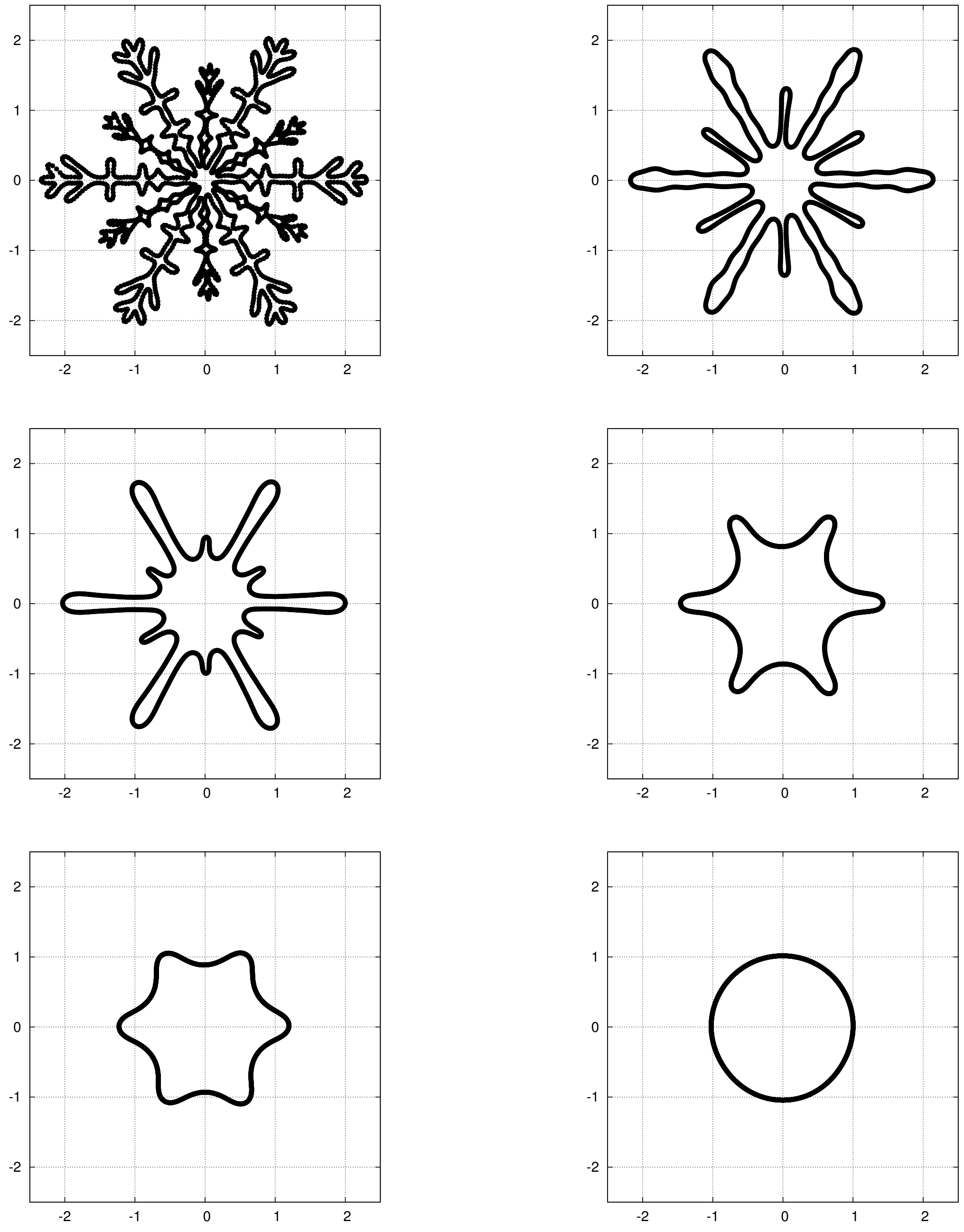


Figure 2: Time evolution of closed curve with initial shape of snowflake. The curve evolves according to area preserving mean curvature flow (6) and approaches steady state of circular shape.

Multifractal Diffusion Entropy Analysis: Optimal Bin Width of Probability Histograms

Jan Korbel

2nd year of PGS, email: korbeja2@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Jizba, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In the framework of Multifractal diffusion entropy analysis we propose a method for choosing an optimal bin-width in histograms generated from underlying probability distributions of interest. The method presented uses techniques of Rényi's entropy and the mean squared error analysis to discuss the conditions under which the error in the multifractal spectrum estimation is minimal. We illustrate the utility of our approach by focusing on a scaling behavior of financial time series. In particular, we analyze the S&P 500 stock index as sampled at a daily rate in the time period 1950–2013. In order to demonstrate a strength of the method proposed we compare the multifractal δ-spectrum for various bin-widths and show the robustness of the method, especially for large values of q. For such values, other methods in use, e.g., those based on moment estimation, tend to fail for heavy-tailed data or data with long correlations. Connection between the δ-spectrum and Rényi's q parameter is also discussed and elucidated on a simple example of multiscale time series.

This article has been published in *Physica A*, 413:438–458, 2014, and the results have been presented and are part of conference proceedigs of ISCS 2014 held in Florence, Italy.

Keywords: Multifractals, Rényi entropy, Stable distributions, Time series

Abstrakt. V rámci multifraktální diffusion entropy analysis (MFDEA) je odvozena nová metoda pro výběr optimální šířky sloupce histogramu generovaného z modelu řízeného pravděpodobnostním rozdělením daného modelu. Tato metoda užívá technik Rényiho entropie a střední kvadratické chyby k diskuzi, za kterých podmínek je odhad multifraktálního spektra optimální. Užitečnost tohoto přístupu je ilustrována na škálování finančních časových řad, konkrétně na časové řadě denních výnosů indexu S&P 500 v období 1950-2013. Za tímto účelem porovnáváme multifraktální δ spektra pro různé šířky sloupců a ilustrujeme robustnost této metody, hlavně pro velké hodnoty parametru q. Pro tyto hodnoty ostatní používané metody, například ty, které jsou založeny na odhadu momentů daného rozdělení, mají tendenci selhat pro data s těžkými rameny nebo data s dlouhými korelacemi. Spojitost mezi δ spektrem a Rényiho parametrem q je také diskutována na jednoduchém příkladu multiškálové časové řady.

Tento příspěvek byl publikován v *Physica A*, 413:438–458, 2014 a byl přenesen (a je součástí sborníku) na konferenci ISCS 2014 konané ve Florenci v Itálii.

Klíčová slova: multifraktály, Rényiho entropie, stabilní rozdělení, časové řady

Parabolic Strip Telescope[*]

Vladislav Kosejk

1st year of PGS, email: kosejvla@fjfi.cvut.cz.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Goce Chadzitaskos, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This work is focused on a proposal of a new type of telescopes using a rotating parabolic strip as the primary mirror. It is a principal modification of the design of telescopes from the times of Galileo and Newton. In order to demonstrate the basic idea, the image of an artificial constellation observed by this kind of telescope was reconstructed using the techniques described in this article. We present a working model of this new telescope, we have used an assembly of the primary mirror — a strip of acrylic glass parabolic mirror 30 cm long and 10 cm wide shaped as a parabolic cylinder of focal length 1 m — and an artificial constellation, a set of LED diodes in a distance of 15 m. In order to reconstruct the image, we made a series of snaps, each after a rotation of the constellation by 5 degrees. Using three different algorithms we reconstructed the image of this artificial constellation. This contribution is based on (Chadzitaskos 2013) with new telescope designs and new experimental tests.

Full text: V. Kosejk, G. Chadzitaskos, and J. Červený, *Parabolic Strip Telescope*, In 'Proceedings of PIERS 2014 in Guangzhou' (2014), 471–476. Available on: http://piers.org/piersproceedings/piers2014GuangzhouProc.php?start=100.

Keywords: Telescope, angular resolution, image processing,

Abstrakt. Tento článek se zaměřuje na návrh nového typu teleskopu, který využívá rotační parabolický pásek jako primární optický element. Takové řešení je hlavní modifikací v návrhu dalekohledů od dob Galiea a Newtona. K demonstraci principu základní myšlenky je využit obraz umělého souhvězdí, pozorovaného rotačním teleskopem, s využitím rekonstrukčních technik popsaných v tomto článku. Fungující model nového teleskopu pracuje s akrylovým parabolickým páskem o rozměrech 30x10 cm, který je tvarovaný jako parabolický válec s ohniskovou vzdáleností 1 metr. Obraz umělého souhvězdí je reprezentován souborem LED diod ve vzdálenosti 15 metrů. Rekonstrukci obrazu provádíme ze série snímků, kde při každém snímání otočíme desku umělého souhvězdí o 5 st. Pro rekonstrukci obrazu využíváme 3 rozdílné algoritmy zpracování nasnímaných obrazů. Tento příspěvek je založen na (Chadzitaskos 2013) s novým modelem teleskopu a novou sérií testů.

Celý článek: V. Kosejk, G. Chadzitaskos, and J. Červený, *Parabolic Strip Telescope*, In 'Proceedings of PIERS 2014 in Guangzhou' (2014), 471–476. Available on: http://piers.org/piersproceedings/piers2014GuangzhouProc.php?start=100.

Klíčová slova: Teleskop, úhlové rozlišení, zpracování obrazu.

[*] Obrana a bezpečnost CZ.1.05/3.1.00/14.0304...

References

[1] Chadzitaskos G. 2013, Parabolic strip telescope, arXiv: astro-ph/1304.6530

[2] King H. C. 2003, The History of the Telescope, Dover Publication, New York

[3] Crawford F.S. Jr. 1968, Berkeley Physics Course 3. Waves (McGraw-Hill Book Co., New York).

[4] ESO 2012, The Very Large Telescope, `https://www.eso.org/public/teles-instr/vlt.html`

[5] Herman, G. T. 2009, Fundamentals of computerized tomography: Image reconstruction from projection, 2nd ed., Springer,

[6] English, R. J.1996, Single–Photon Emission Computed Tomography: A Primer, Publ. of The Society of Nuclear Medicine,

Markov Chain Testing with Application in Tennis Match Outcomes

Tomáš Kouřim

1st year of PGS, email: kourim@outlook.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Volf, Department of Stochastic Informatics
Institute of Information Theory and Automation, AS CR

Abstract. The demand for proper sport match prediction tools is constantly increasing together with the amount of money put into sports betting. Possible ways of modeling tennis matches and their in-play states using discrete Markov chains are introduced in this paper. The results are based on the 2007-2013 ATP seasons.

Keywords: discrete Markov chain, tennis, in-play modeling

Abstrakt. Celosvětově vzrůstající množství prostředků vložených do sportovních sázek stupňuje poptávku po kvalitních nástrojích k predikci sportovních výsledků. V tomto článku jsou představeny některé možnosti využití diskrétních Markovských řetězců pro modelování situací v průběhu tenisových utkání. Závěry jsou postaveny na výsledcích světové tenisové série ATP z let 2007 až 2013.

Klíčová slova: diskrétní Markovovy řetězce, tenis, modelování herních situací

1 Introduction

The popularity of sports betting and especially the online sports betting has been increasing over the course of past several years. In order to satisfy the demands of bettors, the bookmakers are continuously expanding the betting offer. Therefore it becomes increasingly important to correctly predict not only the match outcomes (i.e. the win of a certain player or a team), but also the different particular results (such as the number of sets played, goals scored), especially for the popular sports such as tennis. Mathematical modeling could be a proper tool to produce such predictions.

To predict the outcome of a single match, most approaches consider time development of individual player's strength from match to match and adjust the pre-game parameters according to the previous results [7]. The obvious drawback of this approach is that the subject of study are real people and their *behavior* or *properties* over the course of time. Without doubt the performance during previous matches and tournaments is a good indicator of the future performance, but there are still many other factors to consider. Some of them could be observed, such as the preference of certain surface or a head to head results with a given opponent, but others, such as small illness between tournaments or irregular support of fans, are extremely hard to even identify, let alone to quantify (especially back in time). Altogether, it is obvious that the player's performance varies not only tournament by tournament and match by match, but probably even point by point. Different approach than considering the development from match to match, is examining the development over the course of one match. One match of tennis can be viewed as a realization of a discrete stochastic process, considering sets, games, points or even single strokes as units of time [5]. In this paper we focus on a set by set tennis match model.

1.1 Input data

Tracking all the individual variables that might influence the outcome of a tennis match (and the probabilities of an outcome) several years backwards is a difficult task. Some parts of it, such as the past results, past player's rankings or even point by point match development can be done with the help of computers. On the other hand, there are some variables that certainly influence a tennis game which are very difficult to track in real time and virtually impossible to track backwards. Those are especially some life events of individual players, such as a small illness or injury, change of a personal physiotherapist or even a baggage lost during a flight. Such variables can (but not necessarily have to) influence match outcomes and without their knowledge, computing probabilities from past results can introduce some kind of bias.

However, professional bookmakers keep (and have always kept) track of all those individual variables and already incorporated those into their match odds. In this paper, it is assumed that all matches with similar starting odds should have similar development, no matter what tournament, surface or players are involved. Therefore the starting odds given by bookmakers are taken as a starting point for match modeling That is, two matches between Roger Federer and Rafael Nadal are not considered to be two observations of the same process (or variable), but rather are the two matches where the odds are same (or similar).

1.2 Odds

The bookmakers' odds are just another form of expressing probabilities. There are three way to represent odds, European, American and fractal, for more detail see for example [4]. The most common is the European (decimal) format. Let o_A be the odds for player A to win a match. When a bet b_A is placed, the payout if successful is $pay_A = o_A \cdot b_A$. Thus the probability associated with the odds o_A is $\tilde{p}_A = \frac{1}{o_A}$. However, in order to generate profit, the bookmakers use some margin, causing that $\frac{1}{o_A} + \frac{1}{o_B} = \tilde{p}_A + \tilde{p}_B > 1$. The margin can be as low as 1% (for the most prestigious games such as Grand Slam finals), but also $> 10\%$ (for some low rank tournaments). That means that $\tilde{p}_A$ and $\tilde{p}_B$ are not actual winning probabilities of the players, but margin-adjusted probabilities. In order to obtain the actual probabilities, $\tilde{p}_A$ and $\tilde{p_B}$ have to be normalized.

Standard normalization distributes the margin evenly between the favorite and the outsider. Empirical results, however, suggest, that such distribution is incorrect and that the bookmakers' margin lies rather on the side of the outsider and that the odd-probability of the favorite is very close of the actual winning probability. Therefore, odd adjusted normalization has to be introduced in order to obtain correct probabilities.

2 Markov Chains

Markov chain is a stochastic process with discrete set of states and discrete time that satisfies the Markov property. That is the probability that at time t the chain is at state i, $p_i^{(t)}$, only depends on the previous state, i.e. the state at time $t-1$ [2]. Markov chain is finite (or infinite) if the corresponding state sets are finite and infinite, respectively. Let $p_{ij}^{(t)}$ denote the conditional probability that the chain will be at state j at the next step, given it is at state i in the current step. The probabilities $p_{ij}^{(t)}$ are called *transition probabilities.* The square matrix P such that

$$P_{i,j}^{(t)} = p_{ij}^{(t)}$$

is called the *transition matrix.* A Markov chain is said to be *homogeneous* if the probability $p_{i,j}^{(t)}$ is time independent for all i, j. A state i of a Markov chain is called *absorbing* if it is impossible to leave it (i.e., $p_{ii} = 1$). A Markov chain is *absorbing* if it has at least one absorbing state. In an absorbing Markov chain, a state which is not absorbing is called *transient* [3]. The states of a Markov chain can be renumbered such that the absorbing states come first and the transient last. Then the transition matrix will have the *canonical form*

$$P = \left(\begin{array}{c|c} I & 0 \\ \hline R & Q \end{array}\right).$$

The *fundamental matrix* is the matrix

$$N = (I - Q)^{-1}.$$

The elements of the fundamental matrix have this meaning. n_{ij} is the expected number of times the Markov chain will be in state j if it started in state i. Let b_{ij} be the probability of the chain to be absorbed in state j given it started in i and let

$$B_{i,j} = b_{ij}$$

be a matrix. Then

$$B = NR,$$

where N is the fundamental matrix and R is the sub-matrix from the canonical form. Matrices N and B allow to compute all necessary information about the Markov chain and the stochastic process that it represents. More information about Markov chains together with proofs of the statements above can be found in [3].

3 Data Description

The application of Markov chains on tennis match simulation is studied on the set of tennis match results from the 2007 thru 2013 ATP[1] seasons, available freely from *http://tennis-data.co.uk/alldata.php*[2]. The data contains basic information about the tournament, the

[1] Association of Tennis Professionals, men tennis association.

[2] The data contains many errors, which were removed manually. Thus, some matches from the mentioned seasons are not included in the dataset.

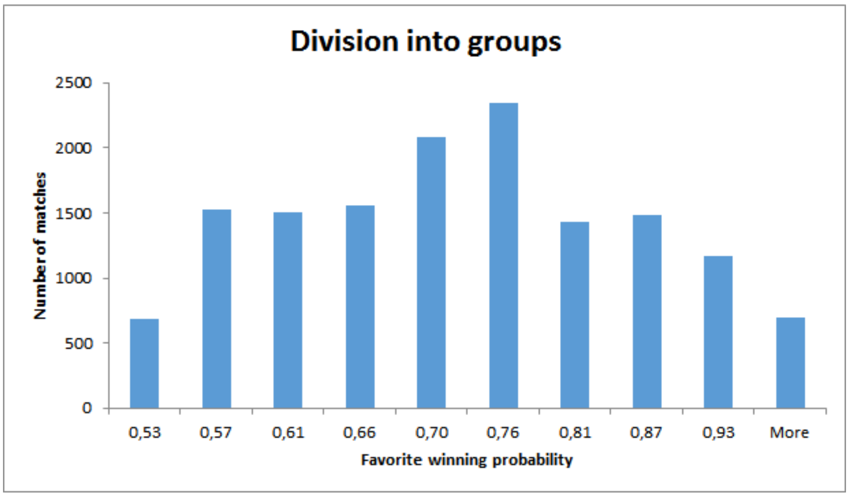


Figure 1: Histogram of the data distribution.

players and set by set results. It also contains winning odds for each player provided by 5 of the top international bookmakers - Bet365, Expect Win, Ladbrokes, Pinnacle Sports and Stan James (Unibet for earlier seasons). Only matches, where there was necessary to win two sets in order to win a match, were considered in our study, therefore the four Grand Slam tournaments were omitted for each season. In this paper, the two tennis players are always labeled as the favorite and the outsider. This is the best way to name the players, as unlike in most other sports, the concept of home and away teams (players) is not used in tennis[3]. In order to avoid confusion, matches without a favorite, i.e. matches where the odds were even, were also not included in our study (242 matches). Altogether, the database contains 14 240 different matches.

The matches were divided into groups according to the winning probability distribution among the two players. In order to obtain groups with enough observations, small intervals were taken instead of individual values. This goes in tact with the fact that the odds given by bookmakers are mere the probability estimates, not the actual probabilities. The data were not uniformly distributed and therefore logarithmic scale was used to create groups with similar number of matches. The histogram showing the division of the matches into respective groups is in Figure 1.

4 Results

4.1 I.i.d. Hypothesis

Given the probabilities of wining a match, the hypothesis can be assumed that the probabilities of winning respective sets are *independent identically distributed* random variables. For the matches played as best of three, there are three possible ways for the favorite to win a match, that is win 2:0, win 2:1 and lose the first set and win 2:1 and lose the second set. The theoretical value of the probability of winning in a set can be obtained by solving the equation

$$p_{match} = p_{set}^2 + 2 \cdot p_{set}^2 \, (1 - p_{set}).$$

[3] Of course, there are some British players playing the Wimbledon and French players playing the French Open etc., but in general, the home away concept is not present in tennis.

This cubic equation can be numerically solved for example using the Newton-Raphson algorithm [6]. Under the i.i.d. hypothesis, a tennis match can be modeled as a Markov chain with six different states. The match starts, is tied, the favorite is leading by one set, the favorite is down by one set, favorite wins and favorite loses. Renumbering the states accordingly, the following transition matrix in the canonical form is obtained, with p standing for the probability of the favorite to win a set under the i.i.d. hypothesis.

$$P = \left(\begin{array}{c|cccccc} & F\,win & F\,lose & F\,lead & F\,back & 1:1 & 0:0 \\ \hline F\,win & 1 & 0 & 0 & 0 & 0 & 0 \\ F\,lose & 0 & 1 & 0 & 0 & 0 & 0 \\ F\,lead & p & 0 & 0 & 0 & 1-p & 0 \\ F\,back & 0 & 1-p & 0 & 0 & p & 0 \\ 1:1 & p & 1-p & 0 & 0 & 0 & 0 \\ 0:0 & 0 & 0 & p & 1-p & 0 & 0 \end{array}\right)$$

It is an absorbing Markov chain which has two absorbing and four transient states. All match outcomes for all in-play situations can be obtained from this transition matrix, see Section 2 for details.

4.2 Empirical Results

The circumstances of a tennis match, as an encounter between two individuals, suggest that the i.i.d. hypothesis might not correspond with the reality and that a different transition matrix has to be constructed. In order to confirm or reject the hypothesis, empirical results were compared with the theoretical values. Tables 1 thru 5 show the results. First two columns show the division into groups, third column contains the set winning probability obtained assuming the i.i.d. hypothesis and the other columns show the values obtained from the data together with the p-values associated with the reality. Namely, if we denote p as the probability derived from the i.i.d hypothesis and $\hat{p}$ relative frequencies computed from data, the test statistics is $Z = \sqrt{\frac{n}{p\cdot(1-p)}} \cdot (\hat{p} - p)$ and, under stated hypothesis, its distribution is taken to be standard normal (due to the central limit theorem) [1].

The results for the first set are almost completely in tact with the i.i.d. hypothesis, see Table 1. This is obvious, as the set number one is the first random variable and there is nothing to be dependent on. The only group where the i.i.d. hypothesis can be rejected is that of the huge favorites with the favorite winning probability over 93 %. This can be caused by the fact that for such a big favorite, the bookmakers' are not that accurate. Tables 2 and 3 on the other hand show that the second set results do not correspond with the i.i.d. hypothesis at all and suggest that winning the first set increases the chances of winning the second set as well. This is in tact with the opinion about tennis and sports in general, which can be expressed as “success breeds success”. Graphical illustration of the situation after the first set can be observed in Figures 2 and 3.

The most interesting part is the third set. The question is whether there is a difference between the beginning of the match (i.e. the state 0:0) and the state 1:1, and between the two ways of getting into the state 1:1. Tables 4, 5 and 6 do not give a definite answer for the question. Not all the p-values speak against the i.i.d. hypothesis and against the

p_match lower bound	p_match upper bound	p_set i.i.d.	first set winning ratio	total matches	p_value
0,50	0,53	0,51	0,51	442	0,491
0,53	0,57	0,54	0,53	1526	0,398
0,57	0,61	0,56	0,57	1501	0,264
0,61	0,66	0,59	0,57	1561	0,123
0,66	0,70	0,61	0,60	2078	0,087
0,70	0,76	0,66	0,68	2348	0,032
0,76	0,81	0,70	0,70	1426	0,361
0,81	0,87	0,75	0,75	1487	0,429
0,87	0,93	0,80	0,79	1173	0,178
0,93	1	0,88	0,91	698	0,006
0,5	1	0,65	0,65	14240	0,476

Table 1: The i.i.d. hypothesis compared to the actual results of first sets.

p_match lower bound	p_match upper bound	p_set i.i.d.	second set winning ratio	total matches	p_value
0,50	0,53	0,51	0,67	226	0,000
0,53	0,57	0,54	0,63	813	0,000
0,57	0,61	0,56	0,65	857	0,000
0,61	0,66	0,59	0,71	892	0,000
0,66	0,70	0,61	0,70	1257	0,000
0,70	0,76	0,66	0,74	1589	0,000
0,76	0,81	0,70	0,78	998	0,000
0,81	0,87	0,75	0,80	1117	0,000
0,87	0,93	0,80	0,87	930	0,000
0,93	1	0,88	0,89	634	0,168
0,5	1	0,65	0,75	9313	0,000

Table 2: The i.i.d. hypothesis compared to the actual results of second sets after the favorite has won the first set.

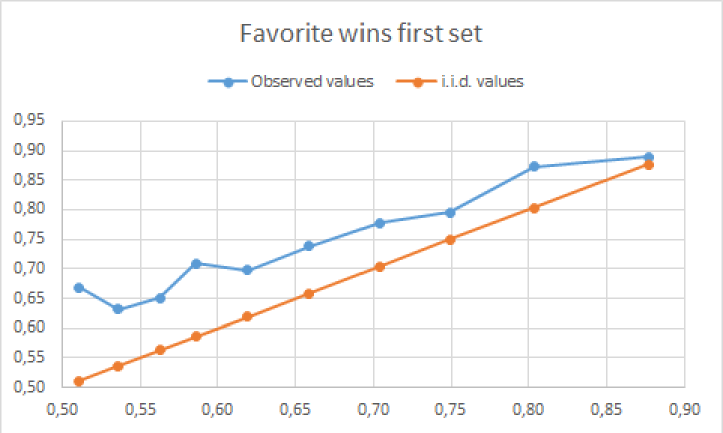


Figure 2: Favorite's winning probabilities after winning first set compared to those computed under i.i.d.

p_match lower bound	p_match upper bound	p_set i.i.d.	second set winning ratio	total matches	p_value
0,50	0,53	0,51	0,36	216	0,000
0,53	0,57	0,54	0,41	713	0,000
0,57	0,61	0,56	0,46	644	0,000
0,61	0,66	0,59	0,48	669	0,000
0,66	0,70	0,61	0,50	821	0,000
0,70	0,76	0,66	0,55	759	0,000
0,76	0,81	0,70	0,60	428	0,000
0,81	0,87	0,75	0,59	370	0,000
0,87	0,93	0,80	0,67	243	0,000
0,93	1	0,88	0,75	64	0,001
0,5	1	0,65	0,51	4927	0,000

Table 3: The i.i.d. hypothesis compared to the actual results of second sets after the favorite has lost the first set.

p_match lower bound	p_match upper bound	p_set i.i.d.	third set winning ratio	total matches	p_value
0,50	0,53	0,51	0,65	78	0,006
0,53	0,57	0,54	0,58	291	0,079
0,57	0,61	0,56	0,59	294	0,189
0,61	0,66	0,59	0,56	320	0,168
0,66	0,70	0,61	0,63	412	0,313
0,70	0,76	0,66	0,66	414	0,486
0,76	0,81	0,70	0,73	255	0,154
0,81	0,87	0,75	0,76	217	0,411
0,87	0,93	0,80	0,79	163	0,348
0,93	1	0,88	0,83	48	0,178
0,5	1	0,65	0,65	2492	0,414

Table 4: The i.i.d. hypothesis compared to the actual results of third sets after the favorite has lost the first set and won the second set.

p_match lower bound	p_match upper bound	p_set i.i.d.	third set winning ratio	total matches	p_value
0,50	0,53	0,51	0,47	75	0,222
0,53	0,57	0,54	0,41	300	0,000
0,57	0,61	0,56	0,50	300	0,010
0,61	0,66	0,59	0,55	260	0,097
0,66	0,70	0,61	0,59	380	0,137
0,70	0,76	0,66	0,65	416	0,379
0,76	0,81	0,70	0,73	222	0,246
0,81	0,87	0,75	0,75	228	0,488
0,87	0,93	0,80	0,76	118	0,132
0,93	1	0,88	0,81	70	0,055
0,5	1	0,65	0,60	2369	0,000

Table 5: The i.i.d. hypothesis compared to the actual results of third sets after the favorite has won the first set and lost the second set.

p_match lower bound	p_match upper bound	favorite loses, then wins	total matches	favorite wins, then loses	total matches	p_value
0,50	0,53	0,65	78	0,47	75	0,009
0,53	0,57	0,58	291	0,41	300	0,000
0,57	0,61	0,59	294	0,50	300	0,012
0,61	0,66	0,56	320	0,55	260	0,375
0,66	0,70	0,63	412	0,59	380	0,130
0,70	0,76	0,66	414	0,65	416	0,404
0,76	0,81	0,73	255	0,73	222	0,421
0,81	0,87	0,76	217	0,75	228	0,444
0,87	0,93	0,79	163	0,76	118	0,285
0,93	1	0,83	48	0,81	70	0,394
0,5	1	0,65	2492	0,60	2369	0,000

Table 6: Comparison of the two ways of getting into the 1:1 state of a tennis match.

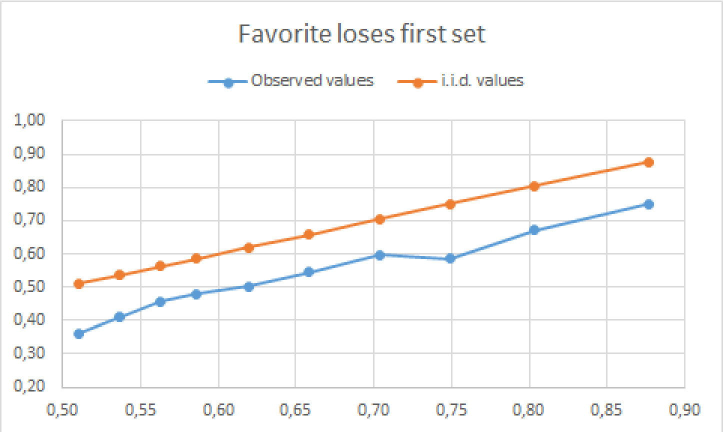


Figure 3: Favorite's winning probabilities after losing first set compared to those computed under i.i.d.

hypothesis that the two ways of getting into 1:1 are equivalent. The mean values of set winning probabilities suggest that the two hypotheses might be incorrect but there is not enough observations to prove it. If we look at all matches as a one group, we can tell that it does matter how is the 1:1 state achieved and that if the outsider ties a match after losing the first set, his chances of winning set number there are better than those computed under the i.i.d. hypothesis.

These results indicate that in order to model a tennis match using a Markov chain we have to improve it according to the observed data by introducing new states. The adjusted transition matrix in the canonical form is then

$$P = \left(\begin{array}{c|ccccccc} & F\,win & F\,lose & 0:0 & F & O & FO & OF \\ \hline F\,win & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ F\,lose & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0:0 & 0 & 0 & 0 & p_{iid} & 1-p_{iid} & 0 & 0 \\ F & p_F & 0 & 0 & 0 & 0 & 1-p_F & 0 \\ O & 0 & 1-p_O & 0 & 0 & 0 & 0 & p_O \\ FO & p_{FO} & 1-p_{FO} & 0 & 0 & 0 & 0 & 0 \\ OF & p_{OF} & 1-p_{OF} & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

where F and O stand for a set won by the favorite and outsider, respectively, and $p_{situation}$ stands for the set winning probability of the favorite under the respective situation. This matrix again has two absorbing states but five transient states. Results from Table 4 suggest that $p_{OF} = p_{iid}$. Again, this matrix can be used to compute all the in-play odds for different match situations. Another possible adjustment to the transition matrix would be to introduce another two absorbing states to differentiate between 2:0 and 2:1 win (or loss).

5 Conclusion

The possibility of modeling a tennis match set by set using Markov chains was studied in this paper. The simple independence hypothesis was rejected by observing actual tennis match results obtained from the ATP series from years 2007 thru 2013. It was proven that the set winning probabilities of the players change throughout the match depending on the outcomes of the past sets. This is in tact with the general belief about sports and tennis in particular. The Markov chain that models the real tennis match set by set development was introduced.

6 Future Work

The presented results constitute to a good starting point for the further study of the implementation of Markov chains in sports and in tennis in particular. The first area of interest is to study whether the presented results apply only for the studied case or can be generalized. Therefore the i.i.d. hypothesis should be tested on different data sets regarding tennis (such as women tennis matches, doubles etc.) and other sports played in sets, both individual (such as table tennis or badminton) and team (such as volleyball).

The other way of research is to study the possible implementation of Markov chains in the modeling of tennis matches in more detail. That is, to study the match not only set by set, but also game by game, point by point or even stroke by stroke.

References

[1] Jiří Anděl. Základy matematické statistiky. 2., praha: Matfyzpress. Technical report, ISBN 80-86732-40-1, 2005.

[2] Jiří Demel. Operační výzkum [online]. Czech Technical University in Prague, 2011. [Accesed: 30-9-2014]. Available at: `http://kix.fsv.cvut.cz/~demel/ped/ov/ov.pdf`.

[3] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 1998.

[4] M. Hejdůšek. Matematické metody pro modelování a predikci výsledků sportovních utkání. *Diplomová práce, ČVUT FJFI*, 2007.

[5] Gordon J.A. Hunter and Krzysztof Zienowicz. Can markov models accurately simulate lawn tennis rallies? In *Proceedings of the Second International Conference on Mathematics in Sport*, volume 1, pages 69–75. Univ. of Groningen, NL, 2009.

[6] Carl T Kelley. *Solving nonlinear equations with Newton's method*, volume 1. Siam, 2003.

[7] Paul K Newton and Kamran Aslam. Monte carlo tennis: A stochastic markov chain model. *Journal of Quantitative Analysis in Sports*, 5(3), 2009.

A Note on Spatial Organisation in Systems with Advection[*]

Michal Kozák

1st year of PGS, email: michal.kozak@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Václav Klika, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Pattern formation due to chemical instability is one of the most important phenomenon in many non-equilibrium systems, ranging from developmental biology to gas-discharge systems, crystal growth in solidifying alloys, plasma or semiconductors. The recognised fundamental symmetry breaking mechanism is a diffusion-driven instability (Turing instability [2]) in a reaction diffusion system (RD system). Turing showed that small perturbation of well-mixed homogeneous system of autocatalic and inhibitory diffusing species could cause instability which leads to an emergence of spatial patterns. Further, motivated by Belousov-Zhabotinsky reaction, Rovinsky and Menzinger proposed other possible mechanism, differential-flow-induced chemical instability (DIFICI) [3] containing a term with advection. As a result, they obtained a range of new spatial patterns. Other mechanism followed, e.g. flow-distributed oscilations (FDO) [4], flow-and-diffusion structures (FDS) [5]. These approaches differ in chosen parameter, importance of each equation element or physical motivation. We will be interested in analysis of such models on the bounded domain where reaction, diffusion and advection are considered (RDA system) and where the spatial pattern formation occurs.

If we compare Turing instability in RD systems and the instabilities in RDA systems described above, the principle of modelling some real situation seems to be very similar, thus we could expect similar results, and yet from the mathematical point of view they are very different. The operator corresponding to RD system is self-adjoint, the operator corresponding to RDA system is not, thus behaviour of RDA operator is not characterized by sum of its eigenfunctions and hence it is more complex (a very helpful theory is the theory about pseudospectrum [8, 9]). Additionally, the presence of advection leads to a difference in concept of "system without diffusion", between setting $D = 0$ and letting limit $\lim_{D\to 0}$. In this work, we consider the latter, we suppose various boundary conditions (Dirichlet's, zero-flux, no outflow due to diffusion, periodic and Danckwert's) and analyse two RDA systems of two equations in one-dimensional spatial variable that are both well outside of the classical diffusion-driven instability regime; in the first case, one species is attached to a fixed substrate (one equation contains neither diffusion nor advection term), in the second case both equations have the same coefficient of diffusion and the same coefficient of advection.

We use concept of linearised stability, thus we are interested in a sign of real parts of roots of the so-called dispersion relation. We employ that we are able to compute eigenvalues of appropriate operators analytically with each type of boundary condition and by standard methods of model analysis we derive conditions when the diffusion driven instability occurs.

[*]This work has been supported by the grant SGS12/198/OHK4/3T/14.

The obtained results differ from naive intuition but also qualitatively differ from those presented in literature where the results were obtained by different methods. In the first studied case, with one attached species, the homogeneous steady state for dominant advection is always stable, which is in contrast with studies of unbounded systems [6, 7]. In the second case, both species moving, significant relaxation of conditions comparing to Turing instability occurs - arbitrary small advection causes that stability of homogeneous system always holds, therefore there are no binding conditions in analysis of instability of the system.

The contribution of this article lies in demonstration and highlighting a crucial difference between conditions of emergence of spatial pattern in RD and RDA systems but also that other (and more complex) methods of stability analysis have to be used for analysis of RDA operator. As the influence of arbitrary small advection on system's behaviour is striking, a natural question arises: is advection present in real applications or can it be safely assumed that advection is not considered? To recognize whenever RD od RDA model fits better to real situation, we shall also look closely to the physical essence of modelled phenomenon in our future work.

Keywords: reaction-diffusion-advection system, diffusion driven instability

Abstrakt. Vznik prostorových struktur je jedním z nejdůležitějších jevů v mnoha nerovnovážných systémech, počínaje vývojovou biologií přes růst krystalů v tuhnoucích slitinách, konče plasmou nebo polovodiči. Základním mechanismem k narušení symetrie je nestabilita způsobená difuzí (diffusion driven instability; Turingova nestabilita [2]) reakčně-difuzních systémů (RD systém). Turing ukázal, že malá perturbace homogenního systému autokatalyticky a inhibičně difundujících druhů mohou způsobit nestabilitu, která vede ke vzniku prostorových struktur. Dále, motivováni Belousov-Zhabotinského reakcí, Rovinsky a Menzinger navrhli jiný možný mechanismus, differential-flow-induced chemical instability (DIFICI) [3] obsahující člen s advekcí. Jako výsledek dostali řadu nových prostorových struktur. Další mechanismy následovaly, například flow-distributed oscilations (FDO) [4], flow-and-diffusion structures (FDS) [5]. Tyto přístupy se liší volbou parametrů, důležitostí jednotlivých členů v rovnici nebo fyzikálními motivacemi. Nás bude zajímat analýza takových modelů uvažovaných na omezených oblastech, kde je přítomna reakce, difuze a advekce (RDA systém) a kde dochází ke vzniku prostorových struktur.

Porovnáme-li Turingovu nestabilitu v RD systémech s nestabilitami v RDA systémech popsanými výše, podstata modelování konkrétní reálné situace vypadá velmi podobně, tedy bychom předpokládali i podobné výsledky, leč z matematického hlediska jsou velmi odlišné. Operátor příslušný RD sysému je samoadjungovaný, operátor příslušný RDA systému ne, tudíž chování RDA operátoru nejde charakterizovat součtem svých vlastních funkcí, jde tedy o komplexnější chování (nápomocnou teorií je teorie pseudospekter [8, 9]). Dále, přítomnost advekce vede k rozdílům v konceptu "systému bez difuze", mezi dosazením $D = 0$ a limitním přechodem $\lim_{D\to 0}$. V této práci uvažujeme posledně jmenovaný, předpokládáme řadu okrajových podmínek (Dirichletovy, nulový tok, nulový tok vzhledem k difuzi, periodické a Danckwertovy) a analyzujeme dva RDA systémy o dvou rovnicích v jedné prostorové proměnné, oba přesahující rámec klasické nestability způsobené difuzí; v prvním případě je jeden druh přichycen k pevnému podloží (jedna rovnice neobsahuje ani difuzní ani advekční člen), v druhém případě obě rovnice mají stejný difuzní koeficient a stejný advekční koeficient. Používáme koncepci linearizované stability, zajímají nás tedy znaménka reálných částí kořenů takzvané disperzní relace. Využíváme, že jsme schopni analyticky spočítat vlastní čísla příslušných operátorů pro každý typ okrajových podmínek a standardními metodami analýzy odvodíme podmínky, za kterých dojde k nestabilitě způsobené difuzí.

Obdržené výsledky se liší od prosté intuice, ale také se kvalitativně liší od těch prezentovaných v literatuře, ve které byly výsledky získány jinými metodami. V prvně studovaném případě, s

jedním nepohyblivým druhem, homogenní stav je v případě dominantní advekce vždy stabilní, což je v kontrastu se studiem neomezených systémů [6, 7]. V druhém případě, s oběma druhy pohybujícími se, dochází v porovnání s Turingovou nestabilitou k významné relaxaci podmínek - libovolně malá advekce zapříčiní, že homogenní systém je vždy stabilní, a tedy do analýzy nestability nejsou přeneseny žádné svazující podmínky.

Přínosem tohoto článku je demonstrace a zdůraznění zásadního rozdílu mezi podmínkami vzniku prostorových struktur v RD a RDA systémech, ale také nutnosti použít jiné (a komplexnější) metody analýzy stability pro analýzu RDA operátorů. Protože vliv libovolně malé advekce na chování systému je markantní, vyvstává přirozená otázka: je advekce přítomna v reálných aplikacích nebo můžeme advekci bezpečně neuvažovat? K rozpoznání, zda RD nebo RDA modely lépe pasují na reálné situace, se budeme v naší budoucí práci podrobněji zabývat fyzikální podstatu modelovaných jevů.

Klíčová slova: reakčně-advekčně-difuzní systém, nestabilita způsobená difuzí

References

[1] Klika, V. and Kozák, M. and Gaffney, E. A., *A Note on Spatial Organisation in Systems with Advection.* will be submitted in Physical Review E.

[2] Turing, A., *The chemical basis of morphogenesis.* Phil. Trans. R. Soc. Lond. B, vol. 237, (1952),37–72.

[3] Rovinsky, A.B. and Menzinger, M., *Chemical instability induced by a differential flow*, Phys rev lett, vol. 69, num.8, (1992), 1193–1196.

[4] Andresén, P. and Bache, M. and Mosekilde, E. and Dewel, G. and Borckmanns, P., *Stationary space-periodic structures with equal diffusion coefficients*, Phys Rev E, vol. 60, num. 1, (1999), 297–301.

[5] Satnoianu, R. A. and Menzinger, M., *Non-Turing stationary patterns in flow-distributed oscillators with general diffusion and flow rates*, Phys Rev E, vol. 62, num. 1, (2000), 113.

[6] Nekhamkina, Olga and Sheintuch, Moshe, *Asymptotic solutions of stationary patterns in convection-reaction-diffusion systems*, Physical Review E, vol. 68, num. 3, (2003), 036207.

[7] Sheintuch, Moshe and Smagina, Yelena, *Stabilizing the absolutely or convectively unstable homogeneous solutions of reaction-convection-diffusion systems*, Physical Review E, vol. 70, numb. 2, (2004), 026221.

[8] Reddy, Satish C. and Trefethen, Lloyd N., *Pseudospectra of the convection-diffusion operator*, SIAM Journal on Applied Mathematics, vol. 54, num. 6, (1994), 1634–1649.

[9] Davis, E. B., *Pseudospectra of differential operators*, Journal of Operator Theory, vol. 43, num. 2, (2000), 243–262.

Distributed Data Processing: Constraint-Based Coscheduling of Computational Jobs and Data Placement in Distributed Environments[*]

Dzmitry Makatun

3rd year of PGS, email: dzmitry.makatun@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Jérôme Lauret, STAR, Brookhaven National Laboratory, USA
Hana Rudová, Faculty of Informatics, MU
Michal Šumbera, Nuclear Physics Institute, AS CR

Abstract. When running data intensive applications on distributed computational resources long I/O overheads may be observed as access to remotely stored data is performed. Latencies and bandwidth can become the major limiting factor for the overall computation performance and can reduce the CPU/WallTime ratio to excessive IO wait. Reusing the knowledge of our previous research, we propose a constraint programming based planner that schedules computational jobs and data placements (transfers) in a distributed environment in order to optimize resource utilization and reduce the overall processing completion time. The optimization is achieved by ensuring that none of the resources (network links, data storages and CPUs) are oversaturated at any moment of time and either (a) that the data is pre-placed at the site where the job runs or (b) that the jobs are scheduled where the data is already present. Such an approach eliminates the idle CPU cycles occurring when the job is waiting for the I/O from a remote site and would have wide application in the community. Our planner was evaluated and simulated based on data extracted from log files of batch and data management systems of the STAR experiment. The results of evaluation and estimation of performance improvements are discussed in this paper.

Keywords: constraint programming, Grid, Cloud, data processing , data transferring, data production, planning, scheduling, optimization, computational jobs, batch system.

Abstrakt. Při běhu datově náročných aplikací na distribuovaných výpočetních systémech se mohou vyskytnout dlouhé I/O prodlevy při vzdáleném přístupu k uloženým datům. Latence a propustnost se mohou stát hlavními limitujícími faktory pro celkový výkon výpočtu a mohou snížit poměr CPU/walltime díky nadměrnému čekání na I/O. Na základě poznatků z našeho předchozího výzkumu navrhujeme Plánovač využívající programování s omezujícími podmínkami, který rozvrhuje výpočetní úlohy a datová umístění (převody) v distribuovaném prostředí s cílem optimalizovat využití zdrojů a snížit celkový čas zpracování úloh. Optimalizace je dosaženo tím, že se zajistí, že žádný ze zdrojů (síťové spojení, datové úložiště a CPU) není přesycený v žádném časovém okamžiku a buď (a) data jsou předem umístěna tam, kde se

[*] This work has been supported by the Czech Science Foundation (GA-CR) and the Office of Nuclear Physics within the U.S. Department of Energy's. We would like to acknowledge and thank Brookhaven National Laboratory for resources that they made available.

úloha spustí, nebo (b) úlohy jsou spuštěny na strojích, na kterých jsou již data přítomna. Takový přístup eliminuje prostoje cyklů CPU vyskytujících se v případech, kdy úloha čeká na I/O ze vzdáleného místa a bude mít široké uplatnění v dané oblasti. Náš Plánovač byl vyzkoušen a simulován na základě údajů získaných ze záznamových souborů dávkových systémů experimentu STAR. Výsledky hodnocení a odhadu zvýšení výkonu jsou popsány v tomto článku.

Klíčová slova: Programování s omezujícími podmínkami, Grid, Cloud, zpracování dat, přenos souborů, plánování, optimalizace, výpočetní úlohy.

1 Introduction

Previous collaborative work between BNL and NPI/ASCR showed that the global planning of data transfers within the Grid can outperform widely used heuristics such as Peer-to-Peer and Fastest link (used in Xrootd)[1, 2]. Those results became the ground for continuation of research and extension of global planning to the entire data processing workflow, i.e., scheduling of CPU allocation, data transferring and placement at storage.

Long I/O overheads when accessing data from remote site can significantly reduce the application's CPUtime/WallTime ratio [3, 4]. For this reason, when setting up a data production at remote sites one has to consider the network throughput, available storage and CPU slots. When there are few remote sites involved in the data processing, the load can be tuned manually and simple heuristic may work, but, as the number of sites grows and the environment is constantly changing (site outage, fluctuations of network throughput and CPU availability), an automated planning of workflows becomes needed.

As an intuitive example of optimization let us consider a situation when a given dataset can be either processed locally, or can be sent to a remote site. Depending on transfer overhead it may appear to be optimal to wait for free CPU slots at the local site and process all the data there, or send a smaller fraction of the dataset for remote processing. Commonly used heuristics such as "*Pull a job when a CPU slot is free*" will not provide an optimization with respect to an overall processing makespan.

Another example arises from a workflow optimization which was done for inclusion of the ANL computational facility into the Cloud based data production of the STAR experiment [5]. In this case, and due to the lack of local storage at the site for buffering, the throughput of a needed direct on-demand network connection between BNL and ANL was not sufficient to saturate all the available CPUs at the remote site. An optimization was achieved by feeding CPUs at ANL from two sources: directly from BNL and through an intermediate site (PDSF) having large local caching and with better connectivity to ANL. This example illustrates an efficient use of indirect data transfers which cannot be guessed by simple heuristics. A general illustration of distributed resources used for data production and their interconnection is given at Figure 1.

Scheduling of computational jobs submitted by users (user analysis) has even more degrees of possible optimization: selection between multiple data sources, grouping of jobs that use the same input files. This case becomes even more complex due to a poor predictability of the user analysis jobs. However, the main question for optimization remains the same as for the examples above: How to distribute a given set of tasks over the available set of resources in order to complete all the tasks within minimal time?

Problems of scheduling, planning and optimization are being commonly solved with

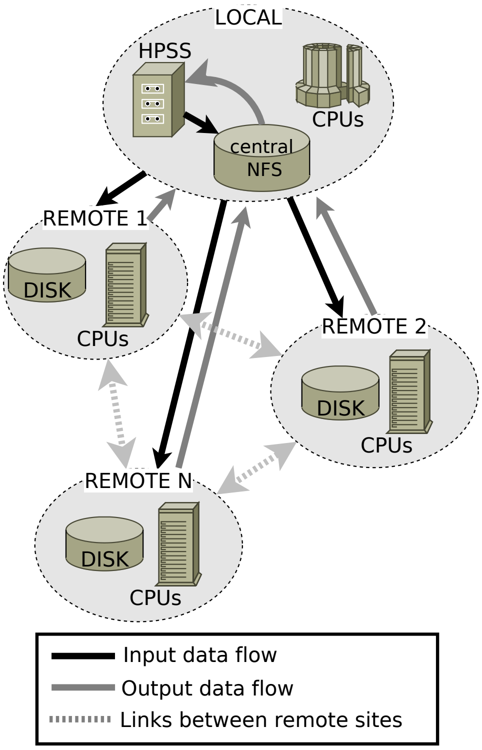


Figure 1: Schema of data production in the Cloud.

the help of Constraint Programming (CP) [6]. It is a form of declarative programming which is widely used in scheduling, logistics, network planning, vehicle routing, production optimization etc... In the next sections we will introduce our Constraint Satisfaction Problem (CSP) formulation for a data production at multiple sites and provide a simulation-based evaluation of the proposed model.

2 Model formulation, assumptions and search approach

A Constraint Satisfaction Problem (CSP) consists of domain variables, domains (a set of possible values of a variable) and constraints in form of mathematical expressions over variables. A solution to CSP is a complete assignment of values to variables which satisfies all the constraints. An optimal solution is the one with minimal/maximal value of a target function of variables.

We will introduce only the core concepts of our CSP formulation and search algorithms, omitting detailed mathematical expressions. The following input parameters are necessary to define our CSP:

Computational Grid (see Figure 1) is described by directed weighted graph where nodes are computational sites c with a given number of CPUs cpu_c and storage space $disk_c$; edges are network links l with weight $slowdown_l$ which is the time required to transfer a unit of data ($slowdown_l = \frac{1}{throughput_l}$). A dedicated storage facility, such as HPSS, can also be modeled as a node with $cpu_c = 0$.

Set of jobs. Each job j has a $duration_j$, it needs one input file of $inputSize_j$, produces one output file of $outputSize_j$, input file is placed at $inputSourceNodes_j$ and output file must be transferred to one of $outputDestinationNodes_j$.

Our goal is to create a schedule of jobs at computational sites, transfers over links and a placement of files at storages for a given computational Grid and a set of jobs. In order to solve this problem the variables of our model define the *resource selection* and *timing* of each task:

Resource selection variables define a node $ProcessingNode_j$ where the job j will be executed and a transfer path for each file f (either input or output of a job). The transfer path is described by a set of boolean variables X_{fl} where $true$ means that a file f will be transferred over a link l and $false$ means the opposite.

Time variables are: Js_j is a start time of a job j, Ts_{fl} is a start time of a transfer of a file f over a link l, Fs_{fc} is a start time of a placement of a file f at a node c, $Fdur_{fc}$ is a duration of a placement of a file f at a node c.

2.1 Model assumptions

Two important assumptions which are reused in the current model were proven in a previous work on global planning of data transferring in Grid [1].

The first assumption states that the entire set of jobs (queue) can be incrementally scheduled by subsets (chunks) without significant lose of optimality. Such an approach helps to reduce a search space and thus improve the planner performance. Moreover, planning for shorter periods and more frequent generation of plans (or replanning) provides the required level of adaptability to changing environment (outage of resources, fluctuating network bandwidth, etc).

The second assumption states that a network link can be modeled as an unary resource with no loss of generality. In other words, in our model we consider that only one file can be transferred over a link at a time. The measurements [1] have shown, that a sequential transfer of a set of files does not require more time then a parallel transfer of the same set of files over the same link.

2.2 Search overview

We use an incomplete search which can provide a suboptimal solution of required quality within a given time limit because the final goal is to create a planner that can process requests online. For a better search performance the overall problem is divided into two subproblems and the search is performed in two stages:

1. Planning Stage: instantiate a part of variables in order to assign resources for each task.

   (a) Assign jobs to computational nodes.

   (b) Select transfer paths for input and output files.

   (c) Estimate a makespan for a given resource assignment $estMakespan$.

   (d) Find a solution for the subproblem with a minimal estimated makespan.

2. Scheduling stage: define a start time for each operation.

   (a) Define the order of operations.

   (b) Put cumulative constraints on resources in order to avoid their oversaturation at any moment of time.

   (c) Find a solution with a minimal *makespan* which is the end time of the last task.

2.3 Constraints at the planning stage

At the planning stage the problem is to assign tasks (computational jobs and file transfers) to resources (computational nodes and links) in such a way that the set of tasks could be completed within minimal time. For this goal, an estimated makespan $estMakespan$ is a target function for minimization. It is defined as maximal time required by each resource to process all the tasks assigned to it.

For each job we have to assign a transfer path for an input and an output file which can be defined by the following constraints (see Figure 2):

1. An input file has to be transferred from one of its sources over exactly one link.

2. An output file has to be transferred to one of its destinations over exactly one link.

3. An intermediate node (neither source, destination nor selected for the job execution) either has exactly one incoming and outgoing transfer or is not on a transfer path: $\exists$ incoming transfer $\Leftrightarrow$ $\exists$ outgoing transfer.

4. There must exist exactly one incoming transfer of an input file and exactly one outgoing transfer of an output file at the node which was selected for the job execution.

5. A file can be transferred from/to each node at most once.

In addition, we use constraints for loop elimination similarly as it is described in [7].

2.4 Constraints at the scheduling stage

At the scheduling stage the problem is to assign a start time for each task. The following constraints on order of tasks are implemented:

- An outgoing transfer of a file from a node can start only after an incoming transfer to that node is finished. The first transfer of an input file from its source and the first transfer of an output file from the processing node are exceptions from this constraint.

- A job can start only after the input file is transferred to the selected processing node.

- An output file can be transferred only after the job is finished.

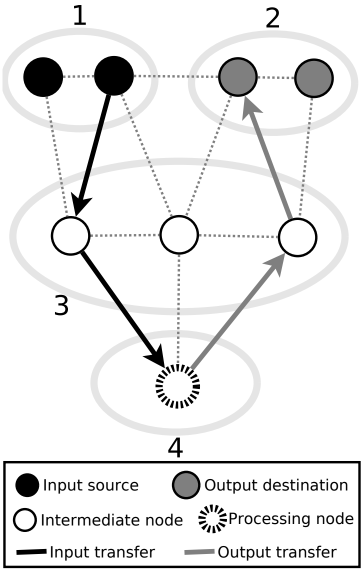


Figure 2: An example of a transfer path. Illustration for constraints 1-4 in section 2.3.

Table 1: Variables and parameters used in cumulative constrains on resources.

Task	Start	Duration	Usage	Limit
Job	Js_{jc}	$duration_j$	1	cpu_c
Transfer	Ts_{fl}	$size_f \cdot slowdown_l$	1	1
File placement	Fs_{fc}	$Fdur_{fc}$	$size_f$	$disk_c$

- A reservation of space for a file at a node is made when a transfer to that node starts.
- A file can be deleted from the start node of a link after the transfer is finished.
- A reservation of space for an output file is made at the processing node when the job starts.
- An input file can be deleted from a processing node after the job is finished.

Cumulative constraints are widely used in Constraint Programming for description of resource usage by tasks. Each cumulative constraint requires that a set of tasks given by *start times*, *durations* and *resource usage*, never require more than a *resource limit* at any time. In our case we use three sets of cumulative constraints: for CPUs, storages and links (see Table 1).

3 Simulations

The constraint satisfaction problem was implemented using MiniZinc [8] and Gecode [9] was used as a solver. The timelimit was set to 3 minutes for both planning and scheduling stages. The simulations were running under Windows 8 64-bit on a computer with Intel i5 (4 cores) 2.50 GHz processor and 6 GB of memory installed. The Gecode solver was running in a parallel mode using 4 threads.

Two sets of simulations were performed for testing of the proposed model. In both cases the simulated environment consisted of 3 nodes: a central storage HPSS ($cpu_{HPSS} = 0$) which was the single source for input files and the single destination for output files, a local processing site and a remote processing site. The slowdown of links between the central HPSS and the local site was set to 0, which means that transfer overheads to/from the local site are negligible comparing to a job duration in both sets of simulations. The number of CPUs at processing nodes and slowdown of links between the central HPSS and the remote node were set-up differently for each set of simulations. Storage constraints were not considered in these simulations. Four different scheduling strategies were compared:

Local: All the jobs are submitted to the local site only. This strategy was used as a base line for comparison against other strategies.

Equal CPU load: Jobs are distributed between nodes with the goal to maintain an equal ratio of job duration per CPU. Each input file is transferred prior to the start of a job. At each node jobs are executed in input order.

Data transferred by job: Each CPU pulls a job from the queue when it is idle, then it has to wait for an input transfer before the job execution starts.

Optimized: This strategy is based on the model proposed in this paper.

In the first set of simulations the main idea was to evaluate different scheduling strategies in a setup where overheads of an input and an output transfers to a remote site taken together are comparable to the job duration. Obviously, in such environment transfer overhead can significantly influence the overall makespan. The number of CPUs at both local and remote sites was set to 10 and the slowdown from/to remote node was set to 1 (one time unit to transfer one unit of size). Several testing sets of jobs were created using a random number generator. For each job an input size was equal to a random value in interval 1..20 of size units. An output size and a job duration were proportional to the input size with a random factor close to 1 and 2 respectively.

Results of the first set of simulations are presented at Figure 3. The plot shows the dependence of a makespan on a number of jobs (bunch) scheduled in one experiment. As it can be seen at the plot, maintaining equal CPU load at local and remote sites (Equal CPU load) increases the makespan more then twice; while scheduling with consideration of a transfer overhead (Optimized) reduces the makespan by 15% compared to local only processing (Local).

In the second set of simulations the slowdown of the links to/from the remote site was increasing in each simulation proportionally to a slowdown factor. The parameters of jobs were taken from logging system of the STAR experiment's data production at computational site KISTI (South Korea) [10]. The average job duration was 3,000 minutes and average time of transfer was 5 and 10 minutes to/from the remote site respectively (in the simulations where the slowdown factor = 1). Then, in further simulations the transfer

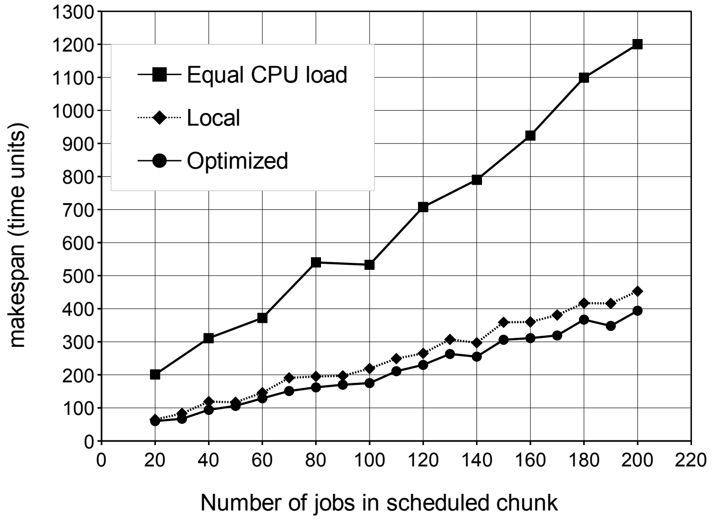


Figure 3: Results of testing simulations. In the simulated environment the I/O transfer overheads to a remote site are comparable to a job duration. The "Equal CPU load" heuristics failed to decrease the makespan using remote resources, while the proposed global planning approach (Optimized) has decreased the makespan by 15%.

times increase proportionally to the slowdown factor. In the simulated environment 80% of CPUs were available at the local site and 20% at the remote site. 2,000 of jobs were scheduled stepwise by subsets (chunks) of 200.

The plot at Figure 4 shows the gain in a makespan delivered by different scheduling policies compared to the job execution at the local site only. The curves shows the performance of the scheduling policies when an overhead of transfer to the remote site increases proportionally to the slowdown factor. When the transfer overhead becomes significant both heuristics ("Equal CPU load" and "Data transferred by job") fail to provide an efficient usage of the remote resources (the makespan improvement goes below zero). Negative makespan improvement means that, in this case, it would be faster to process all the data locally than to distribute it between several sites relying on the heuristic. The proposed global planning approach (Optimized) systematically provides a smaller makespan and adapts to the increase of transfer overheads better then the other simulated heuristics. It was able to provide a positive gain in makespan by using remote resources even when the transfer overhead is comparable to a job duration.

4 Conclusion

A model for scheduling of data production over Grid was formulated in form of constraint satisfaction problem and solved using constraint programming.

Testing simulations has shown that in an environment, where a remote site has the same CPU number as a local site, but the data transfer overhead is comparable to a job duration, maintaining equal CPU load at local and remote sites increases the makespan more then twice; while scheduling with consideration of a transfer overhead can reduce the makespan by 15% compared to local only processing.

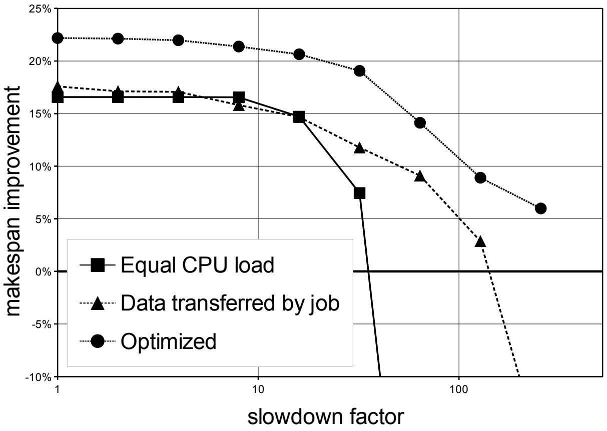


Figure 4: Results of simulations for real data production. Three strategies were evaluated and compared to a ideal local production. The optimized solution (our model) clearly provides the highest gain.

The simulations based on data extracted from log files of batch and data management systems of the STAR experiment has shown that the proposed global planning approach systematically provides a smaller makespan and adapts to the increase of transfer overheads better then the other simulated heuristics.

The proposed approach can provide an *optimization* and an *automatic adaptation* to fluctuating resources with no need for manual adjustment of a workflow at each site or tuning of heuristics.

The future development of global planning for data processing in Grid is ongoing. In future we plan to test this approach on problems of larger size (more nodes, CPU's and links) and improve the search performance in order to enable online scheduling in real environment.

