DOKTORANDSKÉ DNY 2013

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

15. a 22. listopadu 2013

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2013
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 340, Vydání 1.

ISBN 978-80-01-05379-9

Seznam příspěvků

Předmluva

Již osmý ročník workshopu Doktorandské dny se koná ve dnech 15. a 22. listopadu 2013 na katedře matematiky Fakulty jaderné a fyzikálně inženýrské ČVUT v Praze. Na této konferenci, organizované s finanční podporou Studentské grantové soutěže ČVUT, se každoročně představují doktorandi oboru Matematické inženýrství s příspěvky pokrývajícími širokou škálu témat. Jedná se zejména o deterministické a stochastické modely fyzikálních, medicínských a ekonomických procesů, tvorbu a analýzu výpočetních algoritmů, ale i o témata základního výzkumu v teoretické informatice a matematické fyzice.

Možnost prezentace před odborným publikem nejen z řad školitelů a členů Oborové rady je pro naše doktorandy neocenitelnou zkušeností, která je připravuje k účasti na mezinárodních konferencích. Tento sborník je souborem příspěvků, který průběžně dokumentuje práci doktorandů a slouží jako podklad pro hodnocení studia.

Děkujeme všem, kteří se na zdárném průběhu této akce podílejí.

Editoři

Stability of Point Spectrum for Three-state Quantum Walks on a Line*

Iva Bezděková†

2nd year of PGS, email: bezdekova.iva@gmail.com
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Igor Jex, Department of Physics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Evolution operators of certain quantum walks posses, apart from the continuous part, also point spectrum. The existence of eigenvalues and the corresponding stationary states lead to partial trapping of the walker in the vicinity of the origin. This feature was found in the three-state walk on a line with the Grover coin operator [1],[2], where the evolution operator has one eigenvalue equal to unity. Similarly, Grover walk on a square lattice also has a point spectrum [3]. We analyze the stability of this feature for three-state quantum walks on a line subject to homogenous coin deformations. We find two classes of coin operators that preserve the point spectrum. These new classes of coins are generalization of coins found previously by different methods [4] and shed light on the rich spectrum of coins that can drive discrete-time quantum walks.

Keywords: quantum walk, localization

Abstrakt. U jistých typů kvantových procházek může mít evoluční operátor, mimo spojitého spektra, také spektrum bodové. Existence vlastních hodnot a příslušných stacionárních stavů vede k částečnému uvěznění chodce v okolí počátku. Tato vlastnost byla nalezena pro Groverovu procházku o třech možných stavech posunu. Operátor časového vývoje zde má vlastní hodnotu rovnou jedné. Podobně i Groverova procházka na čtvercové síti má bodové spektrum. V naší práci analyzujeme stabilitu této vlastnosti vzhledem k homogenním deformacím mince. Zabýváme se přitom kvantovou procházkou na přímce o třech možných stavech. Výsledkem je nalezení dvou tříd mincí, které zachovávají bodové spektrum. Tyto nové třídy jsou zobecněním předchozích výsledků, které však byly nalezeny jinými metodami. Práce vrhá světlo na široké spektrum mincí, které mohou řídit diskrétní kvantovou procházku.

Klíčová slova: kvantová procházka, lokalizace

References

[1] N. Inui, N. Konno and E. Segawa. *One-dimensional three-state quantum walk.* Phys. Rew. E **72** (2005), 056112.

[2] N. Inui and N. Konno. *Localization of multi-state quantum walk in one dimension.* Physica A **353** (2005), 133.

*The full article can be found on the website http://arxiv.org/abs/1309.7835.

†This work has been done with co-authors Martin Štefaňák, Igor Jex and Stephen M. Barnett.

[3] N. Inui Y. Konishi and N. Konno. *Localization of two-dimensional quantum walks.* Phys. Rev. A **69** (2004), 052323.

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

COMPASS Database Upgrade Proposal

Martin Bodlák

1st year of PGS, email: martin.bodlak@cern.ch
Department of Mathematics
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 focuses on a proposal of the new online database structure for the COMPASS experiment at CERN. Several incidents that happened during the 2012 COMPASS physics run indicated that due to lack of hardware during the development of the current structure the system is not safe in case of a critical failure of one of its components. However, as the new FPGA-based data acquisition system for the COMPASS experiment is currently being developed, there is a possibility to use some of the computers from the old DAQ architecture for different purposes.

Keywords: CERN, COMPASS, database, MySQL

Abstrakt. Tento článek se zabývá návrhem nové architektury online databáze pro experiment COMPASS v CERN. Během fyzikálního programu experimentu COMPASS v roce 2012 došlo k několika incidentům, které poukázaly na to, že kvůli nedostatku dodaných hardwarových komponent během implementace současné databázové architektury není celý systém bezpečný v případě výpadku jednoho z uzlů. Avšak díky vývoji nového systému pro sběr dat, který počítá s využitím FPGA karet, bude možné uvolnit některé počítače pro jiné účely.

Klíčová slova: CERN, COMPASS, databáze, MySQL

1 Introduction

Modern particle physics experiments produce data in quantities never seen before. This poses very strong requirements on the quality of data acquisition systems (both hardware and software). A critial part of every data acquisition system is the online database. The online database at the COMPASS experiment [1] uses the MySQL relational database management system [2]. It contains meta-information about the run of the experiment. These meta-information include beam parameters, detector configuration, software logs, and additional information recorded by the shift crews during the run of the experiment.

Information from this database are needed to be quickly retrieved during the data acquisition and data analysis. This means that all the machines connected to the COMPASS inner network should be able to write to the database and read from it at any time. To ensure this, the database structure should withstand failures of its components without limiting the access of clients.

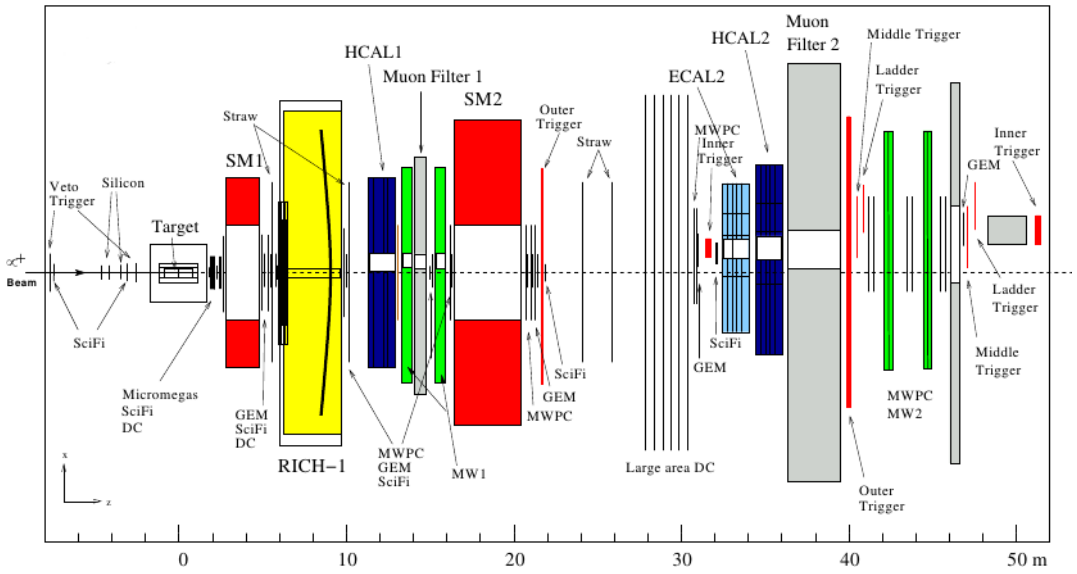


Figure 1: COMPASS spectrometer

2 CERN and COMPASS experiment

The European Organization for Nuclear Research (CERN) is an international scientific organization situated in the northwerst suburbs of Geneva, Switzerland. It was founded in 1954 by 12 European countries and as of 2013 has 20 member states and 7 observers. Its main purpose is to operate the largest particle physics laboratory complex in the world.

The main mission of CERN is to study the basic constituents of matter. The main instruments used are particle accelerators, which boost beam of particles to high speeds before they are made to collide with each other or with particles in fixed targets, and detectors which detect and record results of these collisions. As of 2013 CERN operates the largest particle accelerator in the world – the Large Hadron Collider (LHC).

COMPASS (Common Muon and Proton Apparatus for Structure and Spectroscopy) is a fixed-target high-energy physics experiment located at the Super Proton Synchrotron (SPS) particle accelerator. The main purpose of COMPASS is to investigate the nucleon spin structure and hadron structure and spectroscopy using high intensity hadron and muon beams.

During the long shutdown of CERN accelerators in 2013 and 2014 (LS1) the main plans for the COMPASS experiment include development of the new data acquisition system together with the upgrade of the online database structure.

The COMPASS experiment has around 250.000 detector channels along the 60 m long spectrometer setup (see Figure 1). Data from the detectors are produces via the frontend electronics which feeds the data into 9U VME concentrator modules called CATCH or into HGeSiCA boards. The readout is triggered by the Trigger control system (TCS). The trigger decision is based on the energy deposited by charged particles on hadronic and electromagnetic calorimeters and on signals from some other detectors.

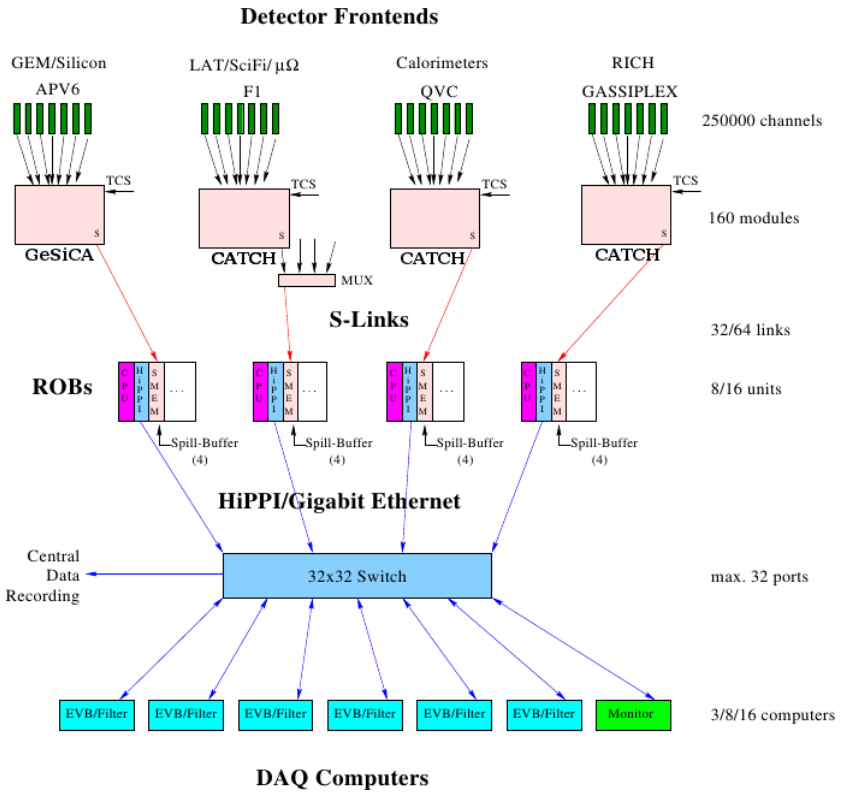


Figure 2: Current COMPASS DAQ system

3 COMPASS DAQ system

The data from CATCH and HGeSiCa boards are transfered through optical links to the DAQ computers – readout buffers (ROB). ROB computers are equipped with spillbuffer PCI cards that buffers the transmitted data. Last layer of computers (event builders) then combines the detector data to complete blocks (events), prepare meta-information, and after 24 hours transfers the data to the Central Data Recording System. The data are then compressed and stored on magnetic tapes in a permanent storage (CASTOR – CERN Advanced Storage) for further processing and analysis.

Several aspects of the experiment are constantly monitored (e.g. operation of the frontend electronics, rate of different triggers, and beam stability). The monitoring is performed on the fly by the DAQ software.

The software for the COMPASS DAQ system is based on the DATE package [3] written for the LHC experiment ALICE.

DATE performs data acquisition in a distributed environment. It provides framework for the detector readout, software for run control, event building, information logging, and event sampling. It also allows for the interactive configuration.

After the physics program in 2011 the COMPASS experiment was approved for 6 more years and it was decided to build a new data acquisition system [5].

The new data acquisition system uses FPGA modules in two different modes: 15 to 1 multiplexer to reduce the number of links from one hundred to 8 and 8x8 switch to combine data belonging to one event. These custom made modules collect and build

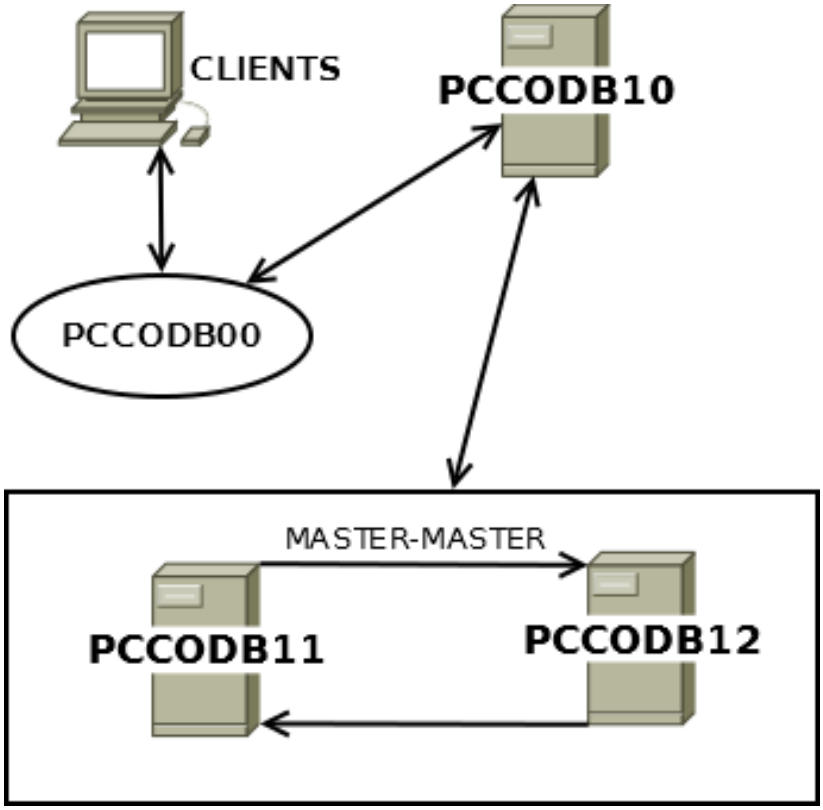


Figure 3: Current COMPASS database structure

complete events without use of any event building software and provide complete events to the readout computers. Deployment of these modules significantly reduces the amount of components involved in the COMPASS DAQ chain, which allows for simplification of the software architecture.

3.1 Current COMPASS online database

The current COMPASS online database (Figure 3) was desinged and implemented by Vladimir Jary in 2010 [4]. It uses three physical servers – all databases are stored on two servers (named *pccodb11* and *pccodb12*) which are synchronized using the master-master replication, i.e. each query executed on *pccodb11* is immediately executed also on *pccodb12* and vice versa. The third server (*pccodb10*) serves as a proxy server and is accessible via *pccodb00* virtual address.

The replication is implemented by three processes (one on the maser server and two on slave server) that read and write binary log files containing changes made to the database tables. On the master server the process reads contents of the binary log and sends updates to the slave server. On the slave server the first process connets to its master server, receives the updates of the binary log, and writes them into a relay log. The second process reads the modification stored in the relay log and executes them. The replication is an asynchronous process, the slave servers do not have to be connected permanently.

Pccodb10, *pccodb11*, and *pccodb12* servers are located in the COMPASS experimental hall in the French part of CERN and are connected to the COMPASS internal network. To increase the safety of the data, *pccodb11* is replicated also to the *compass02* server which is located in the CERN computing center. *Compass02* is also replicated to computer centers of participating institutes to provide a kind of geographical backup in case of problems on the COMPASS internal network. All three servers have the same configuration – 8 core Intel Xeon processor at 2.5 GHz with 16 GB of memory. They are running 64 bit Scientific Linux CERN 5.4 and MySQL server version 5.1.45.

3.2 Nagios monitoring system

The Nagios monitoring software is used to watch over the database system. It is able to monitor available resources on a remote host and present the results in a graphical web interface. Furthermore, the Nagios system is able to perform a predefined action in case of an accident. For example if Nagios detects that *pccodb11* server is down, it reconfigures the proxy server to redirect all clients to the *pccodb12* server. It can notify a system operator by an e-mail or a SMS.

Nagios is very flexible and customizable by plugins. Each Nagios plugin is a small application or script that monitors a state of service or resource and returns an integer value which represents the state itself. A plugin can also print multiple lines of text describing the state in more detail. Nagios periodically executes the plugins and displays the output in a graphical web interface.

3.3 Database incidents

3.3.1 May 2012 incident

A serious problem appeared in May 2012, just few hours before end of the winter shutdown and start of the data taking, the *pccodb11* has crashed as a result of hardware failure. Database experts were notified by the e-mail message sent by the Nagios system. After the *pccodb11* server had been restarted, the replication to the *pccodb12* stopped working. The same problem appeared also on the *compass02* server.

After a short investigation following problem was identified – the thread responsible for storing events to the binary log on the *pccodb11* machine was not running and the *"Client requested master to start replication from impossible position"* error was reported. Thus, the master server failed to write all events into its binary log before the crash and after the restart and the slave server was trying to receive them. Several attempts were made to force the slave process to skip the unwritten events, but it kept crashing.

To ensure the full synchronization, the replication process had to be restarted and data synchronized manually. After the operation, the replication was started again without any problems.

3.3.2 October 2012 incident

During the data taking in October 2012, the COMPASS DAQ system tried to execute the following query during the night shift: *UPDATE tb_run SET title="possibly ok"*. If

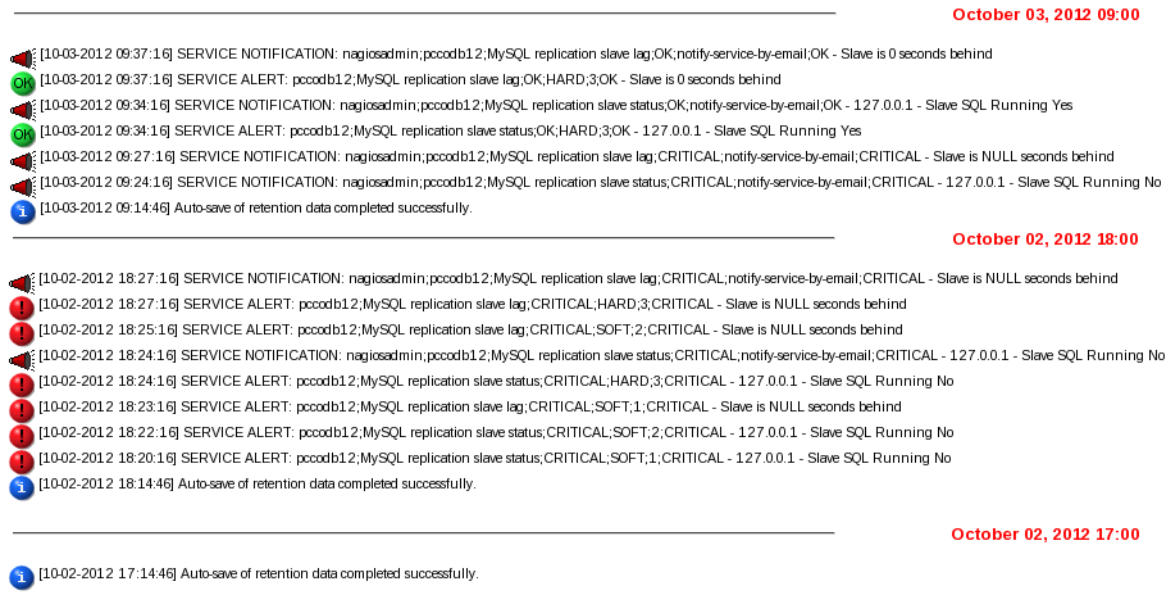


Figure 4: Nagios monitoring system log during the October 2012 incident

executed, this query would rewrite the *title* information of all COMPASS runs to *possibly ok*. While trying to execute the query, the replication process crashed and the query was fortunately not executed at all. The Nagios system sent an e-mail notification about the replication process containing error messages *Slave SQL Running No* and *Slave is NULL seconds behind* stating that the slave replication process was not running and was unable to calculate the replication delay. As the replication was not working, there was no redundancy during the night – i.e. the *pccodb11* was processing queries but not replicating them to *pccodb12* nor to *compass02*. In the morning the query was manually skipped, the replication to *pccodb12* and *compass02* was restored.

During further investigation a bug in the DATE software was found. The shift crew operating the COMPASS spectrometer is responsible for log keeping and for evaluating finished runs. To do this the DATE software provides a graphical interface to fill in the comment and some specific flags, the current run number is automatically pre-filled. However, it has been discovered that if the run number is erased and the comment saved, the DATE software tries to apply the change to all the runs. This bug was immediately fixed to prevent further problems.

3.3.3 Outcome of the incidents

After these two incidents, it was decided that the database structure should change to provide more redundancy in case of a failure. With the current structure, when one of the two servers fails, there is no backup until the problem is fixed. Also some more serious problems might cause limitations during the data taking. For example when the replication crashes in a similar manner as in May 2012, the full database backup is needed which makes the database unavailable during the task, i.e. the data taking cannot proceed. One of the possible solutions is to build the master$\rightarrow n$slaves replication.

To increase the redundancy of the whole system, more nodes should be added and

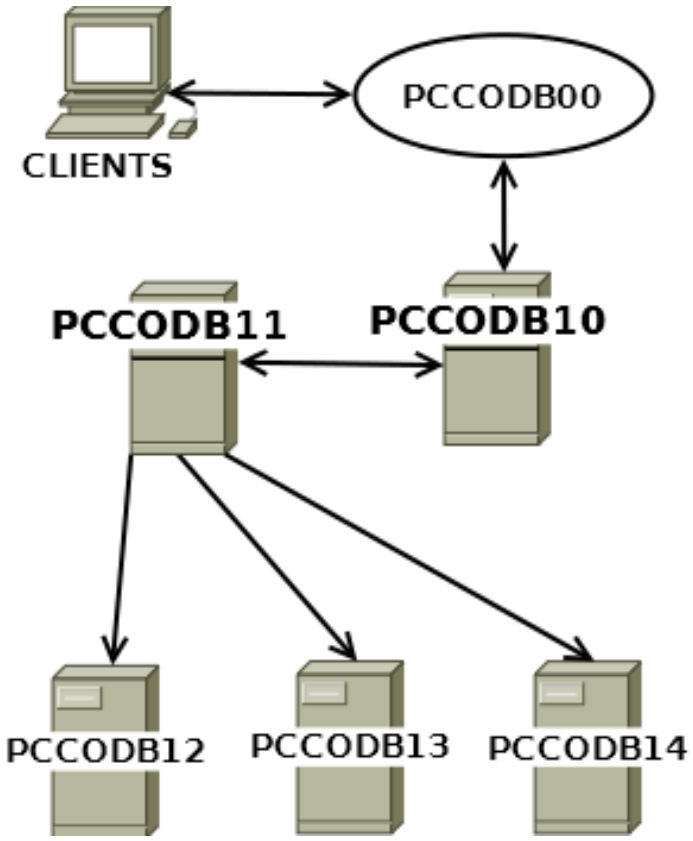


Figure 5: New COMPASS database structure proposal

the master→n slaves replication should be used (see Figure 5).

4 New database structure proposal

As in the current structure, all the clients access the MySQL proxy server located on the *pccodb10* machine via *pccodb00* virtual address. The *pccodb11* machine becomes the only master node and every query executed on this node is then executed on all the slave nodes (*pccodb12*, *pccodb13*, *pccodb14*) as well.

In case of a failure of one of the slave nodes, the master node keeps replicating to the rest of the nodes. The problem of one node can then be fixed without causing any limitations or without losing the backup.

In case of a failure of the master node, one slave node can immediately take its place by redirecting the proxy server to it. Again, this process cuses no limitations during the data taking.

This upgrade should be performed together with upgrading the Scientific Linux CERN and MySQL to the most up-to-date versions on all the machines.

5 Conclusion

The new database structure of the COMPASS online database was proposed and was preliminarily approved on a meeting of COMPASS front-end group. The change should be implemented as soon as required hardware is available (i.e. at the end of 2013 or during 2014).

References

[1] Adolph Ch. et al. (The COMPASS Collaboration): *COMPASS-II Proposal*. CERN-SPSC-2010-014; SPSC-P-340, May 2010.

[2] Widenius M. et al.: *MySQL Reference Manual*. O'Reilly Media, June 2002.

[3] Anticic T. et al. (ALICE DAQ Project): *ALICE DAQ and ECS User's Guide*. CERN, EDMS 616039, January 2006.

[4] Fleková L., Jarý V., Liška T., Virius M.: *Využití databází v rámci fyzikálního experimentu COMPASS*. In: Konference Tvorba softwaru 2010, Ostrava: VŠB - Technická univerzita Ostrava, 2010, ISBN 978-80-248-2225-9 pp. 68–75.

[5] Bodlák M., Frolov V., Jary V., Huber S., Konorov I., Levit D., Mann A., Novy J., Paul S., Virius M. *New data acquisition system for the COMPASS experiment*. 2013 JINST 8 C02009 doi:10.1088/1748-0221/8/02/C02009.

Java Source Code Structures Scanning by Graph Matching Algorithm

Tomáš Bublík

4th year of PGS, email: tomas.bublik@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 introduces a tool which is able to recognize both the properties of a code snippet and even its structure. This tool uses the Sripthon language to describe a snippet. An abstract syntax tree is created from the given Java source, and it is compared with the tree created from a Scripthon source code. During this process, many tree optimizations take place. Therefore, the complete recognition process is very fast, and can be used to scan the large programs.

Keywords: Java, graph matching, AST, Scripthon

Abstrakt. Tato práce představuje nástroj sloužící k detekci struktur zdrojového kódu a jejich vlastností. K popisu vlastností je použit jazyk Scripthon. Při porovnávání se používá strom abstraktní syntaxe získaný ze zdrojového kódu. Tento strom je porovnán se stromem, který je vytvořen ze zdrojového kódu jazyka Scripthon. Během tohoto procesu dochází k mnoha optimalizacím. Proto je vyhledávání velice rychlé a tedy i použitelné pro programy větší velikosti.

Klíčová slova: Java, porovnávání grafu, strom abstraktní syntaxe, Scripthon

1 Introduction

It is an easy task to search the source code. Nevertheless, this applies only in the case of a simple text or simple structure names. This feature is supported in most of the current Java development environments. Some IDEs support an advanced searching with regular expressions. But, what if a user wants to know, whether a program contains the singleton design pattern? Or, whether the specific method (with three concrete parameters) is somewhere in a program?
It is very difficult to find such information; however, with using the mathematical and programming knowledge, it is possible. When using the Scripthon language, these special structures can be described very precisely. On the other side, by using the Java Compiler API, the abstract syntax trees (hereinafter AST) can be obtained and compared with the Scripthon's output. This paper is on the using these trees for searching the desired code snippet. This task is similar to the graph matching and isomorphic subgraphs finding in a large set of trees. A number of solutions for all of these tasks have been proposed [6], but they all suffer from the high computational complexity inherent to the graph matching. An additional problem arises in the applications where an input graph needs to be matched not only to another graph, but to an entire database of the graphs

under the given matching paradigm. Therefore, some complexity reducing algorithms are proposed in this paper.
The first section introduces necessary graph theory concepts. There can be found the definitions of a graph, a subgraph and a graph isomorphism. The next two sections are about the graphs generation, optimizations, and the comparison of the graphs generated by the Compiler API. The Scripthon language is introduced briefly in the next chapter. Because the language has been described already in another paper [1], only the important properties are mentioned here. Finally, some results are presented in the conclusion.

There are several reasons to consider graphs as a very advantageous tool for the representation of a source code of some language. One the reason is, that there is no unnecessary material like spaces, comments etc. Another reason is, that there are many well described mathematical algorithms to work with graphs. Some of the algorithms are known for decades. Representing a code as a graph has also the disadvantage: it has a large demands on a computer power and memory; especially for larger programs.

2 Basic graph theory concepts

A graph is defined as a four-tuple $\mathtt{g} = (\mathtt{V}, \mathtt{E}, \alpha, \beta)$, where $\mathtt{V}$ denotes a finite set of nodes, $\mathtt{E} \subseteq \mathtt{V} \times \mathtt{V}$ is a finite set of edges, $\alpha : \mathtt{V} \to \mathtt{L_V}$ is a node labeling function, and $\beta : \mathtt{E} \to \mathtt{L_E}$ is an edge labeling function. $\mathtt{L_V}$ and $\mathtt{L_E}$ are finite of infinite sets of node and edge labels, respectively. All the graphs in this work are considered to be directed.
A subgraph $\mathtt{g_s} = (\mathtt{V_s}, \mathtt{E_s}, \alpha_\mathtt{s}, \beta_\mathtt{s})$ of a graph $\mathtt{g}$ is a subset of its nodes and edges, such that $\mathtt{V_s} \subseteq \mathtt{V}, \mathtt{E_s} = \mathtt{E} \cap (\mathtt{V_s} \times \mathtt{V_s})$
Two graphs $\mathtt{g}$ and $\mathtt{g'}$ are isomorphic to each other if there exists a bijective mapping $\mathtt{u}$ from the nodes of $\mathtt{g}$ to the nodes of $\mathtt{g'}$, such that the structure of the edges as well as all node and edge labels are preserved under $\mathtt{u}$. Similarly, an isomorphism between a graph $\mathtt{g}$ and a subgraph $\mathtt{g'_s}$ of a graph $\mathtt{g'}$ is called subgraph-isomorphism from $\mathtt{g}$ to $\mathtt{g'}$.
A tree is a connected and undirected graph with no simple circuits. Since a tree cannot have a circuit, a tree cannot contain multiple edges or loops. Therefore, any tree must be a simple graph. An undirected graph is a tree if and only if there is a unique simple path between any two of its vertices.
The two graphs matching problem is actually the same as the finding the isomorphism between them. Moreover, matching the parts of a graph with a pattern is the same challenge as the finding the isomorphic subgraph.

3 Graph generation with Java Compiler API

The Java Compiler API is used to get a graph for the searching algorithm. This API is free, and it is included in a Java distribution. Basically, the Java Compiler API serves to the advanced control of a compilation process. This API uses AST in the form of the visitor design pattern. Unfortunately, this design pattern is no so convenient for the searching purposes. This is because the Scripthon language is unable to describe so many structures, and also because the searching algorithm is difficult to implement with the visitor design pattern. Therefore, the more advanced graph is created from Java AST.

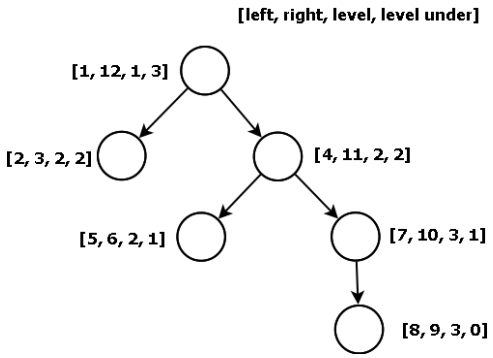


Figure 1: Tree with optimizations

This graph is very similar to AST, but it has several benefits. The first benefit is the replacement of the visitor pattern with the classic approach. And the second one is the enrichment of some additional information which significantly facilitates the searching. While browsing a source code, the tree with nodes enhanced by four numbers, is created. These numbers are the natural numbers named left, right, level and level under. The first and the second number (left, right) denotes the order index of a node after the tree preorder traversal. The level number denotes the level in a tree hierarchy of vertices, and the level under number denotes a number of levels under the current node. (Compare with the method described in [4]). The following rules are valid for these values.
Suppose that `x` and `y` are two nodes from a tree.

- The `y` node is an ancestor of `x` and `x` is a descendant of `y` if $\texttt{y.left} < \texttt{x.left} < \texttt{y.right}$
- The `y` node is an parent of `x` and `x` is a child of `y` if 1) $\texttt{y.left} < \texttt{x.left} < \texttt{y.righ}$ and 2) $\texttt{y.level} = \texttt{x.level} - 1$

All these data are acquired during a single pass through the tree. Obtaining this information is not a time consuming operation, because it is made during the tree production process. On the other hand, the number of comparisons can be significantly reduced with these numbers. Moreover, while comparing the trees, it is very easy to detect:

- How many elements have a given structure
- If a node is a leaf
- How many sub-statements are included in a given structure

Without this information, the comparison of two trees becomes much more time consuming operation. In summary, this information is used in the cases where the shape of the

given structures and its coupling is considered more than its properties.
A line reference to a source code is an important information which is also added to the tree as a metadata. Therefore, it is easy to link the results with the original source position and show it to the user. There are some more elements in a node metadata. For example, some of the other metadata information is a filename of the source file.
Because the number of the comparisons is a key indicator for the algorithm speed, it is necessary to keep the number of nodes as small as possible. Therefore, while creating a tree from a source code, only the supported structures and its properties are considered . Thus, the same Scripthon definition set is used during the tree creation process. Other elements are omitted.

4 Scripthon description

The Scripthon language is described in [1], [2]. The following text will present just the summarized and important properties of this language. Scripthon is a simple-to-lean language which is able to describe a Java source code structure. Because of its simple syntax, it is very easy to learn. The syntax of the Scripthon language is similar to the syntax of Java, and it is very intuitive. Basically, keywords represent the structures in Java language. Thus, a Scripthon program is built only with these words and its properties. Each keyword has a special set of own properties. For example, a class is represented by the `Class()` keyword. The parameters of this structure can be in the parentheses, however, if the brackets includes no parameters, each class is a candidate for searching and each class of a given program corresponds to this structure. For example, the following command:

```
Class(Name = "Main";Rest = public)
```

means that the wanted structure is a public class with the name *Main*. Each option of all parameters is specified in the Scripthon documentation. It is denoted only by the line separators or by tabs, how the structures are nested together and how the searched hierarchy looks like.

```
Meth(Rest = private;ParamsNum = 2)
    Block()
        Init(Type = int;Value = "";Name = "sum")
            Return(Value = "sum")
```

This example means that the searched structure is a private method with two parameters. Inside the method is a block with two statements. The first statement is a variable named *sum* of type int. The second statement is a return statement with a parameter of the previously specified variable.

The big advance of the Scrithon language is that it is able to describe the elements with a variable depth of details. It means, that the searched structures can be described in a detail or very loosely. For example, this is a very detailed description:

```
Class(Name = "TestDecompile"; Rest = public)
   Meth(Name = "main"; Ret = void; Rest = public)
      Init(Name = "toPrintValue"; Type = String)
         MethCall(Name = "System.out.println")
```

The same script without details follows:

```
Class()
   Meth()
      Init()
         MethCall()
```

Therefore, a searched subject can be found on the base of a very inaccurate description. The results can be obtained with the iterative refinement of the input conditions. In the end, a user can get the better results.
The level of detail which can be described by the current version of Scripthon is up to the expression. In addition, Scripthon can describe a lot of Java structures, but it cannot describe the individual elements of an expression statement. For example, while describing an *if* statement, it is possible to address the inner block, or the else block with inner statements, however, the *if* expression in the parentheses cannot be described. Moreover, Scripthon is not able to describe the mathematical operations. If a variable is this way:

```
int i = a + b/45;
```

The most accurate statement in Scripthon is, that finds it, is:

```
Init(Name = "i"; Type = int)
```

In the current version of the Scripthon language, nothing more cannot be described yet.

5 Graph matching

The simple and many times described backtracking algorithm is used for the graph matching. Basically, it is the problem of finding an isomorphic tree to the given tree from a large database of trees. The only difference is, that the node properties need to be considered during the process.
The source trees are created from the corresponding classes. The classes and the trees are mapped one-by-one. Each tree corresponds to exactly one class. In the first step, the algorithm checks whether the shape of the structure match, and then the properties are compared. This is because the properties matching is much more time consuming operation than shape detection. Many structures are eliminated from the process very quickly in the case that the shape does not fit.
If the shape of the structure corresponds to the required shape, the structure parameters are compared. All the parameters of a given node must be met. The node properties are provided by the Java compiler. Unfortunately, because of the backtracking algorithm, each node needs to be compared one-by-one. It has $O(N^3)$ complexity (according to [3]). On the contrary, with the above outlined optimizations, the number of node comparisons is significantly decreased. More on the graph matching techniques can be found in [5].

6 Conclusion

The used algorithm modifications substantially reduced the time needed to find the requested Java structures. Moreover, also the time of the tree generation procedure has been shortened. According to the measurements, the meta-information counting does not significantly affect the time of a graph creation.
The searching with optimization is much faster. The tables I-III show the measured time results. The small program means a program consisting of approximately 20 to 30 classes, while the larger program is a program with approximately 100 to 150 classes. There are also the results before and after the described optimizations.

Table 1: Graph creation

Program type	**Time**
Small program (no optimizations)	412 ms
Larger program (no optimizations)	4 423 ms
Small program (optimized)	132 ms
Larger program (optimized)	337 ms

Table 2: Searching

Program type	**Time**
Small program (no optimizations)	2 345 ms
Larger program (no optimizations)	11 236 ms
Small program (optimized)	753 ms
Larger program (optimized)	1 986 ms

Table 3: Total time

Program type	**Time**
Small program (no optimizations)	2 757 ms
Larger program (no optimizations)	15 659 ms
Small program (optimized)	886 ms
Larger program (optimized)	2 323 ms

The measurements were performed on a quite common computer. The computer configuration was: 4GB of memory, an Intel Core I5 processor with a frequency of 2,4 GHz and Windows 7 as an operating system. The individual results represent the averages of several consecutive measurements. The first column indicates the time needed to AST generation, while the second one represents the time required to find a piece of the sample code described by the Scripthon language. The last column is the sum of both times. The lines represent the sizes of programs on which the measurements were performed.
As you can see from the tables, in the case of the small program, the graph assembling is not significantly different. On contrary, better results can be obtained in the case of larger programs. Probably, this is because the time needed for the overhead services related to the starting and initializing the own search.

References

[1] Bublík, T., Virius, M., *New language for searching Java code snippets* In 'ITAT 2012. Proc. of the 12th national conference ITAT. Ždiar, Sep 17 – 21 2012. Pavol Jozef Šafárik University in Košice. p. 35 – 40.

[2] Bublík, T., Virius, M., *Automatic detecting and removing clones in Java source code* In 'Software Development 2011. Proc. of the 37th national conference Software Development. Ostrava', May 25 – 27 2011. Ostrava: Technical University of Ostrava 2011. ISBN 978-80-248-2425-3. p. 10 – 18.

[3] Ira D. Baxter and Andrew Yahin and Leonardo Moura and Marcelo Sant' Anna and Lorraine Bier, *Clone Detection Using Abstract Syntax Trees* ICSM '98 Proceedings of the International Conference on Software Maintenance, IEEE Computer Society Washington, DC, USA 1998.

[4] J. T. Yao and M. Zhang, *A Fast Tree Pattern Matching Algorithm for XML Query* In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI '04). IEEE Computer Society, Washington, DC, USA, 235-241, 2004.

[5] Horst Bunke, Christophe Irniger, and Michel Neuhaus, *Graph matching - challenges and potential solutions* In Proceedings of the 13th international conference on Image Analysis and Processing (ICIAP'05), Fabio Roli and Sergio Vitulano (Eds.). Springer-Verlag, Berlin, Heidelberg, 1-10, 2005.

[6] Chanchal K. Roy, James R. Cordy, and Rainer Koschke, *Comparison and evaluation of code clone detection techniques and tools: A qualitative approach* Sci. Comput. Program. 74, 7 (May 2009), 470-495, 2009.

Generating $(\pm\beta)$-integers by Conjugated Morphisms*

Daniel Dombek

4th year of PGS, email: dombedan@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Zuzana Masáková, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In this paper we study the expansions of real numbers in positive and negative real base. In particular, we consider the sets $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ of nonnegative β-integers and $(-\beta)$-integers respectively. It is well known that, in numerous cases, $\mathbb{Z}_{-\beta}$ can be completely unrelated to $\mathbb{Z}_\beta^+$. We precisely describe all bases $\pm\beta \in \mathbb{R}$ for which $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ can be coded by infinite words which are fixed points of two conjugated morphisms.

Full version of this contribution, *Generating $(\pm\beta)$-integers by Conjugated Morphisms*, was published in Local Proceedings of WORDS 2013, [4].

Keywords: $(-\beta)$-expansions, $(-\beta)$-integers, antimorphism, conjugation

Abstrakt. Tento příspěk se zabývá rozvoji reálných čísel v kladné a záporné bázi, konkrétně množinami $\mathbb{Z}_\beta^+$ a $\mathbb{Z}_{-\beta}$ nezáporných β-celých a $(-\beta)$-celých čísel. Je známo, že množiny $\mathbb{Z}_\beta^+$ a $\mathbb{Z}_{-\beta}$ se obecně mohou značně lišit. Přesně popíseme všechny báze $\pm\beta \in \mathbb{R}$, pro které lze množiny $\mathbb{Z}_\beta^+$ a $\mathbb{Z}_{-\beta}$ kódovat nekonečnými slovy, které jsou pevnými body konjugovaných morfismů.