References

[1] Zerola M, Lauret J, Barták R and Šumbera M 2012 One click dataset transfer: toward efficient coupling of distributed storage resources and CPUs *J. Phys.: Conf. Series* **368** 012022

[2] Makatun D, Lauret J and Šumbera M 2014 Study of cache performance in distributed environment for data processing, *J. Phys.: Conf. Series* **523** 012016

[3] Horký J, Lokajíček M and Peisar J 2013 *Influence of Distributing a Tier-2 Data Storage on Physics Analysis* (Beijing: 15th Int. Workshop on Advanced Computing and Analysis Techniques in Phys. Res.)

[4] Betev L, Gheata A, Gheata M, Grigoras C and Hristov P 2014 Performance optimisations for distributed analysis in ALICE *J. Phys.: Conf. Series* **523** 012014

[5] Balewski J, Lauret J, Olson D, Sakrejda I, Arkhipkin D, Bresnahan J, Keahey K, Porter J *et al.* 2012 Offloading peak processing to virtual farm by STAR experiment at RHIC *J. Phys.: Conf. Series* **368** 012011

[6] Rossi F, Beek P and Walsh T 2006 *Handbook of Constraint Programming* (Amsterdam: Elsevier)

[7] Troubil P and Rudová H 2011 Integer Linear Programming Models for Media Streams Planning. (Int. Conf. on Applied Operational Res.) *Lecture Notes in Management Sc.* **3** 509-22

[8] Nethercote N, Stuckey P, Becket R, Brand S, Duck G and Tack G 2007 MiniZinc: Towards a Standard CP Modelling Language *Lecture Notes in Comp. Sc.* **4741** 529-543

[9] Tack G 2008 *Gecode: an Open Constraint Solving Library* (Paris: OSSICP 08 Workshop at CP-AI-OR 08)

[10] Korea Institute of Science and Technology Information (KISTI) http://en.kisti.re.kr/

Reaching Quantum Consensus with Partial Swap Interactions*

Jiří Maryška

2nd year of PGS, email: maryska.jiri@gmail.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Igor Jex, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Jaroslav Novotný, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Recently developed formalism of random unitary operations enables us to study the asymptotic dynamics of a large class of open quantum systems with unprecedented accuracy. In this work we focus on the general solution of the asymptotic dynamics of an ensemble of identical quantum systems, each associated with finite-dimensional Hilbert space and local free Hamiltonian. These systems interact with each other by random binary collisions, which are well separated in time. We derive equations which determine the attractor space for such type of open quantum system for a general binary interaction Hamiltonian and discuss the case in which the free Hamiltonian commutes with the interaction Hamiltonian. Next, we choose the interaction between systems to have a form of a well known partial swap quanum gate. We show, that the asymptotic state satisfies the requirements of SSC (symmetric state consensus) defined in [7]. For a factorizable initial state we next show that the reduced asymptotic state of single system has the form of an weighted avarage of the initial states of systems

Keywords: Quantum network, asymptotic evolution, quantum consensus, collision model

Abstrakt. Nedávno vyvinutý formalismus náhodných unitárních operací nám dovoluje studium asymptotické dynamiky velké třídy otevřených kvantových systémů s dříve nevídanou přesností. V této práci se zaměřujeme na obecné řešení asymptotické dynamiky souboru identických systémů, každý s příslušným Hilbertovým prostorem a lokálním volným Hamiltoniánem. Tyto systémy spolu náhodně interagují binárními kolizemi, které jsou oddělené v čase. Odvodíme rovnice, které určují atraktorový prostor pro tento typ otevřených kvantových systémů pro obecný binární interakční Hamiltonián a diskutujeme případ, kdy volný Hamiltonián komutuje s interakčním Hamiltoniánem. V dalším kroku zvolíme za konkrétní interakci dobře známé kvantové hradlo partial swap. Ukážeme, že asymptotický stav splňuje požadavky SSC definovaného v [7]. Pro faktorizovatelný počáteční stavu dále ukážeme, že redukovaný asymptotický stav příslušející jednomu systému má tvar váženého průměru počátečních stavů systémů.

Klíčová slova: Kvantové sítě, asymptotická evoluce, kvantová shoda, kolizní model

[*] Tato práce byla podpořena grantem SGS13/217/OHK4/3T/14.

1 Introduction

Classical networks received a lot of attention in past decades as they are able to describe complex systems as the internet or social groups [1, 2]. Classical network is consisted of a set of nodes, often called agents, each node representing a single system. These nodes are connected by edges, which represent a certain type of interaction between agents. Quantum network [3, 4, 5] is a generalization of classical network, in which classical systems, which are represented by nodes, are replaced by quantum systems.

A recently developed formalism of random unitary operations (RUO) [6] enables us to analytically study open quantum systems, i.e. quantum systems which undergo a non-unitary time evolution (usually due to interaction with an external system or because of lack of knowledge about system), after sufficiently large amount of time of evolution. As quantum networks are open quantum systems, we are thus able to study their asymptotic properties with unprecedented accuracy.

In many application of both classical and quantum systems, one needs to reach a consensus within a given network, i.e. the situation, in which all agents have certain properties identical [7]. These could be the expectation values of some observable, the state of agent etc.

Starting with a network consisted of $N \geq 2$ quantum systems equipped with Hilbert space $\mathscr{H} = \mathbb{C}^d$, $d \in \mathbb{N}$ (so-called qudits) with identical free Hamiltonian H, we let qudits to intract with each other by a sequence of random binary interactions. These interactions have the form of partial-swap interactions. We show that for an arbitrary initial state $\rho(0)$, the asymptotic state after $n \gg 1$ interactions $\rho_\infty(n)$ meets the conditions of SSC (Symmetric State Consensus) introduced in [7]. Moreover, if the initial state is the product state, the reduced single qudit asymptotic state is found as the homogenization of the initial states of all qudits with free evolution.

This work is organized as follows. In section 2 we introduce random unitary operations (RUO), we sum up their basic properties and show the method of finding their asymptotic evolution. In section 3, we focus on the case of binary collisions in an ensemble of identical systems. In section 4 we apply results from section 3 on the case of partial swap quantum gate. We find the asymptotic state and show that it fulfils the requirements of SSC. Furthermore we discuss the form of the reduced asymptotic state of a single system for the factorized initial state. The summary of this work is given in section 5.

2 Attractor method for dynamics generated by RUO

Assume a quantum system associated with a finite dimensional Hilbert space $\mathscr{H}$ and let us denote $\mathcal{B}(\mathscr{H})$ the Hilbert space of all operators acting on the Hilbert space $\mathscr{H}$ (equipped with Hilbert-Schmidt scalar product. One step of evolution is given by the RUO $\Phi : \mathcal{B}(\mathscr{H}) \rightarrow \mathcal{B}(\mathscr{H})$ whose action on the system initially prepared in a general mixed state ρ can be written in the form

$$\Phi(\rho) = \sum_{i=1}^{n} p_i U_i \rho U_i^\dagger \tag{1}$$

with unitary operators U_i and probability distribution $\{p_i\}_{i=1}^n$. RUO Φ is a quantum operation with Kraus operators defined by $K_i = \sqrt{p_i}U_i$. It belongs to the class of trace-preserving unital quantum operations leaving the maximally mixed state invariant. From a physical point of view, RUO takes into account our lack of classical knowledge which unitary evolution the system undergoes and incorporates in incoherent manner all exclusive unitary paths of evolution represented by different unitary operators U_i properly weighted with probabilities associated with all these paths.

The evolution of the system results from repeated applications of RUO Φ. Starting from the initial state $\rho(0)$, the n-th step of the iterated dynamics reads $\rho(n) = \Phi(\rho(n-1))$. In general, the operator Φ is neither hermitian nor normal linear map and consequently a diagonalization of RUO Φ in some orthonormal basis is not guaranteed. Fortunately, the latter does not apply to the asymptotic part of evolution and one can exploit the fact that the asymptotic regime of the evolution takes place in the so-called attractor space $\mathrm{Atr}(\Phi) \subset \mathcal{B}(\mathscr{H})$ constructed as

$$\mathrm{Atr}(\Phi) = \bigoplus_{\lambda\in\sigma_{|1|}} \mathrm{Ker}(\Phi - \lambda I). \tag{2}$$

The attractor spectrum $\sigma_{|1|}$ denotes the set of all eigenvalues λ of Φ with $|\lambda| = 1$. For a given $\lambda \in \sigma_{|1|}$ the corresponding kernel $\mathrm{Ker}(\Phi - \lambda I)$ is constructed as the solution of attractor equations

$$U_i X_{\lambda,j} U_i^\dagger = \lambda X_{\lambda,j} \quad \forall i. \tag{3}$$

Any state in the asymptotic regime, i.e. for large number of iterations n, takes the form

$$\rho_\infty(n) = \sum_{\lambda\in\sigma_{|1|},i=1}^{D_\lambda} \lambda^n \mathrm{Tr}\left[\rho(0) X_{\lambda,i}^\dagger\right] X_{\lambda,i}, \tag{4}$$

where $D_\lambda = \dim[\mathrm{Ker}(\Phi - \lambda I)]$ and $\{X_{\lambda,i} | i \in \{1,\ldots,D_\lambda\}\}$ forms an orthonormal basis of the subspace $\mathrm{Ker}(\Phi - \lambda I)$. Apparently, attractors solving equations (3) are not affected by the particular form of the probability distribution $\{p_i\}$ provided that $p_i \neq 0$. As a direct consequence, the asymptotics of given RUO is independent on particular values p_i as long as they are different from zero.

At this point we have to stress that attractors are not, in general, density operators, i.e. they do not represent states. This can cause difficulties in analysis of asymptotic dynamics like fixed points, decoherence-free subspaces. To overcome this obstacle one may employ the so-called pure-state method [8]. Without going into details we recapitulate its main points. Let us denote $\{|\phi_{\alpha,j_\alpha}\rangle\}$ the orthonormal basis of common eigenstates of unitaries $\{U_i\}$, i.e.

$$U_i|\phi_{\alpha,j_\alpha}\rangle = \alpha|\phi_{\alpha,j_\alpha}\rangle \qquad \forall i, \tag{5}$$

where index j_α takes into account a degeneracy of common eigenvalue α. Any matrix from the span of $\{|\phi_{\alpha,j_\alpha}\rangle\langle\phi_{\beta,j_\beta}|\}$ with fixed product $\alpha\overline{\beta} = \lambda$, i.e.

$$X = \sum_{\alpha\overline{\beta}=\lambda,j_\alpha,j_\beta} A_{\beta,j_\beta}^{\alpha,j_\alpha} |\phi_{\alpha,j_\alpha}\rangle\langle\phi_{\beta,j_\beta}|, \tag{6}$$

satisfies equations

$$U_i X U_j^\dagger = \lambda X \qquad \forall i, j \tag{7}$$

and consequently belongs to the subspace of attractors corresponding to the eigenvalue λ of RUO Φ. On the contrary, any operator satisfying (7) can be decomposed into common eigenvectors as (6). Attractors which can be constructed from common eigenvectors are called p-attractors. As they satisfy more restricting set of equations (7) they do not constitute the whole attractor space. In particular, the identity operator is an attractor but not a p-attractor (except the trivial case of an unitary evolution). The space of attractors always contains, as a minimal subspace, the span of p-attractors and the identity operator. Surprisingly, in some nontrivial cases this minimal subspace forms the whole attractor set and asymptotic dynamics can be analyzed easier. Indeed, assume the orthogonal projection π onto the subspace of common eigenstates of unitaries $\{U_i\}$. Let $\tilde{\pi}$ be its orthogonal complement projection satisfying $\pi + \tilde{\pi} = I$. Both projections are fixed points of quantum operation (1), they both reduce the random unitary operation (1) and the asymptotic dynamics after sufficiently number of iterations of the state $\rho(0)$ can be written as

$$\rho(n \gg 1) = U_i^n \pi \rho(0) \pi \left(U_j^\dagger\right)^n + \frac{\mathrm{Tr}[\rho(0)\tilde{\pi}]}{\mathrm{Tr}\,\tilde{\pi}} \tilde{\pi}, \tag{8}$$

for any pair of indices i, j. In this case the asymptotic evolution can be understood as an incoherent mixture of the unitary dynamics inside the subspace of common eigenstates and the maximally mixed state living on the orthogonal complement of this subspace. We often encounter the special case $\sigma_{|1|} = \{1\}$, which further simplifies the asymptotic evolution given by (8) to the form

$$\rho_\infty = \pi \rho(0) \pi + \frac{\mathrm{Tr}\left[\rho(0)\tilde{\pi}\right]}{\mathrm{Tr}[\tilde{\pi}]} \tilde{\pi}. \tag{9}$$

In such situation any initial quantum state $\rho(0)$ evolves towards the stationary state (9).

3 Binary interactions within scope of RUO

Formalism introduced in the last section is in many situations too general to work with, it is thus convenient to adapt it to fit a particular case. Suppose we have at our disposal an ensemble of N qudits. These qudits interact with each other by binary interaction. The nature of these interactions is such that there exists a time resolution Δt within which maximally one (random) pair of qudits interacts. All interactions are thus separable in time, but otherwise they are uncontrollable. Let the interaction Hamiltonian during collision between systems i and j be H_{ij}^{int}. Our goal in this section is to find connection of the asymptotic evolution of such ensamble in which each qudit is equipped with free Hamiltonian H_0 and the same ensamble without considering the free Hamiltonian of qudits.

Let us first study the ensemble without considering the free Hamiltonian of qudits. The total Hamiltonian during collision of qudits i and j can be thus written as

$$H_{ij} = H_{ij}^{int}.$$

A single evolution step is according to previous section described by a RUO Φ_1 defined as

$$\Phi_1(\rho) = \sum_{ij} p_{ij} V_{ij}^{int} \rho (V_{ij}^{int})^\dagger + p_0 \rho \tag{10}$$

with

$$V_{ij}^{int} = \exp\left[i\Delta t H_{ij}^{int}\right].$$

The last component of (10) reflects the fact, that no collision must occur within the time resolution Δt. For the solution of the asymptotic dynamics of such system, one needs to find the attractor space $\mathrm{Atr}(\Phi_1)$ by solving equations (3). According to [6], all attractors of such system correspond to the eigenvalue $\lambda = 1$. We thus have

$$\mathrm{Atr}(\Phi_1) = \mathrm{Ker}[\Phi - I] = \mathrm{Span}\{X_{1,1}, \dots X_{1,m}\}.$$

The asymptotic evolution has thus form

$$\rho_\infty = \sum_{i=1}^{m} \mathrm{Tr}\left[\rho(0) X_{1,i}^\dagger\right] X_{1,i}. \tag{11}$$

In the second case, the total Hamiltonian during collision of qudits i and j has the form

$$H_{ij} = \sum_{k=1}^{N} H_k + H_{ij}^{int},$$

where H_k is defined by

$$H_k = I^{\otimes k-1} \otimes H \otimes I^{\otimes N-k}$$

with $H = H^\dagger$ being the operator on $\mathbb{C}^d$ and I being the identity operator on $\mathbb{C}^d$. A single evolution step is according to previous section described by a RUO Φ_2 defined as

$$\Phi_2(\rho) = \sum_{ij} p_{ij} V_{ij} \rho (V_{ij})^\dagger + p_0 U_0 \rho U_0^\dagger \tag{12}$$

with

$$V_{ij} = \exp\left[i\Delta t H_{ij}\right],\ U_0 = \exp\left[\sum_{k=1}^{N} H_k\right].$$

Let us define following operators:

$$\begin{aligned} u_0 &= \exp[i\Delta t H], \\ U_{ij} &= \exp[i\Delta t(H_i + H_j + H_{ij}^{int})] u_0^{\dagger\otimes 2}. \end{aligned}$$

These operators satisfy $V_{ij} = U_{ij} U_0$. It is easy to show, that operator $X \in \mathcal{B}(\mathscr{H})$ satisfies equations (3) with $\lambda = 1$ if and only if it satisfies the following equations:

$$\begin{aligned} U_{ij}^\dagger X U_{ij} &= X, \quad \forall i, j, \\ U_0 X U_0^\dagger &= \lambda X \end{aligned} \tag{13}$$

for certain $|\lambda| = 1$. All solutions of these equations form the attractor space $\mathrm{Atr}(\Phi_2)$. Unlike in previous case, the particular attractors correspond to different eigenvalues λ. We thus denote all solutions of equations (13) with fixed λ as $X_{\lambda,k}$. The asymptotic state then satisfies (4). One could think that the first set of these equations determines the attractor space and the latter set determines the eigenvalues of particular attractors. This is however true only if the free Hamiltonians comute with the interaction Hamiltonian, i.e. when $\left[H_i + H_j, H_{ij}^{int}\right] = 0$. If this is not the case, the latter set of equations can narrow the set of operators found by the first set of equations.

The situation is furthermore simplified, when the free Hamiltonian commutes with the interaction Hamiltonian, i.e. when $\left[H_i + H_j, H_{ij}^{int}\right] = 0$. One can notice, that in this situation, we have

$$U_{ij} = V_{ij}^{int}. \tag{14}$$

This implies that the attractor spaces of Φ_1 and Φ_2 are identical. One can thus make use of the solution for the asymptotic evolution (11) of Φ_1 and write the asymptotic state in the form

$$\rho_\infty(n) = \sum_{i=1}^{m} \mathrm{Tr}\left[\rho(0)X_{1,i}^\dagger\right] U_0^n X_{1,i} (U_0^\dagger)^n. \tag{15}$$

By solving the asymptotic dynamics of Φ_1 one thus in general solves the asymptotic dynamics of Φ_2, which is given as a free evolution of the asymptotic state of Φ_1.

4 Asymptotic evolution of partial swap interactions

In this section we study the asymptotic dynamics of an ensemble of N qudits equipped with local free Hamiltonian H, which interact with each other via partial-swap interactions [9, 10].

Partial swap interaction was introduced by Bužek et. al. in [9, 10] as a mean to study thermodynamical properties of system of qubits. Althought it was defined as the operator acting on pair of qubits, its generalization to the operator acting on pair of qudits is straightforward. We consider a one-parameter family of operators $U_{jk}^{(\varphi)}$ acting non-trivially on pair of qudits j and k according to the following formula:

$$U_{jk}^{(\varphi)} = e^{i\varphi} \cos\varphi \; I_{jk} + i e^{-i\varphi} \sin\varphi \; S_{jk} \tag{16}$$

with I_{jk} being the identity operator on qudits j and k and S_{jk} being the swap operator on qudits j and k acting as

$$\begin{aligned} &S_{jk}(|\psi_1\rangle \otimes \cdots \otimes |\psi_j\rangle \otimes \cdots \otimes |\psi_k\rangle \otimes \cdots \otimes |\psi_N\rangle) = \\ &= |\psi_1\rangle \otimes \cdots \otimes |\psi_k\rangle \otimes \cdots \otimes |\psi_j\rangle \otimes \cdots \otimes |\psi_N\rangle . \end{aligned}$$

Operator (16) thus partially swaps the state of qudits j and k, hence its name. Partial swap possesses a number of interesting properties. For our purpose, the most important is the fact that partial swap commutes with an arbitrary local free Hamiltonian H [10].

The previous section enables us to use this fact to solve the asymptotic dynamics of RUO $\Phi^{(\varphi)}$ defined as

$$\Phi^{(\varphi)}(\rho) = \sum_{jk} p_{jk} V_{jk}^{(\varphi)} \rho (V_{jk}^{\varphi})^{\dagger} + p_0 U_0 \rho U_0^{\dagger}. \tag{17}$$

with $V_{jk}^{(\phi)} = U_0 U_{jk}^{(\phi)}$ To find the asymptotic state $\rho_{\infty}(n)$, we first find the attractor space of the RUO $\Phi_{PS}^{(\varphi)}$ defined as

$$\Phi_{PS}^{(\varphi)}(\rho) = \sum_{jk} p_{jk} U_{jk}^{(\varphi)} \rho (U_{jk}^{\varphi})^{\dagger} + p_0 \rho. \tag{18}$$

RUO (18) represents partial swap interactions without considering local free Hamiltonian H. The attractor equations (3) are significantly simplified when transfered to the equations for attractor elements. For this purpose, we define the following notation:

$$X_{kl}^{ij} := \langle ij|X|kl\rangle$$

with $\{|i\rangle \,|i \in \{0, \ldots, d-1\}\}$ forming orthonormal basis of $\mathbb{C}^d$. This notation takes into account only those qudits, on which the operator (16) acts nontrivially. The indicies of other $N-2$ qudits are omitted for sake of simplicity. With the help of this notation, the attractor equations (3) for operators (16) can be tranfered to equations for matrix elements:

$$X_{ij}^{kl} = X_{ji}^{lk}. \tag{19}$$

Any operator $X \in \mathcal{B}(\mathscr{H})$, whose elements for any pair of qudits satisfy (19) is an attractor corresponding to eigenvalue $\lambda = 1$ of RUO (18). From the form of equations (19) one can see, that the attractors are a totally symmetric operators with respect application of the swap operators S_{jk} for any j, k. Let us denote $\mathbf{i} = (i_1, \ldots, i_N)$ with $i_k \in \{0, \ldots, d-1\}$. Operators constituting the orthogonal basis of the attractor space corresponding to the RUO (18) can be denoted as $X_{\mathbf{j}}^{\mathbf{i}}$. They are defined as

$$X_{\mathbf{j}}^{\mathbf{i}} = \frac{1}{N!} \sum_{\pi \in S_N} |\pi(\mathbf{i})\rangle \langle \pi(\mathbf{j})|$$

with S_N being the set of all permutations on the set $\{1, \ldots, N\}$. However, in the set $\{X_{\mathbf{j}}^{\mathbf{i}} | \mathbf{i}, \mathbf{j}\}$ it contains a lot of duplicate operators. For fixed $\mathbf{i}$ and $\mathbf{j}$, the operators $X_{\mathbf{j}}^{\mathbf{i}}$ and $X_{\pi(\mathbf{j})}^{\pi(\mathbf{i})}$ are identical for arbitrary $\pi \in S_N$. A quick calculation reveals, that dimension of the attractor space $\mathrm{Atr}(\Phi_{PS})$ is

$$\dim\left(\mathrm{Atr}(\Phi_{PS})\right) = \binom{d^2 + N - 1}{N}.$$

After discarding the duplicates and orthonormalization, we arrive to the orthonormal basis of the attractor space $\{Y_1, \cdots, Y_{\binom{d^2+N-1}{N}}\}$. The state for $n \gg 1$ has thus the form

$$\rho_{\infty}^{PS}(n) = \sum_{i=1}^{\binom{d^2+N-1}{N}} \mathrm{Tr}[\rho(0) Y_i^{\dagger}] Y_i \tag{20}$$

From the formalism developed in the previous section, we are immediately able to write the state of the RUO (17) for $n \gg 1$ as

$$\rho_\infty(n) = \sum_{i=1}^{\binom{d^2+N-1}{N}} \mathrm{Tr}[\rho(0)Y_i^\dagger]U_0^n Y_i U_0^{\dagger n}. \tag{21}$$

As all operators Y_i are symmetric with respect to an arbitrary permutation and the second set of equations (13) holds, this state is clearly symmetric with respect to an arbitrary permutation and it thus satisfies the definition of SSC [7] for any initial state $\rho(0)$. If the initial state is in the form

$$\rho(0) = \rho_1 \otimes \cdots \otimes \rho_N,$$

then after a lenghty, but not complicated calculation, one arives to the result

$$\rho_\infty^{(k)}(n) = \frac{1}{N} u_0^n \left(\sum_{i=1}^{N} \rho_i \right) u_0^{\dagger n} \tag{22}$$

with $\rho_\infty^{(k)}(n) = \mathrm{Tr}_k[\rho_\infty(n)]$, where $\mathrm{Tr}_k[\cdot]$ is a partial trace over all systems excluding system k. The reduced asymptotic state of a single system is thus found as a free evolution of a state, which arises as an one system reduction of homogenization of an initial state $\rho(0)$.

5 Conclusion

In the previous sections we introduced the RUO model which enables us to calculate the asymptotic dynamics of a large class of open quantum systems analytically. Afterwards we focused on the case of an ensemble of qudits, which evolve acording to local free Hamiltonian H and undergo random binary collisions, which are described by interaction Hamiltonian H_{ij}^{int}. Main result of this section are the attractor equations (13) which describe the asymptotic state of the ensemble. For the case, when the free Hamiltonian commutes with the interaction Hamiltonian, these results imply, that the asymptotic state of such ensemble is givem by free evolution of the asymptotic state (11).

Regarding partial swap interactions, we derived the form of the asymptotic state of such interactions (21) and showed, that it satisfies the definition of SSC introduced in [7]. For factorizable initial state, we have found a simple expression (22) for the asymptotic state of a single system.

There are still open questions regarding partial swap concerning its connection to thermalization processes in interacting quantum system. It is easy to see, that the operators which form an orthonormal basis of the attractor space of arbitrary RUO form an complete set of the integrals of motion of the corresponding RUO. From the properties of the attractors, one can easily see, that the asymptotic state can be written as a generalized thermal state (at least in the limit of its parameters). Thermalization of qudits and its connection with partial swap will be focused in the future research.

References

[1] M. Ballerini et. al., *Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study* PNAS **105**, 1232-1237, 2007.

[2] C. Huepe, G. Zschaler, A. Do, T. Gross, *Adaptive-network models of swarm dynamics*, New J. Phys. **13**, 073022, 2011,

[3] J. Gough and M. R. James, *The Series Product and Its Application to Quantum Feedforward and Feedback Networks*, IEEE Trans. Autom. Control, **54**, 2530-2544, 2009.

[4] X. Wang, P. Pemberton-Ross and S. G. Schirmer, *Symmetry and Subspace Controllability for Spin Networks With a Single-Node Control* IEEE Trans. Autom. Control, **57**, 1945-1956, 2012.

[5] A. Acín, I. Cirac, M. Lewenstein, *Entanglement percolation in quantum networks*, Nature Physics **3**, 256-259, 2007

[6] J. Novotný, G. Alber, and I. Jex, *Asymptotic evolution of random unitary operations*, Central European Journal of Physics, **8** 1001, 2010.

[7] L. Mazzarella, A. Sarlette, F. Ticozzi, *Consensus for Quantum Networks: From Symmetry to Gossip Iterations*, arXiv:1303.4077

[8] B. Kollár, J. Novotný, T. Kiss, and I. Jex, *Discrete time quantum walks on percolation graphs* Eur. Phys. J. Plus **129**, 103 (2014),

[9] V. Scarani, M. Ziman, P. Štelmachovič, N.Gisin and V. Bužek, *Quantum homogenization*, arXiv:quant-ph/0110164

[10] V. Scarani, M. Ziman, P. Štelmachovič, N.Gisin and V. Bužek, *Thermalizing Quantum Machines: Dissipation and Entanglement*, Phys. Rev. Lett. **9**, 88, 2001

Bifurcation of the Laplace Equation with an Interior Unilateral Condition[*]

Josef Navrátil

2nd year of PGS, email: pepa.navratil@gmail.com
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Milan Kučera, Institute of Mathematics, AS CR

Abstract. This article concerns with the bifurcation problem for an equation $-\triangle u + \lambda u + g(\lambda, u) = 0$ with an interior unilateral condition. A form of a solution of the variation inequality is found and regularity of the solution around the obstacle is discussed.

Keywords: bifurcation, Laplace equation, unilateral condition

Abstrakt. Tento článek se zabývá bifurkací pro rovnici $-\triangle u + \lambda u + g(\lambda, u)$ s jednostrannou podmínkou na části vnitřku oblasti, na které rovnici řešíme. Je nalezen tvar řešení a dokázána regularita v okolí vnitřní hranice.

Klíčová slova: bifurkace, Laplaceova rovnice, jednostranna podminka

1 Introduction

Let $\Omega, \Omega_U \subset \mathbb{R}^2$ be open sets and suppose $\Omega_U \subset\subset \Omega$, i.e. $\overline{\Omega_U} \subset \Omega$. Moreover, let $\partial\Omega_U$ be of a class C^2 and Ω, Ω_U be a simply connected sets. Situation is scetched on the Figure 1.

Figure 1: Area

We define a map $g(\lambda, \xi) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ which satisfies for $p > 1$ following growth conditions:

$$\begin{aligned} \exists C_1, C_2 \quad \forall \lambda \in \mathbb{R} \quad \forall \xi \in \mathbb{R} : \quad & |g(\lambda, \xi)| \le C_1(1 + |\xi|^{\frac{p}{2}}), \\ & \left|\frac{\partial g}{\partial \lambda}(\lambda, \xi)\right| + \left|\frac{\partial g}{\partial u}(\lambda, \xi)\right| \le C_2(1 + |\xi|^{\frac{p}{2}}). \end{aligned} \tag{1}$$

Then $\mathcal{N} : (\lambda, u) \to g(\lambda, u)$ and $\tilde{\mathcal{N}} : (u, \lambda) \to g'(\lambda, u)$ are continuous operators from the space $\mathbb{R} \times L^p(\Omega)$ to the space $L^2(\Omega)$. For the derivatives of g with respect to u, a symbol $g'(\lambda, u)$ will be used. Moreover, we suppose that $g(\lambda, u)$ and its derivatives satisfy additional conditions:

$$\forall \lambda \in \mathbb{R} : \quad g(\lambda, 0) = g'(\lambda, 0) = 0.$$

[*]This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS13/217/OHK4/3T/14.

We consider a problem with unilateral interior conditions on the set Ω_U:

$$\triangle u + \lambda u + g(\lambda, u) = 0 \text{ on } \Omega\backslash\Omega_U, \quad u = 0 \text{ on } \partial\Omega, \tag{2}$$

$$u \geq 0, \quad \triangle u + \lambda u + g(\lambda, u) \leq 0 \text{ on } \Omega_U, \quad u \cdot (\triangle u + \lambda u + g(\lambda, u)) = 0 \text{ on } \Omega_U.$$

A weak formulation of this problem is a variational inequality:

$$u \in K, \quad \lambda \in \mathbb{R} : \quad \int_\Omega \nabla u \cdot \nabla(\varphi - u) - (\lambda u + g(\lambda, u))(\varphi - u) \geq 0, \quad \forall \varphi \in K, \tag{3}$$

where the set K is a convex cone:

$$K := \{\varphi \in W_0^{1,2}(\Omega) | \ \varphi \geq 0 \text{ a.e. on } \Omega_U\}.$$

In the following text, we will work with this weak formulation of the problem.

2 Idea of this paper

We will prove that the first eigenvalue of the Laplace operator on the set $\Omega\backslash\Omega_U$ with the homogenous Dirichlet boundary condition is a bifurcation point of the inequality (3). Using the Dancer theorem, we will find a sequence (λ_n, u_n) which satisfy

$$(\lambda_n, u_n) \in \mathbb{R}\times W_0^{1,2}(\Omega\backslash\Omega_U) : \int_{\Omega\backslash\Omega_U} \nabla u \cdot \nabla v - (\lambda_n u_n + g(\lambda_n, u_n))v \,\mathrm{dx} = 0 \ \forall v \in W_0^{1,2}(\Omega\backslash\Omega_U), \tag{4}$$

where

$$\lambda_n \to \lambda_0, \quad u_n \to u_0, \quad \frac{u_n}{\|u_n\|} \to -u_0,$$

and $-u_0$ is the eigenfunction corresponding to the eigenvalue λ_0. We will show that the functions u_n extended by zero on the set Ω_U are for sufficiently large n solutions of the inequality (3) and consequently, λ_0 is a bifurcation point of the inequality (3).

3 Bifurcation of the Laplace equation

3.1 Theorems and lemma used for the proof of the main theorem

We formulate the problem (4) as an operator equation on the space $W_0^{1,2}(\Omega\backslash\Omega_U)$ with the scalar product and norm defined as:

$$\langle u, v\rangle = \int_{\Omega\backslash\Omega_U} \nabla u \cdot \nabla v \,\mathrm{dx} \qquad \forall u, v \in W_0^{1,2}(\Omega\backslash\Omega_U),$$

$$\|u\| = \int_{\Omega\backslash\Omega_U} |\nabla u|^2 \,\mathrm{dx} \qquad \forall u \in W_0^{1,2}(\Omega\backslash\Omega_U).$$

We consider a nonlinear problem on the set $\Omega\backslash\Omega_U$.

$$\triangle u + \lambda u + g(\lambda, u) = 0 \quad \text{on } \Omega\backslash\Omega_U, \tag{5}$$

$$u = 0 \ \text{ on } \ \partial\Omega \cup \partial\Omega_U.$$

We will find a weak solution of this equation which is an element of the space $W_0^{1,2}(\Omega\backslash\Omega_U)$:

$$u \in W_0^{1,2}(\Omega\backslash\Omega_U): \quad \int_{\Omega\backslash\Omega_U} -\nabla u \cdot \nabla\varphi + \lambda u\varphi + g(\lambda, u)\varphi \,\mathrm{dx} = 0 \quad \forall\varphi \in W_0^{1,2}(\Omega\backslash\Omega_U). \tag{6}$$

When $u \in W_0^{1,2}(\Omega\backslash\Omega_U)$ is a solution of (6), then $\triangle u \in L^2(\Omega)$. It is because the equation (5) holds in the distributive sense. We rewrite this equation in the form $-\triangle u = \lambda u + g(\lambda, u)$. The function $\lambda u + g(\lambda, u) \in L^2(\Omega\backslash\Omega_U)$, which implies that $\triangle u \in L^2(\Omega\backslash\Omega_U)$.

We define two new operators A and N as follows:

$$A : W_0^{1,2}(\Omega\backslash\Omega_U) \to W_0^{1,2}(\Omega\backslash\Omega_U), \qquad N : \mathbb{R} \times W_0^{1,2}(\Omega\backslash\Omega_U) \to W_0^{1,2}(\Omega\backslash\Omega_U),$$

$$\langle Au, v\rangle = \int_{\Omega\backslash\Omega_U} u\cdot v \,\mathrm{dx}, \quad \langle N(\lambda, u), v\rangle = \int_{\Omega\backslash\Omega_U} g(\lambda, u)\cdot v \,\mathrm{dx} \quad \forall u, v \in W_0^{1,2}(\Omega\backslash\Omega_U). \tag{7}$$

Using these operators the equation (6) can be formulated as an operator equation:

$$u - \lambda Au - N(\lambda, u) = 0 \tag{8}$$

Definition 1. *Characteristic value of the operator $A : W_0^{1,2}(\Omega) \to W_0^{1,2}(\Omega)$ is a number $\lambda \in \mathbb{R}$ such that*

$$\exists u \in W_0^{1,2}(\Omega\backslash\Omega_U), \ u \neq 0 :$$

$$\lambda Au = u.$$

The function u is said to be an eigenfunction to the characteristic value λ.

Lemma 1. *The first characteristic value of the operator A is simple and positive.*

Proof. The equation for characteristic value can be written in an integral form:

$$\int_{\Omega\backslash\Omega_U} \lambda u \cdot v \,\mathrm{dx} = \int_{\Omega\backslash\Omega_U} \nabla u \cdot \nabla v \,\mathrm{dx} \quad \forall v \in W_0^{1,2}(\Omega\backslash\Omega_U).$$

This is weak formulation of the eigenvalue equation for the Laplace operator. The lowest number λ for which the solultion exists is positive and for this value the equation has a unique solution [3]. □

We denote the first characteristic value of the operator A as λ_0. The first eigenfunction correspondent to this characteristic value will be denoted as u_0 and is a solution of the equation:

$$\int_{\Omega\backslash\Omega_U} -\nabla u_0 \cdot \nabla\varphi + \lambda_0 u_0\varphi \,\mathrm{dx} = 0 \quad \forall\varphi \in W_0^{1,2}(\Omega\backslash\Omega_U).$$

This eigenfunction does not change sign for a.a. $x \in \Omega\backslash\Omega_U$ [3].

Definition 2. *The point $\hat{\lambda}$ is called a bifurcation point of the equation (8) if for any neighbourhood of $(\hat{\lambda}, 0) \in \mathbb{R} \times W_0^{1,2}(\Omega \backslash \Omega_U)$ there is a solution $(\lambda, u) \in \mathbb{R} \times W_0^{1,2}(\Omega \backslash \Omega_U)$ of (8) with $u \neq 0$.*
The point $\hat{\lambda}_1$ is called a bifurcation point of the inequality (3) if for any neighbourhood of $(\hat{\lambda}_1, 0) \in \mathbb{R} \times W_0^{1,2}(\Omega)$ there is a solution $(\lambda, u) \in \mathbb{R} \times W_0^{1,2}(\Omega)$ of (3) with $u \neq 0$.

In the following theorem we consider general operators A and N on a Hilbert space $\mathcal{H}$:

Theorem 1. *Let the operators $A : \mathcal{H} \to \mathcal{H}$ be a compact linear operator, $N : \mathbb{R} \times \mathcal{H} \to \mathcal{H}$ be a nonlinear compact operator, λ_0 be a simple characteristic value of the operator A, u_0 be the eigenfunction corresponding to the characteristic value λ_0. Moreover let for any bounded set $\mathcal{M} \subset \mathbb{R}$ the operator N satisfy a condition:*

$$\lim_{\|u\| \to 0} \frac{N(\lambda, u)}{\|u\|} = 0 \quad \textit{uniformly for all } \lambda \in \mathcal{M}.$$

Denote S the closure of all solutions of the equation (8) with $u \neq 0$, i.e.

$$S = \overline{\{(\lambda, u) \mid u \neq 0,\ u \textit{ is a solution of } (8)\}}.$$

Then $(\lambda_0, 0) \in S$, i.e. λ_0 is a bifurcation point of the equation (8). Denote C the component of S which contains $(\lambda_0, 0)$. Then C consists of two connected sets C^+, C^-, $C = C^+ \cup C^-$ such that

$$C^+ \cap C^- \cap B((\lambda_0, 0); \rho) = \{(\lambda_0, 0)\} \textit{ and } C^\pm \cap \partial B((\lambda_0, 0); \rho) \neq 0,$$

where $B((\lambda_0, 0); \rho)$ is a ball with sufficiently small radius ρ. The sets C^+ and C^- are either both unbounded or

$$C^+ \cap C^- \neq \{(\lambda_0, 0\}.$$

For more information see e.g. [1], Dancer theorem. During a proof of main theorem the Hopf lemma will be also used:

Lemma 2. *Let $\hat{\Omega}$ be a bounded domain in R^n and u is a function which satisfies $\triangle u \geq 0$ on the set $\hat{\Omega}$. Let $x_0 \in \partial\hat{\Omega}$ be such that*

- *u is continuous at x_0,*
- *$u(x_0) > u(x)$ for all $x \in \hat{\Omega}$,*
- *u satisfy the interior sphere condition, i.e there exists a ball $B \subset \hat{\Omega}$ with $x_0 \in \partial\Omega$.*

Then the outer normal derivative u at x_0, if it exists, satisfies the strict inequality

$$\frac{\partial u}{\partial \vec{n}}(x_0) > 0.$$

For proof see [2]. For the boundary $\partial\hat{\Omega}$ at least of a class C^2 the interior sphere condition is automatically fulfilled for all points $x_0 \in \partial\hat{\Omega}$.
To prove a local regularity of the solution of the equation (6) the regularity theorem will be used.

Theorem 2. *Let $\hat{\Omega} \subset R^2$ be an open subset and u is a solution of the equation $\triangle u = f$ on the set $\hat{\Omega}$ with the homogenous Dirichlet condition. Moreover, let $\partial\hat{\Omega} \in C^{k+1}$ and $f \in W^{k,2}(\hat{\Omega})$. Then $u \in W^{k+2,2}(\hat{\Omega})$ and*

$$\|u\|_{k+2,2} \leq C(\|u\|_{k,2} + \|f\|_{k,2}).$$

Proof can be found in e.g. [2].

3.2 Main theorem

Theorem 3. *The first eigenvalue $\lambda_0 \in \mathbb{R}$ of the Laplace operator on the domain $\Omega\backslash\Omega_U$ with homogenous Dirichlet boundary condition is a bifurcation point of the inequality (3). Let u_0 be an eigenfunction correspondent to the eigenvalue λ_0. There exists a sequence (λ_n, u_n) of solutions of (6) such that*

$$u_n \to 0, \quad \lambda_n \to \lambda_0, \quad \frac{u_n}{\|u_n\|} \to -u_0,$$

and for sufficiently large index n the functions u_n extended with zero on the set Ω_U are solutions of the inequality (3).

Proof. The operators A, N defined in (7) satisfy assumption of the Theorem 1, λ_0 is a simple characteristic value of the operator A. The meaning of the sets $C^{\pm}$ in the Dancer theorem 1 is the following. In the first set C^+ there are solutions of the equations (6) bifurcating in the direction $+u_0$, in the second set C^- bifurcating in the direction $-u_0$. Concretely, let us consider a two sequences $(\lambda_n,\ u_n^+) \in C^+$ resp. $(\lambda_n,\ u_n^-) \in C^-$ such that

$$\lim_{n\to\infty} u_n^+ = 0, \quad \lim_{n\to\infty} u_n^- = 0, \quad \lim_{n\to\infty} \lambda_n = \lambda_0.$$

Then the limits of the normalized functions $u_n^{\pm}/\|u_n\|$ are

$$\lim_{n\to\infty} \frac{u_n^+}{\|u_n\|} = +u_0, \quad \lim_{n\to\infty} \frac{u_n^-}{\|u_n\|} = -u_0.$$

We will work with the functions from the set C^- and denote them simply as u_n, i.e. $u_n := u_n^-$.

We define the prolongation $\tilde{u}_n$ for the functions u_n

$$\tilde{u}_n = \begin{cases} u_n & \text{if } x \in \Omega\backslash\overline{\Omega_U} \\ 0 & \text{if } x \in \Omega_U. \end{cases}$$

The functions $\tilde{u}_n$ are elements of the Sobolev space $W_0^{1,2}(\Omega)$. Moreover, because $\tilde{u}_n = 0$ on Ω_U for all $n \in \mathbb{N}$, it holds $\tilde{u}_n \in K$ for all $n \in \mathbb{N}$. We will prove that the functions $\tilde{u}_n$ are solutions of the inequality (3), i.e. that it holds

$$\tilde{u} \in K : \quad \int_\Omega \nabla\tilde{u}_n \cdot \nabla(\varphi - \tilde{u}_n) - (\lambda_n\tilde{u}_n + g(\lambda_n, \tilde{u}_n))(\varphi - \tilde{u}_n) \geq 0 \quad \forall\varphi \in K.$$

We divide the integral into two parts - integration over the set $\Omega\backslash\Omega_U$ and integration over the set Ω_U. The functions $\tilde{u}_n$ have a support on the set $\Omega\backslash\Omega_U$ therefore the integral over the set Ω_U is zero and we can write:

$$\tilde{u}_n \in K : \int_\Omega \nabla\tilde{u}_n \cdot \nabla(\varphi - \tilde{u}_n) - (\lambda_n \tilde{u}_n + g(\lambda_n, \tilde{u}_n))(\varphi - \tilde{u}_n) =$$

$$= \int_{\Omega\backslash\Omega_U} \nabla\tilde{u}_n \cdot \nabla(\varphi - \tilde{u}_n) - (\lambda_n \tilde{u}_n + g(\lambda_n, \tilde{u}_n))(\varphi - \tilde{u}_n)$$

$$\forall \varphi \in K.$$

Because $\triangle u \in L^2(\Omega\backslash\Omega_U)$ we can use Green theorem and write the integral as

$$\int_{\Omega\backslash\Omega_U} -(\triangle\tilde{u}_n + \lambda_n\tilde{u}_n + g(\lambda_n, \tilde{u}_n))(\varphi - \tilde{u}_n) + \int_{\partial\Omega\cup\partial\Omega_U} \frac{\partial\tilde{u}_n}{\partial\vec{n}}(\varphi - \tilde{u}_n)\ \mathrm{dS}.$$

The functions $\tilde{u}_n$ solve the equations $\triangle\tilde{u}_n + \lambda_n\tilde{u}_n + g(\lambda_n, \tilde{u}_n) = 0$ almost everywhere on the set $\Omega\backslash\Omega_U$, hence, the first integral is equal to zero. Furthermore, the functions $\tilde{u}_n$ and φ satisfy:

$$\tilde{u}_n = 0 \text{ on } \partial\Omega \cup \partial\Omega_U \quad \wedge \quad \varphi = 0 \quad \text{on} \quad \partial\Omega \quad \wedge\, \varphi \geq 0 \quad \text{on} \quad \partial\Omega_U.$$

The integral over the boundary then reduces to the expression:

$$\int_{\partial\Omega\cup\partial\Omega_U} \frac{\partial\tilde{u}_n}{\partial\vec{n}}(\varphi - \tilde{u}_n)\ \mathrm{dS} = \int_{\partial\Omega_U} \frac{\partial\tilde{u}_n}{\partial\vec{n}}\varphi\ \mathrm{dS}.$$

We will prove later that the functions $\chi\tilde{u}_n$ are elements of the space $W^{3,2}(\Omega')$, where Ω' is a subset of $\Omega\backslash\Omega_U$:

$$\Omega' \subset\subset (\Omega\backslash\Omega_U) \quad \wedge \quad \partial\Omega' \cap \Omega_U = \Omega_U,$$

and has a boundary of the class C^2 and χ is a cut-off function on the set Ω'. We will also prove later that

$$\left\| \frac{u_n}{\| u_n\|} + u_0 \right\|_{3,2,\Omega'} \to 0. \tag{9}$$

Then $\frac{\partial u_n}{\partial\vec{n}}$ has a meaning in the classical sense for all $x \in \Omega'$ and converges uniformly on the set Ω' to the normal derivative of the function $-u_0$. To determine the sign of the normal derivative $\frac{\partial\tilde{u}_0}{\partial\vec{n}}$ on the set $\partial\Omega_U$ we will use the Hopf lemma 2. Fulfilment of the first assumption of this lemma will be proved later. We will prove now the fulfilment of the second assumption. Because $-u_0 = 0$, and for all $x \in \Omega'$ is $-u_0 < 0$ the second assumption is fulfilled. Furthermore,

$$\triangle(-u_0) = -\lambda_0(-u_0) \geq 0.$$

The interior ball condition is fulfilled automatically, because $\partial\Omega_U \in C^2$. Now we can use the Hopf lemma to conclude that

$$\exists\varepsilon > 0 : \frac{\partial(-u_0)}{\partial\vec{n}}(x) > \varepsilon \quad \forall x \in \partial\Omega_U. \tag{10}$$

Because the normal derivative

$$\frac{1}{\|\tilde{u}_n\|}\frac{\partial \tilde{u}_n}{\partial \vec{n}} \to \frac{(-\partial u_0)}{\partial \vec{n}},$$

uniformly on the set Ω_U, for sufficiently large index n it holds

$$\frac{\partial \tilde{u}_n}{\partial \vec{n}}(x) > 0, \quad \forall x \in \partial\Omega_U \quad \Rightarrow \quad \int_{\partial\Omega_U} \frac{\partial \tilde{u}_n}{\partial \vec{n}} \varphi \text{ dS} \geq 0. \tag{11}$$

Thus the functions $\tilde{u}_n$ are solutions of the variational inequality (3) and λ_0 is a bifurcation point of the inequality (3).

To finish the proof, it remains to prove (9). For an arbitrary Lipschitz set $\tilde{\Omega} \subset \mathbb{R}^2$ the space

$$W^{1,2}(\tilde{\Omega}) \hookrightarrow L^p(\tilde{\Omega}), \ \forall p \in (1, \infty).$$

Let us remind that when $g(\lambda, u)$ satisfy first growth condition in (1) then $\mathcal{N} : u \to g(\lambda, u)$ is a continuous operator from the space $L^q(\Omega)$ to $L^2(\Omega)$.
We consider a set Ω'' such that

$$\Omega'' \subset\subset \Omega' \subset\subset (\Omega \backslash \Omega_U) \ \wedge \ \partial\Omega'' \cap \partial\Omega_U = \partial\Omega_U.$$

Moreover, let $\partial\Omega'' \in C^2$. We define a smooth function χ which has the property

$$(\forall x \in \Omega'')(\chi(x) := 1) \wedge (\forall x \in (\Omega\backslash\Omega'))(\chi := 0).$$

The function χ does not depend on the index n. We consider on the set Ω' the functions

$$(u_n\chi) : \Omega' \to \mathbb{R} : \quad u_n\chi(x) := u_n(x)\chi(x) \ \text{ for a.a.} x \in \Omega'.$$

which are solutions of the equation

$$\triangle(\chi u_n) = u_n \triangle\chi + \triangle u_n \chi + 2\nabla u_n \cdot \nabla\chi u_n = \triangle\chi - \lambda_n u_n \chi - g(\lambda_n, u_n)\chi + 2\nabla u_n \cdot \nabla\chi, \tag{12}$$

$$u_n\chi = 0 \text{ on } \partial\Omega_U \cup \partial\Omega'.$$

The r.h.s. of this equation is an element of the space $L^2(\Omega\backslash\Omega_U)$. For further purposes, let us denote

$$\hat{f}_n(\hat{u}_n, \lambda) := u_n\triangle\chi - g(\lambda_n, u_n)\chi + 2\nabla u_n \cdot \nabla\chi.$$

On the set Ω'' it holds $g(\lambda_n, u_n) = \hat{f}(\lambda_n, \hat{u}_n)$.
The functions $\hat{u}_n = u_n\chi$ are elements of the set $W_0^{1,2}(\Omega\backslash\Omega_U)$ and are defined on the set with the boundary of a class C^2. Theorem 2 about regularity can be applied to the equation (12) to obtain

$$\triangle\hat{u}_n = \hat{f}_n \in L^2(\Omega') \Rightarrow \ \triangle\hat{u}_n \in L^2(\Omega') \Rightarrow \hat{u}_n \in W^{2,2}(\Omega').$$

Because $W^{2,2}(\Omega') \hookrightarrow C^{0,\alpha}(\overline{\Omega'})$ for a suitable $\alpha \in (0, 1)$, the functions $\hat{u}_n$ are on the set Ω' Hölder continous and bounded.

Now we suppose that $g'(\lambda, u)$ satisfies the growth conditions (1). Then derivative of function $g(\lambda, u(x))$ with respect to spatial variables x is an element of the space $L^2(\Omega')$

$$\begin{aligned}
&\forall u \in W^{2,2}(\Omega') : \\
&\int_{\Omega'} \left| \frac{\partial}{\partial x_i} g(\lambda, u) \right|^2 \mathrm{dx} = \int_{\Omega'} \left| \frac{\partial g}{\partial u}(\lambda, u) \right|^2 \left| \frac{\partial u}{\partial x_i}(x) \right|^2 \leq \int_{\Omega'} \left| (1 + |u|^{r+1}) \frac{\partial u}{\partial x_i} \right|^2 \mathrm{dx} = \\
&= \int_{\Omega'} \left| (1 + |u|^{r+1})^2 \left| \frac{\partial u}{\partial x_i} \right|^2 \right| \mathrm{dx} \leq C \int_{\Omega'} \left| \frac{\partial u}{\partial x_i} \right|^2 \mathrm{dx} \leq C \|u\|_{1,2,\Omega'} < \infty \ \Rightarrow \\
&\Rightarrow (\forall i \in \{1, 2\}) \left(\frac{\partial}{\partial x_i}(g(\lambda, u(x))) \in L^2(\Omega') \right) \Rightarrow \ g(\lambda, u(x)) \in W^{1,2}(\Omega'). \qquad (13)
\end{aligned}$$

This result can be used to prove higher regularity of solution. Because $\hat{u}_n \in W^{2,2}(\Omega')$ and $g(\lambda_n, \hat{u}_n) \in W^{1,2}(\Omega')$, it holds that

$$\hat{f}_n = (u_n \triangle \chi - \lambda_n u_n \chi - g(\lambda_n, u_n)\chi + 2\nabla u_n \cdot \nabla \chi) \ \in W^{1,2}(\Omega').$$

As χ has a support on the set Ω', the functions $\hat{f}_n$ have supports also on the set Ω'. The function $\hat{u}_n$ is a solution of the equation $\triangle \hat{u}_n = \hat{f}_n$, and from Theorem (2) it results that $\hat{u}_n \in W^{3,2}(\Omega')$. Because $W^{3,2}(\Omega') \hookrightarrow C^{1,\alpha}(\overline{\Omega'})$ the normal derivative is on the set Ω' well defined in the classical sense. Similarly it can be proved that $\hat{u}_0 \in W^{3,2}(\Omega')$.
The last step is to prove that the functions $\hat{u}_n / \|\hat{u}_n\|$ converge to the function u_0 in the norm of the space $W^{3,2}(\Omega')$. We can use the estimate from Theorem 2

$$\|u_n\|_{W^{3,2}(\Omega')} \leq C \left(\|u_n\|_{W^{1,2}(\Omega')} + \|f\|_{W^{1,2}(\Omega')} \right).$$

The function u_0 must satisfy the equation

$$\triangle(u_0 \chi) = \lambda_0 u_0 + \triangle \chi u_0 + \nabla u_0 \cdot \nabla \chi, \quad u_0 = 0 \text{ on } \partial\Omega'. \qquad (14)$$

Similarly the functions $\hat{u}_n$ are solutions of the equations

$$\triangle(u_n \chi) = \lambda_n u_n + g(\lambda_n, u_n) + \triangle \chi u_n + \nabla u_n \cdot \nabla \chi, \quad u_n = 0 \text{ on } \partial\Omega'. \qquad (15)$$

Our goal is to find that $\hat{u}_n / \|\hat{u}_n\| \to u_0$ in the norm $W^{3,2}(\Omega')$. We will divide the equation (15) by $\|\hat{u}_n\|$ and add to the equation (14) to obtain

$$\begin{aligned}
\triangle \left[\left(\frac{u_n}{\|u_n\|} + u_0 \right) \chi \right] - (\triangle \chi) \left(\frac{u_n}{\|u_n\|} + u_0 \right) - \lambda_n \left(\frac{u_n}{\|u_n\|} + u_0 \right) \chi + (\lambda_n - \lambda_0) u_0 \chi - \\
- \frac{g_n(\lambda_n, u_n)}{\|u_n\|} \chi - \nabla \left(\frac{u_n}{\|u_n\|} + u_0 \right) \cdot \nabla \chi = 0, \quad (16)
\end{aligned}$$

$$\hat{u}_n = 0 \text{ on } \partial\Omega'.$$

The equation can be written in the form

$$\triangle \left[\left(\frac{u_n}{\|u_n\|} + u_0 \right) \chi \right] = f_n, \quad u_n = 0 \text{ on } \partial\Omega_U \cup \partial\Omega',$$

$$f_n := \lambda_n \left(\frac{u_n}{\|u_n\|} + u_0\right)\chi - (\lambda_n - \lambda_0)u_0\chi + \frac{g_n(\lambda_n, u_n)}{\|u_n\|}\chi + \nabla\left(\frac{u_n}{\|u_n\|} + u_0\right)\cdot\nabla\chi.$$

To prove the convergence, we will use the estimate from Theorem (2)

$$\left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\chi\right\|_{3,2,\Omega'} \le C\left(\left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\chi\right\|_{1,2,\Omega'} + \|f_n\|_{1,2,\Omega'}\right).$$

The first term on the r.h.s. is simple, because the sequence $u_n/\|u_n\|$ converges in $W_0^{1,2}(\Omega)$ norm to the function $-u_0$. The function χ is smooth and can be estimated by a constant (we will omit the subscript 1,2,Ω')

$$\left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\chi\right\| \le C_1\left\|\frac{u_n}{\|u_n\|} + u_0\right\| \to 0.$$

The proof that $\|f_n\| \to 0$ is more complicated

$$\left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\triangle\chi - \lambda_n\left(\frac{u_n}{\|u_n\|} + u_0\right)\chi + (\lambda_n - \lambda_0)u_0 - \frac{g_n(\lambda_n, u_n)}{\|u_n\|} - \right.$$
$$\left. -\nabla\left(\frac{u_n}{\|u_n\|} + u_0\right)\cdot\nabla\chi\|\right\| \le$$
$$\le \left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\triangle\chi\right\| + \lambda_n\left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\chi\right\| + |\lambda_n - \lambda_0|\|u_0\chi\| -$$
$$-\frac{\|g_n(\lambda_n, u_n)\|}{\|u_n\|} + \left\|\nabla\left(\frac{u_n}{\|u_n\|} + u_0\right)\cdot\nabla\chi\right\| \le$$
$$\le C_2\left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\chi\right\| + \lambda_n C_3\left\|\left(\frac{u_n}{\|u_n\|} + u_0\right)\chi\right\| + C_4|\lambda_n - \lambda_0|\|u_0\| -$$
$$-C_5\frac{\|g_n(\lambda_n, u_n)\|}{\|u_n\|} + C_6\left\|\nabla\left(\frac{u_n}{\|u_n\|} + u_0\right)\right\|.$$

First three term converge to zero. For the fourth term we have to prove that nonlinear term containing $g(\lambda, u)$ converges to zero. We only know, that

$$\|u_n\| \to 0 \quad \Rightarrow \quad \int_{\Omega\setminus\Omega_U} \frac{g(\lambda_n, u_n)}{\|u_n\|} v \text{ dx} \to 0, \quad \forall v \in W_0^{1,2}(\Omega).$$

We want to prove that also derivatives of $g(\lambda, u(x))$ with respect to x converges to zero.

$$\int_{\Omega'} \frac{\partial}{\partial x_i}\left(\frac{g(\lambda_n, u_n)}{\|u_n\|}\right) v\text{dx} = \int_{\Omega'} \frac{\partial}{\partial u}\left(\frac{g(\lambda_n, u_n)}{\|u_n\|}\right)\frac{\partial u}{\partial x_i} v\text{dx} = \int_{\Omega'} \frac{\frac{\partial u}{\partial x_i}}{\|u_n\|}\frac{\partial}{\partial u}(g(\lambda, u_n))v \text{ dx}.$$

The L^2 norm of a fraction

$$\left\|\frac{\partial u}{\partial x_i}\frac{1}{\|u_n\|}\right\|_2 \le \frac{\|\nabla u_n\|_2}{\|u_n\|} \le C,$$

is bounded. We can estimate

$$\int_{\Omega'} \frac{\frac{\partial u}{\partial x_i}}{\|u_n\|} \frac{\partial}{\partial u}(g(\lambda_n, u_n))v \text{ dx} \leq \left(\operatorname*{ess\,sup}_{x\in\Omega'} g'(\lambda_n, u_n(x)) \right) \int_{\Omega'} \frac{\|\frac{\partial u}{\partial x_i}\|_2}{\|u_n\|} \|v\|_2 \text{ dx} \leq$$
$$\leq C \left(\operatorname*{ess\,sup}_{x\in\Omega'} g'(\lambda_n, u_n(x)) \right) \|v\|_{1,2} \to 0,$$

where we used the condition $(\forall \lambda \in \mathbb{R})\, (g'(\lambda, 0) = 0)$ and Hölder and Poincare inequality. The convergence to zero of the last term in (16) can be proved as follows

$$\left\| \nabla \left(\frac{u_n}{\|u_n\|} + u_0 \right) \right\| \leq \lambda_n \left\| \left(\frac{u_n}{\|u_n\|} + u_0 \right) \chi \right\| + |\lambda_n - \lambda_0| \|u_0\| + \frac{\|g(\lambda_n, \hat{u}_n)\|}{\|u_n\|} \to 0.$$

The Poincare inequality was used again. The conclusion is that the normalized solutions on the set Ω' are convergent in the $W^{3,2}(\Omega')$ norm and their limit is the function $-u_0$. The convergence in the Hölder norm follows from the Morrey's inequality

$$\left\| \frac{u_n}{\|u_n\|} + u_0 \right\|_{C^{1,\alpha}(\Omega')} \leq \left\| \frac{u_n}{\|u_n\|} + u_0 \right\|_{3,2,\Omega'} \to 0.$$

The functions u_n converge in the $C^{1,\alpha}(\overline{\Omega'})$ norm for a suitable α to the function $-u_0$, hence, (9) is proved which finishes the proof. □

References

[1] P. Drabek, J. Milota: *Methods of Nonlinear Analysis: Applications to Differential Equations*, *Springer, 2013*, ISBN: 978-3034803861

[2] D. Gilbarg, N.S. Trudinger: *Elliptic Partial Differential Equations of Second Order*, *Springer, 2001*, ISBN:3-540-08007-4, pg.177

[3] D.S. Grebenkov, Bihn-Thanh Nguyen: *Geometrical structure of Laplacian eigenfunctions*, available online: http://arxiv.org/abs/1206.1278

On the Pseudospectrum of the Harmonic Oscillator with Imaginary Cubic Potential

Radek Novák*

2nd year of PGS, email: novakra9@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: David Krejčiřík, Department of Theoretical Physics
Nuclear Physics Institute, AS CR

Abstract. We study the Schrödinger operator with a harmonic oscillator and imaginary cubic oscillator potential and focus on its pseudospectral properties. The summary of known results about the operator and its spectrum is provided and we emphasize the importance of examining its pseudospectrum as well. This is achieved with employing scaling techniques and treating the operator using semiclassical methods. The existence of pseudomodes very far from the spectrum is proven and as a consequence, the spectrum of the operator is unstable with respect to small perturbations. It is shown that its eigenfunction form a complete set in the Hilbert space, however, they do not form a Riesz basis.

Keywords: pseudospectrum, harmonic oscillator, imaginary qubic potential, $\mathcal{PT}$-symmetry, semiclassical method

Abstrakt. Studujeme Schrödingerův operátor s potenciálem harmonického oscilátoru a imaginárního kubického oscilátoru a zaměřujeme se na jeho pseudospektrální vlastnosti. Shrnujeme známé výsledky o tomto operátoru a jeho spektru a zdůrazňujeme důležitost studia i jeho pseudospektra. Toho je dosaženo aplikací škálovacích technik a zkoumáním operátoru využitím semiklasických metod. Je dokázána existence pseudomódů velmi daleko od spektra a jako důsledek je spektrum operátoru nestabilní vůči malým poruchám. Ukazujeme, že jeho vlastní funkce tvoří úplnou množinu v Hilbertově prostoru, avšak netvoří Rieszovu bázi.

Klíčová slova: pseudospektrum, harmonický oscilátor, imaginární kubický potenciál, $\mathcal{PT}$-symetrie, semiklasická metoda

*This contribution is based on the forthcoming submission to the International Journal of Theoretical Physics

Determination of Stop-Criterion for Incremental Methods Constructing Camera Sensor Fingerprint[*]

Adam Novozámský[†]

4th year of PGS, email: novozamsky@utia.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Stanislav Saic, Department of Image Processing
Institute of Information Theory and Automation, AS CR

Abstract. This paper aims to find the minimum sample size of the camera reference image set that is needed to build a sensor fingerprint of a high performance. Today's methods for building sensor fingerprints do rely on having a sufficient number of camera reference images. But, there is no clear answer to the question of how many camera reference images are really needed? In this paper, we will analyze and find out how to determine the minimum needed number of reference images to remove the mentioned uncertainty. We will introduce a quantitative measure (a stop-criterion) stating how many photos should be used to create a high-performance sensor fingerprint. This stop-criterion will directly reflect the confidence level that we would like to achieve. By considering that the number of digital images used to construct the camera sensor fingerprint can have a direct impact on performance of the sensor fingerprint, it is apparent that this, so far underestimated, topic is of major importance.

Keywords: image ballistics, source camera verification, pattern noise, PRNU, fingerprint performance, laplace distribution

Abstrakt. Tento článek si klade za cíl nalézt minimální velikost množiny referenčních fotografií, která je potřeba k vybudování silného otisku fotoaparátu. Dnešní metody se spoléhají na dostatečný počet referenčních snímků kamery, ale nepřináší jasnou odpověď na otázku, kolik snímků je ve skutečnosti potřeba. V tomto článku budeme analyzovat a zjišťovat, jak určit minimální potřebný počet referenčních snímků a tím odpovědět na tuto otázku. Zavedeme kvantitativní opatření (stop-kritérium) určující, kolik fotografií by mělo být použito k vytvoření silného otisku snímače. Toto stop-kritérium bude přímo odrážet úroveň spolehlivosti, které chceme dosáhnout. Toto téma je velmi důležité, ačkoli je v literatuře velmi podceňováno. Poněvadž to z kolika snímků je otisk senzoru vytvořen má vliv na úspěšnost jeho detekce na testovaných obrázcích.

Klíčová slova: obrazková balistika, ověření zdrojové kamery, šum senzoru, PRNU, úspěšnost otisku, Laplaceova distribuce

[*] This work has been supported by the grant: *Analýza pravosti digitálního videa* GAČR 13-28462S

[†] This work is published in Proceedings of the IWDW 2014:Babak Mahdian, Novozámský Adam and Stanislav Saic. *Determination of stop-criterion for incremental methods constructing camera sensor fingerprint.* In: 13th International Workshop on Digital-Forensics and Watermarking. Taipei, Taiwan , October 1-4, 2014. Paper 81.

References

[1] H. T. Sencar, M. Ramkumar, and A. N. Akansu, *Data Hiding Fundamentals and Applications: Content Security in Digital Multimedia.* Orlando, FL, USA: Academic Press, Inc., 2004.

[2] M. Arnold, M. Schmucker, and S. D. Wolthusen, *Techniques and Applications of Digital Watermarking and Content Protection.* Norwood, MA, USA: Artech House, Inc., 2003.

[3] N. Nikolaidis and I. Pitas, "Robust image watermarking in the spatial domain," *Signal Processing*, vol. 66, no. 3, pp. 385–403, May 1998.

[4] B. Mahdian and S. Saic, "A bibliography on blind methods for identifying image forgery," *Image Commun.*, vol. 25, no. 6, pp. 389–399, 2010.

[5] T.-T. Ng and M.-P. Tsui, "Camera response function signature for digital forensics - part i: Theory and data selection," in *IEEE Workshop on Information Forensics and Security*, Dec. 2009, pp. 156–160.

[6] Z. Lint, R. Wang, X. Tang, and H.-Y. Shum, "Detecting doctored images using camera response normality and consistency," in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1.* Washington, DC, USA: IEEE Computer Society, 2005, pp. 1087–1092.

[7] A. Popescu and H. Farid, "Exposing digital forgeries in color filter array interpolated images," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3948–3959, 2005. [Online]. Available: www.cs.dartmouth.edu/farid/publications/sp05a.html

[8] B. Mahdian and S. Saic, "Blind authentication using periodic properties of interpolation," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 529–538, September 2008.

[9] B. Mahdian and S. Saic, "Detection of copy–move forgery using a method based on blur moment invariants," *Forensic science international*, vol. 171, no. 2–3, pp. 180–189, 2007.

[10] A. E. Dirik, S. Bayram, H. T. Sencar, and N. Memon, "New features to identify computer generated images," in *IEEE International Conference on Image Processing, ICIP '07*, vol. 4, 2007, pp. 433 – 436.

[11] J. Fridrich and T. Pevny, "Detection of double–compression for applications in steganography," *IEEE Transactions on Information Security and Forensics*, vol. 3, no. 2, pp. 247–258, June 2008.

[12] M. Chen, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, March 2008.

[13] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, June 2006.

[14] I. Amerini, R. Caldelli, V. Cappellini, F. Picchioni, and A. Piva, "Analysis of denoising filters for photo response non uniformity noise extraction in source camera identification," in *Proceedings of the 16th international conference on Digital Signal Processing*, ser. DSP'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 511–517. [Online]. Available: http://dl.acm.org/citation.cfm?id=1700307.1700392

[15] E. J. Alles, Z. J. M. H. Geradts, and C. J. Veenman, "Source camera identification for heavily jpeg compressed low resolution still images," *Journal of Forensic Sciences*, vol. 54, no. 3, pp. 628–638, 2009. [Online]. Available: http://www.science.uva.nl/research/publications/2009/AllesJFS2009

[16] C. J. Yongjian Hu, Binghua Yu, "Source camera identification using large components of sensor pattern noise," in *Computer Science and its Applications, 2009. CSA '09. 2nd International Conference on*, Jeju Island, Korea, 2009.

[17] Y. Li and C.-T. Li, "Decomposed photo response non-uniformity for digital forensic analysis," in *e-Forensics*, 2009, pp. 166–172.

[18] Y. Hu, C. Jian, and C.-T. Li, "Using improved imaging sensor pattern noise for source camera identification," in *ICME*, 2010, pp. 1481–1486.

[19] J. Lukas, J. Fridrich, and M. Goljan, "Detecting digital image forgeries using sensor pattern noise," in *In Proceedings of the SPIE.* West, 2006, p. 2006.

[20] M. Chen, J. Fridrich, M. Goljan, and J. Lukáš, "Source digital camcorder identification using sensor photo-response nonuniformity," in *Proc. of SPIE Electronic Imaging, Photonics West*, 2007.

[21] M. Chen, J. Fridrich, and M. Goljan, "Digital imaging sensor identification (further study," in *In Security, Steganography, and Watermarking of Multimedia Contents IX. Edited by Delp, Edward J., III; Wong, Ping Wah. Proceedings of the SPIE, Volume 6505*, 2007.

[22] D. Williams, V. Codreanu, P. Yang, B. Liu, F. Dong, B. Yasar, B. Mahdian, A. Chiarini, X. Zhao, and J. Roerdink, "Evaluation of autoparallelization toolkits for commodity graphics hardware," in *10th International Conference on Parallel Processing and Applied Mathematics.* Warsaw, Poland: Springer, 2013, to appear.

[23] D. Williams, V. Codreanu, J. B. Roerdink, P. Yang, B. Liu, F. Dong, and A. Chiarini, "Accelerating colonic polyp detection using commodity graphics hardware," in *Proceedings of the International Conference on Computer Medical Applications*, Sousse, Tunisia, 2013, pp. 1–6.

Finalization of New Data Acquisition System for COMPASS Experiment*

Josef Nový

2nd year of PGS, email: josef.novy@cern.ch
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Tomáš Liška, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This paper discusses finalization of the new data acquisition system (DAQ) of the COMPASS experiment at CERN and mainly focuses on description of development process and interaction with users. The new DAQ is developed to replace old DAQ written originally for the ALICE experiment and works together with upgrade of readout hardware. It uses extensively possibilities of state of the art field programmable gate arrays (FPGA) technology. The new DAQ software is based on state machines and C++ with usage of the QT framework, the DIM library, and the IPBus library. System is presently in its final stage of preparation.

Keywords: data acquisition, Qt, GUI, FPGA

Abstrakt. Tento článek se věnuje dokončovacím pracem na systému pro sběr dat experimentu COMPASS v CERN a zaměřuje se hlavně na popis interakce s uživatelem. Nový systém je vyvíjen s cílem nahradit starý systém původně vytvořený pro experiment ALICE a pracovat společně s vylepšením vyčítávacího hardwaru. Rozsáhle využívá možnosti nejmodernějších FPGA technologií. Nový systém sběru dat je postaven na stavových automatech a jazyce C++ s použitím Qt frameworku, knihoven DIM a IPBus. Systém je momentálně ve finálním stádiu příprav.

Klíčová slova: sběr dat, Qt, GUI, FPGA

1 Introduction

This paper presents structure and user interfaces of the new data acquisition software package designed to work together with new upgraded structure of readout chain of the COMPASS experiment at CERN, but focuses mostly on its interaction with user, development process as whole and last steps of development in particular. For more information about design please see [6, 8, 7]. COMPASS [10] is a fixed target experiment at CERN which in previous years had a usual data rate of approximately 1500 MB/s during approximately 10 second on-spill with 45 second off-spill. Its present DAQ system was built during years 1999-2001. The Data Acquisition and Test Environment(DATE) [2], originally developed for the ALICE at CERN, was used to control DAQ and event building in old system and its graphical user interface was used as base for designing of the new one.

[*] This work has been supported by grants LA08015 and SGS11/167/OHK4/3T/14

Development of the new DAQ software and hardware was started to improve reliability and speed of system. Both parts has been developed in parallel, but full cooperation and regular communication was kept as they are closely dependent on each other. Main idea of hardware upgrade is to use FPGA technology for event building purposes and thus reducing number of used computers to just eight. Hardware event building was previously tried out by the CDF experiment [3] at Fermilab and the NA48 experiment [12, 13] at CERN. Both these experiment went back to software event-building due to problems either with reliability or flexibility,but now we have possibility to make reliable, flexible and cost-effective hardware event-building thanks two improvements in FPGA technology. The new software have to cope with challenges linked to control of such new hardware event-building network and allow user to operate whole system efficiently.

2 Used technologies

DAQ software package of big experiments can be fairy complex system and as such uses many different kind of technologies. Used technologies are, in this paper, divided to three groups for sake of lucidity. Hardware technologies are in the first group. The second group is composed from programing languages and frameworks. The last group contains communication libraries.

2.1 Hardware technologies

The FPGA technology is key feature of the new DAQ. FPGA chips are special integrated circuits whose behavior can be changed in the field by uploading a new firmware. These chips are usually equipped with many high speed serial links, they are cost effective, and reliable in these days. This, together with high speed DDR3 memory, optical fiber transceivers, and fast Ethernet for control purposes, has made possible to create DAQ module for event-building.

2.2 Used programing languages and frameworks

DAQ system have to address many different aspects, thus there are many different languages used. C++ was chosen as language for core processes as they need to be fast and have good control of used resources. It is supported by MySQL for database access and Python with bash scripts for minor tasks. PHP, HMTL, javascript, and AJAX have been used for creation of web-based configuration interface. The Qt framework, a cross-platform application framework, has been used for all main graphical user interfaces (GUIs) and to speed up development of core applications. Some support GUIs, written in Tool Command Language (TCL), were taken from old DAQ and reused in the new one.

2.3 Used communication libraries

There are two kinds of communication in new DAQ and for each different libraries are needed. The first one is communication between processes. The Distributed Management

System (DIM) library has been used for this purpose. Request to use this library came out from initial studies as it is widely used in the COMPASS experiment. The DIM is a multi-platform library that serves for an asynchronous 1 to many communication through the Ethernet. It was originally developed for the DEPHI experiment at CERN. The second type is communication with FPGA DAQ modules. The IPBus package is used for this communication. It was developed for the level one trigger update of the CMS experiment. This package consists of firmware part and software part. The firmware part mediates access to registers and memory of a FPGA card through Ethernet when properly loaded. The software part is implemented in C++ and contains all classes needed for a connection to the interface of the firmware part.

3 Design of the new DAQ

The new DAQ system, as is shown in figure 1, can be divided to three main sections:

- detectors, frontends and preprocessing modules,
- main DAQ hardware event-building network,
- readout computers, DAQ software and data storage.

3.1 Detectors, frontends and preprocessing modules

Detector setup of the COMPASS compose from different kinds of tracking detectors, calorimeters and detectors for particle identification. These detectors have around 300 000 channels which are read out by various frontend cards. Frontend cards concentrate these channels to approximately 1000 links which are connected to CATCH, HGeSiCa or GANDALF data-concentrator module. Part of modules are then connected directly to next stage and part are connected to Slink multiplexer or TIGER VXS modules for future data concentration. This part of system is in the same form as it was in previous DAQ.

3.2 Main DAQ hardware event-building network

This part was the most challenging one from hardware point of view. Software event-building network running on 50 computers has been exchanged for hardware event-building on eight new DAQ FPGA modules with multiplexer firmware and one module with switch firmware. New DAQ module is based on VIRTEX6 XC6VLX130T FPGA middle size chip. It is equiped with 4 GB of DD3 memory, 1 Gb ethernet, and 16 serial links.

3.3 Readout computers, DAQ software and data storage

The last part incorporate eight servers with special PCI-e card called spillbuffer to which the optical fiber from the switch is connected. Event data are temporary stored in buffer of spillbuffer. Then they are read out by slave readout process and stored on RAID10 array

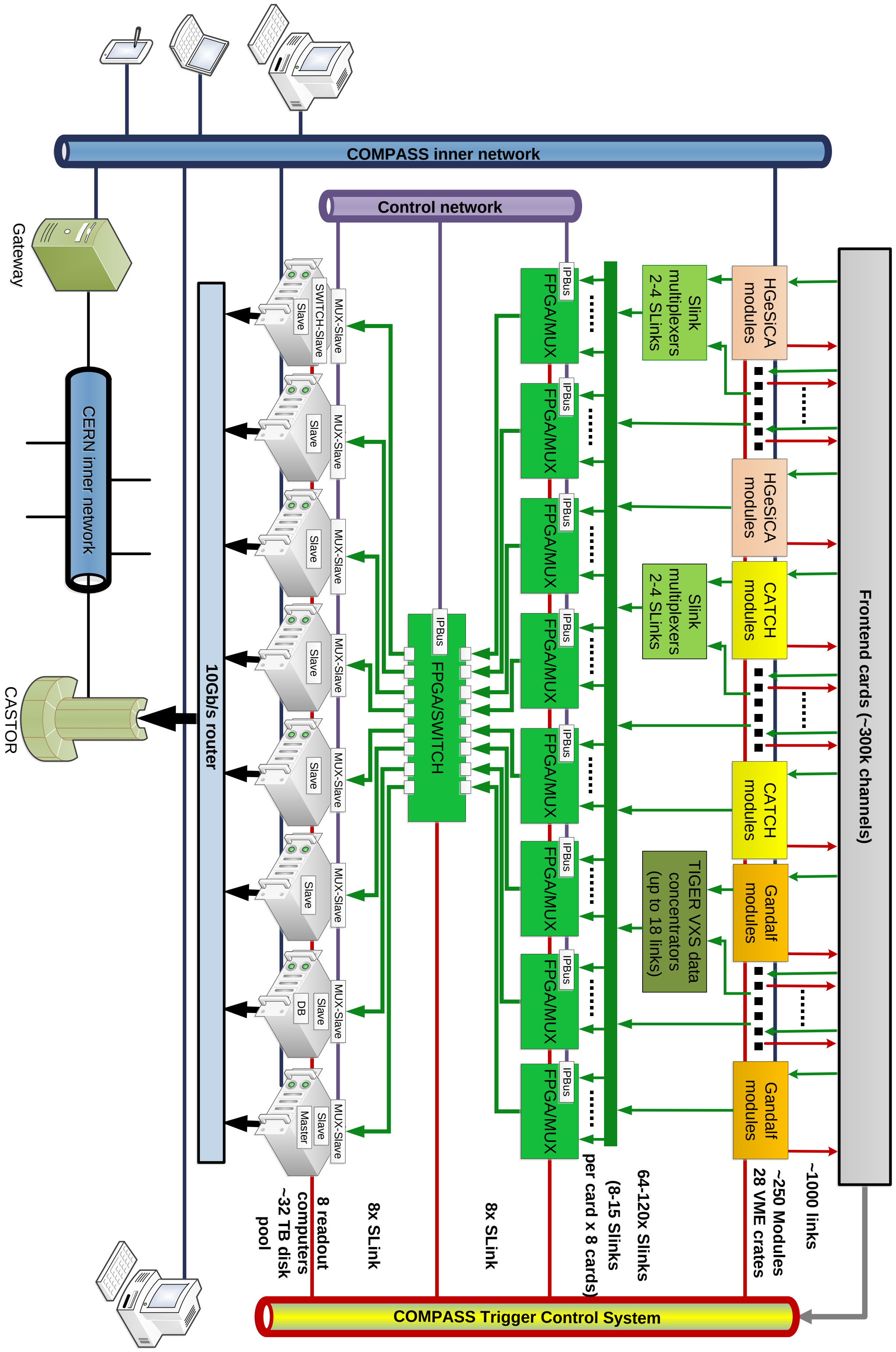


Figure 1: The new DAQ architecture

of 8 harddisks, before transferee to The CERN Advanced STORage manager (CASTOR) for long term storage. All main DAQ processes are running on these computers.

DAQ software is divided to six processes: Master, Slave-readout, Slave-control, Control GUI, MSGLogger, and MSGBrowser.

3.4 Sections interconnections and synchronization

All sections are connected to trigger control system (TCS), which is responsible for sharing informations about collected events and synchronization. Inner COMPASS Ethernet network is used for control and communication between first and third section. This network is used for all other computers in the experimental too. Dedicated network is used for all IPBus communication between second and third section.

4 Development process of the new DAQ

Development process of DAQ for bigger experiments is complicated process, thus it is necessary to divide it to several steps and set some milestones. The first milestone was creation of basic communication prototype. The next step was simple readout test without additional layers. One DAQ module with multiplexer firmware has been added in the third step. One more layer has been added to setup in fourth step. Last milestone is the full system test. All intermediate steps are shown in figure 2. More information about intermediate steps can be found in [6, 8, 7].

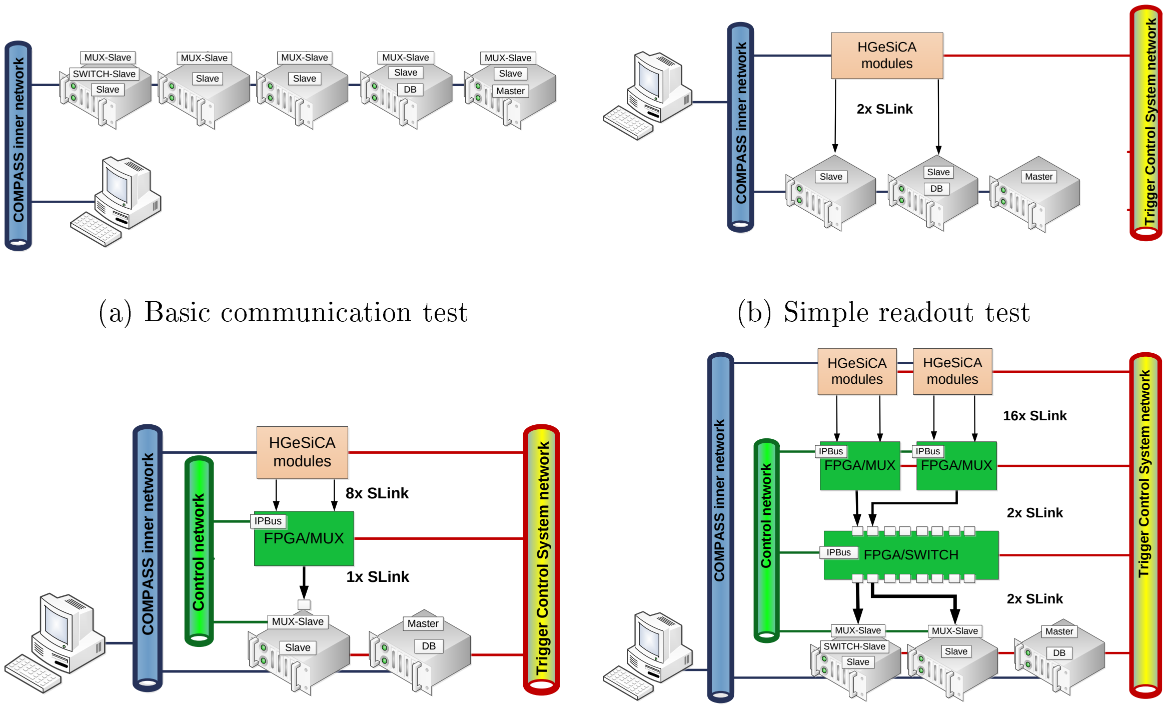
5 Finalization of the new DAQ

Finalization is done with full setup, but instead with smaller one composed of 5 MUXes, one SWITCH and 2 readout engine computers. It have to prove that system is reliable and can readout messages with high enough rate. In this phase detector experts are incorporated to development process as they are needed both for check of consistency of processed date and testing of graphical user interface(GUI). Input from users is extremely important for finalization of GUI design. The DAQ was in this phase at time of finalization of this paper.

6 User interfaces

This section is dedicated to description of new user interfaces. They are used to provide access to different aspects of a DAQ. Those aspects are:

- configuration,
- control,
- monitoring.



(a) Basic communication test

(b) Simple readout test

(c) One middle layer test

(d) Two middle layers test

Figure 2: Prototype steps

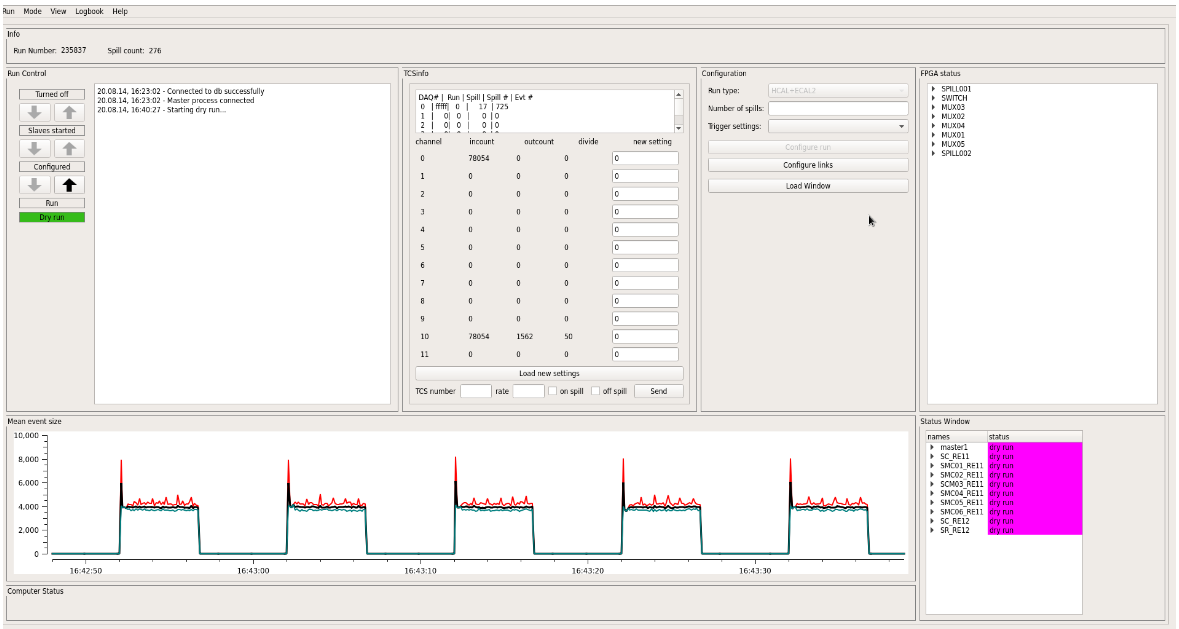
The first interface is web interface written in php with use of javascript and AJAX. This interface is used to change configuration of system. It is divided to several pages with different tasks from small thinks like description of FPGA registers up to definition of connections between detectors, multiplexers, switch, and computers.

The second interface for users is the one of the run control GUI program. It is the most important one because it will be used by both experts and normal collaboration members on a shift duty. It is written in C++ with use of Qt framework. It compose mainly from main run control window shown in figure 3, link overview window shown in figure 4, and load window shown in figure 5.

The main run control window serves for control of the state of system and for monitoring of processes' state, FPGA's status, trigger control system status, DAQ computers status, and event size.

Link overview window is representation of system layers. It is composed from subwidgets which are dynamically created during start of the window. One can track errors to their origin, check details of data flow, or activate/deactivate ports from this window.

The last of the listed windows is load window in which user can look at specific detector and see all information about its connection chain in DAQ, but main task of this window is different. It is used to issue load command to service named config server. This service than loads all necessary settings to selected device. These settings are extracted from the frontend database.



Run Number: 235837   Spill count: 276

```
20.08.14, 16:23:02 - Connected to db successfully
20.08.14, 16:23:02 - Master process connected
20.08.14, 16:40:27 - Starting dry run...
```

```
DAQ# |  Run | Spill | Spill # | Evt #
 0   | fffff|   0   |    17   | 725
 1   |    0|   0   |     0   | 0
 2   |    0|   0   |     0   | 0
```

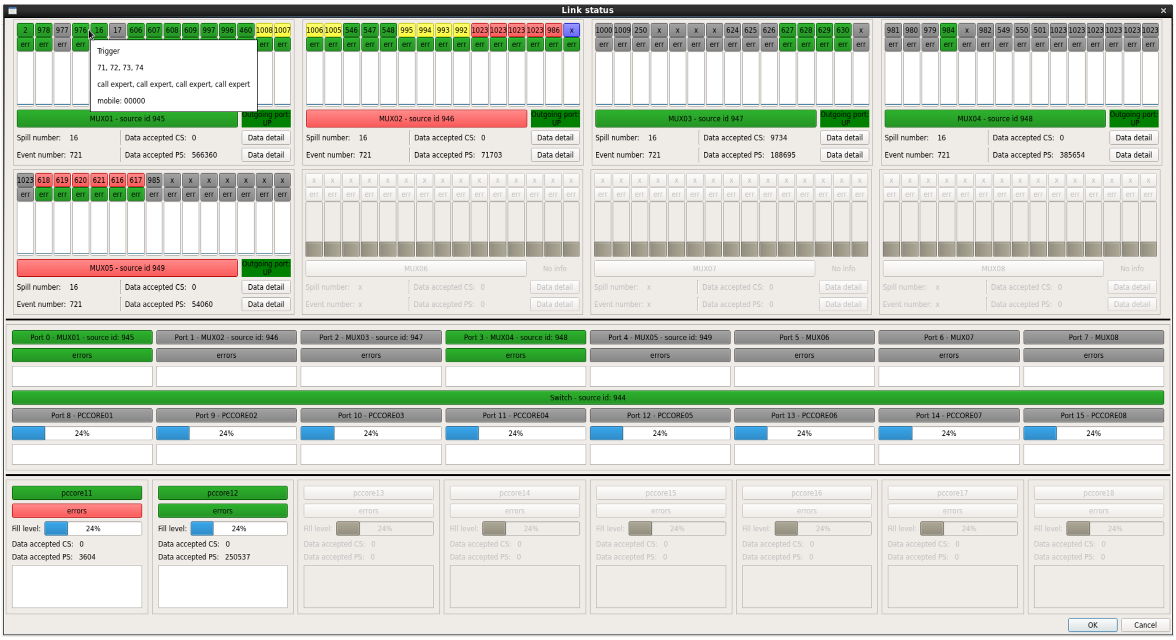
channel	incount	outcount	divide	new setting
0	78054	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	78054	1562	50	0
11	0	0	0	0

Run type: HCAL+ECAL2
Number of spills:
Trigger settings:

SPILL001
SWITCH
MUX03
MUX02
MUX04
MUX01
MUX05
SPILL002

names	status
master1	dry run
SC_RE11	dry run
SMC01_RE11	dry run
SMC02_RE11	dry run
SCM03_RE11	dry run
SMC04_RE11	dry run
SMC05_RE11	dry run
SMC06_RE11	dry run
SC_RE12	dry run
SR_RE12	dry run

Figure 3: Main run control window



Trigger
71, 72, 73, 74
call expert, call expert, call expert, call expert
mobile: 00000

MUX01 - source id 945
Spill number: 16   Data accepted CS: 0
Event number: 721   Data accepted PS: 566360

MUX02 - source id 946
Spill number: 16   Data accepted CS: 0
Event number: 721   Data accepted PS: 71703

MUX03 - source id 947
Spill number: 16   Data accepted CS: 9734
Event number: 721   Data accepted PS: 188695

MUX04 - source id 948
Spill number: 16   Data accepted CS: 0
Event number: 721   Data accepted PS: 385654

MUX05 - source id 949
Spill number: 16   Data accepted CS: 0
Event number: 721   Data accepted PS: 54060

MUX06
Spill number: x   Data accepted CS: 0
Event number: x   Data accepted PS: 0

MUX07
Spill number: x   Data accepted CS: 0
Event number: x   Data accepted PS: 0

MUX08
Spill number: x   Data accepted CS: 0
Event number: x   Data accepted PS: 0

Port 0 - MUX01 - source id: 945	Port 1 - MUX02 - source id: 946	Port 2 - MUX03 - source id: 947	Port 3 - MUX04 - source id: 948	Port 4 - MUX05 - source id: 949	Port 5 - MUX06	Port 6 - MUX07	Port 7 - MUX08

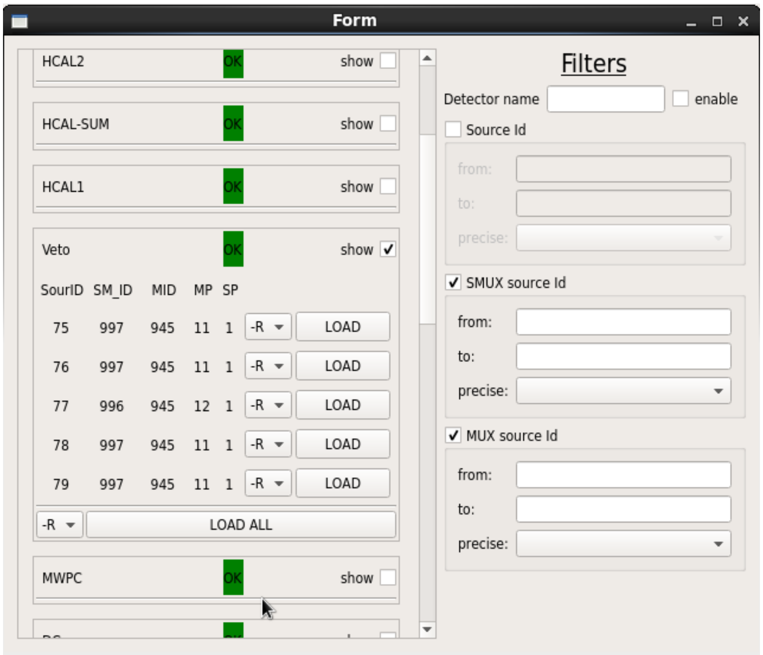
Switch - source id: 944

Port 8 - PCCORE01	Port 9 - PCCORE02	Port 10 - PCCORE03	Port 11 - PCCORE04	Port 12 - PCCORE05	Port 13 - PCCORE06	Port 14 - PCCORE07	Port 15 - PCCORE08
24%	24%	24%	24%	24%	24%	24%	24%

pccore11	pccore12	pccore13	pccore14	pccore15	pccore16	pccore17	pccore18
Fill level: 24%	Fill level: 24%	Fill level: 24%	Fill level: 24%	Fill level: 24%	Fill level: 24%	Fill level: 24%	Fill level: 24%
Data accepted CS: 0	Data accepted CS: 0	Data accepted CS: 0	Data accepted CS: 0	Data accepted CS: 0	Data accepted CS: 0	Data accepted CS: 0	Data accepted CS: 0
Data accepted PS: 3604	Data accepted PS: 250537	Data accepted PS: 0	Data accepted PS: 0	Data accepted PS: 0	Data accepted PS: 0	Data accepted PS: 0	Data accepted PS: 0

Figure 4: Link overview window

Main monitoring interface shown in figure 6 is called message browser. It is written in C++ with use of Qt framework. This interface is used only for monitoring purposes. It can work in online mode, in which it gathers messages directly from running processes, or in offline mode, in which it gets messages from database. Users can use filters to select just specific messages.

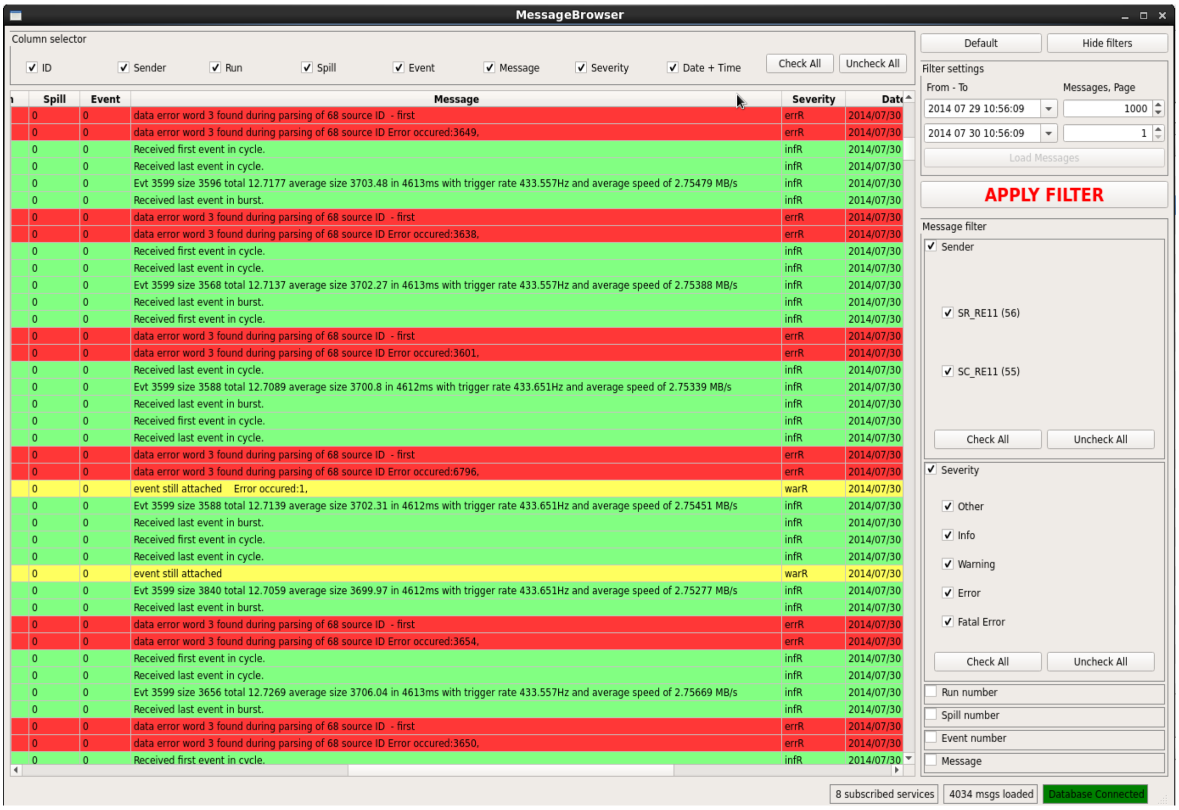


SourID	SM_ID	MID	MP	SP		
75	997	945	11	1	-R	LOAD
76	997	945	11	1	-R	LOAD
77	996	945	12	1	-R	LOAD
78	997	945	11	1	-R	LOAD
79	997	945	11	1	-R	LOAD

-R LOAD ALL

Detector name: □ enable
□ Source Id — from: to: precise:
☑ SMUX source Id — from: to: precise:
☑ MUX source Id — from: to: precise:

Figure 5: Load window



Spill	Event	Message	Severity	Date
0	0	data error word 3 found during parsing of 68 source ID - first	errR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID Error occured:3649,	errR	2014/07/30
0	0	Received first event in cycle.	infR	2014/07/30
0	0	Received last event in cycle.	infR	2014/07/30
0	0	Evt 3599 size 3596 total 12.7177 average size 3703.48 in 4613ms with trigger rate 433.557Hz and average speed of 2.75479 MB/s	infR	2014/07/30
0	0	Received last event in burst.	infR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID - first	errR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID Error occured:3638,	errR	2014/07/30
0	0	Received first event in cycle.	infR	2014/07/30
0	0	Received last event in cycle.	infR	2014/07/30
0	0	Evt 3599 size 3568 total 12.7137 average size 3702.27 in 4613ms with trigger rate 433.557Hz and average speed of 2.75388 MB/s	infR	2014/07/30
0	0	Received last event in burst.	infR	2014/07/30
0	0	Received first event in cycle.	infR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID - first	errR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID Error occured:3601,	errR	2014/07/30
0	0	Received last event in cycle.	infR	2014/07/30
0	0	Evt 3599 size 3588 total 12.7089 average size 3700.8 in 4612ms with trigger rate 433.651Hz and average speed of 2.75339 MB/s	infR	2014/07/30
0	0	Received last event in burst.	infR	2014/07/30
0	0	Received first event in cycle.	infR	2014/07/30
0	0	Received last event in cycle.	infR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID - first	errR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID Error occured:6796,	errR	2014/07/30
0	0	event still attached Error occured:1,	warR	2014/07/30
0	0	Evt 3599 size 3588 total 12.7139 average size 3702.31 in 4612ms with trigger rate 433.651Hz and average speed of 2.75451 MB/s	infR	2014/07/30
0	0	Received last event in burst.	infR	2014/07/30
0	0	Received first event in cycle.	infR	2014/07/30
0	0	Received last event in cycle.	infR	2014/07/30
0	0	event still attached	warR	2014/07/30
0	0	Evt 3599 size 3840 total 12.7059 average size 3699.97 in 4612ms with trigger rate 433.651Hz and average speed of 2.75277 MB/s	infR	2014/07/30
0	0	Received last event in burst.	infR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID - first	errR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID Error occured:3654,	errR	2014/07/30
0	0	Received first event in cycle.	infR	2014/07/30
0	0	Received last event in cycle.	infR	2014/07/30
0	0	Evt 3599 size 3656 total 12.7269 average size 3706.04 in 4613ms with trigger rate 433.557Hz and average speed of 2.75669 MB/s	infR	2014/07/30
0	0	Received last event in burst.	infR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID - first	errR	2014/07/30
0	0	data error word 3 found during parsing of 68 source ID Error occured:3650,	errR	2014/07/30
0	0	Received first event in cycle.	infR	2014/07/30

Filter settings
From - To: 2014 07 29 10:56:09 — 2014 07 30 10:56:09
Messages, Page: 1000 — 1

Message filter
☑ Sender
☑ SR_RE11 (56)
☑ SC_RE11 (55)

☑ Severity
☑ Other
☑ Info
☑ Warning
☑ Error
☑ Fatal Error

□ Run number
□ Spill number
□ Event number
□ Message

8 subscribed services | 4034 msgs loaded | Database Connected

Figure 6: Message browser

7 Conclusion

Demands and restriction on the new data acquisition system were extracted from initial studies and discussion in collaboration of the COMPASS experiment at CERN. The new DAQ software and hardware has been prepared based on these demands and restrictions. The first full version of software package has been tested and used during preparation for winter 2014 data taking. Complete DAQ setup is in final stage of preparation at time of finalization of this paper. Tests performed so far proved viability of the new system, thus the system was approved for usage in winter 2014 data taking.

References

[1] V. Jarý: *Analysis and proposal of the new architecture of the selected parts of the software support of the COMPASS experiment* Prague, 2012, Doctoral thesis, Czech Technical University in Prague

[2] T. Anticic, et al. (ALICE DAQ Project): *ALICE DAQ and ECS User's Guide* CERN, EDMS 616039, January 2006.

[3] T. M. Shaw, et al.: *Architecture and development of the CDF hardware event builder* IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 36, NO. 1, AUGUST 1989

[4] J. Nový: *COMPASS DAQ - Basic Control System* Prague, 2012, Master thesis, Czech Technical University in Prague

[5] M. Bodlák: *COMPASS DAQ Database Architecture and Support Utilities* Prague, 2012, Master thesis, Czech Technical University in Prague

[6] M. Bodlák, et al. *Developing Control and Monitoring Software for the Data Acquisition System of the COMPASS Experiment at CERN.* Acta polytechnica: Scientific Journal of the Czech Technical University in Prague. Prague, CTU, 2013, issue 4. Available at: http://ctn.cvut.cz/ap/

[7] M. Bodlak, et al. *FPGA based data acquisition system for COMPASS experiment.* Journal of Physics: Conference Series. 2014-06-11, vol. 513, issue 1, s. 012029-. DOI: 10.1088/1742-6596/513/1/012029. Available at: http://stacks.iop.org/1742-6596/513/i=1/a=012029?key=crossref.78788d23de2b4a6a34d127c361123b8c

[8] M. Bodlak, et al. *New data acquisition system for the COMPASS experiment.* Journal of Instrumentation. 2013-02-01, vol. 8, issue 02, C02009-C02009. DOI: 10.1088/1748-0221/8/02/C02009. Available at: http://stacks.iop.org/1748-0221/8/i=02/a=C02009?key=crossref.a76044facdf29d0fb21f9eefe3305aa5

[9] M. Bodlák, V. Jarý, J. Nový: *Software for the new COMPASS data acquisition system.* In: COMPASS collaboration meeting, Geneva, Switzerland, 18 November 2011

[10] P. Abbon, et al.(the COMPASS collaboration): *The COMPASS experiment at CERN.* In: Nucl. Instrum. Methods Phys. Res., A 577, 3 (2007) pp. 455–518

[11] L. Schmitt, et al.: *The DAQ of the COMPASS experiment.* In: 13th IEEE-NPSS Real Time Conference 2003, Montreal, Canada, 18–23 May 2003, pp. 439–444

[12] E. Bal, et al.: *The NA48' Data Acquisition System* IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 45, NO. 4, AUGUST 1998

[13] M. Wittgen, et al.: *The NA48' Event-Building PC Farm* IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 47, NO. 2, APRII. 2000

Parametric Study for Kernel Based Classification[*]

Jiří Palek

2nd year of PGS, email: jiri.palek@gmail.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The variability in kernel methods is done by kernel functions. These functions are parameterized and, therefore, in application it is necessary to deal with the choice of parameter. This article deals with scheme for the choice of parameter in kernel PCA with Gaussian and exponential kernel function. Our theoretical approach comes from the study of properties of rounding kernels in finite arithmetics used in software implementation. This theoretical concept is then tested on two-level classification system which is represented by kernel principal component analysis and quadratic discriminant analysis. It is used for diagnostics of Alzheimer's disease.

Keywords: Gaussian kernel, exponential kernel, parametric study, rounding, kernel PCA, whitening, QDA, Alzheimer's disease, diagnostics, leave-one-out cross validation

Abstrakt. Jádrové metody poskytují uživateli velkou variabilitu díky možnosti volby jádrové funkce. Tyto funkce jsou parametrizované a proto je nutné při jejich použití vyřešit otázku vhodné volby těchto parametrů. Tento článek se zabývá schématem pro volbu parametru v jádrové PCA s Gaussovskou a exponenciální jádrovou funkcí. Teoretický přístup vychází ze studie vlastností jader při zaokrouhlování v konečné aritmetice, která se používá v softwarových implementacích výpočetních prostředí. Teoretický koncept je následně testován na dvouúrovňovém klasifikačním modelu, který je složen z analýzy hlavních komponent a kvadratické diskriminační analýzy. Model je použit pro diagnostiku Alzheimerovy choroby.

Klíčová slova: Gaussovská jádrová funkce, exponenciální jádrová funkce, parametrická studie, zaokrouhlování, jádrová PCA, whitening, kvadratická diskriminační analýza, diagnostika, leave-one-out křížová validace

1 Introduction

Kernel-based methods represent popular and well established tools for various data mining tasks.

Let $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ is a set of observations $\mathbf{x}_i \in \mathcal{X}$ and modeled property $y_i \in \mathcal{Y}$, $n \in \mathbb{N}$. The principle of kernel methods is to embed another Hilbert space $\mathcal{H}$ between the input space $\mathcal{X}$ and the output space $\mathcal{Y}$ and perform a dot-product-based methods there. A connection between distances in spaces $\mathcal{X}$ and $\mathcal{H}$ is done via so called kernel functions $\mathrm{k} : \mathcal{X} \times \mathcal{X} \to \mathcal{H}$. The Hilbert space $\mathcal{H}$ and the kernel function k are so constructed that

[*] The paper was created under the support of grant SGS11/165/OHK4/3T/14 CTU in Prague

the followig equation holds $\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbb{x}_i, \mathbb{x}_j \rangle$, where $\langle \cdot, \cdot \rangle$ is a dot-product in $\mathcal{H}$. The original set of data $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ is consequently transform into so called kernel matrix $\mathbb{K}$, which is defined as $(\mathbb{K})_{ij} = \mathrm{k}(\mathbf{x}_i, \mathbf{x}_j), \forall i, j \in \{1, ..., n\}$. For more detailed description see [1] and [2].

Exaples of kernel functions can be found in [1] and [2]. In general, these functions are parameterized and, therefore, in application it is necessary to deal with the choice of this parameter. In this paper we restrict our attention to the following two exponential-based kernel functions

- exponential kernel with parameter $\sigma > 0$

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}\right), \tag{1}$$

- Gaussian kernel with parameter $\sigma > 0$

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right), \tag{2}$$

and find some general rules for choosing their parameter σ.

There is a finite arithmetic in computers: fixed and floating point systems with finite representation. We suppose finitness of floating point mantisa and create a model for choice of parameter σ on it. The main idea is to use the following expansion of exponential function

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

for the kernel functions (1), (2) as model for rounding error in software finite arithmetics and contract the set of possible values for parameter $\sigma > 0$ to relevant ones. The algorithmic details are left out.

The final stage is testing our theoretical approach on Alzheimer's disease diagnostics. The two-level classification system, kernel principal component analysis (PCA) and quadratic discriminant analysis (QDA), is used for this task.

2 Properties of Exponential and Gaussian Kernel

Necessary theoretical background is developed in this section. First of all, the consideration is done for exponential kernel, and then the results are reformulated for the Gaussian kernel.

For simplicity of notation, we write d_{ij} instead of $\|\mathbf{x}_i - \mathbf{x}_j\|_2$. Additionaly, from now on, $\hat{n}$ denotes the set $\{1, ..., n\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\sigma \in \mathbb{R}$, $\sigma > 0$, $d_{\max} = \max\{d_{ij} | i, j \in \hat{n}, i \neq j\}$ and $d_{\min} = \min\{d_{ij} | i, j \in \hat{n}, i \neq j\}$.

2.1 Properties of exponential kernel function

Let us begin with properties of simple exponential kernel function.

Definition 1. *The finite exponential kernel* $\mathrm{k_e}$ *of order* $N \in \mathbb{N}_0 \cup \{\infty\}$ *is given by*

$$\mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; N, \sigma) = \sum_{k=0}^{N} \frac{(-1)^k}{k!} \left(\frac{d_{ij}}{\sigma}\right)^k .$$

Definition 2. *The accuracy of the finite exponential kernel of order* $N \in \mathbb{N}_0$ *is defined by*

$$\begin{aligned} &\mathrm{a}(\mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; N, \sigma)) = \\ &= |\,\mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; N+1, \sigma) - \mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; N, \sigma)| = \\ &= \frac{1}{(N+1)!} \left(\frac{d_{ij}}{\sigma}\right)^{N+1} \end{aligned} \tag{3}$$

Proposition 1. *Let* $\sigma > d_{ij}$*. Then for the following series* $\sum\limits_{k=0}^{\infty} (-1)^k a_k = \sum\limits_{k=0}^{\infty} \frac{(-1)^k}{k!} \left(\frac{d_{ij}}{\sigma}\right)^k$, *it holds*

(i) *the sequence* $(a_k)_{k=0}^{\infty}$ *is positive and strictly monotonically decreasing,* $a_0 = 1$,

(ii) *the series* $\sum\limits_{k=0}^{\infty} (-1)^k a_k$ *is convergent,*

(iii) *the partial sum satisfies*

$$\left| \sum_{k=0}^{\infty} (-1)^k a_k - \sum_{k=0}^{N} (-1)^k a_k \right| \leq \frac{1}{(N+1)!} \left(\frac{d_{ij}}{\sigma}\right)^{N+1} .$$

Proof.

(i) When $\sigma > d_{ij}$ then $\frac{d_{ij}}{\sigma} < 1$ and $\frac{a_{k+1}}{a_k} = \frac{d_{ij}}{\sigma} \frac{1}{(k+1)} < 1$ for all $k \in \mathbb{N}_0$.

(ii) Convergence implies from Leibnitz criterion and (i).

(iii) In the case of convergence, we can directly estimate $\left| \sum\limits_{k=0}^{\infty} (-1)^k a_k - \sum\limits_{k=0}^{N} (-1)^k a_k \right| = \left| a_{N+1} + \sum\limits_{k=1}^{\infty} (-1)^k a_{N+1+k} \right| \leq a_{N+1}$. The last inequality holds from the fact that the sum $\sum\limits_{k=1}^{\infty} (-1)^k a_{N+1+k} < 0$.

□

Proposition 1 states that the finite exponential kernel $\mathrm{k_e}$ is an approximation of proper exponential kernel (1) and that the accuracy (3) is upper estimation of an error of this approximation.

Definition 3. *Let $\sigma > d_{ij}$ and $1 > \epsilon > 0$ be a counting error. The minimum order of* $\mathrm{k_e}$ *is given by $\widetilde{N} = \min\{N \in \mathbb{N}_0 | a(\mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; N, \sigma)) < \epsilon\}$.*

Theorem 1. *Let $1 > \epsilon > 0$ be a counting error, $\sigma > d_{ij}$, $\widetilde{N} \in \mathbb{N}_0$, and $\mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; \widetilde{N}, \sigma)$ be a finite exponential kernel.*

(i) The $\mathrm{k_e}$ *has a minimum order $\widetilde{N} = 0$ if and only if $\sigma \in \left(d_{ij}\frac{1}{\epsilon}; \infty\right)$.*

(ii) The $\mathrm{k_e}$ *has a minimum order $\widetilde{N} > 0$ if and only if $\sigma \in \left(d_{ij} \sqrt[\widetilde{N}+1]{\frac{1}{\epsilon(\widetilde{N}+1)!}}; d_{ij} \sqrt[\widetilde{N}]{\frac{1}{\epsilon \widetilde{N}!}}\right\rangle$.*

Proof. The members $(a_k)_{k=\widetilde{N}+1}^{\infty}$ of the series $\mathrm{k_e}(\mathbf{x}_i, \mathbf{x}_j; \infty, \sigma) = \sum_{k=0}^{\infty} (-1)^k a_k = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!}\left(\frac{d_{ij}}{\sigma}\right)^k$ are nonnegative and strictly monotonically decreasing, and therefore they are not counted if and only if it holds

$$\left|(-1)^{\widetilde{N}+1} a_{\widetilde{N}+1}\right| = \frac{1}{(\widetilde{N}+1)!}\left(\frac{d_{ij}}{\sigma}\right)^{\widetilde{N}+1} < \epsilon.$$

By solving this inequality we obtain

$$\sigma > d_{ij} \sqrt[\widetilde{N}+1]{\frac{1}{\epsilon(\widetilde{N}+1)!}}$$

The above formula is lower bound of each interval; the first one is $\left(d_{ij}\frac{1}{\epsilon}; \infty\right)$, the second is $\left(d_{ij}\sqrt{\frac{1}{\epsilon 2!}}; d_{ij}\frac{1}{\epsilon}\right\rangle$ and so on.

According to the assumption on σ and ϵ, the intervals are meaningful and thus we have the assertion of the Theorem 1. □

2.2 Properties of exponential Kernel Matrix

Next, let us turn to kernel matrix $\mathbb{K}$. In what follows, $\mathbb{K}_\mathrm{e}$ stands for the kernel matrix $\mathbb{K}$ based on finite exponential kernel, $(\mathbb{K}_\mathrm{e})_{ij} = \mathrm{k_e}(\mathbf{x}_i, \mathbf{x}_j; N_{ij}, \sigma)$

Definition 4. *A kernel matrix $\mathbb{K}_\mathrm{e}$, $(\mathbb{K}_\mathrm{e})_{ij} = \mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; \widetilde{N}_{ij}, \sigma)$ for all $i, j \in \hat{n}$, has a minorder $N \in \mathbb{N}$ if*

(i) $(\forall i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij}$ is a minimum order of $\mathrm{k_e}$*),*

(ii) $(\forall i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij} \geq N)$,

(iii) $(\exists i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij} = N)$.

Definition 5. *A kernel matrix $\mathbb{K}_\mathrm{e}$ has a maxorder $N \in \mathbb{N}_0$ if*

(i) $(\forall i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij}$ is a minimum order of $\mathrm{k_e}$*),*

(ii) $(\forall i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij} \leq N)$,

(iii) $(\exists i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij} = N)$.

Definition 6. *A kernel matrix* $\mathbb{K}_{\mathrm{e}}$ *has an exaorder* $N \in \mathbb{N}_0$ *if*

(i) $(\forall i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij}$ *is a minimum order of* $\mathrm{k_e})$*,*

(ii) $(\forall i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij} = N)$.

Theorem 2. *Let* $1 > \epsilon > 0$ *be a counting error,* $\sigma > d_{\max}$*,* $\mathbb{K}_{\mathrm{e}}$*,* $(\mathbb{K}_{\mathrm{e}})_{ij} = \mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; \widetilde{N}_{ij}, \sigma)$ *for all* $i, j \in \hat{n}$*, is a kernel matrix, and* $\widetilde{N}_{ij}$ *are minimum orders.*

$\mathbb{K}_{\mathrm{e}}$ *has a minorder* $N \in \mathbb{N}$ *if and only if* σ *is in interval* $\left(d_{\min} \sqrt[N+1]{\frac{1}{\epsilon(N+1)!}}; d_{\min} \sqrt[N]{\frac{1}{\epsilon N!}}\right\rangle$ $\cap\, (d_{\max}; \infty)$.

Proof. The proof is straightforward verification that σ from the introduced interval fulfils the conditions (i), (ii) and (iii) from the Definition 4:

(i) Directly implies from the assuption of the theorem.

(ii) Property $(\forall i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij} \geq N)$ can be equivalently formulated using Theorem 1 into $(\forall i, j \in \hat{n})(i \neq j)\left(\sigma \leq d_{ij} \sqrt[N]{\frac{1}{\epsilon N!}}\right)$. Therefore $\sigma \leq d_{\min} \sqrt[N]{\frac{1}{\epsilon N!}}$.

(iii) We apply Theorem 1 on the condition $(\exists i, j \in \hat{n})(i \neq j)(\widetilde{N}_{ij} = N)$ and rewrite it as follows $(\exists i, j \in \hat{n})(i \neq j)\left(\sigma \in \left(d_{ij} \sqrt[N+1]{\frac{1}{\epsilon(N+1)!}}; d_{ij} \sqrt[N]{\frac{1}{\epsilon N!}}\right\rangle\right)$. To fulfill this condition, σ has to be in the interval $\left(d_{\min} \sqrt[N+1]{\frac{1}{\epsilon(N+1)!}}; d_{\max} \sqrt[N]{\frac{1}{\epsilon N!}}\right\rangle$.

Combining conditions from (ii) and (iii) with the assupmtion $\sigma > d_{\max}$, we obtain the assertion of the Theorem. □

By analogy, we can formulate and prove following Theorems.

Theorem 3. *Let* $1 > \epsilon > 0$ *be a counting error,* $\sigma > d_{\max}$*,* $\mathbb{K}_{\mathrm{e}}$*,* $(\mathbb{K}_{\mathrm{e}})_{ij} = \mathrm{k_e}(\boldsymbol{x}_i, \boldsymbol{x}_j; \widetilde{N}_{ij}, \sigma)$ *for all* $i, j \in \hat{n}$*, is a kernel matrix, and* $\widetilde{N}_{ij}$ *are minimum orders.*

$\mathbb{K}_{\mathrm{e}}$ *has a maxorder* $N \in \mathbb{N}$ *if and only if* σ *is in interval* $\left(d_{\max} \sqrt[N+1]{\frac{1}{\epsilon(N+1)!}}; d_{\max} \sqrt[N]{\frac{1}{\epsilon N!}}\right\rangle \cap$ $(d_{\max}; \infty)$.

Theorem 4. *Let* $1 > \epsilon > 0$ *is a counting error.* $\mathbb{K}_{\mathrm{e}}$ *has an exaorder 0 if and only if* σ *is in interval* $\left(d_{\max} \frac{1}{\epsilon}; \infty\right)$.

Theorem 5. *Let* $1 > \epsilon > 0$ *is a counting error.* $\mathbb{K}_{\mathrm{e}}$ *has an exaorder 1 if and only if* σ *is in interval* $\left(d_{\max} \sqrt{\frac{1}{\epsilon 2!}}; d_{\min} \frac{1}{\epsilon}\right\rangle$.

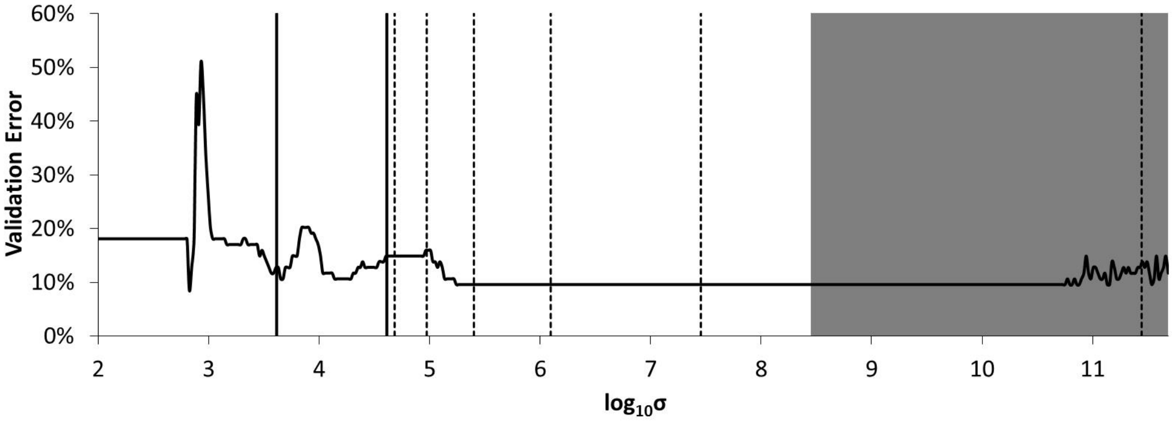


Figure 1: Leave-one-out cross validation error for QDA with Gaussian Kernel PCA with two (uppet) and seven (bottom) components. The dashed vertical lines represents the intervals based on $d_{\min}$.

2.3 Properties of Gaussian kernel matrix

The thoughts are the same as in Section 2.2, and therefore only the conclusion is presented in the following part.

Theorem 6. *Let $1 > \epsilon > 0$ be a counting error, $\sigma > \frac{d_{\max}}{\sqrt{2}}$, $\mathbb{K}_{\mathrm{g}}$, $(\mathbb{K}_{\mathrm{g}})_{ij} = \mathrm{k_g}(\boldsymbol{x}_i, \boldsymbol{x}_j; \widetilde{N}_{ij}, \sigma)$ for all $i, j \in \hat{n}$, is a kernel matrix, and $\widetilde{N}_{ij}$ are minimum orders. $\mathbb{K}_{\mathrm{g}}$ has*

(i) a minorder $N \in \mathbb{N}$ if σ is in interval

$$\left(\frac{d_{\min}}{\sqrt{2}} \sqrt[2N+2]{\frac{1}{\epsilon(N+1)!}}; \frac{d_{\min}}{\sqrt{2}} \sqrt[2N]{\frac{1}{\epsilon N!}}\right\rangle \cap \left(\frac{d_{\max}}{\sqrt{2}}; \infty\right) ,$$

(ii) a maxorder $N \in \mathbb{N}$ if σ is in interval

$$\left(\frac{d_{\max}}{\sqrt{2}} \sqrt[2N+2]{\frac{1}{\epsilon(N+1)!}}; \frac{d_{\max}}{\sqrt{2}} \sqrt[2N]{\frac{1}{\epsilon N!}}\right\rangle \cap \left(\frac{d_{\max}}{\sqrt{2}}; \infty\right) ,$$

(iii) an exaorder 0 if σ is in interval $\left(\frac{d_{\max}}{\sqrt{2}}\sqrt{\frac{1}{\epsilon}}; \infty\right)$,

(iv) an exaorder 1 if σ is in interval $\left(\frac{d_{\max}}{\sqrt{2}}\sqrt[4]{\frac{1}{2\epsilon}}; \frac{d_{\min}}{\sqrt{2}}\sqrt{\frac{1}{\epsilon}}\right\rangle$.

3 Two Level Classification System

The theory build in previous section will be tested on two-level classification system. A kernel-based principal component analysis (PCA) [1] is used for the reduction of dimensionality of the problem in the first part, whereas the quadratic discriminant analysis (QDA) [3] is performed to the analysis itself. We will touch only a few aspects of the kernel-based PCA and QDA in this section.

3.1 Kernel PCA

Kernel PCA is extension of classical PCA [6]. In general, PCA is a transformation of data into new coordinate system which is defined by eigenvectors.

The eigenvalue decomposion is in classical PCA done from covariance matrix $\mathbb{C} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$ of centered data satysfying

$$\sum_{i=1}^{n}\mathbf{x}_i = \mathbb{O}_{p,1}, \mathbf{x}_i \in \mathcal{X} = \mathbb{R}^{p,1}. \tag{4}$$

Let $(\mathbf{v}_1, ..., \mathbf{v}_p)$ denote the eigenvectors of matrix $\mathbb{C}$ and $(\lambda_1, ..., \lambda_p)$ detones the consecutive eigenvalues. The transformation matrix $\mathbb{A}$ of old coordinates $\mathbb{X}$ into new one $\mathbb{Z} = \mathbb{A}\mathbb{X}$, where $\mathbb{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, is defined as $\mathbb{A} = (\mathbf{v}_1, ..., \mathbf{v}_p)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ holds for eigenvalues. Instead of using matrix $\mathbb{A}$, normalized matrix $\mathbb{W} = \left(\frac{\mathbf{v}_1}{\sqrt{\lambda_1}}, ..., \frac{\mathbf{v}_p}{\sqrt{\lambda_p}}\right)$ can be used in the case of data whitening [1].