Nezkrácená verze tohoto příspěvku, *Generating $(\pm\beta)$-integers by Conjugated Morphisms*, vyšla v Local Proceedings of WORDS 2013, [4].

Klíčová slova: $(-\beta)$-rozvoje, $(-\beta)$-celá šísla, antimorfismus, konjugace

1 Introduction

Inspired by the work of Ito and Sadahiro [7], numerous papers have been recently dedicated to the study of numeration systems with negative base from various perspectives. Typically, properties of $(-\beta)$-expansions are examined in comparison with their well-known positive base counterparts. The dynamical properties of $(-\beta)$-transformations were studied for example in [3], [6] and [9]. For the results on the set of $(-\beta)$-integers, see [2] and [15] while a related topic, arithmetics on $(-\beta)$-integers, was studied for instance in [10]. Recently, an effort was made in identifying numbers β for which β- and $(-\beta)$-numerations are the "most similar". In particular, Kalle in [8] characterizes $\beta \in (1, 2)$ for which there exists a measurable isomorphism between β- and $(-\beta)$-transformations. In [11], the authors focus on comparison of languages of infinite words u_β and $v_{-\beta}$ coding

*This work was supported by the Czech Science Foundation, grant GAČR 13-03538S and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS11/162/OHK4/3T/14.

the β- and $(-\beta)$-integers, respectively, in case of quadratic $\beta > 1$. The present contribution extends the result of [11] by providing the characterization of numbers β for which the languages of u_β and $v_{-\beta}$ coincide. For $\beta \in (1,2)$, this happens exactly for the class of multinacci numbers, distinguished in [8].

2 Rényi β-expansions

In 1957, Rényi [13] defined the positional numeration system with positive (in general real) base. Let $\beta > 1$, then any $x \in [0,1)$ has a unique expansion of the form $d_\beta(x) = x_1 x_2 x_3 \cdots$ defined by

$$x_i = \lfloor \beta T_\beta^{i-1}(x) \rfloor, \text{ where } T_\beta(x) = \beta x - \lfloor \beta x \rfloor .$$

For any $x \in [0,1)$ we then get an infinite word, more precisely an element of $\mathcal{A}^{\mathbb{N}} = \{0, 1, \ldots, \lceil \beta \rceil - 1\}^{\mathbb{N}}$. On the other hand, not every infinite word over $\mathcal{A}^{\mathbb{N}}$ does play the role of $d_\beta(x)$ of some $x \in [0,1)$. Those who do, are called admissible (or β-admissible) and their characterization is due to Parry [12]. He proved that a digit string $x_1 x_2 \cdots \in \mathcal{A}^{\mathbb{N}}$ is admissible iff it fulfills the lexicographic condition

$$0^\omega \preceq_{\text{lex}} x_i x_{i+1} x_{i+2} \cdots \prec_{\text{lex}} d_\beta^*(1) = \lim_{y \to 1_-} d_\beta(y) \text{ for all } i \geq 1 . \tag{1}$$

Here, $\prec_{\text{lex}}$ stands for standard lexicographic ordering and the limit is taken over the usual topology on $\mathcal{A}^{\mathbb{N}}$. Recall that the so-called Rényi expansion of unity is defined as

$$d_\beta(1) = d_1 d_2 d_3 \cdots , \text{ where } d_1 = \lfloor \beta \rfloor, d_2 d_3 \cdots = d_\beta(\beta - \lfloor \beta \rfloor) .$$

The infinite Rényi expansion of unity $d_\beta^*(1)$ can then be obtained as

$$d_\beta^*(1) = \begin{cases} (d_1 \cdots d_{m-1}(d_m - 1))^\omega & \text{if } d_\beta(1) = d_1 \cdots d_m 0^\omega \text{ with } d_m \neq 0, \\ d_\beta(1) & \text{otherwise,} \end{cases}$$

where the notation $w^\omega = www\cdots$ stands for infinite repetition of the word w. Let us point out that the lexicographic ordering on admissible strings corresponds to the ordering on the unit interval $[0,1)$, i.e. $x < y \Leftrightarrow d_\beta(x) \prec_{\text{lex}} d_\beta(y)$.

The notion of β-expansions can be naturally extended from $[0,1)$ to all reals.

Definition 1. *Let $\beta > 1$, $x \in \mathbb{R}^+$. Let $k \in \mathbb{N}$ be minimal such that $\frac{x}{\beta^k} \in [0,1)$ and $d_\beta\left(\frac{x}{\beta^k}\right) = x_1 x_2 x_3 \cdots$. Then the β-expansion of x is defined as*

$$\langle x \rangle_\beta = \begin{cases} x_1 \cdots x_{k-1} x_k \bullet x_{k+1} x_{k+2} \cdots & \text{if } k \geq 1, \\ 0 \bullet x_1 x_2 x_3 \cdots & \text{if } k = 0, \end{cases}$$

where the symbol $\bullet$ separates integer and fractional parts of $\langle x \rangle_\beta$. The β-expansion of a negative real number is then defined as $\langle x \rangle_\beta = -\langle |x| \rangle_\beta$.

As a natural generalization of $\mathbb{Z}$, the set $\mathbb{Z}_\beta$ of β-integers can be defined using the notion of $\langle x \rangle_\beta$.

Definition 2. *Let $\beta > 1$. Then the set of nonnegative β-integers is defined as*

$$\mathbb{Z}_\beta^+ = \{x \in \mathbb{R} \ : \ \langle x \rangle_\beta = x_k \cdots x_1 x_0 \bullet 0^\omega\} \ = \ \bigcup_{i \geq 0} \beta^i T_\beta^{-i}(0)\,.$$

The set of all β-integers is then obtained by symmetrization around zero,

$$\mathbb{Z}_\beta = \mathbb{Z}_\beta^+ \cup (-\mathbb{Z}_\beta^+)\,.$$

Recall that a number $\beta > 1$ is called a Parry number, if $d_\beta^*(1)$ is eventually periodic. Note that every Parry number is necessarily an algebraic integer. If it is also purely periodic (i.e. the Rényi expansion of unity $d_\beta(1)$ is finite), then it is called a simple Parry number. The remaining Parry numbers are called non-simple Parry numbers.

Thurston [16] showed that the distances between consecutive elements of $\mathbb{Z}_\beta$ (let us denote them as $\Delta_0, \Delta_1, \Delta_2, \ldots$) are equal to

$$\Delta_k = \sum_{i \geq 1} \frac{d_{i+k}}{\beta^i}, \quad k = 0, 1, 2, \ldots\,, \tag{2}$$

where $d_\beta^*(1) = d_1 d_2 d_3 \cdots$. One can easily see that the set $\mathbb{Z}_\beta$ contains gaps of finitely many different lengths iff β is a Parry number. Moreover, since $\Delta_0 = \sum_{i \geq 1} \frac{d_i}{\beta^i} = 1$ and any suffix of $d_\beta^*(1)$ either fulfills (1) or is equal to $d_\beta^*(1)$ itself, we get $\Delta_k \leq 1$ for all k.

We can encode $\mathbb{Z}_\beta^+$ by an infinite word in the following manner. Starting with number 0 (which is always a β-integer) and continuing through all elements of $\mathbb{Z}_\beta^+$ in increasing order, encoding each gap between consecutive β-integers by a number $\Delta_k \to k$ (where k is the greatest index, at which the β-expansions of the two neighbors differ) will give us an infinite word $u_\beta = u_0 u_1 u_2 \cdots$ over the infinite alphabet $\mathbb{N}$. We can generate this encoding by a certain morphism, having u_β as its fixed point.

If β is a Parry number, the distances between consecutive elements of $\mathbb{Z}_\beta^+$ take only finitely many values and it is known, that both u_β and φ can be projected onto a finite alphabet of the form $\{0, 1, \ldots, n\}$. The explicit form of the morphism φ fixing u_β was originally given by Fabre [5].

Theorem 1 ([5])**.** *Let $\beta > 1$ be a Parry number. Then the morphism $\varphi : \{0, \ldots, n\}^* \to \{0, \ldots, n\}^*$ encoding $\mathbb{Z}_\beta^+$ has the following form:*

- *If β is a simple Parry number, $d_\beta(1) = d_1 \cdots d_k 0^\omega$ ($d_\beta^*(1) = (d_1 \cdots d_{k-1}(d_k - 1))^\omega$), then $n = k - 1$ and*

$$\begin{aligned} \varphi(i) &= 0^{d_{i+1}}(i+1) \quad \text{for } i \leq k-2\,, \\ \varphi(k-1) &= 0^{d_k}\,. \end{aligned}$$

- *If β is a non-simple Parry number, $d_\beta(1) = d_\beta^*(1) = d_1 \cdots d_k (d_{k+1} \cdots d_{k+p})^\omega$, then $n = k + p - 1$ and*

$$\begin{aligned} \varphi(i) &= 0^{d_{i+1}}(i+1) \quad \text{for } i \leq k+p-2\,, \\ \varphi(k+p-1) &= 0^{d_{k+p}} k\,. \end{aligned}$$

3 Ito-Sadahiro $(-\beta)$-expansions

In 2009, Ito and Sadahiro [7] introduced an analogous numeration system to Rényi β-expansions which uses a negative base, the so-called $(-\beta)$-expansions. Instead of defining the expansions of numbers from $[0,1)$ first, the unit interval $[\ell,\ell+1)$ with $\ell=\frac{-\beta}{\beta+1}$ fixed was chosen. Let $-\beta<-1$, then any $x\in[\ell,\ell+1)$ has a unique expansion of the form $d_{-\beta}(x)=x_1x_2x_3\cdots$ defined by

$$x_i=\lfloor -\beta T_{-\beta}^{i-1}(x)-\ell\rfloor,\ \text{where } T_{-\beta}(x)=-\beta x-\lfloor -\beta x-\ell\rfloor\,.$$

As in Rényi numeration system, we get for any $x\in[\ell,\ell+1)$ an infinite word from $\mathcal{A}^{\mathbb{N}}=\{0,1,\ldots,\lfloor\beta\rfloor\}^{\mathbb{N}}$.

Another analogous concept is the $(-\beta)$-admissibility, which characterizes all digit strings over $\mathcal{A}$ being the $(-\beta)$-expansion of some number. The lexicographic condition, similar to the one by Parry, was also proved in [7]. Ito and Sadahiro proved that a digit string $x_1x_2x_3\cdots\in\mathcal{A}^{\mathbb{N}}$ is $(-\beta)$-admissible (or, if no confusion is possible, just admissible) iff it fulfills the lexicographic condition

$$d_{-\beta}(\ell)\preceq_{\text{alt}}x_ix_{i+1}x_{i+2}\cdots\prec_{\text{alt}}d^*_{-\beta}(\ell+1)=\lim_{y\to l+1_-}d_{-\beta}(y)\ \text{for all } i\geq 1\,. \tag{3}$$

Here, $\prec_{\text{alt}}$ stands for alternate lexicographic ordering defined as follows:

$$u_1u_2\cdots\prec_{\text{alt}}v_1v_2\cdots\ \Leftrightarrow\ (-1)^k(u_k-v_k)<0\ \text{for } k \text{ smallest such that } u_k\neq v_k\,.$$

The reference digit strings $d_{-\beta}(\ell)$ and $d^*_{-\beta}(\ell+1)$ play the same role for $(-\beta)$-expansions as Rényi expansions of unity for β-expansions. While $d_{-\beta}(\ell)$ is obtainable directly from the definition, the following rule (proved in [7]) is to be used for determining $d^*_{-\beta}(\ell+1)$:

$$d^*_{-\beta}(\ell+1)=\begin{cases}(0l_1\cdots l_{q-1}(l_q-1))^\omega & \text{if } d_{-\beta}(\ell)=(l_1l_2\cdots l_q)^\omega \text{ for } q \text{ odd},\\ 0d_{-\beta}(\ell) & \text{otherwise.}\end{cases}$$

In the rest of the paper, the notation $d_{-\beta}(\ell)=l_1l_2l_3\cdots$ will be used. We can now recall the definition of $(-\beta)$-expansions for all reals.

Definition 3. *Let $-\beta<-1$, $x\in\mathbb{R}$. Let $k\in\mathbb{N}$ be minimal such that $\frac{x}{(-\beta)^k}\in(\ell,\ell+1)$ and $d_{-\beta}\left(\frac{x}{(-\beta)^k}\right)=x_1x_2x_3\cdots$. Then the $(-\beta)$-expansion of x is defined as*

$$\langle x\rangle_{-\beta}=\begin{cases}x_1\cdots x_{k-1}x_k\bullet x_{k+1}x_{k+2}\cdots & \text{if } k\geq 1,\\ 0\bullet x_1x_2x_3\cdots & \text{if } k=0.\end{cases}$$

Definition 4. *Let $-\beta<-1$. Then the set of $(-\beta)$-integers is defined as*

$$\mathbb{Z}_{-\beta}=\{x\in\mathbb{R}\ :\ \langle x\rangle_{-\beta}=x_k\cdots x_1x_0\bullet 0^\omega\}\ =\ \bigcup_{i\geq 0}(-\beta)^iT_{-\beta}^{-i}(0)\,.$$

In order to describe the distances between consecutive $(-\beta)$-integers, we will recall some notation from [2]. Let

$$\min(k) = \min\{a_{k-1}\cdots a_1a_0 \ : \ a_{k-1}\cdots a_1a_00^\omega \text{ is admissible}\},$$

where min is taken with respect to the alternate order on finite strings. Similarly we define $\max(k)$. Furthermore, let γ be the "value function" mapping finite digit strings to real numbers,

$$x_{k-1}\cdots x_1x_0 \quad \to \quad \gamma(x_{k-1}\cdots x_1x_0) = \sum_{i=0}^{k-1} x_i(-\beta)^i\,.$$

With this notation we can recall the results concerning the distances in $\mathbb{Z}_{-\beta}$ and, later on, encoding of $\mathbb{Z}_{-\beta}$ by infinite words. It was shown in [2] that the distances between consecutive elements $x < y$ of $\mathbb{Z}_{-\beta}$ take the values $y - x \in \{\Delta'_k,\, k \in \mathbb{N}\}$ (not necessarily pairwise distinct) with

$$\Delta'_k := \Big|(-\beta)^k + \gamma\big(\min(k)\big) - \gamma\big(\max(k)\big)\Big|\,, \tag{4}$$

where k is the greatest index at which $\langle x\rangle_{-\beta}$ and $\langle y\rangle_{-\beta}$ differ. In contrast with the result of Thurston describing the distances in $\mathbb{Z}_\beta$, it is difficult to provide a similar explicit result on $\mathbb{Z}_{-\beta}$, due to tedious discussions arising from the alternate ordering. Nevertheless, the formula (4) will be sufficient for our needs.

If we want to encode $\mathbb{Z}_{-\beta}$ by an infinite word, the procedure is similar to the encoding of $\mathbb{Z}^+_\beta$. But since $\mathbb{Z}_{-\beta}$ contains both positive and negative numbers, we directly get a biinfinite word

$$v_{-\beta} = \cdots v_{-3}v_{-2}v_{-1}|v_0v_1v_2\cdots\,, \quad v_i \in \{0,1,2,\ldots\}\,,$$

i.e. the letter $v_j = k$ means that the gap between j-th and $(j+1)$-th $(-\beta)$-integer is Δ'_k. As the following theorem shows, there exists an antimorphism ψ (although not explicitly given), which generates $v_{-\beta}$ as its fixed point, i.e. $\psi(v_{-\beta}) = v_{-\beta}$.

Theorem 2 ([2]). *Let $v_{-\beta}$ be the word associated with $(-\beta)$-integers. There exists an antimorphism $\psi : \mathbb{N}^* \to \mathbb{N}^*$ such that ψ^2 is a non-erasing non-identical morphism and $\psi(v_{-\beta}) = v_{-\beta}$.*

Moreover, ψ is of the form

$$\psi(k) = \begin{cases} S_k(k+1)\widetilde{R_k} & \text{for } k \text{ even},\\ R_k(k+1)\widetilde{S_k} & \text{for } k \text{ odd}, \end{cases}$$

where $\widetilde{u}$ denotes the mirror image of the word u. The word S_k codes the distances between consecutive $(-\beta)$-integers in $\{\gamma(\min(k)0),\ldots,\gamma(\min(k+1))\}$ (in given order) and similarly R_k in $\{\gamma(\max(k)0),\ldots,\gamma(\max(k+1))\}$.

Similarly to Parry numbers, another subclass of algebraic integers, the so-called Yrrap numbers are defined. A real number β is an Yrrap number, if $d_{-\beta}(\ell)$ is eventually periodic. Moreover, let us recall, that β is called a Pisot number, if it is an algebraic integer greater than 1 with all algebraic conjugates less than 1 in modulus. From [6], [9], [14], it is known that every Pisot number is both Parry and Yrrap, while the converse does not hold. Moreover, the sets of Parry and Yrrap numbers do not coincide.

Remark 1. *Although both the encoding $v_{-\beta}$ of $(-\beta)$-integers and the antimorphism ψ generating it were originally defined over the infinite alphabet $\mathbb{N}$, it is not difficult to see that whenever β is an Yrrap number, $v_{-\beta}$ and ψ can be projected to a finite alphabet as the distances of the same length can be coded by the same letter (periodic $d_{-\beta}(\ell) \Rightarrow$ periodic patterns in extremal strings $\min(k)$ and $\max(k) \Rightarrow$ periodic repetition of lengths of distances in $\mathbb{Z}_{-\beta}$). Several examples of antimorphisms over finite alphabet coding $\mathbb{Z}_{-\beta}$ are presented in [2] and [15].*

4 Comparing the Structure of $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$

A natural question to ask is: for given $\beta > 1$, are the sets $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ similar in any way? From our point of view, the "similarity" can be expressed by three properties (ordered in such a way that each one implies all of the previous):

1. both $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ contain only distances of length ≤ 1 (not true for $\mathbb{Z}_{-\beta}$ in general)
2. the sets of distances in $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ are the same
3. u_β and $v_{-\beta}$ are fixed points of conjugated morphisms (which implies that u_β and $v_{-\beta}$ have the same language)

Given an infinite word u (one- or two-directional), its language $\mathcal{L}(u)$ is the set of all its factors. i.e. finite words of the form $u_k u_{k+1} \cdots u_l$ for some $k, l \in \mathbb{Z}$. Note that we cannot just compare maps φ and ψ generating u_β and $v_{-\beta}$ respectively, as one of them is a morphism and the other an antimorphism. Nevertheless, if we take φ^2 and ψ^2 then the comparison makes sense, as both are morphisms.

Definition 5. *Let $\mathcal{A}$ be an alphabet (finite or infinite) and $\pi, \rho : \mathcal{A}^* \to \mathcal{A}^*$ be morphisms on $\mathcal{A}$. We say that π and ρ are conjugated, if there exists a word $w \in \mathcal{A}^*$ such that either*

$$w\pi(a) = \rho(a)w, \text{ for all } a \in \mathcal{A}, \text{ or } \pi(a)w = w\rho(a), \text{ for all } a \in \mathcal{A}.$$

We denote $\pi \sim \rho$.

4.1 Observations for special Pisot Bases

Note that we consider only non-integer bases. The case $\beta \in \mathbb{Z}$ is contained in the main result (Theorem 3) as a trivial subcase (integers are Pisot numbers of degree 1) which can be easily proved separately. In Proposition 1 we recall results of [11] characterizing quadratic bases for which both $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ are encoded by infinite words with the same language.

Proposition 1. *Let $\beta > 1$ be a quadratic Pisot number with minimal polynomial $p(x)$.*

1. *If $p(x) = x^2 - mx - m$, $m \geq 1$:*
   the distances in $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ are the same and $\varphi^2 \sim \psi^2$.
2. *In all other cases, $\mathbb{Z}_{-\beta}$ contains distances > 1.*

In order to provide similar characterization for cubic Pisot units, and later for the general result in Theorem 3, it is useful to consider the following lemma.

Lemma 1. *Let $\beta > 1$, denote $m = \lfloor \beta \rfloor$. Then $\Delta'_1 \leq 1$ implies that*

$$d_{-\beta}(\ell) = m0^{2k-1}l_{2k+1}l_{2k+2}\cdots(l_{2k+1} > 0) \quad or \quad d_{-\beta}(\ell) = m0^{\omega}\,.$$

Note that the case $d_{-\beta}(\ell) = m0^{\omega}$ happens if β is a root of $x^2 - mx - m$, which is treated in Proposition 1. In the following proposition we use results of Akiyama [1] giving characterization of cubic Pisot units.

Proposition 2. *Let $\beta > 1$ be a cubic Pisot unit with minimal polynomial $p(x) = x^3 - ax^2 - bx - c$, $c = \pm 1$.*

1. *If $p(x) = x^3 - mx^2 - mx - 1$, $m \geq 1$:*
   the distances in $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ are the same and $\varphi^2 \sim \psi^2$.

2. *If $p(x) = x^3 - mx^2 + x - 1$, $m \geq 2$ or $p(x) = x^3 - mx^2 + 1$, $m \geq 3$:*
   the distances in $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ are the same but $\varphi^2 \nsim \psi^2$.

3. *In all other cases, $\mathbb{Z}_{-\beta}$ contains distances > 1.*

4.2 Main Result

Theorem 3. *Let $\beta > 1$. Morphisms φ^2 and ψ^2 generating $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ respectively are conjugated iff β is a Pisot number with minimal polynomial $x^k - m(x^{k-1} + \ldots + x) - n$ with $m \geq n \geq 1$, such that k is odd or $m = n$.*

Remark 2. *Let β be a Pisot number of even degree $k \geq 2$ with minimal polynomial $p(x) = x^k - m(x^{k-1} + \ldots + x) - n$, $m > n \geq 1$. The sets of distances in $\mathbb{Z}_\beta^+$ and $\mathbb{Z}_{-\beta}$ do not coincide, hence $\varphi^2 \nsim \psi^2$. Nevertheless, certain level of similarity can still be found, as was observed for $k = 2$ in [11]. Recall that $v_{-\beta}$ is an infinite word coding $\mathbb{Z}_{-\beta}$. If we take the longer distance in $\mathbb{Z}_{-\beta}$ and "cut" it into $\Delta'_{k-1} = 1 + \frac{n}{\beta} = \Delta_0 + \Delta_{k-1}$, we are in fact applying a morphism $\pi(i) : \{0, \ldots, k-1\}^* \to \{0, \ldots, k-1\}^*$ on $v_{-\beta}$, where*

$$\pi(i) = \begin{cases} i & \text{if } i \in \{0, \ldots, k-2\}, \\ 0(k-1) & \text{if } i = k-1. \end{cases}$$

Then one could use similar approach as in previous examples and as in [11] to verify that the words u_β and $\pi(v_{-\beta})$ have the same language. For u_β is a fixed point of φ^2, $\pi(v_{-\beta})$ is a fixed point of ψ'^2 (which is the unique morphism for which $\pi \circ \psi = \psi' \circ \pi$) and $\varphi \sim \psi'^2$.

References

[1] S. Akiyama, *Pisot units with finite beta expansions*, Algebraic Number Theory and Diophantine Analysis, de Gruyter (2000), 11–26.

[2] P. Ambrož, D. Dombek, Z. Masáková, E. Pelantová, *Numbers with integer expansion in the numeration system with negative base*, Funct. Approx. Comment. Math. 47 (2012), 241–266. doi:10.7169/facm/2012.47.2.8.

[3] K. Dajani, S. D. Ramawadh, *Symbolic Dynamics of* $(-\beta)$*-Expansions*, Journal of Integer Sequences 15 (2012), Article 12.2.6.

[4] D. Dombek, *Generating* $(\pm\beta)$*-integers by Conjugated Morphisms*, Local Proceedings of WORDS 2013, TUCS Lecture Notes 20 (2013), 14–25.

[5] S. Fabre, *Substitutions et* β*-systèmes de numération*, Theoret. Comput. Sci. 137 (1995), 219–236. doi:10.1016/0304-3975(95)91132-A.

[6] Ch. Frougny and A.C. Lai, *Negative bases and automata*, Discrete Math. Theor. Comput. Sci., 13 (1) (2011), 75–94.

[7] S. Ito and T. Sadahiro, *Beta-expansions with negative bases*, INTEGERS 9 (2009), 239–259. doi:10.1515/INTEG.2009.023.

[8] Ch. Kalle, *Isomorphisms between positive and negative beta-transformations*, to appear in Ergodic Theory Dynam. Systems.

[9] L. Liao and W. Steiner, *Dynamical properties of the negative beta-transformation*, Ergodic Theory Dyn. Syst. 32 (2012), 1673–1690. doi:10.1017/S0143385711000514.

[10] Z. Masáková, E. Pelantová, T. Vávra, *Arithmetics in number systems with negative base*, Theor. Comp. Sci. 412 (2011), 835–845.

[11] Z. Masáková, T. Vávra, *Integers in number systems with positive and negative quadratic Pisot base*, preprint (2013), 20pp. arXiv:1302.4655

[12] W. Parry, *On the* β*-expansions of real numbers*, Acta Math. Acad. Sci. Hung. 11 (1960), 401–416.

[13] A. Rényi, *Representations for real numbers and their ergodic properties*, Acta Math. Acad. Sci. Hung. 8 (1957), 477–493.

[14] K. Schmidt, *On periodic expansions of Pisot numbers and Salem numbers*, Bull. London Math. Soc. 12 (1980), 269–278.

[15] W. Steiner, *On the structure of* $(-\beta)$*-integers*, RAIRO - Theor. Inform. Appl. 46 (2012), 181–200. doi:10.1051/ita/2011115.

[16] W. P. Thurston, *Groups, tilings, and finite state automata*, AMS Colloquium Lecture Notes, American Mathematical Society, Boulder (1989).

Time Reversal Based NEWS Methods for Localization and Classification of Nonlinear Vibrations in Bio-mechanical Media*

Zuzana Dvořáková

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

advisors:

Václav Kůs, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Zdeněk Převorovský, Institute of Thermomechanics, AS CR

Serge Dos Santos, UMR-S930 "Imaging and Brain" Inserm, Université François Rabelais, Tours, France

Abstract. Non Destructive Testing (NDT) and harmonic medical imaging methods have been widely developed thanks to the use of the symbiosis of Time Reversal (TR) based signal processing tools and Nonlinear Elastic Wave Spectroscopy (NEWS) methods. Improvement of TR-NEWS has been conducted with coded excitation using chirp frequency excitation and the concept was presented and validated in the context of NDE imaging. The chirp-coded TR-NEWS method uses TR for the focusing of the broadband acoustic chirp-coded excitation. The method consist in the successive steps :

- emission of a linear frequency sweep signal (the chirp-coded excitation),
- recording of the response to the emitted signal (the chirp-coded coda),
- computation of the pseudo-impulse response, which is the correlation between the chirp-coded excitation and its response,
- recording of the response to the time-revered emitted pseudo-impulse excitation (chirp-coded TR-NEWS coda).

The resulting responses coming from nonlinearities in material are processed by means of statistical classification methods and signal processing. The classification of nonlinear vibrations in this paper is performed by means of a fuzzy classification method, in which parameters are extracted from the ultrasonic response containing acoustic nonlinearities. Parameters based on ϕ-divergence measure are used in this work. Because ϕ-divergence comes from theory of probability, the spectra are normalized in the sense that sum of the spectrum is always equal to 1. For normalized spectrum S_p the ϕ-divergence D_ϕ is defined as

$$D_\phi = \sum_{i=0}^{l} S_p(i)\phi\left(S_p^{refer}(i)S_p(i)\right), \tag{1}$$

*This work is published in Proceedings of the ICSV20, International Institute of Acoustics and Vibration, 2013, ISSN 2329-3675, ISBN 978-616-551-682-2 and was presented on 20th International Congress on Sound and Vibration (ICSV20), Bangkok, Thailand

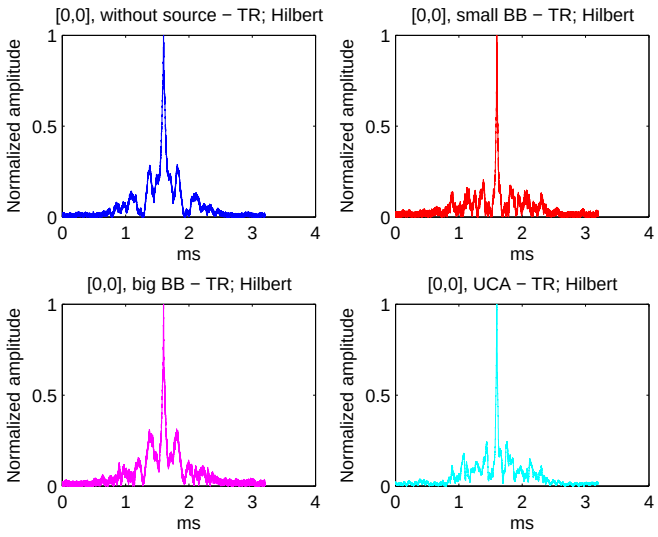


Figure 1: Envelopes of the TR pseudo impulse responses extracted with Hilbert transform. The response for system without source has quite good symmetry with respect to the focusing located at $t_f = 3200/\Delta f = 1.6\mu s$ in comparison with the other signal, where the left part differs from the right part of TR response.

where $\phi : (0, \infty) \rightarrow R$ is convex with $\phi(1) = 0$. There are many possibilities for choosing of fuction ϕ [10], Hellinger divergence [10] is used in this paper. The variable $\tilde{S}^{refer}$ denotes a normalized reference spectrum $\tilde{S}^{refer}(f) = \sum_{i=1}^{m} |\tilde{S}^i(f)|/m$, where m is the number of observations from one type of signals, $\tilde{S}^i(f)$ are individual realizations of the normalized spectrum $\tilde{S}(f)$.

Two experiments are performed to verify suitability of the connection between TR-NEWS process and the classification technique. In the first one, different sources of nonlinearity are measured, analysed and classified. In the second, the same source of nonlinearity is investigated, but in different positions. Consequently, the analysis and classification is conducted in order to reveal different positions by means of the classification. For signal responses (including TR responses), a Hilbert envelop is performed in order to verify better a presence of nonlinearity. Fig.1 shows dissymmetry in Hilbert envelops of TR signal responses when, in the system, either no nonlinearity or different sources of nonlinearity (bubbles or UCA) are presenting. Nonlinearities in the system can be revealed by means of TR-NEWS if the dissymetry is observed. The classification results in experiments were very satisfactory, because only by using simply fuzzy method in combination with parameter ϕ-divergence coming from theory of probability, separation of signals coming from different scatterers or signals of scatter coming from different positions can be done very well. Hence, we are able to decide how many scatterers or positions of scatterer there are in a system. A disadvantage of the fuzzy method is that the number of clusters has to be adjusted previously and in presence of outliers there is possibility of misclassified data. This can be eliminate by analysis of dependency between classification and number of clusters.

Keywords: time reversal, chirp excitation, fuzzy classification, ϕ-divergence

Abstrakt. Nedestruktivní testování (NDT) a lékařské zobrazovací metody dosáhly značného rozvoje díky využití spojení zpracování signálu založeném na časové reverzaci (TR) a metody nelineární elastické vlnové spektroskopie (NEWS). Další zlepšení spojení TR-NEWS spočívá v použití chirp excitace, jenž bylo ověřeno a prezentováno několikrát ve spojení s NDT zobrazováním. Metoda TR-NEWS s chirpově kódovanou excitací se skládá z několika kroků:

- vyslání kmitočtově rozmítaného signálu (chirp excitace),
- nahrání odezvy na vyslaný signál,
- výpočet pseudo-impulsní odezvy, která představuje korelaci mezi chirp excitací a její odezvou,
- nahrání odezvy na vyslanou časově reverzovanou psedo-impulsní excitaci.

Výsledné odezvy pocházející z nelinearit v materiálu jsou zpracovány pomocí statistických klasifikačních metod a metod zpracování signálu. Klasifikace nelineárních vibrací v tomto článku je provedena pomocí metody fuzzy klasifikace, ve které jsou parametry získány z ultrazvukové odezvy obsahující akustické nelinearity. V tomto článku používáme parametry založené na ϕ-divergenční míře mezi spektry. Protože ϕ-divergence pochází z teorie pravděpodobnosti, jsou spektra signálů normovaná v tom smyslu, že suma spektra je vždy rovna 1. Pro normované spektrum S_p je ϕ-divergence D_ϕ definovaná jako

$$D_\phi = \sum_{i=0}^{l} S_p(i)\phi\left(S_p^{refer}(i)S_p(i)\right), \tag{2}$$

kde $\phi : (0, \infty) \to R$ je konvexní a $\phi(1) = 0$. Výběr funkce ϕ je široký [10], v tomto článku byla použita Hellingerova divergence [10]. Proměnná $\tilde{S}^{refer}$ označuje normované referenční spektrum $\tilde{S}^{refer}(f) = \sum_{i=1}^{m} |\tilde{S}^i(f)|/m$, kde m počet pozorování z jednoho typu signálu (nelinearity), $\tilde{S}^i(f)$ jsou jednotlivé realizace normovaného spektra $\tilde{S}(f)$. Pro ověření vhodnosti spojení TR-NEWS metody a klasifikační techniky byly provedeny dva experimenty. V prvním byly naměřeny různé typy nelinearit, které byly analyzovány a klasifikovány. Ve druhém byla měřena ta samá nelinearita, ale v různých pozicích v materiálu. Následně byla provedena analýza a klasifikace pro odhalení různých pozic nelinearity. Aby se lépe určila přítomnost odezvy nelinearity v signálu byly pro signálové odezvy (včetně TR odezev) zkonstruovány Hilbertovy obálky, viz obrázek 1, kde je ukázanána nesymetrie Hilbertových obálek TR signálů v případě žádné nelinearity a různých typů nelinearit (bublinek v kapalině nebo UCA). Nelineratity v systému jsme schopni odhalit pomocí TR-NEWS v případě výskytu jisté nesymetrie. Výsledky klasiface v provedených experimentech byly velmi uspokojivé, protože pouze pomocí jednoduché fuzzy metody v kombinaci s parametrem ϕ-divergence se nám podařilo velmi dobře klasifikovat různé typy nelinearit a rovněž různé pozice konkrétní nelinearity v materiálu. Tudíž jsme schopni rozhodnout na základě využití TR-NEWS a statistické klasifikace o tom, kolik nelinearit je přítomno v materiálu nebo na kolika místech se daná nelinearita vyskytuje. Nevýhodou fuzzy metody je nutnost aprioní znalosti počtu shluků (počtu typů nelinearit či pozic) a další nevýhodou metody je značné ovlivnění výsledků odlehlými pozorováními. Tyto problémy mohou být eliminovány analýzou závisloti mezi klasifikací a počtem shluků.

Klíčová slova: technika časové reverzace, chirp excitace, fuzzy klasifikace, ϕ-divergence

References

[1] Anderson, B. E., Griffa, M., Ulrich, T. J., Johnson, P. A. Time reversal reconstruction of finite sized sources in elastic media, *The Journal of the Acoustical Society of America*, **130** (4), EL219–EL225, (2011).

[2] De Rosny, J., Fink, M. Overcoming the diffraction limit in wave physics using a time-reversal mirror and a novel acoustic sink, *Phys. Rev. Lett.*, **89**, 124301, (2002).

[3] Dos Santos, S., Choi, B., Sutin, A., Sarvazyan, A. Nonlinear imaging based on time reversal acoustic focusing, *Proc. of the 8éme Congrés Francais d'Acoustique*, Tours, France, (2006).

[4] Dos Santos, S., Plag, C. Excitation symmetry analysis method (esam) for calculation of higher order nonlinearities, *Int. Journal of Non-Linear Mechanics*, **43**, 164–169, (2008).

[5] Dos Santos, S., Prevorovsky, Z. Imaging of human tooth using ultrasound based chirp-coded nonlinear time reversal acoustics, *Ultrasonics*, **51** (6), 667–674, (2011).

[6] Dos Santos, S., Vejvodova, S., Prevorovsky, Z. Nonlinear signal processing for ultrasonic imaging of material complexity, *Proceedings of the Estonian Academy of Sciences*, **59**, 301–311, (2010).

[7] Farova, Z., Kus, V., Dos Santos, S. Information-Divergence Based Methods for Acoustic Micro-Defect Identification in Materials, *Proceedings of Forum Acusticum 2011*, Aalborg, Denmark, 26 June–1 July, (2011).

[8] Goursolle, T., Callé, S., Dos Santos, S., Bou Matar, O. A two dimensional pseudospectral model for time reversal (TR) and nonlinear elastic wave spectroscopy (NEWS), *J. Acoust. Soc. Am.*, **122** (6), 3220–3229, (2007).

[9] Guyer, R. A., Johnson, P. A., Nonlinear mesoscopic elasticity: evidence for a new class of materials, *Physics Today*, **52**, 30–36, (1999).

[10] Kus, V., Morales, D., Vajda, I. Extension of the Parametric Families of Divergences Used in Statistical Inference, *Kybernetika*, **44** (1), 95–112, (2008).

[11] Pedrycz, W., *Knoledge-Based Clustering, From Data to Information Granules*, Wiley-Interscience, New Jersey, (2005).

[12] Sarvazyan, A. P., Fillinger, L., Sutin, A. Focusing of broadband acoustic signals using time-reversed acoustics, US Patent 7,587,291 B1, US patent 7,587,291 B1 (2009).

[13] Van Den Abeele, K., Sutin, A., Carmeliet, J., Johnson, P. A. Micro-damage diagnostics using nonlinear elastic wave spectroscopy (news), *NDT E International*, **34**, 239–248, (2001).

Configuration Dynamics Verification Using UPPAAL

David Fabian

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

advisor: Radek Mařík, Department of Cybernetics, Faculty of Electrical Engineering, CTU in Prague

Abstract. Software applications become more and more complicated, nowadays. The complexity of the internal dynamics of a modern software application can be hard to maintain. Software configuration is one of the areas where the internal dynamics can become very complicated because there usually exists a huge amount of states that the system can be in. Of course, implementation of configuration tools is a hard task, especially in imperative-style languages, since the programmer must take into consideration all special combinations of states and implement the appropriate behavior of the program for all of them. It is very easy to make a mistake or to omit a special condition in such a code. There are two ways to solve this problem. One is to use declarative programming which is suitable for these classes of problems (but the programmer must be familiar with an unusual approach of this style of programming) or to use system verifiers to check whether the application behaves correctly under all circumstances.

In [1], a multi-platform configuration tool Freeconf is described. This tool has different types of configuration keys and for each key it uses an extra set of Boolean properties that extend the semantics of the key [3]. The development of values of these semantic properties is highly dynamic since the values change according to the user's actions and one change usually propagates further and induce more changes. Freeconf in its core implements this dynamic behavior in Python. The code is not very maintainable since it is complex and adding more properties or changing some rules of propagation is particularly non-trivial.

In [2], attempts have been made to abstract away from Freeconf and design a formalism that would allow us to describe the general dynamic processes in a compact way and to be able to verify whether the implementation is sound and the model itself does not have any deadlocks. In the paper, the *configuration hierarchical model* is introduced and the propagation dynamics in Freeconf is encoded in it. The model has two parts, one is a description of a static structure of properties that must form a tree and the second is a list of propagation rules which have the form of implications (i.e. condition-action rules). This compact declarative description is then translated to UPPAAL, a powerful model-checking verification software written in Java. UPPAAL expects the to be verified model in the form of a set of finite-state machine automata and provides a GUI for designing them. In section 5 of [2], some of the difficulties and workarounds of this encoding are mentioned. Soundness of the Freeconf instance of the hierarchical model was successfully verified in UPPAAL even though some errors in the propagation rules were found and corrected rules were proposed. UPPAAL, however, turned out not to be the best for verifying the configuration hierarchical model because of its visual modeling. In the future, the better way seems to be to use the Spin verifier that uses the Promela language (similar to C) to model the system.

Keywords: software, configuration, hierarchy, model, verification, UPPAAL

Abstrakt. Komplexita softwarových aplikací v současnosti stále roste. Implementace vnitřní dynamiky moderních aplikací se těžko spravuje. Typickým příkladem jsou programy pro softwarovou konfiguraci, ve kterých je vnitřní dynamika zpravidla velmi komplikovaná, neboť systém může mít velké množství různých stavů, mezi kterými přechází. Programování takových nástrojů je netriviální, zvláště v imperativních jazycích, protože programátor musí vzít v úvahu všechny okrajové stavy systému a implementovat chování aplikace pro každý z nich. Je velmi snadné udělat v takovém kódu chybu nebo vynechat některý speciální případ. Existují dva přístupy k řešení této situace. Za prvé je možné programovat tyto nástroje v deklarativních jazycích, které se zvláště hodí pro tyto třídy úloh (i když programátor musí přijmout zvláštní styl takového programování). Za druhé je možné použít tzv. systémové verifikátory (system verifier) pro kontrolu toho, že se aplikace chová správně za všech okolností.

V článku [1] je popsán multiplatformní konfigurační nástroj Freeconf. Tento nástroj používá různé typy konfiguračních klíčů a pro každý z nich udržuje sadu booleovských proměnných sloužících k zachycení sémantiky klíče [3]. Vývoj hodnot těchto proměnných je značně dynamický, neboť závisí na akcích uživatele a jedna změna se často propaguje dále a způsobuje další změny. Jádro Freeconfu implementuje toto dynamické chování v Pythonu. Tato implementace není příliš udržitelná, protože je komplikovaná a např. přidání nové dynamické proměnné nebo změna pravidel propagace je značně netriviální.

V článku [2] je popsán pokus o abstrakci konkrétních dynamických procesů ve Freeconfu a vytvoření formalismu, který by umožnil zapsat v kompaktní podobě obecné dynamické procesy a usnadnil verifikaci implementace systému. Článek zavádí tzv. *konfigurační hierarchický model* a popisuje jeho konkrétní instanci pro popis dynamiky Freeconfu. Model má dvě části, a to popis statické struktury sémantických proměnných, která musí tvořit strom, a seznam propagačních pravidel ve tvaru implikací (neboli pravidla typu podmínka-akce). Tento kompaktní zápis je pak převeden do výkonného verifikátoru modelů UPPAAL, který je napsán v Javě. UPPAAL očekává svůj vstup v podobě množiny konečných stavových automatů a poskytuje GUI pro jejich zadávání. V sekci 5 článku [2] jsou popsány některé problémy při vytváření vstupu UPPAALu. Instance hierarchického modelu pro Freeconf byla poté úspěšně ověřena, i když byly v průběhu nalezeny některé chyby v propagačních pravidlech a byla navržena oprava. UPPAAL se nicméně ukázal jako ne příliš vhodný nástroj pro ověřování konfiguračního hierarchického modelu, neboť vyžaduje vizuální modelování vstupu. V budoucnosti se zdá být výhodnější použití verifikátoru Spin, který očekává vstup v jazyku Promela, který je blízký jazyku C.

Klíčová slova: software, konfigurace, hierarchie, model, ověřování, UPPAAL

References

[1] D. Fabian. Freeconf: A general-purpose multi-platform configuration utility. In 'Doktorandské dny 2012', 21–30. ČVUT v Praze, (2012).

[2] D. Fabian and R. Mařík. Towards a formalism of configuration properties propagation. In '15th International Configuration Workshop', 35–42. École des Mines d'Albi-Carmaux, (2013).

[3] D. Fabian, R. Mařík, and T. Oberhuber. Towards a formalism of configuration properties propagation. In 'ConfWS'12', 15–20. CEUR Workshop Proceedings, (2012).

On Construction of Solutions of Yang-Baxter Equation

Jan Fuksa

2nd 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

Alexey P. Isaev, Bogolyubov Laboratory of Theoretical Physics, JINR Dubna

Abstract. The problem of constructing the $GL(n)$-invariant solutions of Yang–Baxter equation is considered. The degeneration of R-matrices to projectors is used to build new solutions.

Keywords: Yang–Baxter equation, R-matrix, representation, Lie group

Abstrakt. Tento příspěvek se zabývá konstrukcí $GL(n)$-invariantních řešení Yang–Baxterovy rovnice. Degenerace R-matic je využita ke konstrukci nových řešení.

Klíčová slova: Yang–Baxterova rovnice, R-matice, reprezentace, Lieova grupa

1 Introduction

Let $V_i,\ i = 1, 2, 3$ be vector spaces. The Yang-Baxter equation (YBE) is the following equation in the tensor product $V_1 \otimes V_2 \otimes V_3$

$$R_{12}(u)R_{13}(u+v)R_{23}(v) = R_{23}(v)R_{13}(u+v)R_{12}(u). \tag{1}$$

The basic object $R_{ij}(u)$ arising in eq. (1) is a canonical embedding of a parameter-dependent linear operator acting in the tensor product of spaces $V_i \otimes V_j$ into $V_1 \otimes V_2 \otimes V_3$. The parameter $u \in \mathbb{C}$ is called the spectral parameter. The spaces V_i can be of arbitrary dimension. Solutions of the Yang-Baxter equation are called R-matrices.

Let the spaces V_i be modules of representations T_i of a group G. Solutions of (1) are called G-invariant, if the following equality is satisfied

$$T_i(g) \otimes T_j(g)R_{ij}(u) = R_{ij}(u)T_i(g) \otimes T_j(g) \tag{2}$$

for all $g \in G$. It turns out that (2) is a very restrictive constraint on the solutions of (1).

Proposition 1. *Let T_i, T_j are representations of a group G. Let $T_i \otimes T_j$ is a completely reducible representation with a Clebsch-Gordan decomposition containing no multiplicities. Then*

$$R_{ij}(u) = \sum_k \rho_k(u) P_{\Lambda_k}$$

where operators P_{Λ_k} are projectors on the k-th representation in the Clebsch-Gordan decomposition of $T_i \otimes T_j$ and $\rho_k(u)$ are corresponding eigenvalues depending on spectral parameter u.

Proof is a simple application of the Schurr lemmas.

There is an unproved assertion about $\mathfrak{g}$-invariant solutions of the YBE (1) where $\mathfrak{g}$ is a Lie algebra: for each triplet of $\mathfrak{g}$-moduli $V^{\Lambda_a}, V^{\Lambda_b}, V^{\Lambda_c}$ the YBE

$$R_{ab}^{\Lambda_a\Lambda_b}(u)R_{ac}^{\Lambda_a\Lambda_c}(u+v)R_{bc}^{\Lambda_b\Lambda_c}(v) = R_{bc}^{\Lambda_b\Lambda_c}(v)R_{ac}^{\Lambda_a\Lambda_c}(u+v)R_{ab}^{\Lambda_a\Lambda_b}(u)$$

is satisfied and the solution is unique up to a scalar factor.

Throughout this text, we are interested in $GL(n)$-invariant solutions of (1). The group $GL(n)$ has n^2 generators, denoted $e_{\alpha\beta}$, defined as a matrix identity, i.e.

$$(e_{\alpha\beta})^i_j = \delta^i_\alpha \delta_{\beta j} \tag{3}$$

where $\alpha, \beta, i, j = i, \ldots, n$. These generators satisfy the following commutation relation

$$[e_{\alpha\beta}, e_{\mu\nu}] = \delta_{\beta\mu}e_{\alpha\nu} - \delta_{\alpha\nu}e_{\mu\beta}. \tag{4}$$

The fundamental space of $GL(n)$ is $\mathbb{C}^n$.

Let us take the simplest case and solve equation (1) in the tensor product of three fundamental vector spaces $V_1 \otimes V_2 \otimes V_3$. The R-matrix in $V_i \otimes V_j$ is called the fundamental R-matrix and will be denoted as $R^{11}_{ij}(u)$. It can be proved that only two matrices satisfying the invariance condition (2) for $GL(n)$ are the identity matrix I and the permutation matrix P_{ij} which permutes vectors of the u-th and the j-th space in the tensor product $V_i \otimes V_j$. Therefore,

$$R^{11}_{12}(u) = uI + P_{12}. \tag{5}$$

The lower indexes denote which space is dealing with. The upper will be explained below.

Let us denote $E_{\alpha\beta} = T(e_{\alpha\beta})$ a representation of generators $e_{\alpha\beta}$. Then, of course, the operators $E_{\alpha\beta}$ must satisfy the same commutation relation (4).

2 Constructing R-matrices

Remark 1. *Let us explain the notation used throughout the text. A finite-dimensional irreducible representation T^Λ of $GL(n)$ is characterized by its highest weight $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ where λ_i are integers satisfying the dominance condition $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Using this characterization we denote the space corresponding to the representation T^Λ as V^Λ. Upper indexes are reserved to denote which representation is used. Lower indexes will be used for a better orientation on vector spaces, as a kind of coordinates.*

The most important representation is the fundamental one with the carrying space $\mathbb{C}^n$. All the other highest weight representations can be obtained out of it. The highest weight of fundamental representation is $(1, 0, \ldots, 0)$.

We use the following notation: the highest weight $(m, 0, \ldots, 0)$ will be denoted as $m+$ and $(\overbrace{1, 1, \ldots, 1}^{m}, 0, \ldots, 0)$ as $m-$. For the fundamental representation we use $(1, 0, \ldots, 0) \equiv 1+ \equiv 1- \equiv 1$.

As a consequence, a tensor product of k fundamental spaces V^1 can be denoted as $V^{\overbrace{1 \otimes 1 \otimes \cdots \otimes 1}^{k}} \equiv V^{\otimes k}$.

Using this notation we will denote a R-matrix acting in $V^{\Lambda_a} \otimes V^{\Lambda_b}$ as $R_{ab}^{\Lambda_a \Lambda_b}$ where the indices a resp. b denote the first resp. the second space of the tensor product. Moreover, if, for example, $\Lambda_a = 2+$ then index a has two components $a = \{a_1, a_2\}$.

The starting point of our construction will be the R-matrix in the product of two fundamental spaces $V_i \otimes V_j$, $V_i = V_j = V^1 \equiv \mathbb{C}^n$. As mentioned above, it has to be of the form $R_{ij}^{11} = uI + P_{ij}$. This R-matrix is invertible for all $u \in \mathbb{C}$ with two exceptions $u = \pm 1$. In this case R_{ij}^{11} degenerates to a multiple of the projectors $\mathcal{P}^+$ resp. $\mathcal{P}^-$ on the symmetric resp. the antisymmetric subspace of $V \otimes V$

$$R_{ij}^{11}(1) = 2\mathcal{P}^+ = 2\left[\frac{1}{2}(I + P_{12})\right], \quad R_{ij}^{11}(-1) = -2\mathcal{P}^- = -2\left[\frac{1}{2}(I - P_{12})\right].$$

Another solutions of the YBE (1) in bigger spaces than in the fundamental one can be obtained using this degenerative property of R_{ij}^{11} which, as we will see below, can be generalized to more complicated R-matrices.

Let us solve the YBE in the space $V_{12}^{\otimes 2} \otimes V_3 \otimes V_4$, i.e.

$$R_{12,3}^{\otimes 2,1}(u) R_{12,4}^{\otimes 2,1}(u+v) R_{34}^{11}(v) = R_{34}^{11}(v) R_{12,4}^{\otimes 2,1}(u+v) R_{12,3}^{\otimes 2,1}(u). \tag{6}$$

We obtain the solution $R_{12,3}^{\otimes 2,1}(u) = R_{13}^{11}(u) R_{23}^{11}(u)$ and $R_{12,4}^{\otimes 2,1}(u+v) = R_{14}^{11}(u+v) R_{24}^{11}(u+v)$.

As known, the tensor product $V_1 \otimes V_2$ can be decomposed into two subspaces corresponding to two irreducible representations, called the symmetric resp. the antisymmetric. The symmetric is denoted as V^{2+} and the antisymmetric is denoted as V^{2-}. Here we use the notation in accordance with remark 1. Let us denote the projectors projecting on these spaces as $\mathcal{P}^+$ resp. $\mathcal{P}^-$. A simple idea how to obtain R-matrices acting on these spaces is to restrict solutions of equation (6) $R_{12,3}^{\otimes 2,1}(u)$ to the irreducible subspaces V^{2+} resp. V^{2-} of $V^{\otimes 2}$.

Using this idea, the solution of YBE on the space $V_{12}^{2+} \otimes V_3 \otimes V_4$

$$R_{12,3}^{2+,1}(u) R_{12,4}^{2+,1}(u+v) R_{34}^{11}(v) = R_{34}^{11}(v) R_{12,4}^{2+,1}(u+v) R_{12,3}^{2+,1}(u) \tag{7}$$

should be easily obtained by restriction of solution of eq. (6) to its irreducible subspace V^{2+}. But this does not work. A small modification is needed. The solutions are

$$R_{\{12\}3}^{2+,1}(u) = \mathcal{P}_{12} R_{13}^{11}(u+1) R_{23}^{11}(u) \mathcal{P}_{12}$$

with only small change in the argument of R_{13}^{11} where instead of the term u appears $u+1$. This small change in u is, in fact, very important for the success of this construction.

The essence of above mentioned construction is expressed in the following theorem which is, in fact, much more general than special case above.

Theorem 1 ("on reproduction", [2])**.** *Let the YBE's*

$$R_{12}(u) R_{1c}(u+v) R_{2c}(v) = R_{2c}(v) R_{1c}(u+v) R_{12}(u) \tag{8}$$

are satisfied for $c = 3, 4$. Let the matrix $R_{12}(u)$ degenerates at the point $u = x$. Let

$$R_{12}(x)\mathcal{P}_{12} = R_{12}(x),\ \mathcal{P}_{12}^2 = \mathcal{P}_{12},\ R_{12}(x)\mathcal{P}_{12}^{\perp} = 0,\ \mathcal{P}_{12}^{\perp} = I - \mathcal{P}_{12}$$

where $\mathcal{P}_{12}$ is the projector on the complementary space to the kernel of $R_{12}(x)$. Let the following YBE's are satisfied for $a = 1, 2$

$$R_{a3}(u)R_{a4}(u+v)R_{34}(v) = R_{34}(v)R_{a4}(u+v)R_{a3}(u).$$

Then the matrices acting in the spaces $\mathcal{P}_{12}(V_1 \otimes V_2) \otimes V_b$ resp. $\mathcal{P}_{12}^{\perp}(V_1 \otimes V_2) \otimes V_b$, $b = 3, 4$

$$R_{(12),b}(u) = \mathcal{P}_{12}R_{1b}(u+x)R_{2b}(u)\mathcal{P}_{12}$$

resp.

$$R_{<12>,b}(u) = \mathcal{P}_{12}^{\perp}R_{1b}(u+x)R_{2b}(u)\mathcal{P}_{12}^{\perp}$$

satisfy the YBE

$$R_{(12),3}(u)R_{(12),4}(u+v)R_{34}(v) = R_{34}(v)R_{(12),4}(u+v)R_{(12),3}(u)$$

resp.

$$R_{<12>,3}(u)R_{<12>,4}(u+v)R_{34}(v) = R_{34}(v)R_{<12>,4}(u+v)R_{<12>,3}(u).$$

Because R_{12}^{11} degenaretes in $u = 1$ into the projector $\mathcal{P}_{12} = 2\mathcal{P}^+$ we immediately obtain using this theorem solutions for YBE (7) in $V_{12}^{2+} \otimes V_3 \otimes V_4$

$$R_{12,3}^{2+,1}(u) = \mathcal{P}^+R_{13}^{11}(u+1)R_{23}^{11}(u)\mathcal{P}^+$$

and simiralry for $R_{12,4}^{2+,1}(u+v)$. At the same time, we obtain the solution of YBE in the space $V_{12}^{2-} \otimes V_3 \otimes V_4$ because of the fact that the projector $\mathcal{P}^-$ onto V^{2-} is the complement to $\mathcal{P}^+$ in $V \otimes V$. The solution is the following

$$R_{12,3}^{2-,1}(u) = \mathcal{P}^-R_{13}^{11}(u+1)R_{23}^{11}(u)\mathcal{P}^-.$$

2.1 Generalization of the theorem on reproduction

There is a straightforward generalization of the reproduction theorem to the case of a tensor product of several spaces. As we have seen, the R-matrix $R^{11}(\pm 1)$ degenerates into the projectors

$$\mathcal{P}_{12}^{+} = \frac{1}{2}R_{12}^{11}(1), \quad \mathcal{P}_{12}^{-} = -\frac{1}{2}R_{12}^{11}(-1).$$

There is a more general form of this fact which can be proved using induction

$$\mathcal{P}_{1\ldots m+1}^{\pm} = \pm\frac{1}{m+1}\mathcal{P}_{1\ldots m}^{\pm}R_{1,m+1}^{11}(\pm m)\mathcal{P}_{2\ldots m+1}^{\pm}, \tag{9}$$

$$\mathcal{P}_{1\ldots m+1}^{\pm} = (\pm)^m\frac{1}{(m+1)!}R_{m,m+1}^{11}(\pm 1)R_{m-1,m+1}^{11}(\pm 2)\ldots R_{1,m+1}^{11}(\pm m)\mathcal{P}_{1\ldots m}^{\pm}, \tag{10}$$

$$\mathcal{P}_{1\ldots m+1}^{\pm} = (\pm)^{1/2m(m+1)}\prod_{l=1}^{m}\frac{1}{(l+1)!}\prod_{k=1}^{m}\prod_{l=1}^{k}R_{m+1-l,m+2-k}^{11}(\pm(1+l-k)) \tag{11}$$

$$= (\pm)^{1/2m(m+1)}\prod_{l=1}^{m}\frac{1}{(l+1)!}\prod_{k=1}^{m}\prod_{l=k+1}^{m}R_{k,l}^{11}(\pm(l-k)). \tag{12}$$

We can generalize theorem 8 in this form

Theorem 2. *Let the generalization of YBE's* (8) *in* $V^{\otimes m}\otimes V$

$$\prod_{k=1}^{m}\prod_{l=k+1}^{m}R^{11}_{kl}(u_k-u_l)\prod_{j=1}^{m}R^{1,1}_{m+1-j,c}(v+u_j)=\prod_{j=1}^{m}R^{1,1}_{j,c}(v+u_j)\prod_{k=1}^{m}\prod_{l=k+1}^{m}R^{11}_{kl}(u_k-u_l)$$

is satisfied for $c=I,II$. *Here,* $u_1\equiv 0$ *is only an auxiliary variable. Let the matrix* $\mathcal{Q}_{1\dots m}(u_2,\dots,u_m)=\prod_{k=1}^{m}\prod_{l=k+1}^{m}R^{11}_{kl}(u_k-u_l)$ *degenerates at the point* $(u_2,\dots,u_k)=(x_2,\dots,x_k)$

$$\mathcal{Q}_{1\dots m}(x_2,\dots,x_k)\mathcal{P}_{1\dots m}=\mathcal{Q}_{1\dots m}(x_2,\dots,x_k),\quad \mathcal{P}^2_{1\dots m}=\mathcal{P}_{1\dots m},$$
$$\mathcal{Q}_{1\dots m}(x_2,\dots,x_k)\mathcal{P}^{\perp}_{1\dots m}=0,\quad \mathcal{P}^{\perp}_{1\dots m}=I-\mathcal{P}_{1\dots m}$$

where $\mathcal{P}_{1\dots m}$ *is the projector on the complementary space to the kernel of* $\mathcal{Q}_{1\dots m}(x_2,\dots,x_m)$. *Let the following YBE's are satisfied for* $a=1,2,\dots,m$

$$R_{a,I}(u)R_{a,II}(u+v)R_{I,II}(v)=R_{I,II}(v)R_{a,II}(u+v)R_{a,I}(u).$$

Then the matrices acting in the spaces $\mathcal{P}_{1\dots m}(V_1\otimes V_2\otimes\dots\otimes V_m)\otimes V_b$ *resp.* $\mathcal{P}^{\perp}_{1\dots m}(V_1\otimes V_2\otimes\dots\otimes V_m)\otimes V_b$, $b=I,II$,

$$R_{(1\dots m),b}(u)=\mathcal{P}_{1\dots m}\prod_{j=1}^{m}R^{1,1}_{m+1-j,c}(v+x_j)\mathcal{P}_{1\dots m}$$

resp.

$$R_{<1\dots m>,b}(u)=\mathcal{P}^{\perp}_{1\dots m}\prod_{j=1}^{m}R^{1,1}_{m+1-j,c}(v+x_j)\mathcal{P}^{\perp}_{1\dots m}$$

satisfy the YBE

$$R_{(1\dots m),I}(u)R_{(1\dots m),II}(u+v)R_{I,II}(v)=R_{I,II}(v)R_{(1\dots m),II}(u+v)R_{(1\dots m),I}(u)$$

resp.

$$R_{<1\dots m>,I}(u)R_{<1\dots m>,II}(u+v)R_{I,II}(v)=R_{I,II}(v)R_{<1\dots m>,II}(u+v)R_{<1\dots m>,I}(u).$$

After this, we obtain the expression for the R-matrix in $V^{m+}\otimes V$ resp. $V^{m-}\otimes V$

$$R^{m+,1}_{(1,\dots,m),a}(u)=\mathcal{P}^{+}_{1,\dots,m}R^{11}_{1a}(u+m-1)\dots R^{11}_{ma}(u)\mathcal{P}^{+}_{1\dots m},\tag{13}$$
$$R^{m-,1}_{(1,\dots,m),a}(u)=\mathcal{P}^{-}_{1,\dots,m}R^{11}_{1a}(u-(m-1))\dots R^{11}_{ma}(u)\mathcal{P}^{-}_{1\dots m}.\tag{14}$$

2.2 R-matrices of the form $R^{\Lambda,1}$

Proposition 2. *Let* $e_{\alpha\beta}$ *be the generators* (3) *of* $GL(n)$ *in the fundamental representation and* $E_{\alpha\beta}$ *the same generators in an representation* T^{Λ}. *Then the operator* $L^{\Lambda,1}$ *in* $V^{\Lambda}\otimes V$ *defined by*

$$L^{\Lambda,1}(u)\equiv uP_{\Lambda}\otimes I+\sum_{\alpha,\beta=1}^{n}E_{\alpha\beta}\otimes e_{\beta\alpha}\tag{15}$$

satisfies the YBE in $V^{\Lambda}\otimes V\otimes V$ *where* P_{Λ} *is the projector on the space* V^{Λ}.

There is a beautiful statement joining the L-operator (15) with the results (13), (14) obtained above for $\Lambda = m\pm$, cf. [1].

Proposition 3. *If T^Λ is an irreducible representation with the highest weight of the form $\Lambda = m\pm$ then the solution* (15) *coincides up to a scalar factor with the solutions* (13), (14)

$$L^{\Lambda,1}(u) = \prod_{k=1}^{m-1}(u \pm k)^{-1} R^{m\pm,1}(u). \tag{16}$$

2.3 More general

Using the theorem 2, the same construction can be used to obtain $R^{m\pm,n\pm}_{(1\dots m)(1\dots n)}(u)$. In fact, the R-matrix $R^{m\pm,\Lambda}$ for a general representation Λ can be obtained. It is evident that the construction of the theorem 2 is independent on the spaces $V_c,\ c = I, II$, i.e. these spaces can be arbitrary modules of the group $GL(n)$. Therefore, if we take for $V_a = V^\Lambda$ in (13) and (14), we obtain general R-matrices $R^{m+,\Lambda}$ resp. $R^{m-,\Lambda}$

$$R^{m+,\Lambda}_{(1,\dots,m),a}(u) = \mathcal{P}^+_{1,\dots,m} R^{1\Lambda}_{1a}(u+m-1)\dots R^{1\Lambda}_{ma}(u)\mathcal{P}^+_{1\dots m}, \tag{17}$$

$$R^{m-,\Lambda}_{(1,\dots,m),a}(u) = \mathcal{P}^-_{1,\dots,m} R^{1\Lambda}_{1a}(u-(m-1))\dots R^{1\Lambda}_{ma}(u)\mathcal{P}^-_{1\dots m}. \tag{18}$$

If $\Lambda = (n, 0, \dots, 0)$, then using the R-matrix $R^{1,n+}(u)$ (13) we obtain

$$\begin{aligned} R^{m+,n+}_{(1\dots m),(1\dots n)}(u) &= \mathcal{P}^+_{1\dots} R^{1n^+}_{1a}(u+m-1)\dots R^{1n+}_{ma}(u)\mathcal{P}^+_{1\dots m} \\ &= \mathcal{P}^+_{1,\dots,m}\mathcal{P}^+_{1,\dots,n} R^{11}_{11}(u+m+n-2)\dots R^{11}_{1n}(u+n-1) \\ &\quad R^{11}_{21}(u+m+n-3)\dots R^{11}_{2n}(u+n-2) \\ &\quad \vdots \\ &\quad R^{11}_{m1}(u+m-1)\dots R^{11}_{mn}(u)\mathcal{P}^+_{1\dots n}\mathcal{P}^+_{1\dots m} \end{aligned} \tag{19}$$

All the R-matrices $R^{m+,n-}(u), R^{m-,n+}(u), R^{m-,n-}(u)$ can be representaed in terms of fundamental matrices in a similar way as (19).

3 Spectral decomposition of R-matrices

Let us consider arbitrary representations Λ_a, Λ_b of $GL(n)$ and $\Lambda_c = 1$ and assume the existence of corresponding R-matrices. Using explicit form of $R^{\Lambda,1}$ (15) we obtain the following YBE

$$\begin{aligned} &R^{\Lambda_a\Lambda_b}_{ab}(u)\left(u+v+\sum_{i,j=1}^n E^a_{ij}\otimes e_{ji}\right)\left(v+\sum_{i,j=1}^n E^b_{ij}\otimes e_{ji}\right) \\ &= \left(v+\sum_{i,j=1}^n E^b_{ij}\otimes e_{ji}\right)\left(u+v+\sum_{i,j=1}^n E^a_{ij}\otimes e_{ji}\right) R^{\Lambda_a\Lambda_b}_{ab}(u). \end{aligned} \tag{20}$$

Using the following notation

$$d_{ij} = E^a_{ij} - E^b_{ij}, \qquad d^2_{ij} = \sum_{k=1}^{n} (E^a_{ik} - E^b_{ik})(E^a_{kj} - E^b_{kj}), \tag{21}$$

$$C_2(E) = \sum_{i,j=1}^{n} E_{ij} E_{ji} \tag{22}$$

and relation

$$\sum_{k=1}^{n} (E^a_{ik} E^b_{kj} - E^b_{ik} E^a_{kj}) = \frac{1}{4} [C(E^a - E^b), d^2_{ij}], \tag{23}$$

and excluding matrices e_{ij} from YBE (20) we obtain the equation

$$\left[R^{\Lambda_a \Lambda_b}_{ab}(u), u d_{ij} - \frac{1}{2} d^2_{ij} \right] = \frac{1}{4} \left\{ [C(E^a - E^b), d_{ij}], R^{\Lambda_a \Lambda_b}_{ab}(u) \right\} \tag{24}$$

where [,] is a commutator and { , } is an anticommutator.

The group invariance of $R^{\Lambda_a \Lambda_b}_{ab}(u)$ implies the spectral decomposition of the form

$$R^{\Lambda_a \Lambda_b}_{ab}(u) = \sum_k \rho_k(u) P_{\Lambda_k} \tag{25}$$

where P_{Λ_k} is the projector on the space V^{Λ_k} in the Clebsch-Gordan series $V^{\Lambda_a} \otimes V^{\Lambda_b} = \sum_k V^{\Lambda_k}$.

The equations (24) and (25) allow, in principle, to determine the eigenvalues $\rho_k(u)$ up to a scalar factor, cf. [1].

The spectral decomposition of R-matrices is of a great importance for construing solutions of the YBE (1) because of theorem 1, on repreduction, where a partial knowledge of spectral decomposition of R-matrices is necessary. Therefore, if we know a spectral decomposition of some R-matrix, we are able to construct another solutions of YBE in the spaces corresponding to the spectral decomposition.

4 Conclusions

In the paper [1], Kulish, Reshetikhin and Sklyanin were successful in constructing of all $GL(2)$-invariant R-matrices for arbitrary representations Λ_a, Λ_b of $GL(2)$.

In [1, 2] are shown many solutions of $GL(3)$-invariant R-matrices. The authors also mention that $GL(3)$-invariant R-matrices can be constructed for all finite-dimensional irreducible representations but only special cases are shown.

Nevertheless, the method shown in section 2 cannot be directly applied to all representations in the general case $GL(n), n > 3$, because of multiplicities in Clebsch-Gordan series.

The set of relations (24) and (25) imposes big constraints on the eigenvalues of $GL(n)$-invariant R-matrix. The spectral decomposition of $R^{m\pm,n\pm}(u)$ were obtained in [1]. Another results were obtained in [2]. But, for general Λ_a, Λ_b is the system of equations (24) and (25) overdetermined and the question of its consistency is under consideration.

References

[1] P. Kulish, N. Reshetikhin and E. Sklyanin. *Yang-Baxter equation and representation theory.* Lett. Math. Phys., **5** (1981), pp. 393–403.

[2] P. Kulish and N. Reshetikhin. *On GL(3)-invariant solutions of the Yang-Baxter equation and associated quantum systems.* J. Soviet Math., **34** (1986), pp. 1948–1971.

[3] V. Kazakov and P. Vieira, *From Characters to Quantum (Super)Spin Chains via Fusion*, arXiv:0711.2470 [hep-th], April 2008.

Dynamic Textures Modelling with Temporal Mixing Coefficients Approximation

Michal Havlíček*

4th 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. Appearance of many real world materials is not static but changes in time. In case of spatially and temporally homogeneous changes the material can be represented by means of dynamic texture. Dynamic texture modelling is a challenging problem. In this article we present possible solution based on eigen analysis of input data and subsequent processing and modelling of temporal interpolation eigen coefficients using a combination of piecewise linear approximation and normal distribution sampling. The proposed method shows good performance, enables compress significantly the original data and extremely fast synthesis of arbitrarily long extension of the original texture.

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

Abstrakt. Vzhled mnoha skutečných materiálů není statický, ale mění se v čase. V případě prostorově a časově homogenních změn může být materiál reprezentován pomocí dynamické textury. Modelování dynamických textur představuje složitý problém. V tomto článku uvádíme možné řešení založené na vlastní analýze vstupních dat a následném zpracování a modelování časových interpolačních vlastních koeficientů pomocí kombinace po částech lineární aproximace a vzorkování z normálního rozdělení. Navržená metoda dosahuje dobrých výsledků, umožňuje výraznou kompresi původních dat a velmi rychlou syntézu libovolně dlouhého rozšíření původní textury.

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

1 Introduction

Dynamic textures (DT) can be understood as spatially repetitive motion patterns exhibiting homogenous temporal properties. Good examples might be smoke, fire or liquids. Also waving trees or straws or some moving mechanical objects can be considered as dynamic textures. A sequence of images which are called frames is a basic representation of DT. Original data are always represented by finite length sequence. This property may limit the use of DTs in virtual reality systems so temporally unconstrained modelling of DT is an interesting problem in research such as computer vision, pattern recognition and computer graphics.

*Pattern Recognition Department, Institute of Information Theory and Automation, ASCR.

Already published works dealing with DTs can be divided according to the application to: recognition, representation and synthesis [1]. The DT synthesis is apparently the most difficult task and there are only few papers on this topic available [2]. For example: spatio temporal causal auto regressive model [7], auto regressive moving average model applied on responses of dimensionality reduction filter based on singular value decomposition [6], generative mono spectral DT model based on moving object structure modelling and trajectory modelling by means of dictionary containing Gabor bases for particle elements and Fourier bases for wave elements [8], combination of spatial steerable pyramid and temporal wavelet transformation [3]. All of them are limited by time consuming synthesis algorithm. In addition method [7] requires some high level of temporal homogeneity of the input and method [3] is restricted on monospectral DTs.

Another possibility is utilize so called video editing techniques, 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. Several examples: video textures generation based on searching for transition points for looping with additional blending and morphing [5], further extended in [4], or tree structured vector quantization [9]. These techniques are also time demanding, but some of them produce very high visual quality results [9].

The contribution of this paper is to propose straightforward colour DT modelling method with low computational demands enabling extremely fast synthesis of arbitrarily long DT sequence and in addition compression of original data. The method is based on combination of input data dimensionality reduction using eigen analysis and modelling of resulted temporal coefficients by means of combination of piece wise linear interpolation and uncorrelated noise sampling. It was inspired by the method described in [2] and represents interesting alternative.

The rest of paper is organized as follows: Section 2 explains input data dimensionality reduction using eigen analysis, Section 3 describes temporal coefficients modelling, Section 4 deals with DT synthesis, Section 5 presents some achieved results and Section 6 summarizes the article with a discussion.

2 Dynamic Texture Eigen Analysis

The first step is so called normalization of analysed DT in which average frame from all frames in the sequence is computed and then this frame is subtracted from each frame in this sequence. Values corresponding to pixels 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 $\times$ frame height $\times$ number of spectral planes in the frames and t is a number of frames. Then a covariance $(t \times t)$ matrix A is computed as: $A = C^T C$. 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.

Only $k < t$ eigen vectors corresponding to eigen numbers representing the most of the information are saved. The number k can be determined by several techniques. The threshold selecting vectors which are not used may be computed from the values of the eigen numbers. Assuming that the eigen numbers i.e. the elements $D_{(i,i)}$ are ordered by

their value then the threshold δ can be computed as for example:

$$\delta = \frac{1}{t}\sum_{i=1}^{t} D_{(i,i)} \quad or$$

$$\delta = D_{(i,i)} \quad where \quad i = argmin_{j\in\{1,\ldots,k-1\}}(|D_{(j,j)} - D_{(j+1,j+1)}|) \ .$$

Only eigen vectors which fulfill that their corresponding eigen number is higher than the treshold δ are saved. The effects of selecting the threshold δ and therefore the number of preserved vectors k and the other possibilities are further discussed in Section 5 and Section 6.

Eigen images $(n \times k)$ matrix I is computed as: $I = CT$, where T is $(t \times k)$ matrix with elements: $T_{(i,j)} = \frac{U_{(i,j)}}{\sqrt{D_{(j,j)}}}$. Finally a matrix of temporal mixing coefficients of individual eigen images I for all frames from the sequence is computed as: $M = I^T C$. The $(k \times t)$ matrix M is a subject of further processing described in following section.

3 Temporal Mixing Coefficients Processing

A threshold α is computed first: $\alpha = \frac{1}{k}\sum_{i=1}^{k}(\sigma_i)$ where

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}(|M_{(j,i)} - M_{(j,i+1)}| - \mu_j)(|M_{(j,i)} - M_{(j,i+1)}| - \mu_j)} \quad ,$$

$$\mu_j = \frac{1}{n-1}\sum_{i=1}^{n-1}(|M_{(j,i)} - M_{(j,i+1)}|) \ .$$

Then the matrix M is processed following manner: if j-th row of M fulfils $\sigma_j > \alpha$ then mean $\hat{\mu_j}$ and dispersion $\hat{\sigma_j}$ of normal distribution from elements of this row are estimated as:

$$\hat{\mu_j} = \frac{1}{t}\sum_{i=1}^{t} M_{(j,i)} \ , \quad \hat{\sigma_j} = \frac{1}{t}\sum_{i=1}^{t}(M_{(j,i)} - \hat{\mu_j})^2 \ .$$

The row which is under $\sigma_j \leq \alpha$ is disjoint into several sub intervals. We denote the set of the indices representing end points of the rows as L . The right edge i_1 of the block is detected by the threshold μ_j applied to $|M_{(j,i_1)} - M_{(j,i_1+1)}|$ so that at least one row $j_0 \in L$ satisfies $|M_{(j_0,i_1)} - M_{(j_0,i_1+1)}| > \mu_{j_0}$. Then values of $M_{(j,i_0)}$ and $M_{(j,i_1)}\, \forall j \in L$, where i_0 is the left edge of the block, are saved instead of all values in corresponding interval. In addition blocks with less than two elements are not saved at all. The set of all saved blocks will be denoted as B . The division is driven by the row j^* which both fulfils $\sigma_{j^*} \leq \alpha$ and the average value of all elements of this row is the higher than any other such value of the rest of the rows $j \in \hat{k}$ under $\sigma_j \leq \alpha$.

Another possibility is to disjoint rows into the sub intervals with the same length. The length of intervals affects overall dynamics of the synthesized sequence and it appears that each DT need different division to achieve the best result. Although we have not

developed any technique to detect this optimal division yet it is apparent that some semi optimal division sufficient enough exists and it was verified by many experiments that this semi optimal length equals to two percents of the total length of the original sequence.

4 Synthesis

The goal of the synthesis is to create certain number of DT frames so that overall visual appearance is close enough to the original. Unfortunately there does not exist any applicable criterion to decide if the synthesized DT is close enough to the original as explained in Section 5.

During the synthesis a matrix $(k \times t^\dagger)$ of temporal mixing coefficients $M^\dagger$, where $t^\dagger$ is a length of the synthesized sequence, in general different from t, is created block wise from the blocks occurring the set B . Element $M^\dagger_{i,j}$ is linearly interpolated if $j \in L$ or sampled from uncorrelated noise with mean $\hat{\mu_j}$ and dispersion $\hat{\sigma_j}$ otherwise. Blocks may be chosen even non deterministically but $|M_{i_1,j} - M_{i_0,j}| < \mu_j$ must hold for all $j \in L$, i_1 is the right edge of previously used block and i_0 is the left edge of the following one.

New DT sequence $C^\dagger$ which is $(n \times t^\dagger)$ matrix can be then computed simply as: $C^\dagger = M^\dagger U$. Final step is addition of the average frame to each frame in the synthesized sequence. Since only matrix operations occur in this step it can be easily performed on contemporary graphics hardware which considerably increases the synthesis speed.

5 Results

We used the dynamic texture data sets from DynTex texture database [1] as a source of test data. Each dynamic texture from this sets is typically represented by a 250 frames long video sequence, that is equivalent to ten second long video. An analysed DT is processed frame by frame. Each frame is 400×300 RGB colour image. As a test DT were chosen: smoke, steam, streaming water, sea waves, river, candle light, close shot of moving escalator, sheet, waving flag, leaves, straws and branches.

Some results can be seen on Figures 1 and 2, showing selected synthesized frames and corresponding frames from original sequence. In this case the deterministic version of the algorithm with fixed length intervals were use to reproduce the sequence.

From the shown results can be seen that although there are some differences between original and synthesized frames the overall dynamic stayed preserved. Unfortunately it is really hard to express this similarity exactly. Robust and reliable similarity comparison between two static textures is still unsolved problem up to now. Moreover, when we switch to the dynamic textures the complexity of comparison between original and synthesized DT sequence increase even more.

In some cases the synthesized DT is visually similar to the original except for less details (for example: river and straws on Figure 1, sea waves on Figure 2), sometimes the moves in synthesized sequence are blurred (for example: waving leaves and sheet on Figure 2). Less detailed appearance is mainly caused by information loss during the dimensionality reduction phase when only about 15% of the original information is saved.

[1] http://www.cwi.nl/projects/dyntex/

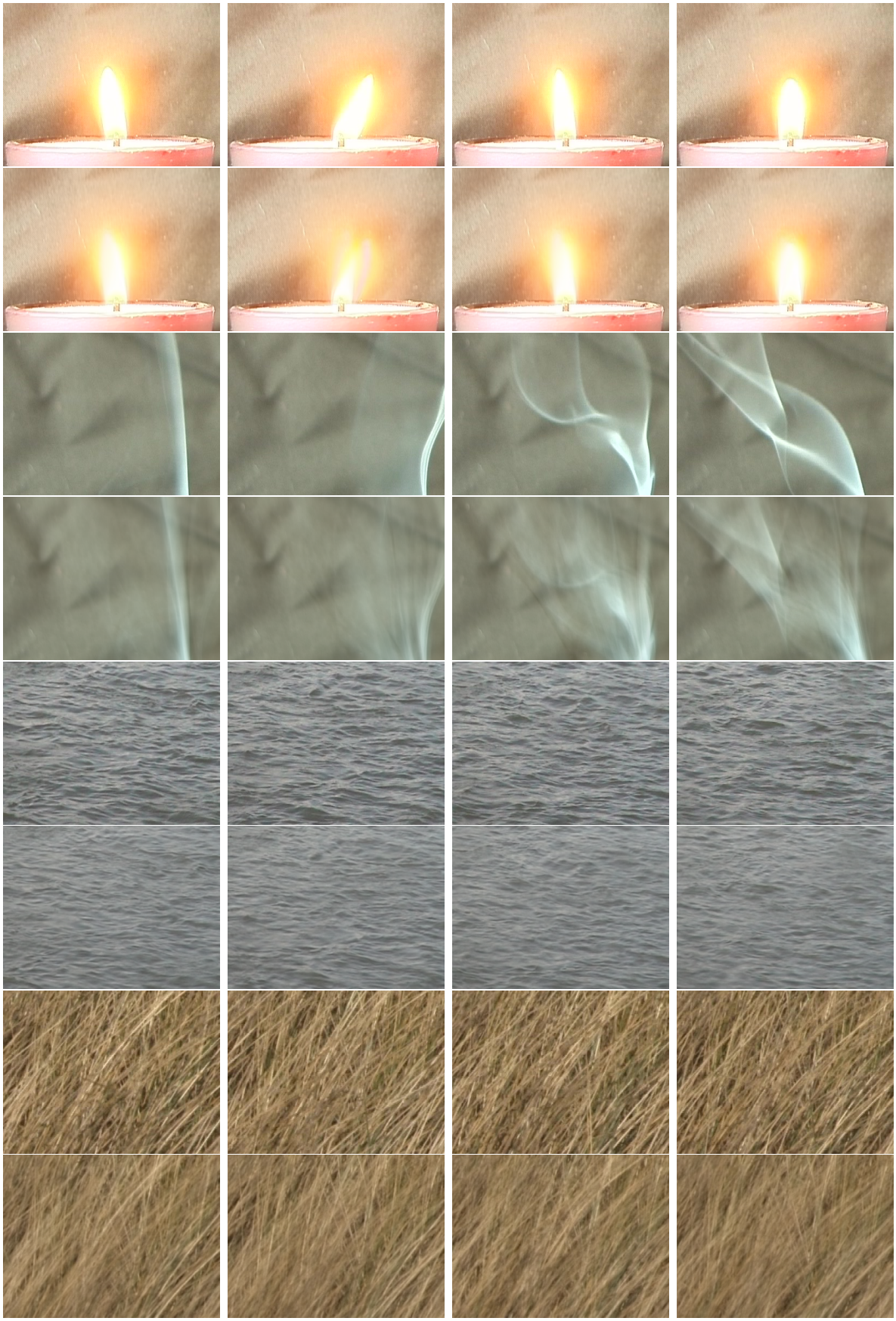


Figure 1: Original frames (odd rows) versus corresponding synthesized ones (even rows), sequences: candle light, smoke, river, straws.