There is shown in [1] that doing kernel PCA is equivalent to doing classical PCA with kernel matrix $\mathbb{K}$ instead of covariance matrix $\mathbb{C}$. The centering operation (4) can be done in kernel PCA by following transformation of kernel matrix $\mathbb{K}$ called kernel whitening

$$\hat{\mathbb{K}} = \mathbb{K} - \frac{1}{n}\mathbb{I}_{n,n}\mathbb{K} - \frac{1}{n}\mathbb{K}\mathbb{I}_{n,n} + \frac{1}{n^2}\mathbb{I}_{n,n}\mathbb{K}\mathbb{I}_{n,n}.$$

Finally, the kernel PCA is done by eigenvalue decomposion of the matrix $\hat{\mathbb{K}}$.

3.2 Quadratic Discriminant Analysis

The principal of QDA algorithm [3] is to approximate the data from different classes by normal distributions. The classification of new observation is then made by calculating the probability of pertinence to every class and choosing the one with the maximum value.

Let us assume that we have M classes C_i, $i \in \hat{M}$, with distributions $f_i(\mathbf{x}), \mathbf{x} \in \mathcal{X} = \mathbb{R}^{p,1}$. The goal is to decompose $\mathcal{X}$ into N sets A_i, $\mathcal{X} = \cup_{i=1}^{M} A_i$. The classification rule is then given by formula

$$\mathbf{x} \in C_i \Leftrightarrow \mathbf{x} \in A_i. \tag{5}$$

Finding the optimal decomposition is equivalent to finding the minimum of functional

$$L = \sum_{i=1}^{M}\int_{A_i}\sum_{j=1}^{M}\pi_j f_j(\mathbf{x})d\mathbf{x}, \tag{6}$$

where π_j is a priori probability of class C_j. There is shown in [3] that the classification rule (5) with respect to the functional (6) gives following condition for classification

$$(\mathbf{x} \in c_j) \Leftrightarrow \left(\left(\forall t \in \hat{M}\right)(j \neq t)\left(\pi_t f_t(\mathbf{x}) > \pi_j f_j(\mathbf{x})\right)\right)$$

Where normal probability distribution $f_j \sim N(\mu_j, \sum_j)$ is used in the case of QDA.

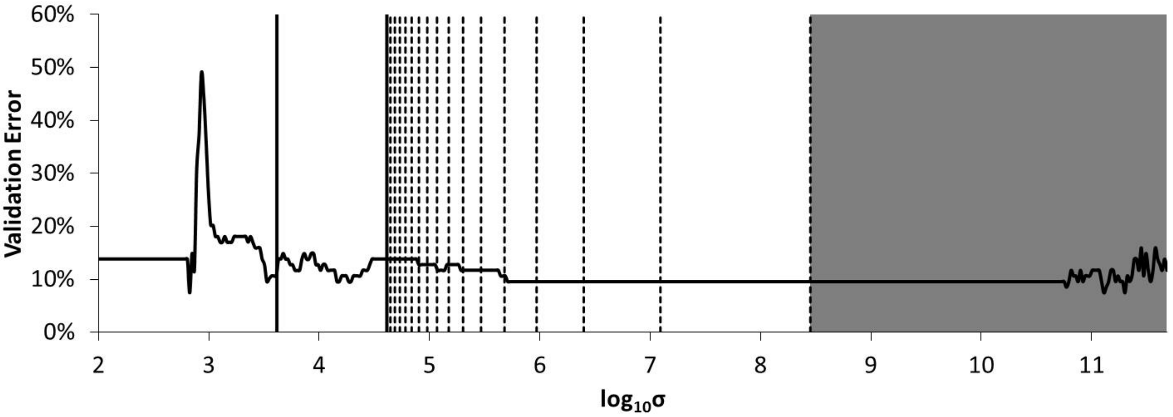


Figure 2: Leave-one-out cross validation error for QDA with Gaussian Kernel PCA with five and eight components. The dashed vertical lines represents the intervals based on $d_{\min}$.

4 Experimental Part

4.1 Alzheimer's Disease Diagnostics

With increasing life expectancy across the world, the number of elderly people with dementia is growing rapidly. Dementia is characterized by irreversible and progressive decline of cognitive functions interfering with common activities of daily living and social and working skills. It has many causes. The most common neurodegenerative disorder is Alzheimer's disease (AD). Other frequent diseases with dementia include vascular dementia, frontotemporal lobar degeneration and dementia with Lewy bodies.

The treatment of AD is the most effective in the initial phase. Therefore it is of a great importance to identify patients with AD among the whole spectrum of dementia diseases accurately and early. The correct diagnosis of AD in the incipient stages is difficult. Clinicians can diagnose probable AD based on clinical findings. AD is considered definite if both clinical and histopathological evidence are present. Clinical accuracy for AD according to the National Institute of Neurological and Communicative Sisorders and Stroke and Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria is 80 % at the experienced centers and decreases in less trained physicians. There is no single diagnostic test for AD or most of other types of dementia. The new revised research criteria for AD have introduced biological markers as supportive features in the diagnosis of AD [5]. They include magnetic resonance imaging, cerebrospinal fluid biomarkers and positron emission tomography (PET). However, single photon emission tomography (SPECT) of the brain is more widely available and cheaper than PET. Since the diagnostic accuracy estimates is bellow prerequisite 80 % level required by the Reagan Biomarker Working Group. 99M Tc-HMPAO SPECT identifies diagnosed AD with moderate sensitivity (77 - 80%) and specificity (65 - 93%) [5]. Therefore SPECT of the brain is not recommended investigation for AD according the European and US guidelines [5],[4]. This modality should be used in an unclear case after clinical and structural imaging work up. Therefore there is space to increase the diagnostic potential of this functional neuroimaging to detect AD correctly, especially in the earliest stages.

4.2 Data Description

Two groups of patients (AD/CN) were investigated via 19m F-deoxyglucose radiomarker of brain aktivity using SPECT technique. Patient scans were represented as 3D matrces of size 79x95x69 which were space normalized using SPM-7 techique [7],[8]. Intensity maps were also normalized to obtain nonnegative intensities with unit patient sums. First group conists of 38 Alzheimer's diseased patients (AD). The second group consists of 56 Control normal patients (CN). Space and intensity normalizations enable to use voxel-by-voxel analysis of patient and group differences, where every patient is represented by intensity vector of length 517 845, which makes statistical analysis difficult in general. But our methodology is based only on patient-by-patient distances, which form a single distance matrix of size 94 x 94.

4.3 Methodology of Data Processing

The diagnostics of 3D SPECT images was divided into two parts. In the first step, the Kernel PCA was processed on the whole dataset to obtain principal components and new coordinates of the data. Calculations were performed for a wide variety of parameters of exponential and Gaussian kernels. This allowed us to study the dependence of result on choice of parameters. The next step was the own classification performed by QDA with leave-one-out cross validation. All calculations were performed in MATLAB environment.

Finally, we study the dependency of the leave-one-out cross validation error on the choice of the sigma parameter.

4.4 Analysis of Results

We used from one to eleven components for QDA. Results for selected number of components used are on Figures 1 and 2.

All figures have the same structure. The x-axis represents the $\log_{10}\sigma$, whereas the y-axis represents the leave-one-out cross validation error for QDA with the selected number of Kernel PCA components with Gaussian kernel. The grey area denotes the exaorder 0 and 1 from Theorem 6 *(iii)* and *(iv)*. The vertical solid lines are bouindaries $d_{\min}$ and $d_{\max}$ and the dashed lines represents the intervals based on Theorem 6. The minorder intervals *(i)* are used in Figure 1 and the maxorder intervals *(ii)* in the Figures 2.

5 Conclusion

It can be seen from the Figures, that our model for the choice of parameter is able to describe changes of results. The general model for the choice of the parameter can be state as follows:

(i) The most interestin region is interval $\left\langle d_{\min}/\sqrt{2}, d_{\max}/\sqrt{2}\right\rangle$ for Gaussian kernel and interval $\langle d_{\min}, d_{\max}\rangle$ for exponential kernel. We recomend to equidistantly search it with respect to the size of the dataset.

(ii) Otherwise, we do not recomend to use sigma from the intervals with exaorders one and zero. This is $\left(\frac{d_{\max}}{\sqrt{2}}\sqrt[4]{\frac{1}{2\epsilon}};\infty\right)$ for Gaussian kernel and $\left(d_{\max}\sqrt{\frac{1}{2\epsilon}};\infty\right)$ for exponential kernel.

(iii) For the choice of σ from the area between (i) and (ii) we can state that higher the σ, lower the interesting of result. Because of it we recomend to

(a) start with smaller σ,

(b) take as many of them as you can with respect to the maxorder or minorder intervals and stress the smaller ones.

For example, you can take geometric means of intervals *(i)* or *(ii)* from Theorem 6 for Gaussian kernel and geometric means of intervals from Theorem 2 or Theorem 3.

(iv) For the rest interval use again the equidistant search with respect to the size of dataset.

References

[1] J. Shawe-Taylor, N. Cristianini. *Kernel Methods for Patter Analysis.* Cambridge University Press, 2004. http://dx.doi.org/10.1017/CBO9780511809682

[2] B. Schölkopf, A. Smola. *Learning with Kernels.* Cambridge: MIT Press, 2002.

[3] J. Anděl. *Matematická statistika.* SNTL, Praha, 1978.

[4] D. S. Knopman. *Practice parameter: diagnosis of dementia (an evidence-based review).* Report of the Quality Standards Subcommittee of the American Academy of Neurology. In: Neurology, 2001 8;56(9) :1143-53.

[5] G. Waldemar, B. Dubois, M. Emre, J. Georges, I.G. McKeith, M. Rossor, P. Scheltens, P. Tariska, B. Winblad. *Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia.* EFNS guideline. Eur J Neurol 2007;14(1):e1-26. http://dx.doi.org/10.1111/j.1468-1331.2006.01605.x

[6] H. Řezanková. *Shluková analýza dat.* Professional Publishing, Praha, 2009.

[7] W. Jagust, R. Thisted, M. D. Devous. *SPECT perfusion imaging in the diagnosis of Alzheimer's disease: a clinicalpathologic study.* In: Neurology, 2001; 56: 950–56. http://dx.doi.org/10.1212/WNL.56.7.950

[8] N. J. Dougall, S. Bruggink, K. P. Ebmeier. *Systematic review of the diagnostic accuracy of 99mTc-HMPAO-SPECT in dementia.* In: Am J Geriatr Psychiatry 2004; 12: 554-70. http://dx.doi.org/10.1176/appi.ajgp.12.6.554

Numerical Simulation of NAPL Vapor Transport in Soil*

Ondřej Pártl

3rd year of PGS, email: partlond@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Beneš, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This paper deals with the simulation of NAPL vapor transport driven by gas flow in porous medium. The mathematical model describing this phenomenon combines mass balance equations with the Darcy law and the ideal gas equation of state. In order to solve the governing equations, a numerical scheme based on the finite volume method is derived. Finally, some of our numerical results computed by this scheme are presented.

Keywords: porous medium, vapor transport, finite volume method

Abstrakt. Tento příspěvek se zabývá simulací transportu par látek typu NAPL proudícím plynem v porézním prostředí. Tento jev je popisován matematickým modelem, který spojuje zákon zachvání hmoty s Darcyho rychlostí a stavovou rovnicí ideálního plynu. Pro řešení získaných rovnic je metodou konečných objemů odvozeno numerické schéma, jehož výsledky jsou v závěru prezentovány.

Klíčová slova: porézní prostředí, transport par, metoda konečných objemů

1 Introduction

Flow of gases in porous medium and transport of contaminants driven by this flow is a part of a variety of complicated natural processes and, for this reason, it has been researched and simulated for years. In our research, this contaminant is NAPL (Non-Aqueous Phase Liquids) vapor. The NAPLs are liquids that do not easily dissolve in water, e.g., gasoline or TCE.

The conservation laws describing the previous phenomenon cannot be solved numerically simply by applying, for example, the standard Galerkin finite element method because such an approach results in non-physical behavior of the numerical solution. Therefore, we test an approach combining the finite element method with finite volume method that is described in [4] for a different type of problems; it seems, however, to work in our case as well.

*This work is partly supported by the project "Development and Validation of Porous Media Fluid Dynamics and Phase Transitions Models for Subsurface Environmental Application" Kontakt II LH14003 of Czech Ministry of Education, Youth and Sports.

2 Mathematical Model

We consider NAPL vapor transport driven by gas flow in soil. By the term 'gas', we denote the first component of the whole mixture of the two gases. Typically, it will be air. The governing equation for the flow of the mixture in a rectangular domain $\Omega \subset \mathbb{R}^2$ is derived ([7], [9]) by substituting the Darcy velocity of the mixture,

$$\boldsymbol{u} = -\frac{1}{\mu}\boldsymbol{k}\left(\nabla p - \rho \boldsymbol{g}\right), \tag{1}$$

into the continuity equation of the mixture

$$\phi \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = F, \tag{2}$$

where μ $[\frac{\mathrm{kg}}{\mathrm{m \cdot s}}]$ is the dynamic viscosity, $\boldsymbol{k} = \begin{pmatrix} k_1 & k_2 \\ k_3 & k_4 \end{pmatrix}$ $[\mathrm{m}^2]$ the permeability tensor, $\boldsymbol{g} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$ $[\frac{\mathrm{m}}{\mathrm{s}^2}]$ the gravitational acceleration vector, p [Pa] the pressure, ρ $[\frac{\mathrm{kg}}{\mathrm{m}^3}]$ the density, ϕ $[-]$ the porosity, t [s] the time and F $[\frac{\mathrm{kg}}{\mathrm{m}^3 \cdot \mathrm{s}}]$ the sink/source term of the mixture.

The state variables of the NAPL vapor as well as the gas are assumed to be related by the ideal gas equation of state. Therefore, the pressure and density of the mixture satisfy the ideal gas equation of state in the form

$$\rho = p\frac{M}{RT}, \; M = \left(\frac{X_{\mathrm{n}}}{M_{\mathrm{n}}} + \frac{X_{\mathrm{g}}}{M_{\mathrm{g}}}\right)^{-1}, \tag{3}$$

where M_{n} and M_{g} $[\frac{\mathrm{kg}}{\mathrm{mol}}]$ are the molar weights of the NAPL vapor and gas, respectively; X_{n} and X_{g} $[-]$ the mass fractions of the NAPL vapor and gas ($X_{\mathrm{n}} + X_{\mathrm{g}} = 1$), respectively, in the mixture. R $[\frac{\mathrm{J}}{\mathrm{K \cdot mol}}]$ denotes the gas constant and T [K] the thermodynamic temperature.

Equation (3) will be used in the form

$$\rho = p\frac{M_{\mathrm{g}}}{RT}\frac{1}{1 + X_{\mathrm{n}}\left(\frac{M_{\mathrm{g}}}{M_{\mathrm{n}}} - 1\right)}. \tag{4}$$

Carrying out the time differentiation in (2), we get the following equation for the unknown pressure p and mass fraction X_{n}

$$\phi\frac{\partial \rho}{\partial p}\frac{\partial p}{\partial t} + \phi\frac{\partial \rho}{\partial X_{\mathrm{n}}}\frac{\partial X_{\mathrm{n}}}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = F, \tag{5}$$

where ρ is defined by (4) and $\boldsymbol{u}$ by (1).

The NAPL vapor transport within the mixture is assumed to be governed by the continuity equation in the form

$$\phi\frac{\partial \left(X_{\mathrm{n}}\rho\right)}{\partial t} + \nabla \cdot \left(X_{\mathrm{n}}\rho\,\boldsymbol{u} - D\rho\nabla X_{\mathrm{n}}\right) = R_{\mathrm{n}},$$

where D $[\frac{\mathrm{m}^2}{\mathrm{s}}]$ denotes the diffusion coefficient and R_n $[\frac{\mathrm{kg}}{\mathrm{m}^3\cdot\mathrm{s}}]$ the sink/source term of the NAPL vapor. Carrying out the time differentiation and substituting for the derivative of the density from (2), we obtain

$$\phi\rho\frac{\partial X_\mathrm{n}}{\partial t} + \nabla\cdot(X_\mathrm{n}\rho\,\boldsymbol{u} - D\rho\nabla X_\mathrm{n}) - X_\mathrm{n}\nabla\cdot(\rho\boldsymbol{u}) = R_\mathrm{n} - FX_\mathrm{n}. \tag{6}$$

Again, ρ is defined by (4) and $\boldsymbol{u}$ by (1).

The mass concentration c_n $[\frac{\mathrm{kg}}{\mathrm{m}^3}]$ of the NAPL vapor and mass fraction X_n are related by the equation $c_\mathrm{n} = X_\mathrm{n}\rho$.

Equations (5) and (6) are considered for $t \in I = [t_\mathrm{ini}, t_\mathrm{fin}]$, and they are subject to the initial conditions

$$p(x, t_\mathrm{ini}) = p_\mathrm{ini}(x),\ x \in \overline{\Omega}, \tag{7}$$

$$X_\mathrm{n}(x, t_\mathrm{ini}) = X_\mathrm{n,ini}(x),\ x \in \overline{\Omega} \tag{8}$$

and boundary conditions

$$p|_{\Gamma_\mathrm{p,Dir}} = p_\mathrm{p,Dir} \text{ and } (\rho\boldsymbol{u})|_{\Gamma_\mathrm{p,Neu}}\cdot\boldsymbol{n} = q_\mathrm{p,Neu}; \tag{9}$$

$$X_\mathrm{n}|_{\Gamma_\mathrm{X,Dir}} = X_{\mathrm{n}\mathrm{X,Dir}},\ (X_\mathrm{n}\rho\,\boldsymbol{u} - D\rho\nabla X_\mathrm{n})|_{\Gamma_\mathrm{X,Neu}}\cdot\boldsymbol{n} = q_\mathrm{X,Neu} \text{ and } \nabla X_\mathrm{n}|_{\Gamma_\mathrm{X,per}}\cdot\boldsymbol{n} = 0; \tag{10}$$

where $\Gamma_\mathrm{p,Dir}\cup\Gamma_\mathrm{p,Neu} = \partial\Omega$ and $\Gamma_\mathrm{p,Dir}\cap\Gamma_\mathrm{p,Neu} = \emptyset$; $\Gamma_\mathrm{X,Dir}\cup\Gamma_\mathrm{X,Neu}\cup\Gamma_\mathrm{X,per} = \partial\Omega$, $\Gamma_\mathrm{X,Neu}\subset\Gamma_\mathrm{p,Neu}$, and $\Gamma_\mathrm{X,Dir}$, $\Gamma_\mathrm{X,Neu}$, $\Gamma_\mathrm{X,per}$ are pairwise disjoint. The symbol $\boldsymbol{n}$ stands for the unit outward normal with respect to the boundary.

In this contribution, the points in Ω are denoted by (x, y) if the spatial coordinates need to be distinguished (g_1 and g_2 correspond to x and y, respectively); otherwise, they are denoted simply by $x \in \Omega$.

3 Numerical Solution

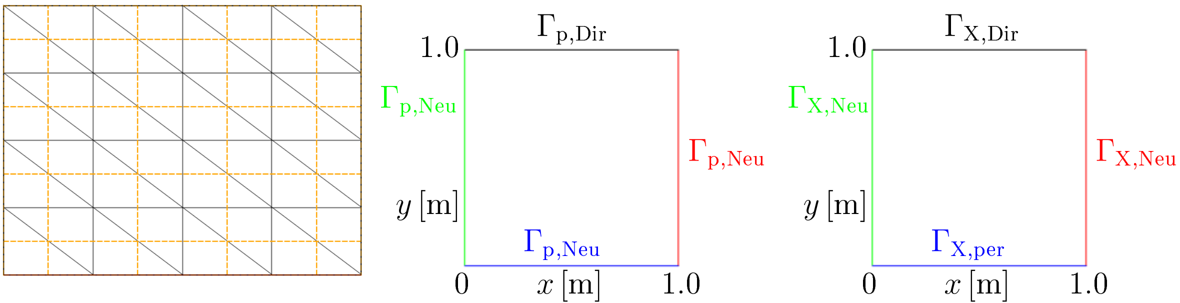
In order to solve problem (5)–(10), the author derived two different numerical schemes based on the finite volume method, the explicit and semi-implicit one. In this contribution, however, only the second one is discussed. Deriving it, we follow the ideas in [4] and [5].

The unknown functions p and X_n are approximated employing the classical finite element space based on the linear Lagrange elements ([2]), where the domain Ω is covered by the triangulation $\mathcal{T} = \{T^e\}_{e=1}^{N_\mathcal{T}}$ depicted in Figure 1a, where $N_\mathcal{T}$ is the number of triangles in $\mathcal{T}$. Each vertex x_i of the triangulation is associated with the basis function φ_i. Further, we use the node-centered dual mesh of finite volumes $\mathcal{V} = \{V_i\}_{i=1}^{N_\mathcal{V}}$ based on the Voroni diagrams ([4], [5] and [8]), where $N_\mathcal{V}$ denotes the number of nodes in $\mathcal{T}$. This mesh will be described later on. Finally, the time interval I is divided by a strictly increasing sequence $(t_n)_{n=0}^{N_t}$, where $t_0 = t_\mathrm{ini}$ and $t_{N_t} = t_\mathrm{fin}$.

We shall use the following notation:

- $\mathcal{X} = \{x_i\}_{i=1}^{N_\mathcal{V}}$ is the set of all vertices in the triangulation $\mathcal{T}$;

- $\Lambda^e = \{i | x_i \in T^e\}$;
- $\Lambda_i = \{j | (\exists T^e \in \mathcal{T})(x_i \in \Lambda^e \wedge x_j \in \Lambda^e)\} \setminus \{i\}$;
- $\Lambda_i^e = \Lambda^e \cap \Lambda_i$;
- $\Lambda_i^b = \Lambda_i \cap \{j | x_j \in \partial\Omega\}$;
- $\Lambda_{i,j} = \{e | i \in \Lambda^e \wedge j \in \Lambda^e\}$;
- $\Lambda_i^n = \{e | x_i \in \Lambda^e\}$;
- $x_{i,j}$ is the midpoint of the line segment connecting the vertices x_i and x_j;
- x_e is the circumcenter of the triangle T^e;
- $\Gamma_{i,j}^e$ is the line segment connecting the vertices x_e and $x_{i,j}$;
- $\Gamma_{i,j}^b$ is the line segment connecting the boundary vertices x_i and $x_{i,j}$;
- $\Gamma_i = \bigcup_{j \in \Lambda_i} \bigcup_{e \in \Lambda_{i,j}} \Gamma_{i,j}^e$;
- $\Gamma_i^b = \bigcup_{j \in \Lambda_i^b} \Gamma_{i,j}^b$ for $x_i \in \partial\Omega$;
- $x_{i,j}^e$ is the midpoint of $\Gamma_{i,j}^e$;
- $x_{i,j}^b$ is the midpoint of $\Gamma_{i,j}^b$;
- $\Lambda_{\mathrm{p,Neu},i}^b = \left\{j \in \Lambda_i^b | x_{i,j}^b \in \Gamma_{\mathrm{p,Neu}}\right\}$;
- $\Lambda_{\mathrm{X,Neu},i}^b = \left\{j \in \Lambda_i^b | x_{i,j}^b \in \Gamma_{\mathrm{X,Neu}}\right\}$;
- $V_i^e = V_i \cap T^e$;
- $f(x_i) = f_i$, where the possible time coordinate is omitted;
- $f(x_{i,j}) = f_{i,j}$, where the possible time coordinate is omitted;
- $f(x_{i,j}^e) = f_{i,j}^e$, where the possible time coordinate is omitted;
- $f(x_{i,j}^b) = f_{i,j}^b$, where the possible time coordinate is omitted;
- f_e is the constant value of f on $T^e \in \mathcal{T}$;
- f_B^e is the value of f in the barycenter of $T^e \in \mathcal{T}$, where the possible time coordinate is omitted;
- $f(t_n) = f^n$, where $f = f(t)$;
- $X_{i,j}^e$ denotes the special upwind term defined in Section 3.3;
- $\tau = \frac{t_{\mathrm{fin}} - t_{\mathrm{ini}}}{N_t}$ if the sequence $(t_n)_{n=0}^{N_t}$ is arithmetic.


$\Gamma_{\mathrm{p,Dir}}$
1.0
$\Gamma_{\mathrm{p,Neu}}$
$\Gamma_{\mathrm{p,Neu}}$
y [m]
$\Gamma_{\mathrm{p,Neu}}$
0
x [m]
1.0
$\Gamma_{\mathrm{X,Dir}}$
1.0
$\Gamma_{\mathrm{X,Neu}}$
$\Gamma_{\mathrm{X,Neu}}$
y [m]
$\Gamma_{\mathrm{X,per}}$
0
x [m]
1.0

(a) Primary (solid) and dual (dashed) mesh.

(b) Boundary conditions for p.

(c) Boundary conditions for X_n.

Figure 1: Meshes and boundary conditions.

The preceding notation is used for scalar (f) as well as for vector-valued ($\boldsymbol{f}$) functions.

The finite volume V_i associated with the vertex x_i is defined as the open set surrounded by the piecewise linear curve Γ_i (i.e., $\partial V_i = \Gamma_i$) for $x_i \notin \partial\Omega$ and by the piecewise linear curve $\Gamma_i \cup \Gamma_i^b$ (i.e., $\partial V_i = \Gamma_i \cup \Gamma_i^b$) for $x_i \in \partial\Omega$. The dual mesh of finite volumes is depicted in Figure 1a.

The numerical schemes are based on the following local mass balance equations derived by integrating equations (5) and (6) over the volume V_i and applying the Green formula:

$$\int_{V_i} \phi \frac{\partial \rho}{\partial p}\frac{\partial p}{\partial t} + \int_{V_i} \phi \frac{\partial \rho}{\partial X_{\mathrm{n}}}\frac{\partial X_{\mathrm{n}}}{\partial t} + \int_{\partial V_i} \rho \boldsymbol{u} \cdot \boldsymbol{n} = \int_{V_i} F, \tag{11}$$

$$\int_{V_i} \phi\rho \frac{\partial X_{\mathrm{n}}}{\partial t} - \int_{\partial V_i} X_{\mathrm{n}} \rho \boldsymbol{u} \cdot \boldsymbol{n} + \int_{V_i} \rho \boldsymbol{u} \cdot \nabla X_{\mathrm{n}} + \int_{\partial V_i} (X_{\mathrm{n}} \rho \boldsymbol{u} - \rho D \nabla X_{\mathrm{n}}) \cdot \boldsymbol{n} = \int_{V_i} R_{\mathrm{n}} - \int_{V_i} F X_{\mathrm{n}}. \tag{12}$$

In order to compute the integrals in (11) and (12), we substitute the functions $\frac{\partial \rho}{\partial p}$, p, $\frac{\partial \rho}{\partial X_{\mathrm{n}}}$, X_{n}, ρ, $\boldsymbol{u}$, F and R_{n} by approximations from our finite element space (except for $\boldsymbol{u}$, they are standard, e.g., $p = \sum_i p_i \varphi_i$) and employ the following approximation techniques (the possible time coordinate is omitted):

- $\int_{V_i} f(x)\,\mathrm{d}x \doteq \sum_{e \in \Lambda_i^n} |V_i^e|\, f_i$, where $|V_i^e|$ denotes the area of V_i^e;
- $\int_{V_i} \boldsymbol{f}(x) \cdot \boldsymbol{g}(x)\,\mathrm{d}x \doteq \sum_{e \in \Lambda_i^n} |V_i^e|\, \boldsymbol{f}_B^e \cdot \boldsymbol{g}_B^e$;
- $\int_{\Gamma_i} \boldsymbol{f}(x) \cdot \boldsymbol{n}\,\mathrm{d}x \doteq \sum_{j \in \Lambda_i} \sum_{e \in \Lambda_{i,j}^e} \left|\Gamma_{i,j}^e\right| \boldsymbol{f}_{i,j}^e \cdot \boldsymbol{n}_{i,j}^e$, where $\left|\Gamma_{i,j}^e\right|$ denotes the length of the line segment $\Gamma_{i,j}^e$ and $\boldsymbol{n}_{i,j}^e$ the unit outward normal with respect to $\Gamma_{i,j}^e$;
- $\int_{\Gamma_i^b} \boldsymbol{f}(x) \cdot \boldsymbol{n}\,\mathrm{d}x \doteq \sum_{j \in \Lambda_i^b} \left|\Gamma_{i,j}^b\right| \boldsymbol{f}_{i,j}^b \cdot \boldsymbol{n}_{i,j}^b$ for $x_i \in \partial\Omega$, where $\left|\Gamma_{i,j}^b\right|$ denotes the length of the line segment $\Gamma_{i,j}^b$ and $\boldsymbol{n}_{i,j}^b$ the unit outward normal with respect to $\Gamma_{i,j}^b$.

The approximation of the Darcy velocity $\boldsymbol{u}$ and the function X_{n} in the first part of the last integrand on the left-hand side of equation (12) requires more careful treatment; it will be discussed further. We also assume that the permeability tensor and porosity take constant values $\boldsymbol{k}_e$ and ϕ_e, respectively, on each triangle $T^e \in \mathcal{T}$.

Now, we can put together the aforementioned numerical scheme.

3.1 Semi-Implicit Scheme

If equations (11) and (12) are considered at time t_{n+1}, the time derivatives are approximated by the backward finite differences, some terms (they are chosen heuristically) from time t_{n+1} are approximated at time t_n in order to get a system of linear algebraic equations for the unknown values p_i^{n+1} and $X_{\mathrm{n},i}^{n+1}$, the approximation techniques above mentioned are applied, and the boundary conditions (9) and (10) are considered, we obtain the following system of equations for $n = 0, 1, \ldots, N_t - 1$:

$$\begin{aligned}\sum_{e\in\Lambda_i^n} |V_i^e|\, \phi_e \left(\frac{\partial\rho}{\partial p}\right)_i^n \frac{p_i^{n+1} - p_i^n}{\tau} = &-\sum_{e\in\Lambda_i^n} |V_i^e|\, \phi_e \left(\frac{\partial\rho}{\partial X_\mathrm{n}}\right)_i^n \frac{X_{\mathrm{n},i}^{n+1} - X_{\mathrm{n},i}^n}{\tau} + \sum_{e\in\Lambda_i^n} |V_i^e|\, F_i^{n+1} \\ &- \sum_{j\in\Lambda_i}\sum_{e\in\Lambda_{i,j}} \left|\, \Gamma_{i,j}^e \,\right| \rho_{i,j}^{e,n} \boldsymbol{u}_{i,j}^{e,n+1} \cdot \boldsymbol{n}_{i,j}^e - \sum_{j\in\Lambda^b_{\mathrm{p,Neu},i}} \left|\, \Gamma^b_{i,j} \,\right| q^{b,n+1}_{\mathrm{p,Neu},i,j}\end{aligned} \tag{13}$$

for $i = 1, 2, \ldots, N_\mathcal{V}$, $x_i \notin \Gamma_\mathrm{p,Dir}$;

$$p_i^{n+1} = p^{n+1}_{\mathrm{p,Dir},i} \text{ for } i = 1, 2, \ldots, N_\mathcal{V}, x_i \in \Gamma_\mathrm{p,Dir}; \tag{14}$$

$$\begin{aligned}\sum_{e\in\Lambda_i^n} |V_i^e|\, \phi_e \rho_i^n \frac{X_{\mathrm{n},i}^{n+1} - X_{\mathrm{n},i}^n}{\tau} = &-\sum_{e\in\Lambda_i^n} |V_i^e|\, F_i^{n+1} X_{\mathrm{n},i}^{n+1} + \sum_{j\in\Lambda_i}\sum_{e\in\Lambda_{i,j}} \left|\, \Gamma_{i,j}^e \,\right| \rho_{i,j}^{e,n} X_{\mathrm{n},i,j}^{e,n+1} \boldsymbol{u}_{i,j}^{e,n} \cdot \boldsymbol{n}_{i,j}^e \\ &+ \sum_{j\in\Lambda^b_{\mathrm{X,Neu},i}} \left|\, \Gamma^b_{i,j} \,\right| X^{b,n+1}_{\mathrm{n},i,j} q^{b,n+1}_{\mathrm{p,Neu},i,j} - \sum_{e\in\Lambda_i^n} |V_i^e|\, \rho_i^n \boldsymbol{u}_B^{e,n} \cdot \nabla X_\mathrm{n}^{e,n+1} \\ &- \sum_{j\in\Lambda_i}\sum_{e\in\Lambda_{i,j}} \left|\, \Gamma_{i,j}^e \,\right| \rho_{i,j}^{e,n} \boldsymbol{n}_{i,j}^e \cdot \left(X_{i,j}^{e,n+1} \boldsymbol{u}_{i,j}^{e,n} - D_{i,j}^e \nabla X_{\mathrm{n},i,j}^{e,n+1} \right) \\ &- \sum_{j\in\Lambda^b_{\mathrm{X,Neu},i}} \left|\, \Gamma^b_{i,j} \,\right| q^{b,n+1}_{\mathrm{X,Neu},i,j} + \sum_{e\in\Lambda_i^n} |V_i^e|\, R_{\mathrm{n},i}^{n+1}\end{aligned} \tag{15}$$

for $i = 1, 2, \ldots, N_\mathcal{V}$, $x_i \notin \Gamma_\mathrm{X,Dir}$;

$$X_{\mathrm{n},i}^{n+1} = X^{n+1}_{\mathrm{nX,Dir},i} \text{ for } i = 1, 2, \ldots, N_\mathcal{V}, x_i \in \Gamma_\mathrm{X,Dir}. \tag{16}$$

The upwind term $X_{i,j}^{e,n+1}$ and the approximation of the velocity $\boldsymbol{u}$ are defined in Sections 3.3 and 3.2, respectively. Computing the term $\boldsymbol{u}_{i,j}^{e,n+1}$, $\rho_{i,j}^{e,n+1}$ is approximated by $p_{i,j}^{e,n+1}\left(\frac{\partial\rho}{\partial p}\right)_{i,j}^{e,n}$. Remark that the sums over the boundary nodes are correct because $\Gamma_\mathrm{X,Neu} \subset \Gamma_\mathrm{p,Neu}$ and $\Lambda^b_{\mathrm{p,Neu},i} = \Lambda^b_{\mathrm{X,Neu},i} = \emptyset$ for $x_i \notin \partial\Omega$.

The initial conditions are

$$p_i^0 = p_{\mathrm{ini},i},\ X_{\mathrm{n},i}^0 = X_{\mathrm{n,ini},i} \text{ for } i = 1, 2, \ldots, N_\mathcal{V}. \tag{17}$$

This system is solved by the DGESVX subroutine from LAPACK ([1]). This subroutine equilibrates the matrix of the system in order to reduce its condition number first, then solves the system via the LU decomposition, and finally applies the iterative refinement.

3.2 Approximation of Velocity u

The approximation of the Darcy velocity $\boldsymbol{u}$ should be carried out very carefully because it results in an additional numerical flux. If, for example, the pressure p and density ρ are both approximated in the aforementioned finite element space and substituted to formula (1), than, due to (3), the velocity which is nonzero under hydrostatic conditions is obtained.

In our numerical schemes, we employ the method ([3] and [6]) that approximates the velocity $\boldsymbol{u}$ on a triangle $T^e \in \mathcal{T}$ as

$$\boldsymbol{u}|_{T^e} = -\frac{1}{\mu}\boldsymbol{k}_e \left[\nabla\left(\tilde{p} + \tilde{h}\right) - \tilde{h}_x \begin{pmatrix}1\\0\end{pmatrix}\right], \tag{18}$$

where

$$h(t,x,y) = \int_{y_1}^{y} -g_2\tilde{\rho}(t,x,s)\mathrm{d}s \text{ and } h_x(t,x,y) = \int_{y_1}^{y} -g_2\frac{\partial\tilde{\rho}}{\partial x}(t,x,s)\mathrm{d}s. \tag{19}$$

In these formulas, the tilde denotes the standard finite element approximation of the function, and y_1 is the y-coordinate of the vertex that corresponds to the point $(0,0)$ if the triangle T^e is mapped to the reference triangle with the vertices $(0,0)$, $(0,1)$ and $(1,0)$ and the local coordinates (ξ,η) by a map defined as the inversion of the mapping

$$x = x(\xi,\eta) = \sum_{i=1}^{3} x_i\varphi_i(\xi,\eta) \text{ and } y = y(\xi,\eta) = \sum_{i=1}^{3} y_i\varphi_i(\xi,\eta),$$

where (x_i,y_i) is the vertex of T^e, and φ_i is the basis function associated with the i-th vertex of the reference triangle.

Using this approximation, we also assume that $g_1 = 0$.

3.3 Types of Upwind

If the NAPL vapor spreads mainly by convection, the term X_n in $X_\mathrm{n}\rho\,\boldsymbol{u}$ in equation (6) requires a special treatment in order to prevent the numerical solution from oscillating non-physically. This treatment is carried out by means of a suitable definition of the term $X_{i,j}^{e,n+1}$ in equation (15). We test the options mentioned in [4] and [5] (the exponential upwind was modified by the author), where the number $\gamma_{i,j}^e$ is defined as

$$\gamma_{i,j}^e = D_{i,j}^e \nabla\varphi_j(x_{i,j}^e) \cdot \sum_{k\in\Lambda_i^e} \mid \Gamma_{i,k}^e \mid \boldsymbol{n}_{i,k}^e,$$

and

$$P_{i,j}^e = \frac{\mid \Gamma_{i,j}^e \mid \boldsymbol{u}_{i,j}^e \cdot \boldsymbol{n}_{i,j}^e}{\gamma_{i,j}^e}$$

is an analogy of local Peclet number. Here, we list only the options used in Section 4. All of the values in the definitions of $\gamma_{i,j}^e$ and $P_{i,j}^e$ are from time t_n.

- Full upwind

$$X^e_{i,j} = \begin{cases} X_{\mathrm{n},i}, & \boldsymbol{u}^e_{i,j} \cdot \boldsymbol{n}^e_{i,j} \geq 0 \\ X_{\mathrm{n},j}, & \boldsymbol{u}^e_{i,j} \cdot \boldsymbol{n}^e_{i,j} < 0 \end{cases} .$$

- Exponential upwind

  For $\gamma^e_{i,j} > 0$, we define

$$X^e_{i,j} = X_{\mathrm{n},i}(1+\theta) - X_{\mathrm{n},j}\theta,$$

  where

$$\theta = \begin{cases} -\frac{\omega(P^e_{i,j})}{1+P^e_{i,j}\omega(P^e_{i,j})}, & \left| P^e_{i,j} \right| \leq 10^{-5} \\ -\frac{1}{P^e_{i,j}} + \frac{1}{\exp(P^e_{i,j})-1}, & \left| P^e_{i,j} \right| > 10^{-5} \end{cases}$$

  and

$$\omega(x) = \frac{1}{2!} + \frac{x}{3!} + \frac{x^2}{4!} + \frac{x^3}{5!};$$

  otherwise, the full upwind is used.

4 Numerical Results

In this section, the results of one of our numerical simulations are presented. The simulation was performed by the scheme described in Section 3, where the term $X^{e,n+1}_{i,j}$ in equation (15) was defined by the exponential upwind option in Section 3.3.

The domain Ω was of the form $(0,1) \times (0,1)$, where the units are [m], and there were 41 nodes on each side of the spatial mesh (see Figure 1a). The time step was $\tau = 0.01$. The permeability tensor $\boldsymbol{k}$ was always a scalar multiple of the identity, i.e., $\boldsymbol{k} = \tilde{k}\boldsymbol{I}$, $\tilde{k}$ is spatial dependent. The values of the physical constants used are listed in Table 1. The values of porosity and permeability are from [10].

The following initial and boundary conditions and $\tilde{k}$ and ϕ were considered (the division of $\partial\Omega$ is depicted in Figure 1):

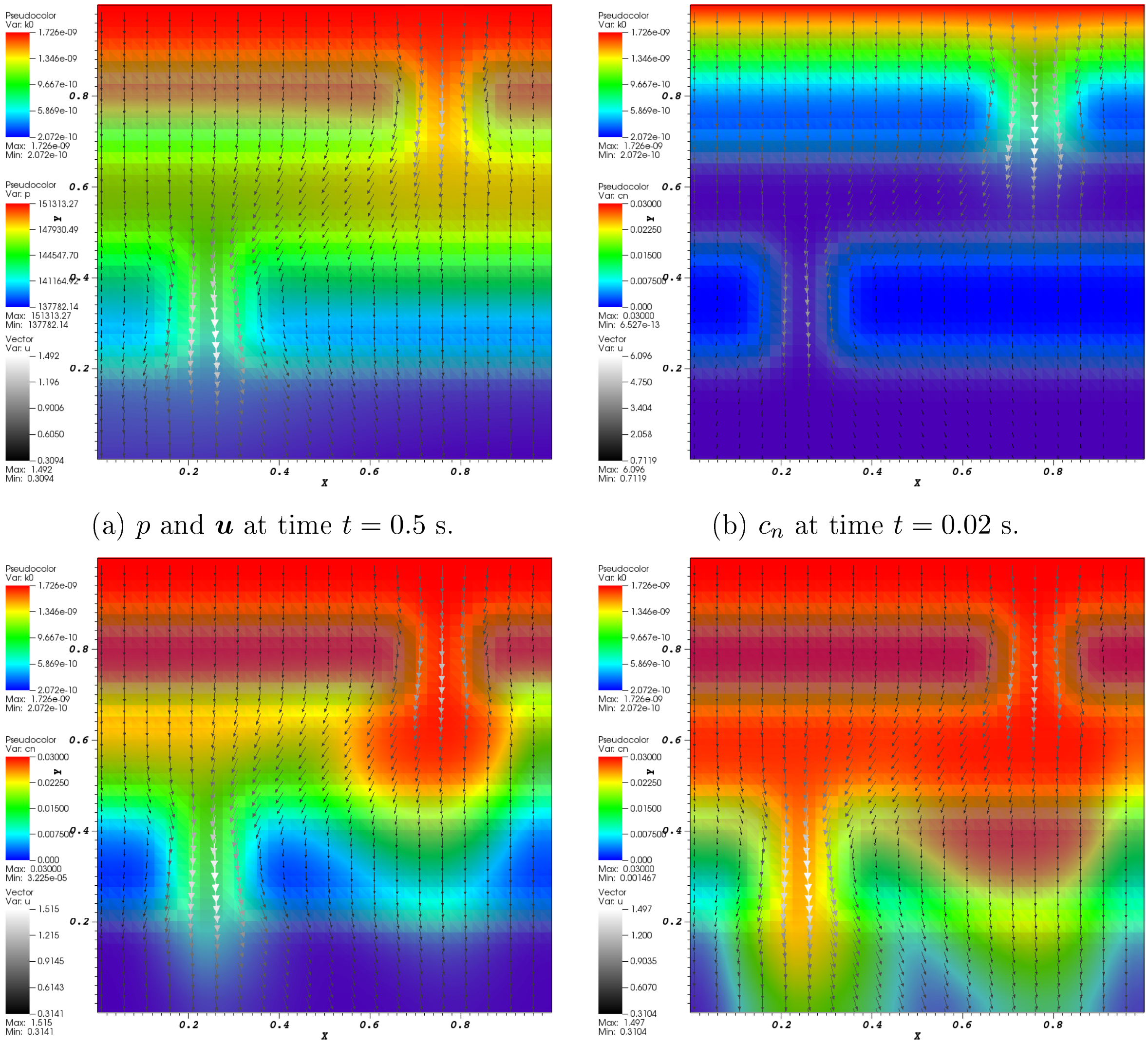
- $p_{\mathrm{p,Dir}}(t,x,y) = p_{\mathrm{ref}} + 5 \cdot 10^4$;
- $q_{\mathrm{p,Neu}}(t,x,y) = \begin{cases} 1, & \text{if } y = 0 \\ 0, & \text{otherwise} \end{cases}$;
- $X_{\mathrm{nX,Dir}}$ is computed from $c_{\mathrm{n,ref}}$ and $p_{\mathrm{p,Dir}}$;
- $q_{\mathrm{X,Neu}}(t,x,y) = 0$;
- $p_{\mathrm{ini}}(x,y) = p_{\mathrm{ref}} \exp\left(\frac{M_{\mathrm{g}} g_2 y}{RT}\right)$;
- $X_{\mathrm{n,ini}}(x,y) = 0$;
- values of $\tilde{k} = k_0$ are depicted in the background of Figures 2a–2d;
- $\phi(x,y) = \frac{\phi_2 - \phi_1}{\tilde{k}_2 - \tilde{k}_1}(\tilde{k} - \tilde{k}_1) + \phi_1$.

The numerical results are shown in Figures 2a–2d. We can see that the NAPL vapor really spreads like a wave.

parameter	value	unit
μ	$1.81 \cdot 10^{-5}$	$\mathrm{kg} \cdot \mathrm{m}^{-1} \cdot \mathrm{s}^{-1}$
M_{g}	0.02896	$\mathrm{kg} \cdot \mathrm{mol}^{-1}$
M_{n}	0.13139	$\mathrm{kg} \cdot \mathrm{mol}^{-1}$
R	8.3144621	$\mathrm{J} \cdot \mathrm{K}^{-1} \cdot \mathrm{mol}^{-1}$
T	288.15	K
g_1	0	$\mathrm{m} \cdot \mathrm{s}^{-2}$
g_2	-9.81	$\mathrm{m} \cdot \mathrm{s}^{-2}$
F	0	$\mathrm{kg} \cdot \mathrm{m}^{-3} \cdot \mathrm{s}^{-1}$

parameter	value	unit
p_{ref}	101325	Pa
R_{n}	0	$\mathrm{kg} \cdot \mathrm{m}^{-3} \cdot \mathrm{s}^{-1}$
$c_{\mathrm{n,ref}}$	$3 \cdot 10^{-2}$	$\mathrm{mol} \cdot \mathrm{m}^{-3}$
ϕ_1	0.339	–
ϕ_2	0.433	–
$\tilde{k}_1$	$1.726 \cdot 10^{-9}$	m^2
$\tilde{k}_2$	$2.012 \cdot 10^{-10}$	m^2
D	10^{-5}	$\mathrm{m}^2 \cdot \mathrm{s}^{-1}$

Table 1: Values of physical parameters.



(a) p and $\boldsymbol{u}$ at time $t = 0.5$ s.

(b) c_n at time $t = 0.02$ s.

(c) c_n at time $t = 0.2$ s.

(d) c_n at time $t = 0.4$ s.

Figure 2: Numerical results. The arrows indicate the direction and magnitude of u. The shades in the background of the figures are the values of $\tilde{k}$ ('k0').

5 Conclusions

The numerical scheme derived in Section 3 seems to solve the governing equations without producing non-physical oscillations in the NAPL vapor concentration. Therefore, the results can be compared with experimental data, and the approach on which the scheme is based may be employed on more complex equations.

References

[1] Manual pages for lapack, (November 2013).

[2] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods.* Springer, 3rd edition, (2008).

[3] P. Frolkovič. *Consistent velocity approximation for density driven flow and transport.* In 'Advanced Computational Methods in Engineering, Part 2: Contributed papers', 603–611, Maastricht, (1998). Shaker Publishing.

[4] P. Frolkovič. *Discretization in d3f - A Simulator for Density-Driven Flow.* Gesellschaft fuer Anlagen-und Reaktorsicherheit (mbH), Braunschweig, (1998).

[5] P. Frolkovič. *Maximum principle and local mass balance for numerical solutions of transport equation coupled with variable density flow.* Acta Mathematica Universitatis Comenianae **67** (1998), 137–157.

[6] P. Frolkovič and P. Knabner. *Consistent velocity approximations in finite element or volume discretizations of density driven flow.* In 'Computational Methods in WaterResources XI', A. A. Aldama, (ed.), Computational Mechanics Publication (19996), 93–100.

[7] V. Giovangigli. *Multicomponent Flow Modeling.* Birkhäuser Boston, 1st edition, (1999).

[8] T. Ikeda. *Maximum Principle in Finite Element Models for Convection-Diffusion Phenomena.* North-Holland Publishing Company, 1st edition, (1983).

[9] N. I. Kolev. *Multiphase Flow Dynamics 1.* Springer-Verlag Berlin Heidelberg, 2nd edition, (2005).

[10] M. H. Schroth, S. J. Ahearn, J. S. Selker, and J. D. Istok. *Characterization of miller-similar silica sands for laboratory hydrologic studies.* Soil Science Society of America Journal **60** (1996), 1331–1339.

New Approach to Electricity Markets: Best Response of Producer

Miroslav Pištěk

3rd year of PGS, email: miroslav.pistek@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Didier Aussel, Laboratoire PROMES, Université de Perpignan

Jiří Outrata, Department of Decision-Making Theory
Institute of Information Theory and Automation, AS CR

Abstract. A new way to treat the problem of electricity markets analytically is proposed here. We consider several electricity producers and a central authority of an independent system operator (ISO). We model such conflict situation in a standard way as a bi-level non-cooperative Nash game, where ISO is a leader player and producers are considered as followers. Using analytical formula for a solution to the ISO problem [1], we provide a detailed analysis of the problem of a producer. We conclude by providing the conditions for existence of the best response, which is then described by an explicit formula. We note that the topology of the electricity dispatch network is not considered at the moment.

Keywords: electricity markets, bi-level Nash games

Abstrakt. V této práci je představen nový přístup k modelování trhu s elektřinou. Uvažujeme několik producentů elektřiny a nezávislého systémového operátora (ISO). Tuto konfliktní situaci modelujeme standardně jako dvouúrovňovou nekooperativní Nashovu hru, kde ISO je uvažován jako lídr a producenti jako jeho následovníci. S použitím analyického řešení ISO problému [1], jsme provedli detailni rozbor problému producenta. Získali jsme podmínky pro existenci jeho optimální akce, která je pak popsána explicitním vzorcem. Poznamenáváme, že topologie elektrické rozvodné sítě není zde není uvažována.

Klíčová slova: trhy s elektřinou, dvouúrovňové Nashovy hry

1 Introduction

The modelling of the electricity networks is a very current topic, since in the last two decades they were privatized in many countries. The ultimate aim of such movement was to enhance the effectiveness of electricity production and distribution, and so naturally also electricity markets were founded, typically at the national level. Later, these markets were consolidated; soon there will be just one pan-European electricity market. Moreover, also an operational requirements of the so-called *smart grids*, i.e., electricity dispatch networks with non-stable wind and solar power plants of various scales, are newly considered. Thus, many practical and at the same time scientifically interesting questions arose within this area.

Further, we consider only the electricity market itself, omitting all the problems concerning electricity dispatch network. We may observe that such market can not run in the same way as, for instance, stock market. Indeed, electricity is a special kind of commodity which is hard to store effectively. Thus, either all the produced electricity is consumed at the very same moment, or we undergo high economic losses (either by overproduction, or by possible black-out). On that account market has to be regulated by an *Independent System Operator* (denoted by ISO in the sequel), which is typically a state company. Then, all the electricity producers and consumers participating in the market have to obey the decisions of ISO. This fact is the very novelty when modelling such market and has important mathematical consequences.

From the point of view of producers and consumers, the electricity market may be modelled as a non-cooperative Nash game. However, the presence of ISO makes this problem much more complicated. In general, such bi-level problem is a special kind of Equilibrium Problem with Equilibrium Constraint (EPEC), where the lower-level leader problem, i.e., ISO problem in our case, is considered as an equilibrium constraint for the upper-level problem, which is a Nash game of producers and consumers [5]. Since this explicit dependence on the solution of ISO problem does not preserve any convexity, we can not use the classical Nash theorem for existence of solution to EPEC in general. Then, some more assumptions are needed [2], or only a more specific setting with just two players may be considered [3].

In this article, we avoid the general problem of EPEC, and analyse the problem of the electricity market directly. We have already shown [1] that under a very natural assumptions the ISO problem possesses one solution on general. In this article, we substitute this solution of lower level problem directly into the upper level problem, avoiding all these previously mentioned difficulties. Finally, a discussion of the obtained results is provided. Further, we denote

* $D > 0$ the overall energy demand.

* $\mathcal{N}$ be the set of producers (N being its cardinal, $N > 1$).

* $q_i \geq 0$ represents the non-negative production of i-th producer, $i \in \mathcal{N}$

* $a_i, b_i \geq 0$ are coefficients of i-th producer bid function $a_i q_i + b_i q_i^2$

For $q \in \mathbb{R}_+^N$ we denote by $q_{-i} \in \mathbb{R}_+^{N-1}$ vector $q_{-i} = (q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_N)$.

2 ISO's Problem

Based on the bids of all producers, the aim of the ISO is to minimize the total cost of production, taking into account that the demand has to be satisfied. Each producer provides to the ISO a quadratic bid function $a_i q_i + b_i q_i^2$ given by non-negative parameters $a_i, b_i \geq 0$. This bid cost function may differ from the real cost function of producer i. The ISO, knowing the bid vectors $a = (a_1, \cdots, a_N) \in \mathbb{R}_+^N$ and $b = (b_1, \cdots, b_N) \in \mathbb{R}_+^N$ provided by producers, computes $q = (q_1, \ldots, q_N) \in \mathbb{R}_+^N$ in order to minimize the total generation cost, that is to solve the following optimization problem

$$\text{ISO(a,b)} \qquad \begin{array}{ll} \min\limits_{q} & \sum\limits_{i\in\mathcal{N}}(a_i q_i + b_i q_i^2) \\ \text{s.t.} & \begin{cases} q_i \geq 0, \ \forall i \in \mathcal{N} \\ \sum\limits_{i\in\mathcal{N}} q_i = D \end{cases} \end{array}$$

for positive overall demand $D > 0$. Then, it is a well-known fact that this problems admits at least one solution. Nevertheless, the market problem can be ill-posed if the solution set of ISO(a,b) contains more than one point, see e.g. [4]. In [2, 3] the uniqueness of the response of the ISO(a,b) comes from the hypothesis that producers are bidding true quadratic function with $b_i > 0$, thus implying the strict convexity of the objective function of ISO(a,b) problem. Since in our work, we allow linear bid of a producer, even eventually of all of them, an additional assumption is needed to guarantee uniqueness of solution of ISO(a,b) problem. On that account, we add *equity property* assumption

$$\text{(H)} \qquad (a_i, b_i) = (a_j, b_j) \Longrightarrow q_i = q_j$$

which is supposed to hold for all $i, j \in \mathcal{N}$. This assumption acctualy formalize that ISO makes no difference among producers. Let us remark that the optimization problem ISO(a,b) assuming (H) is as follows

$$\text{ISO(a,b)+(H)} \qquad \begin{array}{ll} \min\limits_{q} & \sum\limits_{i\in\mathcal{N}}(a_i q_i + b_i q_i^2) \\ \text{s.t.} & \begin{cases} q_i \geq 0, \ \forall i \in \mathcal{N} \\ (a_i, b_i) = (a_j, b_j) \Rightarrow q_i = q_j, \forall i, j \in \mathcal{N} \\ \sum\limits_{i\in\mathcal{N}} q_i = D \end{cases} \end{array}$$

and therefore all the following results concerns this formulation of the problem, even though we will speak about the problem ISO(a,b) and hypotesis (H) separately.

To analyse this problem further, we introduce index set mapping $\mathcal{N}_a(\lambda)$

$$\mathcal{N}_a(\lambda) = \{i \in \mathcal{N} | a_i < \lambda\} \subset \mathcal{N}.$$

This set represents, for a given price λ, the subset of producers being "in the money". Then we define several critical parameters of ISO(a,b), namely a critical market price $\lambda^c(a, b)$, a critical value of the overall demand $D^c(a, b)$, and a set of producers bidding critical (linear) bids $\mathcal{N}^c(a, b) \subset \mathcal{N}$

$$\begin{aligned} \lambda^c(a,b) &= \min_{i\in\mathcal{N}, b_i=0} a_i \\ \mathcal{N}^c(a,b) &= \{i \in \mathcal{N} |\, a_i = \lambda^c(a,b), b_i = 0\} \qquad (1) \\ D^c(a,b) &= \sum_{i\in\mathcal{N}_a(\lambda^c(a,b))} \frac{\lambda^c(a,b) - a_i}{2b_i} \end{aligned}$$

For the case of $\mathcal{N}_a(\lambda^c(a,b)) = \emptyset$, i.e., $a_i \geq \lambda^c(a,b)$ for all $i \in \mathcal{N}$, we put $D^c(a,b) = 0$. If there is not any producer bidding linear function, i.e., we have $b_i > 0$ for all $i \in \mathcal{N}$, we set $\lambda^c(a,b) = D^c(a,b) = +\infty$. For the cardinality of $\mathcal{N}^c(a,b)$ we use the notation $N^c(a,b) = |\mathcal{N}^c(a,b)|$.

These critical parameters have clear economic meaning. First, $\lambda^c(a,b)$ denotes the minimum price such that at least one linearily bidding producer ($b_i = 0$) will participate in the market. Since such producer can provide arbitrary amount of electricity at this price, $\lambda^c(a,b)$ is also the highest possible price in the market. Then, $D^c(a,b)$ will be later identified with the overall amount of electricity produced by sub-critical producers, i.e., those participating in the market having $b_i > 0$, see the proof of Theorem 2.3. Finally, $\mathcal{N}^c(a,b)$ is the set of all the critical producers that may possibly participate in the market. Next, we denote $\lambda^m(a) = \min_{i\in\mathcal{N}} a_i$.

Remark 2.1. *(a) From the definition of $\lambda^c(a,b)$ we clearly have that $a_i < \lambda^c(a,b)$ immediately implies $b_i > 0$. This means that if the linear term of the bid of producer i is strictly smaller than the critical market price, then this producer is bidding quadratically.*

(b) We note that condition $D^c(a,b) = 0$ means that no sub-critical producer, i.e. producer bidding $b_i > 0$, will participate in the market, cf. the meaning of $D^c(a,b)$ discussed above. Moreover, this condition may be equivalently stated as $\lambda^m(a) = \lambda^c(a,b)$.

Next, we define $\Delta = \left\{(a,b,\lambda) \in \mathbb{R}_+^{2N+1} | \lambda^m(a) < \lambda \leq \lambda^c(a,b)\right\}$ (considering sharp inequality for the case of $\lambda^c(a,b) = +\infty$) and function $F : \Delta \to \mathbb{R}_+$ as

$$F(a,b,\lambda) = \sum_{i\in\mathcal{N}_a(\lambda)} \frac{\lambda - a_i}{2b_i}, \tag{2}$$

We note that for $\lambda > \lambda^c(a,b)$ formula (2) is ill-posed because there exists $i \in \mathcal{N}^c(a,b) \subset \mathcal{N}_a(\lambda)$ such that $b_i = 0$, and that by the definition of Δ we have $\mathcal{N}_a(\lambda) \neq \emptyset$.

Consider any $(a,b) \in \mathbb{R}_+^{2N}$ fixed. As an immediate consequence of the definition of F we have

$$\begin{aligned} \lim_{\lambda\to\lambda^m(a)} F(a,b,\lambda) &= 0, && , \\ \lim_{\lambda\to+\infty} F(a,b,\lambda) &= +\infty && \text{if} \quad \lambda^c(a,b) = +\infty, \\ F(a,b,\lambda^c(a,b)) &= D^c(a,b) && \text{if} \quad \lambda^c(a,b) < +\infty \end{aligned}$$

Moreover, for any $(a,b) \in \mathbb{R}_+^{2N}$ function $\lambda \to F(a,b,\lambda)$ is continuous and piece-wise linear on $[\lambda^m(a), \lambda^c(a,b)[$ and aditionally it possesses monotonicity property playing an important role in the sequel.

Lemma 2.2. *For any $(a,b) \in \mathbb{R}_+^{2N}$ function $\lambda \to F(a,b,\lambda)$ is strictly increasing.*

A technical proof of this Lemma is included in [1]. This lemma justifies the following definition of function $\lambda(a,b,D) : \mathbb{R}_+^{2N} \times]0,+\infty[\to \mathbb{R}_+$

$$\lambda(a,b,D) = \begin{cases} \lambda \in \mathbb{R}_+ \text{ s.t. } F(a,b,\lambda) = D \text{ if } D \in]0, D^c(a,b)[\\ \lambda^c(a,b) \text{ if } D \geq D^c(a,b) \end{cases} \tag{3}$$

For any $(a, b) \in \mathbb{R}_+^{2N}$ function $\lambda(a, b, D)$ is continuous and piece-wise linear in D owing to the same properties of $F(a, b, \lambda)$. Next, we state a convenient implicit formula for the unique solution $q(a, b, D)$ to the convex minimization problem ISO(a,b) assuming (H).

Moreover, in [1] we shown that for any fixed configuration of bids of producers $(a, b) \in \mathbb{R}_+^{2N}$, function $\lambda(a, b, D)$ assign to each demand $D > 0$ the respective market marginal price of the production.

Theorem 2.3. *Let $D > 0$, then for $(a, b) \in \mathbb{R}_+^{2N}$ such that $\lambda^c(a, b) > 0$, the regulator's problem ISO(a,b) admits a unique solution $q(a, b)$ obeying the equity property (H). Moreover, this optimal solution is given by*

$$q_i(a, b, D) = \begin{cases} \frac{\lambda - a_i}{2b_i} \text{ if } a_i < \lambda \\ \frac{D - D^c(a,b)}{N^c(a,b)} \text{ if } a_i = \lambda, b_i = 0 \\ 0 \text{ if } a_i > \lambda, \text{ or } a_i = \lambda, b_i > 0 \end{cases} \tag{4}$$

with $\lambda = \lambda(a, b, D)$ determined by (3).

This theorem was shown in a detail in [1]. The main idea is to use KKT system corresponding to ISO(a,b), and by a detailed analysis show that it possesses only one solution. In general, $\lambda(a, b, D)$ is not a smooth function, but we may compute several directional derivatives easily. This is our main technique to tackle the problem of a producer in the next section.

3 Producer's problem

In this section we stress the point of view of a particular producer denoted by $i \in \mathcal{N}$. We assume that the set of all producers $\mathcal{N}$ is fixed and we suppose that the true production cost function of producer $i \in \mathcal{N}$ is given by $A_i q_i + B_i q_i^2$ with coefficients $A_i \geq 0$ and $B_i > 0$ being known only to producer i. All the following results may be extended to $B_i = 0$, but to avoid several technical issues we omit it here. We argue that $B_i = 0$ is not realistic since the real marginal cost of electricity production is increasing in q_i. Now, producer $i \in \mathcal{N}$ aims to maximize his profit $\pi_i(a, b, D)$ given by

$$\pi_i(a, b, D) = (a_i - A_i)\, q_i(a, b, D) + (b_i - B_i)\, q_i(a, b, D)^2 \tag{5}$$

manipulating his own strategic variables $a_i, b_i \geq 0$ with the rest of variables $(a_{-i}, b_{-i}) \in \mathbb{R}_+^{2N-2}$ kept fixed. In other words, the i-th producer's problem $P_i(a_{-i}, b_{-i}, D)$ reads

$$P_i(a_{-i}, b_{-i}, D) \qquad\qquad \tilde{\pi}_i := \sup_{a_i, b_i \geq 0} \pi_i(a_i, a_{-i}, b_i, b_{-i}, D).$$

Then, the solution to this problem, i.e., the best response of producer $i \in \mathcal{N}$, provides him with a clear instruction how to bid in the modelled market situation. We consider the overall demand D as a parameter and our aim is to provide a full discussion of solution to $P_i(a_{-i}, b_{-i}, D)$ with respect to the value of this parameter. This closely corresponds to the actual needs of producers in the real-world electricity markets. Generally they have only some expectations on the overall demand D, and so they consider several possible scenarios with various values of D thus yielding different optimal bid functions. To this end the statement of the forthcomming and concluding Theorem 3.7 is presented in terms of the overall electricity demand D.