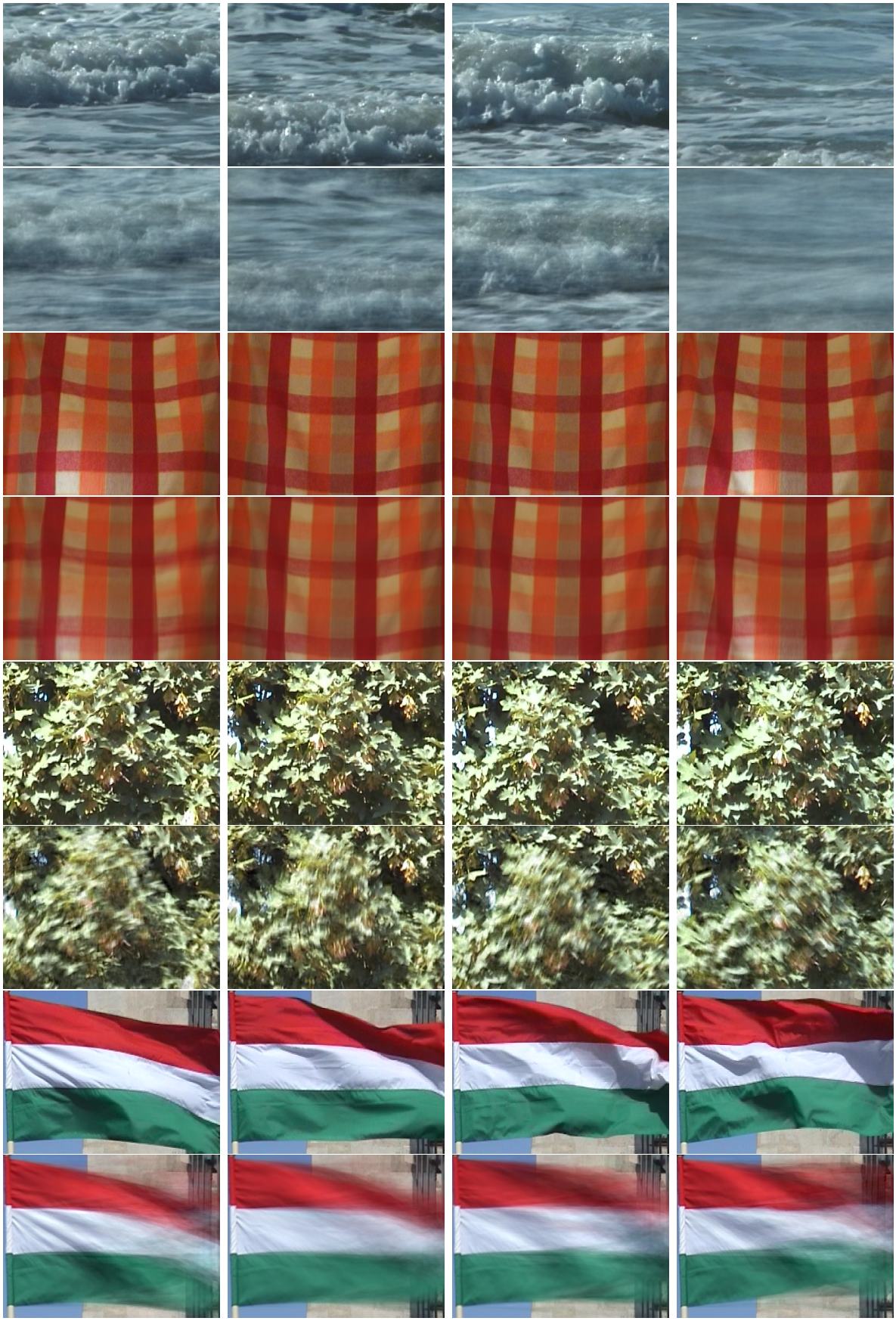


Figure 2: Original frames (odd rows) versus corresponding synthesized ones (even rows), sequences: sea waves, sheet, waving leaves, flag.

Figure 3: The synthesis of several textures (candle light, river, straws and waving leaves) 300th and 400th frames.

The approximation of coefficients is reflected in the blurring. The worst result is the flag sequence synthesis (Figure 2), maybe it is because this is not real DT but rather dynamic scene and this method is limited to DTs.

Main advantage of this method to the solution published in [2], where Causal Auto Regressive (CAR) model is used to process matrix M, is its stability in the synthesis step. Another issue of using CAR model is that the overall dynamics of synthesized sequence decreases with time which is serious problem in case of sequences longer than original one. The general dynamic of the sequence is preserved in time in case of our method as presented on some results on Figure 3 showing selected frames from synthesized sequence longer than original one. The computational demands are identical for both methods.

6 Conclusion and discussion

We presented a novel method for fast synthesis of dynamic multispectral textures in this article. The main part of the approach is based on modelling of temporal coefficients resulted from input data dimensionality reduction step. This solution enables extremely fast synthesis of arbitrary number of multispectral DT frames, which can be even more efficiently performed by contemporary graphical hardware. There are still some unsolved tasks. The detection of optimal number of component which should be saved is still discussed, because this step is essential and affect overall performance and resulting visual quality. The division of temporal matrix is not always the best solution and sometimes the fixed length sub intervals serves as the universal semi optimal solution. On the other we have not developed any method for optimal fixed sub interval length detection yet but many experiments demonstrated that for most DTs 2% of the total length of the sequence is sufficient. Although this method is still under development it represents interesting alternative to the existing approaches.

References

[1] D. Chetverikov, 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.

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

[3] Z. Joseph, R. El-Yaniv, D. Lischinski, M. Werman. *Texture mixing and texture movie synthesis using statistical learning.* IEEE Transactions on Visualization and Computer Graphics, **7**(2), 2001, 120–135.

[4] V. Kwatra, A. Schödl, I. Essa, A. Bobick. *Graphcut textures: image and video synthesis using graph cuts.* ACM SIGGRAPH 2003, ACM Press, **22**(2), 2003, 277–286.

[5] A. Schödl, R. Szeliski, D. Salesin, I. Essa. ACM SIGGRAPH 2000, ACM Press, 489–498.

[6] S. Soatto, G. Doretto, Y. Wu. *Dynamic textures.* In ICCV, 2001, 439–446.

[7] M. Szummer and R. Picard. *2Temporal texture modeling.* In Proceedings of IEEE International Conference on Image Processing (ICIP 96), 1996.

[8] Y. Wang and S. Zhu. *Analysis and synthesis of textured motion: Particle, wave and cartoon sketch.* IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2004.

[9] L. Wei, M. Levoy. *Fast texture synthesis using treestructured vector quantization.* ACM SIGGRAPH 2000, ACM Press, pages 479–488.

Delone Characteristics of Spectra of Cubic Complex Pisot Units*

Tomáš Hejda†

2nd year of PGS, email: tohecz@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Edita Pelantová, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. For $q \in \mathbb{R}$, $q > 1$, Erdős, Joó and Komornik study distances of the consecutive points in the set

$$X^m(q) = \left\{ \sum_{j=0}^{n} a_j \beta^j : n \in \mathbb{N},\ a_k \in \{0, 1, \ldots, m\} \right\}.$$

The Pisot numbers play a crucial role for properties of $X^m(q)$.

We follow work of Zaïmi who consideres $X^m(\gamma)$ with γ being a complex Pisot number. For a class of cubic complex Pisot units we show that $X^m(\gamma)$ is a Delone set in the plane $\mathbb{C}$ and for γ the complex root of $Y^3 + Y^2 + Y - 1$ we determine two parameters of the Delone set $X^m(\gamma)$ which are analogous to minimal and maximal distance for the real case $X^m(q)$.

Keywords: beta-numeration, Delone set, cut-and-project scheme

Abstrakt. Erdős, Joó a Komornik studují, pro zadané $q \in \mathbb{R}$, $q > 1$, mezery mezi sousedy v množině

$$X^m(q) = \left\{ \sum_{j=0}^{n} a_j \beta^j : n \in \mathbb{N},\ a_k \in \{0, 1, \ldots, m\} \right\}.$$

Ukazuje se, že pisotovskost q má zásadní vliv na vlastnosti množiny $X^m(q)$.

Navazujeme na práci Zaïmiho, který studuje množinu $X^m(\gamma)$ pro complexní pisotovské číslo γ. Pro jistou třídu kubických complexních pisotovských čísel jsme ukázali, že $X^m(\gamma)$ je delonovská množina v $\mathbb{C}$. Pro komplexní kořen polynomu $Y^3 + Y^2 + Y - 1$ jsme určili dva parametry delonovské množiny $X^m(\gamma)$, které jsou obdobou minimální a maximální mezery v reálném případě $X^m(q)$.

Klíčová slova: beta-numerace, delonovská množina, průměty mřížky

1 Introduction

In articles [EJK, EJK′], Erdős, Joó and Komornik study the set

$$X^m(\beta) \coloneqq \left\{ \sum_{j=0}^{n} a_j \beta^j : n \in \mathbb{N},\ a_k \in \{0, 1, \ldots, m\} \right\},$$

*This is an extract of the work [HP].

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

where $\beta > 1$. Clearly this set has no accumulation points, hence we can find an increasing sequence

$$0 = x_0 < x_1 < x_2 < \cdots < x_k < \cdots$$

such that $X^m(\beta) = \{x_k : k \in \mathbb{N}\}$. Their research aims to describe distances between consecutive points of $X^m(\beta)$, i.e. the sequence $(x_{k+1} - x_k)_{k\in\mathbb{N}}$. Properties of this sequence depend on the value $m \in \mathbb{N}$. It is easy to show that when $m \geq \beta - 1$, we have $x_{k+1} - x_k \leq 1$; and when $m < \beta - 1$, the distances $x_{k+1} - x_k$ can be arbitrarily large.

Properties of $X^m(\beta)$ are dependent on β being a *Pisot number* (i.e. an algebraic integer > 1 such that all its Galois conjugates are in modulus < 1). Bugeaud [B] showed that

$$\ell_m(\beta) \coloneqq \liminf_{k\to\infty}(x_{k+1} - x_k) > 0$$

for all $m \in \mathbb{N}$ if and only if the base β is a Pisot number. Recently, Feng [F] proved a stronger result that the bound $\beta - 1$ for the alphabet size is crucial. In particular, $\ell_m(\beta) = 0$ if and only if $m > \beta - 1$ and β is not a Pisot number.

Therefore, let us focus on the case β Pisot and $m > \beta - 1$. From the approximation property of Pisot numbers we know that for a fixed β and $m > \beta - 1$ the sequence $(x_{k+1} - x_k)$ takes only finitely many values. Feng and Weng [FW] used this fact to show that the sequence of distances $(x_{k+1} - x_k)$ is substitutive, roughly speaking, can be generated by a system of rewriting rules over a finite alphabet. This allows, for a fixed β and m, to determine values of all distances $(x_{k+1} - x_k)$. An algorithm for obtaining the minimal distance $\ell_m(\beta)$ was as well proposed by Borwein and Hare [BH].

The first formula which determines the value of $\ell_m(\beta)$ for all m at once appeared in 2000 where Komornik, Loreti and Pedicini [KLP] study the base golden mean. The generalization of this result to all quadratic Pisot units was provided by Takao Komatsu [K] in 2002.

Zaïmi [Z] started to study the set $X^m(\gamma)$ where he considered γ a complex number of modulus > 1, and he put

$$\ell_m(\gamma) \coloneqq \inf\Big\{|x - y| : x \neq y,\ x, y \in X^m(\gamma)\Big\}.$$

He proved an analogous result to the one for real bases by Bugeaud, namely that $\ell_m(\gamma) > 0$ for all m if and only if γ is a *complex Pisot number*, which is defined as a non-real algebraic integer of modulus > 1 whose Galois conjugates except its complex conjugate are of modulus < 1.

To study $X^m(\gamma)$ in $\mathbb{C}$, we need to define characteristics analogous to $\ell_m(\beta)$ and $L_m(\beta)$ for the real case. Let us inspire by the notions used in the definition of Delone sets.

We say that a set Σ is: *uniformly discrete* if there exists $d > 0$ such that $|x - y| > d$ for all distinct $x, y \in \Sigma$; *relatively dense* if there exists $D > 0$ such that every ball $B(x, D/2)$ of radius $D/2$ contains a point from Σ; and *Delone* if it is both uniformly discrete and relatively dense.

Clearly, if $\ell_m(\gamma)$ is positive, then $X^m(\gamma)$ is uniformly discrete and $\ell_m(\gamma)$ is the minimal d in the definition of uniform discreteness.

Let us define

$$L_m(\gamma) \coloneqq \inf\Big\{D > 0 : B(x, D/2) \cap X^m(\gamma) \neq \emptyset \text{ for all } x \in \mathbb{C}\Big\}.$$

In particular, $L_m(\gamma) = +\infty$ if and only if $X^m(\gamma)$ is not relatively dense.

The question for which pairs (γ, m) the set $X^m(\gamma)$ is uniformly discrete, and for which (γ, m) it is relatively dense is far from being solved. It is not even clear what maximal allowed digit m ensures the relative denseness.

The aim of this work is to study the sets $X^m(\gamma)$ simultaneously for all $m \in \mathbb{N}$, for a certain class of cubic complex Pisot numbers with a positive conjugate γ'. For such γ the Rényi expansions in the base $1/\gamma'$ have nice properties which will be crucial in the proofs. When the base $1/\gamma' \in \mathbb{R}$ has so-called Property (F), we show that $X^m(\gamma) \subseteq \mathbb{C}$ is a cut-and-project set. Roughly speaking, $X^m(\gamma)$ is formed by projections of points from the lattice $\mathbb{Z}^3$ which lie in a sector bounded by two parallel planes in $\mathbb{R}^3$, see Theorem 3.1. From that, we deduce the asymptotic behaviour of $\ell_m(\gamma)$ and $L_m(\gamma)$:

Theorem 1.1. *Let γ be a cubic complex Pisot unit such that it has a positive real conjugate γ', whose inverse $1/\gamma'$ has Property (F). Then*

$$\ell_m(\gamma) = \Theta(\sqrt{m}) \quad \textit{and} \quad L_m(\gamma) = \Theta(\sqrt{m}).$$

The method of inspection of Voronoi cells for a specific cut-and-project set, as established by Masáková, Patera and Zich [MPZ, MPZ′, MPZ″], enables us to give a general formula for both $\ell_m(\gamma)$ and $L_m(\gamma)$. In the case that γ is the complex Tribonacci constant, i.e. the complex root of $Y^3 + Y^2 + Y - 1$, we get:

Theorem 1.2. *Let $\gamma = \gamma_T \approx -0.771 + 1.115\mathrm{i}$ be the complex root of the polynomial $Y^3 + Y^2 + Y - 1$. Let $m \in \mathbb{N}$. Find a maximal $k \in \mathbb{Z}$ such that $m \geq (1 - \gamma')\left(\frac{1}{\gamma'}\right)^k$, where γ' is the real Galois conjugate of γ. Then we have*

$$\ell_m(\gamma) = |\gamma|^{-k} \quad \textit{and} \quad L_m(\gamma) = A|\gamma|^{-k}, \quad \textit{where } A = 2\sqrt{\frac{1 - (\gamma')^2}{3 - (\gamma')^2}}.$$

The article is organized as follows. In the preliminaries, we recall certain notions from algebraic number theory. In section 3 we prove that $X^m(\gamma)$ is a cut-and-project set in certain cases. In section 4 we provide algorithms for computing $\ell_m(\gamma)$ and $L_m(\gamma)$, and we prove Theorem 1.1. These algorithms are applied on the complex Tribonacci number in section 5, providing the proof of Theorem 1.2. The conclusions are in section 6. We should remark that we omit proofs of all statements.

2 Preliminaries

We will widely use the algebraic properties of cubic complex Pisot numbers γ. Such γ has two Galois conjugates. One of them is the complex conjugate $\overline{\gamma}$. The second one is real and in modulus < 1, we will denote it γ'; we have either $-1 < \gamma' < 0$ or $0 < \gamma' < 1$. In general, for $z \in \mathbb{Q}(\gamma)$ we denote by $z' \in \mathbb{R}$ its image under the Galois isomorphism that maps $\gamma \mapsto \gamma'$.

As usual, we denote $\mathbb{Z}[\gamma]$ the set of integer combinations of positive powers of γ. When γ is a unit (i.e. the absolute term of its minimal polynomial is ± 1), we know that $\mathbb{Z}[1/\gamma] = \mathbb{Z}[\gamma] = \gamma\mathbb{Z}[\gamma]$.

We use some notions from β-expansions. For a real base $\beta > 1$, and for a number $x \geq 0$, there exist unique integer coefficients $a_N, a_{N-1}, a_{N-2}, \ldots$ such that

$$0 \leq x - \sum_{j=n}^{N} a_j \beta^j < \beta^n \quad \text{for all } n \leq N$$

(unique up to leading zeros). Then the string $a_N a_{N-1} \cdots a_1 a_0 . a_{-1} a_{-2} \cdots$ is called the *Rényi expansion* of x in the base β. If only finitely many a_j's are non-zero, we speak about *finite Rényi expansions*. The set of numbers $\pm x$ such that x has finite Rényi expansion is denoted $\mathrm{Fin}(\beta)$. We say that $\beta > 1$ satisfies the *Property (F)* if $\mathrm{Fin}(\beta)$ is an algebraic ring, i.e. $\mathrm{Fin}(\beta) = \mathbb{Z}[\beta] + \mathbb{Z}[1/\beta]$.

Akiyama [A] described the real cubic units having Property (F) in terms of the coefficients of the minimal polynomial. From this result, and using Cardano's formula to determine whether a cubic polynomial has complex roots, we can deduce that a non-real γ satisfies the hypothesis of Theorem 3.1 if and only if

$$\gamma^3 + b\gamma^2 + a\gamma - 1 = 0, \quad \text{where } a, b \in \mathbb{Z} \text{ satisfy:}$$
$$a \geq 0, \quad -1 \leq b \leq a + 1 \quad \text{and} \quad 18ab + 4a^3 - a^2b^2 - 4b^3 + 27 > 0. \quad (2.1)$$

In particular, the *complex Tribonacci constant* $\gamma_T \approx -0.771 + 1.115\mathrm{i}$ (the root of $Y^3 + Y^2 + Y - 1$) and the *minimal cubic complex Pisot unit* $\gamma_M \approx -0.877 + 0.744\mathrm{i}$ (the root of $Y^3 + Y^2 - 1$) fall into this scheme. More generally, for all $a \geq 0$ and $b = -1, 0, 1$ the polynomial $Y^3 + bY^3 + aY - 1$ is good.

3 Cut-and-project sets versus $X^m(\gamma)$

A cut-and-project scheme in dimension $d + e$ comprises two linear maps $\Psi : \mathbb{R}^{d+e} \to \mathbb{R}^d$ and $\Phi : \mathbb{R}^{d+e} \to \mathbb{R}^e$ satisfying that $\Psi(\mathbb{R}^{d+e}) = \mathbb{R}^d$ and restriction of Ψ to the lattice $\mathbb{Z}^{d+e}$ is injective and the set $\Phi(\mathbb{Z}^{d+e})$ is dense in $\mathbb{R}^e$.

Let $\Omega \subset \mathbb{R}^e$ be a nonempty bounded set such that its closure equals the closure of its interior, i.e. $\overline{\Omega} = \overline{\Omega^\circ}$. Then the set

$$\Sigma(\Omega) \coloneqq \left\{ \Psi(v) : v \in \mathbb{Z}^{d+e}, \Phi(v) \in \Omega \right\} \subseteq \mathbb{R}^d$$

is called *cut-and-project* set with the acceptance window Ω. Cut-and-project sets can be defined in a slightly more general way, c.f. [M].

It is well known that $\Sigma(\Omega)$ is a Delone set with finite local complexity. Morover, in case $e = 1$, the form of acceptance window $\Omega = [l, r)$ or $\Omega = (l, r]$ guarantees that $\Sigma(\Omega)$ is repetitive, i.e. for every $x \in \Sigma(\Omega)$ and $\varrho > 0$ the patch $(\Sigma(\Omega) - x) \cap B(0, \varrho)$ occurs infinitely many times in $\Sigma(\Omega)$.

We will use the concept of cut-and-project sets for $d = 2$ and $e = 1$. With a slight abuse of notation, we will consider $\Psi : \mathbb{R}^3 \to \mathbb{C} \simeq \mathbb{R}^2$. Then it is straightforward that for a cubic complex number γ, the set defined by

$$\Sigma_\gamma(\Omega) = \left\{ z \in \mathbb{Z}[\gamma] : z' \in \Omega \right\}, \quad \text{where } \Omega \subseteq \mathbb{R} \text{ is an interval,} \quad (3.1)$$

is a cut-and-project set; really, we have

$$\Psi_\gamma(v_0,v_1,v_2) = v_0 + v_1\gamma + v_2\gamma^2 \quad \text{and} \quad \Phi_\gamma(v_0,v_1,v_2) = v_0 + v_1\gamma' + v_2(\gamma')^2.$$

We will often omit the index γ in the sequel.

The set $X^m(\gamma)$ is described in terms of algebra, whereas the set $\Sigma(\Omega)$ has a geometric description. We show that in certain cases, these sets coincide:

Theorem 3.1. *Let γ be a cubic complex Pisot unit with a positive conjugate $0 < \gamma' < 1$. Suppose that $1/\gamma'$ has the Property (F). Let $m \geq \gamma\overline{\gamma} - 1$ be an integer. Then $X^m(\gamma)$ is a cut-and-project set, namely*

$$X^m(\gamma) = \Sigma(\Omega) = \Big\{ z \in \mathbb{Z}[\gamma] : z' \in \Omega \Big\} \quad \textit{with} \quad \Omega = \Big[0, m/(1-\gamma')\Big). \tag{3.2}$$

In general, the cut-and-project sets are not self-similar. However, in our special case (3.1), we can prove a nice self-similarity property that will be useful later:

Proposition 3.2. *Let γ be a cubic non-real unit. Then we have*

$$\Sigma\Big((\gamma')^k\Omega\Big) = \gamma^k\Sigma(\Omega) \quad \textit{for any interval } \Omega \textit{ and any } k \in \mathbb{Z}.$$

4 Voronoi tessellation of $X^m(\gamma)$

In a Delone set Σ, the *Voronoi cell* of a point $x \in \Sigma$ is the set of points that are closer to x than to any other point in Σ, formally

$$\mathcal{T}(x) = \Big\{ z \in \mathbb{C} : |z-x| \leq |z-y| \text{ for all } y \in \Sigma \Big\}.$$

The cell is a convex polygon having x as an interior point. For every cell $\mathcal{T}(x)$ we define two characteristics:

- $\delta(\mathcal{T}(x))$ is the maximal diameter $d > 0$ such that $B(x, d/2) \subseteq \mathcal{T}(x)$;
- $\Delta(\mathcal{T}(x))$ is the minimal diameter $D > 0$ such that $\mathcal{T}(x) \subseteq B(x, D/2)$.

These δ and Δ allow us to compute the values of $\ell_m(\gamma)$ and $L_m(\gamma)$, namely

$$\ell_m(\gamma) = \inf_x \delta\Big(\mathcal{T}(x)\Big) \quad \text{and} \quad L_m(\gamma) = \sup_x \Delta\Big(\mathcal{T}(x)\Big), \tag{4.1}$$

where x runs the whole set $X^m(\gamma)$.

A *protocell* of a point x is the set $\mathcal{T}(x) - x$. We can define δ, Δ analogously for the protocells. The set of all protocells of the tessellation of $\Sigma(\Omega)$ is called *palette* of $\Sigma(\Omega)$ and is denoted $\mathrm{Pal}(\Omega)$.

Cut-and-project sets have finite local complexity. This means there are only finitely many protocells, i.e. the palette is finite. For any $y \in \Sigma(\Omega)$, the local configuration of size L around y is

$$\Sigma(\Omega) \cap B(y, L) = y + \Sigma(\Omega - y') \cap B(0, L). \tag{4.2}$$

Therefore, there exists $L > 0$ such that if $\Sigma(\Omega - y_1') \cap B(0, L) = \Sigma(\Omega - y_2') \cap B(0, L)$, then the protocells of y_1 and y_2 are identical. From the theory of Voronoi tessellations we know that $L = \sup_{x \in \Sigma(\Omega)} \Delta(\mathcal{T}(x))$ is a good estimate. In the rest of the section, we will consider L with such property.

Since $\Sigma(\Omega)$ is repetitive in our case, we have that $\ell_m(\gamma) = \delta(\mathcal{T}(x))$ for infinitely many $x \in X^m(\gamma)$, and $L_m(\gamma) = \Delta(\mathcal{T}(x))$ for infinitely many $x \in X^m(\gamma)$.

The algorithm to compute all protocells of the set $\Sigma(\Omega)$ for $\Omega = [0, c)$ is based on the following claim about them. Not only that the palette is final, we are even able to arrange the points of $\Sigma(\Omega)$ by their protocell:

Lemma 4.1. *Let $\Omega = [0, c)$ be an interval. Then there exists a finite set $\Xi = \{\xi_0 < \xi_1 < \cdots < \xi_{N-1}\} \subset (0, c)$ such that the protocell of $y \in \Sigma(\Omega)$ as a function*

$$[0, c) \cap \mathbb{Z}[\gamma'] \to \mathrm{Pal}(\Omega), \qquad y' \mapsto \mathcal{T}(y) - y$$

is constant on each of the intervals $[0, \xi_0), [\xi_0, \xi_1), \ldots, [\xi_{N-2}, \xi_{N-1}), [\xi_{N-1}, c)$.

The proof is constructive and gives

$$\Xi \coloneqq \Big(\big\{x' : x \in \Sigma(\Omega) \cap B(0, L)\big\} \cup \big\{c - x' : x \in \Sigma(\Omega) \cap B(0, L)\big\}\Big) \setminus \{0\}. \tag{4.3}$$

The lemma allows us to compute all the protocells of the Voronoi tessellation of $\Sigma(\Omega)$ for a fixed $\Omega = [0, c)$:

Algorithm 4.2. Input: γ satisfying (2.1), $\Omega = [0, c)$, $L \geq 0$.
Output: The pallete of $\Sigma(\Omega)$.

1. Compute the set $\Xi = \{\xi_0 < \xi_1 < \cdots < \xi_{N-1}\}$ given by (4.3).
2. Choose abritrary points $y_0, \ldots, y_N \in \Sigma(\Omega)$ such that $0 \leq y_0 < \xi_0 \leq y_1 < \cdots \leq y_{N-1} < \xi_{N-1} \leq y_N < c$.
3. Compute the local configuration of size L around each y_j.
4. Compute the corresponding protocells to each of these points.
5. Remove possible duplicates in the list of protocells.

Remark. In the real algorithm, we do not need to get the points y_j, we can consider directly ξ_j as the value of y_j' and compute the local neighborhood as $\Sigma([\xi_j, \xi_j + c)) \cap B(0, L)$.

The output of this algorithm for $\gamma = \gamma_T \approx -0.771 + 1.115\mathrm{i}$, the complex Tribonacci constant, and for $X^2(\gamma) = \Sigma(\Omega)$, where $\Omega = [0, 2/(\gamma' - 1))$, can be seen in Figure 1.

The self-similararity property (cf. Proposition 3.2) allows us, when we study $\Sigma(\Omega)$ with $\Omega = [0, c)$, to fix aribtrary $c_0 > 0$ and consider only values of c such that $\gamma' c_0 \leq c < c_0$.

Lemma 4.3. *Let us fix $c_0 > 0$. Then there exists a finite set $\Theta = \{\theta_0 < \theta_1 < \cdots < \theta_{N-1}\} \subseteq (\gamma' c_0, c_0)$ such that the pallete $\mathrm{Pal}([0, c))$ as a function*

$$c \mapsto \mathrm{Pal}\big([0, c)\big)$$

is constant on each of the intervals $(\gamma' c_0, \theta_0), (\theta_0, \theta_1), \ldots, (\theta_{N-2}, \theta_{N-1}), (\theta_{N-1}, c_0)$.

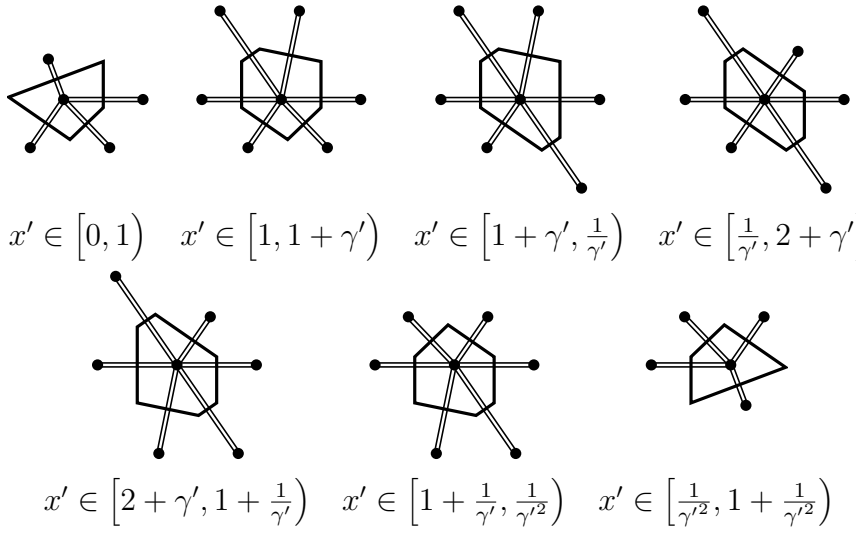


Figure 1: Voronoi protocells for $X^2(\gamma_T) = \Sigma_{\gamma_T}(\Omega)$, where $\Omega = [0, 2/(1-\gamma'_T))$. For a given point $x \in \Sigma(\Omega)$, its protocell $\mathcal{T}(x) - x$ is determined by the value of x'.

As in the previous lemma, the proof is constructive and leads

$$\Theta \coloneqq (\Pi_0 - \Pi_0) \cap (\gamma' c_0, c_0), \quad \text{where} \quad \Pi_0 \coloneqq \Pi \cap (-c_0, c_0) \\ \text{and} \quad \Pi \coloneqq \left\{ x' : x \in \Sigma(\mathbb{R}) \cap B(0, L) \right\}. \quad (4.4)$$

The lemma gives us all possible cut-points of the interval $[\gamma' c_0, c_0)$ into sub-intervals on which the palette is stable. However, unlike Lemma 4.1, this one gives no clue on what happens directly at the cut-points, and the cases $c \in \Theta$ have to be studied seperately. Therefore, we can find all the palettes by the following algorithm:

Algorithm 4.4. Input: γ satisfying (2.1), $c_0 > 0$, $L > 0$.
Output: All possible palettes $\mathrm{Pal}(\Omega)$ of $\Sigma(\Omega)$ for $\Omega = [0, c)$ and $\gamma' c_0 \le c < c_0$.

1. Compute the set $\Theta = \{\theta_0 < \theta_1 < \cdots < \theta_{N-1}\}$.
2. Choose abritrary points $y_0, \ldots, y_N \in \mathbb{R}$ such that $\gamma' c_0 < y'_0 < \theta_0 < y'_1 < \cdots < y'_{N-1} < \theta_{N-1} < y'_N < c_0$.
3. Using Algorithm 4.2, compute the palettes $\mathrm{Pal}(\Omega)$ for all $\Omega = [0, c)$ with $c = \gamma' c_0, \theta_0, \ldots, \theta_{N-1}, y_0 \ldots, y_N$. (We need $2N + 2$ steps.)
4. Remove possible duplicates in the list of palettes.

The output of this algorithm in a certain case is in the section 5, namely in Table 1.

In the previous, we assumed that we know an estimate $L \ge \sup_{x \in \Sigma(\Omega)} \Delta(\mathcal{T}(x))$, and we have yet not provided a way how to find such number. The following procedure enables to find a good estimate:

Algorithm 4.5. Input: γ satisfying (2.1), $\Omega = [0, c_0)$.
Output: An upper bound L such that $L \ge \sup_{x \in \Sigma(\Omega)} \Delta(x)$.
We will denote $\tilde{L} \coloneqq 2|\gamma|^p$, where p is minimal such that $\Im(\gamma^p)$ and $\Im\gamma$ have the opposite signs.

Interval for c	**The palette of** $\Sigma(\Omega)$**, where** $\Omega = [0, c)$													
β^2							$\mathcal{T}_4$	$\mathcal{T}_5$			$\mathcal{T}_8$			$\mathcal{T}_{10}$
$(\beta^2, 2\beta)$				$\mathcal{T}_2$			$\mathcal{T}_4$	$\mathcal{T}_5$			$\mathcal{T}_8$			$\mathcal{T}_{10}$
$(2\beta, \beta+2)$				$\mathcal{T}_2$			$\mathcal{T}_4$	$\mathcal{T}_5$			$\mathcal{T}_8$		$\mathcal{T}_9$	
$(\beta+2, \beta^2+1)$				$\mathcal{T}_2$			$\mathcal{T}_4$		$\mathcal{T}_6$		$\mathcal{T}_8$		$\mathcal{T}_9$	
$(\beta^2+1, 2\beta+1)$		$\mathcal{T}_1$		$\mathcal{T}_2$	$\mathcal{T}_3$				$\mathcal{T}_6$		$\mathcal{T}_8$		$\mathcal{T}_9$	
$(2\beta+1, \beta^2+\beta)$		$\mathcal{T}_1$		$\mathcal{T}_2$	$\mathcal{T}_3$				$\mathcal{T}_6$	$\mathcal{T}_7$			$\mathcal{T}_9$	
$(\beta^2+\beta, \beta^2+2)$	$\frac{1}{\gamma}\mathcal{T}_4$	$\mathcal{T}_1$		$\mathcal{T}_2$	$\mathcal{T}_3$	$\frac{1}{\gamma}\mathcal{T}_8$				$\mathcal{T}_7$			$\mathcal{T}_9$	
$(\beta^2+2, 2\beta+2)$	$\frac{1}{\gamma}\mathcal{T}_4$	$\mathcal{T}_1$	$\frac{1}{\gamma}\mathcal{T}_5$		$\mathcal{T}_3$	$\frac{1}{\gamma}\mathcal{T}_8$				$\mathcal{T}_7$			$\mathcal{T}_9$	
$(2\beta+2, \beta^3)$	$\frac{1}{\gamma}\mathcal{T}_4$	$\mathcal{T}_1$	$\frac{1}{\gamma}\mathcal{T}_5$		$\mathcal{T}_3$	$\frac{1}{\gamma}\mathcal{T}_8$				$\mathcal{T}_7$		$\frac{1}{\gamma}\mathcal{T}_{10}$		
Tile	$\frac{1}{\gamma}\mathcal{T}_4$	$\mathcal{T}_1$	$\frac{1}{\gamma}\mathcal{T}_5$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\frac{1}{\gamma}\mathcal{T}_8$	$\mathcal{T}_4$	$\mathcal{T}_5$	$\mathcal{T}_6$	$\mathcal{T}_7$	$\mathcal{T}_8$	$\frac{1}{\gamma}\mathcal{T}_{10}$	$\mathcal{T}_9$	$\mathcal{T}_{10}$
Value of δ	β^{-1}	β^{-1}	β^{-1}	β^{-1}	β^{-1}	β^{-1}	$\beta^{-\frac{1}{2}}$	$\beta^{-\frac{1}{2}}$	$\beta^{-\frac{1}{2}}$	$\beta^{-\frac{1}{2}}$	$\beta^{-\frac{1}{2}}$	$\beta^{-\frac{1}{2}}$	$\beta^{-\frac{1}{2}}$	1
Value of Δ [(a)]	A	B	A	B	B	A	B	B	B	B	B	A	B	B

[(a)] $A = 2\sqrt{\frac{\beta^2-1}{3\beta^2-1}}$, $B = A\sqrt{\beta}$.

Table 1: The protocells for the complex Tribonacci system for abritrary alphabets. For $m \in \mathbb{N}$ get minimal $k \in \mathbb{Z}$ such that $c \coloneqq \beta^{k+1}m/(\beta-1) \geq \beta^2$, where $\beta = 1/\gamma'$. Take the corresponding row in the table. Then the protocells of $X^m(\gamma)$ are tiles in this row deflated (and rotated) by the factor $1/\gamma^k$. Each but the last tile in the list appears rotated by 180° as well, we omit these to make the table shorter. We omitted the palettes for the cut-points. However, a palette for a cut-point is the intersection of the palettes for the surrounding intervals, i.e., for instance $\mathrm{Pal}([0, \beta^2+1)) = \{\mathcal{T}_2, \mathcal{T}_6, \mathcal{T}_8, \mathcal{T}_9\}$.

1. Compute the 'palette' $\mathcal{P}$ of $\Sigma(\tilde{\Omega})$, where $\tilde{\Omega} = [0, 2)$, using Algorithm 4.2, where we input $\tilde{L}$ in the algorithm.
2. For this palette, compute the maximal value of Δ and denote it $L_1 \coloneqq \max_{\mathcal{T}\in\mathcal{P}} \Delta(\mathcal{T})$.
3. Let k be minimal integer such that $c_0(\gamma\overline{\gamma})^k \geq \tilde{c}$.
4. Output $L \coloneqq |\gamma|^k L_1$.

All the above considerations lead to Theorem 1.1.

5 Complex Tribonacci number exploited

In this section, we will describe the details of the proposed workflow on an example – the complex Tribonacci base $\gamma = \gamma_T$. We aim at the proof of Theorem 1.2. We put $\beta \coloneqq \gamma\overline{\gamma} = 1/\gamma'$ in the sequel.

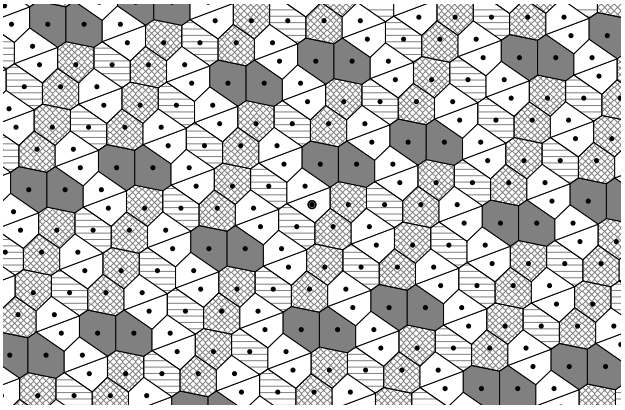


Figure 2: Part of the Voronoi tessellation of $X^2(\gamma)$, where $\gamma = \gamma_T$ is the complex Tribonacci constant. The point 0 is encircled, tiles of the same shape are drawn in the same colour. The case $m = 2$ is one of the special cases when $c = 2/(1-\gamma')$ hits a cut-point, namely $\frac{2}{1-\gamma'} = (\gamma')^{-2} + 1$. The palette of $X^2(\gamma)$ is the intersection of the 4th and the 5th row of Table 1.

We will choose $c_0 = \beta^3$. Algorithm 4.5 gives for $\tilde{L} = 2|\gamma|^2 = 2\beta \approx 3.6786$ a value $L_1 = \beta\sqrt{\frac{\beta^2-1}{3\beta^2-1}} \approx 1.8774$. We have that $c_0/\beta > 2 > c_0/\beta^2$, therefore $k = 1$ and we get an estimate $L = \sqrt{\beta}\sqrt{\frac{\beta^2-1}{3\beta^2-1}} \approx 1.3843$.

Using this L, we run Algorithm 4.4. This gives Θ of size 14. The number of cases in step 3 of this algorithm is then 31.

This means that we have to run Algorithm 4.2 exactly 31 times to obtain all the possible palettes. Amongst these 31 cases, there are many duplicates, and we end with only 18 cases. Moreover, we observe that for cut-points θ_i, the palette is the intersection of palletes of the two surrounding intervals. All the palette for the intervals are depicted in Table 1.

At the bottom of the table, the values of $\delta(\mathcal{T})$ and $\Delta(\mathcal{T})$ are written out for each protocell. It turns out that every row of the table but the special case $c = \beta^2$ has minimal value of δ equal to $1/\beta$ and maximal value of Δ equal to $\sqrt{\beta}\sqrt{\frac{\beta^2-1}{3\beta^2-1}}$. However, it cannot happen that $m/(1-\gamma') = (\gamma\overline{\gamma})^k$.

Theorem 1.2 summarizes the results in this section.

6 Conclusions and open problems

In this paper, we prove that

$$\ell_m(\gamma) = \Theta(\sqrt{m}) \quad \text{and} \quad L_m(\gamma) = \Theta(\sqrt{m})$$

for a wide class of cubic complex Pisot numbers. For a given γ satisfying (2.1), we give an algorithm for computing $\ell_m(\gamma)$ and $L_m(\gamma)$ simultaneously for all m.

The question whether this asymptotic behaviour is true for all complex Pisot γ remains open, as well as the question which is the minimal m (depending on γ) such that $L_m(\gamma) < +\infty$.

References

[A] Shigeki Akiyama. *Cubic Pisot units with finite beta expansions.* In 'Algebraic number theory and Diophantine analysis (Graz, 1998)', de Gruyter (2000), 11–26.

[BH] Peter Borwein and Kevin G. Hare. *Some computations on the spectra of Pisot and Salem numbers.* Math. Comp. **71** (2002), 767–780.

[B] Yann Bugeaud. *On a property of Pisot numbers and related questions.* Acta Math. Hungar. **73** (1996), 33–39.

[EJK] Pál Erdős, István Joó, and Vilmos Komornik. *Characterization of the unique expansions* $1 = \sum_{i=1}^{\infty} q^{-n_i}$ *and related problems.* Bull. Soc. Math. France **118** (1990), 377–390.

[EJK′] Paul Erdős, István Joó, and Vilmos Komornik. *On the sequence of numbers of the form* $\epsilon_0 + \epsilon_1 q + \cdots + \epsilon_n q^n,\ \epsilon_i \in \{0, 1\}$. Acta Arith. **83** (1998), 201–210.

[F] De-Jun Feng. *On the topology of polynomials with bounded integer coefficients.* Preprint, (2013).

[FW] De-Jun Feng and Zhi-Ying Wen. *A property of Pisot numbers.* J. Number Theory **97** (2002), 305–316.

[HP] Tomáš Hejda and Edita Pelantová. *Spectral properties of cubic complex Pisot units.* Submitted, (2013).

[K] Takao Komatsu. *An approximation property of quadratic irrationals.* Bull. Soc. Math. France **130** (2002), 35–48.

[KLP] Vilmos Komornik, Paola Loreti, and Marco Pedicini. *An approximation property of Pisot numbers.* J. Number Theory **80** (2000), 218–237.

[MPZ] Zuzana Masáková, Jiří Patera, and Jan Zich. *Classification of Voronoi and Delone tiles in quasicrystals. I. General method.* J. Phys. A **36** (2003), 1869–1894.

[MPZ′] Zuzana Masáková, Jiří Patera, and Jan Zich. *Classification of Voronoi and Delone tiles of quasicrystals. II. Circular acceptance window of arbitrary size.* J. Phys. A **36** (2003), 1895–1912.

[MPZ″] Zuzana Masáková, Jiří Patera, and Jan Zich. *Classification of Voronoi and Delone tiles of quasicrystals. III. Decagonal acceptance window of any size.* J. Phys. A **38** (2005), 1947–1960.

[M] Robert V. Moody. *Meyer sets and their duals.* In 'The mathematics of long-range aperiodic order (Waterloo, ON, 1995)', volume 489 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, Kluwer Acad. Publ. (1997), 403–441.

[Z] Toufik Zaïmi. *On an approximation property of Pisot numbers. II.* J. Théor. Nombres Bordeaux **16** (2004), 239–249.

A Novel Approach to Silicon Chips Classification

Martin Hejtmánek

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

advisor: Václav Vrba, Institute of Physics, AS CR

Abstract. This article is focusing on automatic electronic chip classification depending on their quality directly after fabrication. The mass production of chips can not be 100% efficient, therefore the defects on chip may occur. These defects are usually not random and could be divided into classes. The aim of the method proposed in this article is to explore these classes of defects and automatically recognize them.

Keywords: chip classification, Timepix detector, clusterization, principal component analysis

Abstrakt. Tento článek se zabývá automatickým rozřazováním elektronických čipů do skupin v závislosti na jejich kvalitě. Masová produkce čipů nemůže zabezpečit 100% výtěžnost, a tak musí být počítáno s tím, že některé z čipů budou defektní. Tyto defekty však typicky nebývají náhodné a jsou pozorovány opakovaně. Cílem popisované metody je odhalit tyto časté defekty, rozřadit je do tříd a automaticky je rozpoznávat.

Klíčová slova: klasifikace čipů, detektor Timepix, shluková analýza, principal component analysis

1 Introduction

In microelectronics industry, integrated chips are produced on a thin circular slices of semiconductor called wafer. The wafer serves as a basis for microchip fabrication on which photolitographic, ion implantation and etching operations are performed. Finalized semiconductor chips of one or more designs (so called MPW - Multi-Project Wafer) are placed side by side, and finally before packaging ther are diced into individual electronic circuits.

During these production phases, many defects can arise, mainly because of imperfections in wafer processing or impurities such as dust particles. Some defects are fatal for the final chip, while others do not significantly influence the operation of a chip. Therefore it is necessary to test (probe) the chips before distributing them and then use the results of the test to classify them into categories depending on their quality.

In this article, we present the method of fast and efficient chip classification using patterns which utilizes digital-to-analog converter voltage trends and the knowledge of commonly occurring defects and their impact on the quality of the chip. The results of this method is a set of patterns which is often observed and which may be used for assigning a chip to a defined quality class. Although we are focusing solely on the Timepix

chip in the following text, the method is general enough to be applied to an arbitrary integrated circuit which is produced with specific structural elements.

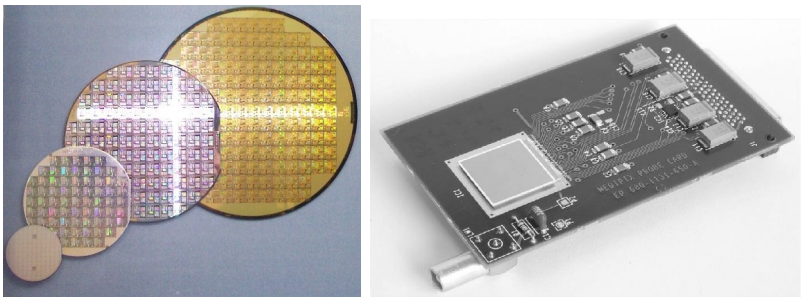


Figure 1: On the left side, there is an example of a silicon wafer. Timepix chip bounded on a probe card can be seen on the right side.

2 Timepix chip and its characteristics

The Timepix chip is a member of Medipix2 chip family developed at CERN. It is a hybrid pixel detector consisting of 256×256 pixel matrix, with pixel size $55 \times 55\,\mu\text{m}^2$. The chip is designed in $0.25\,\mu$m CMOS technology and is intended for medical diagnostics, defectoscopy, etc.

The main characteristics determining the chip quality after fabrication are its 13 digital analog converters (DACs). Five of them convert voltage, the remaining eight current. The Timepix chip contains special testing structures which allow to measure responses of the DACs to input current, resp. voltage directly on the wafer. The testing is performed using probe station which is able to precisely connect each chip on the wafer with special needles.

In our measurement, the responses of all 13 DACs from chips on 3 wafers were collected and analysed. Each wafer contains 107 chips, i.e. totally we have $107 \cdot 3 \cdot 13 = 4173$ data sets. In table 1, detailed technical properties of 13 DACs are shown.

For further information on the Timepix chip, please refer to [4] and [5].

3 Mathematical tools

In this section, the mathematical methods and algorithms used for chip classification will be briefly described.

3.1 Principal component analysis

The Principal component analysis (PCA) is a mathematical method used to reduce the dimensions of data set in order to simplify further data processing while keeping as much information as possible. It is based on linear transformations performed in a way that the resulting data sets have the most variability in the first coordinate, second most in the

DAC	Range	Bits	Mid range value	V/I	LSB size	Values
IKrum	0-40nA	8	20nA→1.497V	I	157pA	26
Disc	0-1.67μA	8	840nA→1.005V	I	6.57nA	26
Preamp	0-2μA	8	1.0μA→966.4mV	I	7.89nA	26
BuffAnalogA	0-10.2μA	8	5.04μA→924.1mV	I	39.4nA	26
BuffAnalogB	0-391μA	8	197μA→1.168V	I	1.54μA	26
Hist	0-200nA	8	100nA→582mV	I	780pA	26
THL	0-2.2V	10+4	1.16V	V	398μV	102
Vcas	0-2.2V	8	1.16V	V	398μV	2
FBK	0-2.2V	8	1.18V	V	9.19mV	26
GND	0-2.2V	8	1.18V	V	9.19mV	26
THS	0-40nA	8	20nA→1.47V	I	156pA	26
BiasLVDS	0-382μA	8	197μA→1.603V	I	1.54μA	26
RefLVDS	0-817mV	8	417mV	V	3.19mV	26

Table 1: Detailed electronical characteristics of the Timepix DACs. In the last column, a number of measured values for each chip is presented.

second coordinate, etc. This allows us to cut off the data sets and continue the analysis with significantly reduced amount of data.

More precisely, given data vectors ordered in a matrix $\mathbb{X}$ by rows (each row represents one data set), we are looking for such a matrix transformation $\mathbb{Y} = \mathbb{X} \cdot \mathbb{W}^T$ that $\operatorname{var} \mathbf{y}_1$ is maximal, $\operatorname{var} \mathbf{y}_2$ is maximal while $\mathbf{y}_1$ and $\mathbf{y}_2$ are uncorrelated, and similarly for other components, $\operatorname{var} \mathbf{y}_k$ is maximal while keeping the condition $\mathbf{y}_k$ is uncorrelated to $\mathbf{y}_i$, $i = 1, 2, \ldots k-1$.

It can be shown [1] that this problem is equivalent to finding eigenvalues and eigenvectors of a sample covariance matrix $\mathbb{C} = (\mathbb{X} - \overline{\mathbf{x}} \cdot \mathbf{1_n^T})^T \cdot (\mathbb{X} - \overline{\mathbf{x}} \cdot \mathbf{1_n^T})$ of data sets, i.e. solving the equation

$$\mathbb{C}\mathbf{v} = \lambda \mathbf{v}\,, \quad \|\mathbf{v}\| = 1\,,$$

ordering the resulting eigenvalues λ_i in descending order, and setting the rows of the matrix $\mathbb{W}$ as

$$\mathbb{W} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \end{pmatrix},$$

where $\mathbf{v}_i$ is eigenvector corresponding to i-th largest eigenvalue and n is dimension of data samples.

Furthermore, when we define the set of variables

$$\phi_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \cdot 100 \quad \text{for} \quad k = 1, 2, \ldots, n\,,$$

we obtain the percentages of variability contained in first k components. This allows us to decide how many components are sufficient to use for further analysis.

For more detailed information on PCA, please see [1] or [3].

3.2 Hierarchical cluster analysis

Cluster analysis is an important statistical method for data classification. Since the term cluster can not be precisely defined, there are many approaches to perform cluster analysis, each giving different results. Therefore the appropriate algorithm must be carefully chosen, according to the specific application. The one chosen for our purposes is hierarchical clustering using Ward's criterion [2].

Hierarchical clustering is a class of algorithms based on recursive agglomeration, resp. division of data into clusters. In our case, agglomerative clustering was used starting with n clusters formed by n experimental samples. In each step of the algorithm, two clusters are merged together. To determine which two clusters are going to be merged in step i, Ward's criterion

$$\min_{k,l \in C_i} \delta_{kl} = \min_{k,l \in C_i} \frac{n_k \cdot n_l}{n_k + n_l} \cdot \|\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_l\|^2$$

is evaluated. C_i denotes the set of clusters in step i, n_k and n_l are the number of points in clusters k and l respectively, $\overline{\mathbf{x}}_k$ and $\overline{\mathbf{x}}_l$ are the centroid coordinates of clusters computed with respect to the Euclidean distance of data sets in cluster.

Algorithm ends after $n-1$ steps when all points are forming one cluster. The result is often presented as dendrogram, a tree diagram illustrating the process of merging clusters. by selecting the level k of the tree, division into k classes can be obtained.

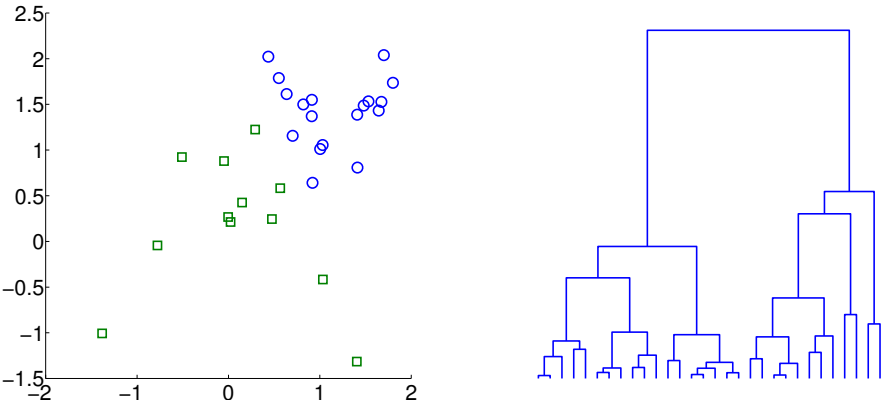


Figure 2: Example of hierarchical clustering. In this model case two Gaussian distributed clusters consisting of 15 points were created around points $[0, 0]$ and $[1.5, 1.5]$ with different variance. After performing Ward's clustering, all but three points were properly matched. The plot on the right represents the resulting dendrogram.

4 Method proposal

The presented method of chip classification involves two steps. First, the significant trends are extracted from sufficiently large amount of experimental data. After that, the trends are assigned to chip quality classes and the desired patterns are used for chip classification.

4.1 Significant trend extraction

For significant trend extraction, the PCA and hierarchical clustering discussed in section 3 are used. The PCA is performed on all of the DACs data sets (i.e. 13 times for each DAC separately). Then, it is sufficient to analyse just a few of the transformed components (in our case, 2 or 3). With these reduced data sets, hierarchical clustering is performed. The number of resulting clusters should be chosen experimentally. However, assuming that statistically the most of chips are without defects, the number of clusters can be relatively high in order to achieve finer resolution. This approach has a little drawback, since many of clusters would contain a small number of outlying points (i.e. defected in the worst way), but these points can be omitted. The most interesting clusters are those with high of points – we can select these clusters as significant trends.

After trends extraction, detailed discussion with competent electronics designers is necessary. Some trends may be non-defective, some DACs may be more important than others, etc. Finally, the patterns are assigned to the quality classes, eventually other criteria can be set.

4.2 Chip classification

To obtain a specific pattern for each DAC, all points from a cluster are averaged, eventually fitted with a curve.

To perform a chip classification, the PCA of the selected patterns is computed. Chips under test are then identified with quality class, if their characteristics are close enough to specific patterns, e.g. with respect to the Euclidean distance in the PCA transformed coordinates.

5 Results

In this section, an example of the proposed method usage is presented. The measured 13 DACs of Timepix chips were used as data sets. Alltogether, data from 321 chips were used. Data corresponding to each DAC were analyzed separately using the PCA and cluster analysis of their first two components. In this special case, the variability included in the first two components was approximately 95%, which turned out to be satisfactory.

Each DAC data set was divided into 12 clusters. Everytime, minimum of two or three distinctive clusters with large amount of data appeared. Furthermore, we assumed that the cluster with the most members is the optimal class (however, it is not always the case). To illustrate how the exact patterns can be determined, we took the cluster with

	Fit of Preamp	Fit of BuffAnalogA	Fit of BuffAnalogB
a	0.8759	0.9445	5.04
b	0.01279	0.005273	-1.976
c	-0.4078	-0.3934	1.531
d	-0.2632	-0.1678	-0.02372

Table 2: Computed coefficients for exponential fitting averaged 'ideal' patterns of Preamp, BuffAnalogA, BuffAnalogB DACs.

the maximal number of elements and declared it as the 'ideal' behavior. These ideal clusters were averaged and, in three cases (Preamp, BuffAnalogA, BuffAnalogB), fitted with exponential function

$$f(x) = a \cdot \mathrm{e}^{b \cdot x} + c \cdot \mathrm{e}^{d \cdot x}\,.$$

This step was necessary due to the fact that the averaged curves were not smooth enough, possibly because of the relatively small data set of 321 chips. The list of the computed coefficients for these fits can be seen in table 2.

6 Conclusions

In this article, we have introduced the method for automatic recognizing of commonly occurring defects in silicon chips. The results using Timepix chip were presented. It was shown that this method is efficient and that it could be used in real application. In present time, the chip classification is complicated task which could be performed often only by experienced electronics designers. The presented method could be useful for acceleration of the process of defining and recognizing chip quality classes.

References

[1] I.T. Jolliffe. *Principal Component Analysis.* Second Edition, Springer-Verlag, New York, (2002)

[2] F. Murtagh, P. Legendre. *Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithms.*

[3] S. Wold. *Principal Component Analysis.* Chemometrics and Intelligent Laboratory Systems **2** (1987) 37–52

[4] X. Llopart. *TIMEPIX Manual v1.0.* Medipix2 Collaboration, `http://medipix.web.cern.ch/medipix/pages/medipix2/documentation.php`

[5] *Medipix homepage.* `http://medipix.web.cern.ch/MEDIPIX`

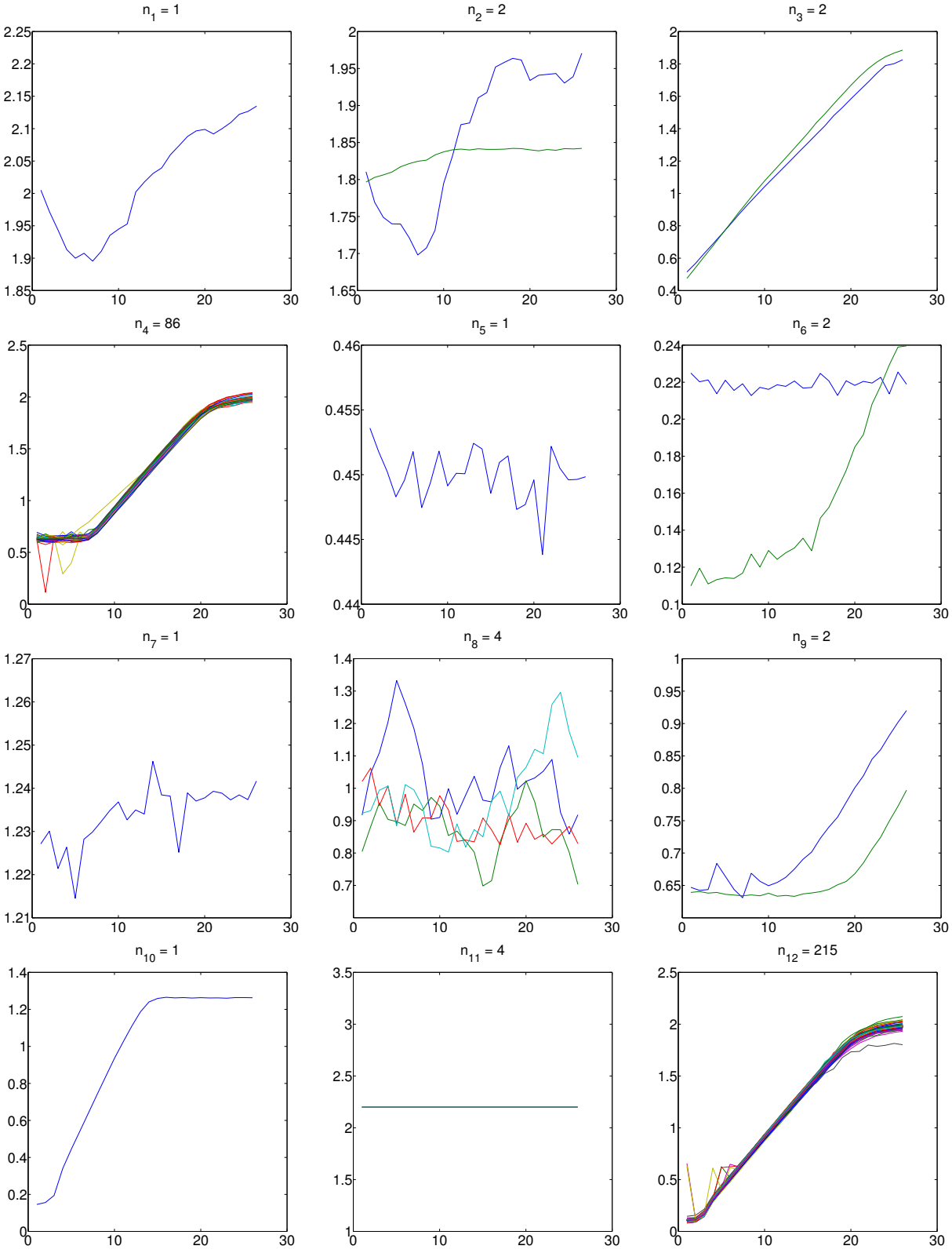


Figure 3: Example of the Vcas DAC trend exploration. As can be seen, two major groups are dominant – 12 and 4 with number of points in the cluster of 215, resp. 86. The group 12 was used for ideal DAC computing. Most of the other groups are outliers, e.g. 6 or 7 and should be treated as absolutely insuficient.

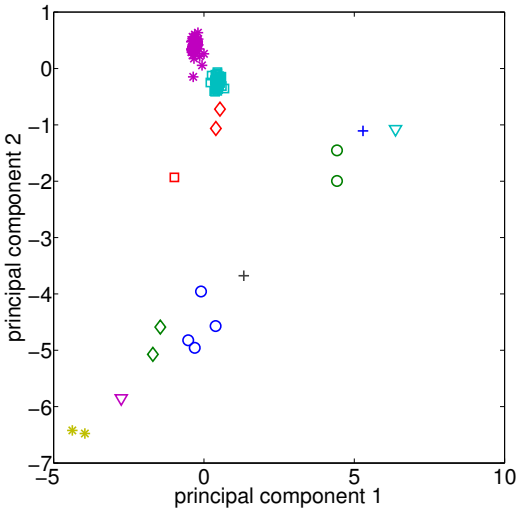


Figure 4: Example of the Vcas DAC trend exploration. In figure above, the result of hierarchical clustering can be found. The largest clusters in the top of plot correspond to groups 12 and 4 from previos figure.

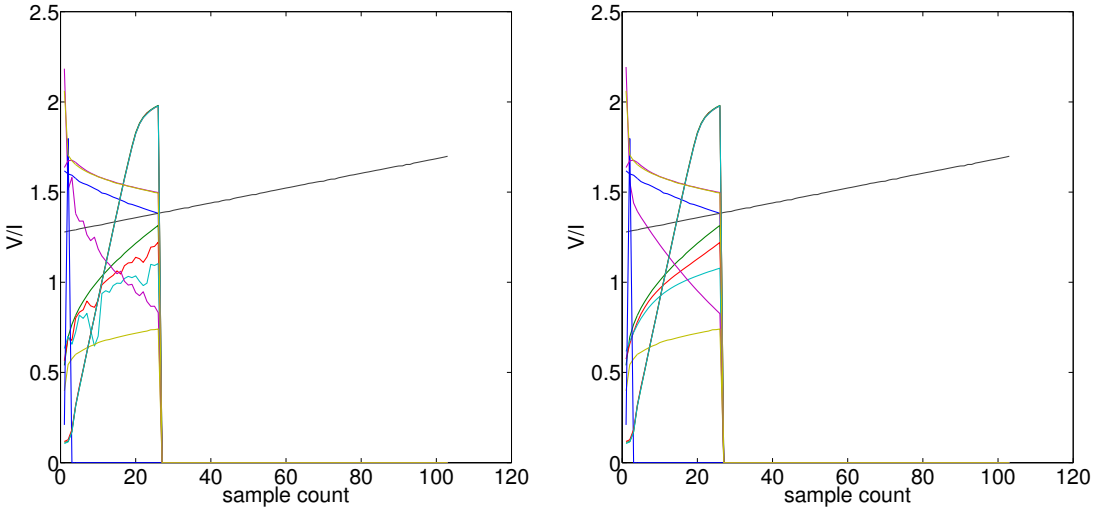


Figure 5: Ideal DACs reconstruction. On the left side, the results obtained by simple averaging of the most significant trends of each DAC is shown. On the right side, the same results with three DACs Preamp, BuffAnalogA, and BuffAnalogB fitted with exponential functions are presented.

Dynamical Decoupling and Bent Networks*

Antonín Hoskovec

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

advisor: Igor Jex, Department of Physics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Dynamical decoupling is a known and successful method of eliminating undesired environmental effects on quantum systems. We present another application of the dynamical decoupling by Pauli pulses, namely using it to eliminate a specific additional coupling added to a working linear qubit network. We assume the additional coupling to arise from bending the network, which is a step towards more dimensional arrangements than one dimensional linear networks.

Keywords: dynamical decoupling, quantum networks and spin chains

Abstrakt. Dynamical decoupling je známá, úspěšně používaná metoda pro eliminaci nežádoucích efektů prostředí na kvantové systémy. Ukážeme další aplikaci dynamical decoupling pomocí Pauliho pulsů a to eliminaci nežádoucí interakce v jinak fungující kvantové síti. Předpokládáme, že příčinou přidané interakce je fyzické přiblížení qubitů při ohybu lineárního řetízku qubitů. Ohyb sítě je první krok k sítím fungujícím ve více dimenzích než jedné.

Klíčová slova: dynamical decoupling, kvantové sítě a spinové řetízky

1 Introduction

Quantum communication was first introduced by means of transfer of a qubit quantum state between the two ends of a linear spin chain by Bose, Nikolopoulos *et al.*, and Christandl *et al.* independently [1, 2, 3], but for a fully operational quantum computer more advanced techniques of quantum information manipulation are needed [4]. It is natural to assume two dimensional arrangements are the next possible step towards constructing an operational quantum computer.

In this article we consider a general formalism summarized for example in [5], which is independent of the specific physical implementation of a quantum network at hand and is, therefore, very general. The advantage of the formalism is that one can describe all the qubits in the network with Hilbert spaces $\mathbb{C}^2$ and the respective Hamiltonians can be

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS13/217/OHK4/3T/14.

expressed using only the Pauli matrices

$$\sigma^x = \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{1a}$$

$$\sigma^y = \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \tag{1b}$$

$$\sigma^z = \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{1c}$$

and the identity

$$I = \sigma^0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{1d}$$

The method of dynamical decoupling by Pauli pulses that we are going to use has been described in great detail in [6]. We will be giving a brief summary of the method in Sec. 3. In order to use the method, we have to assume that the individual systems can be manipulated by very fast, even instantaneous, operations – the assumption of the so called *bang-bang* control. The premiss of the bang-bang control enables us to represent the manipulation by a set of unitary operators.

2 Formalism

We will be using the dynamical decoupling to eliminate the effect of an unwanted part of the Hamiltonian $\mathcal{H}$ and turn it into $\mathcal{H}_{\text{ideal}}$, which we assume is one of the Hamiltonians known to facilitate the perfect state transfer [5]. Without the loss of generality we assume the $\mathcal{H}_{\text{ideal}}$ to be a Heisenberg Hamiltonian.

Let us assume the network consists of N qubits, with the Hilbert space $(\mathbb{C}^2)^{\otimes N}$ and the Hamiltonian $\mathcal{H}_{\text{ideal}}$ being some general Heisenberg Hamiltonian

$$\mathcal{H}_{\text{ideal}} = \sum_i B_i \sigma_i^z - \sum_{i,j} J_{i,j} \left(\sigma_i^x \sigma_j^x + \sigma_i^y \sigma_j^y \right), \tag{2}$$

which, however, facilitates the transfer of a single excitation (that is a condition on choosing the appropriate $J_{i,j}$). The situation we are investigating can be then described by the Hamiltonian

$$\mathcal{H} = \mathcal{H}_{\text{ideal}} + g \left(\sigma_{\alpha-1}^x \sigma_{\alpha+1}^x + \sigma_{\alpha-1}^y \sigma_{\alpha+1}^y \right), \tag{3}$$

where $g \in \mathbb{R}$ and $\alpha \in \{2, \ldots, N-1\}$ is an index of the corner site under consideration from Figure 1.

We expect the interaction between the qubits to have some sort of spatial dependence and the additional interaction g then to arise in the system naturally. It makes good sense to study g from 0 to the interaction magnitude between the corner site and its neighbors only, otherwise the additional interaction could not be considered a perturbation. This type of perturbation and its effects have been previously studied in [7] and it has been shown that the interaction has a severe negative effect on the performance of the network. It is therefore of interest to us to attempt to eliminate the perturbation via the dynamical decoupling method.

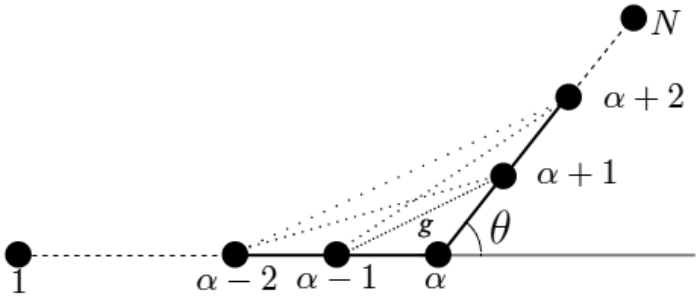


Figure 1: Bent network and additional interaction

3 Methods

The dynamical decoupling we will be using divides the time evolution over time t governed by the Hamiltonian $\mathcal{H}$ into m sub-intervals Δt. Before and after each interval we apply an instantaneous unitary operation. The goal of the procedure is to eliminate the effect of the undesired part of the Hamiltonian after time t by chosing a suitable sequence of the unitaries.

Let the sequence of the unitary operations be denoted by $p_0, \ldots, p_m$, if $\hbar = 1$, the time evolution is

$$U\left(m \cdot \Delta t\right) = p_m e^{-i\mathcal{H}\Delta t} p_{m-1} e^{-i\mathcal{H}\Delta t} \ldots p_1 e^{-i\mathcal{H}\Delta t} p_0, \tag{4}$$

where

$$p_i = \sigma_i^{k_1} \otimes \ldots \otimes \sigma_i^{k_N}, \tag{5}$$

$$k_j \in \{0, \ldots, 3\}. \tag{6}$$

We can then introduce new operators derived from p_i by

$$g_k = p_k \cdot p_{k-1} \cdot \ldots \cdot p_0. \tag{7}$$

If we now notice that

$$p_k = g_k g_{k-1}^\dagger, \tag{8}$$

we can rewrite the time evolution into the form

$$U\left(m \cdot \Delta t\right) = g_m \left(g_{m-1}^\dagger e^{-i\mathcal{H}\Delta t} g_{m-1}\right) \ldots \left(g_0^\dagger e^{-i\mathcal{H}\cdot t} g_0\right) \tag{9}$$

$$= g_m e^{-i\left(g_{m-1}^\dagger \mathcal{H} g_{m-1}\right)\Delta t} \ldots e^{-i\left(g_0^\dagger \mathcal{H} g_0\right)\Delta t}, \tag{10}$$

where in (10) we have used the fact that g_i are unitary. We would now like to use the Magnus expansion [9], in order to do that we identify the time evolution with the one resulting from an average Hamiltonian $\overline{\mathcal{H}}$

$$U(m\Delta t) = g_m e^{-i\overline{\mathcal{H}} m \cdot \Delta t}.$$

The operator g_m can be chosen to be the identity and then it is possible to perform the Magnus expansion in

$$\overline{\mathcal{H}} = \overline{\mathcal{H}}^{(0)} + \overline{\mathcal{H}}^{(1)} + \ldots, \tag{11}$$

to find in the lowest order

$$\overline{\mathcal{H}}^{(0)} = \frac{1}{m}\sum_{i=0}^{m-1} g_i^\dagger H g_i. \tag{12}$$

If we can choose the sequence p_i so that

$$\overline{\mathcal{H}}^{(0)} = \frac{1}{m}\sum_{i=0}^{m-1} g_i^\dagger H g_i ! = \frac{1}{D}\mathcal{H}_{\text{ideal}}, \tag{13}$$

where we allow for the scaling factor D, just by rescaling the time we would effectively eliminate the additional coupling from the system. That is in the lowest order, one needs to remember that the dynamical decoupling is only an approximate method.

4 Decoupling Scheme

We propose a decoupling scheme in Table 1.

	$\sigma^{i_1}\otimes$	...	$\sigma^{i_{\alpha-1}}\otimes$	$\sigma^{i_\alpha}\otimes$	$\sigma^{i_{\alpha+1}}\otimes$	...	σ^{i_N}
g_0	I	I	I	I	I	I	I
g_1	σ^x	σ^x	σ^x	I	I	I	I
g_2	I	I	I	I	σ^y	σ^y	σ^y
g_3	alter σ^z and I		σ^z	σ^x	I	alter σ^z and I	

Table 1: Decoupling scheme

That the scheme is actually a decoupling scheme can be easily shown by direct calculation of the condition (13) for $m = 4$ and $D = 2$ if one uses the properties of the Pauli matrices.

The procedure goes as follows:

1. Let the system evolve for $\frac{1}{4}t$, where t is the time of the unperturbed state transfer.
2. Apply the σ^x Pauli pulse to all the qubits in front of the bending, repeat Step 1.
3. Apply the $(\sigma^x)^\dagger$ Pauli pulse to all the qubits in front of the bending and σ^y on all the qubits behind the bending.
4. Repeat Step 1 and apply $(\sigma^y)^\dagger$ to all the qubits behind the bending.
5. Apply the altering sequence of σ^z and I to all the qubits but the corner, where you should apply σ^x, repeat Step 1.
6. Apply the altering sequence of $(\sigma^z)^\dagger$ and I to all the qubits but the corner, where you should apply $(\sigma^x)^\dagger$.
7. Repeat Steps 1-6.

The procedure was derived using equation (8) defining the Pauli pulses from g_i. The repetition of the decoupling scheme is necessary because of the scaling factor $\frac{1}{2}$.

Usually, the intermediate state during the process is not considered, as it is assumed to change rapidly. However, this scheme uses the σ^x and σ^y matrices on many of the qubits and these matrices create excitations in the sites they act on. For that reason, during the procedure many excitations are created and annihilated. It is an important question of stability of the system that arises and should be answered.

5 Conclusions

We were able to find a decoupling scheme that eliminates in the first order the additional coupling introduced into the system by bending the network. We hope this is a first step toward manipulation of information in more dimensions.

Because the procedure may excite the network to a great extent, simulations need to be performed in the future to find out if — on average — it is an issue or not. Simulations are also desirable because the procedure is imperfect, it is working only in the first order and it might be the case that the remaining terms in the expansion are too large to neglect.

The method we propose is one that relies on being able to instantly apply Pauli matrices to all the qubits in a very rapid sequence. On various systems this could be done differently, but it is a question that needs to be addressed for every computational system individually. On trapped ions, for example, the sequence of σ^x pulses can be achieved by illuminating all the ions with an electromagnetic pulse.

References

[1] S. Bose, "Quantum communication through an unmodulated spin chain", Phys. Rev. Lett. **91**, 207901 (2003).

[2] G. M. Nikolopoulos, D. Petrosyan, and P. Lambropoulos, "Coherent electron wavepacket propagation and entanglement in array of coupled quantum dots", Europhys. Lett. **65**, 297 (2004).

[3] M. Christandl, N. Datta, A. Ekert, and A. J. Landahl, "Perfect state transfer in quantum spin networks", Phys. Rev. Lett. **92**, 187902 (2004).

[4] D. P. DiVincenzo, "The physical implementation of quantum computation", arXiv:quant-ph/0002077 (2000).

[5] A. Kay, "A review of perfect, efficient, state transfer and its application as a constructive tool", int. J. Quantum Inf. **8**, 641 (2010).

[6] H. Frydrych, Master thesis, Technische Universität Darmstadt, 2011.

[7] G. M. Nikolopoulos, A. Hoskovec, and I. Jex , "Analysis and minimization of bending losses in discrete quantum networks", Phys. Rev. A **85**, 062319 (2012).

[8] L. Viola, E. Knill, and S, Lloyd, "Dynamical decoupling of open quantum systems", Phys. Rev. Lett. **82**, 2417–2421 (1999).

[9] W. Magnus, "On the exponential solution of differential equations for a linear operator", Comm. Pure and Appl. Math. **VII(4)**, 649–673 (1954).

Headway Distribution in Interacting Particle Systems Used for Traffic Modeling*

Pavel Hrabák

4th 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. This extended abstract serves as a summary of the study of microscopic behavior of interacting particle systems published in [1], [2], and to be published in [3], [5], which extend works [4], [6]. The aim of the study is the time- and distance- headway distribution of interacting particle systems used for traffic modeling. A class of models is introduced, for which a mapping to zero-range processes is a useful tool to obtain the analytical derivation of stated quantities using car oriented mean field approximation.

Keywords: headway distribution, TASEP, zero-range process

Abstrakt. Tento rozšířený abstrakt slouží jako shrnutí studie mikroskopického chování systémů interagujících částic prezentovaných v [1], [2] a v přijatých článcích [3], [5]. Tato studie rozšiřuje články [4], [6]. Cílem studie jsou časové a prostorové rozestupy v systémech interagujících částic užívaných pro modelování dopravy. Je představena třída modelů, jejichž zobrazení na zero-range procesy je užitečným nástrojem pro analytické odvození zmíněných veličin při požití tzv. car oriented mean field aproximace.

Klíčová slova: rozdělení rozestupů, TASEP, zero-range proces

1 Introduction

This extended abstract introduces a concept of interacting particle systems which are used for traffic modeling. The study focuses on exclusion processes [4], [6], [1] and zero-range processes [2], [5] mainly. The goal is tu study the headway distributions, which is considered to be a microscopic characteristics of traffic-like models. To derive such quantities analytically, it is useful to follow the concept of car oriented mean field approximation. For the Totally asymmetric simple exclusion process (TASEP) this has been done in [4], [6]. The idea is to use the grand-canonical measures $P^{\bullet}_{\varrho}(n) = \Pr[\circ \bullet \bullet \bullet \circ]$, $P^{\circ}_{\varrho}(m) = \Pr[\bullet \circ \circ \circ \bullet]$ of the system on an infinite line for investigation of the system on a large system of $L \gg 1$ sites and $\lfloor \varrho L \rfloor$ particles. The first measure is referred to as probability of a cluster of size n, the second as probability of a gap of length m.

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

2 Model background

This article focuses on these measurers as it is the first step for headway analyzes. The use of zero-range processes for this purpose is motivated by the simplicity of the steady state probability measure, which has factorized form

$$P(n_1, \ldots, n_M) = \frac{1}{\mathcal{Z}} \prod_{k=1}^{M} f(n_k)\,, \tag{1}$$

where $n_k \in \mathbb{N}_0$ is the number of particles in site k, M is the number of sites, and $\mathcal{Z}$ the normalization constant. The dynamics of considered ZRP models is given by the hopping rates $g(n)$, denoting the intenzity of a particle to hop from a cluster of size n to the neighboring site. As we consider the totally asymmetric processes, the particle in site x hops to $x+1$ with intenzity $g(n_x)$. In such case, the marginal measure f can be calculated as

$$f(n) = \prod_{k=1}^{k} g(k)^{-1}\,. \tag{2}$$

This means, that the hopping rates are crucial for investigating the steady state of the system. Moreover, the marginal measure $f(n)$ is closely related to the measures $P^{\bullet}_{\varrho}(n) = \Pr[\circ\bullet\bullet\bullet\circ]$ and $P^{\circ}_{\varrho}(m) = \Pr[\bullet\circ\circ\circ\bullet]$. A particle hopping model as depicted in Figure1 can be understood as the ZRP in two different ways. Both of them are depicted in Figure 2. Firstly, the sites (containers) of ZRP are associated with empty sites (denoted by numbers in Figure 1). The state of each container corresponds to the number of particles in the compact block behind the empty site. Analogically, we can associate the containers with particles (denoted by letters in Figure 1) and the state variable denotes the number of empty sites in front of the particle.

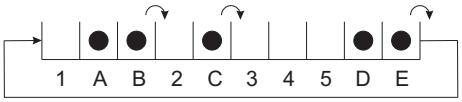


Figure 1: Particle-hopping process with periodic boundary. 5 particles A, B, C, D, E are moving along the lattice of 10 sites; 5 empty sites 1, 2, 3, 4, 5 are "moving" in opposite direction

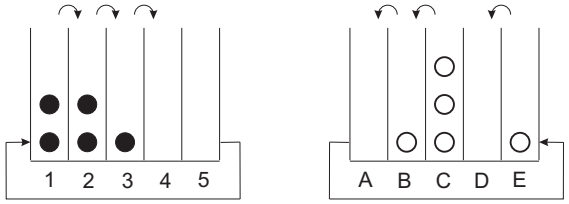


Figure 2: Two different mappings of particle-hopping process from Figure 1 to the ZRP.

The aim of the study to be published in [5] was to investigate the possibilities of extracting appropriate hopping rates from the "real" system. As a reference model, the

car-following Intelligent driver model (IDM) with randomized velocity, acceleration parameter, and deceleration parameter has been used.

3 Results

The simulation experiment showed that the mapping of IDM to the particle hopping model of the first type is not applicable, because the repulsive force between particles disables the particles to form a cluster in the sense of ZRP. The mapping to the zero range process of the second type, i.e., the process associated with hopping holes in opposite direction, is more straightforward.

As will be shown in [5], resulting hopping rates $g(n)$ for the corresponding ZRP are presented in the Figure 3.