Finally, we explicitly state that we search only for a solution to $P_i(a_{-i}, b_{-i}, D)$ such that $\pi_i(a, b, D) > 0$, that is we assume that producer will not participate in the market otherwise. Indeed, since we model only one time period here, it makes no sense to participate in the market having non-positive profit.

At the moment, since the strategic variables $(a_{-i}, b_{-i}) \in \mathbb{R}^{2N-2}$ of the other producers are supposed to be fixed, we have to abandon the previous symmetry of the notation. There are several variables describing the (potential) situation in a market without producer $i \in \mathcal{N}$., i.e., a market consisting only of producers in $\mathcal{N} \setminus \{i\}$: we define

$$\lambda^c(a_{-i}, b_{-i}) = \min_{j \in \mathcal{N} \setminus \{i\}, b_j = 0} a_j,$$

and similarly also the other critical parameters $\mathcal{N}^c(a_{-i}, b_{-i})$, $D^c(a_{-i}, b_{-i})$ of E-ISO(a_{-i}, b_{-i}, D). In the same manner, we may define function $F(a_{-i}, b_{-i}, \lambda)$ and derive market price $\lambda(a_{-i}, b_{-i}, D)$ in analogy to (3). Meaning of all these reduced variables fully corresponds to the case of the full market definitions. We note that also Theorem 2.3 is valid for the setting of E-ISO(a_{-i}, b_{-i}, D). Having such a notation established, we illustrate the influence of the i-th producer's bid on the market price $\lambda(a, b, D)$.

Lemma 3.1. *Consider demand $D > 0$ and bid vector $(a, b) \in \mathbb{R}^{2N}_+$. Then*

(a) $\lambda(a, b, D) \leq \lambda(a_{-i}, b_{-i}, D)$,

(b) $a_i \leq \lambda(a, b, D)$ if and only if $a_i \leq \lambda(a_{-i}, b_{-i}, D)$,

(c) if $b_i > 0$, then, $a_i < \lambda(a, b, D)$ if and only if $a_i < \lambda(a_{-i}, b_{-i}, D)$.

Note that all the statements in this section will be given without a proof, an interested reader can found the details in the forthcomming publication.

Although this lemma can appear to be only a technical issue, it has some straightforward economical interpretations:

(a) part (a) states that the price in the market including producer i is always less or equal to the price in the market without producer i

(b) part (b) enlightens that if producer i considers to enter the market with a bid (linear or quadratic) lower than the present marginal price (in the market without him) then the modification of the market price due to his participation to the market can not make him out of the money.

(c) part (c) means that if producer i is in the money with a quadratic bid in the market including him, then the price in the market without him would be strictly higher than the linear coefficient of his bid.

Next, we show what values of $(a_i, b_i) \in \mathbb{R}^2_+$ are of potential interest for the i-th producer.

Theorem 3.2. *Assume $D > 0$ and take $(a_{-i}, b_{-i}) \in \mathbb{R}_+^{2N-2}$. Then, considering the unique solution $q(a, b, D)$ to the regulator's problem E-ISO(a,b,D), the i-th producer profit $\pi_i(a, b, D)$ satisfies one of the following statements:*

(a) for $a_i \leq \lambda(a_{-i}, b_{-i}, D)$ and $b_i > 0$,

$$\pi_i(a, b, D) = \frac{\lambda(a, b, D) - a_i}{4b_i^2}\big[a_i b_i - 2A_i b_i + a_i B_i + \lambda(a, b, D)(b_i - B_i)\big], \tag{6}$$

(b) for $a_i < \lambda(a_{-i}, b_{-i}, D)$ and $b_i = 0$ (and so $a_i = \lambda^c(a, b)$ and $\mathcal{N}^c(a, b) = \{i\}$),

$$\pi_i(a, b, D) = (\lambda^c(a, b) - A_i)(D - D^c(a, b)) - B_i(D - D^c(a, b))^2, \tag{7}$$

(c) for $a_i = \lambda(a_{-i}, b_{-i}, D)$ and $b_i = 0$ (and so $a_i = \lambda^c(a, b)$ and $i \in \mathcal{N}^c(a, b)$),

$$\pi_i(a, b, D) = (\lambda^c(a, b) - A_i)\frac{D - D^c(a, b)}{N^c(a, b)} - B_i\left(\frac{D - D^c(a, b)}{N^c(a, b)}\right)^2, \tag{8}$$

(d) for $a_i > \lambda(a_{-i}, b_{-i}, D)$ it holds $\pi_i(a, b, D) = 0$

Note that the different cases of Theorem 3.2 are described in terms of comparison between a_i and $\lambda(a_{-i}, b_{-i}, D)$ thus independently of the value of $\lambda(a, b, D)$, which is not known when producer i wants to decide his bid (a_i, b_i). Let us now emphasize, through the following corollary, that as soon as the linear coefficient A_i of the production cost function of the i-th producer is greater than the price $\lambda(a_{-i}, b_{-i}, D)$ in the market without producer i, then there is no bid (a_i, b_i) for producer i ensuring him positive profit.

Corollary 3.3. *For any $D > 0$, $(a_{-i}, b_{-i}) \in \mathbb{R}_+^{2N-2}$ and $A_i \geq \lambda(a_{-i}, b_{-i}, D)$, the i-th producer's profit is non-positive, that is $\pi_i(a, b, D) \leq 0$.*

Next we introduce a level of production

$$\tilde{q}_i(a_{-i}, b_{-i}) = \frac{\lambda^c(a_{-i}, b_{-i}) - A_i}{2B_i} \tag{9}$$

having a significant economic meaning for producer $i \in \mathcal{N}$.

Remark 3.4. *Let $(a_{-i}, b_{-i}) \in \mathbb{R}^{2N-2}$, $a_i = \lambda^c(a_{-i}, b_{-i})$ and $b_i = 0$ be fixed for some $i \in \mathcal{N}$. Then, if we consider q_i in (5) as a free variable for the moment, the profit of producer i is given by $\pi_i^c(q_i) : q_i \to (\lambda^c(a_{-i}, b_{-i}) - A_i)\, q_i - B_i\, q_i^2$. Then, the maximum of $\pi_i^c(q_i)$ is attained for $q_i = \tilde{q}_i(a_{-i}, b_{-i})$, thus corresponding to a kind of ideal production rate for producer i. This follows from $B_i > 0$, then for production quantity higher than $\tilde{q}_i(a_{-i}, b_{-i})$ the additional production cost will be higher than the respective additional gain. Finally, we note that $\tilde{q}_i > 0$ and $\pi_i^c(\tilde{q}_i) > 0$ provided $A_i < \lambda^c(a_{-i}, b_{-i})$.*

Further, we investigate only values of $(a_i, b_i) \in \mathbb{R}_+^2$ such that assumptions of Theorem 3.2 (a), (b) and (c) are satisfied. Otherwise, the i-th producer's profit would be non-positive and we assume that under such conditions the producer will not enter the market at all. Then, we characterize conditions for the existence of a solution to $P_i(a_{-i}, b_{-i}, D)$, determine this solution and show that it is unique. Some more preliminary notation is necessary. We introduce two more quantities of electricity production being significant for producer $i \in \mathcal{N}$:

$$q_i^m(a_{-i}, b_{-i}) = \frac{\lambda^m(a_{-i}) - A_i}{2B_i + m^+(a_{-i}, b_{-i}, \lambda^m(a_{-i}))}, \tag{10}$$

$$q_i^c(a_{-i}, b_{-i}) = \frac{\lambda^c(a_{-i}, b_{-i}) - A_i}{2B_i + m^-(a_{-i}, b_{-i}, \lambda^c(a_{-i}, b_{-i}))}. \tag{11}$$

Lemma 3.5. *For any $(a_{-i}, b_{-i}) \in \mathbb{R}_+^{2N-2}$ it holds $q_i^m(a_{-i}, b_{-i}) < \tilde{q}_i(a_{-i}, b_{-i})$ and $q_i^c(a_{-i}, b_{-i}) < \tilde{q}_i(a_{-i}, b_{-i})$. Moreover, we have $q_i^m(a_{-i}, b_{-i}) < q_i^c(a_{-i}, b_{-i})$ provided $A_i < \lambda^c(a_{-i}, b_{-i})$.*

Now, recall that function $f : \mathbb{R} \to \mathbb{R}$ is quasiconcave if for all $x, y \in \mathbb{R}$ and all $u \in [x, y]$ it holds $f(u) \geq \min\{f(x), f(y)\}$. Moreover, function $f : \mathbb{R} \to \mathbb{R}$ is strictly quasiconcave if it is quasiconcave and for all $x, y \in \mathbb{R}$, $x \neq y$ and all $z \in]x, y[$ we have $f(z) > \min\{f(x), f(y)\}$.

Proposition 3.6. *Let $(a_{-i}, b_{-i}) \in \mathbb{R}_+^{2N-2}$, $D > 0$ and $b_i = 0$ be fixed. Then, $\pi_i(a_i, a_{-i}, 0, b_{-i}, D)$ is strictly quasiconcave in a_i on $[0, \lambda(a_{-i}, b_{-i}, D)[$, and problem*

$$\hat{P}_i(a_{-i}, b_{-i}, D) \qquad \sup_{a_i \in [0, \lambda(a_{-i}, b_{-i}, D)[} \pi_i(a_i, a_{-i}, 0, b_{-i}, D)$$

admits a solution if and only if one of the following alternatives holds:

(a) $A_i < \lambda(a_{-i}, b_{-i}, D) < \lambda^c(a_{-i}, b_{-i})$ (implying $\lambda^m(a_{-i}) < \lambda(a_{-i}, b_{-i}, D)$),

(b) $\lambda^m(a_{-i}) < \lambda(a_{-i}, b_{-i}, D) = \lambda^c(a_{-i}, b_{-i})$ and $q_i^c(a_{-i}, b_{-i}) > D - D^c(a_{-i}, b_{-i})$.

Moreover, if a solution exists, it is unique. Denoting it by $\tilde{a}_i$, it is given by

$$\begin{cases} \tilde{a}_i = \lambda^m(a_{-i}) & \text{if} \quad D \leq q_i^m(a_{-i}, b_{-i}), \\ \dfrac{\tilde{a}_i - A_i}{2B_i + m^-(a_{-i}, b_{-i}, \tilde{a}_i)} \leq D - F(a_{-i}, b_{-i}, \tilde{a}_i) \leq \dfrac{\tilde{a}_i - A_i}{2B_i + m^+(a_{-i}, b_{-i}, \tilde{a}_i)} & \text{if} \quad D > q_i^m(a_{-i}, b_{-i}), \end{cases} \tag{12}$$

and satisfies $\tilde{a}_i \in [\lambda^m(a_{-i}), \lambda^c(a_{-i}, b_{-i})[$. Moreover, the respective maximal profit is positive, $\pi_i(\tilde{a}_i, a_{-i}, 0, b_{-i}, D) > 0$. Additionally, if a solution does not exist, then $\pi_i(a, b, D)$ is strictly increasing in a_i on $[0, \lambda(a_{-i}, b_{-i}, D)[$.

It may occur that there is no maximizer in problem $P_i(a_{-i}, b_{-i}, D)$, i.e., the best response of producer $i \in \mathcal{N}$ does not exist. However, if the supremum of the profit $\tilde{\pi}_i$ defined in $P_i(a_{-i}, b_{-i}, D)$ is positive, a sequence of bids $(\tilde{a}_i^k, \tilde{b}_i^k)_k$ is said to be a limiting best response of producer i if it yields the optimal profit $\tilde{\pi}_i$, i.e.,

$$\lim_{(\tilde{a}_i^k, \tilde{b}_i^k) \to (\tilde{a}_i, \tilde{b}_i)} \pi_i(\tilde{a}_i^k, a_{-i}, \tilde{b}_i^k, b_{-i}, D) = \tilde{\pi}_i. \tag{13}$$

In such a situation we will present in the forthcomming theorem one bounded limiting best response, thus providing a limiting best response strategy to producer $i \in \mathcal{N}$. Then, we call $\tilde{\pi}_i$ a limiting profit, and the respective production quantity will be referred to as a limiting production quantity. Next, we introduce the final theorem of this article discussing (existence of) the best response of producer $i \in \mathcal{N}$ with respect to various values of the overall electricity demand $D > 0$ and bid functions of other producers $(a_{-i}, b_{-i}) \in \mathbb{R}_+^{2N-2}$ kept fixed. In this setting we define $q_i^0(a_{-i}, b_{-i}) = F(a_{-i}, b_{-i}, A_i)$ provided $A_i \le \lambda^c(a_{-i}, b_{-i})$, thus allowing reformulation of Corollary 3.3 in terms of production quantity.

Theorem 3.7. *Let $D > 0$, $(a_{-i}, b_{-i}) \in \mathbb{R}_+^{2N-2}$ for some $i \in \mathcal{N}$ and consider the problem*

$$P_i(a_{-i}, b_{-i}, D) \qquad\qquad \tilde{\pi}_i := \sup_{a_i, b_i \ge 0} \pi_i(a_i, a_{-i}, b_i, b_{-i}, D). \tag{14}$$

If $D^c(a_{-i}, b_{-i}) > 0$ then either $A_i \ge \lambda^c(a_{-i}, b_{-i})$ and $\tilde{\pi}_i \le 0$, or one of the following alternatives holds:

(a) if $D \in]0, q_i^0(a_{-i}, b_{-i})]$ then $\tilde{\pi}_i \le 0$,

(b) if $D \in]q_i^0(a_{-i}, b_{-i}), D^c(a_{-i}, b_{-i}) + q_i^c(a_{-i}, b_{-i})[$ then $\tilde{\pi}_i > 0$ and there is a unique best response $(\tilde{a}_i, \tilde{b}_i)$ given by $\tilde{a}_i \in [\lambda^m(a_{-i}), \lambda^c(a_{-i}, b_{-i})[$ satisfying (12) and $\tilde{b}_i = 0$,

(c) if $D \in [D^c(a_{-i}, b_{-i}) + q_i^c(a_{-i}, b_{-i}), D^c(a_{-i}, b_{-i}) + \tilde{q}_i(a_{-i}, b_{-i})]$ then $\tilde{\pi}_i > 0$ and a limiting best response $(\tilde{a}_i^k, \tilde{b}_i^k)_k$ is given by $\tilde{a}_i^k \nearrow \lambda^c(a_{-i}, b_{-i})$ and $\tilde{b}_i^k = 0$,

(d) if $D \in]D^c(a_{-i}, b_{-i}) + \tilde{q}_i(a_{-i}, b_{-i}), +\infty[$ and $D \ne D^c(a_{-i}, b_{-i}) + (N^c(a_{-i}, b_{-i}) + 1)\, \tilde{q}_i(a_{-i}, b_{-i})$ then $\tilde{\pi}_i > 0$ and a limiting best response $(\tilde{a}_i^k, \tilde{b}_i^k)_k$ is given by $\tilde{a}_i^k \nearrow \lambda^c(a_{-i}, b_{-i})$ and $\tilde{b}_i^k \searrow 0$ satisfying

$$\tilde{a}_i^k = \lambda^c(a_{-i}, b_{-i}) - \frac{2B_i b_i^k}{B_i + b_i^k}\, \tilde{q}_i(a_{-i}, b_{-i}), \tag{15}$$

(e) if $D = D^c(a_{-i}, b_{-i}) + (N^c(a_{-i}, b_{-i}) + 1)\, \tilde{q}_i(a_{-i}, b_{-i})$ then $\tilde{\pi}_i > 0$ and there is a unique best response $(\tilde{a}_i, \tilde{b}_i) = (\lambda^c(a_{-i}, b_{-i}), 0)$.

For $D^c(a_{-i}, b_{-i}) = 0$ all these alternatives are still valid provided $q_i^c(a_{-i}, b_{-i}) := 0$.

Having $D^c(a_{-i}, b_{-i}) = 0$, i.e., $\lambda^m(a_{-i}) = \lambda^c(a_{-i}, b_{-i})$ due to Remark 2.1 (b), and $A_i < \lambda^c(a_{-i}, b_{-i})$, it holds $q_i^0(a_{-i}, b_{-i}) = F(a_{-i}, b_{-i}, A_i) \le F(a_{-i}, b_{-i}, \lambda^c(a_{-i}, b_{-i})) = F(a_{-i}, b_{-i}, \lambda^m(a_{-i})) = 0$ with regards to Lemma 2.2, thus alternatives (a) and (b) can not occur since we put $q_i^c(a_{-i}, b_{-i}) := 0$. Thus, only alternatives (c), (d) and (e) of the theorem have to be considered for the case of $D^c(a_{-i}, b_{-i}) = 0$. Note that $q_i^c(a_{-i}, b_{-i})$ was not previously defined once $\lambda^m(a_{-i}) = \lambda^c(a_{-i}, b_{-i})$, see (11).

We note that the presentation of the statement of Theorem 3.7 closesly corresponds to a real-world needs of electricity producers, which look for the optimal bid function considering several scenarios with vairous values of electricity demand D.

4 Conclusion

This article closely follows the ideas developed already in [1]. We found a new way how to treat the modelling of the electricity markets. Using the analytic formula for the ISO problem, we are able to finally resolve analytically even the problem of a producer, thus obtaining Theirem 3.7. This way we have a complete picture of the best bidding strategy for the considered producer.

The proposed way of research seems to be truly promissing. Further extensions of this model may directly lead to more realistic results, whereas such an analytical approach will still make these model ameanable to a detailed examination. This is however beyond the scope of this article.

References

[1] M. Pištěk, *New approach to electricity markets: analytic solution of ISO problem*, Proceedings of PhD workshop, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Czech Republic, 15,22 November 2013, 217-225.

[2] X. Hu & D. Ralph, *Using EPECs to Model Bilevel Games in Restructured Electricity Markets with Locational Prices*, Operations Research 55 (2007), 809-827.

[3] D. Aussel, M. Cervinka and M. Marechal, Day-ahead electricity market with production bounds, (2012), 24 pp.

[4] D. Aussel, R. Correa & P. Marechal *Spot electricity market with transmission losses*, J. Indust. Manag. Optim. (2012), 18 pages.

[5] R. Henrion, J.V. Outrata, and T. Surowiec, Analysis of M-stationary points to an EPEC modeling Oligopolistic Competition in an Electricity Spot Market, ESAIM: COCV 18 (2012) 295-317

Adaptive Knowledge Testing with Bayesian Networks

Martin Plajner

1st year of PGS, email: plajnmar@fjfi.cvut.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jiří Vomlel, Department of Decision-Making Theory
Institute of Information Theory and Automation, AS CR

Abstract. This paper focuses on the topic of adaptive knowledge testing. The concept of testing is briefly reviewed as well as the structure of Bayesian networks, which are used to control test. Special attention is given to the data collection with a knowledge test. The test is analysed here and the paper brings its psychometrics evaluation that proves it is a correct tool to collect data. Test results are summarized and data obtained are used to model a Bayesian network for this case. In the end we demonstrate adaptive testing with this network.

Keywords: adaptive, testing, Bayesian network

Abstrakt. Tento článek se věnuje tématu adaptivního testování znalostí. Je přiblížen koncept testování a struktura bayesovských sítí, které jsou k němu využívány. Značná pozornost je věnována způsobu sběru dat pomocí testu znalostí, který je zde analyzován. Článek přináší psychometrický rozbor testu a prokazuje, že je vhodným nástrojem pro sběr dat. Výsledky testu jsou shrnuty a následně jsou sesbíraná data využita k tvorbě bayesovké sítě modelující náš případ. Na závěr na této síti demonstrujeme způsob adaptivního testování.

Klíčová slova: adaptivní, testování, bayesovká síť

1 Introduction

Educational testing plays an important role in the modern society. Every person participates in a large number of tests that are used to prove different qualities through his/her life. While tests vary in their questions, style, thoroughness, and improve over time with more elaborate questions and better sets of questions, the way of testing itself does not change much. There is usually a large set of possible questions which are adequate to be asked but only a fixed subset is selected for a single version of the test. Reasons for this are obvious as it is not possible to force an examinee to answer hundreds or even thousands of questions. It is possible to pick questions at random but that could lead to "lucky" and "unlucky" selections of differing difficulties. Other methods for selecting questions can be used but there will always be a large amount of questions which were not included in the test.

One way of solving problems mentioned above is Computerized Adaptive Testing (CAT). It is a knowledge-based testing concept where the examinee is not required to answer all questions from the question pool. Questions for the examinee are selected by the computer based on his/her previous answers and the data set modelled from many previous respondents. It means that there is no fixed version of the test and every individual gets his/her test crafted while answering questions. This concept provides a way to obtain a measurement of student's abilities with the reasonably high accuracy and confidence, but require fewer responses, less time and have positive effect on the examinee's morale as, typically, the student should answer about a half of questions correctly. It is also possible to take the advantage of a large question pool as every questions which has a significant information value is asked for some examinees and on the other hand question with no or small information value are asked less often.

Examples of a successful adaptive testing deployment are TOEFL language exam and GRE – graduate record examination [5].

2 Adaptive Testing and Bayesian Networks

CAT can be divided into two phases: model creation and testing. In the first one the model of the system is created while in the second one the model is used to actually test the examinees. There are many different model types which can be used for adaptive testing but in this paper we are going to focus on the Bayesian networks only.

Bayesian network is a conditional independence structure. It consists of the following: A set of variables and a set of directed edges between them which form a directed acyclic graph (DAG). Each variable has a list of mutually exclusive possible states. The conditional probability distribution is defined for each variable with its parents in the conditioning part (variable A with parents B_1,B_2,...,B_n has the conditional probability table $P(A|B_1, B_2, ..., B_n)$)[1]. For example the structure in the Figure 2 is a graphical representation of a Bayesian network (probability tables are not shown).

The goal of the model creation is to describe the relations between questions and student abilities. In order to construct a Bayesian network it is necessary to identify random variables. There are two types of variables. The first type is an observed variable which is a response to an individual question. Collection of these variables is called a test model. The second type is an unobserved, student model, variable. This type of variables corresponds to abilities of the examinee and since there is no way of direct measurement of these variables they cannot be observed. For the first type, there is a variable for every question. For the second type, there is no exact rule how to create these variables. One way of creating them is expert knowledge where an expert describes a set of abilities and what abilities he/she is expecting to play a part in each question. This forms connections to create a DAG. To finish a Bayesian network creation it is necessary to add initial values into conditional probability tables. It is again usually done be an expert. In order to reflect data it is necessary to perform a fitting. This step is called learning and it can be done with different machine learning algorithms.

When the network is set up as described it is possible to use it for predictions. The theory behind the use of Bayesian networks is quite extensive and we will bring only a brief overview of the most important mechanism[2]. There are two main key points in the calculations with Bayesian network. The first one is evidence. Evidence e is

[1] Note that the variables with no parents have the table in the form $P(A)$

[2] Detailed explanations and further reading can be found for example in [3] or [1]

a piece of information we obtained (in our case as an answer to a question) about the variable A. This information fixes the state of the variable A. This corresponds to setting probabilities of states of the variable A to zeros except for the state i. The second key point is a belief updating (also called inference). It allows us to propagate the information obtained as evidence through the network and also to evaluate marginal probabilities for a single variable. There are different methods for inference but in general they all consists of repeated steps of multiplications of potentials[3] and marginalizations of variables. It is usually performed over the junction tree constructed from the network. One of the inference mechanisms is the lazy propagation which takes advantage of local structures in the network to improve the efficiency of calculations [4].

3 Test design

In order to perform our research with CAT we collected data as an input for a further analysis and the test creation. The paper test for high school students was designed to serve this need. The test focuses on mathematical knowledge and it is intended for students attending last two years of high schools. The test was revised and updated several times before it reached its final state.

It is essential to note that this test is not meant for student school qualification and no grades which would influence student's school results are given.

During the assessment of the test the analysis of common mistakes was performed. Based on this analysis questions were further broken down into sub-questions and every sub-question was graded separately. This step was performed to provide a neater scope of separation and better connection of responses to student's abilities. Each sub-question is graded with 0-4 points and each question consist of 1-3 (in one case 8) sub-questions.

In addition to answers to problems information about students is collected. This includes mostly some personal factors as sex, age, and grades from mathematics, physics, and chemistry from the recent period. These factors will be used to better differentiate between students and to better predict their performance as well as to verify the validity of the test.

Feedback for students

To increase the motivation of students the test results are stored online and every student can view his/her performance. The comparison with the rest of the test sample is provided (in the form of quantiles). The idea is to provide students with reflection of their abilities. A website was designed to meet these needs and if the student enters his/her email in the test he/she is notified when the result is uploaded. It is then possible for the student to display the result. There is also a utility on the website which allows one to enter additional information missing in the test (or unavailable during the filling time) such as the final exam grades.

The final shape of the test, respondents

There are 29 questions in the final version of the test which are divided into 55 sub-questions. The maximal score is 100 points. There is a time limit of 45 minutes (it has to

[3] To clarify: probability tables are special case of potentials, tables during the course of operation need not be probability tables (sum to 1 etc.).

fit one high school lesson). To this date the test was answered by 110 students from two Prague's high-schools in the age of 16 - 20. More students should attend the test in the year 2014/2015. There were 33 men and 63 women in total, the remaining 11 students did not provide the sex.

4 Test assessment

In the following section psychometrics analysis methods from [6] are summarized. Results of the analysis is presented as well.

True scores and reliability

The goal of the test is to measure variables from the student model described in the section 2 and to use these to predict answers to other questions. As always the measurement process is obstructed with measurement errors. The error is caused by many different factors (the examinee could have a bad day, be ill, guess the answer, or get distracted while solving a single problem,...) and it is reasonable to expect them to have an important influence. The value obtained as a measurement x of the variable X is called a raw score and is in the form

$$x = \tau + e$$

where τ is the true score and e is an additive error.

There is an obvious question whether the raw score is influenced more by the true score or the error. For many measurements the maximum-likelihood estimator of the error is the variance of many consecutive measurements of the same factor. In our case it proves to be impractical to measure one person multiple times and it is not as well possible to use the variance of many different examinees as their true values most likely differ. The variability of scores in the data set is then caused by actual differences between examinees (different true scores) as well as errors. It is usually expected that the data set satisfies homoskedasticity condition[4]. With this assumption true scores and errors are statistically independent and thus the observed variance σ_x is a sum of variances of true scores σ_τ and errors σ_e.

$$\sigma_x = \sigma_\tau + \sigma_e$$

The best possible situation is that the variance of the measured variable X is fully modelled by true scores. This situation is very unlikely to happen. To determine the level of the relationship we introduce the value called reliability which is defined as follows:

$$r_{xx} = \frac{\sigma_\tau}{\sigma_x} = \frac{\sigma_\tau}{\sigma_\tau + \sigma_e}$$

The higher the value the better. Unfortunately variables σ_τ in the nominator as well as σ_e in the denominator of the second fraction are hidden (unobservable) variables and as such we are unable to evaluate their variance. The reliability has to be estimated with a different approach.

There are many possible approaches and we will elaborate more into one of them which is known as Cronbach's alpha coefficient only. The idea is that the items of the

[4] Homoskedasticity means that the size of an error is not correlated with the size of the measured variable

test are measuring the same factor and thus they should correlate with each other. The amount of pair wise correlations for q questions is $k = \frac{q(q-1)}{2}$. All these correlations are put together in the Cronbach's alpha coefficient which can be calculated as

$$r_{xx} \approx \alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_t^2}\right)$$

where σ_i is the variance of the ith item of the test, σ_t is the variance of the whole test and n is the number of items in the test. The coefficient should reach high values. According to [2] any value below 0.5 means the test is of no use. To provide reasonable comparison results it should be over 0.9.

For our data set the following values were calculated:
Cronbach's alpha for numeric classification: $\alpha = 0.92$
These values show reasonably high reliability of the test.

Normalization and standard scores

In order to use scores to distinguish between different examinees raw scores have to be transformed to standard scores. There are many different types of standard scores and all of them are obtained by a linear transformation of raw scores (note that it means that the order of examinees is not changed by this kind of transformation) by the following formula

$$x' = \mu' + \sigma' \frac{(x - \mu)}{\sigma}$$

Where x' is the transformed score, μ' and σ' are desired mean and variance values of the standardized score, μ and σ are previous mean and variance values and x is the raw score. To apply these transformations it is required that the raw score belong to the Gaussian distribution (ideally with the mean value in the middle of possible scores). Standardized scores differ in the chosen parameters of μ' and σ' and some special selections are generally recognized. The most commonly used is the z-score with the mean value 0 and the variance 1. Another well known standard score is the IQ score ($\mu' = 100$, $\sigma' = 15$) used mostly for intelligence testing. Other well known scores are also stens, stenines, percentiles, and t-scores.

The set of scores obtained from our data set did not belong to Gaussian distribution. The proof is displayed in the Figure 1 where it can be clearly seen that it does not even fit the Gaussian distribution with the mean value 44.182 (instead of middle 50 points) due to very low p-value. The solution to this problem is provided by the McCall's area standardization [6] which transforms raw scores to the Gaussian distribution. This step was performed at first and then scores were transformed to standardized score scales. To illustrate these scales a short excerpt from whole scale tables for the z-score and the IQ score is shown in the Table 1.

Table 1: Standardized scores

raw	0	10	20	30	33	36	46	56	66	76	86	93
z	-2.61	-1.36	-0.59	-0.05	0	0.10	0.51	0.94	1.16	1.46	1.74	2.61
IQ	61	80	91	99	100	102	108	114	117	122	126	139

Validity

Another question it is important to ask is whether the test is actually measuring the factor

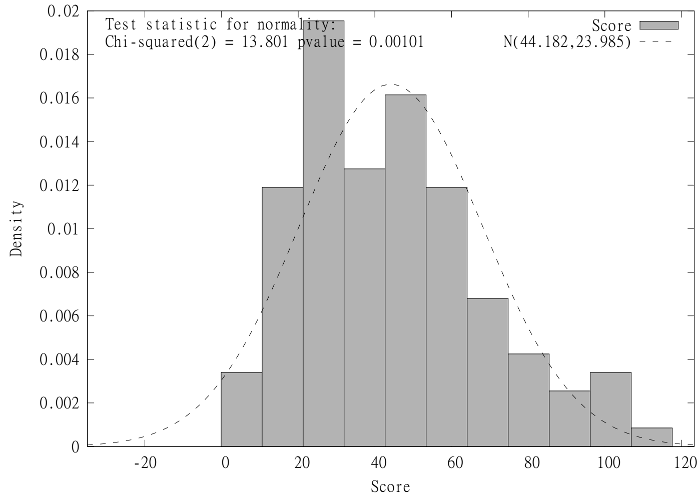


Figure 1: Score frequency

it is supposed to measure (i.e. in our case if the score obtained reflects mathematical skills rather than for example the ability to read the question or the writing skill of the examinee). This characteristic is called validity and there are many different ways of proving the test is valid. Most validity proofs come from outside the test. One way is to let examinee to answer a new different test measuring the same factor (ideally a test which is already well established). Another way is to consult other factors known about the examinee, which is what was performed in our case.

in addition to answers to problems student's grades from subjects (mathematics, physics, and chemistry) were obtained. It is reasonable to expect a correlation between these grades and the score reached. The correlation is present and its values are shown in the following paragraphs. Because of this fact, although the complete validation would require more thorough examination, it is expected that the test is valid.

Results

The Table 2 shows the reached scores divided by gender and school. It seems that there is no obvious correlation between sex of the student and the score the student obtained. This is supported by the calculation of the correlation which evaluated to $c(score, sex) = 0.03$. With the null hypothesis that there is a correlation the statistical test results in the $p - value = 0.7572$ disallowing the rejection of independence. Interestingly there is a connection between the filling of the name cell (some students did not fill their name) and their score where the correlation coefficient is $c(score, name) = 0.33$ with $p - value = 0.0004$ which means that a person who filled his/her name was likely to score better in the test.

As mentioned above, the correlation between grades of the student and achieved score was measured. Obtained values are shown in the Table 3. It is clear that these correlations are as expected (a better grade (smaller) leads to a higher score - negative sign in the correlation). Also the correlation with mathematics is the highest and chemistry is the lowest with physics in the middle. This fact was not predicted and it is an interesting although not surprising one.

Table 2: Average score achieved in two high schools ("Na Pražačce" and "Arcus")

	Na Pražačce	**Arcus**	**Total**
Total	46.68	42.76	43.86
Males	40.08	51.40	46.94
Females	54.86	42.53	45.27

Table 3: Correlations of the score with grades

	Mathematics	**Physics**	**Chemistry**
Correlation with score	-0.58	-0.44	-0.36
p-value	0.0000	0.0000	0.0004

Some questions were in the form of real life problems rather than mathematical problems. These questions were correlated with the score independently as well. The result is displayed in the Table 4. In the first column it is possible to see that there is a strong correlation with the total score. Also in this case there is not a strong correlation with the sex of the student even though the value is a bit higher (in favour for men). Moreover the trend of correlations with grades is preserved but values are lower. It leads to an assumption that students with worse grades from these subjects answered correctly rather this kind of questions than other questions.

Table 4: Correlations of the real life problems with other factors

	Score	**Female**	**Mathematics**	**Physics**	**Chemistry**
Correlation	0.72	-0.11	-0.36	-0.18	-0.17

5 Current Bayesian Network

Based on the data collected and the experience from the assessment of tests a Bayesian network was created[5]. Its structure is displayed in the Figure 2. As explained above there is a node for each sub-question (yellow/white) and there are 7 ability nodes (red/grey). The abilities correspond to different mathematical skills and are described in the Table 5.

Each question is connected to at least one ability (groups in the top part of the graph) or more (groups in the bottom part of the graph). This network's design was based on our expert knowledge and there were also initial probabilities inserted (not shown in the Figure 2). The network is then learned using the Hugin's EM algorithm to update

[5] We use the Hugin environment to model the network and to perform calculations (www.hugin.com).

Table 5: Skills present in the student model

S1	graphs (recognize functions, draw,...)
S2	points in the graph (find points, plot points, read from graph,...)
S3	monotonic functions (analysing, using for calculations,...)
S4	domains of functions
S5	function formulas (create, use)
S6	equations
S7	real-world problems

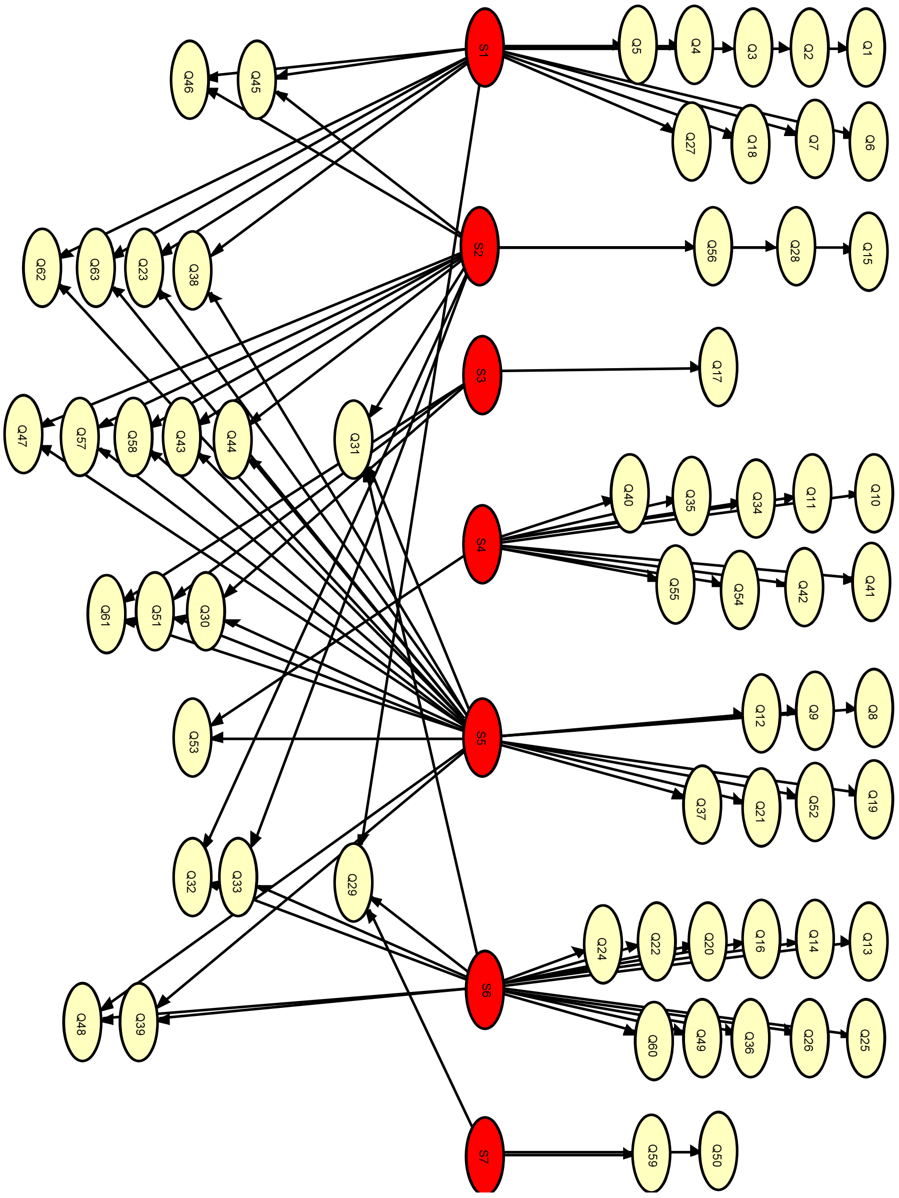


Figure 2: Bayesian network constructed with our expert knowledge

probabilities to reflect our data set. After this step probabilities associated with each node, corresponding question/skill for an average student, can be viewed in Hugin.

The main goal is to use the Bayesian network for the adaptive testing. Once we have the learned network we should be able to do this. The testing is performed in turns which correspond to answers to questions. In every turn following steps are performed:

1. The question with the best information value is selected
2. The student responds to the question
3. The answer is inserted into the network as new evidence
4. Probabilities in the network are updated (inference is performed)

This sequence is repeated until a criterion is reached. Either we test for a fixed number of questions or a determination goal could be set: the test is stopped when it reaches a state where an examinee can be assigned a score (or group of students) with a certain confidence. Steps 3 and 4 from the list above are performed with the mechanisms outlined in the Section 2. The second step is the task of the examinee. The remaining step 1 could prove quite hard to perform. It is necessary to select the question which gives us the best information value and allows us to differentiate the examinee from the rest of the group as much as possible. The best candidates to select are questions which in the current state of the network have probabilities close to the uniform distribution. Since these questions are the hardest to predict and they provide the largest information gain as well.

Next we provide an example simulation of the testing process with the network in Figure 2. First, the network is learned with the EM algorithm. Then, we select one question with high information value - Q12 ($P(Q12 = 0) = 0.4273$, $P(Q12 = 1) = 0.0909$, $P(Q12 = 2) = 0.4818$). In our simulation the student answers correctly to this question. Inserted evidence modifies the probabilities of other questions as well as the skill it is attached to (S5). Updated probabilities of some of the variables are shown in the top part of the Table 6. Probabilities are updated only for variables connected to the skill S5 where the probability of the student having this skill increased. The next question is Q32 ($P(Q32 = 0) = 0.4633,\ P(Q32 = 1) = 0,\ P(Q32 = 2) = 0.5367$). The student answers this question incorrectly. The bottom part of the Table 6 shows updated probabilities. In this case two skill nodes updated and as well are variables connected to them. In both cases observed changes correspond with our expectations - correct answers yielding higher probabilities for other correct answers and vice versa.

For more credible verification it is necessary to collect more data and then perform additional tests. It is planned to run the leave-one-out cross-validation which consist of the following steps. First, a single observation (examinee's result) is removed from the data set and the network is learned from the remaining data. The network then simulates the testing. Questions are selected as described above and answers are fed from the previously removed observation. The failure ratio is recorded as the relative ratio of wrong predictions. This procedure is repeated for every observation (n times). The goal is to have the average ratio over all examinees as low as possible. It would mean that the network predicts examinees' answers correctly (makes only a small number of mistakes).

Table 6: Changes in probabilities after inserting evidence $e_1(Q12 = 2)$ and then $e_2(Q32 = 0)$. In skill variables (S2, S5, S6) states are h-have/hn-have not the skill.

P	$P(S5 = h\|e_1)$	$P(Q9 = 1\|e_1)$	$P(Q37 = 0\|e_1)$	$PQ30 = 4\|e_1)$
Before	18.35	58.60	61.82	32.93
After	34.30	66.69	53.85	41.05
P	$P(S2 = h\|e_1, e_2)$	$P(S6 = hn\|e_1, e_2)$	$P(Q24 = 0\|e_1, e_2)$	$PQ29 = 4\|e_1, e_2)$
Before	0.4313	0.6264	0.8000	0.3770
After	0.3695	0.9258	0.9490	0.2874

6 Conclusion

The most important result presented in this paper is the empirical proof that the test brings valid data about examinees. It is possible to continue collecting data in the same way. It is necessary to increase data volume to continue our work. Nevertheless, it was already possible to construct a Bayesian network which seems to provide reasonable predictions. The following step is to prove this assumption with more elaborate procedures. An additional software tool to perform the inference and to manage more advanced tasks with the network is also being developed. It will allow us to do operations outside of the Hugin environment in a controlled and specific way. This tool will be later used in the implementation of a computerized version of the test in its adaptive form.

References

[1] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer, (1999).

[2] G. C. Helmstadter. *Principles of Psychological Measurement.* New York: Appleton-Century-Crofts, (1964).

[3] F. V. Jensen. *Bayesian Networks and Decision Graphs.* Springer, (2001).

[4] A. L. Madsen and F. V. Jensen. *Lazy propagation: A junction tree inference algorithm based on lazy evaluation.* Artificial Intelligence **113** (1999), 203–245.

[5] R. J. Mislevy and R. G. Almond. *Graphical models and computerized adaptive testing.* CSE Technical Report **434** (1997).

[6] T. Urbanek, D. Denglerova, and J. Sirucek. *Psychometrika.* Portál, (2011).

Molecular Simulation of Water Vapor–Liquid Phase Interfaces Using TIP4P/2005 Model

Barbora Planková

3rd year of PGS, email: barbora.plankova@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jan Hrubý, Department of Thermodynamics
Institute of Thermomechanics, AS CR

Abstract. Molecular dynamics simulations for water were run using the TIP4P/2005 model for temperatures ranging from 250 K to 600 K. The density profiles and the surface tension were calculated as a preliminary results. The surface tension values matched quite nicely with the IAPWS correlation over wide range of temperatures. As a partial result, DL_POLY Classis was successfully used for tests of the new computing cluster in the Institute of Thermomechanics.

This text is a short version of the one that will be presented at **Experimental fluid mechanics 2014** in Český Krumlov (18.11.2014 - 21.11.2014). Whole text subsequently published in **The European Physical Journal**.

Keywords: Density gradient theory, nucleation, PC-SAFT, Cahn-Hilliard theory

Abstrakt. Provedli jsme simulace molekulární dynamikou pro vodu s použitím modelu TIP4P/2005 pro teplotu od 250 K do 600 K. Jako předběžné výsledky jsme spočítali profily hustot a povrchová napětí. Hodnoty povrchového napětí dobře korespondují s hodnotami IAPWS pro širokou škálu teplot. Jako mezivýsledek jsme rozběhali a otestovali program DL_POLY na novém výpočetním klastru Institutu termomechaniky.

Tento text je krátkou verzí práce, která bude prezentována na konferenci **Experimental fluid mechanics 2014** v Českém Krumlově (18.11.2014 - 21.11.2014). Celý text následně publikován v žurnálu **The European Physical Journal**.

Klíčová slova: Gradientní teorie, nukleace, PC-SAFT, Cahn-Hilliardova teorie

1 Introduction

Water is perhaps the most studied substance in the world due to its importance in daily life, industry or physical, chemical or biological processes. Due to its many anomalies and non-standard behavior it is however very hard to model. The non-trivial phenomena are caused by its polar character and consequently by its formation of hydrogen bonds.

Motivation of this work is to shed light on the discrepancies between experiments and simulations, e.g. the second inflection point of water [4], to reproduce measurements of the surface tension of supercooled water as well as to enhance our theoretical work concerning nucleation rates predictions [10] and capillary waves modeling [3].

In this paper the preliminary results are published as part of our newly formed simulation group. Primary objective was to get the software executing the simulation procedure working in our new cluster in the direction of our interests. We used DL_POLY Classic

Table 1: Simulation parameters of TIP4P/2005 water molecule atoms: oxygen O, hydrogen H, mass-less charge M. m is molar mass, M charge in units of elementary charge, ϵ and σ are Lennard–Jones parameters.

atom	m (g/mol)	M (e)	ϵ (kJ/mol)	σ (Å)
O	15.99940	0.000	0.77490	3.1589
H	1.00794	0.5564	–	–
M	0.00000	-1.1128	–	–

program on 4 × 24 Intel(R) Xeon(R) CPU E5645 @ 2.40GHz CPUs, debugging and early computations were executed on computer with 4 AMD Phenom(tm) 9600 Quad-Core GPUs.

A water slab for various temperatures from 250 K towards the critical point (which was not exceeded) stopping at T = 600 K was simulated. To have a relevant reference point, we followed specifications of papers [11] and [12].

2 TIP4P/2005 model

There are many water models that are simple and ridig but describe the water properties quite well. Perhaps the most universal one is the so-called TIP4p/2005 model, based on TIP4P [6], which introduces an auxilliary atom M. This atom carries the negative charge, is massless and close to oxygen atom (0.15 Å). The model is TIP4P/2005 was proposed by Vega and Abascal [1] in 2005. They tried to combine good phase diagram of TIP4P with target properies of SPC/E improving the melting point. It has been shown [14, 13] that TIP4P/2005 behavior is even better than that of SPC/E. Therefore, in this work we use the TIP4P/2005 model. Another extension TIP5P was proposed by Mahoney *et al.* [8] to carry the negative charge on two auxiliary atoms. However, the performance is not better than of TIP4P/2005.

TIP4P/2005 model includes Lennard–Jones interactions between oxygen atoms only; hydrogens have negligible mass compared to them, which makes the simulation easier. Other interaction is electrostatic which occurs between hydrogen H and charge M atoms. TIP4P/2005 parameters are listed in Tab. 1.

3 Simulation methods

The simulation was performed as follows: first, a liquid cubic box of 1372 molecules was run for 50 ps, then the z-size of the box was expanded to approximately 3× the original proportion and run for 10 ns to provide reliable data for surface tension determination. Sizes of the box were calculated depending on the NIST [7] values of the water saturated liquid density for particular temperature. For supercooled region, a constant box size corresponding to 300 K system was used. Periodic boundary conditions were used in all directions. Timestep of the simulation was chosen as 2 fs (same as in [11]) with velocity Verlet integrator. To maintain constant temperature, the Nosé–Hoover thermostat was used with relaxation constant 100 fs. Cutoffs of Lennard–Jones interactions and van der Waals forces were set to 14.5 Å. For electrostatic interactions, direct Ewald method was used, with automatic parameter optimization constant set to 10^{-5}. Density was computed as a histogram in z-direction in every step and averaging through the time. Examples of the density profiles converted to g/cm^3 for 300 and 500 K can be seen in Fig. 1 depicted by the solid lines. Snapshots of the simulations for two temperatures are shown in Fig. reffig2.

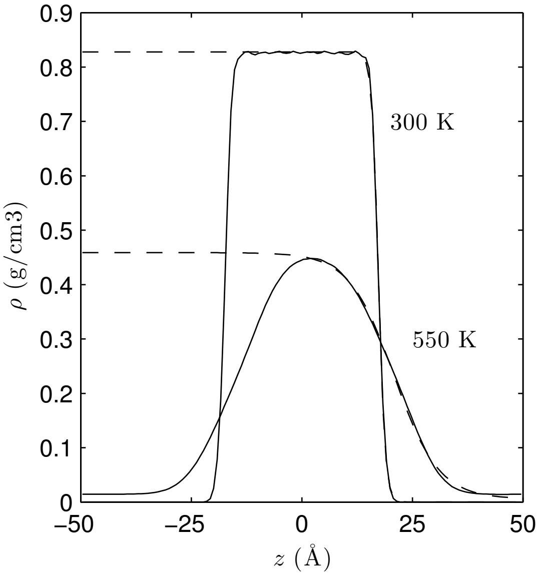


Figure 1: Density profiles ρ as functions of z-coordinate for systems having temperatures $T = 300$ K and 500 K. Solid lines are time-averaged profiles obtained from the simulations, dashed lines are fits of the right-hand profiles (liquid - vapor) to the hyperbolic tangent profile, Eq. (1).

The density profiles were divided to two halves approximately in the centre of the simulation box. One half of the density profile was subsequently fitted to a hyperbolic tangent density profile model,

$$\rho(z) = \frac{\rho_{\mathrm{L}} + \rho_{\mathrm{V}}}{2} - \frac{\rho_{\mathrm{L}} - \rho_{\mathrm{V}}}{2} \tanh\left(\frac{z - z_0}{d}\right), \tag{1}$$

where ρ_{L} and ρ_{V} are fitted bulk densities, z_0 is the position of the Gibbs dividing surface of the interface, d is the parameter for the thickness of the interface. The results of the fit are depicted as the dashed lines in Fig. 1. The fitted values were used used to evaluate the surface tension γ in the following manner:

$$\begin{aligned}\gamma &= \frac{L_{\mathrm{z}}}{2}\left(P_{\mathrm{zz}} - \frac{P_{\mathrm{xx}} + P_{\mathrm{yy}}}{2}\right) + 12\pi\epsilon\sigma^6(\rho_{\mathrm{L}} - \rho_{\mathrm{V}})^2 \\ &\times \int_0^1 \int_{r_{\mathrm{c}}}^{\infty} \coth\left(\frac{rs}{d}\right) \frac{3s^3 - s}{r^3} \mathrm{d}r \mathrm{d}s. \end{aligned} \tag{2}$$

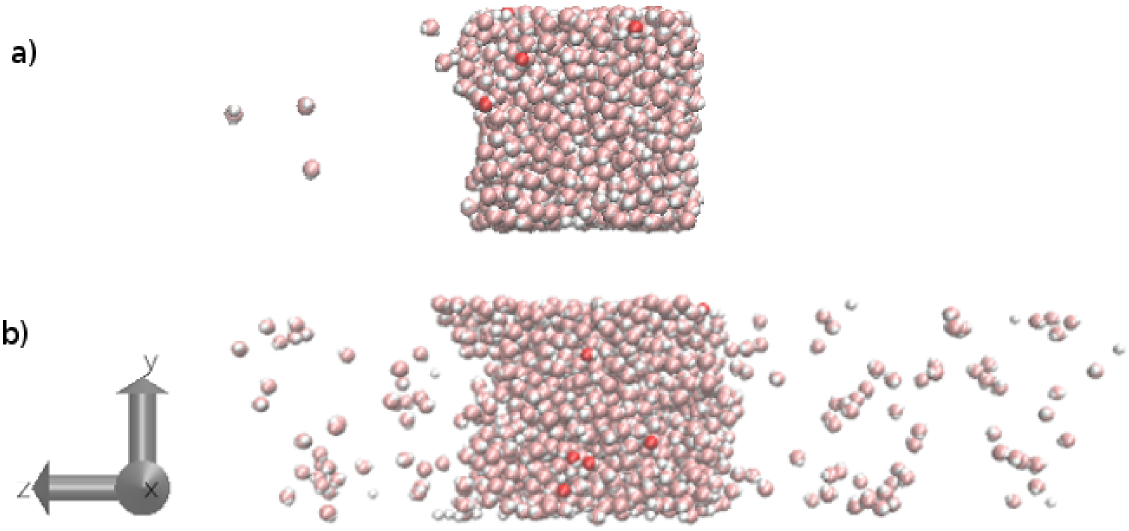


Figure 2: Snapshots of two configurations during the simulations for two temperatures, a) $T = 300$ K and b) 550 K. Liquid cubes are in the middle, vapor is on the left and right side (not so apparent for lower temperature).

In Eq. (2), L_z denotes the box size in z direction, P_ii is the ii-th component of the pressure tensor, ϵ and σ are the Lennard–Jones parameters for oxygen atom, and r_c is the cutoff for the Lennard–Jones potential. Their values as well as other model parameters are summarized in Tab. 1. Second term in Eq. (2) corresponds to the Lennard–Jones tail correction [2].

4 Results

Insipired by the work of Sakamaki *et al.* [11], molecular dynamics simulations of a water slab enclosed by the vapor were performed for temperatures T = 250 K, 270 K, 275 K, 300 K, 350 K, 400 K, 450 K, 500 K, 550 K, and 600 K.

Given by the parameters stated in previous section, computations ran approximately 3 days if running on all 24 CPUs. Melting temperature for TIP4P/2005 is $T_\mathrm{m} = 249$ K, therefore even for lowest temperature of 250 K we did not encounter any of the liquid water during the simulation.

Figure 2 shows an example of two configurations for two temperatures (300 K, 550 K). The liquid phase persisted in the centre of the simulation box, while the vapor phase gradually expanded into the vacuum space after the box got stretched in the z-direction. As can be seen at low temperatures, the molecule escaping from the liquid phase into the vapor phase was rather rare event. On the other hand at the elevated temperatures, the vapor phase got significantly denser.

Surface tension computed using Eq. (2) is shown in Fig. 3 as circles, the IAPWS values [5, 7] are shown as a solid line. Simulated values nicely describe the reference data, despite the disagreement with the bulk density values.

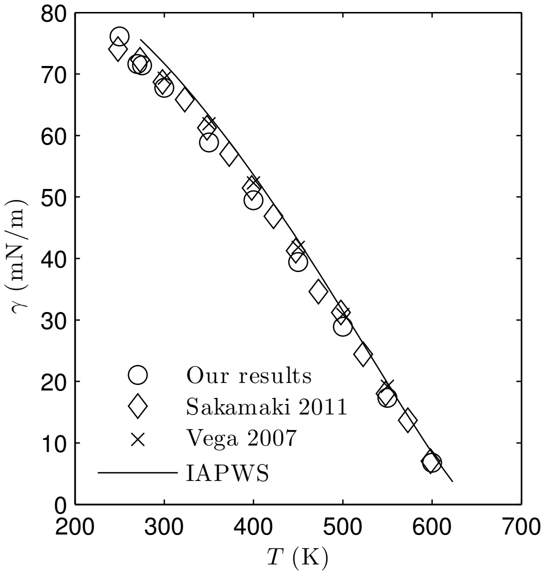


Figure 3: Surface tension γ as a function of temperature T as predicted by this work, computed using Eq. (2) (circles), compared to the IAPWS values [5, 7] (solid line).

5 Conclusions

In this work MD simulations were performed for water for various temperatures ranging from 250 K to 600 K. As a water model TIP4P/2005 was used which is probably the best rigid non-polarizable water model available at the moment. The surface tension was computed as a preliminary result.

In future, we would like to perform more simulations in the supercooled region of liquid water using the TIP4P/2005 and the SPC/E models to compare simulated results with our recent experiments. [4].

Also, we would like to model the so-called capillary waves contribution to the surface tension, i.e. to simulate, how the thermal motion of molecules affect (lower) the surface tension for planar phase interface. New molecular simulations will support our theoretical work [15, 9, 16, 10, 3].

References

[1] J. L. Abascal and C. Vega. *A general purpose model for the condensed phases of water: Tip4p/2005.* The Journal of chemical physics **123** (2005), 234505.

[2] G. A. Chapela, G. Saville, S. M. Thompson, and J. S. Rowlinson. *Computer simulation of a gas–liquid surface. part 1.* J. Chem. Soc., Faraday Trans. II **73** (1977), 1133–1144.

[3] J. Hrubý, B. Planková, and V. Vinš. *Corrections to the classical work of formation of critical clusters.* In 'Nucl. and Atmos. Aerosols: 19th Int. Conf.', (2013).

[4] J. Hrubý, V. Vinš, R. Mareš, J. Hykl, and J. Kalová. *Surface tension of supercooled water: No inflection point down to -25°c.* J. Phys. Chem. Lett. **5** (2014), 425–428.

[5] International Association for the Properties of Water and Steam. *IAPWS release on surface tension of ordinary water substance*, (1994). URL: `http://www.iapws.org/`.

[6] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics **79** (1983), 926–935.

[7] E. Lemmon, M. McLinden, and D. Friend. *Thermophysical properties of fluid systems.* In 'NIST Chemistry WebBook, NIST Standard Reference Database Number 69', P. Linstrom and W. Mallard, (eds.), National Institute of Standards and Technology (2014 (retrieved September 10)).

[8] M. W. Mahoney and W. L. Jorgensen. *A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions.* The Journal of Chemical Physics **112** (2000), 8910–8922.

[9] B. Planková, J. Hrubý, and V. Vinš. *Homogeneous droplet nucleation modeled using the gradient theory combined with the pc-saft equation of state.* In 'EPJ Web of Conferences', volume 45, 01076. EDP Sciences, (2013).

[10] B. Planková, J. Hrubý, and V. Vinš. *Prediction of the homogeneous droplet nucleation by the density gradient theory and pc-saft equation of state.* In 'Nucl. and Atmos. Aerosols: 19th Int. Conf.', (2013).

[11] R. Sakamaki, A. K. Sum, T. Narumi, and K. Yasuoka. *Molecular dynamics simulations of vapor/liquid coexistence using the nonpolarizable water models.* J. Chem. Phys. **134** (2011).

[12] C. Vega and E. De Miguel. *Surface tension of the most popular models of water by using the test-area simulation method.* J. Chem. Phys. **126** (2007), 154707.

[13] C. Vega and J. L. Abascal. *Simulating water with rigid non-polarizable models: a general perspective.* Physical Chemistry Chemical Physics **13** (2011), 19663–19688.

[14] C. Vega, J. L. Abascal, M. Conde, and J. Aragones. *What ice can teach us about water interactions: a critical comparison of the performance of different water models.* Faraday discussions **141** (2009), 251–276.

[15] V. Vinš, J. Hrubý, and B. Planková. *Proceedings of the international conference.* In 'Experimental Fluid Mechanics 2010', Liberec, (November 24-26 2010).

[16] V. Vinš, J. Hrubý, and B. Planková. *Surface tension of binary mixtures including polar components modeled by the density gradient theory combined with the pc-saft equation of state.* International Journal of Thermophysics **34** (2013), 792–812.

Definition of New Features for Automated Java Source Code Classification*

Michal Rost

4th year of PGS, email: rost.michal@gmail.com
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Miroslav Virius, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This paper reports on progress in the development of patterns and feature space for automated recognition of well-designed data types in software source code. It summarizes work from previous year, deals with formulation of new patterns, and describes new testing data set. Last but not least, new set of features is proposed at the end of this paper, together with reasons that led to formulation of this proposal.

Keywords: Java, code analysis, design patterns

Abstrakt. Článek informuje o postupu práce na vytváření vzorů a příznakového prostoru pro automatickou detekci známých struktur ve zdrojovém kódu softwarových projektů, shrnuje kroky provedené v posledním roce a představuje nové vzory a testovací datovou sadu. Ve svém závěru se článek věnuje formulaci nové skupiny příznaků a uvádí důvody, které k tomuto kroku vedly.

Klíčová slova: Java, analýza zdrojového kódu, návrhové vzory

1 Introduction

Automated source code patterns recognition efforts date back to 1998, when Antoniol et al. [1] focused on detection of five structural design patterns [3, 9] in C++ source code. This approach consisted of four steps: AOL (Abstract Object Language) representation extraction, AOL representation parsing, class metrics extraction and pattern recognition. In the third step, metrics were collected from AOL representation of each particular class of the analyzed source code. These metrics comprised number of private/public/protected attributes and operations, as well as number of associations, aggregations and inheritance. Last step represented a multi-stage process in which a set of constraints were gradually applied to collected observations in order to filter required patterns.

Later In 2004, Guéhéneuc et al., inspired by Antoniol's work, continued on improvement of feature space [4]. Guéhéneuc's team presented an improved metric space divided into four categories: size (number of methods or fields), filiation (number of parents or children), cohesion (degree to which class features belong together), and coupling (strength of associations among classes). Metrics were collected for individual classes that participate in design patterns, with regard to the fact that each class can act in different role in various patterns. Based on metrics fingerprints were introduced; each

[*] This work has been supported by the grants SGS 11/167/OHK4/3T/14 and LA08015

fingerprint corresponded to a specific role of class in a particular design pattern, and together they formed a set of rules for individual patterns. These rules were mined from a repository, created from source codes of various software projects.

Ferenc et al., in contrast to Guéhéneuc, used machine learning algorithms to detect design pattern as a whole instead of individual classes [2]. They formulated a list of predictors for each design pattern to depict its unique properties. During the learning process predictor values were collected from ASG representation of source code, consequently decision trees and backpropagation neural networks were used to create model files with acquired knowledge.

Lerthathairat and Prompoon came up [5] with the idea to classify a given source code to clean, bad, or ambiguous using standard software metrics [6] and fuzzy logic. Moreover they presented a design of tool for suggestion of refactoring techniques for ambiguous code.

Our approach [7, 8, 10, 11] is to classify individual classes as patterns that could (but not neccessarily do) represent fragments of design patterns. We believe that if a part of given software project is composed from well designed data types (classes), then these particular classes could together form a higher-level stucture - a design pattern, which might be revealed in the subsequent analysis. In other words, we intend to detect well designed data types (design pattern fragments, UML stereotypes) in source code over poorly designed classes (a noise); However, in contrast to [5] we have more than one classification pattern for the “clean” (well designed) code.

2 Materials and methods

2.1 Previous work

2.1.1 Pattern identification

In the year 2013, a number of software projects was examined and representatives of well-designed classes (data types) were identified. Consequently, a list of patterns was introduced. This list contained 11 patterns including: adapter, bean, builder, composite, constant, dao, decorator, factory, proxy, utility, and worker. For detailed explanation of these patterns see [10].