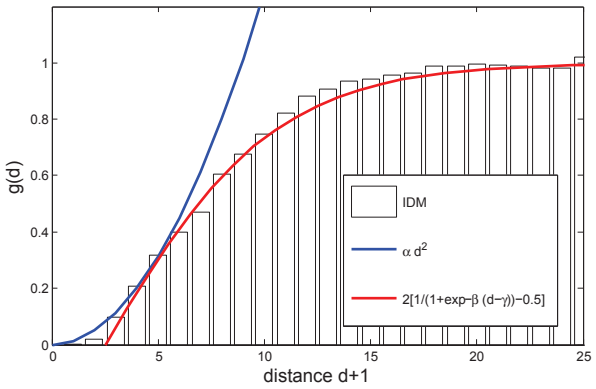


Figure 3: The experimental hopping rates of ZPR in traffic.

The temporal and headway distribution for various densities obtained via simulations of corresponding ZRP are given in Figure 4. Such distribution qualitatively correspond to those observed in real traffic.

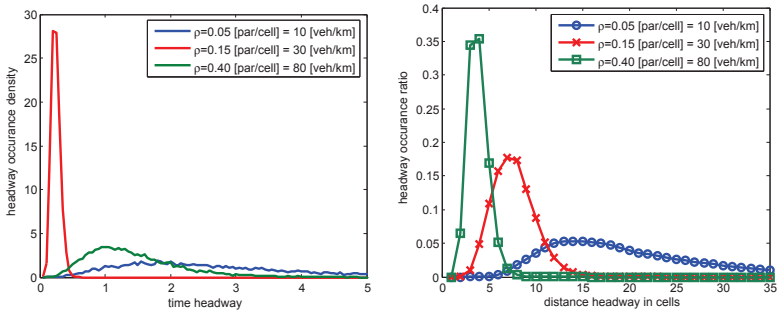


Figure 4: Time and headway distribution.

References

[1] P. Hrabák. Time-headway distribution of Totally Asymmetric Exclusion Process with Nearest-Particle Interaction. In 'SPMS 2011 Proceedings', 77–84, Prague, (2011). CTU.

[2] P. Hrabák. Zero-Range Processes in Traffic Flow Modeling, Microstructural Study. In 'SPMS 2012 Proceedings', 73–80, Prague, (2012). CTU.

[3] P. Hrabák. Time-headways for interacting particle systems in stationary state. In 'SPMS 2013 Proceedings, To be published in 2013', (To be published in 2013).

[4] P. Hrabák and M. Krbálek. *Distance- and Time-headway Distribution for Totally Asymmetric Simple Exclusion Process.* Procedia - Social and Behavioral Sciences **20** (2011), 406–416.

[5] P. Hrabák and M. Krbálek. *Microscopic Traffic-Like Characteristics of Zero-Range Processes, Comparison with Car-Following Models.* In 'Traffic and Granular Flow 2013, To be published in 2014', Springer (To be published in 2014).

[6] 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.

Noninvasive Study of Skin Viscoelastic Properties In-vivo Using Ultrasonic Time Reversal Technique and Mapping of Human Skin Anisotropy Using Ultrasound*

Jana Hradilová

2nd year of PGS, email: hradilova.jana@it.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Zdeněk Převorovský, Department of Impact and Waves in Solids, Institute of Thermomechanics, AS CR

Abstract. Mechanical behavior of human skin is of great interest in dermatology, plastic surgery, regenerative therapies, and cosmetics. Small changes in viscoelastic properties of skin are very sensitively reflecting some skin- and also internal diseases. This work deals with ultrasonic noninvasive investigation of viscoelastic properties of human skin under stepwise tensile loading in-vivo. A small skin-loading device with built-in ultrasonic transmitting and receiving probes [4] is used to observe elastic wave propagation changes during the complex short time step loading and relaxation history. Chirp coded ultrasonic signals of variable amplitude in the frequency range 0.1 - 1 MHz are transmitted along the forearm of several all-aged persons. Ultrasonic wave propagation along the human skin tissue is influenced by many external factors, for example by temperature, humidity, etc. Moreover, mechanical properties of the skin depend on the whole time history of loading. Linear ultrasonic parameters like velocity and attenuation are evaluated from direct propagating waves, and Time Reversal (TR) procedure is used to reveal amplitude-dependent spectral changes and nonlinear effects during the wave propagation at different loading and relaxation stages. Instantaneous complex elastic modules are obtained from ultrasonic measurements, and viscoelastic 5-element rheologic model parameters are evaluated from relaxation curves. The influence of external factors like local temperature, humidity, and others (gender-, age-dependency) on resulting skin characteristics have been discussed in our previous work [2]. TR signal reconstruction helps to partial elimination of dispersion effects. Further, we investigate the anisotropic behavior of the skin using a small multi-axial device which is equipped with ultrasonic transmitting and receiving probes [3]. In our previous work [6] we investigated locally the anisotropic behavior of the forearm and back skin in-vivo. Basic anisotropy characteristics are determined from temporal changes of ultrasonic velocity and attenuation after defined skin loading in various directions, and compared with viscoelastic properties of the skin evaluated from tensile test curves. Using those methods, [5] and [6], we expect to detect some nonlinearities which refer to a pathological behaviour of the skin tissue. Contrary to current state-of-the art, e.g. [1], [7], a great merit of this approach is the possibility to measure instantaneous changes caused by relaxation behavior of biopolymers, including skin.

*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 (Ultrasonic testing of a mechanically loaded human skin tissue - experiments and modeling).

Partial results of this work were presented at International Congress on Ultrasonics 2013 [2], Human Skin Engineering and Reconstructive Surgery 2013 [3], POSTER 2013 - 17th International Student Conference on Electrical Engineering [5] and Stochastic and Physical Monitoring Systems 2013 [6]. The whole text will be published in the Ultrasonics journal.

Keywords: Anisotropy, in-vivo methods, ultrasonic testing, viscoelasticity, time-reversal method

Abstrakt. Znalost mechanického chování lidské kožní tkáně je důležitá pro oblasti jako dermatologie, plastická chirurgie, regenerativní terapie a kosmetika. Malé změny ve viskoelastických vlastnostech velmi citlivě odrážejí některá kožní a také vnitřní onemocnění. Tato práce se zabývá ultrazvukovým neinvazivním vyšetřováním viskoelastických vlastností lidské kožní tkáně při skokovitém tahovém zatěžování in-vivo. K vyšetřování používáme malý přípravek na zatěžování kůže, který je opatřený ultrazvukovými sondami [4], jednou přijímací a dvěma vysílacími. Pomocí něj můžeme sledovat změny v šíření elastických vln kůží během komplexních cyklů zatěžování a relaxace v krátkých časových krocích. Ultrazvukové pulzy typu Chirp s různou amplitudou a frekvencí 0,1 – 1 MHz jsou vysílány podél předloktí několika dobrovolníků různého věku. Šíření ultrazvuku podél lidské kožní tkáně je ovlivněno mnoha vnějšími faktory, například teplotou, vlhkostí, atd. Mechanické vlastnosti kůže navíc zavisí na celém průběhu zatěžování. Z šíření ultrazvukových vln vyhodnocujeme lineární ultrazvukové parametry jako rychlost šíření a útlum. K vyhodnocení časové závislosti spektrálních změn a nelineárních efektů při různých zatěžovacích a relaxačních stavech používáme metodu časové reverzace (TR). Z ultrazvukových měření získáme okamžité komplexní elastické moduly. Z relaxačních křivek jsou vyhodnoceny viskoelastické parametry reologického 5-prvkového modelu. V naší předchozí práci [2] je diskutován vliv vnějších faktorů jako teplota, vlhkost a jiných (závislost na pohlaví, věku) na charakteristiky kůže. Rekonstrukce TR signálů pomáhá částečné eliminaci disperzních efektů. Dále vyšetřujeme anizotropní chování kůže pomocí malého kruhového přístroje opatřeného ultrazvukovými vysílacími a přijímacími sondami [3]. V naší předchozí práci [6] jsme zkoumali lokální anizotropní chování kůže zad a předloktí in-vivo. Základní charakteristiky anizotropie jsou určeny z časových změn rychlosti a útlumu ultrazvuku při určitém zatížení a v různých směrech. Následně jsou porovnány s viskoelastickými vlastnostmi kůže získanými z křivek namáhání v tahu. Použitím těchto metod, [5] a [6], předpokládáme zjištění nelinearit, které budou poukazovat na patologické chování kožní tkáně. Oproti současnému stavu v tomto oboru, např. [1], [7], je přínos této práce v možnosti měřit okamžité změny způsobené relaxačním chováním biopolymerů, tedy i kůže.

Částečné výsledky této práce byly prezentovány na konferencích International Congress on Ultrasonics 2103 [2], Human Skin Engineering and Reconstructive Surgery 2013 [3], POSTER 2013 - 17th International Student Conference on Electrical Engineering [5] and Stochastic and Physical Monitoring Systems 2013 [6]. Celý text bude publikován v časopise Ultrasonics.

Klíčová slova: Anizotropie, metody in-vivo, ultrazvukové testování, viskoelasticita, metoda časové reverzace

References

[1] S. Gahagnon, Y. Mofid, G. Josse, F. Ossant. *Skin anisotropy in vivo and initial natural stress effect: A quantitative study using high-frequency static elastography.* Journal of Biomechanics. **45** (2012), 2860–2865.

[2] J. Hradilová, D. Tokar, Z. Převorovský, S. Dos Santos. *Ultrasonic time reversal technique used to in-vivo investigation of human skin under loading.* In 'Proceedings of the 2013 International Congress on Ultrasonics', May 2–5 2013, Singapore. Singapore: Research Publishing Services. (2013), 978-981-07-5938-4.

[3] J. Hradilová, D. Tokar, Z. Převorovský. *A noninvasive in-vivo study of the skin anisotropy using multi-directional ultrasonic probe.* In 'International Conference Human Skin Engineering and Reconstructive Surgery Proceedings', May 20–22 2013, Prague, Czech Repubic. Technical University of Liberec. (2013), 978-80-7372-956-1.

[4] Z. Převorovský, M. Chlada, J. Krofta, D. Varchon, P. Vescovo. *Nonlinear ultrasonic characterization of human skin under tension.* In 'Ultrasonics International 2003 " U I ' 0 3 " Abstract Book', June 30 - July 3, 2003, Granada, Spain. Elsevier. (2003), P191.

[5] D. Tokar, J. Hradilová. *Device for viscoelastic properties evaluation of human skin in-vivo.* In 'POSTER 2013 - 17th International Student Conference on Electrical Engineering', May 16 2013, Prague, Czech Republic. Czech Technical University in Prague. (2013), 978-80-01-05242-6.

[6] D. Tokar, J. Hradilová, Z. Převorovský. *In-vivo mapping of human skin anisotropy using multi-directional ultrasonic probe.* In 'Stochastic and Physical Monitoring Systems 2013,Proceedings ', June 24–29 3013, Nebřich, Czech Republic. Prague: CTU. (2013). (accepted)

[7] A. Vexler, I. Polyansky, R. Gorodetsky. *Evaluation of skin viscoelasticity and anisotropy by measurement of speed of shear wave propagation with viscoelasticity skin analyze.* Journal of Investigative Dermatology. **113** (1999), 732–739.

Principal Component and Economic Data

Radek Hřebík

2nd 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 paper deals with principal component analysis in sphere of economic data. The aim is not to deal primary with principal component analysis but to introduce the possible use in interpreting economic indicators. As it is well known principal component analysis reduce the dimensionality of origin data set. The input for this research is simple, statistic data about economic situation of more than thirty states during twenty two years. Paper present three ways of interpreting these data as input to principal component analysis and show the results.

Keywords: principal component, analysis, economic time series, objects

Abstrakt. Příspěvek se zabývá analýzou hlavní komponenty v oblasti ekonomických dat. Cílem příspěvku není se primárně zabývat samotnou analýzou hlavní komponenty, ale její aplikací na data z ekonomické oblasti. Cílem analýzy hlavní komponenty je snížení dimenze původního souboru s daty. Vstupem analýzy pro tento účel jsou statistická data popisující ekonomickou situaci ve více než třiceti zemích po dobu dvaceti dvou let. Cílem je prezentace tří přístupů k analýze hlavní komponenty těchto ekonomických časových řad.

Klíčová slova: hlavní komponenta, analýza, ekonomické časové řady, objekty

1 Introduction

The contribution is focused on principal component analysis (PCA). The aim is not to describe the principal component analysis itself in detail. The main idea of principal component analysis is reduction of dimensionality of some data set that consists of a large number of interrelated variables. The reduction retains as much as possible of the variation present in the data set. The aim is achieved by transforming to a new set of variables called the principal components. These principal components are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables. [2] In this research is the aim the reduction to two principal components (PC1 and PC2).

Paper deals with the basic economic data and shows the ways of possible interpretation to serve as input for principal component analysis. The aim is to search the main indicators, monitor the potential trend of concrete objects and finding objects having something in common. It goes hand in hand with principal component analysis goal defined by Abdi and Williams – extracting the important information from the table to represent it as a set of new orthogonal variables called principal components and to display the pattern of similarity of the observations and of the variables as points in maps. [1]

Paper presents three basic ways of using principal component analysis to interpret economic data. The way means to interpret the data set as objects. As the reason for doing such research can be also trying to predict the future development of some country and find the position of state if we know the basic economic prediction. There is also very interesting to capture some progress in time.

1.1 Used data

To do such research play the key role the input data set. As already said it should be some economic time series. Used economic data has been selected from Statistical Annex of European Economy presented by European Commission in spring 2013. [3]

As input to analysis serve the thirty five countries from the whole world, majority are the European countries. The observation take place in years 1993 to 2014. Selected indicators are the total population, unemployment rate, gross domestic product at current market prices, private final consumption expenditure at current prices, gross fixed capital formation at current prices, domestic demand including stocks, exports of goods and services, imports of goods and services and gross national saving. So totally nine indicators are monitored. As the time series go to year 2014 it is clear that years 2013 and 2014 represent predictions.

2 State in year as object

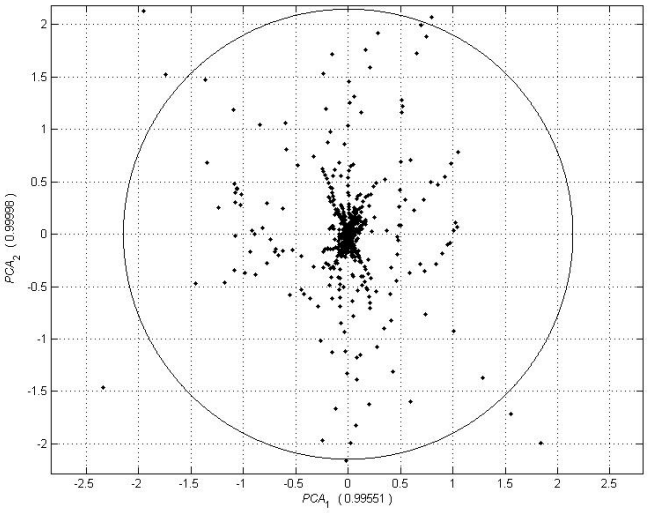
As first possible interpretation of the data set is the object represented by a state in a given year. So the number of objects is relatively high. The total number of object is in this case seven hundred and eighty, it represents number of states multiplied by the number of observed years.

As the number of object is high, the origin data set dimensionality is relatively small. It is created just by nine indicators. The result of principal component analysis is that two principal components are created mainly by combination of population and gross domestic product as shown the indicators weights in table 1.

Table 1: PCA – State in Year as Object

Indicator	PC1	PC2
Total population	-1,106	1,683
Unemployment rate	-0,000	0,025
Gross domestic product	-0,113	-16,492
Private final consumption expenditure	-0,000	0,013
Gross fixed capital formation	0,000	-0,021
Domestic demand including stocks	-0,000	0,004
Exports of goods and services	-0,000	-0,062
Imports of goods and services	-0,000	-0,060
Gross national saving	0,000	-0,026

Figure 1: PCA – State in Year as Object



The figure 2 shows that main points are concentrated by vertical axis. As representative state of vertical line can be selected for example Germany. As the state represented by movement also in horizontal line can be mentioned for example France. Because the number of objects is quite high, for better interpretation there are the objects grouped by the same colour for a given year in figure **??**. The weights of components are in table 1. The first principal component explains almost all of the variance.

The detailed view on values of principal components for three selected countries is shown in table 2. As already mentioned Germany is represented by points in vertical line as can be seen in figure 3.

In case of France there is the result of principal component analysis shown in figure 4 from which is evident that growing gross domestic product is connected with growing population. So in this case the growing gross domestic product goes hand in hand with growing population. That is the different between France and Germany, where the gross domestic product is growing in conditions of almost the same population.

The example of Czech Republic shows that the population is almost constant as in case of Germany, but the potential to grow the gross domestic product is much smaller. The differences between years are very small.

3 States as objects

In second case of possible use of principal component analysis there are the object represented by each state. So the properties are made of indicators in selected years. The number of object is thirty five.

In comparison to first case of use the number of objects is dramatically fallen down. So the representation will be very simple and it will be clear which states are closed to each other. From graphic representation are easily noticed the groups of states. When

Figure 2: PCA – State in Year as Object with legend

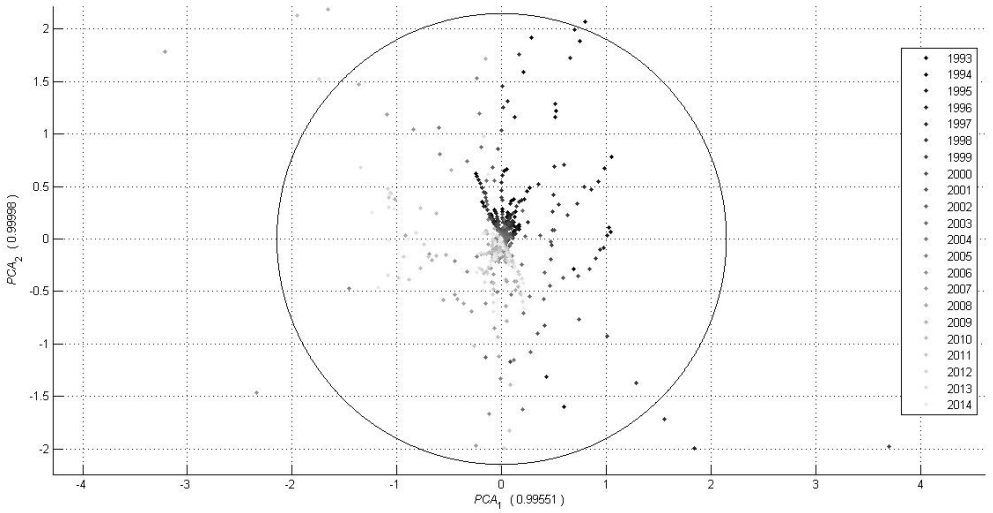


Table 2: Values of principal components of selected countries

Year	CZ - PC1	CZ - PC2	DE - PC1	DE - PC2	FR - PC1	FR - PC2
1993	0,0067	0,2861	0,2827	1,9173	1,0522	0,7847
1994	0,0050	0,2647	0,2052	1,5870	0,9852	0,6696
1995	0,0063	0,2370	0,1282	1,1589	0,9195	0,5430
1996	0,0110	0,1980	0,0564	1,3097	0,8555	0,4740
1997	0,0144	0,1865	0,0091	1,4556	0,7919	0,4969
1998	0,0170	0,1607	0,0148	1,2527	0,7238	0,3324
1999	0,0206	0,1501	-0,0047	1,0328	0,6289	0,2273
2000	0,0235	0,1200	-0,0371	0,8603	0,5009	0,0847
2001	0,0383	0,0600	-0,0854	0,6842	0,3648	0,0330
2002	0,0450	-0,0031	-0,1299	0,6133	0,2278	0,0244
2003	0,0446	-0,0073	-0,1421	0,5579	0,0936	0,0207
2004	0,0429	-0,0389	-0,1377	0,3294	-0,0474	-0,0789
2005	0,0342	-0,0848	-0,1273	0,1840	-0,1925	-0,1445
2006	0,0236	-0,1322	-0,1000	-0,2735	-0,3284	-0,3074
2007	0,0060	-0,1683	-0,0720	-0,8476	-0,4502	-0,5334
2008	-0,0275	-0,2207	-0,0296	-1,1202	-0,5596	-0,5795
2009	-0,0458	-0,1366	0,0486	-0,7816	-0,6633	-0,2043
2010	-0,0540	-0,1618	0,0810	-1,3902	-0,7716	-0,2752
2011	-0,0481	-0,1983	0,0712	-1,8237	-0,8814	-0,3854
2012	-0,0523	-0,1764	0,0276	-1,9930	-0,9842	-0,3722
2013	-0,0549	-0,1612	-0,0166	-2,1568	-1,0835	-0,3459
2014	-0,0566	-0,1736	-0,0425	-2,5564	-1,1822	-0,4637

Figure 3: PCA – State in Year as Object – Germany

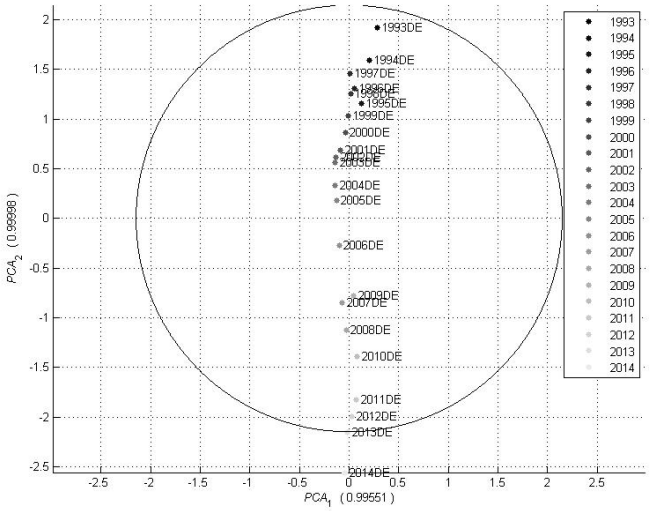


Figure 4: PCA – State in Year as Object – France

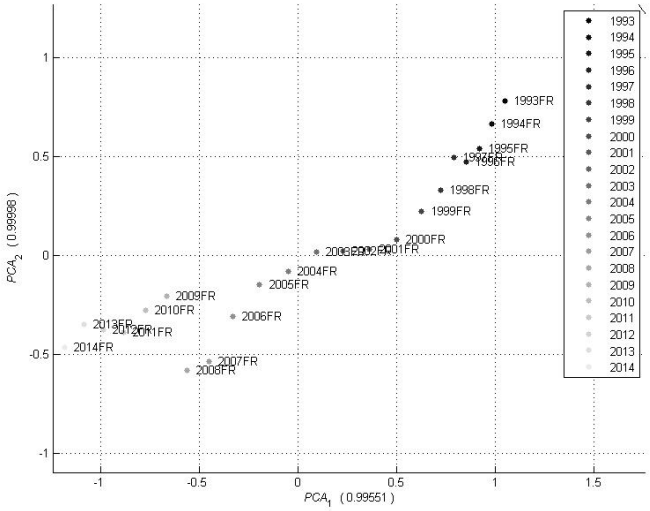
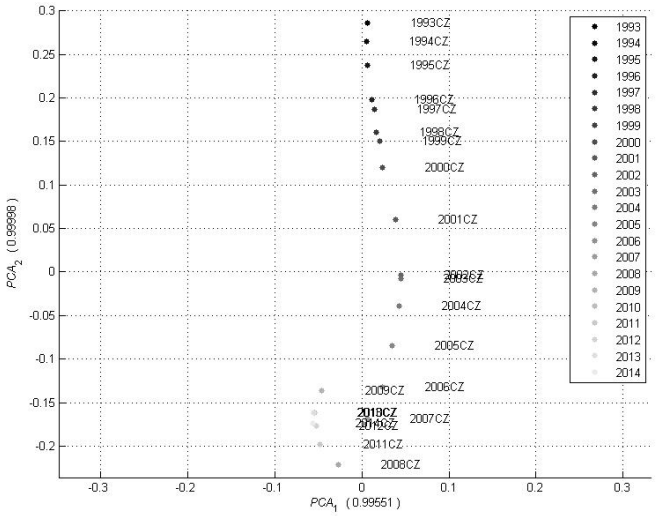


Figure 5: PCA – State in Year as Object – Czech Republic



one point represent one state there is very easily seen the groups of states with similar type of economy. The result of principal component analysis is shown in figure 6. The first principal component explains almost ninety nine percent of variance in origin data set. The values of principal components of each state are summarized in table 3.

The principal components are in this case counted from nearly two hundred indicators. So the reduction of dimensionality is high in this case. These values are created by the nine economy indicators in twenty two years. As in the first case of using principal component analysis also here are the biggest weights on gross domestic product and population. In case of first principal component is the population values included with bigger weight than in case of gross domestic product. Second principal component is preferring the values of gross domestic product in years.

The values of first principal component are in most cases very close to zero, following the weights that implies that the population is without big changes having affect to component values. Second principal component is mostly counted from gross domestic product values. There also apparent the bigger range in values.

4 Years as objects

The third kind of data interpretation is by objects representing calendar year. So there is only twenty four objects in this case. As the number of objects is decreasing, the number of properties of each object is increasing. The total number of indicators of each object is created by number of countries mal number of describing properties. The number of properties is totally over three hundreds. The result showing principal component values is shown in figure 7. The first principal component explains almost ninety nine percent of variance in origin data set.

The advantages of such approach is the very clearly seen the progress in time. The

Table 3: PCA – States as objects

State	PCA 1	PCA 2
Belgium	0,195	0,376
Germany	0,258	-0,115
Estonia	0,317	0,090
Ireland	0,179	0,044
Greece	0,223	-0,118
Spain	-0,547	-0,509
France	-0,443	0,558
Italy	-0,202	1,178
Cyprus	0,279	0,075
Luxembourg	0,290	0,061
Malta	0,299	0,020
Netherlands	0,148	0,089
Austria	0,239	0,141
Portugal	0,230	-0,205
Slovenia	0,296	0,062
Slovakia	0,296	0,047
Finland	0,267	0,129
Bulgaria	0,431	-0,001
Czech Republic	0,280	0,240
Denmark	0,262	0,090
Latvia	0,356	0,022
Lithuania	0,377	-0,032
Hungary	0,347	0,067
Poland	0,279	0,388
Romania	0,468	0,333
Sweden	0,219	0,361
United Kingdom	-0,318	1,921
France	0,324	0,118
F.Y.R. of Macedonia	0,294	0,016
Iceland	0,297	0,022
Turkey	-1,189	-4,953
Montenegro	0,304	0,021
Serbia	0,322	-0,169
United States	-5,431	1,033
Japan	0,056	-1,400

Figure 6: PCA – States as objects

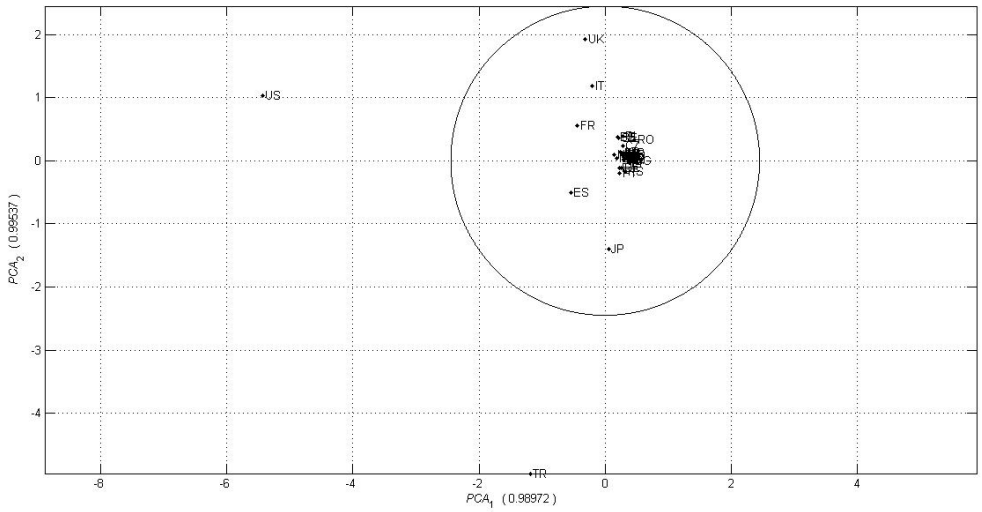


Figure 7: PCA – Years as objects

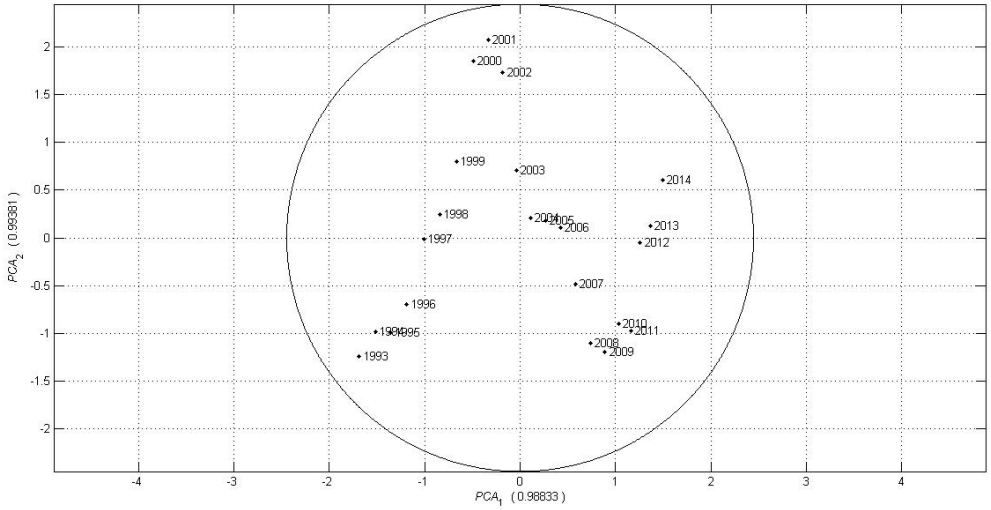
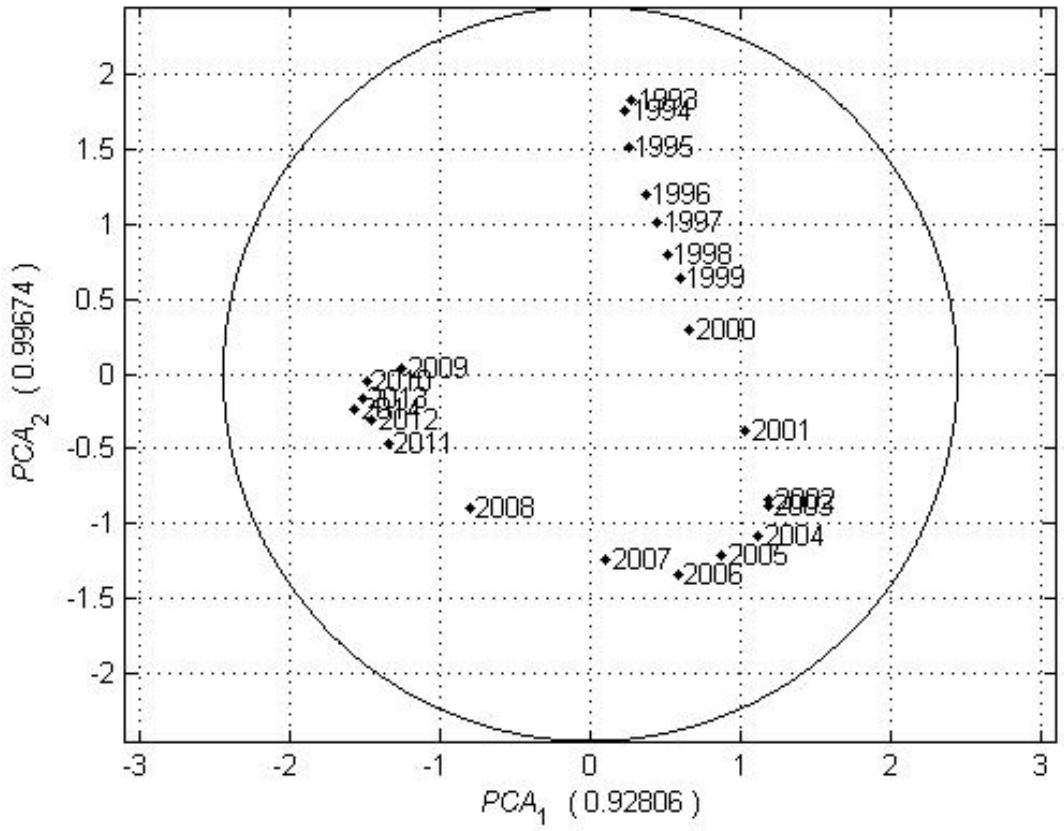


Figure 8: PCA – Years as objects – Czech Republic



next possible use of this approach is to do the analysis just for national data and see the development of separate country. Example of Czech Republic is shown in figure 8. In this case explains the first principal component almost ninety three percent of variance in origin data set. Both principal components explain almost all variance in origin data set.

5 Summary

It was shown that principal component analysis can be also very useful in interpreting the economic data. It represents some other way of interpreting time series and shows how the states position in comparison to others. To fully interpret the results there is need to study the weights of principal components to know what stands behind the components values. The third case of use – the years as objects – gives very clear representation of changing economic situation.

References

[1] H. Abdi and L. J. Williams. *Principal Component Analysis*. (2010). [online]. [cited 2012-08-21]. http://www.utdallas.edu/~herve/abdi-awPCA2010.pdf.

[2] I. T. Jolliffe. *Principal Component Analysis – 2nd Ed.*. Springer (2002).

[3] European Comission. *Statistical Annex of European Economy: Spring 2013*. Economic and Financial Affairs (2013). [online]. [cited 2012-08-21]. http://ec.europa.eu/economy_finance/publications/european_economy/2013/pdf/2013_05_03_stat_annex_en.pdf.

Permutation Entropy*

Václav Hubata-Vacek

3rd year of PGS, email: hubatvac@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. EEG signal of healthy patient can be recognized as output of a chaotic system. There are many measures of chaotic behavior: Hurst and Lyapunov exponents, various dimensions of attractor, various entropy measures, etc. We prefer permutation entropy of equidistantly sampled data. The novelty of our approach is in bias reduction of permutation entropy estimates, memory decrease, and time complexities of permutation analysis. Therefore, we are not limited by EEG signal and permutation sample lengths. This general method was used for channel by channel analysis of Alzheimer diseased (AD) and healthy (CN) patients to point out the differences between AD and CN groups.

Keywords: EEG, Alzheimer's disease, permutation entropy, unbiased estimation, hash table

Abstrakt. EEG sinály zdravých pacientů jsou podobné chaotickému systému. Existuje mnoho měr pro chaotické chování: Hurstův a Lyapunův exponent, attraktory, entropie atd. V tomto článku preferujeme permutační entropii ekvidistantně vzdálených dat. Výhodou tohoto nového postupu je redukce vychýlení odhadu permutační entropie, snížení paměťové a časové náročnosti. Díky tomu nejsme limitováni délkou EEG signálu a délkou permutačního vzorku. Tato metoda byla použita pro analýzu jednotlivých kanálů EEG u pacientů s Alzheimerovou chorobou (AD) a zdravých jedinců. Nasledně byly tyto dvě skupiny porovnány.

Klíčová slova: EEG, Alzheimerova choroba, permutační entropie, nestranný odhad, hašovací tabulka

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia, which gradually destroys the host's brain cells. Recent findings estimate that 35 million people worldwide currently suffer from AD. Clinically, AD manifests itself as a slowly progressing impairment of mental functions whose course lasts several years prior to the death of the patient. Structural changes in AD are related to the accumulation of amyloid plaques between nerve cells in the brain and with the appearance of neurofibrillary tangles inside nerve cells, particularly in the hippocampus and the cerebral cortex. Although a definite diagnosis is possible only by necropsy, a differential diagnosis with other types of dementia and with major depression should be attempted. Magnetic resonance imaging and computerized tomography can be normal in the early stages of AD, but a diffuse cortical atrophy is the main sign in brain scans. Mental status tests are also useful. Electroencephalography

*This work has been supported by the grant SGS11/165/OHK4/3T/14

(EEG) is a non-invasive technique that was first used by Hans Berger in 1929 to record electrical activity of the human brain. The EEG has been used as a tool for investigating dementias for several decades. The conventional spectral analysis of EEG has mainly been concerned with spectral features in several frequency bands. Although the spectral analysis has been successful in AD studies, nonlinear dynamic analysis is crucial if trying to capture higher order dynamic properties of the brain. In particular, several authors have analyzed the EEG in AD patients with non-linear methods. It has been shown that AD patients have lower correlation dimension (D_2) values as a measure of the underlying system dimensional complexity - than control subjects [9]. Furthermore, AD patients also have significantly lower values of the largest Lyapunov (λ_1) exponent than controls in almost all EEG channels. However, estimating the non-linear dynamic complexity of physiological data using measures such as D_2 and λ_1 is problematic, as the amount of data required for meaningful results in their computation is beyond the experimental possibilities for physiological data [10]. One alternative solution lies in computing the entropy of the EEG [8]. The concept of entropy has achieved a large consensus as an indicator of complexity of nonlinear signals [7], [11]. Dauwels et al. [12] and many other authors have shown that Alzheimer's disease increases power in the delta and theta-bands in the case of EEG analysis in frequency domain but the power spectrum is a global characteristics of EEG signal which disables to study local events in the signal. A number of variants of this notion have been proposed in the literature which show different degrees of flexibility, relevance to different problems, efficiency in their computation, as well as theoretical foundations. This work investigates the potential of complexity analysis of multidimensional EEG as indicator of AD onset through permutation entropic modeling.

2 Permutation entropy

2.1 Shanon entropy and its estimation

Definition. Shannon entropy [5] H_{S} of a discrete random variable X with possible values $x_1, ..., x_m$ and probability mass function $p(X)$ is defined as

$$H_{\mathrm{S}} = -\sum_{i=1}^{m} p_i \ln p_i, \tag{1}$$

where $p_i = p(x_i)$.

If the probability function is unknown for an experimental data set, and the number of possible values is finite for random variable X, we estimate probalility function p_i by relative frequency $p_{j,\mathrm{N}}$ and number of events k_{N} as

$$p_{j,\mathrm{N}} = \frac{n_j}{n}, \tag{2}$$

$$k_{\mathrm{N}} = \sum_{n_j>0} 1 \leq k, \tag{3}$$

where n_j is the number of occurrences x_i of random variable X, and n the total number of measurement results. Then we get *naive estimate* of Shannon entropy as

$$H_{\mathrm{N}} = -\sum_{j=1}^{k_{\mathrm{N}}} p_{j,\mathrm{N}} \ln p_{j,\mathrm{N}}. \tag{4}$$

This estimate is biased, and therefore it has a systematic error.

Miller [2] modified *naive estimate* H_{N} using first order Taylor expansion, which produces better estimation

$$H_{\mathrm{M}} = H_{\mathrm{N}} + \frac{k_{\mathrm{N}} - 1}{2n}. \tag{5}$$

2.2 Application to permutation analysis

Entropy estimates can be easily applied to permutation event analysis [3],[4]. Methodology from [2] estimates a smaller bias. Let time series be $\{a_k\}_{k=1}^{T}$ and sliding window $\{b_k\}_{k=1}^{w}$ of length w, then we can substitute signal values b_k in the window with their orders and then obtain permutation pattern $\{\pi_k\}_{k=1}^{w}$. The process of pattern conversion is depicted in Fig. 1.

The universe of random variable X is a set of all permutation of length w. Therefore, the number of possible permutations is

$$m = w!, \tag{6}$$

but the number of various permutations in given signal cannot exceed the number of sliding samples as

$$k_n \leq n = T - w + 1. \tag{7}$$

The number of occurences of j^{th} permutation pattern corresponds with n_j, and n is the total number of samples. Now, we can directly use (4) and calculate the biased naive estimation H_{N} as in [5]. Our methodology is based on Miller's approach [2] and direct application of (5) to permutation patterns. The difference between estimates (4) and (5) varies according to number of distinct patterns and time series length.

3 Permutation analysis for large samples

The main disadvantage of the original procedure of permutation analysis [3] is in its memory and time complexities. They realized permutation memory as a matrix of w columns and $w!$ rows together with counter vector of length $w!$. It enables permutation analysis only for $w < 13$ on a typical computer. The time complexity of single permutation counting is also $w!$, in the worst case. Therefore, we decided to use more sophisticated data structure for permutation analysis. There are many data structures and algorithms for realizing of *look-up table* as a kind of memory with fast access. Our memory has to be optimized only for two operations: FIND and INSERT. We used *hash table* with open

addressing and linear probe strategy [6] as a model, which is easy to realize. Let $P > n$ be the optional prime number. Then the *loading factor* is defined as a ratio $\alpha = n/P < 1$. The mean number of permutation vector comparisons during successful FIND operation was determined [6] as

$$\mathrm{E}T_{\mathrm{OPT}} = \frac{1}{2}\left(1 + \frac{1}{1-\alpha}\right). \tag{8}$$

In the case of unsuccessful FIND operation and INSERT operation, the mean number of permutation vector comparisons is higher [6] than in the previous optimistic case

$$\mathrm{E}T_{\mathrm{PES}} = \frac{1}{2}\left(1 + \frac{1}{(1-\alpha)^2}\right). \tag{9}$$

Our tiny and fast implementation of permutation memory is a matrix of occurred permutations with w columns and $P > n$ rows together with counter vector of length P. The time complexity of single permutation counting is constant and dependent only on loading factor in the best (8) and worst (9) cases. It enables very fast permutation analysis for higher sample length w and long EEG sequences. The last implementation detail is how to realize hash function $index = \mathrm{h}(\boldsymbol{\pi})$ for given permutation pattern $\boldsymbol{\pi}$. By substracting vector of units from vector $\boldsymbol{\pi}$, we obtain digital form $\boldsymbol{y} = \boldsymbol{\pi} - 1$ in the first step. Let $R = w$ be the base of digital system. In the second step, we calculate the value v of $\boldsymbol{y}$ according to base R. The resulting index into hash table has a value $index = v \bmod P$. In the case of Matlab environment, we must increase the index by one. In the case when $P > 3n$, we have $\alpha < 1/3$ and then the mean number of trials is less than 1.25 in the optimistic case (8) and less than 1.625 in the pessimistic (9).

4 Application to EEG

Permutation entropy was applied to EEG signals obtained from two groups of patients. In our prospective study, EEG data were obtained during examinations of 10 patients with moderate dementia (MMSE score 10-19). All subjects underwent brain CT, neurological and neuropsychological examinations. The other group is a control set consisting of 10 age-matched, healthy subjects who had no memory or other cognitive impairments. The average MMSE of the AD group is 16.2 (SD of 2.1). The ages of the two groups are 69.4 ± 9.2 in Alzheimer's group and 68.7 ± 7.7 in normal group, respectively. The first group included 5 men and 5 women, the second group 4 men and 6 women. Informed consent was obtained from all included subjects and the study was approved by the local ethics committee. All recordings were performed under similar standard conditions. The subjects were in a comfortable position, on a bed, with their eyes closed. Electrodes were positioned according to the 10-20 system of electrode placement; the recording was conducted on a 21-channel digital EEG setup (TruScan 32, Alien Technik Ltd., Czech Republic) with a 22-bit AD conversion and a sampling frequency of 200 Hz. The linked ears were used as references. Stored digitized data were zero-phase digitally filtered using a bandpass FIR filter (100 coefficients, Hamming window) of 0.5-60 Hz and a bandstop filter of 49-51 Hz [6]. The analysis started by manual artifact removal. Time series length T varies between 70000 and 120000. We tried to separate these two groups of patients

by two-sample t-test with null hypotheses and alternative hypothesis as

$$\mathrm{H}_0 : \mathrm{E}\hat{H}(\mathrm{AD}) = \mathrm{E}\hat{H}(\mathrm{CN}), \tag{10}$$

$$\mathrm{H}_\mathrm{A} : \mathrm{E}\hat{H}(\mathrm{AD}) \neq \mathrm{E}\hat{H}(\mathrm{CN}). \tag{11}$$

The window length w is the only one parameter of permutation entropy evaluation. We investigated its influence in the case of 8^{th} channel in the range $w = 4$ to 13. Results are collected in Tab. 1 related to the separation power in two-sampled t-test and its p-value. Optimum value of window length (embedded dimension) is $w = 14$ which is in contradiction to statistical conventions. Our interpretation is based on supposition that EEG permutation patterns are not as diverse as they theoretically should be. This hypothesis is illustrated on Fig. 2 where ten most frequent permutation patterns of two patients are added into two distinct plots. Locally monotonic behavior of EEG signal has relatively high probability on the case of AD, while CN exhibits rather chaotic behavior. This phenomenon is difficult to investigate using shorter window or performing analysis in frequency domain.

The final results for permutation entropy estimators H_N and H_M are in Tabs. 2 and 3.

First, we evaluated separation ability of naive estimate H_N of Shannon etropy H_S. Using False Discovery Rate (FDR) [1] methodology of multiple testing for 19 channels and $\alpha = 0.05$ together with t-test, we obtained $\alpha_{\mathrm{FDR}} = 0.0413$ from p_{value} in the Tab. 2. But the differences are significant over the whole front and medial part of the skull for $ch < 18$ in the sense of FDR.

Then we evaluated separation ability of Miller estimate H_M of Shannon etropy H_S. Using the same method as above, we obtained $\alpha_{\mathrm{FDR}} = 0.0216$ from p_{value} in Tab. 3 and the differences are significant mostly over the front half of the skull for $ch = 1..12, 14, 17$.

The difference between naive and Miller estimates is not constant because both EEG signal length and the number of occurring patterns vary within patient groups. Therefore, Miller estimate of permutation entropy causes results which differ from naive approach. Fortunately, novel estimate generates results with more clear biomedical interpretation. Separation power of permutation entropy is depicted on Fig. 3 for 8^{th} channel and optimum pattern length $w = 14$ for naive (left) and Miller (right) approaches.

5 Conclusion

Using Miller's approach instead of direct calculation of Shannon's entropy from permutation frequencies, we have developed a novel method of ECG analysis via permutation entropy. The second advantage of our method is in its very fast permutation analysis and low consumption of computer memory which enables analysis of large time series with greater length of permutation patterns. When the method was applied to diagnose Alzheimer's disease from 19 channel EEG, we recommended pattern length $w = 14$ and Miller estimate of permutation entropy to achieve the best separation between AD and CN groups in standard two-sided two-sampled t-test.

References

[1] Benjamini Y., Hochberg Y., *Multiple hypotheses testing with weights*. Scandinavian Journal of Statistics, 24 407-418.

[2] Miller G., *Note on the bias of information estimates*. Information Theory in Psychology: Problems and Methods, pp. 95-100 (1955).

[3] Bandt C., Pompe B., *Permutation Entropy: A Natural Complexity Measure for Time Series*. Physical Review Letterss, 88, 174102 (2002).

[4] Cao Y., Tung W., Gao J. B., Protopopescu V. A., Hively L. M., *Detecting dynamical changes in time series using the permutation entropy*. Physical review E 70, 046217 (2004).

[5] Shannon C. E., *A mathematical theory of communication*. Bell System Technical Journal (1948)

[6] Knuth D. E., *Art of Programming*. Volume 3: Sorting and Searching, Addison-Wesley, Reading (1998).

[7] Pincus S. M., *Approximate entropy as a measure of system complexity*. Proc. Nati. Acad. Sci. USA, Vol. 88, pp. 2297-2301 (March 1991).

[8] Abásolo D., Hornero R., Espino P., *Approximate Entropy of EEG Background Activity in Alzheimer's Disease Patients*. Intelligent Automation and Soft Computing, Vol. 15, No. 4, pp. 591-603 (2009).

[9] Tang S., Jiang X., Liu Z., Ma L., Zhang Z., Zheng Z., *Entropy Analysis in Interacting Diffusion Systems on Complex Networks*. International Journal of Mathematics and Computers in Simulation, ISSN: 1998-0159 (2012).

[10] Morabito F. C., Labate D., Foresta F. L., Bramanti A., Morabito G., Palamara I., *Multivariate Multi-Scale Permutation Entropy for Complexity Analysis of Alzheimer's Disease EEG*. ISSN 1099-4300, 14, 1186-1202; doi:10.3390/e14071186 (2012).

[11] Jeong-Hyeon Park, Sooyong Kim, Cheol-Hyun Kim, Andrzej Cichocki, Kyungsik Kim, *Multiscale entropy analysis of EEG from patients under different Pathological Conditions*. Fractals, Vol. 15, No. 4 399-404 (2007).

[12] Dauwels et al., *Diagnosis of Alzheimer's Disease from EEG signals: Where are we standing?*, Curr Alzheimer Res. 2010 Sep; 7(6):487-505.

Table 1: Naive estimate of permutation entropy for 8^{th} channel

Window	Mean H_{N}		p_{value}
	AD	CN	
4	2.6227	2.6763	0.289642
5	3.7898	3.8901	0.245163
6	5.0485	5.2067	0.210272
7	6.3754	6.6048	0.178109
8	7.7024	8.0207	0.136442
9	8.8811	9.3133	0.070555
10	9.7614	10.2749	0.022015
11	10.3455	10.8547	0.004363
12	10.6971	11.1372	0.001093
13	10.8891	11.2568	0.001305

Table 2: Naive estimate of permutation entropy ($w = 14$)

Channel	Mean H_{N}		p_{value}
	AD	CN	
1	10.9509	11.2344	0.016177
2	10.9288	11.2340	0.008799
3	10.9993	11.2730	0.013094
4	10.9439	11.2670	0.006146
5	10.9060	11.2483	0.004253
6	10.9520	11.2611	0.005397
7	10.9841	11.2793	0.009685
8	10.9866	11.3035	0.003957
9	10.9596	11.2858	0.005039
10	10.9461	11.2645	0.005418
11	10.9514	11.2629	0.009163
12	11.0033	11.2973	0.011947
13	10.9875	11.2294	0.041253
14	10.9350	11.2227	0.017088
15	10.9433	11.2043	0.032689
16	10.9311	11.1979	0.038126
17	10.9410	11.2494	0.013556
18	10.9690	11.1694	0.132795
19	10.9643	11.1649	0.120322

Table 3: Miller estimate of permutation entropy ($w = 14$)

Channel	Mean H_{M}		p_{value}
	AD	CN	
1	11.4235	11.7096	0.018250
2	11.3954	11.7084	0.008843
3	11.4808	11.7570	0.013002
4	11.4095	11.7476	0.005664
5	11.3629	11.7228	0.003964
6	11.4196	11.7390	0.004630
7	11.4621	11.7632	0.009132
8	11.4643	11.7943	0.002798
9	11.4278	11.7702	0.003966
10	11.4110	11.7424	0.004780
11	11.4184	11.7399	0.009315
12	11.4858	11.7863	0.011526
13	11.4636	11.6979	0.053263
14	11.3966	11.6882	0.021538
15	11.4063	11.6662	0.045093
16	11.3920	11.6574	0.054132
17	11.4048	11.7225	0.015627
18	11.4407	11.6232	0.203424
19	11.4349	11.6188	0.193535

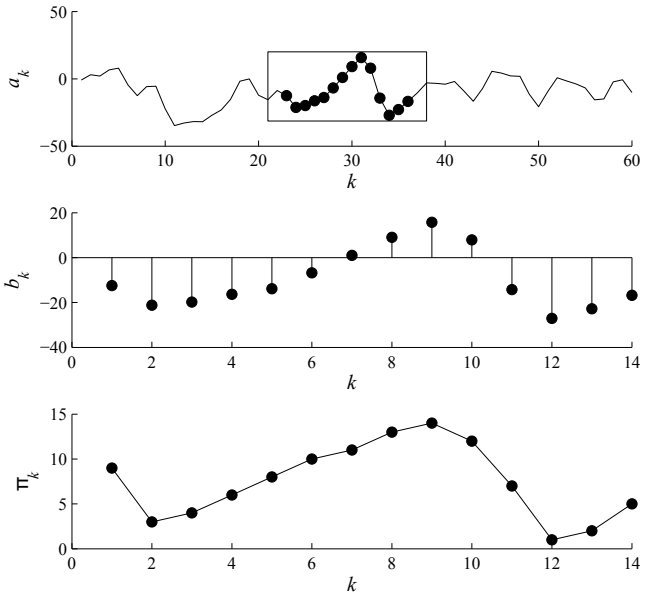


Figure 1: Permutation analysis of EEG: original EEG (top), windowed signal for $w = 14$ (middle), permutation pattern(bottom)

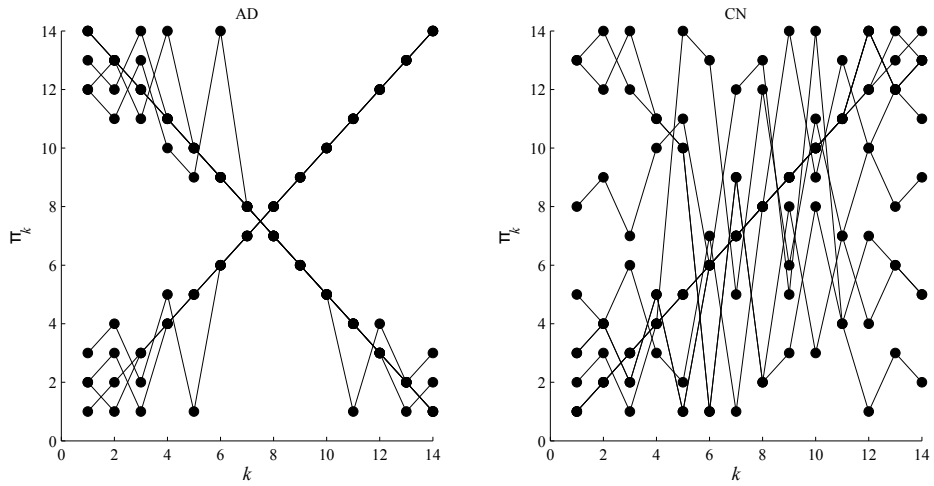


Figure 2: Ten most frequent permutation patterns as union plot for 8^{th} EEG channel and $w = 14$ for typical AD patient (left) and CN patient (right)

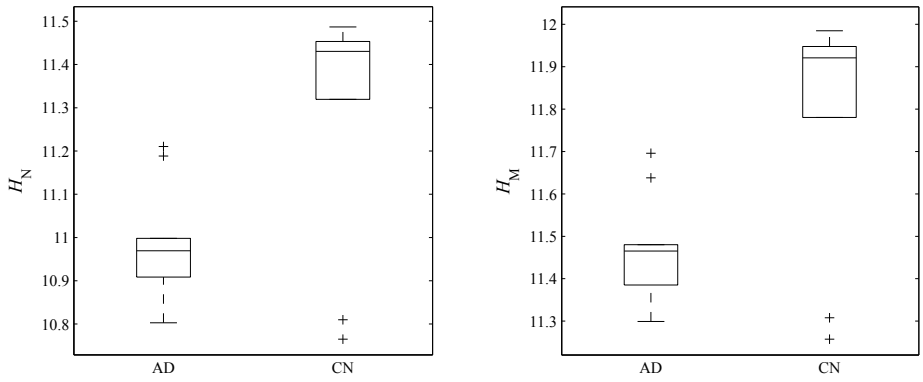


Figure 3: Permutation entropies for AD and CN (w=14, ch=8): naive (left) and Miller (right) approaches

Spectral Asymptotics of a Strong δ' Interaction on a Planar Loop*

Michal Jex

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

advisor: Pavel Exner, Nuclear Physics Institute, AS CR

Abstract. We consider a generalized Schrödinger operator in $L^2(\mathbb{R}^2)$ with an attractive strongly singular interaction of δ' type characterized by the coupling parameter $\beta > 0$ and supported by a C^4-smooth closed curve Γ of length L without self-intersections. It is shown that in the strong coupling limit, $\beta \to 0_+$, the number of eigenvalues behaves as $\frac{2L}{\pi\beta} + \mathcal{O}(|\ln\beta|)$, and furthermore, that the asymptotic behaviour of the j-th eigenvalue in the same limit is $-\frac{4}{\beta^2} + \mu_j + \mathcal{O}(\beta|\ln\beta|)$, where μ_j is the j-th eigenvalue of the Schrödinger operator on $L^2(0, L)$ with periodic boundary conditions and the potential $-\frac{1}{4}\gamma^2$ where γ is the signed curvature of Γ.

This paper was published in Journal of Physics A: Mathematical and Theoretical within the volume 46, number 34 and it was presented at the conference Mathematical result in Quantum Mechanics QMath12 in Berlin from September 10th to 13th, 2013.

Keywords: quantum graphs, singular interactions of δ' type, point spectrum

Abstrakt. V této práci se zabýváme Schrödingerými operátory působícími na $L^2(\mathbb{R}^2)$ popisujícími singulární interakce typu δ' charakterizované vazebným parametrem $\beta > 0$ lokalizované na C^4-hladké uzavřené prosté křivce Γ o délce L. Je spočteno chování bodového spektra v limitě silné vazby, která odpovídá situaci $\beta \to 0_+$. Počet vlastních hodnot lze spočíst jako $\frac{2L}{\pi\beta} + \mathcal{O}(|\ln\beta|)$. Dále j-tá vlastní hodnota v rámci stejné limity silné vazby lze zapsat pomocí výrazu $-\frac{4}{\beta^2} + \mu_j + \mathcal{O}(\beta|\ln\beta|)$, kde μ_j je j-tá vlastní hodnota Schrödingera operátoru na prostoru $L^2(0, L)$ s periodickými hraničními podmínkami a s potenciálem ve tvaru $-\frac{1}{4}\gamma^2$, kde γ je křivost křivky Γ.

Plná verze příspěvku byla otištěna v časopise Journal of Physics A: Mathematical and Theoretical within the volume 46, number 34 a byla přednesena na konferenci Mathematical result in Quantum Mechanics QMath12 in Berlin from September 10th to 13th, 2013.

Klíčová slova: kvantové grafy, singularní interakce typu δ', bodové spektrum

References

[AGH05] 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.

*The research was supported by the Czech Science Foundation within the project P203/11/0701 and by by Grant Agency of the Czech Technical University in Prague, grant No. SGS13/217/OHK4/3T/14.

[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]

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

[BK13] G. Berkolaiko, P. Kuchment: *Introduction to Quantum Graphs*, Amer. Math. Soc., Providence, R.I., 2013.

[BT92] J.F. Brasche, A. Teta: Spectral analysis and scattering for Schrödinger operators with an interaction supported by a regular curve, in *Ideas and Methods in Quantum and Statistical Physics*, (S. Albeverio, J.E. Fenstadt, H. Holden, T. Lindström, eds.), Cambridge Univ. Press 1992; pp. 197–211.

[CE07] C. Cacciapuoti, P. Exner: Nontrivial edge coupling from a Dirichlet network squeezing: the case of a bent waveguide, *J. Phys. A: Math. Theor.* **40** (2007) F511–F523.

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

[CS99] T. Cheon, T. Shigehara: Some aspects of generalized contact interaction in one-dimensional quantum mechanics, in *Mathematical Results in Quantum Mechanics* (QMath7 Proceedings; J. Dittrich, P. Exner, M. Tater, eds.), Birkhäuser, Basel 1999; pp. 203–208.

[CAZ03+] P.L. Christiansen, H.C. Arnbak, V.N. Zolotaryuk, V.N. Ermakova, Y.B. Gaididei: On the existence of resonances in the transmission probability for interactions arising from derivatives of Dirac's delta function, *J. Phys. A: Math. Gen.* **36** (2003), 7589–7600.

[Ex08] P. Exner: Leaky quantum graphs: a review, in [EKK08+], pp. 523–564.

[EI01] P. Exner, T. Ichinose: Geometrically induced spectrum in curved leaky wires, *J. Phys. A: Math. Gen.* **34** (2001), 1439-1450.

[EKK08+] P. Exner, J.P. Keating, P. Kuchment, T. Sunada, A. Teplyaev, eds.: *Analysis on Graphs and Applications*, Proceedings of a Isaac Newton Institute programme, January 8–June 29, 2007; 670 p.; AMS "Proceedings of Symposia in Pure Mathematics" Series, vol. 77, Providence, R.I., 2008

[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.

[EP12] P. Exner, K. Pankrashkin: Strong coupling asymptotics for a singular Schrödinger operator with an interaction supported by an open arc, arXiv: 1207.2271 [math-ph]

[EŠ89] P. Exner, P. Šeba: Bound states in curved quantum waveguides, *J. Math. Phys.* **30** (1989), 2574–2580.

[EY02] P. Exner, K. Yoshitomi: Asymptotics of eigenvalues of the Schrödinger operator with a strong δ-interaction on a loop, *J. Geom. Phys.* **41** (2002), 344–358.

[GH10] Yu. Golovaty, S. Hryniv: On norm resolvent convergence of Schrödinger operators with δ'-like potentials, *J. Phys. A: Math. Theor.* **15** (2010), 155204.

[Ku78] Y.V. Kurylev: Boundary conditions of a curve for a three-dimensional Laplace operator, *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **78** (1978), 112–127; English transl. *J. Soviet Math.* **22** (1983), 107–1082.

[Še86] P.Šeba: Some remarks on the δ'-interaction in one dimension, *Rep. Math. Phys.* **24** (1986), 111–120.

Numerical Simulations of Lunar Plasma Environment*

Martin Jílek

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

advisor: Pavel Trávníček, Astronomical Institute, Institute of Atmospheric Physics, AS CR

Abstract. We study the phenomena associated with the solar wind protons reflections on the lunar surface. We perform several two-dimensional global hybrid simulations with proton particles and fluid electrons. The ambient interplanetary magnetic field is perpendicular to the simulation plane. The results show the formation of a wake structure behind the obstacle with plasma depleted cavity in the center surrounded by low-frequency waves propagating away from the cavity. The properties and generating mechanism of these waves are discussed. We also study the dynamics of reflected protons and its influence to lunar plasma environment.

Keywords: solar wind, Moon, hybrid simulations

Abstrakt. Studujeme jevy spojené s odrazy protonů slunečního větru na měsíčním povrchu. Provádíme několik dvourozměrných globálních simulací s protony jako částicemi a elektrony jako kontinuem. Meziplanetární magnetické pole je kolmé na simulační rovinu. Výsledky ukazují vznik struktury lunárního chvostu za překážkou, s dutinou bez plazmatu uprostřed obklopenou nízkofrekvenčními vlnami šířícími se směrem od dutiny. Popisujeme vlastnosti a mechanismus vzniku těchto vln. Zabýváme se také dynamikou odražených protonů a jejich vlivem na prostředí plasmatu v okolí Měsíce.

Klíčová slova: sluneční vítr, Měsíc, hybridní simulace

1 Introduction

The Moon has no atmosphere nor significant global dipolar magnetic field. Therefore the solar wind particles directly impact its surface forming a lunar wake structure on the nightside of the Moon. Studying the interaction between the solar wind and the Moon is important for understanding the lunar plasma environment.

The beginning of the research of lunar plasma environment is, naturally, associated with first flights of space satellites. First in-situ data from the lunar wake were based on the measurements made by Explorer 35 trough the years 1967 – 1973 [6]. Also the Apollo surface and orbital experiments [10] made some measurements of the lunar wake, but, as in the case of Explorer 35, with a very low resolution. However, all these experiments were able to detect a significant depletion of solar wind density behind the Moon.

*This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS13/146/OHK4/2T/14

Simulation run	Ambient magnetic field	Reflection model
a	$\mathbf{B} = \left(\sqrt{2}/2, \sqrt{2}/2, 0\right)$	no reflections
b	$\mathbf{B} = (0, 0, 1)$	no reflections
c	$\mathbf{B} = (0, 0, 1)$	specular
d	$\mathbf{B} = (0, 0, 1)$	inverting

Table 1: List of performed simulations.

Then several decades passed with relatively little interest in lunar research. On December 27, 1994, the WIND spacecraft used the Moon for a gravitational assist. It passed at a distance of 6.5 lunar radii through the lunar wake and made several measurements in this area with all its instruments switched on. The data showed a number of interesting plasma physical processes. They were described in many papers published mostly in 1996 [2, 8]. Several numerical simulations were performed to explain observed phenomena [1, 12]. Until then, it has been believed, that all particles hitting the Moon are absorbed by the lunar surface.

A systematical research of the lunar plasma environment started in 2007 by Japanese spacecraft SELENE (Kaguya) [7, 9] followed by Indian spacecraft Chandrayaan-1 [4, 13]. Their measurements have indicated that the simplified picture of the Moon as a passive solar wind absorber is incomplete. The instruments onboard these spacecrafts detected solar wind ions reflected on the lunar dayside surface. These ions were also detected inside the near lunar wake. Let us note that Apollo 12 and 14 experiments observed energetic ion fluxes at the nightside surface [3].

In fact, the lunar plasma environment seems to be more complicated. Chandrayaan-1 discovered that up to 20% of the impinging solar wind protons are reflected from the lunar surface back to space as neutral hydrogen atoms [13]. Moreover, the bomboardment of the lunar surface by charged particles may also lead to charging and mobilization of lunar dust. In this paper we present results from global hybrid simulation in the plane perpendicular to IMF. We focus to periodic kinetic effects caused by ion gyration and generated wave structure.

The Japanese spacecraft Kaguya was orbiting the Moon at $\sim$ 100 km altitude. The low energy up-going ions measurements by MAP-PACE onboard discovered that about one percent of SW ions is scattered at the lunar surface [9] In situ observations during one revolution was presented in [7]. The trajectory plane was perpendicular to IMF. The autors explain the unexpected detection of upgoing positive ions deep in nightside using simple numerical model - particle trajectory calculations in prescripted magnetic field. In order to get more realistic picture of lunar wake, we implement proton reflections on the surface into global simulation. Then we compare the data from virtual spacecraft flight through the simulation plane with real in-situ observations from [7].

2 Simulations

We have performed four simulations with different conditions. A 2.5-D version of the hybrid code is used [5]. It has 2 spatial dimensions and 3 velocity dimensions. We use

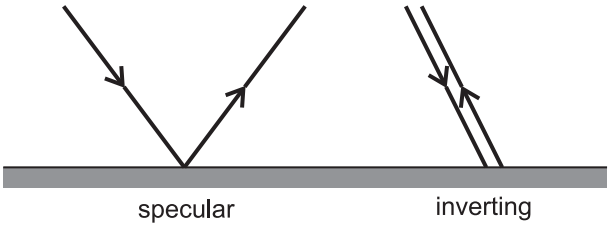


Figure 1: Illustrations of the two different reflection models used in simulations.

selenocentric coordinate system in the simulation X - Y plane, where the x-axis points tailward.

At the beginning of the simulation, the Moon represented by a disk of radius R_L is surrounded by isotropic Maxwellian protons with the constant solar wind speed $v_{sw} = 5v_A$. The initial ambient magnetic field is also introduced in the simulation plane. It is scaled to one, and its orientation differs in different simulation runs. Since the electric field is proportional to the factor $1/\rho$, we must avoid the plasma density dropping out below the value $n_{min} = 0.05$.

The simulation units are derived from the properties of unperturbed plasma. The time is given in inversed proton gyrofrequency, $\Omega_p^{-1} = \omega_{gp}^{-1}$. The unit of space distances is proton inertial length, $\Lambda_p = c/\omega_{pp}$. It follows that the velocities are scaled by the Alfvén velocity v_A. The values of protonic and electronic betas are chosen to be $\beta_p = \beta_e = 1$.

We use spatial resolutions $\Delta x = \Delta y = 0.2\Lambda_p$ and the temporal resolution $\Delta t = 0.01\Omega_p^{-1}$. For calculations of electromagnetic fields we use substepping $\Delta t_B = \Delta t/10$. The simulation plane contains $N_x = 3200$ meshpoints in x-direction and $N_y = 2100$ meshpoints in y-direction. We use 200 superparticles per cell. Total simulation time is $t_{tot} = 90\Omega_p^{-1}$. Since the proton gyroradius $r_{gp} = 2\sqrt{\beta_p/\pi}\Lambda_p = 1.13\Lambda_p$, the selected space resolution $0.2\Lambda_p$ is sufficient to exhibit effects of proton gyromotion. Assuming the density $n_p = 5cm^{-3}$, the proton inertial length is $\Lambda_p = 102$ km. Since the Moon radius is 1738 km, we can set $R_L = 17\Lambda_p$. Note that it is possible to model the solar wind-Moon interaction on real scales. This is not true for example in Mercury simulations [11], where the ratio between the planet radius and λ_p is much higher and scaling down of the sizes is needed. The total sizes of our simulation box are $L_x = 38R_L$ and $L_y = 25R_L$.

Up to here, the simulation parameters are the same for all configurations. The list of performed simulations is given in Table 1. We denote different simulation runs by letters *a*, *b*, *c*, and *d*. They differ in the orientation of the ambient magnetic field and in the behavior of proton superparticles that hit the lunar surface.

In simulation *a* the vectors of the ambient magnetic field $\mathbf{B}$ lie in the simulation plane ($\mathbf{B} = (\sqrt{2}/2, \sqrt{2}/2, 0)$). Thus the angle between $\mathbf{B}$ and solar wind velocity v_{sw} is 45°. Such configuration was widely investigated [12] and we present it in this thesis only in the reason of comparison. All other simulations (*b*, *c*, and *d*) have been performed with the ambient magnetic field perpendicular to the X - Y plane, *i. e.*, $\mathbf{B} = (0, 0, 1)$.

When the superparticle in simulations *a* or *b* hit the lunar surface, it is removed from the simulation. In fact, it is removed with the probability $1 - n_{min} = 0.95$ in order to avoid very low plasma densities resulting to singularities of the electric field, as discussed above. Simulations *c* and *d* are extended by implementation of proton reflections. Thus,

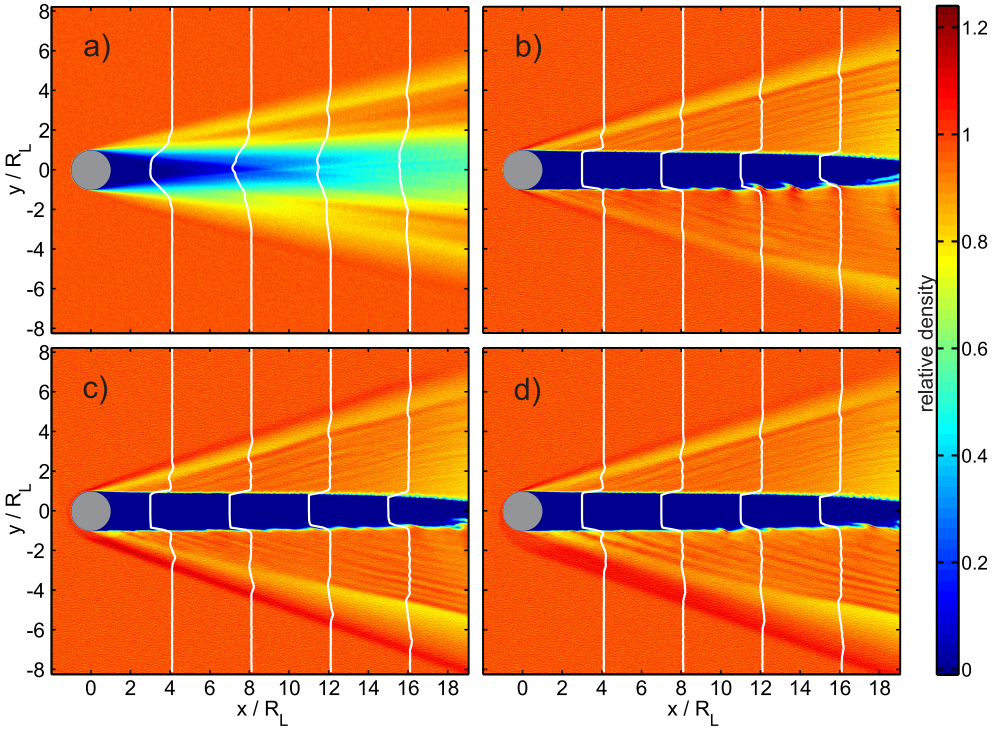


Figure 2: Proton density plots for various simulation runs: Panel a represents the case with ambient magnetic field vectors lying in the simulation plane (i.e. initially $B_z = 0$). Other panels show the situation with ambient magnetic field vectors perpendicular to the simulation plane (i.e. initially $B_x = B_y = 0$) and different reflection models: without reflection (b), specular (c), and inverting (d). Right panel shows the corresponding scale. Overlaid curves denote several profiles of the density in arbitrary units.

when the superparticle hit the surface of the Moon, it is reflected with the probability $f_R = 0.01$. We assume no velocity loss during reflection. The proportion of particles not removed from the simulation remains unchanged, *i. e.*, $n_{min} = 0.05$.

The dependence of reflection angle on the incidence angle is still unknown. Moreover, since it probably depends on the microstructure of the lunar surface, it can also vary in time. We use two extremal reflection models (Figure 1).

Simulation c use a *specular* model, which assumes the Moon to be an ideal sphere. The incidence and reflection angle are equal in this case. Another model, which we have called *inverting*, is used in simulation d. The superparticle hitting the surface changes the sign of all velocity components. This model corresponds to very uneven surface.

3 Results

3.1 Densities

Let us start with the distribution of the proton density (Figure 2). In all cases, we can see a vacuum region formed downstream the Moon. It is surrounded by waves propagating away from the center and forming edges of lunar tail structure. At both edges of the tail,

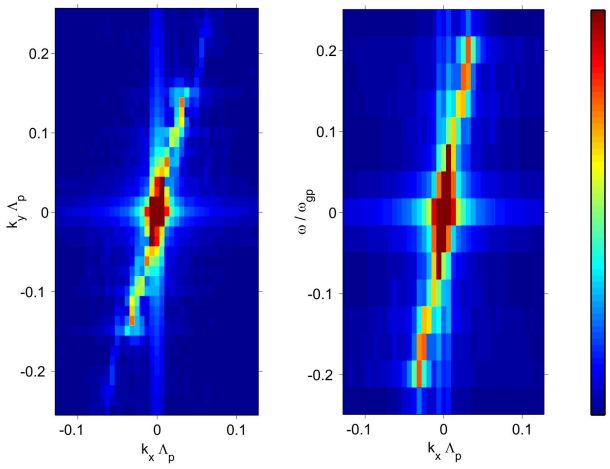


Figure 3: Fourier analysis of the magnetic fluctuations $\delta\mathbf{B}$ in region $15R_L < x < 25R_L$, $-6R_L < y < -1R_L$ and time interval $80\omega_{qp}^{-1} < t < 110\omega_{qp}^{-1}$ The left panel shows the fluctuations as a function of k_x and k_y (averaged over all frequencies ω). The right panel shows the fluctuations as a function of k_x and ω (averaged over all k_y). The frequency is given in the plasma rest frame.

large rarefied plasma regions are observed.

Although the overall structure of the lunar wake is similar in all configurations, the influence of the ambient magnetic field orientation on the lunar wake environment is crucial. In case *a*, plasma refills the cavity along magnetic field lines. Thus, the cavity is refilled relatively fast.

When the ambient magnetic field is perpendicular to the simulation plane (cases *b*, *c*, and *d*), the situation is absolutely different. The plasma particles cannot move across magnetic field lines and the plasma-depleted cavity is of larger size as compared with the previous case (see Figure 2b-d). The waves propagating away from the cavity will be discussed in separate Section 3.2.

Introduction of proton reflections on the lunar surface leads to further changes in the lunar wake environment. Let us focus first on the specular model (Figure 2c). We observe a dense plasma region at the bottom edge of the lunar wake followed by a rarefied plasma region, which is larger than in previous cases. Another difference is the compression of bottom wave-dominated region in y-direction.

Using the inverting model (Figure 2d) leads to further changes in lunar wake. Namely, the region with relatively high density at the bottom edge of the wake is larger. Thus, the selected reflection model influences the global plasma environment. In other words, the global solar wind - Moon interaction is influenced by changes in the local microphysics of the reflection process.

3.2 Waves

In Figure 2 we observe low-frequency waves propagating away from the lunar cavity. In order to process Fourier analysis of these waves, we have performed the simulation *d* (the most realistic one, as we will see in Section 3.4) with a longer total time, $t_{tot} =$

$110\omega_{qp}^{-1}$. The reason of such a long time is that we need to analyze larger waves-dominated rectangular area for longer time. We study the waves in the solar wind rest frame, in the bottom part of the lunar wake. Namely, we get the region $15R_L < x < 25R_L$, $-6R_L < y < -1R_L$ and the time interval $80\omega_{qp}^{-1} < t < 110\omega_{qp}^{-1}$. The time resolution is chosen to be $0.5\omega_{qp}^{-1}$.

The results of the analysis are shown in Figure 3. Left panel shows the fluctuations $\delta\mathbf{B}$ as a function of k_x and k_y averaged over all frequencies ω. The dependence of the fluctuations on k_x and the frequency ω is plotted in the right panel. The plot is averaged over all k_y. It follows that the frequency and the wavenumber of observed low-frequency waves are $\omega \approx 0.19\omega_{gp}$ and $k \approx 0.14$, respectively. The resulting phase velocity $\omega/k \approx 1.35v_A$ and the fact that the waves propagate perpendicular to the magnetic field enable us to assume that the waves are magnetosonic waves with dispersion relation

$$\omega^2 = (v_s^2 + v_a^2)k^2, \tag{1}$$

where v_s is the speed of ion acustic wave which in simulation units is

$$v_s = \sqrt{\frac{k_B\, T_e}{m_p}} = 1. \tag{2}$$

Thus, the phase speed of magnetosonic waves is $\omega/k = \sqrt{2}v_A$. It corresponds to the results of the Fourier analysis.

The generating mechanism is related to Larmor radius and thus it is a kinetic effect. We will describe it from the view of the rest frame. For illustration, let us focus on the protons having the guiding center at the level $y = -R_L$. According to the phase of the Larmor motion, the proton is at the given moment located above or below this level. This location is important at the position $x = 0$ of the proton trajectory. Whereas the protons located here above the guiding center hit the lunar surface and are removed from the simulation, another protons continue in the motion. Note that the process is, in fact, more complex, because the Moon is placed not only at the position $x = 0$.

This Larmor phase filtering effect of the obstructing Moon leads to formation of a periodic strucure along the cavity boundaries with the period $2\pi v_{sw}$. This periodicity leads to propagation of magnetosonic waves.

The wavenumber can be expressed by

$$k = \frac{\sqrt{2}}{4\pi v_A}. \tag{3}$$

The components of wavevector $\mathbf{k}$ are then

$$k_x = \frac{1}{2\pi v_{sw}} \quad , \; k_y = \frac{\sqrt{2}}{4\pi v_A}\sqrt{1 - \frac{2v_A^2}{v_{sw}^2}} \tag{4}$$

and the angle between equiphase lines of propagating waves and the x-axis

$$\cos\vartheta = \frac{\sqrt{2}v_A}{v_{sw}}. \tag{5}$$

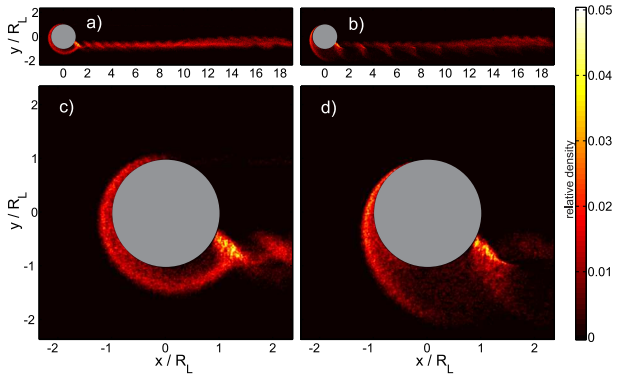


Figure 4: The reflected protons density for specular (a, c) and inverting (b, d) model. The values are scaled to the SW proton density.

If we put the parameters of our simulation to these relations, we get $k = 0.12$, $\mathbf{k} = (0.032, 0.11)$, and $\vartheta = 16.4°$. These values corespond almost exactly to to the simulation data (see Figure 3). Note that from the rest frame, view the waves look stationary and they do not propagate.

3.3 Proton reflections

As explained in section 2, we have implemented proton reflections on the lunar surface. One percent of the protons impacting the lunar dayside are reflected without loss of kinetic energy (in the rest frame). Two different models prescribing how the reflection angle depends on the incidence angle were used. Let us now focus on the reflections in more detail.

Figure 4 shows the distribution of the density of reflected protons in both reflection models. There are significant differences in these plots according to used model. We see that the protons enter the near Moon wake and give rise to a strong asymmetry in this region. Whereas the southern hemisphere of the near-Moon wake is dominated by the reflected protons, they cannot reach the northern part. Note the cloud with a relatively high density of reflected protons approximately at the position $[R_L, -R_L]$. The formation of this region is explained below.

The dynamics of the reflected protons can be described in the following way. For simplicity, we assume the reflection at the equator, in the direction normal to the surface. First we look at the situation from the rest frame. The proton moves in the solar wind with the velocity of $\mathbf{v}_{sw}$. Then it is reflected on the lunar dayside surface (without loss of energy) and its velocity changes to $-\mathbf{v}_{sw}$. In the plasma frame, the velocity of its motion is equal to $-2\mathbf{v}_{sw}$ and the proton starts to gyrate counterclockwise with $r_g = (v_{sw}/v_A)\Lambda_p$. At the bottom part of its gyration, it has the velocity of $2\mathbf{v}_{sw}$. If we now return to the rest frame, the velocity in that area reaches $3\mathbf{v}_{sw}$. Thus, the reflected protons obtain 9 times the original kinetic energy at this position.

When the proton reach the first loop of the trajectory, the magnitude of its velocity is minimal. This leads to the creation of regions with relatively high density of reflected protons observed in Figure 4.