2.1.2 Data sets collection

Later, training and testing data sets were prepared. Training data set consisted of 175 java source files that were selected from different open source projects or design pattern tutorials. A special effort was devoted to cover a wide range of possible implementations of each pattern. Testing data sets were represented by three open source projects: *JaHoCa*, *JHotDraw*, and *AndEngine*. Files in all mentioned sets were manually examined and classified. For detailed explanation of data sets see [7].

2.1.3 Feature space definition and collection

After acquisition of data, a metric space was developed; it contained over 40 features in four categories: *expression*, *statement*, *member*, and *relation*. Brief description of features can be found in [10], for more detailed information see [8]. Collection of features from source code was also a non-trivial task, an abstract syntax tree (AST) representation of source code was utilized and two different approaches for AST querying were implemented. Brief description of feature collection using *XQuery* and *Groovy* can be found in [11].

2.1.4 Classifiers implementation

In 2014, more than ten classifiers were implemented by Matej Mojzeš and Josef Smolka. These classifiers comprised, for example, of: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Parzen Discriminant Analysis (PDA), or Desission Tree (DT). In order to reduce feature space dimensions submodels were introduced. Since there were $2^{40} - 1$ sub-models, it was hardly possible to search systematically for the optimal one. Therefore FSA heuristic has been used for finding the best sub-model; the heuristic has been applied repeatedly for each classifier and best results from each run have been recorded. Finally, a list of frequently used features in the most successful submodels has been created. In the general the member fatures performed the best. The list together with classifiers explanation can be seen in [8].

2.1.5 Software design and development

Paralelly to conducted research a cross-platform tool for source code analysis of Java software projects was developed. Requirements for mentioned tool were formulated in [10] together with tool design in which four main tool components were proposed: *collector*, *classifier*, *validator* and *launcher*.

Currently all four major components are implemented and the tool is functional. User is provided with both text and graphical interface, where he can specify a location of project for validation, or locations of train and test projects for classification; user can also choose a method of feature collection, required validator, or classifier. Data of each project are collected into the *ProjectObservation* class instance which serves as storage for feature values of every class file in project's source code - *TypeObservation*. Moreover, a user classification can be loaded for training purposes. An instance of *ProjectObservation* can be injected into selected classifier or validator for subsequent analysis. Once analysis is performed, results from project's classification are saved into *ProjectClassification* object, while results from project's validation are kept in the *ProjectValidation* instance. Both classes share same interface, so they can be treated in the same way. For example, all results can be exported to CSV in order to be published or archived. Relationships among the mentioned classes are depicted in the Figure 1.

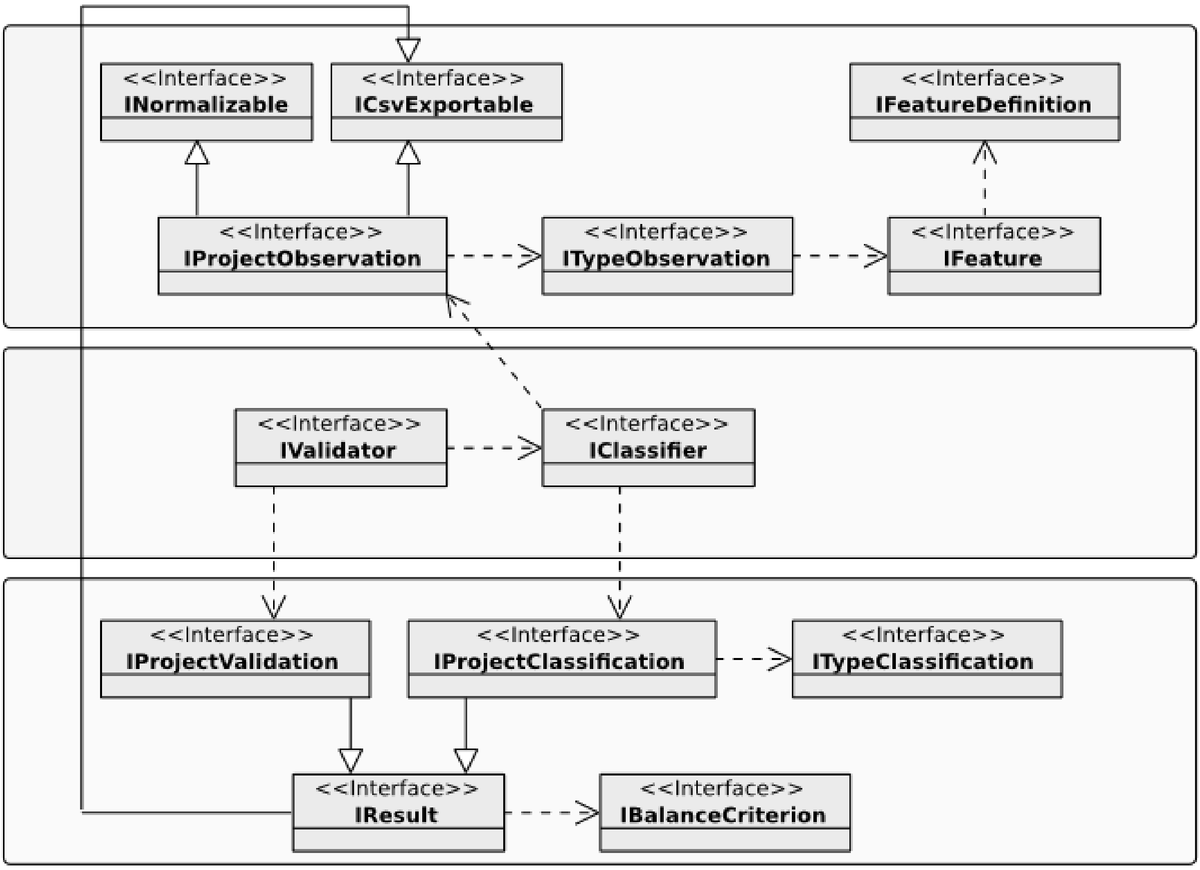


Figure 1: A class diagram of developed tool's interfaces.

2.2 New patterns

2.2.1 Implementation pattern

During analysis of the results from test projects classification, a question has arisen, if worker pattern is not overly generic. There were many classes that were classified as worker pattern, but did not satisfy its definition: "A worker uses other objects in order to perform the core logic of a certain part of the application." After the analysis of badly classified data types, it was decided to introduce a new pattern: implementation. It could be, for instance, a simple abstract class that implements certain interface, or a class that overrides methods of its parent. The important point is that these classes do not introduce new attributes; they only implement or reimplement simple logic.

2.2.2 Read only bean pattern

It was concluded that it is quite common that software projects contain beans that can be accessed solely in read-only mode. These beans generally contain only getter methods and no setters, instead their data are injected through constructors. For this reason a rbean pattern was introduced.

Table 1: Classification results for Neuroph project (SVM).

Pattern	*Success* [%]	*TOP misinterpreted as*	*In* [%]
utility	95	factory	5
worker	87	dao	4
factory	71	utility	29
rbean	67	composite	33
implementation	57	adapter	17
bean	40	worker	53
adapter	0	worker	63

2.3 Data sets update

As new patterns were introduced, data sets had to be updated. Examples of implementation and rbean patterns were added to training data set together with new representants of worker, factory and proxy patterns. Training data set now contains 230 java classes.

2.3.1 Neuroph project

The set of test projects was extended with a new project - Neuroph. Neuroph is neural network framework, which can be used to create and train neural networks in Java applications. Neuroph code is very clear and contains a large number of patterns, including: adapters, beans, rbeans, workers, implementations, utilities, factories and other. Approximately 200 classes of neuroph project were manually classified and prepared for analysis.

3 Results

After numerous classifications of test projects (including Neuroph) were performed, various results were observed. In case of Neuroph project acceptable results were achieved by SVM classifier for utility, worker, and factory patterns. However, results for bean and adapter were unsatisfactory. All results can be seen in Table 1. Dao, builder, proxy and constant patterns are not listed in the table, because of the low incidence of these patterns in the project. Third column of the table contains the most frequent pattern that was misinterpreted as valid one, the frequency is recorded in the last column.

In the training data, bean patterns consisted purely of attributes, setters, getters, and sometimes of constructors. However, in the reality, for example in the Neuroph project, many beans can have additional methods that are used, for instance, for: serialization, minor computations, or attribute derivation. This fact causes that the classification strongly depends on how many pure setters and getters are in the classified bean. If there is a lack of setters/getters, then this bean can be classified as worker.

Similar issue was recorded for the adapter and the proxy patterns. In such patterns, the majority of methods usually mediate access to an adaptee or a proxied object. Important role for the worker pattern detection plays feature *fm#amr* [10] that expresses

Table 2: Interesting method types according to accessibility aspect.

	∃ parameter	*∃ attribute*	*non-void output*
Transformation	Yes	No	Yes
Derivation	No	Yes	Yes
Computation	No	Yes	No
Addition	Yes	Yes	No
Combination	Yes	Yes	Yes

how much a selected class uses its attributes in its methods. This feature reaches high values for both worker's "computation" methods and adapter's "mediator" methods.

4 Discussion

Unsatisfactory results for bean and adapter were caused by fact that the current features do not sufficiently take into account behavior of methods, but mostly depend on their external properties. The current version of feature space tries to cover all possible combinations of method properties like static/non-static, abstract/non-abstract, private/public/protected; nevertheless, purpose of these methods is not considered. That is the reason, why an improvement of feature space was proposed.

4.1 Feature space improvement proposal

Based on repeated analysis of Neuroph classification results an idea was born to divide methods classes by two fundamental aspects: *method accessibility* and *method purpose.*

4.1.1 Method accessibility

Accessibility aspect takes into account inputs and outputs of methods, thus it is based on three pieces of logical information: if a method has at least one parameter, if method uses at least one attribute of its owning class, and if method returns any value. Five frequent combinations of these properties were identified and named according to their role (Table 2). *Transformation* method transforms its parameters to an output without using owning class attributes. *Derivation* method derives an extra information merely from the class attributes and returns it as a result. *Computation* method, in contrast to the derivation, stores computed result locally in the owning class. During *addition* an external information is added to the method through the parameters in order to update the owning class. In the case of *combination* a parameter is combined with local attributes to produce an output. A scope for accessibility aspect are public, non-static, non-abstract methods that are neither setters nor getters.

4.1.2 Method purpose

The second - purpose point of view deals with meaning of a code inside the method definitions. So far, six examples of this aspect were named: *mediation*, *usage*, *recursion*, *production*, *creation* and *build*. If at least one member method of an attribute is called from within other class method, it is called mediation. On the other hand, when a member method invokes at least one other method of same class, it is called usage. A special case when a method calls itself was named as recursion. Production is when a method introduces a local variable which is instantiated within this method and returned as a result. However, creation is when local variable is created a returned without instantiation. Last but not least, if an attribute is instantiated within a method, we call it build. A scope for purpose aspect are public, non-abstract methods that are neither setters nor getters.

Since every important method should be described as combination of one accessibility aspect and one or more purpose aspects, we belive that these new features will help us to better classify our patterns.

5 Conclusion

The paper dealt with automated analysis of software project's source code. At first brief overview of the current state of the art was given; subsequently, our progress in this topic was reported and description of developed software tool was provided. Next, two new patterns the *implementation* and the *read only bean* were introduced, together with the new testing data set - the Neuroph project. In the following section, results from classification of Neuroph project were presented. Finally, in the discussion section, an improvement of current feature space was proposed in order to capture the purpose of class methods.

In the nearest future we will concentrate on the implementation of collectors for the newly proposed features. Once they will be collected, we will continue on reduction of feature space dimensions. Additionally, we would like to employ graph theory algorithms to perform a posterior analysis of improperly designed data types that do not match any standard pattern.

References

[1] G. Antoniol, R. Fiutem, and L. Cristoforetti. *Design Pattern Recovery in Object-Oriented Software.* In 'Proceedings of the 6th International Workshop on Program Comprehension', IEEE Computer Society, (1998), 153–160.

[2] R. Ferenc, Á. Beszédes, L. Fülöp and J. Lele. *Design Pattern Mining Enhanced by Machine Learning.* In 'Proceedings International Conference on Software Maintenance', IEEE Computer Society, (2005), 295–304.

[3] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Pattern Book.* Addison-Wesley Professional, (1994).

[4] Y. G. Gueheneuc, H. Sahraoui, and F. Zaidi. *Fingerprinting Design Patterns* In 'Proceedings of the 11th Working Conference on Reverse Engineering', IEEE Computer Society, (2004), 172–181.

[5] P. Lerthathairat and N. Prompoon. *An Approach for Source Code Classification to Enhance Maintainability.* In 'Proceedings Eighth International Joint Conference on Computer Science and Software', Engineering', IEEE Computer Society, (2011), 319–324.

[6] M. V. Mäntylä and C. Lassenius. *Subjective Evaluation of Software Evolvability Using Code Smells: An Empirical Study.* In 'Journal of Empirical Software Engineering', Springer, volume 11, issue 3, (2006), 395–431.

[7] M. Mojzeš, M. Rost, J. Smolka, and M. Virius. *Tool for Statistical Classification of Java Projects.* In 'Proceedings of the 1st Conference on Software Development and Object Technologies', Praha, (2013), 149–156.

[8] M. Mojzeš, M. Rost, J. Smolka, and M. Virius. *Feature Space for Statistical Classification of Java Source Code Patterns.* In 'Proceedings of the 15th International Carpathian Control Conference (ICCC)', IEEE Computer Society, (2014), 357–361.

[9] R. Pecinovský. *Návrhové vzory.* Computer Press, (2007).

[10] M. Rost. *Feature Definition and Software Design for Java Source Code Classification Tool.* In 'Proceedings Doktorandské dny 2013', Czech Technical University, (2013), 235–243.

[11] J. Smolka. *Feature Collection for Source Code Classification and Pattern Recognition.* In 'Proceedings Doktorandské dny 2013', Czech Technical University, (2013), 255–262.

Ambiguity of Covariantized Noether Identities*

Josef Schmidt

4th year of PGS, email: schmijos@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jiří Bičák, Institute of Theoretical Physics
Faculty of Mathematics and Physics, Charles University in Prague

Abstract. We present a new kind of ambiguity in Noether identity generated by dipheomorphism invariance due to the covariantization procedure. The problem arising from non-commutativity of covariant derivatives is illustrated on simple examples of Lagrangian linear and quadratic in second derivatives of metric tensor. General case is then studied and covariant Klein identities are used for proving the conservation of Noether current.

Keywords: Noether identities, dipheomorphism invariance, Lagrangian field theory, Einstein-Hilbert action, Gauss-Bonnet gravity, Klein identities

Abstrakt. Prezentujeme nový druh nejednoznačnosti identity Noetherové generované invariancí vůči difeomorfismům, který je důsledkem kovariantizační procedury. Problém vyvstávající kvůli obecné nekomutativitě kovariantních derivací je ilustrován na jednoduchých příkladech s Lagrangiány, které jsou lineární a kvadratické v druhých derivacích metriky. Je zkoumán obecný případ a v něm jsou kovariantní Kleinovy identity použity k důkazu zachování Noetherovského proudu.

Klíčová slova: identity Noetherové, invariance vůči difeomorfismům, Lagrangeovská polní teorie, Einstein-Hilbertova akce, Gauss-Bonnetova gravitace, Kleinovy identity

1 Introduction

In [1] a general method for obtaining covariantized Noether identities stemming from dipheomorphism invariance for second order Lagrangian using auxiliary metric was developed. Two classes of conserved currents were introduced based on switching order of the second covariant derivatives which do not commute in general. In this paper we would like to show some conceptual problems with this approach.

2 Linear Lagrangian

Let's consider a Lagrangian with metric as a dynamical field, linear in second derivatives, i.e.

$$\hat{\mathcal{L}}(g_{mn}; g_{mn,a}; g_{mn,ab}) = \sqrt{-g}\left[P^{\mu\nu\alpha\beta}(g_{mn}; g_{mn,a})\, g_{\mu\nu,\alpha\beta} + Q(g_{mn}; g_{mn,a})\right]. \qquad (1)$$

[*] This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS13/217/OHK4/3T/14

$P^{\mu\nu\alpha\beta}$ and Q are arbitrary functions of metric and its first derivatives. Now we introduce auxiliary (background) metric $\bar{g}_{\mu\nu}$ and covariantize the Lagrangian, i.e. rewrite the partial derivatives[1] as

$$\begin{aligned} g_{\mu\nu,\alpha} &= g_{\mu\nu\bar{;}\alpha} + 2\bar{\Gamma}^{\sigma}_{\ \alpha(\mu} g_{\nu)\sigma}, \qquad (2) \\ g_{\mu\nu,\alpha\beta} &= g_{\mu\nu\bar{;}\alpha\beta} + 4\bar{\Gamma}^{\rho}_{\ \alpha)(\mu} g_{\nu)\rho\bar{;}(\beta} + \bar{\Gamma}^{\rho}_{\ \alpha\beta} g_{\mu\nu\bar{;}\rho} + 2g_{\rho(\mu}\bar{\Gamma}^{\rho}_{\ \nu)\alpha,\beta} + 2\bar{\Gamma}^{\rho}_{\ \alpha(\mu}\bar{\Gamma}^{\sigma}_{\ \nu)\beta} g_{\rho\sigma} + 2g_{\rho(\mu}\bar{\Gamma}^{\sigma}_{\ \nu)\alpha}\bar{\Gamma}^{\rho}_{\ \beta\sigma} \\ &= g_{\mu\nu\bar{;}\alpha\beta} + K_{\mu\nu\alpha\beta}(g_{mn}; g_{mn\bar{;}a}; \bar{g}_{mn}; \bar{g}_{mn,a}; \bar{g}_{mn,ab}) \qquad (3) \end{aligned}$$

and we get the new Lagrangian

$$\hat{\mathcal{L}}^{*}(g_{mn}; g_{mn\bar{;}a}; g_{mn\bar{;}ab}; \bar{g}_{mn}; \bar{g}_{mn,a}; \bar{g}_{mn,ab}) = \hat{\mathcal{L}}(g_{mn}; g_{\mu\nu\bar{;}\alpha} + 2\bar{\Gamma}^{\sigma}_{\ \alpha(\mu} g_{\nu)\sigma}; g_{\mu\nu\bar{;}\alpha\beta} + K_{\mu\nu\alpha\beta}), \quad (4)$$

more elaborately

$$\begin{aligned} \hat{\mathcal{L}}^{*} &= \sqrt{-g}\,\big[P^{\mu\nu\alpha\beta}(g_{mn}; g_{\mu\nu\bar{;}\alpha} + 2\bar{\Gamma}^{\sigma}_{\ \alpha(\mu} g_{\nu)\sigma})\,[g_{\mu\nu\bar{;}\alpha\beta} + K_{\mu\nu\alpha\beta}(g_{mn}; g_{mn\bar{;}a}; \bar{g}_{mn}; \bar{g}_{mn,a}; \bar{g}_{mn,ab})] \\ &\qquad + Q(g_{mn}; g_{\mu\nu\bar{;}\alpha} + 2\bar{\Gamma}^{\sigma}_{\ \alpha(\mu} g_{\nu)\sigma})\big] \\ &= \sqrt{-g}\left[\tilde{P}^{\mu\nu\alpha\beta}\, g_{\mu\nu\bar{;}\alpha\beta} + \tilde{P}^{\mu\nu\alpha\beta} K_{\mu\nu\alpha\beta} + \tilde{Q}\right], \qquad (5) \end{aligned}$$

where we denoted

$$\begin{aligned} \tilde{P}^{\mu\nu\alpha\beta}(g_{mn}; g_{mn\bar{;}a}; \bar{g}_{mn}; \bar{g}_{mn,a}) &= P^{\mu\nu\alpha\beta}(g_{mn}; g_{\mu\nu\bar{;}\alpha} + 2\bar{\Gamma}^{\sigma}_{\ \alpha(\mu} g_{\nu)\sigma}), \qquad (6) \\ \tilde{Q}(g_{mn}; g_{mn\bar{;}a}; \bar{g}_{mn}; \bar{g}_{mn,a}) &= Q(g_{mn}; g_{\mu\nu\bar{;}\alpha} + 2\bar{\Gamma}^{\sigma}_{\ \alpha(\mu} g_{\nu)\sigma}). \qquad (7) \end{aligned}$$

The partial derivative with respect to the second covariant derivatives of the covariantized Lagrangian is then simply

$$\frac{\partial \hat{\mathcal{L}}^{*}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = \sqrt{-g}\,\tilde{P}^{\mu\nu\alpha\beta}. \qquad (8)$$

2.1 Properties of $P^{\mu\nu\alpha\beta}$ (or $\tilde{P}^{\mu\nu\alpha\beta}$)

As partial derivatives commutes the only reasonable choice of $P^{\mu\nu\alpha\beta}$ is the one which is symmetrical in (α, β), $P^{\mu\nu\alpha\beta} = P^{\mu\nu(\alpha\beta)}$. No matter how do we choose antisymmetric part $P^{\mu\nu[\alpha\beta]}$ the Lagrangian as a function remains the same because of the trivial identity $P^{\mu\nu[\alpha\beta]} g_{\mu\nu,\alpha\beta} = 0$. Let's say that a physical theory is defined by its Lagrangian. Then $P^{\mu\nu[\alpha\beta]}$ has no effect on the theory – hence it is irrelevant. This has the important consequence on the covariantized Lagrangian. Generally, covariant derivatives do not commute, so it would seem that our covariantization procedure depends on the order of covariant derivatives as $g_{\mu\nu\bar{;}\alpha\beta} = g_{\mu\nu\bar{;}\beta\alpha} + 2g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta}$, but we have $P^{\mu\nu\alpha\beta} = P^{\mu\nu(\alpha\beta)}$ and consequently

$$P^{\mu\nu(\alpha\beta)} g_{\mu\nu,\alpha\beta} = P^{\mu\nu(\alpha\beta)}\,(g_{\mu\nu\bar{;}\alpha\beta} + K_{\mu\nu\alpha\beta}) = P^{\mu\nu\alpha\beta}\,\big(g_{\mu\nu\bar{;}(\alpha\beta)} + K_{\mu\nu(\alpha\beta)}\big)\,. \qquad (9)$$

[1] The auxiliary metric $\bar{g}$ defines the Riemann-Levi-Civita connection $\bar{\Gamma}$ and corresponding covariant derivative $\bar{\nabla}$ or in index notation $\bar{;}$.

So the covariant derivatives appears only in symmetric combination and the ambiguous antisymmetric part, which either can be or need not to be converted to Riemann tensor, vanishes. In the end the partial derivative is

$$\frac{\partial \hat{\mathcal{L}}^*}{\partial g_{\mu\nu;\alpha\beta}} = \frac{\partial \hat{\mathcal{L}}^*}{\partial g_{\mu\nu;\beta\alpha}} = \sqrt{-g}\, \tilde{P}^{\mu\nu(\alpha\beta)} = \sqrt{-g}\, \tilde{P}^{\mu\nu\alpha\beta}.$$

The analogous situation is with symmetricity of metric field $g_{\mu\nu}$: $P^{\mu\nu\alpha\beta} = P^{(\mu\nu)\alpha\beta}$.

Let's review what will happen if we insist on keeping the antisymmetric part of $P^{\mu\nu\alpha\beta}$ in Lagrangian as is done in [1]. We have

$$\begin{aligned} P^{\mu\nu\alpha\beta} g_{\mu\nu,\alpha\beta} &= \left(P^{\mu\nu(\alpha\beta)} + P^{\mu\nu[\alpha\beta]}\right)\left(g_{\mu\nu;\alpha\beta} + K_{\mu\nu\alpha\beta}\right) &(10)\\ &= P^{\mu\nu(\alpha\beta)}\left(g_{\mu\nu;(\alpha\beta)} + K_{\mu\nu(\alpha\beta)}\right) + P^{\mu\nu[\alpha\beta]}\left(g_{\mu\nu;[\alpha\beta]} + K_{\mu\nu[\alpha\beta]}\right). &(11)\end{aligned}$$

It can be easily checked that $K_{\mu\nu[\alpha\beta]} = -g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta}$, hence

$$P^{\mu\nu\alpha\beta} g_{\mu\nu,\alpha\beta} = P^{\mu\nu\alpha\beta}\, g_{\mu\nu;\alpha\beta} + P^{\mu\nu(\alpha\beta)} K_{\mu\nu(\alpha\beta)} - P^{\mu\nu[\alpha\beta]} g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta}. \tag{12}$$

It should be emphasized that antisymmetric part $P^{\mu\nu[\alpha\beta]}$ can be completely arbitrary as it does not contribute into Lagrangian at all. In the end we got partial derivatives (arbitrarily) depending on the order of covariant derivatives

$$\frac{\partial \hat{\mathcal{L}}^*}{\partial g_{\mu\nu;\alpha\beta}} = \sqrt{-g}\, \tilde{P}^{\mu\nu\alpha\beta}, \tag{13}$$

$$\frac{\partial \hat{\mathcal{L}}^*}{\partial g_{\mu\nu;\alpha\beta}} - \frac{\partial \hat{\mathcal{L}}^*}{\partial g_{\mu\nu;\beta\alpha}} = 2\sqrt{-g}\, \tilde{P}^{\mu\nu[\alpha\beta]}. \tag{14}$$

On the other hand, if antisymmetrized covariant derivatives are converted into background Riemann tensor, $g_{\mu\nu;[\alpha\beta]} = g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta}$, the arbitrariness originating from $P^{\mu\nu[\alpha\beta]}$ cancel out

$$P^{\mu\nu\alpha\beta} g_{\mu\nu,\alpha\beta} = P^{\mu\nu(\alpha\beta)} g_{\mu\nu;(\alpha\beta)} + P^{\mu\nu[\alpha\beta]} g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta} + P^{\mu\nu(\alpha\beta)} K_{\mu\nu(\alpha\beta)} - P^{\mu\nu[\alpha\beta]} g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta} \tag{15}$$

leading to unambiguous expression (9).

At last we can switch the order of covariant derivatives in (12) via $g_{\mu\nu;\alpha\beta} = g_{\mu\nu;\beta\alpha} + 2g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta}$ to get

$$P^{\mu\nu\alpha\beta} g_{\mu\nu,\alpha\beta} = P^{\mu\nu\beta\alpha} g_{\mu\nu;\alpha\beta} + P^{\mu\nu(\alpha\beta)} K_{\mu\nu(\alpha\beta)} + P^{\mu\nu[\alpha\beta]} g_{\rho(\mu}\bar{R}^{\rho}{}_{\nu)\alpha\beta}. \tag{16}$$

Again, converting antisymmetric part of covariant derivative leads to unambiguous expression (9).

2.2 Example of Einstein-Hilbert action

The Einstein-Hilbert action has the exact form of our sample Lagrangian (1). In fact it is the only possible choice if we demand the general covariance. Let's find the symbol

$P^{\mu\nu\alpha\beta}$ in this particular case. If one writes the Riemann tensor as

$$\begin{aligned} R^{\lambda}{}_{\tau\rho\sigma} &= \Gamma^{\lambda}_{\tau\sigma,\rho} - \Gamma^{\lambda}_{\tau\rho,\sigma} + \Gamma^{\lambda}_{\rho\eta}\Gamma^{\eta}_{\tau\sigma} - \Gamma^{\lambda}_{\eta\sigma}\Gamma^{\eta}_{\tau\rho} \\ &= \frac{1}{2}g^{\lambda\iota}\left(g_{\iota\tau,\sigma\rho} + g_{\iota\sigma,\tau\rho} - g_{\tau\sigma,\iota\rho} - g_{\iota\tau,\rho\sigma} - g_{\iota\tau,\rho\sigma} + g_{\tau\rho,\iota\sigma}\right) + \Gamma^{\lambda}_{\rho\eta}\Gamma^{\eta}_{\tau\sigma} - \Gamma^{\lambda}_{\eta\sigma}\Gamma^{\eta}_{\tau\rho} \\ &= P^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta} g_{\mu\nu,\alpha\beta} + Q^{\lambda}{}_{\tau\rho\sigma}, \end{aligned} \tag{17}$$

we find that

$$P^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta} = g^{\lambda(\mu}\delta^{\nu)}_{\tau}\delta^{\alpha}_{[\sigma}\delta^{\beta}_{\rho]} + g^{\lambda(\mu}\delta^{\nu)}_{[\sigma}\delta^{\beta}_{\rho]}\delta^{\alpha}_{\tau} - g^{\lambda\alpha}\delta^{(\mu}_{\tau}\delta^{\nu)}_{[\sigma}\delta^{\beta}_{\rho]}. \tag{18}$$

And then the coefficients $P^{\mu\nu\alpha\beta}$ (and Q) would be

$$P^{\mu\nu\alpha\beta} = g^{\tau\sigma}P^{\lambda}{}_{\tau\lambda\sigma}{}^{\mu\nu\alpha\beta}, \qquad (Q = g^{\tau\sigma}Q^{\lambda}{}_{\tau\lambda\sigma}). \tag{19}$$

Giving the result $P^{\mu\nu\alpha\beta} = g^{\alpha(\mu}g^{\nu)\beta} - g^{\mu\nu}g^{\alpha\beta}$ which is already symmetric in (α, β). Nevertheless, as the non-contracted symbol $P^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta}$ will be important in more complicated Lagrangians involving Riemann tensor (such as the case of Einstein-Gauss-Bonnet gravity) the correct way is to consider only symmetrized part $P^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu(\alpha\beta)}$. To understand this let's have a look at "canonical" covariantization of Riemann tensor. Using identity

$$\Gamma^{\lambda}_{\tau\sigma} - \bar{\Gamma}^{\lambda}_{\tau\sigma} = \Delta^{\lambda}_{\tau\sigma} = \frac{1}{2}g^{\lambda\iota}\left(g_{\iota\tau\bar{;}\sigma} + g_{\iota\sigma\bar{;}\tau} - g_{\tau\sigma\bar{;}\iota}\right) \tag{20}$$

we get

$$\begin{aligned} R^{\lambda}{}_{\tau\rho\sigma} &= \Delta^{\lambda}_{\tau\sigma\bar{;}\rho} - \Delta^{\lambda}_{\tau\rho\bar{;}\sigma} + \Delta^{\lambda}_{\rho\eta}\Delta^{\eta}_{\tau\sigma} - \Delta^{\lambda}_{\eta\sigma}\Delta^{\eta}_{\tau\rho} + \bar{R}^{\lambda}{}_{\tau\rho\sigma} \\ &= \frac{1}{2}g^{\lambda\iota}\left(g_{\iota\tau\bar{;}\sigma\rho} + g_{\iota\sigma\bar{;}\tau\rho} - g_{\tau\sigma\bar{;}\iota\rho} - g_{\iota\tau\bar{;}\rho\sigma} - g_{\iota\tau\bar{;}\rho\sigma} + g_{\tau\rho\bar{;}\iota\sigma}\right) + \tilde{Q} \end{aligned} \tag{21}$$

leading to

$$\tilde{P}^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta} = g^{\lambda(\mu}\delta^{\nu)}_{\tau}\delta^{\alpha}_{[\sigma}\delta^{\beta}_{\rho]} + g^{\lambda(\mu}\delta^{\nu)}_{[\sigma}\delta^{\beta}_{\rho]}\delta^{\alpha}_{\tau} - g^{\lambda\alpha}\delta^{(\mu}_{\tau}\delta^{\nu)}_{[\sigma}\delta^{\beta}_{\rho]} = P^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta} \tag{22}$$

The chosen order of covariant derivatives **effectively fixes the antisymmetric part of** $\tilde{P}^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta}$. But again this choice is rather arbitrary as we could switch the order of covariant derivative in each of six second order terms in Riemann tensor getting 2^6 different symbols $\tilde{P}^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta}$. The resolution is again simple. We should realize that only the symmetric part of $P^{\lambda}{}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta}$ is contributing to expression for Riemann tensor

${P^\lambda}_{\tau\rho\sigma}{}^{\mu\nu\alpha\beta} g_{\mu\nu,\alpha\beta} + {Q^\lambda}_{\tau\rho\sigma}$. Covariantization with respect to this leads to

$$
\begin{aligned}
{R^\lambda}_{\tau\rho\sigma} = {} & \frac{1}{2} g^{\lambda\iota}\left(g_{\iota\sigma\bar{;}(\tau\rho)} - g_{\tau\sigma\bar{;}(\iota\rho)} - g_{\iota\tau\bar{;}(\rho\sigma)} + g_{\tau\rho\bar{;}(\iota\sigma)}\right) \\
& + \frac{1}{2}\left[{g^{\lambda\iota}}_{\bar{;}\rho}\left(g_{\iota\tau\bar{;}\sigma} + g_{\iota\sigma\bar{;}\tau} - g_{\tau\sigma\bar{;}\iota}\right) - {g^{\lambda\iota}}_{\bar{;}\sigma}\left(g_{\iota\tau\bar{;}\rho} + g_{\iota\tau\bar{;}\rho} - g_{\tau\rho\bar{;}\iota}\right)\right] \\
& + \Delta^\lambda_{\rho\eta}\Delta^\eta_{\tau\sigma} - \Delta^\lambda_{\eta\sigma}\Delta^\eta_{\tau\rho} + {\bar{R}^\lambda}_{\tau\rho\sigma} \\
& + \frac{1}{2} g^{\lambda\iota}\left(2g_{\kappa(\iota}{\bar{R}^\kappa}_{\tau)\sigma\rho} + g_{\kappa(\iota}{\bar{R}^\kappa}_{\sigma)\tau\rho} - g_{\kappa(\tau}{\bar{R}^\kappa}_{\sigma)\iota\rho} - g_{\kappa(\iota}{\bar{R}^\kappa}_{\rho)\tau\sigma} + g_{\kappa(\tau}{\bar{R}^\kappa}_{\rho)\iota\sigma}\right) \qquad (23) \\
= {} & \frac{1}{2} g^{\lambda\iota}\left(g_{\iota\sigma\bar{;}(\tau\rho)} - g_{\tau\sigma\bar{;}(\iota\rho)} - g_{\iota\tau\bar{;}(\rho\sigma)} + g_{\tau\rho\bar{;}(\iota\sigma)}\right) \\
& + \frac{1}{2}\left[{g^{\lambda\iota}}_{\bar{;}\rho}\left(g_{\iota\tau\bar{;}\sigma} + g_{\iota\sigma\bar{;}\tau} - g_{\tau\sigma\bar{;}\iota}\right) - {g^{\lambda\iota}}_{\bar{;}\sigma}\left(g_{\iota\tau\bar{;}\rho} + g_{\iota\tau\bar{;}\rho} - g_{\tau\rho\bar{;}\iota}\right)\right] + \Delta^\lambda_{\rho\eta}\Delta^\eta_{\tau\sigma} - \Delta^\lambda_{\eta\sigma}\Delta^\eta_{\tau\rho} \\
& + \frac{1}{4}\left[{\bar{R}^\lambda}_{\tau\rho\sigma} + g^{\lambda\iota}\left(g_{\kappa\tau}{\bar{R}^\kappa}_{\iota\sigma\rho} + g_{\kappa\sigma}{\bar{R}^\kappa}_{\rho\tau\iota} + g_{\kappa\rho}{\bar{R}^\kappa}_{\sigma\iota\tau}\right)\right]. \qquad (24)
\end{aligned}
$$

The important thing to remember from this rather complicated expression (in comparison with the "canonical" covariantized one) is that partial derivative with respect to the $g_{\mu\nu\bar{;}\alpha\beta}$ is always symmetrical in (α, β), i.e.

$$
\frac{\partial {R^\lambda}_{\tau\rho\sigma}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = {P^\lambda}_{\tau\rho\sigma}{}^{\mu\nu(\alpha\beta)} = g^{\lambda(\mu}\delta^{\nu)}_{[\sigma}\delta^{(\alpha}_{\rho]}\delta^{\beta)}_{\tau} - \delta^{(\mu}_{\tau}\delta^{\nu)}_{[\sigma}\delta^{(\alpha}_{\rho]}g^{\beta)\lambda}. \qquad (25)
$$

3 Einstein-Gauss-Bonnet gravity

Let's now turn our attention to more complicated example with quadratic terms in second derivatives as is the case in Gauss-Bonnet gravity with Lagrangian

$$
\hat{\mathcal{L}}_{EGB} = \sqrt{-g}\left[R^2 - 4R_{\tau\sigma}R^{\tau\sigma} + R_{\lambda\tau\rho\sigma}R^{\lambda\tau\rho\sigma}.\right] \qquad (26)
$$

The structure of it is as follows

$$
\hat{\mathcal{L}} = \sqrt{-g}\left[P_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}} g_{\mu\nu,\alpha\beta}\, g_{\tilde{\mu}\tilde{\nu},\tilde{\alpha}\tilde{\beta}} + P_1^{\mu\nu\alpha\beta} g_{\mu\nu,\alpha\beta} + P_0.\right] \qquad (27)
$$

with P_0, P_1, P_2 being functions of metric g and its first derivatives.

Once again it holds that $P_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}} = P_2^{\mu\nu(\alpha\beta)\tilde{\mu}\tilde{\nu}(\tilde{\alpha}\tilde{\beta})}$ and $P_1^{\mu\nu\alpha\beta} = P_1^{\mu\nu(\alpha\beta)}$ due to commuting of partial derivatives. The covariantization leads to

$$
\begin{aligned}
\hat{\mathcal{L}}^* &= \sqrt{-g}\left[P_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}}(g_{\mu\nu\bar{;}\alpha\beta} + K_{\mu\nu\alpha\beta})(g_{\tilde{\mu}\tilde{\nu}\bar{;}\tilde{\alpha}\tilde{\beta}} + K_{\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}}) + P_1^{\mu\nu\alpha\beta}(g_{\mu\nu\bar{;}\alpha\beta} + K_{\mu\nu\alpha\beta}) + P_0\right] \\
&= \sqrt{-g}\Big[\tilde{P}_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}} g_{\mu\nu\bar{;}(\alpha\beta)}\, g_{\tilde{\mu}\tilde{\nu}\bar{;}(\tilde{\alpha}\tilde{\beta})} \\
&\qquad + \left(\tilde{P}_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}} K_{\tilde{\mu}\tilde{\nu}(\tilde{\alpha}\tilde{\beta})} + \tilde{P}_2^{\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}\mu\nu\alpha\beta} K_{\tilde{\mu}\tilde{\nu}(\tilde{\alpha}\tilde{\beta})} + \tilde{P}_1^{\mu\nu\alpha\beta}\right) g_{\mu\nu\bar{;}(\alpha\beta)} \\
&\qquad + \left(\tilde{P}_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}} K_{\mu\nu(\alpha\beta)} K_{\tilde{\mu}\tilde{\nu}(\tilde{\alpha}\tilde{\beta})} + \tilde{P}_1^{\mu\nu\alpha\beta} K_{\mu\nu(\alpha\beta)} + \tilde{P}_0\right)\Big] \qquad (28)
\end{aligned}
$$

tildas above P_i denotes functional change due to substitution as in (6) and (7). Then we have unambiguous and symmetrical expression for

$$\frac{\partial \hat{\mathcal{L}}^*}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = \frac{\partial \hat{\mathcal{L}}^*}{\partial g_{\mu\nu\bar{;}\beta\alpha}} = \sqrt{-g}\left[P_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}} g_{\tilde{\mu}\tilde{\nu};(\tilde{\alpha}\tilde{\beta})} + P_2^{\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}\mu\nu\alpha\beta} g_{\tilde{\mu}\tilde{\nu}\bar{;}(\tilde{\alpha}\tilde{\beta})}\right.$$
$$\left. + P_2^{\mu\nu\alpha\beta\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}} K_{\tilde{\mu}\tilde{\nu}(\tilde{\alpha}\tilde{\beta})} + P_2^{\tilde{\mu}\tilde{\nu}\tilde{\alpha}\tilde{\beta}\mu\nu\alpha\beta} K_{\tilde{\mu}\tilde{\nu}(\tilde{\alpha}\tilde{\beta})} + P_1^{\mu\nu\alpha\beta}\right]. \tag{29}$$

3.1 Calculation

Let's perform an explicit calculation of the derivative with respect to the $g_{\mu\nu\bar{;}\alpha\beta}$.

$$\frac{\partial \hat{\mathcal{L}}_{EGB}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = 2\sqrt{-g}\left[\frac{\partial R^{\lambda}{}_{\tau\rho\sigma}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} R_{\lambda}{}^{\tau\rho\sigma} - 4\frac{\partial R_{\tau\sigma}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} R^{\tau\sigma} + \frac{\partial R}{\partial g_{\mu\nu\bar{;}\alpha\beta}} R\right]. \tag{30}$$

We use the symmetrized result

$$\frac{\partial R^{\lambda}{}_{\tau\rho\sigma}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = g^{\lambda(\mu}\delta^{\nu)}_{[\sigma}\delta^{(\alpha}_{\rho]}\delta^{\beta)}_{\tau} - g^{\lambda(\alpha}\delta^{\beta)}_{[\rho}\delta^{(\mu}_{\sigma]}\delta^{\nu)}_{\tau} \tag{31}$$

with corresponding contractions

$$\frac{\partial R_{\tau\sigma}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = g^{\beta)(\mu}\delta^{\nu)}_{(\tau}\delta^{(\alpha}_{\sigma)} - \frac{1}{2}g^{\alpha\beta}\delta^{(\mu}_{\tau}\delta^{\nu)}_{\sigma} - \frac{1}{2}g^{\mu\nu}\delta^{(\alpha}_{\tau}\delta^{\beta)}_{\sigma}, \tag{32}$$

$$\frac{\partial R}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = g^{\alpha(\mu}g^{\nu)\beta} - g^{\mu\nu}g^{\alpha\beta}. \tag{33}$$

Finally, the result is shown below.

$$\frac{\partial \hat{\mathcal{L}}_{EGB}}{\partial g_{\mu\nu\bar{;}\alpha\beta}} = 2\sqrt{-g}\Bigg[R_{\lambda}{}^{\tau\rho\sigma}\left(g^{\lambda(\mu}\delta^{\nu)}_{[\sigma}\delta^{(\alpha}_{\rho]}\delta^{\beta)}_{\tau} - g^{\lambda(\alpha}\delta^{\beta)}_{[\rho}\delta^{(\mu}_{\sigma]}\delta^{\nu)}_{\tau}\right)$$
$$-4R^{\tau\sigma}\left(g^{\beta)(\mu}\delta^{\nu)}_{(\tau}\delta^{(\alpha}_{\sigma)} - \frac{1}{2}g^{\alpha\beta}\delta^{(\mu}_{\tau}\delta^{\nu)}_{\sigma} - \frac{1}{2}g^{\mu\nu}\delta^{(\alpha}_{\tau}\delta^{\beta)}_{\sigma}\right)$$
$$+R\left(g^{\alpha(\mu}g^{\nu)\beta} - g^{\mu\nu}g^{\alpha\beta}\right)\Bigg] \tag{34}$$

$$= 2\sqrt{-g}\left[2R^{\alpha(\mu\nu)\beta} - 4g^{\alpha)(\mu}R^{\nu)(\beta} + 2g^{\alpha\beta}R^{\mu\nu} + 2g^{\mu\nu}R^{\alpha\beta} + \left(g^{\alpha(\mu}g^{\nu)\beta} - g^{\mu\nu}g^{\alpha\beta}\right)R\right]. \tag{35}$$

It is symmetrical both in (μ, ν) and (α, β) and differs from results in [1], [2] and [3].

4 General case

Consider the theory with arbitrary set of tensor densities as dynamical fields with second order scalar density Lagrangian $\hat{\mathcal{L}}(Q_B; Q_{B,\alpha}; Q_{B,\alpha\beta})$. In any case the indices of $Q_{B,\alpha\beta}$ has to be contracted to form scalar density. And this contraction reflects the symmetry of partial derivatives as we demonstrated in preceding sections. The covariantization leads to Lagrangian $\hat{\mathcal{L}}^*(Q_B; Q_{B\bar{;}\alpha}; Q_{B\bar{;}\alpha\beta}; \bar{g}; \bar{R})$ with antisymmetrical part of second covariant derivatives $Q_{B\bar{;}[\alpha\beta]}$ vanishing. Hence the changing of order of covariant derivatives via substituting $Q_{B\bar{;}\alpha\beta} = Q_{B\bar{;}\beta\alpha} + Q_B|^{\rho}_{\sigma}\,\bar{R}^{\sigma}{}_{\rho\alpha\beta}$[2] does not change Lagrangian $\hat{\mathcal{L}}^*$ at all on the contrary to what is suggested in [1].

If one insists on keeping arbitrary antisymmetric parts of tensor contracted with second covariant derivatives we get nonvanishing derivative

$$\frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\beta]}} = \frac{1}{2}\left(\frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}\alpha\beta}} - \frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}\beta\alpha}}\right), \tag{36}$$

then we truly get a Lagrangian $\hat{\mathcal{L}}^{**}$ by switching the order of covariant derivatives.

In general a covariant conserved current has the following form $\hat{\imath}^{\alpha} = \hat{u}^{\alpha}_{\sigma}\xi^{\sigma} + \hat{m}^{\alpha\tau}_{\sigma}\xi^{\sigma}_{\bar{;}\tau} + \hat{n}^{\alpha\tau\beta}_{\sigma}\xi^{\sigma}_{\bar{;}\beta\tau}$ which satisfies the identity $\hat{\imath}^{\alpha}_{\bar{;}\alpha} = 0$ for every vector field ξ. The formulas for coefficients $\hat{u}$, $\hat{m}$ and $\hat{n}$ are given in [1]. For each Lagrangian we get a set of current coefficients and their difference is described by the following formulas

$$\Delta\hat{n}^{\alpha\tau\beta}_{\sigma} = \frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\beta]}}\, Q_B|^{\tau}_{\sigma} + \frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\tau]}}\, Q_B|^{\beta}_{\sigma}, \tag{37}$$

$$\Delta\hat{m}^{\alpha\tau}_{\sigma} = -2\frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\tau]}} Q_{B\bar{;}\sigma} + 2\left(\frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\beta]}}\, Q_B|^{\tau}_{\sigma}\right)_{\bar{;}\beta}, \tag{38}$$

$$\Delta\hat{u}^{\alpha}_{\sigma} = -2\left(\frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\beta]}} Q_{B\bar{;}\sigma}\right)_{\bar{;}\beta} + \frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\tau]}}\, Q_B|^{\beta}_{\lambda}\,\bar{R}^{\lambda}{}_{\sigma\tau\beta}. \tag{39}$$

These differences **satisfy covariant Klein identities** (see appendix A) for arbitrary $\frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\beta]}}$, i.e. the current $\hat{\imath}^{\alpha}$ formed by coefficients (37), (38) and (39) is conserved. The antisymmetric part $\frac{\partial \hat{\mathcal{L}}^*}{\partial Q_{B\bar{;}[\alpha\beta]}}$ is not defined by the original Lagrangian $\hat{\mathcal{L}}$ in any way. In particular, the result cannot be generated by adding a divergence to the original Lagrangian as divergence $\hat{d}^{\alpha}_{,\alpha}$ produces current coefficients in the form

$$\hat{u}^{\alpha}_{\sigma} = 2(\delta^{[\alpha}_{\sigma}\hat{d}^{\beta]})_{\bar{;}\beta}, \qquad \hat{m}^{\alpha\beta}_{\sigma} = 2\delta^{[\alpha}_{\sigma}\hat{d}^{\beta]}, \qquad \hat{n}^{\alpha\beta\gamma}_{\sigma} = 0. \tag{40}$$

5 Conclusion

We demonstrated that the new class of ambiguity in Noether currents appers due to covariantization procedure. It doesn't have the character of adding divergence to the Lagrange function as it doesn't change the original Lagrangian at all and also the resulting current has different structure. The canonical covariantization should take into account the symmetricity of partial derivatives present in the original Lagrangian and this is in conflict with Petrov's approach in [1] suggesting two, rather artificially chosen, classes of currents. This discrepancy was demonstrated in the sketch of current coefficients computation in the case of Gauss-Bonnet gravity.

[2] Quantity $Q_B|^{\rho}_{\sigma}$ is defined via geometrical properties of Q_B (e.g. rank and weight of the tensor field density) and by formula for Lie derivative $\mathcal{L}_{\xi}Q_B = \xi^{\rho}Q_{B,\rho} - Q_B|^{\rho}_{\sigma}\,\xi^{\sigma}_{,\rho}$.

A Covariant Klein identities

We want to make use of the arbitrariness of ξ in conserved current $\hat{\imath}^\alpha$. Let's distribute the derivative in $\hat{\imath}^\alpha_{\bar{;}\alpha}$ and obtain

$$\hat{u}^\alpha_{\sigma\bar{;}\alpha}\xi^\sigma + \left(\hat{u}^\alpha_\sigma \xi^\sigma_{\bar{;}\alpha} + \hat{m}^{\alpha\tau}_{\sigma\bar{;}\alpha}\xi^\sigma_{\bar{;}\tau}\right) + \left(\hat{m}^{\alpha\tau}_\sigma \xi^\sigma_{\bar{;}\tau\alpha} + \hat{n}^{\alpha\tau\beta}_{\sigma\bar{;}\alpha}\xi^\sigma_{\bar{;}\beta\tau}\right) + \hat{n}^{\alpha\tau\beta}_\sigma \xi^\sigma_{\bar{;}\beta\tau\alpha} = 0. \tag{41}$$

Coefficients at corresponding derivatives of vector field should vanish. But not all higher derivatives are linearly independent – it is necessary to choose basis, e.g. $\xi^\sigma_{\bar{;}(\alpha\beta)}$, $\xi^\sigma_{\bar{;}(\alpha\beta\gamma)}$ – i.e. symmetrization of covariant derivatives. Antisymmetric parts are converted into lower order via Riemann tensor. To achieve this we use Young projection operators to decompose tensors $\hat{m}^{\alpha\tau}_\sigma$ and $\hat{n}^{\alpha(\tau\beta)}_\sigma$ (see [4]) into

$$\hat{m}^{\alpha\tau}_\sigma = \hat{m}^{(\alpha\tau)}_\sigma + \hat{m}^{[\alpha\tau]}_\sigma, \qquad \hat{n}^{\alpha(\tau\beta)}_\sigma = \hat{n}^{(\alpha\tau\beta)}_\sigma + \frac{4}{3}\hat{n}^{[\beta\alpha]\tau}_\sigma + \frac{2}{3}\hat{n}^{[\tau\beta]\alpha}_\sigma \tag{42}$$

and using

$$2\xi^\sigma_{\bar{;}[\tau\alpha]} = \bar{R}^\sigma{}_{\rho\tau\alpha}\,\xi^\rho, \tag{43}$$

$$2\xi^\sigma_{\bar{;}\tau[\beta\alpha]} = \bar{R}^\sigma{}_{\rho\beta\alpha}\,\xi^\rho_{\bar{;}\tau} - \bar{R}^\rho{}_{\tau\beta\alpha}\,\xi^\sigma_{\bar{;}\rho}, \tag{44}$$

$$2\xi^\sigma_{\bar{;}[\tau\beta]\alpha} = \bar{R}^\sigma{}_{\rho\tau\beta\bar{;}\alpha}\,\xi^\rho + \bar{R}^\sigma{}_{\rho\tau\beta}\,\xi^\rho_{\bar{;}\alpha} \tag{45}$$

we finally get decompositions using only covariant derivatives of chosen basis

$$\hat{m}^{\alpha\tau}_\sigma \xi^\sigma_{\bar{;}\tau\alpha} = (\hat{m}^{(\alpha\tau)}_\sigma + \hat{m}^{[\alpha\tau]}_\sigma)\xi^\sigma_{\bar{;}\tau\alpha} = \hat{m}^{(\alpha\tau)}_\sigma \xi^\sigma_{\bar{;}(\alpha\tau)} + \frac{1}{2}\hat{m}^{\alpha\tau}_\sigma \bar{R}^\sigma{}_{\rho\alpha\tau}\xi^\rho, \tag{46}$$

$$\hat{n}^{\alpha(\tau\beta)}_\sigma \xi^\sigma_{\bar{;}\tau\beta\alpha} = \hat{n}^{\alpha\tau\beta}_\sigma \xi^\sigma_{\bar{;}(\alpha\tau\beta)} + \hat{n}^{\alpha(\tau\beta)}_\sigma \bar{R}^\sigma{}_{\rho\alpha\tau}\xi^\rho_{\bar{;}\beta} + \frac{2}{3}\hat{n}^{\alpha(\tau\beta)}_\sigma \bar{R}^\rho{}_{\tau\beta\alpha}\xi^\sigma_{\bar{;}\rho} + \frac{1}{3}\hat{n}^{\alpha(\tau\beta)}_\sigma \bar{R}^\sigma{}_{\rho\alpha\tau\bar{;}\beta}\xi^\rho. \tag{47}$$

Rearranging the identity (41) using above written decompositions leads to covariant Klein identities

$$0 = \hat{u}^\alpha_{\sigma\bar{;}\alpha} + \frac{1}{2}\hat{m}^{\alpha\rho}_\lambda \bar{R}^\lambda{}_{\sigma\alpha\rho} + \frac{1}{3}\hat{n}^{\alpha\rho\gamma}_\lambda \bar{R}^\lambda{}_{\sigma\alpha\rho\bar{;}\gamma}, \tag{48}$$

$$0 = \hat{u}^\alpha_\sigma + \hat{m}^{\lambda\alpha}_{\sigma\bar{;}\lambda} + \hat{n}^{\tau\alpha\rho}_\lambda \bar{R}^\lambda{}_{\sigma\tau\rho} + \frac{2}{3}\hat{n}^{\lambda\tau\rho}_\sigma \bar{R}^\alpha{}_{\tau\rho\lambda}, \tag{49}$$

$$0 = \hat{m}^{(\alpha\beta)}_\sigma + \hat{n}^{\lambda(\alpha\beta)}_{\sigma\bar{;}\lambda}, \tag{50}$$

$$0 = \hat{n}^{(\alpha\beta\gamma)}_\sigma. \tag{51}$$

References

[1] A. N. Petrov, R. R. Lompay. *Covariantized Noether identities and conservation laws for perturbations in metric theories of gravity.* Gen Relativ Gravit (2013) 45:545-579

[2] A. N. Petrov. *Three types of superpotentials for perturbations in the Einstein-Gauss-Bonnet gravity.* Class. Quantum Grav. **26** (2009) 135010 (16pp)

[3] A. N. Petrov. *Noether and Belinfante corrected types of currents for perturbations in the Einstein-Gauss-Bonnet gravity.* Class. Quantum Grav. **28** (2011) 215021 (17pp)

[4] R. R. Lompay, A. N. Petrov. *Covariant Differential Identities and Conservation Laws in Metric-Torsion Theories of Gravitation. I. General Consideration.* arXiv:1306.6887v2 [gr-qc]

Klasifikace vzorů ve zdrojových kódech[*]

Josef Smolka

4. ročník PGS, email: smolkjos@fjfi.cvut.cz
Katedra softwarového inženýrství
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Miroslav Virius, Katedra softwarového inženýrství
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. The paper presents results of application of five different classifiers to a problem of pattern classification in Java source codes. Source code perception by tools for automated analysis and transformation can be enhanced by finding a effective method of patterns classification.

Keywords: source code, classification, Java

Abstrakt. Příspěvek prezentuje výsledky aplikace pěti různých klasifikátorů na problém klasifikace vzorů ve zdrojovém kódu programů napsaných v jazyce Java. Nalezení efektivní metody klasifikace je základ pro zlepšení vnímání kódu automatizovanými nástroji pro jeho analýzu a transformaci.

Klíčová slova: zdrojový kód, klasifikace, Java

1 Úvod

Identifikace a klasifikace definovaných vzorů ve zdrojových kódech softwaru může napomoci ke zlepšení vnímání zdrojového kódu ze strany podpůrných nástrojů vývojáře, jako jsou např. refaktorovací a analytické nástroje, inteligentní editory, modelovací nástroje a obecně verifikační nástroje, které se snaží ověřit implementaci vůči specifikaci. Tato práce prezentuje experimentální výsledky aplikace několika běžně používaných klasifikačních metod: k-NN, neuronových sítí, logistické regrese a SVM.

1.1 Návaznost práce

Myšlenka rozpoznávání vzorů ve zdrojovém kódu nabyla na významu s rozmachem objektově orientovaného programování, které vytváření takových vzorů značně usnadňuje. V roce 1998 Antoniol a kolektiv [1, 2] vypracovali metodu pro detekci malé množiny návrhových vzorů [7] ve zdrojových kódech programů napsaných v C++. Ze zdrojových kódu extrahoval hodnoty definovaných metrik, pomocí kterých prováděl samotnou detekci vzorů. Tato práce rovněž zakládá klasifikaci vzorů na hodnotách příznaků získaných ze zdrojového kódu, na rozdíl od [1, 2] však pro klasifikaci využívá jiné metody. Podobně jako [6] a [9] využívá tato práce metody strojového učení. Tato práce přejímá myšlenku uvedenou v [8] a zabývá se analýzou na úrovni tříd.

[*] Tato práce byla podpořena grantem SGS11/167/OHK4/3T/14 a LA08015.

1.2 Východiska práce

Práce se omezuje na zkoumání struktur v programovacím jazyce Java. Důvodem je rozšíření platformy Java, díky čemuž je k dispozici velké množství rozmanitých zdrojových kódů ke zkoumání. Dále jsou volně k dispozici nástroje pro práci se samotným jazykem, které jsou nutné pro analýzu zdrojového kódu a sestavení příznakového vektoru pro klasifikátory. Zde uvedené postupy jsou však aplikovatelné i na jiné objektově orientované silně typované jazyky. Naopak použité metody nelze aplikovat na dynamické jazyky, protože některé použité příznaky se spoléhají na identifikaci datových typů proměnných.

2 Metody klasifikace

Vzhledem k tomu, že hledané struktury, ať už se jedná o primitivní struktury jako bean/datový typ, nebo složitější kompozitní struktury, mohou být těžko popsatelné a jejich konkrétní implementace se liší v závislosti na programátorovi, zaměřuje se tato práce především na metody strojového učení s učitelem, kdy jsou jednotlivé klasifikátory předem trénovány na vybrané množině příkladů. Objektový návrh softwaru, který je stále populární a používaný, dal vzniknout celé řadě znovupoužitelných vzorů. Jedná se především o návrhové vzory [7], které se staly v oboru softwarového inženýrství hlavním zdrojem inspirace při návrhu softwaru. Nicméně ve zdrojovém kódu lze vysledovat i jiné vzory, které byť nejsou tak formalizované, mohou stále pomoct s orientací ve zdrojovém kódu. Návrhové vzory nejsou definovány exaktně, jedná se pouze o jistou šablonu, která popisuje problém a obecný návrh řešení. Konkrétní implementace návrhových vzorů se tedy mohou značně lišit.

2.1 Klasifikační třídy a příznakový prostor

Pro začátek bylo navrženo 11 klasifikačních tříd, které pokrývají některé jednodušší návrhové vzory, UML stereotypy a jiné běžně se vyskytující vzory [10]:

- **Utility** – Pomocná třída obsahující zpravidla statické metody. V UML se taková třída označuje stereotypem Auxiliary [12].

- **Factory** – Návrhový vzor patřící do skupiny vzorů zabývajících se tvorbou objektů. Tento vzor se dotýká problémů, kdy je k vytvoření nového objektu potřeba například jiný zdroj, který však nemá být obsažen ve výsledném objektu. Vytvoření objektu je tedy delegováno na specializovanou třídu – továrnu.

- **Builder** – Návrhový vzor, který patři do stejné skupiny jako Factory. Builder řeší problém inicializace stavu zpravidla kompozitních objektů se složitým vnitřním stavem. Vzor umožňuje postupnou konfiguraci stavu objektu po vytvoření.

- **Adapter**, **Proxy**, **Decorator** – Skupina návrhových vzorů, které obecně řeší zástupnost objektů s různými rozhraními.

- **Bean** – Třída, která pouze zapouzdruje data. Jedná se o návrhový vzor Crate a UML stereotyp Type [12].

- **DAO** – Představuje persistentní data. Jedná se o UML stereotyp Entity [12].
- **Worker** – Třída implementující hlavní logiku pracující s daty. Představuje UML stereotyp Focus [12].
- **Composite** – Třída představující hierarchická data jako jsou seznam, strom, a jiné. Jedná se o návrhový vzor Composite.
- **Constant** – Třída obsahující pouze konstanty. Zpravidla slouží jako nějaká statická konfigurace systému. V UML se taková třída označuje stereotypem Auxiliary [12].

Příznakový prostor je definován jako množina $S = \{F_1, F_2, F_3, \dots, F_p\}$, kde p je počet definovaných příznaků a $F_i : \mathbf{C} \to \mathbb{R}$ funkce, která transformuje deklaraci datového typu $c \in \mathbf{C}$ na reálné číslo. Hodnota F_i typicky vyjadřuje četnost sledovaného fenoménu v deklaraci datového typu a měla by být invariantní k absolutní velikosti deklarace (ať už měřeného pomocí počtu řádků kódu, počtu příkazů, nebo jinou metodou) [10].

2.2 Použité klasifikátory

V rámci této práce byly využity výhradně metody klasifikace s učením s učitelem.

2.2.1 Algoritmus k nejbližších sousedů (k-NN)

V oboru klasifikace je k-NN jedním z nejjednodušších, přesto hojně používaných a účinných neparametrických algoritmů. Bod v příznakovém prostoru je klasifikován podle k nejbližších sousedů v trénovací množině. Algoritmus vychází z [4], jedná se vlastně o speciální případ pro k = 1. Slabinou algoritmu je případ, kdy rozložení pravděpodobnosti výskytu třídy v trénovací množině není rovnoměrné. Pak zástupci nejvíce se vyskytující třídy ovlivňují výsledek klasifikace. Algoritmus lze jednoduše modifikovat volbou k a volbou metriky vzdálenosti.

2.2.2 Logistická regrese

Logistická regrese je pravděpodobnostní statistický klasifikátor. Pravděpodobnost, že prvek patří do určité třídy je modelována na základě vysvětlujících proměnných (vektor příznaků) pomocí logistické funkce (logitové transformace). Tato metoda se začala používat v 60. letech jako alternativa k lineární regresi. Klasická logistická regrese řeší binární problém, kdy může být prvek zařazen pouze do dvou tříd.

2.2.3 Perceptron

Perceptron je prvním a jedním z nejjednodušších modelů umělé neuronové sítě, který představil F. Rosenblatt již v roce 1957 [11]. Jedná se ve své podstatě o váženou síť. Tento výpočetní model ve své době podpořil zájem o metody klasifikace s účením, nicméně tento zájem opětovně opadl v době, kdy M. Minsky ve své knize Perceptrons ukázal, že jednoduchý perceptron není schopen simulovat boolovskou funkci XOR. Pozdější výzkumy však ukázaly, že vícevrstvý perceptron je již tohoto schopen, což vedlo opět k oživení zájmu o umělé neuronové sítě.

2.2.4 Vícevrstvý perceptron

Vícevrstvý perceptron odstraňuje podmínku lineární separovatelnosti klasického perceptronu vložením jedné nebo více skrytých disjunkních vrstev s nelineární aktivační funkcí mezi vrstvu vstupní a výstupní. Výstup neuronů jedné vrstvy slouží jako vstup neuronů vrstvy následující, přičemž každý neuron z jedné vrstvy je zpravidla propojen se všemi neurony následující vrstvy.

2.2.5 Support Vector Machine

Metoda podpůrných vektorů (SVM) řeší problém nelineární separovatelnosti dat přenesením problému do prostoru vyšší dimenze. Základním principem metody je nalezení takové nadroviny, která rozděluje prostor problému na dva podprostory, kde každý podprostor obsahuje převážně zástupce jedné třídy a přitom maximalizuje vzdálenost nejbližších zástupců těchto tříd (podpůrné vektory) od hranice tvořené touto nadrovinou [3].

2.3 Redukce příznakového prostoru

S ohledem na velikost trénovací množiny (cca 170 položek) a velikost příznakového prostoru (40 příznaků) byly v rámci ladění klasifikátoru aplikovány heuristiky (zpětná hladová eliminace, rychlé simulované žíhání) pro nalezení optimálního submodelu.

3 Výsledky

Každý aplikovaný klasifikátor byl podroben analýze, v rámci které byla laděna konfigurace parametrů a architektura klasifikátoru. Tabulka 1 zobrazuje nejúspěšnější konfigurace pro každý typ klasifikátoru.

Klasifikátor	**XV full**	**XV sub-model**	**# příznaků**
k-NN	0,2674	0,2036	25
Logistická regrese	0,2963	0,2408	27
Perceptron	0,2411	0,2205	22
Složený perceptron	0,2572	0,2050	27
SVM	0,2352	0,2018	25

Tabulka 1: Průměrná chyba klasifikace v křížové validaci na plném modelu (XV full) a nejlepším nalezeném submodelu (XV sub-model).

U klasifikátoru k-NN byly zkoušeny různé metriky pro měření vzdálenosti mezi body v příznakovém prostoru: euklidovská, čebyševova, kosínová a síťová. Nejlepších výsledků bylo dosaženo za použití síťové metriky $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$ pro $k = 5$. U logistické regrese byly pomocí heuristiky laděny parametry α a λ. Nejlepších výsledků bylo dosaženo pro hodnoty $\alpha^* = 0,999347$ a $\lambda^* = 5,988688 \cdot 10^{-6}$. U perceptronu učeného pomocí zpětné propagace byly zkoušeny dvě hlavní architektury: perceptron s výstupem pro každou třídu a klasifikátor složeny s perceptronů, kde každý perceptron má jeden výstup. Dále byly zkoušeny různé aktivační funkce a míra učení, aby nedošlo k přeučení klasifikátoru. Jako nejlepší se ukázal klasifikátor složený z jednotlivých perceptronů pro každou třídu s aktivační funkcí ve tvaru $\sigma(\xi) = a \tanh(v\xi)$. U složeného perceptronu byly podobně zkoušeny architektury, kdy je klasifikátor tvořen jedním složeným perceptronem s výstupem pro každou třídu a klasifikátor tvořený složenými perceptrony pro každou třídu. Dále byl zkoumán počet neuronů ve skryté vrstvě, vhodná aktivační funkce a míra natrénování sítě. Nejmenší průměrnou chybu klasifikace měl opět klasifikátor složen z jednotlivých vícevrstvých perceptronů se 4 skrytými neurony a stejnou aktivační funkcí jako v předešlém případě. U SVM klasifikátoru byly zkoumány různé kernely a pomocí heuristiky hledány vhodné hodnoty parametrů ϵ, υ a γ. Celkově nejúspěšnější se ukázal SVM klasifikátor s RBF kernelem $k(x_i, x_j) = (-\gamma||x_i \cdot x_j||^2)$ a hodnotami parametrů $\epsilon = 0,00079$, $\upsilon = 0,32144$ a $\gamma = 0,15617$.

Jednotlivé klasifikátory byly testovány pomocí křížové validace, kdy byla testovací množina rozdělena na 5 rovnoměrných dílů, postupně byly vždy 4/5 použity pro trénování a zbývající 1/5 pro ověření klasifikátoru [5]. Takto byla otestována celá množina. Celý proces byl zopakován 100 krát a chyba zprůměrována.

4 Závěr a další práce

Pomocí postupného ladění klasifikátorů bylo dosaženo přesnosti klasifikace vybraných vzorů téměř 80%. To se dá pokládat za dobrý počáteční výsledek do dalšího zkoumání. V rámci další práce bude rozšířena množina klasifikovaných vzorů, bude dále revidován příznakový prostor, rozšířena množina zkoumaných dat a vyzkoušeny další klasifikační metody.

Literatura

[1] Antoniol, G.; Fiutem, R.; Cristoforetti, L., *Using metrics to identify design patterns in object-oriented software*, Proceedings of Fifth International Software Metrics Symposium, pp.23,34, 1998, doi: 10.1109/METRIC.1998.731224

[2] Antoniol, G.; Fiutem, R.; Cristoforetti, L., *Design pattern recovery in object-oriented software*, Proceedings of IWPC '98, pp.153,160, 1998, doi: 10.1109/WPC.1998.693342

[3] Cortes, C.; Vapnik, V., *Support-vector networks*, Machine Learning vol.20, Springer, pp.273,297, 1995, doi:10.1007/BF00994018.

[4] Cover, T.; Hart, P., *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, vol.13, no.1, pp.21,27, 1967, doi: 10.1109/TIT.1967.1053964

[5] Duda, R. O.; Hart, P. E.; Stork, D. G., *Pattern Classification*, John Wiley & Sons, New York, 2001.

[6] Ferenc, R.; Beszédes, Á.; Fülöp, L.; Lele, J., *Design Pattern Mining Enhanced by Machine Learning*, Proc. International Conference on Software Maintenance (ICSM 05), IEEE Computer Society, pp. 295-304, 2005, doi: 10.1109/ICSM.2005.40.

[7] Gamma, E.; Helm, R.; Johnson, R.; Vlissides, J., *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley Professional, 1994, ISBN: 978-0201633610

[8] Gueheneuc, Y., G.; Sahraoui, H.; Zaidi, F., *Fingerprinting Design Patterns*, Proc. 11th Working Conference on Reverse Engineering (WCRE 04), IEEE Computer Society, pp. 172-181, 2004, doi: 10.1109/WCRE.2004.21.

[9] Maiga, A.; Ali, N.; Bhattacharya, N.; Sabane, A; Gueheneuc, Y.; Antoniol, G.; Aimeur, E., *Support vector machines for anti-pattern detection*, Automated Software Engineering (ASE), 2012 Proceedings of the 27th IEEE/ACM International Conference, pp.278,281, 2012 doi: 10.1145/2351676.2351723

[10] Mojzeš, M.; Rost, M.; Smolka, J.; Virius, M., *Feature Space for Statistical Classification of Java Source Code Patterns*, Proc. 15th International Carpathian Control Conference (ICCC). IEEE Computer Society, pp.357,361, 2014, doi:10.1109/CarpathianCC.2014.6843627.

[11] Rosenblatt, F., *The Perceptron: A Perceiving and Recognizing Automaton*, Aeronautical Lab., Cornell Univ, 1957.

[12] Object Management Group, *OMG Unified Modeling Language (OMG UML) – Superstructure*, verze 2.4.1, 2011

Structure of Nafion*

Lucie Strmisková

5th year of PGS, email: lucka.strmiskova@seznam.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
František Maršík, Institute of Thermomechanics, AS CR
Petr Sedlák, Institute of Thermomechanics, AS CR

Abstract. Nafion is a solid polymer that is used as a proton conducting membrane in hydrogen fuel cells. Proton conducting as well as mechanical properties of Nafion strongly depend on its microstructure. Despite the extensive research, there is no model of Nafion microstructure that would be generally accepted. This contribution describes the mesoscale model of Nafion microstructure, that we are developing, and compares it with the experimental results.

Keywords: Nafion structure, modelling, MesoDyn

Abstrakt. Nafion je polymer používaný jako elektrolyt ve vodíkových palivových článcích. Vodivostní i mechanické vlastnosti Nafionu významně závisí na jeho mikrostruktuře. Navzdory dlouholetému výzkumu stále neexistuje model mikrostruktury, který by byl bez výhrad přijímán. Tento příspěvek popisuje meziškálový model struktury Nafionu, který vyvíjíme, a srovnává ho s experimentálními výsledky.

Klíčová slova: struktura Nafionu, modelování, MesoDyn

1 Introduction

Nafion is the most common material used as a proton conducting membrane in hydrogen fuel cells nowadays. Nafion consists of a polytetrafluoroethylene backbone with the randomly attached perfluorinated side chains ending by a sulfonate ionic group (figure 1).

Despite its wide usage and decades of intensive research, there are still heated discussions about Nafion morphology. There have been presented several models of Nafion microstructure, but none of them is fully accepted in the fuel cell community.

The common feature in these models is the existence of the clustering of hydrophilic domains inside hydrophobic polytetrafluoroethylene, but there is still heated debate over the shape and structure of the ionic clusters. The complicating facts are the randomness of the attachment of side chains to the polymer backbone and the sensitivity of the Nafion microstructure to the processing methods and its history.

This contribution describes the mesoscale model of Nafion microstructure, that we are currently working on, and compares it with the experimental data.

*This work has been supported by the grants CZ.1.07/2.3.00/20.0107 and SGS13/217/OHK4/3T/14.

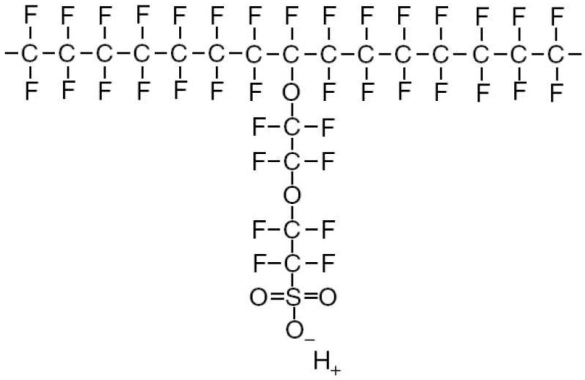


Figure 1: Structure of Nafion.

2 Atomistic modelling

Modelling and simulations are very popular nowadays and they have helped with understanding almost all important aspects of membrane structure, including morphology development, proton transport and the role of side chains in these areas.

Ab initio methods based on quantum mechanics provide the most accurate picture of the local structure and have already given us much information about the proton transport mechanisms, the dissociation of sulphuric acid and the aggregation of the side chains with this acid through the formation of hydrogen bonds. But the accuracy is paid by the size of the region, we are able to model. Ab initio methods compute with all valence electrons in the system to gain the electronic structure, systems with a maximum of one hundred atoms are generally modelled by these methods, so it can give us information only about small sections of Nafion polymer.

Molecular dynamics is not so computationally demanding as ab initio methods as it describes the motions of atoms without taking care about their inner structure. The motion of particles is described by Newton's second law in molecular dynamics. The potential of the molecular system is not a function of electronic wave functions like in ab initio models, but it is a function of the positions of nuclei $U(\vec{R}_j)$. These functions $U(\vec{R}_j)$ are evaluated by methods of quantum mechanics or empirically. Atoms and molecules are considered as classical particles moving in this potential field.

$$m_i \frac{d^2\vec{r}_i}{dt^2} = -\nabla U \tag{1}$$

The good choice of potential U (often called as a force field) is a crucial point in molecular dynamics and it is determined by the bond types, desired accuracy and of course our computational resources. Also comparison with measurements on thermo-physical properties and vibration frequencies is necessary for choosing the most suitable force field.

The result of the molecular dynamics computation are trajectories and velocities of all particles in the system. This also requires a lot of memory capacity and sufficient computer processor speed, so it generally provides motion of system of thousands of atoms on a time scale of a few nanoseconds.

The molecular dynamics supports the idea of irregularly shaped ionic clusters [5], although to model the space distribution of these clusters (2-5 nm clusters distant from each other 12-15 nm) is out of the range of molecular dynamics. The mesoscale model that would form a bridge between the fast molecular kinetics and slow thermodynamics relaxation of macroscale properties is thus necessary.

Mesoscale models gain increasing attention nowadays. One of the popular mesoscale models is MesoDyn which is a simulation code that was generated especially for describing the mesoscale structures in polymer liquids [1].

3 MesoDyn code

MesoDyn is a mesoscale simulation program implemented in Accelrys Materials Studio. It is based on a dynamic variant of self-consistent mean field theory.

The system under study is transformed to its coarse-grained structure, where several atoms are taken as one unit. This unit is called as a bead. The original structure of the bead is forgotten, but MesoDyn finds interactions for these beads corresponding to essential physics of the original system.

Because the beads consist of several atoms, the time and space lengths of simulation can be expand up to 100 nm.

The theory behind the MesoDyn code [1] will be shortly mentioned here, the details of the numerical procedure can be found in the original paper from 1997 [2].

The fluid is described by the density distributions of the individual beads $\rho_I(\vec{r})$. The density distributions dynamically evolve due to the gradient of chemical potential and random thermal noise according to the Langevin equation

$$\frac{\partial \rho_I}{\partial t} = M \mathrm{div}(\rho_I \nabla \mu_I) + \eta_I, \tag{2}$$

where M is a bead mobility parameter, μ_I is the chemical potential of the respective bead (the derivative of the free energy with respect to the density ρ_I), a noise η_I brings the kinetics of Brownian motion to the equation and satisfies the fluctuation-dissipation theorem in the following form

$$\langle \eta_I(\vec{r}, t) \rangle = 0, \tag{3}$$

$$\langle \eta_I(\vec{r}, t) \eta_J(\tilde{\vec{r}}, \tilde{t}) \rangle = -\frac{2Mv}{\beta} \delta(t - \tilde{t}) \times \nabla_r . \delta(\vec{r} - \tilde{\vec{r}}) \rho_I \rho_J \nabla_{\tilde{r}}, \tag{4}$$

where δ is the delta function, v is the average bead volume and β is the inverse temperature $\frac{1}{k_B T}$.

The bead mobility parameter is in a simple relation with bead diffusion coefficient $D = M k_B T$ and its value is same for all species in the system. Tests showed, that changes in this parameter have just small effect on the final structure, however they have an influence on the rate, at which equilibrium is achieved [6]. The default value of the bead mobility parameter is $10^{-7} cm^2 s^{-1}$ in MesoDyn code allowing thus to use time steps of 50 ns for most of the fluids.

The equations (2) can be transformed to the so called dynamic Langevin equations, that are integrated on a cubic lattice using a Crank-Nicholson numerical scheme in MesoDyn code.

The Langevin equations are constructed for incompressible system with a dynamic constraint

$$\frac{1}{v} = \sum_I \rho_I(\vec{r}, t), \tag{5}$$

where v is a constant molar volume of a substance.

So at each step of the computation, the density distribution is calculated, starting from an initially homogeneous mixture in a cube box with periodic boundary conditions. Density distribution $\rho_I(\vec{r})$ evolves via the Langevin equations and forms a slowly changing external potential $U_I(\vec{r})$. The relation between the density and external potential is through the derivative of the partition function Z

$$\rho_I(\vec{r}) = -n_I kT \frac{\partial Z}{\partial U_I(\vec{r})},$$

where n_I is the number of chains and k is the Boltzmann constant.

Such system generates Helmholtz free energy

$$F = -kT \sum_i \ln \frac{Z_i^{n_i}}{n_i!} - \sum_I \int_V U_I(\vec{r})\rho_I(\vec{r})d\vec{r} + \tag{6}$$

$$\frac{1}{2}\sum_{I,J} \int_V \int_V \varepsilon_{IJ}(|\vec{r} - \vec{r}'|)\rho_I(\vec{r})\rho_J(\vec{r}')d\vec{r}d\vec{r}' + \tag{7}$$

$$\frac{k_H}{2} \int_V (\sum_I v_I(\rho_I(\vec{r}) - \rho_I{}^0))^2 d\vec{r}, \tag{8}$$

where the first two terms represent the ideal free energy. The third term represents the interaction between the chains and it has the following form in the MesoDyn code

$$\varepsilon_{IJ}(|r - \tilde{r}|) = \varepsilon_{IJ}^0 \left(\frac{3}{2\pi a^2}\right)^{\frac{3}{2}} \exp[-\frac{3}{2a^2}(r - \tilde{r})^2], \tag{9}$$

where ε_{IJ}^0 is related to the Flory-Huggins mixing parameters via $\chi_{IJ} = \frac{\beta}{v}\varepsilon_{IJ}^0$.

The parameters in the last term (8) are Helfand compressibility parameter k_H, the average density of each bead $\rho_I{}^0$ and bead volume v_I. This term gives a restriction on the size of the density fluctuation that can occur in the system.

Electrostatics may be included in the Flory-Huggins parameter, or may be explicitly included for each bead in the system using the Donnan approximation.

The models using the Donnan approximation showed that the electrostatic effects do not have a high influence on the membrane morphology, so using simple Flory-Huggins theory is sufficient [4].

There are two steps in generating mesoscale model. First a coarse-grained model of the original system has to be determined. The second step is the calculation of the interaction energies ε_{IJ}.

One has to be really careful, while creating the coarse-grained topology of the system, and represent all chemically distinct units of the system by different beads. Otherwise the original chemistry of material would be lost.

After determining all the beads, it is necessary to define the connectivity between them. In the case of many small molecules, single bead can represent the whole molecule. Small molecules with chemically distinct regions should be represented as a short chain of more than one bead. Real polymers are represented by a Gaussian chain of respective beads (intra molecular interactions are described by harmonic oscillator potentials) that exhibit the same response functions as the original chain. It is worth to mention, that the structures of the real and Gaussian chains can be different. Linear polymer can be represented by a Gaussian chain with branches and vice versa.

The next step after finding the coarse-grained topology is to determine the interactions between the species. Because the interactions must correspond to the interactions of real molecules and they should be easily calculated, the interactions in MesoDyn are calculated via the Flory-Huggins interaction parameters χ_{IJ} defined for each pair of the species presented in the system.

There are several ways how to determine the Flory-Huggins mixing parameters χ. It can be gained experimentally from e.g. the partial vapour pressure of solvent-polymer solutions or interfacial tension. It can be also calculated from the energy of mixing.

There is a simple relation between Flory-Huggins mixing parameters between the components I and J and Hildebrand solubility parameters δ_I and δ_J of these components.

$$\chi_{IJ} = \frac{V_{ref}(\delta_I - \delta_J)^2}{RT} + \chi_s, \tag{10}$$

where V_{ref} is a reference volume - a mean molar volume of components I and J and χ_s is the entropy contribution to the mixing energy. This term can be usually neglected, because it is only a small correction to the first term.

Hildebrand solubility parameters can be found in the polymer handbooks or they can be directly calculated from molecular dynamics simulations using the formula

$$\delta = \sqrt{\frac{E_{coh}}{V}}, \tag{11}$$

where $\frac{E_{coh}}{V}$ is a cohesive energy density.

4 Mesoscale model of Nafion

This section describes the mesoscale model of Nafion morphology, that was originally published in [5].

The limitation of the dynamic self-consistent mean field theory, which is implemented in MesoDyn, is that there is only one reference volume. So it means, that all beads should occupy approximately the same volume.

The natural choice is to take the perfluorinated side chain as a single bead S. Its volume is $0.31 nm^3$. This volume corresponds to the four $(-CF_2-CF_2-)$ monomer groups ($0.33 nm^3$) (bead P) and approximately to ten water molecules (bead W). The average between these volumes ($0.32nm^3$) seems then as a good choice for the bead reference volume. The bead reference volume v and the reference volume V_{ref} used in the equation (10) are simply connected through Avogadro number - $V_{ref} = vN_A \doteq 1.9 \times 10^{-4} m^3.mol^{-1}$.

The volumes of the beads were calculated using the SYNTHIA module.

Because of the average length of Nafion chains, a single chain of Nafion with equivalent weight of $1100 \mathrm{g.mol}^{-1}$ is represented by twenty repeating PPS monomers. This choice leads to the coarse-grained structure depicted in figure 2.

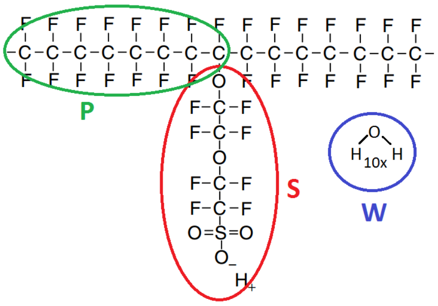


Figure 2: Coarse-grained structure of Nafion

4.1 Molecular modelling of the cohesive energy density

The choice of the interaction energies ε_{IJ} is crucial for obtaining the correct morphologies. Hildebrand solubility parameters (11) will be calculated here.

The value of cohesive energy density is very sensitive to the used force field. The molecular dynamic and ab initio simulations showed that COMPASS force field describes all the important interactions in Nafion polymer with a sufficient accuracy. Wescott et al. [5] slightly modified default COMPASS atom typing and partial charge assignments in order to have higher agreement with experimental results. Their atom typing assignments were successfully used in several studies. The same modified COMPASS force field were used during the calculation.

Three bulk amorphous models of each beads were generated with AMORPHOUS CELL in order to obtain the densities of cohesive energy for each bead.

One amorphous cell was filled with four polymers composed of 100 C_2F_4 monomers, the other amorphous cell was filled with 80 side chains. These two cells were generated at a density of $2.05 \mathrm{g.cm}^{-3}$ that corresponds to the experimental value of Nafion density.

The third amorphous cell was filled with 300 water molecules with at a density of $1 \mathrm{g.cm}^{-3}$.

The minimization of each of the cells with smart algorithm was used after the construction. Then each of the cells was equilibrated by molecular dynamic simulation in NVT ensemble starting at 200 K. The temperature increased to 300 K in steps of 25 K with 20 ns dynamics with 1 ps time step. After equilibration, another 300 ns of NVT dynamics were carried to obtain the average value of the cohesive energy density. Andersen thermostat was used during the calculations.

There is a problem with the value of solubility parameter for water. The calculation gives us the value of 47.2 $\text{MPa}^{\frac{1}{2}}$. However this value is really high and leads to demixing of all Nafion - water mixture at all hydration levels which does not correspond to the experimental observations. Futerko and Hsing solved this problem by defining an effective value of this parameter. They suggested the value $\delta_W = 25 \quad \text{MPa}^{\frac{1}{2}}$ for Hildebrand solubility of water and this value was consistent with their measurements.

The reduction of the value of Hildebrand solubility parameter to the value $\delta_W = 23 \quad \text{MPa}^{\frac{1}{2}}$ was eventually done in order Crank-Nicholson scheme to converge.

This reduction and the values of solubility parameters for Teflon ($\delta_P = 13.3 \quad \text{MPa}^{\frac{1}{2}}$) and for side chain ($\delta_S = 21.2 \quad \text{MPa}^{\frac{1}{2}}$) lead to the following Florry-Huggins parameters used in MesoDyn simulation

$$\chi_{PS} = \frac{11.8}{RT}, \quad \chi_{PW} = \frac{17.8}{RT} \quad \text{and} \quad \chi_{WS} = \frac{0.6}{RT}.$$

4.2 Calculation of mesoscale structure

The details of the MesoDyn calculation of Nafion microstructure followed by the discussion of the results will be described in this subsection. Nafion membrane with equivalent weight 1100 g.mol^{-1} is comprised of chains with the same lengths. Each chain is represented by twenty repeating coarse-grained monomers FF(S), where bead F represents four C_2F_4 groups and bead S represent the whole side chain in the original Nafion structure. Ten water molecules are included in the W bead.

The simulations were carried on the cubic lattice with volume $(29nm)^3$ and grid resolution $0.9nm$. The calculation started from a homogeneous distribution of each bead and the morphologies were equilibrated in $150\mu s$. This relaxation time corresponds to the 5000 time steps. The simulation temperature was 300 K.

Phase separation of beads can be characterized by their order parameters. The order parameter is defined as

$$P_I = \frac{1}{V} \int_V ({\theta_I}^2(r) - {\theta_I}^2)dr, \tag{12}$$

where θ_I is dimensionless density (volume fraction) for bead I. The higher is the value of order parameter, the higher is the phase segregation.

You can see the values of the order parameters for different level of hydration in Nafion in figures 3, 4 and 5.

It seems that no phase segregation occurs for low level of hydration in Nafion (figure 3). The values of order parameters for respective beads increase with the increasing level of hydration inside the membrane. The rate of segregation also grows with increasing hydration which is in agreement with experimental results. However the phase segregation should occur even in dry Nafion, so the model needs to be modified.

You can see no phase segregation for $\lambda = 2$ also in figure 6 and high phase segregation for $\lambda = 10$ also in figure 7. These figures show the density distributions of respective beads.

However these are just intermediate results and the model need to be improved to be in better agreement with experimental results. It seems that the values of the Hildebrand solubility parameters from molecular dynamics do not correspond to the reality perfectly.

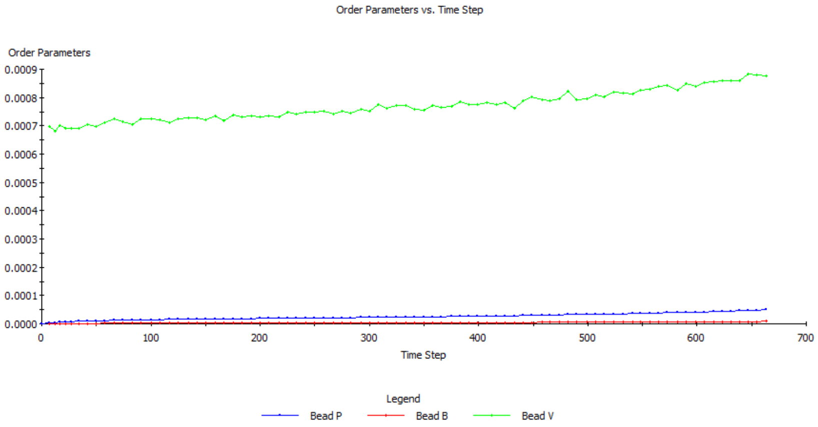


Figure 3: Order parameter for $\lambda = 2$

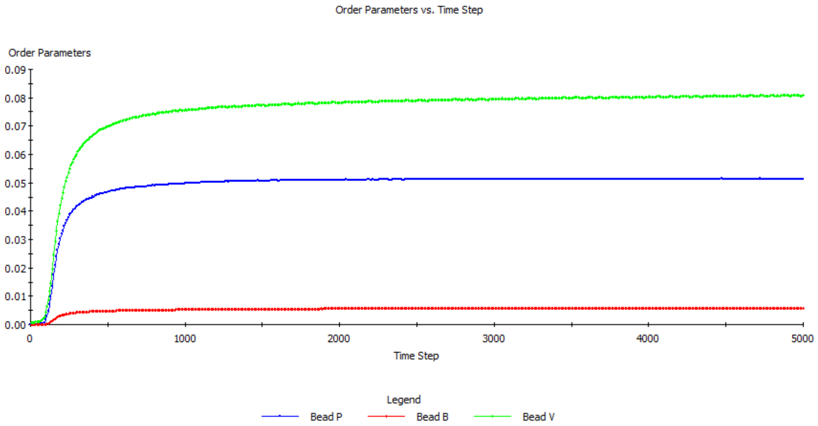


Figure 4: Order parameter for $\lambda = 4$

So it would be better to take them as a first trial and then slightly change these values to obtain higher correspondence with the experimental results.

4.3 Conclusion and further work

The aim of my thesis is to find the model of Nafion microstructure, because we believe that it has to be incorporated to the constitutive relation. The mesoscale model of Nafion morphology was introduced in this study, however this model needs to be further improved.

The values of Hildebrand solubility parameters do not correspond to the reality perfectly. So they will be taken as a first trial and their values will be changed to obtain higher correspondence with the reality.

Also the size of the beads in the original model is too big according to us, so the next step is to create different coarse-graining structure of Nafion.

The sulphonic acid has different properties from the rest of the side chain. So it will

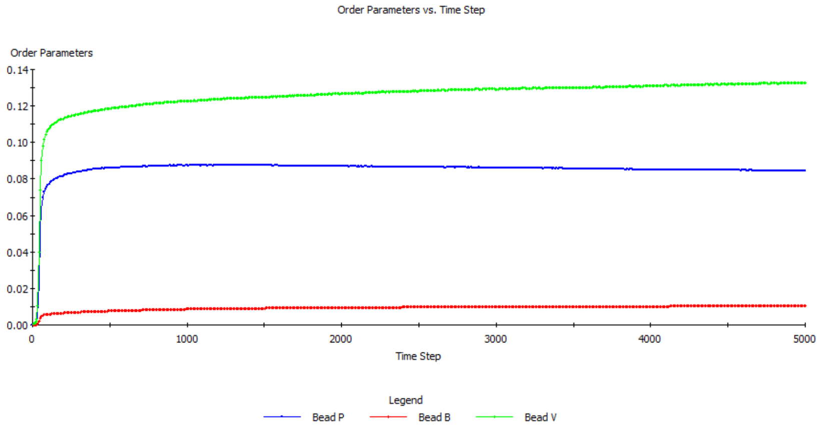


Figure 5: Order parameter for $\lambda = 8$

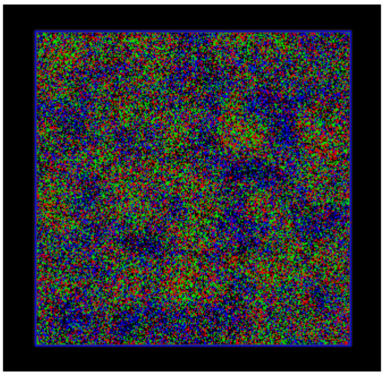


Figure 6: Density distributions for $\lambda = 2$

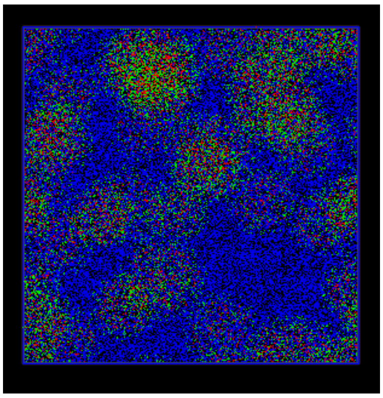


Figure 7: Density distributions for $\lambda = 10$

be taken as one bead and the rest of the structure will be transformed according to this choice of bead volume.

The electrostatics will be not comprised in Florry-Huggins parameters, but it will be calculated explicitly.

Calculating parameters for this model with finer coarse-grained structure will be my task in the following months.

References

[1] P. Altevogt, O.A. Evers, J.G.E.M. Fraaije, N.M. Maurits, B.A.C. van Vlimmeren. *The MesoDyn project: software for mesoscale chemical engineering* Journal of Molecular Structure (Theochem) 463 (1999), 139-–143

[2] J.G.E.M. Fraaije, B.A.C. van Vlimmeren, N.M. Maurits, M. Postma, O.A. Evers, C. Hoffman, P. Altevogt, G. Goldbeck-Wood. *The dynamic mean-field density func-*

tional method and its application to the mesoscopic dynamics of quenched block copolymer melts Journal of chemical physics 106 (1997), issue 10, 4260–4270

[3] K.-D. Kreuer, S.J. Paddison, E. Spohr, M. Schuster. *Transport in Proton Conductors for Fuel-Cell Applications: Simulations, Elementary Reactions, and Phenomenology* Chemical Reviews 2004, 104, 4637–4678

[4] S. McLaughlin. *The Electrostatic Properties of Membranes* Annual Review of Biophysics and Biophysical Chemistry Vol. 18, 113-136 (1989)

[5] J.T. Wescott, Y. Qi, L. Subramanian, T.W. Capehart. *Mesoscale simulation of morphology in hydrated perfluorosulfonic acid membranes* Journal of chemical physics 124 (2006), 134702(14)

[6] Accelrys Materials studio - manual for MESODYN software

Monte Carlo Estimation of EEG Correlation Dimension

Lucie Tylová

3rd year of PGS, email: tylovluc@fjfi.cvut.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Electroencephalography examination (EEG) records the brain activity. That makes it important part of neurological diseases diagnosis, eg. Alzheimer's disease. Modern methods describe EEG signal as a chaos. With such an approach, new characteristics of chaotic systems are calculated. The correlation dimension is one of those properties, however, its estimation requires high time-complexity. The article compares EEG time series to known chaotic system of Brownian motion and describes correlation dimension approach using Monte Carlo method.

Keywords: EEG, Brownian motion, correlation dimension, Monte Carlo

Abstrakt. Elektroencefalografické vyšetření (EEG) slouží k zaznamenávání mozkové aktivity. Proto je důležitou součástí při diagnostice neurologických chorob, např. Alzheimerovy choroby. Moderní postupy nahlížejí na EEG signál jako na chaos. Takový přístup tak přináší nové charakteristiky popisující tento systém. Jednou z nich je i korelační dimenze, jejíž výpočet je však časově náročný. Článek srovnává EEG signál se známým chaotickým systémem Brownova pohybu a popisuje odhad korelační dimenze pomocí metody Monte Carlo.

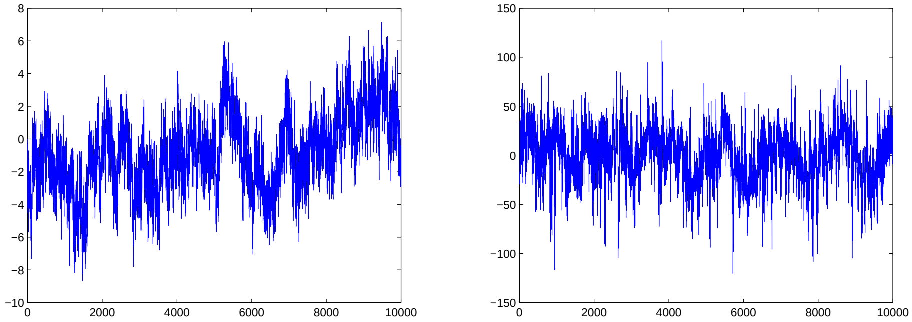
Klíčová slova: EEG, Brownův pohyb, korelační dimenze, Monte Carlo

1 Introduction

Electroencephalography (EEG) signal analysis have been widely adopted. The aim is an early detection of disorders or confirmation of disease diagnosis which change the brain activity in different ways. EEG electrodes record the sum of the graded potentials of the many thousand underlying neurons. The time series of EEG signal seems to have irregular and chaotic progress, at first sight, but we can recognize waves with some periodicity too [1].

Beside time-frequency analysis, EEG signal can be considered to be generated by non-linear dynamic systems with chaotic behaviour. One of the values used for description of chaotic systems is correlation dimension D_2. Its calculation involves some difficulties as algorithms developed from mathematical theorems are valid for noiseless and endless chaotic processes. EEG signal does not meet this condition, EEG times series are not endless, but experimental data are long enough to make D_2 calculation very time-expensive.

The aim of this article is to compare EEG signal data to a chaotic system with known value of correlation dimension and approach this value using Monte Carlo method to eliminate time complexity.



(a) Brownian motion ($H = 0.1$)

(b) EEG signal

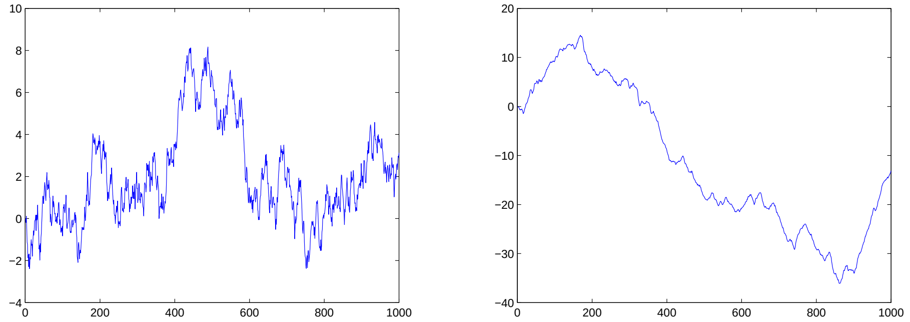
Figure 1: Visualization of chaotic system and real data.

2 Fractional Brownian Motion

Fractional Brownian motion (fBm) is a Gaussian process $\mathrm{B}_H(\mathrm{t})$ with stationary increments and zero mean, which depends on a parameter $H \in (0,\, 1)$ called the Hurst exponent [2]. For $H = \frac{1}{2}$, it is the standard Brownian motion $\mathrm{B}(t)$. For $H > \frac{1}{2}$, the increments of the process are positively correlated and they are negatively correlated for $H < \frac{1}{2}$. The higher value of H leads to a smoother motion as shown in Fig. 2.

Brownian motion is defined as a stochastic process $\mathrm{B}(\mathrm{t})$ that satisfies [3]:

- $\mathrm{B}(0) = 0,\ \forall\ t,$
- random variables $\mathrm{B}(t_2)$ - $\mathrm{B}(t_1)$ and $\mathrm{B}(t_4)$ - $\mathrm{B}(t_3)$ are independent for $0 < t_1 < t_2 < t_3 < t_4$,
- the variable $\mathrm{B}(t + s)$ - $\mathrm{B}(t)$ is a Gaussian variable with zero mean and standard deviation s, $\forall (s, t) \geq 0$,
- $\mathrm{B}(t)$ is a continuous function of t.



(a) $H = 0.5$

(b) $H = 0.9$

Figure 2: Fractional Brownian Motion

3 Correlation Dimension Monte Carlo Approach

The traditional characteristic of chaotic behaviour is called correlation dimension [4]. The correlation dimension measures the dimensionality of the object, usually attractor, formed from N points in some embedding space. Its value lies between topological and Hausdorff dimensions [5] according to inequalities

$$D_{\mathrm{T}} \leq D_2 \leq D_{\mathrm{H}}. \tag{1}$$

Taken as two-dimensional space for ease of vizualization, the case corresponds to a time series X_N where we draw a cirle of radius $r > 0$ around picked point and count number of points inside the cirle. Grassberger and Procaccia [6] suggest to measure the distance between every pair of points. Let $\mathrm{C}(r)$ be the number of points within all these circles. Then $\mathrm{C}(r)$ is called the correlation sum and is calculated by

$$\mathrm{C}(r) = \frac{2}{N(N-1)} \sum_{j=1}^{N} \sum_{i=j+1}^{N} \Theta(r - r_{i,j}) \tag{2}$$

where Θ is the Heaviside function, $r_{i,j} = \|\vec{x}_i - \vec{x}_j\|$, and N is a number of data points. $\mathrm{C}(r)$ converges to the correlation integral for $N \to \infty$ and can be inspected as the probability that two different randomly chosen points will be closer than r [7].

A slope of $\log \mathrm{C}(r)$ versus $\log r$ plot in the limit of small r and large N is the correlation dimension

$$D_2 = \lim_{r \to 0} \lim_{N \to \infty} \frac{\mathrm{d} \log \mathrm{C}(r)}{\mathrm{d} \log r}. \tag{3}$$

For finite N, D_2 can be estimated via LSQ method using a linearized model

$$\log \mathrm{C}(r) = A + D_2 \log r. \tag{4}$$

The main disadvantage of this approach is the time complexity of $\mathrm{C}(r)$ calculation for large N. The opossite problem is bias of D_2 estimate which comes with small N. Another possibility is to use the Monte Carlo approach. This methodology was desribed and tested on past results [8] as follows.

Let $M \in \mathbb{N}$ be number of Monte Carlo simulations [9]. Let $\Delta \in \mathbb{N}$ be given barrier. The approach is based on the Monte Carlo estimation of $\mathrm{C}(r)$ for $k = 1, ..., M$, where r_k are also results of simulation. Single simulation step is based on two random indices $i, j \sim \mathrm{U}(\{1, 2, ..., N\})$ which are repeatedly generated until $|i - j| > \Delta$. Vector distance $d_k = \|x_i - x_j\|$ is stored as result of k^{th} simulation. After M simulations, we sort d_k to obtain non-decreasing series of d_k and then

$$\begin{aligned} r_k &= d_{(k)}, \\ \mathrm{C}(r_k) &= k/M. \end{aligned} \tag{5}$$

Resulting pairs $(r_k, \mathrm{C}(r_k))$ are censored using contrains

$$p_{\min} \leq \mathrm{C}(r_k) \leq p_{\max} \tag{6}$$

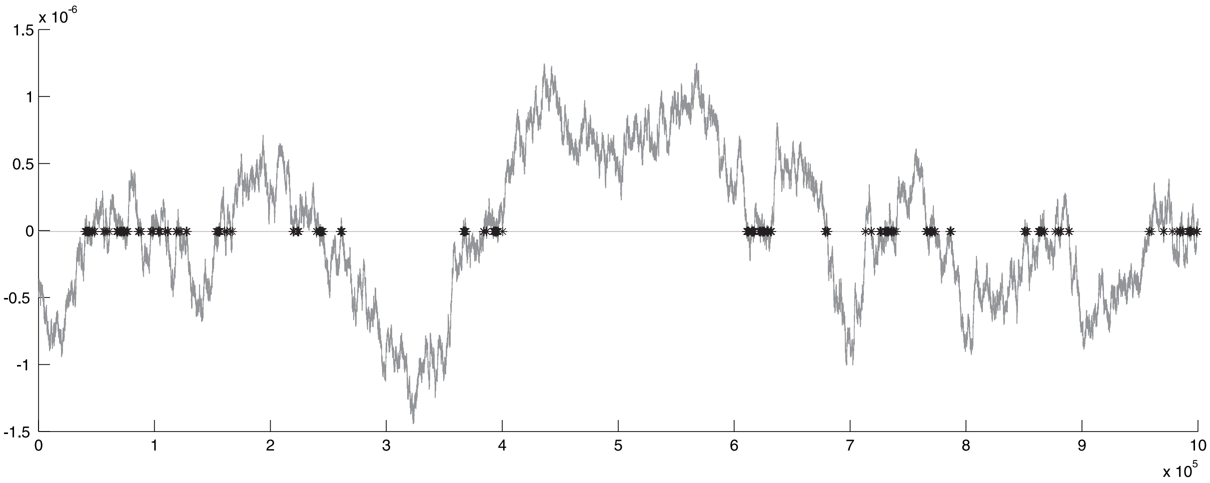


Figure 3: fBm ($H = 0.5$) crossing with a median line

to avoid disturbances for extreme values of r_k. Linear model (4) is then applied only to censored data from Monte Carlo experiment.

The main advantage of this novel approach is in decreasing of time complexity in the case of large time series but the estimation error also depends on $M, \Delta, p_{\min}$, and $p_{\max}$. Meanwhile, the time complexity of original correlation sum calculations is $T(N) = O(N^2)$ per each fixed parameter value r, the time complexity of this approach is only $T(M) = O(M \log M)$ due to sorting complexity.

4 Numerical Experiment

Monte Carlo approach was verified on Fractional Brownian motion levelsets. Levelsets, the generalization of zero sets $Z = \{t \neq 0 | \mathrm{B}_H(t) = 0\}$ [10], are obtained by the crossing of $B_H(t)$ with a constant line $\mathrm{B}_H(t) = c$. Moreover, the levelsets have known fractal dimension $1 - H$ [3].

The fBm time series $X = \{x_k\}_{k=0}^{N}$ for $N = 10^6$ were generated first and the trend was subtracted. The median value was set as c parameter for the levelset line. The time series and the levelset line crossing points were used in Monte Carlo estimation.

The main aim of simulations is to map D_2 approximation on fBm for different Hurst exponent values $H = 0.1, ..., 0.9$ with simulation length $M = 10^4$ which was chosen due to time complexity. Monte Carlo estimation is performed for $L = 10$ loops to obtain its mean value $\mathrm{E}D_2$ and standard deviation s. Knowning only theoretical value of D_2^*, the point estimates in one-sample t-test are used as a kind of statistical pesimism. Results of the estimation are shown in Tab. 1 including $p-$value of one-sample two-sided t-test of H_0 about $\mathrm{E}D_2$ and D_2^* equity.

5 Alzheimer's Disease Testing

Alzheimer's disease is an irreversible neurological disorder that causes dementia. It physically defects neurons and their synapses. The result is a loss of memory, thinking, and

Table 1: D_2 simulation for fBm

H	D_2^*	$\mathrm{E}D_2$	s	$p-$value
0.1	0.9	0.9171	0.0095	0.1058
0.2	0.8	0.8385	0.0263	0.1777
0.3	0.7	0.7213	0.0273	0.4560
0.4	0.6	0.5739	0.0314	0.4298
0.5	0.5	0.5053	0.0378	0.8921
0.6	0.4	0.4300	0.0270	0.2947
0.7	0.3	0.3814	0.0301	0.0242
0.8	0.2	0.2899	0.0203	0.0017
0.9	0.1	0.4185	0.1120	0.0223

language skills. Due to changes in the brain, EEG takes a part in Alzheimer's disease study [11] as a research tool.

For testing, the real biomedical signal data were used. EEG time series were obtained from two examined groups of patients consist of 139 control normals (CN) and 26 with Alzheimer's disease (AD) diagnosis. The signal was recorded in a form of multichannel EEG using the standard 10-20 scheme with nineteen channels and two reference electrodes [12].

Parameters of the simulation were kept same as for the Fractional Brownian motion estimation. Instead of iterations, individiual D_2 values were simulated for each channel for each group of patients. The length of EEG data varied patient by patient starting at a 5 minutes minimum with the sampling frequency 200 Hz. The correlation sum was calculated from $M = 10^4$ pairs of points.

Resulting values of the mean $\mathrm{E}D_2$ and the standard deviation s of correlation dimension approaches are collected in Tab. 2. There is no theoretical number representing ideal correlation dimension value D_2^* for EEG time series. Therefore, for a statistical description of this model, a two-sample t-test was applied on significant level $\alpha = 0.05$. This method tests the hypothesis that two indepedent samples come from distributions with equal means $H_0 : \mathrm{E}X = \mathrm{E}Y$ [13]. Calculated $p-$value are also in Tab. 2.

To avoid false positive results, the standard methodology of False Discovery Rate (FDR) was used [14]. The corrected critical value was determined as $\alpha_{FRD} = 0.0030$. According to this value, the third channel is significant. Other channels with $p-$value < 0.05 are 5, 9, and 11 which has $p-$value equal to α_{FRD}.

6 Conclusions

The Monte Carlo approach of correlation dimension was successfully tested on D_2 of Fractional Brownian motion crossing its median levelset line with various Hurst exponent values. Along with 10^4 simulations, it decreases the time complexity as expected. Acceptable results were obtained also on real data set which was EEG signal obtained from patients with control normal and with confirmed Alzheimer's disease. Statistically significant difference between these two groups was estimated on the third channel. The

Table 2: EEG D_2 simulation results

	CN		AD		
Ch	ED_2	s	ED_2	s	$p-$value
1	0.9622	0.0020	0.9547	0.0039	0.0370
2	0.9564	0.0019	0.9628	0.0034	0.6090
3	0.9563	0.0020	0.9688	0.0039	**0.0018**
4	0.9570	0.0019	0.9573	0.0028	0.7963
5	0.9600	0.0018	0.9620	0.0043	0.0156
6	0.9617	0.0018	0.9557	0.0045	0.3610
7	0.9575	0.0018	0.9571	0.0047	0.8129
8	0.9592	0.0019	0.9665	0.0043	0.6127
9	0.9603	0.0018	0.9548	0.0049	0.0179
10	0.9586	0.0019	0.9550	0.0047	0.0671
11	0.9626	0.0017	0.9644	0.0032	0.0030
12	0.9603	0.0019	0.9578	0.0054	0.1420
13	0.9560	0.0019	0.9706	0.0050	0.2361
14	0.9559	0.0019	0.9610	0.0046	0.8306
15	0.9592	0.0019	0.9644	0.0040	0.6199
16	0.9586	0.0020	0.9604	0.0032	0.9315
17	0.9591	0.0019	0.9564	0.0057	0.1604
18	0.9590	0.0019	0.9545	0.0032	0.7327
19	0.9598	0.0017	0.9512	0.0042	0.1923

third electrod is placed over the frontal lobe as shown in Fig. 4 where also other channels with low $p-$value are highlighted.

7 Discussion

The result of EEG analysis confirms conclusions of past studies and is in accordance with biomedical hypotheses that Alzheimer's disease causes atrophy mainly at frontal and temporal lobes [15]. Due to a kind of statistical pesimism, only the third channel was accepted as significant. The same channel was also significant for EEG linear prediction study [16]. Better results could be obtained by changing model driving parameters as a simulation length. Other important parameter is the length of tested data. The model was tested for crossing only with median. The next improvement could be crossing with more lines on different levels producing more points for comparison.

References

[1] A. Mekler. *Calculation of EEG Correlation Dimension: Large Massifs of Experimental Data.* Computer Methods and Programs in Biomedicine **92** (2008), 154–160.

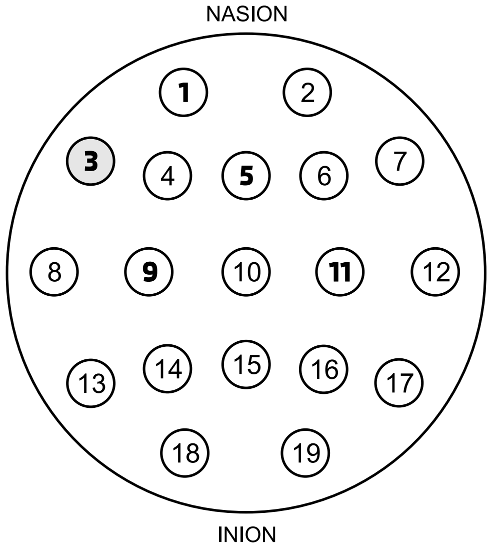


Figure 4: Electrode 10-20 scheme

[2] D. Nualart *Fractional Brownian Motion: Stochastic Calculus and Applications.* International Congress of Mathematicians, Madrid, Spain (2006).

[3] J. Gao, Y. Cao, W. Tung, J. Hu. *Multiscale Analysis of Complex Time Series.* John Wiley & Sons (2007).

[4] P. Grassberger. *Generalized dimensions of strange attractors.* Physics Letters A **97** (1983), 227–230.

[5] J. C. Sprott. *Chaos and Time-Series Analysis.* Oxford University Press (2003).

[6] P. Grassberger, I. Procaccia. *Characterization of Strange Attractors.* Physical Review Letters **50** (1983), 346–349.

[7] G. P. DeCoster, D. W. Mitchell. *The Efficacy of the Correlation Dimension Technique in Detecting Determinism in Small Samples.* Journal of Statistical Computation and Simulation **39** (1991), 221–229.

[8] L. Tylová. *Monte Carlo Estimation of Correlation Dimension for EEG Analysis.* Doktorandské dny 2013 (2013), 293–300.

[9] C. Z. Mooney. *Monte Carlo Simulation.* SAGE Publications (1997).

[10] A. Dahl. *A Rigorous Introduction to Brownian Motion.* Summer Virge REU(2010).

[11] R. Rusina, K. Sheardova, I. Rektorova, P. Ridzon, P. Kulistak, R. Matej. *Amyotrophic Lateral Sclerosis and Alzheimer's Disease – Clinical and Neuropathological Considerations in Two Cases.* European Journal of Neurology **14** (2007), 815–818.

[12] W. O. Tatum. *Handbook of EEG Interpretation.* Demos Medical Publishing (2007).

[13] M. Meloun, J. Militky. *The statistical analysis of experimental data.* Academia (2004).

[14] Y. Benjamini, Y. Hochberg. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society **57** (1995), 289–300.

[15] A. Redolfi, P. Bosco, D. Manset, G. B. Frisoni. *Brain Investigation and Brain Conceptualization.* Functional Neurology **28** (2013), 175–190.

[16] L. Tylova, J. Kukal, O. Vysata. *Predictive Models in Diagnosis of Alzheimer's Disease from EEG.* Acta Polytechnica **53** (2013), 94–97.

Real Functions Computable by Finite State Transducers in Möbius Number Systems*

Tomáš Vávra

3rd year of PGS, email: t.vavra@seznam.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Zuzana Masáková, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Petr Kůrka, Center for Theoretical Study, Charles University in Prague

Abstract. We consider Möbius number systems with sofic expansion subshift. Let $F : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ be an analytic function defined on the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. We show that if F is computable by a finite state transducer, then it is in fact a Möbius transformation, that is, $F(x) = \frac{ax+b}{cx+d}$, $ad - bc \neq 0$. The same problem for a different number system was also studied in [1]. Furthemore, we show that unlike in modular Möbius systems, in bimodular systems, not every rational Möbius transformation is computable by a finite state transducer.

This contribution has been presented at a conference by the author and the results have been published in [2].

Keywords: exact real algorithms, transducers, möbius

Abstrakt. Uvažujeme Möbiovské číselné systémy se sofickým expanzním subshiftem. NechťF : $\overline{\mathbb{R}} \to \overline{\mathbb{R}}$ je analytická funkce definovaná na $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Ukážeme, že pokud je funkce F počitatelná konečným transducerem, potom je F Möbiova transformace, tzn. $F(x) = \frac{ax+b}{cx+d}$, kde $ad - bc \neq 0$. Stejný problém, ale pro jiný systém, byl také studován v [1]. Navíc ukážeme, že narozdíl od modulárních Möbiovských systémů, v bimodulárních systémech není každá Möbiova transformace počitatelná konečným transducerem.

Tento příspěvek byl prezentován autorem na konferenci a výsledky byly publikovány v [2].

Klíčová slova: exaktní reálné algoritmy, transducery, möbius

References

[1] M. Konečný, *Real functions incrementally computable by finite automata,* Theoretical Computer Science, 315(1), 109–133, 2004.

[2] P. Kůrka, T. Vávra, *Analytic functions computable by finite state transducers,* Lecture Notes in Computer Science 8587, 252–263, 2014.

*This work has been supported by the Czech Science Foundation research project GAČR 13-03538S and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS14/205/OHK4/3T/14

Higher Dorfman Bracket, Automorphisms, and Integration of its Infinitesimal Symmetries

Jan Vysoký

4th year of PGS, email: vysokjan@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Branislav Jurčo, Mathematical Institute, Charles University in Prague

Abstract. We define and review basic properties of a higher Dorfman bracket, an extension of a vector field commutator to a direct sum of tangent bundle and a p-fold wedge product of a cotangent bundle. Lie algebra of its derivations and group of its automorphisms is calculated. We introduce a notion of integration of first order differential operators. We find an explicit formula for integrating automorphism of infinitesimal symmetries of higher Dorfman bracket. Examples and an application to infinitesimal isometries of generalized metric are included.

Keywords: Generalized geometry, higher Dorfman bracket, automorphisms, generalized metric

Abstrakt. Definujeme a shrneme vlastnosti vyšší Dorfmanové závorky, rozšíření komutátoru vektorových polí na direktní součet tečného bundlu a p-násobného vnějšího součinu kotečného bundlu. Spočteme Lieovu algebru jejích derivací a grupu jejích automorfismů. Definujeme pojem integrace diferenciálních operátorů prvního řádu a nalezneme explicitní vzorec pro integrující automorfismus infinitesimálních symetrií vyšší Dorfmanové závorky. Na závěr zahrneme příklady a aplikaci na infinitesimální izometrie zobecněné metriky.

Klíčová slova: Zobecněná geometrie, vyšší Dorfmanové závorky, automorfismy, zobecněná metrika

1 Definition, basic properties

Let p be any non-negative integer. Consider a vector bundle $E = TM \oplus \Lambda^p T^*M$, where we identify $\Lambda^0 T^*M \cong M \times \mathbb{R}$. There exists a well-known extension of a vector field commutator to a bracket on $\Gamma(E)$, for $p = 1$ called a Dorfman bracket [3, 8]. It constitutes a simplest example of Courant algebroid. For $p > 1$, there is a bracket on $\Gamma(E)$, usually called a higher Dorfman bracket. Let us denote the sections of E as ordered pairs. For $e \in \Gamma(E)$, we will write $e = (x, a_p)$, where $x \in \mathfrak{X}(M)$ is a vector field on M and $a_p \in \Omega^p(M)$ is a p-form on M. A higher Dorfman bracket is defined as

$$(x, a_p) \circ (y, b_p) = ([x, y], \mathcal{L}_x b_p - i_y da_p), \tag{1}$$

for all $(x, a_p), (y, b_p) \in \Gamma(E)$. For a detailed discussion of topics related to this definition and its skew-symmetric counterpart, see [1, 4]. Since there is no known proper higher analogue of a Courant algebroid, one has to stick to a more general notion of Leibniz algebroid. Leibniz algebroid is a triple $(E, \rho, \circ)$, where E is a vector bundle, $\rho : E \to TM$ is a vector bundle morphism, and $\circ : \Gamma(E) \times \Gamma(E) \to \Gamma(E)$ is an $\mathbb{R}$-bilinear bracket, satisfying the following axioms:

1. $e \circ (fe') = f(e \circ e') + (\rho(e).f)e'$, *(Leibniz rule)*
2. $e \circ (e' \circ e'') = (e \circ e') \circ e'' + e' \circ (e \circ e'')$, *(Leibniz property)*

for all $e, e', e'' \in \Gamma(E)$. To avoid confusion, Leibniz property is in some literature called a Loday identity, and Leibniz algebroid a Loday algebroid.

It is not difficult to show that if one chooses $\rho = pr_{TM}$, a projection onto the first summand of E, then $(E, \rho, \circ)$ from the first paragraph of this section forms a Leibniz algebroid. It is not a Courant algebroid, since there is no canonical $\mathbb{R}$-valued bilinear pairing on E.

Leibniz property establishes a following property. For each $e \in \Gamma(E)$, we can define a $\mathbb{R}$-linear map $\chi(e) : \Gamma(E) \to \Gamma(E)$ using a higher Dorfman bracket:

$$(\chi(e))(e') = e \circ e', \tag{2}$$

for all $e, e' \in \Gamma(E)$. Leibniz property then states that $\chi(e)$ is a derivation of the higher Dorfman bracket:

$$(\chi(e))(e' \circ e'') = (\chi(e))(e') \circ e'' + e' \circ (\chi(e))(e''). \tag{3}$$

This is equivalent to an observation that χ is a bracket homomorphism in a sense that

$$[\chi(e), \chi(e')] = \chi(e \circ e'), \tag{4}$$

for all $e, e' \in \Gamma(E)$. This can be easily proved directly from the definition of χ and a Leibniz property. χ can be thus viewed as a $\mathbb{R}$-linear representation of $\circ$ on $\Gamma(E)$. Note that it is not a representation by $C^\infty(M)$-linear maps (that is vector bundle morphisms of E). Instead, there holds $(\chi(e))(fe') = f(\chi(e))(e') + (\rho(e).f)e'$ for all $e, e' \in \Gamma(E)$ and $f \in C^\infty(M)$. This also shows that $\chi(e)$ is a vector bundle morphism, iff $\rho(e) = 0$.

2 Derivations of the bracket

We will now examine a special class of $\mathbb{R}$-linear endomorphisms of $\Gamma(E)$. Let $\mathcal{F} : \Gamma(E) \to \Gamma(E)$ be an $\mathbb{R}$-linear map, and assume that there is an $\mathbb{R}$-linear operator D on a vector space $C^\infty(M)$, such that

$$\mathcal{F}(fe) = f\mathcal{F}(e) + (D.f)e. \tag{5}$$

Consistency on the products of two functions requires D to satisfy:

$$D.(fg) = (D.f)g + f(D.g). \tag{6}$$

This implies that D is a vector field on M. Denote this vector field as $x \in \mathfrak{X}(M)$. Condition (5) can be called "a locality property", since it ensures that $\mathcal{F}$ does not depend on $e \in \Gamma(E)$ in its entirety, but only on the values of e in a neighborhood of every point. This means that if $e|_U = e'|_U$ in some neighborhood in M, then $\mathcal{F}(e)|_U = \mathcal{F}(e')|_U$.

Said in a slightly different language, see [7], $\mathcal{F}$ satisfying (5) are called *first order differential operators* on E, and they form a space of sections of a vector bundle $\mathcal{D}(E)$.

The map assigning to $\mathcal{F}$ a vector field x can be then viewed as vector bundle morphism $a : \mathcal{D}(E) \to TM$. It follows that $(\mathcal{D}(E), a, [\cdot, \cdot])$ forms a Lie algebroid.

Moreover, assume that $\mathcal{F}$ acts as a derivation of the higher Dorfman bracket (1), that is there holds

$$\mathcal{F}(e \circ e') = \mathcal{F}(e) \circ e' + e \circ \mathcal{F}(e'). \tag{7}$$

Combining the properties (5) and (7) yields the equation

$$\rho(F(e)) = [x, \rho(e)], \tag{8}$$

for all $e \in \Gamma(E)$. Recall that given $x \in \mathfrak{X}(M)$, we have the map $\chi(x, 0)$, defined in the previous section, which satisfies (5), (7) and (8). Define a new map $\mathcal{G} : \Gamma(E) \to \Gamma(E)$ as

$$\mathcal{F} = \chi(x, 0) + \mathcal{G}.$$

It is easy to see that $\mathcal{G}$ is a $C^\infty(M)$-linear endomorphism of $\Gamma(E)$, that is a vector bundle endmorphism of E (over identity). This means that vector field in (5) is zero for $\mathcal{G}$. Moreover, since the space of derivations is a vector space, $\mathcal{G}$ is also a derivation of the bracket (1). It follows from (8) that $\rho(\mathcal{G}(e)) = 0$. We can thus write $\mathcal{G}(e)$ in the block form

$$\mathcal{G}(y, b_p) = \begin{pmatrix} 0 & 0 \\ G_1 & G_2 \end{pmatrix} \begin{pmatrix} y \\ b_p \end{pmatrix},$$

where a division into blocks corresponds to the sum $TM \oplus \Lambda^p T^* M$. Plugging back into the condition (7) gives the conditions:

$$G_1(y) = -i_y C, \; G_2(b_p) = \lambda \cdot 1, \tag{9}$$

where $C \in \Omega^{p+1}_{closed}(M)$ and $\lambda \in \Omega^0_{closed}(M)$. We have thus proved that the most general map $\mathcal{F}$ satisfying (5) and (7) is of the form

$$\mathcal{F}(y, b_p) = (\chi(x, 0))(y, b_p) + (0, -i_y C + \lambda b_p) = ([x, y], \mathcal{L}_x b_p + \lambda b_p - i_y C), \tag{10}$$

for $(y, b_p) \in \Gamma(E)$, $C \in \Omega^{p+1}_{closed}(M)$ and $\lambda \in \Omega^0_{closed}(M)$, and $x \in \mathfrak{X}(M)$. Each map $\mathcal{F}$ is thus determined uniquely by a triple (x, C, λ). The space of all derivations of $\circ$ denoted as $\mathrm{Der}_\circ(E)$ is a Lie algebra. One finds that

$$[(x, C, \lambda), (x', C, \lambda')] = ([x, x'], (\mathcal{L}_{x'} - \lambda')C - (\mathcal{L}_x - \lambda)C', 0). \tag{11}$$

This proves that $\mathrm{Der}_\circ(E) \cong \mathfrak{X}(M) \ltimes (\Omega^{p+1}_{closed}(M) \rtimes \Omega^0_{closed}(M))$, where $\Omega^{p+1}_{closed}(M)$ and $\Omega^0_{closed}(M)$ are viewed as Abelian Lie algebras. To conclude this section, recall that $\chi(e)$ presents an example of an element of $\mathrm{Der}_\circ(E)$.

3 Automorphisms of the bracket

In this section, we will examine the vector bundle automorphisms of E, preserving the higher Dorfman bracket. We will in fact show that the resulting group $\mathrm{Aut}_\circ(E)$ has $\mathrm{Der}_\circ(E)$ as its Lie algebra (not being rigorous, no infinte-dimensional manifolds are discussed). We will now consider vector bundle automorphisms over diffeomorphisms, that is pairs $(\mathcal{F},\psi)$, where $\psi \in \mathrm{Diff}(M)$ is a diffeomorphism, and $\mathcal{F} : E \to E$ is a smooth map, such that $\mathcal{F} : E_m \to E_{\psi(m)}$ is a a linear isomorphism for each $m \in M$. Note that each $(\mathcal{F},\psi)$ induces a $\mathbb{R}$-linear automorphism of $\Gamma(E)$ defined as $(\mathcal{F}(e))(\psi(m)) = \mathcal{F}(e(m))$, for all $e \in \Gamma(E)$ and $m \in M$. Map $\mathcal{F}$ can be characterized by a property similar to (5):

$$\mathcal{F}(fe) = (f \circ \psi^{-1})\mathcal{F}(e). \tag{12}$$

This condition in fact ensures that $\mathcal{F}$ depends only on the values of e at each point $m \in M$, that it if $e(m) = e'(m)$, then $\mathcal{F}(e)(\psi(m)) = \mathcal{F}(e')(\psi(m))$.

Now we will consider the subset of all vector bundle automorphisms of E, preserving the bracket (1). We thus impose a condition

$$\mathcal{F}(e \circ e') = \mathcal{F}(e) \circ \mathcal{F}(e'), \tag{13}$$

for all $e, e' \in \Gamma(E)$. We can again combine (12) and (13) to get a consistency condition:

$$(\rho(e).f) \circ \psi^{-1} = \rho(\mathcal{F}(e)).(f \circ \psi^{-1}). \tag{14}$$

for all $e \in \Gamma(E)$ and $f \in C^\infty(M)$. Let $\psi \in \mathrm{Diff}(M)$ be a diffeomorphism of M. Define a vector bundle automorphism $(T(\psi),\psi)$ of E by defining $T(\psi) : \Gamma(E) \to \Gamma(E)$ as

$$T(\psi)(y, b_p) = (\psi_*(y), \psi^{-1*}(b_p)). \tag{15}$$

It is easy to see that $(T(\psi),\psi)$ satisfies (12) and (13). Now take arbitrary $(\mathcal{F},\psi)$ satisfying (12,13). Since such vector bundle morphisms form a group, define a new automorphism $(\mathcal{G}, Id_M)$ by formula

$$(\mathcal{F},\psi) = (T(\psi),\psi) \circ (\mathcal{G}, Id_M). \tag{16}$$

Condition (14) for $(\mathcal{G}, Id_M)$ than says that $\rho(e) = \rho(\mathcal{G}(e))$. In other words, $\mathcal{G}$ has a block form

$$\mathcal{G}(y, b_p) = \begin{pmatrix} 1 & 0 \\ G_1 & G_2 \end{pmatrix} \begin{pmatrix} y \\ b_p \end{pmatrix},$$

where G_2 is an invertible map. Plugging back into (13) yields the conditions:

$$G_1(y) = -i_y C, \; G_2(b_p) = \lambda \cdot 1, \tag{17}$$

where $C \in \Omega^{p+1}_{closed}(M)$, and $\lambda \in \Omega^0_{closed}(M)$. Invertibility of G_2 implies that $\lambda(x) \neq 0$ for all $x \in M$. Denote the group (with respect to multiplication) of locally constant everywhere non-vanishing functions as $\widetilde{\Omega}^0_{closed}(M)$. We have thus found that $(\mathcal{F},\psi)$ has the form

$$\mathcal{F}(y, b_p) = (\psi_*(y), \psi^{-1*}(\lambda b_p - i_Y C)),$$

where $\psi \in \mathrm{Diff}(M)$, $C \in \Omega^{p+1}_{closed}(M)$ and $\lambda \in \widetilde{\Omega}^0_{closed}(M)$. Every element of $\mathrm{Aut}_\circ(E)$ is thus determined uniquely by a triple (ψ, C, λ). We can then find a composition rule:

$$(\psi, C, \lambda) \circ (\psi', C', \lambda') = (\psi\psi', \psi'^* C + (\psi'^* \lambda) C', (\psi'^* \lambda)\lambda'). \tag{18}$$

This proves the group isomorphism $\mathrm{Aut}_\circ(E) \cong \mathrm{Diff}(M) \ltimes (\Omega^{p+1}_{closed}(M) \rtimes \widetilde{\Omega}^0_{closed}(M))$. We see that this is in a good agreement with the previous results we have found for $\mathrm{Der}_\circ(E)$.

4 Integration of infinitesimal symmetries

We have shown in the first section, that to each $e \in E$, there corresponds a map $\chi(e) \in \mathrm{Der}_{\circ}(E)$. We suspect that there should be a corresponding 1-parameter subgroup of $\mathrm{Aut}_{\circ}(E)$ of the bracket automorphisms, integrating the map $\chi(e)$ in the following sense.

Let $\mathcal{G}$ be any first order differential operator on E, that is (5) holds. We say that 1-parameter subgroup of vector bundle automorphisms $(\mathcal{F}_t, \psi_t)$ *integrates* the map $\mathcal{G}$, if there holds

$$\mathcal{G}(e) = \frac{d}{dt}\Big|_{t=0} \mathcal{F}_{-t}(e), \tag{19}$$

for all $e \in \Gamma(E)$. Since ψ_t is 1-parameter subgroup of $\mathrm{Diff}(M)$, it corresponds to the flow of some vector field. It follows from (19), that it is exactly the vector field D in the locality property (5) for $\mathcal{G}$. This is another reason to consider first order differential operators on E as Lie algebra of $\mathrm{Aut}(E)$. We write formally $\mathcal{F}_{-t} = \exp(t\mathcal{G})$.

Let us start with some simple examples. Consider fist the map $\mathcal{G} = \chi(x, 0)$. We have $\mathcal{G}(y, b_p) = ([x, y], \mathcal{L}_x b_p)$. Let ϕ^x_t be a flow of $x \in \mathfrak{X}(M)$. One can guess that $(\mathcal{F}_t, \psi_t) = (T(\phi^x_t), \phi^x_t)$ will do the trick. We thus have to check that for all $m \in M$:

$$(\mathcal{G}(e))(m) = \frac{d}{dt}\Big|_{t=0} T(\phi^x_{-t})(e(\phi^x_t(m))) = \frac{d}{dt}\Big|_{t=0} (\phi^x_{-t*}(y|_{\phi^x_t(m)}), \phi^{x*}_t(b_p|_{\phi^x_t(m)})).$$

The right-hand side is exactly $([x, y], \mathcal{L}_x b_p)$ at m, as we wanted.

Second example is $\mathcal{G} = \chi(0, a_p)$. We thus have $\mathcal{G}(y, b_p) = (0, -i_y da_p)$. Since $\mathcal{G}$ is $C^\infty(M)$-linear, we get immediately that $\psi_t = Id_M$. It is then easy to guess that $\mathcal{F}_t(y, b_p) = (y, b_p + t i_y da_p)$. In this case it is almost trivial to verify the condition (19).

The main pursue of this section is to find the automorphism integrating the map $\chi(e)$ for general $e \in \Gamma(E)$, and show that it is an automorphism of higher Dorfman bracket (1). Main idea follows from the previous two paragraphs. We know how to integrate $\chi(e)$ for $e = (x, 0)$ and for $e = (0, a_p)$. We can thus guide our steps by suitably using the BCH formula. Define $X = \chi(x, 0)$ and $Y = \chi(0, a_p)$. Now observe that (4) implies that only non-trivial nested commutator of X's and Y's is the one containing single Y, and $(n-1)$-times X:

$$X_n \equiv [X, [X, \dots, [X, Y]] \dots]. \tag{20}$$

We will use a variant of the BCH formula, called a Zassenhaus formula [2]. This formula states that in the above special case, we have

$$e^{t(X+Y)} = e^{tX} \prod_{n=1}^{\infty} C_n(t, X, Y), \tag{21}$$

where $C_n(t, X, Y) = \exp((-1)^{n+1} \frac{t^n}{n!} X_n)$. Note that C_n commute with each other, and the product in (21) becomes an exponential of the sum:

$$e^{t(X+Y)} = e^{tX} \exp(\sum_{n=1}^{\infty} (-1)^{n+1} \frac{t^n}{n!} X_n).$$

Inverting this expression, and interchanging $X \leftrightarrow -X$ and $Y \leftrightarrow -Y$ gives the opposite-order equality:

$$e^{t(X+Y)} = \exp(\sum_{n=1}^{\infty} \frac{t^n}{n!} X_n) e^{tX}. \tag{22}$$

To proceed, recall the definition of X and Y and (4). One gets

$$X_n = \chi(0, \mathcal{L}_x^{n-1} a_p). \tag{23}$$

We thus obtain (a formal expression):

$$e^{t\chi(e)} = \exp(\chi(0, \sum_{n=1}^{\infty} \frac{t^n}{n!} \mathcal{L}_x^{n-1} a_p)) e^{t\chi(x,0)}.$$

Recalling the first two examples, we get the (still slightly formal) expression for $e^{t\chi(e)}$:

$$e^{t\chi(e)}(y, b_p) = \big(\phi^x_{-t*}(y), \phi^{x*}_t(b_p) - i_{\phi^x_{-t*}(y)} \{\sum_{n=1}^{\infty} \frac{t^n}{n!} \mathcal{L}_x^{n-1} da_p\}\big). \tag{24}$$

See that sum over n can be rewritten as an integral of the power series over t:

$$\sum_{n=1}^{\infty} \frac{t^n}{n!} \mathcal{L}_x^{n-1} = \int_0^t e^{t\mathcal{L}_x} dt. \tag{25}$$

But $e^{t\mathcal{L}_x}$ is nothing but a pullback by a flow ϕ^x_t. We thus get

$$e^{t\chi(e)}(y, b_p) = \big(\phi^x_{-t*}(y), \phi^{x*}_t(b_p) - i_{\phi^x_{-t*}(y)} \int_0^t \{\phi^{x*}_t(da_p)\} dt\big). \tag{26}$$

It is easy to see that this map indeed satisfies the integration condition (19), that is

$$\frac{d}{dt}\Big|_{t=0} e^{t\chi(e)}(y, b_p) = \chi(e)(y, b_p) \equiv e \circ (y, b_p). \tag{27}$$

Now, we expect that $e^{t\chi(e)}$ will be an automorphism of the Dorfman bracket. According to the Section 3, this accounts to the verification of the closedness of the form $\int_0^t \{\phi^{x*}_t(da_p)\} dt$. The exterior differential operator commutes both with integration (imagine it as a differentiation with respect to parameter of the integrand) and with the pullback. Closedness of this form thus follows from $d(da_p) = 0$.

Example 4.1. Let us try to show the integration on the example. Let $M = \mathbb{R}^2(y^1, y^2)$. Let $x \in \mathfrak{X}(\mathbb{R}^2)$ be defined as

$$x = y^1 \partial_2 - y^2 \partial_1.$$

Let $p = 1$. 1-form a_1 is defined as $a_1 = y^1 dy^2$. First, we have to find a flow corresponding to x. This is a standard calculation, giving:

$$\phi^x_t(y^1, y^2) = (y^1 \cos(t) - y^2 \sin(t), y^1 \sin(t) + y^2 \cos(t)).$$

We see that x is a complete vector field, with uniform rotation along $\mathbb{R}^2$ origin as its flow. Now, we are supposed to calculate the pullback of the form $da_p = dy^1 \wedge dy^2$. Pullback of a 2-form just multiplies it by a Jacobian of the map, which is of course $|J| = 1$ in this case (rotation is orthogonal). Thus $\phi_t^{x*}(dy^1 \wedge dy^2) = dy^1 \wedge dy^2$. This also follows from the fact that rotations are symplectomorphisms with respect to the canonical symplectic form da_1 on $\mathbb{R}^2$. We then obtain:

$$I_t(da_1) \equiv \int_0^t \{\phi_t^{x*}(da_1)\} = t \cdot dy^1 \wedge dy^2. \tag{28}$$

Finally, we will calculate the action of $e^{t\chi(e)}$ on the section $(y, b_1) = (y, 0)$, where $y = y^1\partial_1$. Pulling back the vector field y gives

$$\phi^x_{-t*}(y) = \{\cos(t)(y^1\cos(t) - y^2\sin(t))\}\partial_1 + \{\sin(t)(y^2\sin(t) - y^1\cos(t))\}\partial_2.$$

Plugging this into the 2-form (28) gives:

$$-i_{\phi^x_{-t*}(y)} I_t(da_1) = \{t\sin(t)(y^2\sin(t) - y^1\cos(t))\}dy^1 + \{t\cos(t)(y^2\sin(t) - y^1\cos(t))\}dy^2.$$

To conclude this example, see that $\chi(x, a_p)$ acts on $(y, 0)$ as

$$(\chi(x, a_1))(y, 0) = ([x, y], -i_y da_1) = (-y^2\partial_1 - y^1\partial_2, -y^1 dy^2). \tag{29}$$

It is easy to see that there indeed holds

$$\frac{d}{dt}\bigg|_{t=0} (\phi^x_{-t*}(y), -i_{\phi^x_{-t*}(y)} I_t(da_1)) = (-y^2\partial_1 - y^1\partial_2, -y^1 dy^2). \tag{30}$$

5 Example

Let us show the application of the formula (26) in finding a finite transformation corresponding to infinitesimal isometry of generalized metric on E. It is a fiberwise metric $\mathbf{G}$ on vector bundle E, naturally appearing in the Hamiltonian of membrane sigma models. For details, see for example [5]. First recall that given metric g on M, one can define a fiberwise metric $\widetilde{g}$ on the exterior product bundle $\Lambda^p TM$ as

$$\widetilde{g}_{IJ} = \delta_I^{k_1 \dots k_p} g_{k_1 j_1} \dots g_{k_p j_p}. \tag{31}$$

where $I = (i_1 < \dots < j_p)$ and $J = (j_1 < \dots < j_p)$ are strictly ordered p-indices, labeling the local basis of $\Gamma(\Lambda^p TM) \equiv \mathfrak{X}^p(M)$ induced from arbitrary local basis $(e_i)_{i=1}^n$ of $\Gamma(TM) \equiv \mathfrak{X}(M)$ as $e_I = e_{i_1} \wedge \ldots \wedge e_{i_p}$. Not only that $\widetilde{g}$ is a symmetric $C^\infty(M)$-bilinear form on the module $\mathfrak{X}^p(M)$, but it is non-degenerate in a usual sense. Its signature (as a quadratic form) depends only on the signature of g, and for positive definite g, $\widetilde{g}$ is also positive definite.

Writing C as a matrix, we mean a rectangular $n \times \binom{n}{p}$ matrix C_{iJ} defined as $C_{iJ} = C(e_i, e_J)$ for a $(p+1)$-form $C \in \Omega^{p+1}(M)$. Note that the transpose matrix C^T corresponds to the map $x \mapsto i_x C$. A generalized metric $\mathbf{G}$ is defined by a symmetric block matrix:

$$\mathbf{G} = \begin{pmatrix} g + C\widetilde{g}^{-1}C^T & -C\widetilde{g}^{-1} \\ -\widetilde{g}^{-1}C^T & \widetilde{g}^{-1} \end{pmatrix}. \tag{32}$$

Having the fiberwise metric $\mathbf{G}$, we can define a generalized Killing vectors $e \in E$ to be the sections of E satisfying the generalized Killing equation:

$$\rho(e).\mathbf{G}(e', e'') = \mathbf{G}(e \circ e', e'') + \mathbf{G}(e', e \circ e''). \tag{33}$$

Such sections have certain physical significance. They correspond to Noetherian currents conserved in time evolution, for details, see [6]. Thanks to (4) it is easy to prove that given generalized Killing sections $e, e' \in \Gamma(E)$, their higher Dorfman bracket $e \circ e'$ is again a generalized Killing section. It follows that the set of generalized Killing vectors of $\mathbf{G}$ forms a Leibniz algebra. Writing $e = (x, a_p)$, we can extract the content of the equation (33) to get a set of three equations:

$$\mathcal{L}_x g = 0,\ \mathcal{L}_x C = da_p,\ \mathcal{L}_x \widetilde{g} = 0. \tag{34}$$

The last one in fact follows from the first one, although this is quite tedious to show. Interpretation of the first condition is obvious, x is a Killing vector field of g, generating thus an isometry of g. What is a meaning of the second condition?

We can integrate the infinitesimal isometry $\chi(x, a_p)$ to the automorphism of a higher Dorfman bracket using the formula (26). We see that we have to find the $(p+1)$-form $\int_0^t \{\phi_t^{x*}(da_p)\}dt = \int_0^t \{\phi_t^{x*}(\mathcal{L}_x C)\}dt$, where we have used the generalized Killing equation for (x, a_p). Glancing at the original power series expression for the integral, one sees that $\int_0^t \{\phi_t^{x*}(da_p)\}dt = \phi_t^{x*}C - C$. But we know that this is a closed form (see the discussion after (26)). In other words, action of the isometry ϕ_t^x on C only makes a gauge transformation by a closed $(p{+}1)$-form $\int_0^t \{\phi_t^{x*}(da_p)\}dt$. The automorphism corresponding to the map $\chi(e)$ is according to (26):

$$e^{t\chi(e)}(y, b_p) = \big(\phi^x_{-t*}(y), \phi_t^{x*}(b_p) - i_{\phi^x_{-t*}(y)}(\phi_t^{x*}C - C)\big). \tag{35}$$

It is not difficult to verify that $e^{t\chi(e)}$ is moreover an isometry of $\mathbf{G}$, that is there holds

$$\mathbf{G}(e^{t\chi(e)}(e'), e^{t\chi(e)}(e'')) = \mathbf{G}(e', e'') \circ \phi_t^x, \tag{36}$$

for all $e', e'' \in \Gamma(E)$. To conclude, see that one can find the generalized Killing vectors as follows. First find an ordinary Killing vector x of g. Necessary condition for $\mathcal{L}_x C = da_p$ to have some solution is a closedness of the form on the left-hand side. That is there holds

$$0 = d(\mathcal{L}_x C) = \mathcal{L}_x(dC) = \mathcal{L}_x F,$$

where $F = dC$ is a field strength corresponding to C. Generalized Killing equation has the solution for given $x \in \mathfrak{X}(M)$, only if x is an infinitesimal symmetry of F. If we work locally or on a contractible space, we can always find a form a_p, such that $\mathcal{L}_x C = da_p$.

Example 5.1. Let $M = \mathbb{R}^2(y^1, y^2)$, and $g = (dy^1)^2 + (dy^2)^2$ be an Euclidean metric on $\mathbb{R}^2$. Consider $p = 1$, and let $C = (y^1 + y^2)dy^1 \wedge dy^2$. We would like to find all generalized Killing vectors of generalized metric $\mathbf{G}$. Killing algebra of g is generated by translations $t_i = \partial_i$ and a rotation along the origin: $r = y^2\partial_1 - y^1\partial_2$. We may find the potentials for $\mathcal{L}_x C$ separately - the condition is additive. For translations, it is very simple:

$$\mathcal{L}_{t_1} C = \mathcal{L}_{t_2} C = dy^1 \wedge dy^2 = d(y^1 dy^2).$$

The set of possible choices of a_1 for t_i is thus a cohomology class $[y^1dy^2]$. For a rotation generator r, we get

$$\mathcal{L}_rC = (y^2 - y^1)dy^1 \wedge dy^2.$$

We can find a potential for $\mathcal{L}_rC$ easy enough. We can choose

$$a_1 = -\frac{1}{2}\{(y^2)^2dy^1 + (y^1)^2dy^2\}. \tag{37}$$

Set of all such a_1 is again the cohomology class of the above particular solution. The set of all generalized Killing vectors $GK(\mathbf{G})$ can be thus described as

$$\begin{aligned} GK(\mathbf{G}) = \{\big(\alpha_1t_1 + \alpha_2t_2 + \beta r, (\alpha_1 + \alpha_2)y^1dy^2 - \frac{\beta}{2}\{(y^2)^2dy^1 + (y^1)^2dy^2\} + df\big) \\ \mid \alpha_1, \alpha_2, \beta \in \mathbb{R}, f \in C^\infty(M)\}. \end{aligned} \tag{38}$$

We see that Leibniz algebra $GK(\mathbf{G})$ is not finite-dimensional (as is in the ordinary Killing vector algebra), because there is always an infinite-dimensional ambiguity in the condition $\mathcal{L}_xC = da_p$.

References

[1] Y. Bi and Y. Sheng. *On higher analogues of Courant algebroids.* Science in China A: Mathematics **54** (March 2011), 437–447.

[2] F. Casas, A. Murua, and M. Nadinic. *Efficient computation of the Zassenhaus formula.* Computer Physics Communications **183** (November 2012), 2386–2391.

[3] T. Courant. *Dirac manifolds.* Trans. Amer. Math. Soc. **319** (1990), 631–661.

[4] Y. Hagiwara. *Nambu-Dirac manifolds.* J. Phys. A **35** (2002), 1263.

[5] B. Jurčo, P. Schupp, and J. Vysoký. *Extended generalized geometry and a DBI-type effective action for branes ending on branes.* JHEP **1408** (2014), 170.

[6] B. Jurčo, P. Schupp, and J. Vysoký. *p-Brane Actions and Higher Roytenberg Brackets.* JHEP **1302** (2013), 042.

[7] K. C. Mackenzie. *General theory of Lie groupoids and Lie algebroids*, volume 213 of *London Mathematical Society Lecture Note Series.* Cambridge University Press, Cambridge, (2005).

[8] D. Roytenberg. *Courant algebroids, derived brackets and even symplectic supermanifolds.* ArXiv Mathematics e-prints (October 1999).

Path-Integral Approach to the Wigner–Kirkwood Expansion*

Václav Zatloukal

3rd year of PGS, email: zatlovac@gmail.com
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Jizba, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The Wigner–Kirkwood (WK) expansion was originally presented in two seminal papers [2, 3] and since its very inception it has had two important implications. On the one hand, it has been used for studying the equilibrium statistical mechanics of a nearly classical system of particles obeying Maxwell–Boltzmann statistics. WK expansion is in its essence an expansion of the quantum Boltzmann density in powers of Planck's constant $\hbar$, or equivalently of the thermal de Broglie wavelength $\lambda = \hbar\sqrt{\beta/M}$, where β is the inverse temperature and M is the mass of a particle. On the other hand, it has paved a way for new alternative mathematical techniques and practical calculational schemes that are pertinent to the high-temperature regime in quantum systems.

In this paper, we pursue the study of the WK perturbation method by means of the path integral (PI) calculus. The relevance of the PI treatment in a high-temperature context is due to several reasons: PI's allow to connect evolutionary equations (Bloch equation or Fokker–Planck equation) with the underlying stochastic analysis [4], they are tailor-made for obtaining quasi-classical asymptotics [5], they allow to utilize some powerful transformation techniques to simplify the original stochastic process, etc. Besides, PI's also provide an excellent tool for direct numerical simulations of the underlying stochastic dynamics including many-body systems. One of the key advantages of the PI approach is, however, the fact that the techniques and methodologies used can efficiently bypass the explicit knowledge of the exact energy spectrum. In particular, one can progress without relying on the explicit use of approximate expressions or interpolation formulas for the energy eigenvalues which are often difficult to judge due to lack of reliability in their error estimates.

The idea to use PI's as a means of producing various WK-type expansions and related thermodynamic functions is clearly not new. Indeed, the first systematic discussions and analyzes of these issues emerged already during the early 1970's. Among these belong the early attempts of PI treatments of the high-temperature behavior of partition functions for anharmonic oscillators and gradient expansions of free energy [5]. These approaches belong in the class of the so-called *analytic perturbation schemes* which account for an explicit analytic expressions of the coefficient functions. For many practical purposes it is desirable to have explicit analytical expressions for coefficients in the WK perturbation expansion. This is so, for instance, when the symmetry (Lorentz, gauge, global) is supposed to be broken by quantum or thermal fluctuations. Though these issues are more pressing in quantum field theories, they have in recent two decades entered also in the realm of a few-body finite-temperature quantum mechanics. The catalyst has been theoretical investigations and ensuing state-of-the-arts experiments in condensed Bose gases,

[*] This work has been published in [1].

degenerate Fermi gases, quantum clusters or strongly coupled Coulomb systems. It is not only the zero-temperature regime that is of interest in these systems. Many issues revolve also around finite-temperature or “high”-temperature questions. These include, thermal and thermoelectric transport of ultra-cold atomic gases, hydrogen, helium, and hydrogen/helium mixtures and their astrophysical implications, Lennard–Jones ^{3}He and ^{4}He gases, etc.

A serious weakness of existent analytic WK expansions and their various disguises (be they based on PI's or not), resides in their inability to progress very far with the expansion order. This makes it difficult to address thermodynamically relevant intermediate-temperature regions that is particularly pertinent in molecular and condensed matter chemistry (binding energies, self-dissociation phenomena, order-disorder transitions, etc.). The best analytic expansions are presently available within the framework of the world-line path integral method (known also as the string inspired method) [6]. In this approach the expansion coefficients are available up to order $\mathcal{O}(\beta^{12})$, subject to the actual interaction potential. Other more conventional approaches, such as the recursive or non-recursive heat-kernel calculations or higher derivative expansions by Feynman diagrams, achieve at best the order $\mathcal{O}(\beta^7)$. The key problem is a rapid escalation in the complexity of higher-order terms which is difficult to handle without some type of resummation. In the present paper we derive a new resummation formula that provides a rather simple and systematic way of deriving the coefficient functions. Its main advantages rely on both an analytic control of the high-temperature behavior, and on an accurate description over a wide temperature range via numerical calculations that can be simply carried out at the level of an undergraduate exercise.

The structure of the paper is as follows. To set the stage we recall some fundamentals of PI formulations of the Bloch density matrix and the ensuing partition function and Boltzmann density. With the help of the space-time transformation that transforms the Wiener-process sample paths to the Brownian-bridge sample paths we obtain the PI that represents a useful alternative to the original Feynman–Kac representation. Consequently we arrive at a new functional representation of the Boltzmann density which is more suitable for tackling the high-temperature regime than the genuine Wigner–Kirkwood formulation. While the method resembles in principle the Wentzel–Brillouin–Kramers (WKB) solution for the transition amplitude, its details are quite different. In two associated subsections we examine some salient technical issues related to the low-order high-temperature expansion in one dimension. To illustrate the potency of our approach we consider the high-temperature expansion of the one-dimensional anharmonic oscillator. In particular, we perform the Boltzmann density and ensuing partition function expansions and compute the related thermodynamic quantities. The expansions obtained improve over the classic results of Schwarz [7] and Padé-approximation-based expansion of Gibson [8]. We proceed by extending our expansion to the whole Bloch density matrix. The expansion thus obtained is compared with the more conventional Wick's theorem based perturbation expansion based on the Onofri–Zuk Green's functions. There we show that our prescription comprises substantially less (in fact, exponentially less) terms contributing to higher perturbation orders. Also the algebraic complexity of the coefficient functions involved is substantially lower in our approach. The paper is accompanied by Mathematica code that generates the higher-order expansion terms for arbitrary smooth local potentials up to 18th order in β.

Let us add a final note. Most of the presented mathematical derivations are of a heuristic nature — as it should be expected from the mathematical analysis based on the path-integral calculus. The basic purpose of this paper is to find explicit formulas for the coefficient functions, and in doing so to reveal the elaborate algebraic and combinatorial structure present in these functions. A more rigorous treatment of the aforementioned mathematical aspects is possible, but would involve different language and techniques than are employed in this paper.

Keywords: path integral, Gibbs operator, high-temperature expansion, Wigner-Kirkwood expansion

References

[1] P. Jizba and V. Zatloukal, *Path-integral approach to the Wigner-Kirkwood expansion*, Phys. Rev. E **89** (2014), 012135; arXiv:1309.0206 [cond-mat.stat-mech].

[2] E. Wigner, Phys. Rev. **40**, 749 (1932).

[3] J.G. Kirkwood, Phys. Rev. **44**, 31 (1933).

[4] Z. Haba, *Feynman Integral and Random Dynamics in Quantum Physics; A Probabilistic Approach to Quantum Dynamics*, (Kluwer, London, 1999).

[5] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets*, 5-th edition, (World Scientific, London, 2009).

[6] C. Schubert, Physics Report **355**, 73 (2001).

[7] M. Schwartz,Jr., J. Stat. Phys. **15**, 255 (1976).

[8] W.G. Gibson, J. Phys. A: Math. Gen. **17**, 1891 (1984).

Micro-Scale Modeling of Soil Freezing*

Alexandr Žák

3rd year of PGS, email: alexandr.zak@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Beneš, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This contribution deals with modeling of mechanical manifestations within saturated soil, which are induced by freezing of the pore water. A simple 2D mechanical model of the phase transition in a portion of a pore is presented. This model is based on the Navier equations and on the continuity equation and serve mainly for a verification of the dynamics of the mechanical reaction. A basic qualitative computational study of this model is presented. Further, this model is generalized by supplementing it with a heat balance law onto the thermo-mechanical model describing the mutual interaction of all pore components. Thus, the extended model enables more realistic description of the studied dynamics in a more general part of the soil pore material. For this model, some basic qualitative studies, which indicate non-trivial progress of the mechanical interaction, are presented as well.

Keywords: freezing, model, phase-transition, soil, micro-scale

Abstrakt. V příspěvku se zabýváme modelováním mechanických projevů v saturované půdě, které jsou vyvolány zamrzáním vody v pórech. Je zde představen jednoduchý 2D mechanický model fázové změny části saturovaného póru založený na Navierových rovnicích a rovnici kontinuity. Tento model slouží především pro ověření dynamiky mechanické reakce. Je zde představena základní kvalitativní výpočetní studie tohoto modelu. Dále je tento model zobecněn, přidáním tepelného bilančního zákona, na termomechanický model popisující již vzájemnou interakci všech složek porézního materiálu. Rozšířený model tak umožňuje reálnější popis zkoumané dynamiky na obecnější části půdního porézního prostředí. Také pro tento model jsou představeny základní kvalitativní studije naznačující netriviální průběh mechanické interakce.

Klíčová slova: zamrzání, model, fázová změna, půda, mikro škála

1 Introduction

In regions suffering from freezing seasons or the climate change, upper layers of soil ground exhibit structural changes due to the phase transition of a wet component of soil. Therefore these phenomena introduce an unceretaint into designs of building structures in the cold regions or an ambiguity into ecological problems associated with impacts of the climate change.

*Partial support of the project of the “Development and Validation of Porous Media Fluid Dynamics and Phase Transitions Models for Subsurface Environmental Application, project of Czech Ministry of Education, Youth and Sports Kontakt II LH14003, 2014-2016” and of the project “Advanced supercomputing methods for mathematical modeling of natural processes, project of the Student Grant Agency of the Czech Technical University in Prague No. SGS14/206/OHK4/3T/14 2014-16”.

Although there are several macro-scale models of one of the most significant freezing phenomena ([2], [3], [4]), the frost heave, they are not sufficiently general and complex, or are one-sidedly oriented, or are based on some simplified assumptions. One of the reasons for such state of complex understanding of the freezing soil problem is that there is a little studies, experimental or theoretical, concerning with the behavior of the phenomenon on the pore-scale level. Therefore one of the aims of this work is to improve the understanding of the impacts of soil freezing on such level.

This work freely follows the preliminary models of soil freezing described in [1]. In this contribution, firstly, we present a 2D micro-scale continuum model capturing the basic mechanical response within a pore during freezing of its water component. This model helps to reveal the dynamics of this response and serve generally for the study and verification of the induced structural changes. Secondly, we further generalize this model in terms of thermo-mechanical coupling and in terms of a general geometric scenario. This model is capable of providing the dynamics of the freezing processes within an ideal pore region. Computational studies for the both stages of model are also presented.

2 Artificially driven phase and structural model

To get a basic insight into the structural dynamics of freezing water in a nontrivial shaped domain, we have designed a simple two-dimensional mechanical model capturing both the solidification of water in terms of a local change of the stress tensor of water and the structural change induced by the solidification. As the model has been intended to serve purposes of study of the issue, the both changes are driven artificially by a step function Υ, $\Upsilon = \Upsilon(t, x, y)$, where its arguments stand for the temporal and spatial coordinates, respectively.

2.1 Mathematical model

The default mathematical description for this model is the homogeneous isotropic elastic model involving the Navier equations for the displacement vector $\mathbf{u}$ in two dimensions and the mass conservation equation. The latter equation provides an additional relation for the pressure p as for the another dependent variable. Let $\Omega_{\triangleright}$ denote the considered domain, then our system of equation reads

$$\varrho_l \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot \sigma \ \text{ in } \Omega_{\triangleright} , \qquad \frac{p}{\varrho_l E_l} + \nabla \cdot \mathbf{u} = 0 \ \text{ in } \Omega_{\triangleright} ,$$

where ϱ_l stands for the density of water, σ is the stress tensor, and E_l is Young's modulus of water.

In order to be able to capture the forementioned changes, the default description is altered in terms of the modification of the water stress tensor, σ. It is expressed as a temporal-spatial dependent tensor field controlled only by use of functions Υ in the following way

$$\sigma = \sigma(t, x, y) = \Upsilon(t, x, y)\sigma_i + (1 - \Upsilon(t, x, y))\sigma_l , \tag{1}$$

where σ_l stands for the stress tensor for the Newtonian fluid,

$$\sigma_l = -p\mathbb{I} + \mu \left(\nabla \dot{\mathbf{u}} + (\nabla \dot{\mathbf{u}})^{\mathbb{T}} \right) ,$$

and σ_i stands for the stress tensor of an isotropic linear elastic material extended with terms introducing the structural change during the solidification of water,

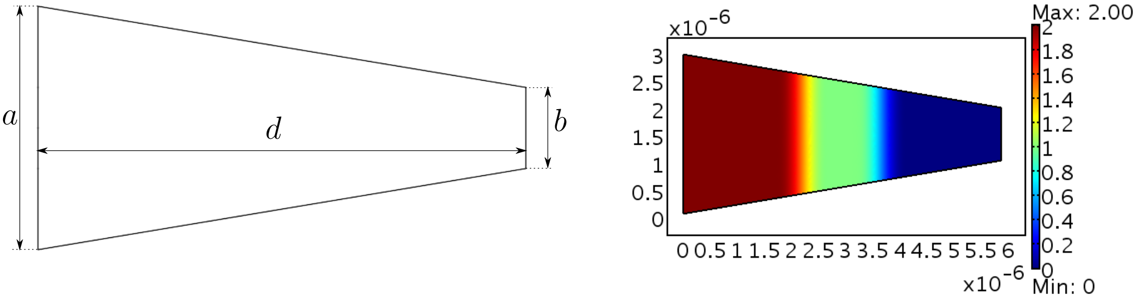
$$\sigma_i = \frac{E_i}{2(1+\nu_i)}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathbb{T}}\right) + \frac{\nu_i E_i \nabla\cdot\mathbf{u}}{(1+\nu_i)(1-2\nu_i)}\mathbb{I} \quad + \frac{\beta_i E_i}{1-2\nu_i}\hat{\Upsilon}\,,$$

where μ is the (dynamic) viscosity of water, E_i and η_i stand for Young's modulus and Poisson's ratio of ice, respectively, and β_i represents the inner stress ratio.

The system of equation is supplemented with the boundary and initial conditions. Their particular form follows from the nature of the shape of a pore and is discussed in following subsection.

2.2 Geometry

When interested in applications of induced forces in soils as in porous media, it is useful to consider the domain of study as a somehow idealized part of a pore area. A convenient shape can thus be an isosceles trapezoid representing the cross-section of a simplified meniscus between two soil grains.



(a) A representative domain $\Omega_{\triangleright}$ ($a = 3e{-}6m$, $b = 1e{-}6m, d = 6e{-}6m$)

(b) A plot of (mollified) $\Upsilon + \hat{\Upsilon}$ at $t = 8s$.

Figure 1

In our study, we use this shape of domain (see Figure 1a) and consider that narrower of the parallel sides represents the contact point of two grains, so points on this side are fixed in the horizontal direction, and that wider side experiences a symmetrical force influence, so points on this side are also fixed in this direction. The remaining sides are considered to be subjected loosely to the stress conditions within domain. Rewriting these conditions on the boundaries in terms of the displacement vector $\mathbf{u}$, we have

$$\mathbf{u}_{(x)} = 0 \ \text{, on the parallel parts of boundary } \partial\Omega_{\triangleright}\,,$$
$$\sigma\cdot\vec{n} = 0 \ \text{, on the remaining parts of boundary } \partial\Omega_{\triangleright}\,.$$

The initial conditions are set to be constant; no initial displacement is assumed and $p_{ini} = 0$.

To simulate the natural process of gradual propagation of ice into a meniscus, we prescribe functions Υ as follows

$$\Upsilon(t,x,y) = \vartheta(v(t-t_1)-x)\,, \qquad \hat{\Upsilon}(t,x,y) = \vartheta(v(t-t_1)-x-h)\,, \tag{2}$$

where v is the artificial velocity of the propagation, t_1 is a delay, and h is the distance between the steps of both functions. The parameter h has the meaning of a test tool for the distinguishing the effects of the phase transition of water and of its structural response. For simulation of the real process of freezing, the simultaneous action of the both effect is assumed, i.e $h = 0$.

2.3 Simulations

To avoid the convergence difficulties, we use a smoothed form (with smoothing parameter ε) of the step functions. An illustration of such a regularization is shown in Figure 1b. For further simplicity, all physical parameters are kept constant and their exact values are shown in Table 2.

Solutions are obtained by use of the FEM method in combination with the BDF solver. Solving process is controlled by an error condition; the step is accepted if the following inequality holds

$$\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|e_i|}{A_i + R|e_i|}\right)^2\right)^{\frac{1}{2}} < 1\,, \tag{3}$$

where u is the solution vector, e is the solver's estimate of the local error, A_i is the absolute tolerance for degree of freedom i, R is the relative tolerance, and N is the number of degree of freedom.

3 Thermally driven phase and structural model

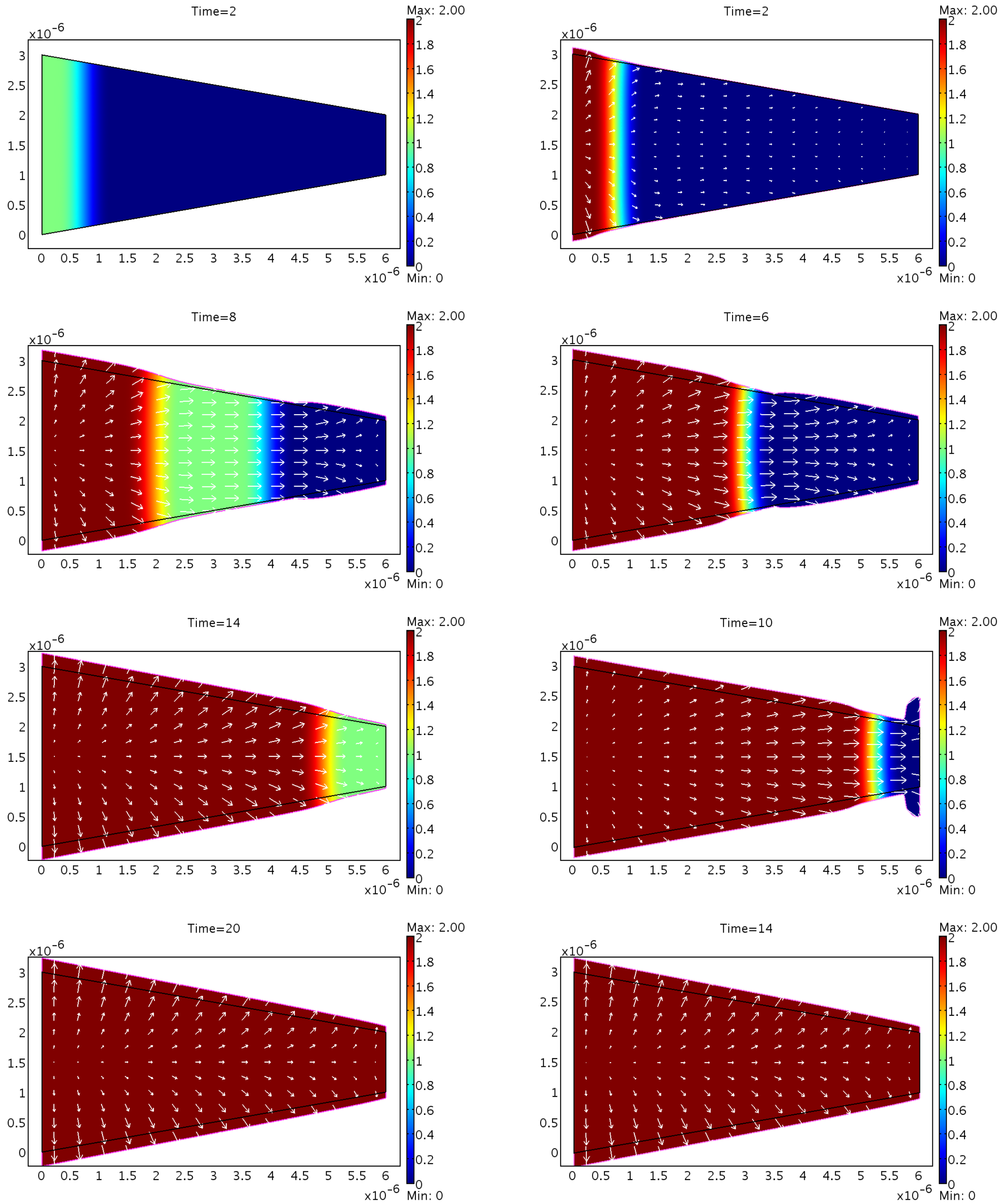
Been focusing on a more realistic problem, we have generalized the previous model in several aspects. We have included the adjacent domains into our considerations, but still the smallest appropriate region of a soil pore structure is considered. We have passed to the temperature, T, as to the natural initiator of phase change and made physical quantities dependent on the current phase.

3.1 Geometry of the problem

Problem scale of the interest is such that dimensions of pores are not negligible with regard to the dimensions of the considered pore region. Therefore, phases contained in soil are clearly distinguished and occupy separated subdomains. Thus generally, a representative domain of freezing saturated soil, Ω, consists of subdomains for liquid water, ice, and skeleton and of all their mutual boundaries. The domain illustration and the particular notation of all domain parts are shown in Figure 3a.

3.2 Mathematical model

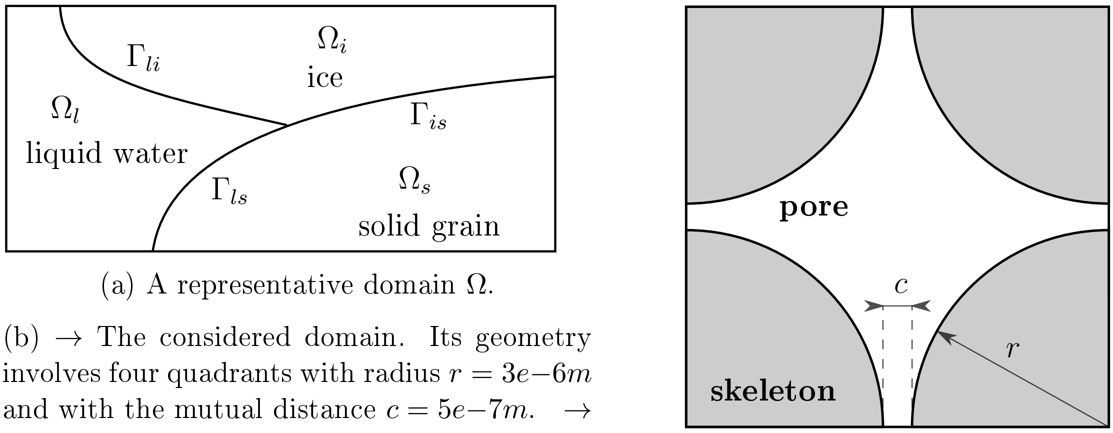
To cover the thermo-mechanical interactions within saturated soils during their cooling (and warming) on pore-scale (micro-scale) level, the previous model for the water domain has been extended with the modified heat equation, capturing the phase transition



(a) Here the inception of the phase change and that of the structural change are shifted, $h = 2e{-}6m$.

(b) Here the inception of the phase change and that of the structural change occur at the same time, $h = 0m$.

Figure 2: Simulations of the phase change propagation in the simplified meniscus. The color signifies the values of function $\Upsilon + \hat{\Upsilon}$. The arrows stand for the displacement, and the displacement of the domain is 10 times scaled.



(a) A representative domain Ω.

(b) → The considered domain. Its geometry involves four quadrants with radius $r = 3e{-}6m$ and with the mutual distance $c = 5e{-}7m$. →

Figure 3

of water and representing a heat balance relation, and with the pair of corresponding equations for the skeleton domain as well.

Further some conditions must be also prescribed on the mutual boundaries between the domains in order to the assembly of the equation could be performed. Therefore we assume that the temperature and the displacement vector are continuous on the inner boundaries, that the heat fluxes are continuous over the boundaries along the grain surfaces, and that the momentum is balanced on these boundaries.

Functions Υ initiating the changes during the phase transition are now naturally defined as functions, which are explicitly dependent on the tempriture; in particular, they are defined in following way: $\Upsilon(T) = \hat{\Upsilon}(T) = \vartheta(T - T_\star)$, where $T_\star$ stands for the local freezing point depression. This value can be obtained from equilibrium condition

$$\varrho_i l \frac{T_\star - T_0}{T_\star} = \gamma\kappa\,, \tag{4}$$

which generally holds on a curved equilibrium interface between liquid and solid phases and where T_0 is the bulk freezing point of water, γ denotes the surface tension, and κ stands for the curvature of the interface.

Under these assumptions, the governing system for the micro-scale model of the phase and structural change within a freezing saturated soil reads:

$$\begin{array}{rr}
\varrho c \dfrac{\partial T}{\partial t} + \varrho l \dfrac{\partial \Upsilon(T)}{\partial t} = \nabla \cdot (k \nabla T) \ \text{ in } \Omega_l \cup \Omega_i \cup \Gamma_{li}\,, & \varrho_s c_s \dfrac{\partial T}{\partial t} = \nabla \cdot (k_s \nabla T) \ \text{ in } \Omega_s\,, \\
\varrho \dfrac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot \sigma \ \text{ in } \Omega_l \cup \Omega_i \cup \Gamma_{li}\,, & \varrho_s \dfrac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot \sigma_s \ \text{ in } \Omega_s\,, \\
\dfrac{p}{\varrho_l E_l} + \nabla \cdot \mathbf{u} = 0 \ \text{ in } \Omega_l\,, &
\end{array}$$

where ϱ is the effective value of the density of water, c is the effective value of the volumetric heat capacity of water, l stands for the volumetric latent heat of water, k is the effective value of the thermal capacity of water, σ is the effective form of the stress tensor, and subscript s signifies the analogous quantities of the skeleton, and where the effective values of the water properties are taken as the convex combinations of the effective values of the corresponding quantity of each pore component; i. e.

$$\varrho = \Upsilon(T)\varrho_i + (1 - \Upsilon(T))\varrho_l\,, \qquad c = \Upsilon(T)c_i + (1 - \Upsilon(T))c_l\,,$$
$$k = \Upsilon(T)k_i + (1 - \Upsilon(T))k_l\,, \qquad \sigma = \Upsilon(T)\sigma_i + (1 - \Upsilon(T))\sigma_l\,,$$

where subscripts i and l signify the quantities of ice and water, respectively.

3.3 Simulations

The following scenario has been designed to provide a basic information on the interaction between the freezing pore water content and the surrounding (uncemented) solid skeleton. The problem scenario considers a vertical cross-section through a small region of saturated soil with an ideal geometry but with the real physical dimensions and properties. The geometry comprises a group of four untouching quadrants, which represents the skeleton grains, and the remaining region, which stands for the pore filled with water. The particular geometry with all sizes is illustrated in Fig. 3b.

The outer boundary conditions for this scenario have been provided in the following manner: Heat flux q has been prescribed on the top boundary; the remaining boundaries have been assumed as thermally isolated; movement of the grain sides points has been allowed only along the geometry sides, and the free condition has been set for displacements of the outer points of the pore domain.

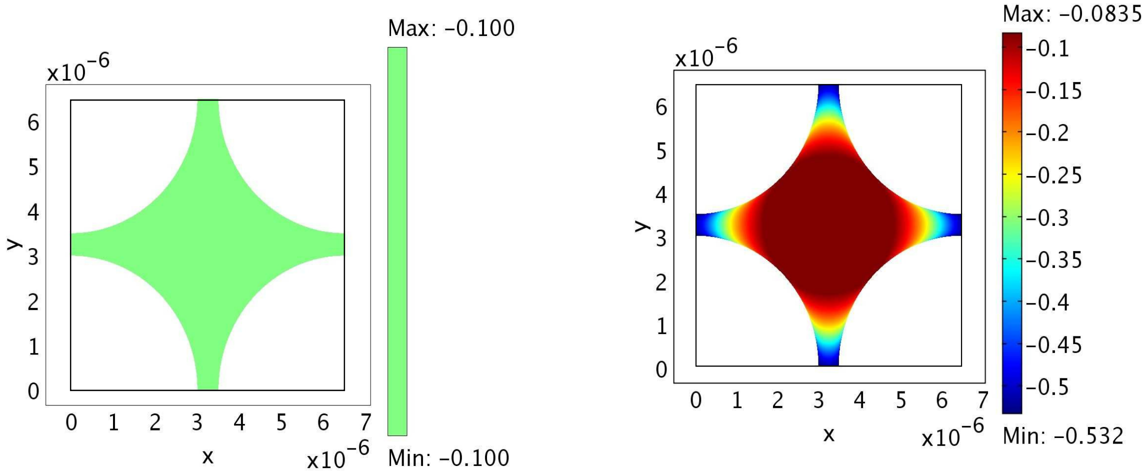
$A_i^{(T)}$	$1e{-}7[1]$	$A_i^{(u)}$	$1e{-}12[1]$	β_i	$1.3044e8[1]$
c_i	$2.1e3[J \cdot kg^{-1} \cdot k^{-1}]$	c_l	$4.2e3[J \cdot kg^{-1} \cdot k^{-1}]$	c_s	$1e3[J \cdot kg^{-1} \cdot k^{-1}]$
E_i	$7.8e9[Pa]$	E_l	$5.33e9[Pa]$	E_s	$7.5e10[Pa]$
γ	$7.5e{-}2[Pa \cdot m^{-1}]$	k_i	$2.18[W \cdot K^{-1} \cdot m^{-1}]$	k_l	$0.6[W \cdot K^{-1} \cdot m^{-1}]$
k_s	$2[W \cdot K^{-1} \cdot m^{-1}]$	l	$3.34e5[J \cdot kg^{-1}]$	μ	$1.8e2[Pa \cdot s]$
ν_i	$0.33[1]$	ν_s	$0.3[1]$	R	$1e{-}2[1]$
ϱ_i	$9.2e2[kg \cdot m^{-3}]$	ϱ_l	$1e3[kg \cdot m^{-3}]$	ϱ_s	$2.5e3[kg \cdot m^{-3}]$

Table 1: Values used in the simulations - the thermally driven model.

To stress the importance of the geometry effect, two simulations have been run. One under the assumption of a constant freezing point of the water in the pore and another under the assumption of the spatially dependent freezing point distribution induced by the equilibrium condition (4). The maps of freezing points are shown in Figure 4a and Figure 4b, respectively. The simulation results are then shown in Figure 5 and Figure 6.

4 Conclusions

The presented micro-scale model includes a basic heat and force balance and has been designed for the purpose of a study of structural change dynamics within saturated soils caused by the phase transition of the water content. Simulations so far provided by



(a) The constant freezing point distribution.

(b) The freezing point distribution induced by the equilibrium condition.

Figure 4

A_i	$1e{-}12[1]$	β_i	$1.3e{-}2[1]$
E_l	$5.33e{-}9[Pa]$	ε	$5e{-}7[1]$
μ	$1.8e3[Pa \cdot s]$	ν_i	$0.33[1]$
R	$1e{-}2[1]$	ϱ_l	$1e3[kg \cdot m^{-3}]$
v	$5e{-}7[m \cdot s^{-1}]$	t_1	$0.5[s]$

Table 2: Values used in the simulations - the artificially driven model.

the model indicate non-trivial progress of the thermo-mechanical interaction, but for a general conclusion wider testing will be needed. Results obtained from existing and future studies on this level are planned to be used for upscaling the relevant information into our macro-scale model ([1]).

References

[1] A. Žák and M. Beneš and T. H. Illangasekare, *Analysis of Model of Soil Freezing and Thawing*, IAENG International Journal of Applied Mathematics, Volume 43 Issue 3, Pages 127-134, Sep. 2013.

[2] R. R. Gilpin, *A Model for the Prediction of Ice Lensing and Frost Heave in Soils*, Water Resources Research, Vol. 16, No. 5, pp. 918–930, 1980.

[3] A. C. Fowler, *Secondary Frost Heave in Freezing Soils*, SIAM J. APPL. Math., Vol. 49, No. 4, pp. 991–1008, 1989.

[4] R. L. Michalowski, *A Constitutive Model of Saturated Soils for Frost Heave Simulations*, Cold Region Science and Technology, Vol. 22, Is. 1, pp. 47–63, 1993.

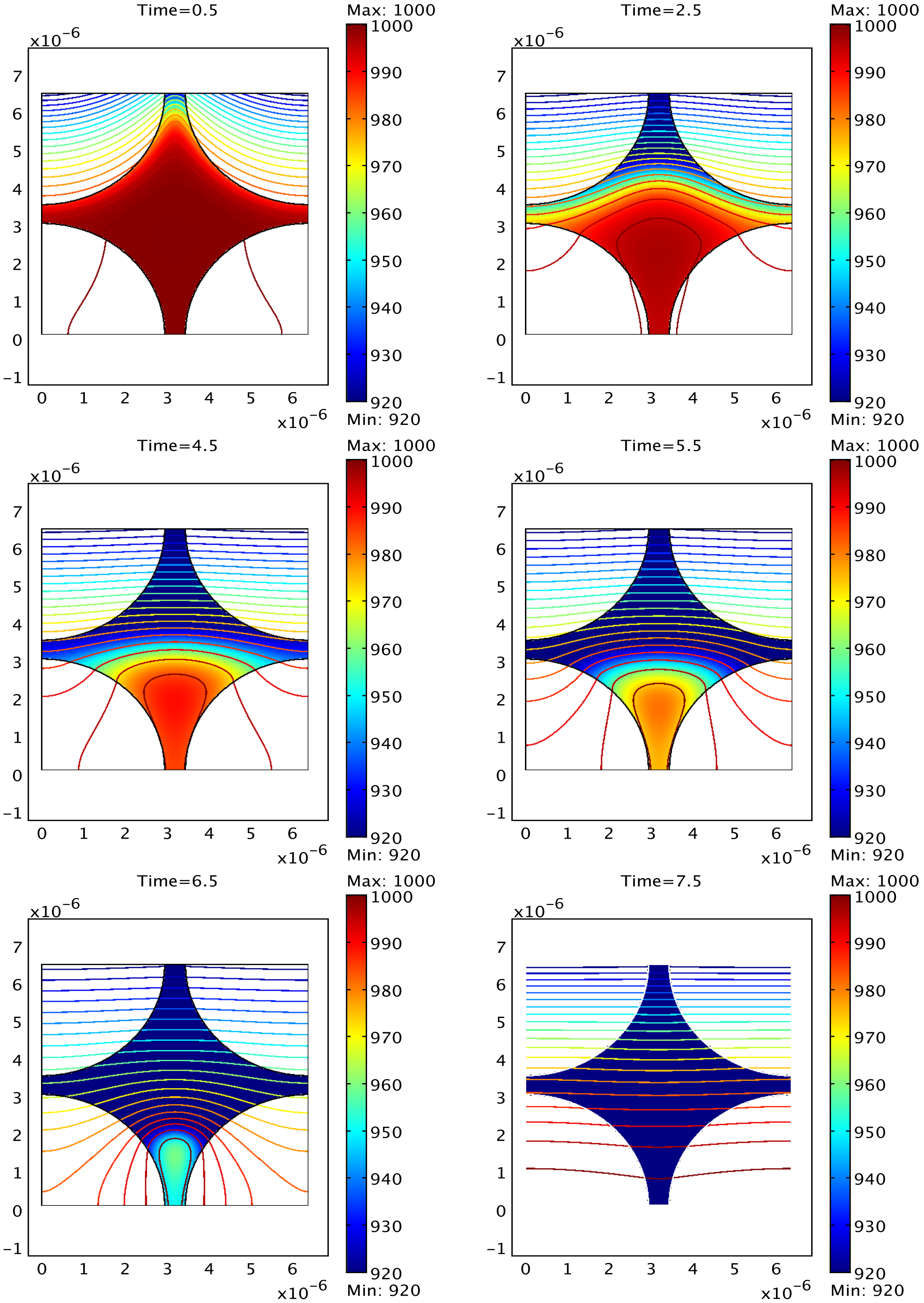


Figure 5: The simulated dynamics of freezing of the considered domain. The freezing point is constant as in Figure 4a. The color stands for the density; the isolines signify 20 current uniformly distributed isotherms - their color legend is not shown.

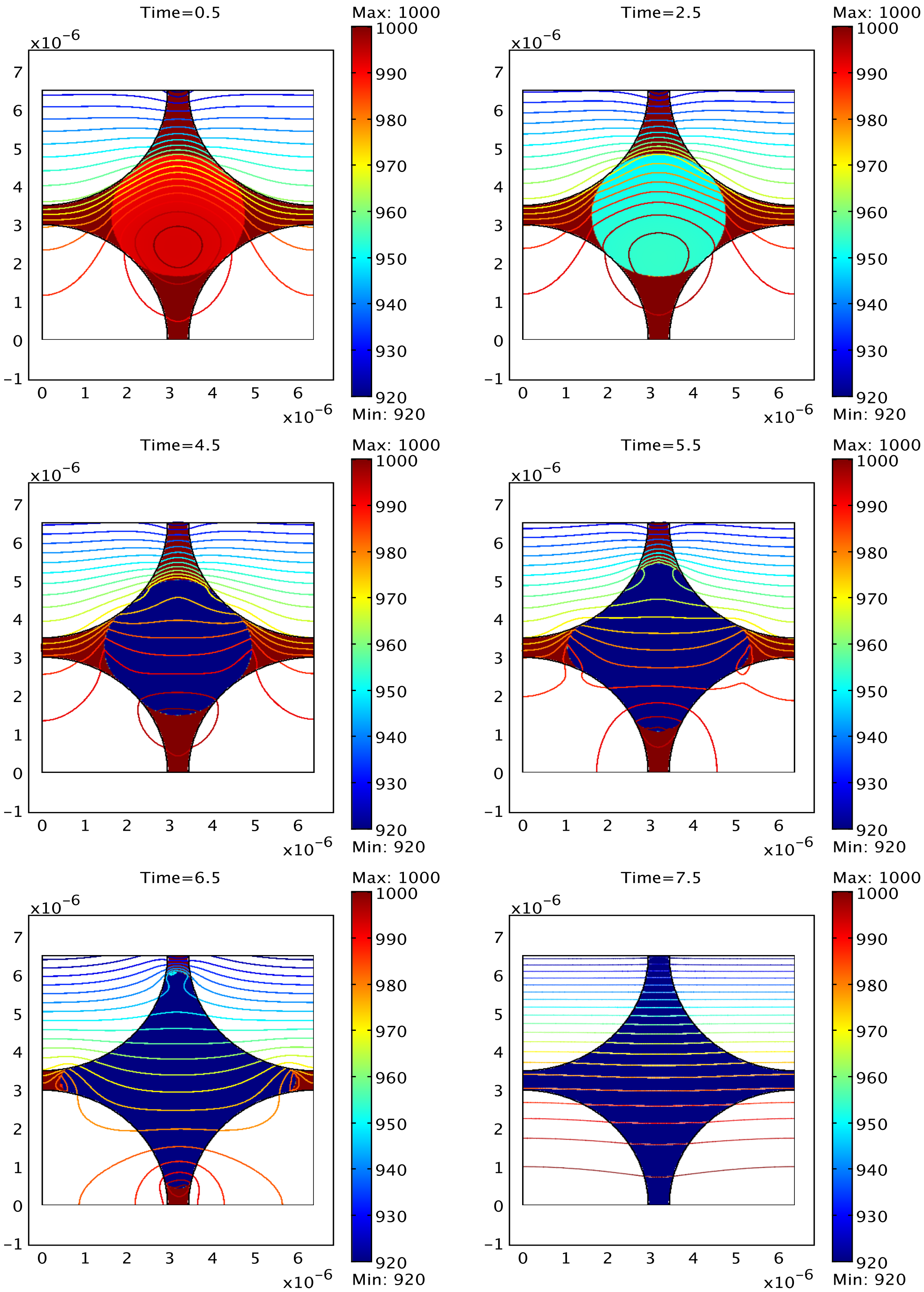


Figure 6: The simulated dynamics of freezing of the considered domain. The freezing point is distributed as in Figure 4b. The color stands for the density; the isolines signify 20 current uniformly distributed isotherms - their color legend is not shown.