There are three possible destinations of reflected protons depending on the position

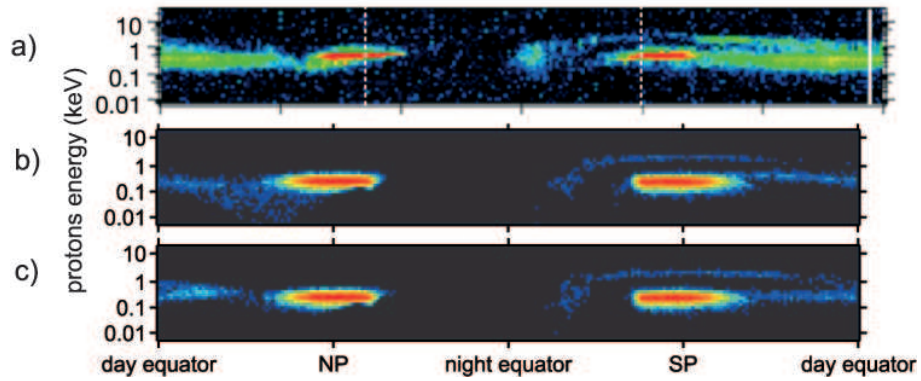


Figure 5: Comparison of experimental data from Kaguya with simulation data. Energetic ion spectrum taken during one revolution by Kaguya (a) (adopted from [7]). Simulation data taken along a virtual spacecraft orbit for specular (b) and inverting (c) reflection model.

and the angle of the reflection: they can hit the dayside again, or they can impact the nightside, or, finally, they can continue in gyro-motion with the solar wind bulk speed.

3.4 Comparison with in-situ data

In order to verify the relevance of our simulation model, we let a virtual spacecraft fly through our simulation plane at ~ 100 km $\approx 1\Lambda_p$ altitude and measure up-going protons. Then we compare the resulting data with real in-situ observations. In fact, we take a ring of inner radius Λ_p and outer radius $1.4\Lambda_p$ and divide it into 220 slices. We sum all protons in each slice and sort it according to their kinetic energy.

The results are plotted in Figure 5. Top panel shows the real in-situ measurement of Kaguya. The flight of our virtual spacecraft begins above the equator on the dayside and continues to the north. Here it detects only protons reflected from the surface. Around the north pole, the solar wind protons are detected. Therefore there is a growth of the measured particles number. No particles reach the north part of the nightside. Then, below the equator, the spacecraft starts to detect protons reflected on the dayside and accelerated by the motional electric field. The detection of solar wind protons around the south pole follows again. In the south part of sunward side we can see both the protons reflected from the surface and those reflected further north from the detection place and accelerated by the electric field.

Middle panel corresponds to the specular reflection model. We observe a detection of low-energy protons in the left part of the spectrum plot. Such protons are not present in real measurement. It indicates that the inverting reflection model (bottom panel) is more realistic than the specular one.

4 Conclusions

We have studied the phenomena associated with the solar wind proton reflections on the lunar surface using different initial and boundary parameters. We have performed several

numerical simulations using a two-dimensional hybrid code. The results provide a global view to the in-situ observations.

We have demonstrated the influence of the magnetic field direction upon the shape of the lunar wake. Some interesting physical effects are exhibited in simulation results. We observe the creation of waves propagating away from the center of the lunar wake. The Fourier analysis showed that the wavenumber $k = 0.13$ and the frequency $\omega = 0.18\omega_{qp}$. Since the waves propagate perpendicular to the magnetic field lines with the velocity $\approx \sqrt{2}v_A$, we expect that they are magnetosonic waves.

We have described the generating mechanism of the waves. It is related to the Larmor phase filtering effect by the obstructing Moon. It results in the periodic structure of plasma along the borders of plasma-depleted cavity downstream the Moon. Such a configuration is unstable and leads to energy dissipation through magnetosonic waves. We have derived the expected wavenumber $k = \sqrt{2}/(4\pi v_A) = 0.12$ and the angle between equiphase lines of propagating waves and the x-axis $\vartheta = \arccos(\sqrt{2}v_A/v_{sw}) = 16.4°$.

In other two simulations we have implemented to the code proton reflections on the lunar surface. One percent of the impacted protons is reflected without loss of energy. We used two reflection models: the specular model and the inverting model. The former assumes that the Moon behaves like an ideal sphere, whereas in the latter all protons are reflected contrary to the incidence direction.

We have showed that the introduction of the proton reflections changes the shape of the lunar wake. The reflected protons are picked-up by the solar wind, accelerated by the motional electric field to obtain up to 9 times the original kinetic energy, and then they penetrate into the near-Moon wake, leading to asymmetry of the lunar plasma environment. We observe a region with relatively high density of reflected protons below the nightside sub-solar point associated with a trajectory loop. We have demonstrated that the protons reflected to the south of the equator may hit the lunar nightside surface.

We have compared the simulation data with in-situ observations of Japanese spacecraft Kaguya. The comparison of two reflection models with real data follows that the inverting model is more realistic than the specular one.

There are several directions how to extend the research of lunar wake. First of all, the full three-dimensional model will give more realistic results. We can also include alpha particles into the simulation, which were neglected in present thesis. It will be interesting to verify the simulation results with real data from ARTEMIS mission, in which two spacecrafts orbit at various altitudes. In order to fit the simulation to the real data, the more appropriate reflection model will be required.

References

[1] Birch, P. C., and S. C. Chapman, *Detailed Structure and Dynamics in Particle-in-Cell Simulations of the Lunar Wake*, Phys. Plasmas **8(10)** (2001), 4551–4559.

[2] Bosqued, J. M., et al., *Moon-solar wind interactions: First results from the WIND/3DP Experiment*, Geophys. Res. Lett. **23(10)** (1996), 1259–1262.

[3] Freeman, J. W., Jr., *Energetic ion bursts on the nightside of the moon*, J. Geophys. Res. **77** (1972), 239–243.

[4] Futaana, Y., S. Barabash, M. Wieser, M. Holmström, A. Bhardwaj, M. B. Dhanya, R. Sridharan, P. Wurz, A. Schaufelberger, K. Asamura, *Protons in the near-lunar wake observed by the Sub-keV Atom Reflection Analyzer on board Chandrayaan-1*, J. Geophys. Res. **115** (2010), A10248.

[5] Matthews, A. P., *Current advance method and cyclic leapfrog for 2D multispecies hybrid plasma simulations*, K. Copmput. Phys., **112** (1994), 102–116.

[6] Ness, N. F., K. W. Behannon, C. S. Scearce, and S. C. Cantarano, *Early Results from the Magnetic Field Experiment on Lunar Explorer 35*, J. Geophys. Res. **72(23)** (1967), 5769–5778.

[7] Nishino, M. N., M. Fujimoto, K. Maezawa, Y. Saito, S. Yokota, K. Asamura, T. Tanaka, H. Tsunakawa, M. Matsushima, F. Takahashi, T. Terasawa, H. Shibuya, and H. Shimizu, *Solar-wind proton access deep into the near-Moon wake*, Geophys. Res. Lett., **36** (2009), L16103.

[8] Oglivie, K. W., J. T. Steinberg, R. J. Fitzenreiter, C. J. Owen, A. J. Lazarus, W. M. Farrell, and R. B. Torbert, *Observations of the lunar plasma wake from the WIND spacecraft on December 27, 1994*, Geophys. Res. Lett. **23** (1996), 1255–1258.

[9] Saito, Y., et al., *Solar wind proton reflection at the lunar surface: Low energy ion measurement by MAP-PACE onboard SELENE (KAGUYA)*, Geophys. Res. Lett., **35** (2008), L24205.

[10] Shubert, G., and B. R. Lichtenstein, *Observations of moon-plasma interactions by orbital and surface experiments*, Rev. Geophys. and Space Phys. **12** (1974), 592–626.

[11] Trávníček, P., P. Hellinger, and D. Schriver, *Structure of Mercury's magnetosphere for different pressure of the solar wind: Three dimensional hybrid simulations*, Geophys. Res. Lett., **34** (2007), L05104.

[12] Trávníček, P., P. Hellinger, D. Schriver, and S. D. Bale, *Structure of the lunar wake: Two-dimensional global hybrid simulations*, Geophys. Res. Lett., **32** (2005), L06102.

[13] Wieser, M., S. Barabash, Y. Futaana, M. Holmstrom, A. Bhardwaj, R. Sridharan, M. B. Dhanya, P. Wurz, A. Schaufelberger, K. Asamura), *Extremely high reflection of solar wind protons as neutral hydrogen atoms from regolith in space*, Planetary and Space Science **57(14–15)** (2009), 2132–2134.

Cohomologies of Lie algebras and extendability*

Dalibor Karásek

3. 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. The connection between extensions of Lie algebras and a cohomology of Lie algebras is described. It is used to reformulate several conjectures and prove them partially.

Keywords: Lie algebra, solvable, extensions, cohomology

Abstrakt. Tento příspěvek si klade za cíl přiblížit souvislost mezi rozšířeními Lieových algeber a jejich kohomologiemi. Ta je následně využita k reformulaci našich hypotéz a jejich částečnému dokázání.

Klíčová slova: Lieova algebra, řešitelná, rozšíření, kohomologie

1 Úvod

Lieovy algebry mají nezanedbatelnou roli v moderní matematické fyzice. Objevují se v mnoha oblastech výzkumu, od teorie strun, přes symetrie diferenciálních rovnic, až po kvantovou mechaniku. Přes jejich nesmírnou důležitost ještě zdaleka nejsou klasifikovány a prozkoumány. Jediná oblast, která je poměrně podrobně zmapovaná, jsou poloprosté Lieovy algebry. Na druhou stranu řešitelné algebry, které jsou díky Leviho teorému [3] druhou podstatnou částí důležitou ke klasifikaci Lieových algeber, jsou klasifikovány kompletně pouze pro dimenze $n < 10$.

Alternativou k podrobné výčtové klasifikaci je přístup Pavla Winternitze, Libora Šnobla a dalších autorů (namátkou třeba [5–7]), kteří zvolili konstrukční přístup. Vybrali si posloupnost nilpotentních algeber a nalezli všechna jejich řešitelná rozšíření. Tím se samozřejmě nezíská kompletní výčet řešitelných algeber, ale dostaneme jich velké množství, a to má své výhody. Například jsme si mohli v průběhu klasifikace všimnout zajímavých vlastností, které měly všechny tyto řešitelné algebry společné, a vyslovit několik hypotéz.

Rozšíření Lieových algeber má velkou souvislost s jejich kohomologiemi. Například druhá komologická grupa má přímou souvislost s centrálními rozšířeními (viz např. [1]). Proto jsme se rozhodli, že se pokusíme přeformulovat náš problém do řeči kohomologií, s čímž nám velmi pomohl článek [4].

Struktura tohoto příspěvku je následující. Předpokládáme, že čtenář má základní znalosti Lieových algeber a jejich kohomologií (pokud ne, jdou nalézt třeba v [1, 2]). V sekci 1 popíšeme souvislost rozšíření algeber s kohomologiemi. V následující sekci upřesníme obecnou konstrukci pro náš případ řešitelného rozšíření. Poté se krátce zastavíme u toho,

*Tato práce byla podpořena grantem SGS10/295/OHK4/3T/14 ze Studentské grantové soutěže ČVUT.

jak klasifikace rozšíření souvisí s exaktními posloupnostmi. Nakonec, v sekci 4, formulujeme dvě hypotézy, podíváme se na ně pomocí aparátu kohomologií, což nám pomůže zjistit, proč jedna z nich platila pro naše případy.

1 Konstrukce rozšíření na direktním součtu algeber

1.1 Operátor pseudokohranice

Pro každou Lieovu algebru máme následující exaktní posloupnost.

$$0 \longrightarrow \mathfrak{C}(\mathfrak{g}) \xrightarrow{i} \mathfrak{g} \xrightarrow{\mathrm{ad}} \mathfrak{Der}(\mathfrak{g}) \xrightarrow{\pi} \mathfrak{Out}(\mathfrak{g}) \longrightarrow 0, \tag{1}$$

Kde $\mathfrak{Out}(\mathfrak{g}) := \mathfrak{Der}(\mathfrak{g}) \,/\, \mathfrak{Inn}(\mathfrak{g})$, i je inkluzivní zobrazení a π je kanonická projekce na faktorprostor.

Chceme rozšířit algebru $\mathfrak{g}$ pomocí algebry $\mathfrak{h}$. K tomu nám poslouží zobrazení $\theta : \mathfrak{h} \to \mathfrak{Out}(\mathfrak{g})$, po kterém chceme, aby to byl homomorfismus Lieových algeber.

Ke každému takovému θ lze zvolit řez σ, t.j. lineární zobrazení (obecně to nebude homomorfismus algeber) takové, že $\pi \circ \sigma = \theta$.

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathfrak{C}(g) & \xrightarrow{i} & \mathfrak{g} & \xrightarrow{\mathrm{ad}} & \mathfrak{Der}(\mathfrak{g}) & \xrightarrow{\pi} & \mathfrak{Out}(\mathfrak{g}) & \longrightarrow & 0 \\ & & & & & & \uparrow \theta & \nearrow \sigma & & & \\ & & & & & & \mathfrak{h} & & & & \end{array} \tag{2}$$

Přestože σ není obecně homomorfismus, a tedy negeneruje reprezentaci na $\mathfrak{g}$, můžeme derivace zúžit na centrum (t.j. $\sigma(\cdot)\restriction_{\mathfrak{C}(g)}$) vytvořit tak $\mathfrak{h}$-modul z $\mathfrak{C}(g)$. Navíc tento $\mathfrak{h}$-modul nezávisí na volbě řezu σ, protože různé řezy se liší pouze o vnitřní derivaci, která je na centru nulová. Podobně část o kterou řez σ není homomorfismus také vymizí na centru. Máme tedy reprezentaci ρ_θ na $\mathfrak{C}(\mathfrak{g})$.

$$\rho_\theta(x)v := \sigma(x)v, \tag{3}$$

pro libovolný řez σ.

Pro každý řez σ jde definovat zobrazení $\mathrm{d}_\sigma : C^n(\mathfrak{h}, \mathfrak{g}) \to C^{n+1}(\mathfrak{h}, \mathfrak{g})$ pomocí „kohraničního předpisu“.

$$\begin{aligned} (\mathrm{d}_\sigma\,\omega)(x_0, \ldots, x_n) := &\sum_{i=0}^{n} (-1)^i \sigma(x_i)\omega(x_0, \ldots, \hat{x}_i, \ldots, x_n) + \\ &+ \sum_{i<j} (-1)^{i+j} \omega([x_i, x_j], x_0, \ldots, \hat{x}_i, \ldots, \hat{x}_j, \ldots, x_n). \end{aligned} \tag{4}$$

Toto zobrazení nazveme operátor pseudokohranice.

Zobrazení d_σ je nilpotentní (t.j. $\mathrm{d}_\sigma^2 = 0$) právě tehdy, když σ je homomorfismus. V každém případě, pokud se omezíme pouze na formy s hodnotami v centru $\mathfrak{g}$, všechna

zobrazení d_σ splynou do jednoho operátoru kohranice (značit ho budeme d_{ρ_θ} nebo d_θ) a vytvoří nám komplex $(C^*(\mathfrak{h}, \mathfrak{C}(\mathfrak{g})), \mathrm{d}_\theta)$, neboť $\sigma \restriction_{\mathfrak{C}(\mathfrak{g})}$ už homomorfismus je.

1.2 Obstrukce rozšíření

Zobrazení σ sice obecně není homomorfismus, ale můžeme „změřit", jak moc se od něj liší. Pro každou dvojici vektorů x, y z $\mathfrak{h}$ musí totiž platit, že $\sigma([x,y]) - [\sigma(x), \sigma(y)]$ je nějaká vnitřní derivace. Pro každou σ můžeme tímto způsobem zvolit $\gamma_\sigma : \mathfrak{h} \wedge \mathfrak{h} \to \mathfrak{g}$, tak aby

$$[\sigma(x), \sigma(y)] - \sigma([x,y]) =: \mathrm{ad}_{\gamma_\sigma(x,y)}\,. \tag{5}$$

Pomocí dvojice σ a γ_σ lze definovat algebru na $\mathfrak{E} := \mathfrak{h} \dotplus \mathfrak{g}$.

$$\begin{aligned} [\langle 0; e_1\rangle, \langle 0; e_2\rangle] &:= \langle 0; [e_1, e_2]\rangle, \\ [\langle x; 0\rangle, \langle 0; e\rangle] &:= \langle 0; \sigma(x)e\rangle, \\ [\langle x; 0\rangle, \langle y; 0\rangle] &:= \langle [x,y]; \gamma_\sigma(x,y)\rangle. \end{aligned} \tag{6}$$

Zbytek relací se dodefinuje tak, aby bylo násobení lineární a antisymetrické. Zbývá ověřit, zda platí Jacobiho identity pro různé volby vektorů $e_i \in \mathfrak{g}$ a $x, y, z \in \mathfrak{h}$.

- Jacobiho identita pro libovolnou trojici e_1, e_2, e_3 je splněna, neboť platí i pro $\mathfrak{g}$.
- Jacobiho identita pro libovolnou trojici e_1, e_2, x je splněna díky faktu, že σ je derivace.
- Jacobiho identita pro libovolnou trojici e, x, y je splněna z definice γ_σ v (5).
- Jacobiho identita pro libovolnou trojici x, y, z neplatí automaticky, ale je splněna právě tehdy, když $\mathrm{d}_\sigma\, \gamma_\sigma = 0$.

Definujme obstrukci $f^{\sigma,\gamma_\sigma} := \mathrm{d}_\sigma\, \gamma_\sigma$ a prozkoumejme ji. Na začátku víme, že se jedná o zobrazení $f^{\sigma,\gamma_\sigma} : \bigwedge^3 \mathfrak{h} \to \mathfrak{g}$.

Věta 1.1. Pro všechny řezy σ a kompatibilní volby γ_σ má f^{σ,γ_σ} hodnoty v $\mathfrak{C}(g)$. A tedy $f^{\sigma,\gamma_\sigma} \in C^3(\mathfrak{h}, \mathfrak{C}(\mathfrak{g}))$.

Proof. Plyne z definice γ_σ v (5). □

Věta 1.2. Pro všechny řezy σ a kompatibilní volby γ_σ je obstrukce f^{σ,γ_σ} kocyklus.

Proof. Až na krok $(*)$, kde je nutno rozepsat delší algebraický výraz, je důkaz přímočarý.

$$\begin{aligned} f^{\sigma,\gamma_\sigma} \in Z^3(\mathfrak{h}, \mathfrak{C}(\mathfrak{g}); \rho_\theta) \qquad &\Leftrightarrow 0 = \mathrm{d}_\theta\, f^{\sigma,\gamma_\sigma} = \mathrm{d}_\theta\, \mathrm{d}_\sigma\, \gamma_\sigma = \mathrm{d}_\sigma\, \mathrm{d}_\sigma\, \gamma_\sigma \\ &\overset{(*)}{=} [\gamma(x_0, x_1), \gamma(x_2, x_3)] + \text{ cyklus v } x_0, \ldots, x_3. \end{aligned}$$

□

Již víme, že pro dané zobrazení θ jsou obstrukce vždy kocykly. Následující tvrzení ukážou, že dokonce pokrývají právě jednu třídu v $H^3(\mathfrak{h}, \mathfrak{C}(\mathfrak{g}); \rho_\theta)$.

Věta 1.3. Pro libovolnou kompatibilní dvojici σ, γ_σ a libovolný další řez σ' existuje $\gamma' \in C^2(\mathfrak{h}, \mathfrak{g})$ kompatibilní s σ' a zachovávající obstrukci, t.j.

$$\begin{aligned} f^{\sigma,\gamma_\sigma} &= f^{\sigma',\gamma'} \\ \mathrm{d}_\sigma\, \gamma_\sigma &= \mathrm{d}_{\sigma'}\, \gamma'. \end{aligned} \tag{7}$$

Proof. Pro $\sigma'(x) = \sigma(x) + \mathrm{ad}_{A(x)}$ lze volit

$$\begin{aligned} \gamma'(x,y) &:= \gamma(x,y) + A([x,y]) + \sigma(x)A(y) - \sigma(y)A(x) + [A(x), A(y)], \\ &= \gamma(x,y) + \mathrm{d}_\sigma\, A(x,y) + [A(x), A(y)] \end{aligned} \tag{8}$$

□

Věta 1.4. Pro dané θ se libovolné dvě obstrukce f^{σ_1,γ_1}, f^{σ_2,γ_2} liší o kohranici.

Proof. Nejprve využijeme předchozí věty a najdeme γ' takové, že $f^{\sigma_1,\gamma_1} = f^{\sigma_2,\gamma'}$ a poté ukážeme, že $\gamma' - \gamma_2$ je hledaná kohranice. □

Nyní víme, že všechny obstrukce patří do jedné třídy ekvivalence. Následující věta ukazuje, že tato třída ekvivalence je tímto způsobem úplně pokryta.

Věta 1.5. Pro libovolnou obstrukci f^{σ,γ_σ} a libovolný ekvivalentní kocykl g platí, že g je také obstrukce.

Proof. Ukážeme, že $g = f^{\sigma,\gamma_\sigma+\beta}$, kde β je prvek jehož kohranice je rozdíl $g - f$. Snadno se ukáže, že $\gamma_\sigma + \beta$ je kompatibilní se σ. □

Jelikož obstrukce pokrývají právě jednu kohomologickou třídu, můžeme definovat „globání obstrukci“

$$f_\theta := [f^{\sigma,\gamma_\sigma}] \tag{9}$$

a shrnout naše poznatky v následujícím tvrzení.

Věta 1.6. Rozšíření $\mathfrak{g}$ pomocí $\mathfrak{h} \xrightarrow{\theta} \mathfrak{Out}(\mathfrak{g})$ existuje právě tehdy, když $f_\theta = 0$.

Proof. Pokud rozšíření existuje, odpovídá mu jedna nulová obstrukce, takže třída ekvivalence f_θ musí být nutně nulová. Pokud je globální obstrukce nulová mohu vzít libovolnou obstrukci f^{σ,γ_σ}, ta je nutně ekvivalentní 0 a pomocí postupu ve větě 1.5 najdeme nulovou obstrukci. □

2 Řešitelné rozšíření

V sekci 1 jsme rozebrali, jak lze definovat struktura lineární algebry na direktním součtu $\mathfrak{h} \dotplus \mathfrak{g}$. V této sekci definujeme, co se myslí obecným rozšířením a poté se omezíme na rozšíření řešitelná.

Definice 2.1. Rozšíření algebry $\mathfrak{g}$ o algebru $\mathfrak{h}$ je uspořádaná trojice $(i, \mathfrak{E}, \pi)$, taková, že

1. $\pi : \mathfrak{E} \to \mathfrak{h}$ je homomorfismus algeber,
2. $i : \mathfrak{g} \to \mathfrak{E}$ je homomorfismus algeber
3. krátká posloupnost

$$0 \longrightarrow \mathfrak{g} \xrightarrow{i} \mathfrak{E} \xrightarrow{\pi} \mathfrak{h} \longrightarrow 0 \tag{10}$$

je exaktní krátkou posloupností (SES).

Tato definice v sobě zahrnuje fakt, že $\mathfrak{g}$ lze interpretovat jako ideál algebry $\mathfrak{E}$, neboť z exaktnosti posloupnosti (10) plyne, že $\mathfrak{g}$ je isomorfní jádru zobrazení π. Navíc lze korektně sestrojit homomorfismus algeber $\theta : \mathfrak{h} \to \mathfrak{Out}(\mathfrak{g})$ předpisem $\theta(x) := i^{-1} \circ \mathrm{ad}_{\pi^{-1}x} \circ i$, což není nic jiného, že se vezme libovolný π-vzor x, najde se odpovídající vnitřní derivace a ta se zúží na $\mathfrak{g}$.

Důsledky definice v předchozím odstavci, spolu s faktem, že $\mathfrak{E}$ je jako vektorový prostor isomorfní $\mathfrak{h} \dotplus \mathfrak{g}$ ukazují, že postupem v sekci 1 opravdu vytvoříme rozšíření algebry $\mathfrak{g}$ o algebru $\mathfrak{h}$. (Dokonce takto dostaneme, až na isomorfismus SES, všechna rozšíření.)

Řešitelné rozšíření $\mathfrak{g}$ je speciální případ rozšíření algebry, ve kterém $\mathfrak{E}$ je řešitelná a $\mathfrak{g}$ je jejím nilradikálem. Požadavek nilradikality nám dá několik podmínek. Zaprvé $\mathfrak{g}$ musí být nilpotentní, potom, jelikož $\mathfrak{h} \simeq \mathfrak{E} / \mathfrak{g}$, vyžadujeme po algebře $\mathfrak{h}$, aby byla abelovská a do třetice chceme, aby $\mathfrak{h}$ působilo na $\mathfrak{g}$ nilindependentně. To znamená, že pokud zvolíme libovolný nenulový vektor x z doplňku $\mathfrak{g}$ do $\mathfrak{E}$, tak vnitřní derivace ad_x nesmí být nilpotentní (jinak by x také patřilo do nilradikálu).

3 Klasifikace

Teď, když víme, že rozšíření nejsou nic jiného než krátké exaktní posloupnosti (SES), můžeme je klasifikovat. Na množině rozšíření algebry $\mathfrak{g}$ o algebru $\mathfrak{h}$ se zavádí dvě relace ekvivalence, první je definována pomoci isomorfie SES a druhá, jemnější, pomocí ekvivalence SES.

Definice 3.1 (Isomorfie SES)**.** Dvě krátké exaktní posloupnosti (SES)

$$0 \longrightarrow A_i \xrightarrow{\alpha_i} B_i \xrightarrow{\beta_i} C_i \longrightarrow 0 \tag{11}$$

jsou isomorfní, právě když existuje trojice isomorfismů (a, b, c) takových, že diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A_1 & \xrightarrow{\alpha_1} & B_1 & \xrightarrow{\beta_1} & C_1 & \longrightarrow & 0 \\ & & \uparrow{\scriptstyle a} & & \uparrow{\scriptstyle b} & & \uparrow{\scriptstyle c} & & \\ 0 & \longrightarrow & A_2 & \xrightarrow{\alpha_2} & B_2 & \xrightarrow{\beta_2} & C_2 & \longrightarrow & 0 \end{array} \tag{12}$$

komutuje.

Požadavky v definici lze samozřejmě zeslabit. Z existence isomorfismů a, b plyne existence isomorfismu c, stejně jako z existence isomorfismů b, c plyne existence a. Případně takzvaná věta o třech morfismech říká, že pokud a, c jsou isomorfismy a existuje homomorfismus b, tak b je také isomorfismus.

Kromě isomorfie krátkých exaktních posloupností se ještě definuje jejich ekvivalence.

Definice 3.2 (Ekvivalence SES)**.** Dvě krátké exaktní posloupnosti (SES), pro které $A_1 = A_2$ a $C_1 = C_2$ jsou ekvivalentní, pokud existuje homomorfismus $\Xi : B_2 \to B_1$ takový, jsou isomorfní pomocí trojice $(\mathbb{1}, \Xi, \mathbb{1})$. Neboli komutuje diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \xrightarrow{\alpha_1} & B_1 & \xrightarrow{\beta_1} & C & \longrightarrow & 0 \\ & & \| & & \uparrow{\scriptstyle \Xi} & & \| & & \\ 0 & \longrightarrow & A & \xrightarrow{\alpha_2} & B_2 & \xrightarrow{\beta_2} & C & \longrightarrow & 0 \end{array} \tag{13}$$

Je vidět, že se jedná o silnější podmínku, než isomorfismus, protože fixujeme první a třetí homomorfismus na identitu.

Obě relace ekvivalence mají své výhody. Hrubší relace se používala ve [2–4], neboť platí následující věta.

Věta 3.3. Mějme dvě řešitelná rozšíření $(i_\varepsilon, \mathfrak{E}_\varepsilon, \pi_\varepsilon)$, kde $\varepsilon = 1, 2$. Algebry $\mathfrak{E}_1$ a $\mathfrak{E}_2$ jsou isomorfní právě tehdy, když jsou rozšíření $(i_\varepsilon, \mathfrak{E}_\varepsilon, \pi_\varepsilon)$ isomorfní jako SES.

Proof. Směr zprava doleva plyne z definice isomorfie SES. Pro opačný směr musíme nalézt automorfismus a z definice (3.1), za předpokladu, že známe b. Existence c je pak zaručena. Využijeme faktu, že obraz $i_\varepsilon(\mathfrak{g})$ je nilradikálem $\mathfrak{g}$, protože uvažujeme řešitelná rozšíření. Nilradikál je jednoznačně určen a můžeme tedy korektně složit $a := i_1^{-1} \circ b \circ i_2$. Snadno je vidět, že levý čtverec v (12) komutuje. □

Pokud nás tedy zajímají kolik máme tříd neisomorfních algeber, jež jsou řešitelnými rozšířeními zadaného $\mathfrak{g}$, stačí nám zajímat se o neisomorfní SES. Na druhou stranu ekvivalence rozšíření nám pomocí následující věty, kterou ponecháme bez důkazu, umožní zapojit kohomologické metody.

Věta 3.4. Pro pevné $\theta : \mathfrak{h} \to \mathfrak{Out}(\mathfrak{g})$ jsou třídy neekvivalentních rozšíření (pokud alespoň jedno rozšíření existuje) 1-1 k $H^2(\mathfrak{h}, \mathfrak{C}(\mathfrak{g}), \rho_\theta)$.

V této sekci jsme se dozvěděli, že abychom klasifikovali řešitelná rozšíření algebry $\mathfrak{g}$ o k vektorů, je potřeba vzít všechny homomorfismy $\theta : \mathfrak{a}_k \to \mathfrak{Out}(\mathfrak{g})$, kde $\mathfrak{a}_k$ je abelovská algebra o dimenzi k. Pro dané pevné θ je potřeba vybrat libovolný řez $\sigma : \mathfrak{a} \to \mathfrak{Der}(\mathfrak{g})$ a nejprve zkontrolovat nilindependenci, t.j. podívat se, zda jediný nilpotentní operátor v $\sigma(\mathfrak{a})$ je ten nulový. Následně vybereme libovolné $\gamma : \mathfrak{h} \wedge \mathfrak{h} \to \mathfrak{g}$ kompatibilní podle vzorce (5). Pokud je obstrukce $f^{\sigma,\gamma} = 0 \bmod B^3(\mathfrak{a}, \mathfrak{C}(\mathfrak{g}), \rho_\theta)$, pak rozšíření existuje a množinu neekvivalentních rozšíření s daným θ lze parametrizovat pomocí $H^2(\mathfrak{a}, \mathfrak{C}(\mathfrak{g}), \rho_\theta)$.

4 Hypotézy

V této sekci budeme prezentovat dvě hypotézy, které jsme vyvodili z výsledků z [2–4] a dalších, v nich citovaných, článků.

Hypotéza 1a Řešitelné rozšíření s maximální dimenzí je jednoznačné ve smyslu, že pro daný řez σ existuje právě jedno kompatibilní γ.

Hypotéza 2a Pokud pro dané θ existuje alespoň jedno rozšíření, lze zvolit σ tak, že jeho hodnoty jsou v centru $\mathfrak{C}(\mathfrak{g})$.

Nyní můžeme využít našich definic a přeformulovat naše hypotézy do řeči kohomologií a krátkých exaktních posloupností. V té má první hypotéza obzvláště jednoduchý tvar.

Hypotéza 1b Pro rozšíření s maximální dimenzí, které je dané zobrazením θ, je $H^2(\mathfrak{a}, \mathfrak{C}(\mathfrak{g}), \rho_\theta) = 0$.

Hypotéza 2b Pokud pro dané θ existuje alespoň jedno rozšíření, existuje rozšíření $(i, \mathfrak{E}, \pi)$ takové, že π má levou inverzi (takzvané splittable rozšíření).

Nyní lze snadno odvodit hlubší důvod, proč první hypotéza platila pro naše zkoumané příklady. K tomu použijeme větu, kterou lze najít dokázanou například v [1].

Věta 4.1. Nechť $\mathfrak{a}$ je nilpotentní algebra, ρ reprezentace této algebry na $\mathfrak{b}$. Potom jsou následující tři tvrzení ekvivalentní.

1. $H^0(\mathfrak{a}, \mathfrak{b}, \rho) = 0$.
2. $H^1(\mathfrak{a}, \mathfrak{b}, \rho) = 0$.
3. $H^n(\mathfrak{a}, \mathfrak{b}, \rho) = 0, \ \forall n \in \mathbb{N}$.

Tuto větu nyní použijeme. Algebra $\mathfrak{a}$ bude abelovská algebra, kterou rozšiřujeme, roli $\mathfrak{b}$ bude hrát $\mathfrak{C}(\mathfrak{g})$ a reprezentace bude ρ_θ. Ve všech případech maximálního rozšíření, které jsme zkoumali byl řez σ vždy regulární, t.j. operátory $\sigma(x)$ pro $x \in \mathfrak{a}$ neměly netriviální společné jádro. Tím více nemají společné jádro, když je zúžíme na $\mathfrak{C}(\mathfrak{g})$. A tvrzení, že $\sigma(\mathfrak{a})\restriction_{\mathfrak{C}(\mathfrak{g})}$ nemají společné jádro není nic jiného než přeformulované tvrzení, že $H^0(\mathfrak{a}, \mathfrak{C}(\mathfrak{g}), \rho_\theta) = 0$. A z věty 4.1 pak plyne, že kohomologická grupa $H^2(\mathfrak{a}, \mathfrak{C}(\mathfrak{g}), \rho_\theta) = 0$, a to je tvrzení naší hypotézy.

Z věty 4.1 plyne také pomocí podobné úvahy, že pro libovolné nilindependentní regulární θ rozšíření existuje ($H^3(\mathfrak{a}, \mathfrak{C}(\mathfrak{g}), \rho_\theta) = 0$) a je jednoznačné (až na ekvivalenci).

Tyto úvahy nám toho samozřejmě neříkají moc o tom, proč bylo pro maximální rozšíření θ regulární. To je předmětem našeho dalšího zkoumání, což platí i pro naši druhou hypotézu.

References

[1] D. W. Barnes. *On the cohomology of soluble Lie algebras.* Math. Z. **101** (1967), 343–349.

[2] 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.

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

[4] L. Šnobl and P. Winternitz. *All solvable extensions of a class of nilpotent Lie algebras of dimension n and degree of nilpotency* $n-1$. J. Phys. A **42** (2009), 105201, 16 pp.

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 the method of algebraic multigrid and its parallelization on GPU. Algebraic multigrid is a sparse matrix iterative solver, which finds the solution by solving also restricted versions of the original problem. The main difference from more widely known geometric multigrid is that it can create the restricted problems without any knowledge about the matrix origin and therefore it can be used for larger range of problems. The article farther presents possibilities how to parallelize this algorithm on multicore CPU architecture and on GPU. Finally it also shows speedup obtained by the GPU parallelization.

Keywords: GPU, Algebraic Multigrid, Parallelization

Abstrakt. Tento článek se zabývá metodou algebraického multigridu a její paralelizací na GPU. Algebraický multigrid je iterační metoda pro řešení soustav rovnic s řídkou maticí, využívající k řešení restrikce problému na menší soustavy. Na rozdíl od geometrického multigridu nepotřebuje k vytvoření podproblémů znalost původní úlohy, ze které matice pochází, což jej činí mnohem více univerzálním. Článek se dále zabývá možnostmi paralelizace tohoto algoritmu a to jednak pro vícejádrové proceseory a druhak pro grafické karty. V závěru je představeno zrychlení, kterého bylo paralelizací na GPU dosaženo.

Klíčová slova: GPU, Algebraický Multigrid, Paralelizace

1 Introduction

Multigrid methods are a group of algorithms for solving differential equations using a hierarchy of discretizations, they can be used as solvers as well as preconditioners. Convergence analysis shows that many standard iterative solvers can quickly eliminate high-frequency parts of errors, but not the low-frequency ones. The main idea of multigrid methods is therefore to solve problem also on hierarchy of coarser grids, where formerly low-frequency parts of error become high-frequency ones. The solution from coarser grids can be then used to improve solution on the original grid, which should yield a significant improvement in convergence speed. The typical application for multigrid is in the numerical solution of elliptic partial differential equations in two or more dimensions.

Main class of this method (the so called geometric multigrid methods) has the issue that coarser grids and transition operators between them must be provided as part of the original problem and are defined based on the used discretization shceme and geometry. This restricts the use of geometric multigrid as black-box solver and limits the types of problems it can be used for.

Algebraic multigrid methods (AMG), on the other hand, construct their hierarchy of operators directly from the system matrix, and the levels of the hierarchy are simply subsets or aggregations of original unknowns without any geometric interpretation. Thus, AMG methods can be used as true black-box solvers for sparse matrices, however their are mostly little less effective, that their geometric counterpart.

As with all iterative methods if solution of really big problems is desired some kind of parallelization must be used. This article deals with parallelization of multigrid methods on multi-core architecture with shared memory, specifically on multi-core CPUs and GPUs. GPU (graphical processing unit) is a special piece of hardware designed to improve visual quality of computer games. It has highly parallel architecture and outperforms processors both in computational power and memory bandwidth, which makes it very suitable for efficient numerical programming.

2 Algebraic multigrid

Main part of AMG is the creation of coarser version of the original problem. This is achieved through following steps:

- Selecting variables which will form coarser grid
- Defining transition operators
- Creating coarse problem matrix

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 Coarse/Fine grid splitting

First part is to choose unknowns which will form the coarser grid. There are two requirements on the coarser grid:

1. Must correctly approximate the problem.
2. Must have substantially fewer points.

First requirement is however quite general so it will have to be explained a bit more. For coarser grid to correctly approximate the finer one it is needed that all unknowns which aren't in the coarser grid can be approximately calculated from the ones that are in the coarser grid. The problem matrix A describes how each of the unknowns depend on it's own value (we expect A to be M-matrix so the diagonal entry should be dominant) and on the values of other unknowns (non-diagonal entries). Therefore it is logical the expect that some good approximation of the missing unknowns can be obtained if one knows the values of unknown which the missing ones depend upon. Let us 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

$$-a_{i,j} \geq \theta max_{k \neq i}(-a_{ik}). \tag{1}$$

This means that variable u_i strongly depends on the variable u_j if the coefficient a_{ij} is comparable in magnitude to the largest off-diagonal coefficient in the ith equation. We can state this also from inverse perspective.

Definition 2: If the variable u_i strongly depends on the variable u_j, then the variable u_j *strongly influences* the variable u_i.

Let us denote C as the set of all unknowns which will be chosen for coarse grid, F as all unknowns that won't and S_i all unknowns, that strongly influence unknown i. Then given the previous definitions we can more exactly specify coarse grid requirements as

1. For each unknown $i \in F$, every unknown $j \in S_i$ (that strongly influence i) either should be in the C or should strongly depend on at least one point in C

2. The set of coarse unknowns C should be a maximal subset of all unknowns with the property that no unknown $i \in C$ strongly depends on any other unknown $j \in C, j \neq i$.

It is not always possible to enforce both these rules. In such cases we prefer to fulfill the first one. While this choice may lead to larger coarse grids than necessary, experience shows that this trade-off between accuracy and expense is generally worthwhile[1].

The basic coarsening algorithm can look as follows:

1. Evaluate all unknowns, based on the number of other variables they strongly influence

2. Take one with biggest score (in case there is more than one, select any of them) and put in C

3. Put all unknowns that strongly depend on it to F

4. Reevaluate all affected unknowns

5. Repeat from 2

2.2 Defining transition operators

When the coarse grid has been selected, the next goal is to define transition operators. Starting with the interpolation one I_{2h}^h(although physical grids may not be present, we continue to denote fine-grid quantities by h and coarse-grid quantities by $2h$) we require that the ith component of $I_{2h}^h e$ be given by

$$(I_{2h}^h e)_i = \begin{cases} e_i & \text{if } i \in C \\ \sum_{j \in C_i} w_{ij} e_j & \text{if } i \in F \end{cases} \tag{2}$$

where the interpolation weights are defined, according to [1] by

$$w_{ij} = -\frac{a_{ij} + \sum_{m \in D_i^s} \left(\frac{a_{im} a_{mj}}{\sum_{k \in C_i} a_{mk}} \right)}{a_{ii} + \sum_{n \in D_i^w} a_{in}}, \tag{3}$$

where C_i is the set of the coarse-grid unknowns $j \in C$ that strongly influence i, D_i^s is the set of the neighboring unknowns $k \in F$ that strongly influence i, and D_i^w is the set of all neighboring unknowns that do not strongly influence i.

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

$$I_h^{2h} = (I_{2h}^h)^T, \tag{4}$$

and restricted matrix is produced by

$$A^{2h} = I_h^{2h} A^h I_{2h}^h. \tag{5}$$

2.3 Multigrid cycle

Main iteration multigrid cycle is same for both algebraic and geometric multigrid. It expects to have the coarse problem matrices and transition operators defined and it finds the solution of given problem by iterative process. Classical *V-cycle* multigrid cycle looks as follows:

1. Start with initial approximate solution x_0^h
2. Relax (do few smoother iterations) the current solution to get new guess x^h
3. Compute the fine-grid residual $r^h = b^h - A^h x^h$
4. Restrict residual to the coarse grid $r^{2h} = I_h^{2h} r^h$
5. Solve $A^{2h} e^{2h} = r^{2h}$
6. Interpolate error correction to fine grid by $e^h = I_{2h}^h e^{2h}$
7. Correct current solution $x^h = x^h + e^h$
8. Repeat from 2 (if needed)

This was case of two grid hierarchy, if more grids are to be used, simply replace the direct solution of the coarse-grid problem with a recursive call to this algorithm on all grids except the coarsest one.

3 Multigrid parallelization

Main part of the computational time is normally spent in iteration cycle. This cycle as was already described, consist of transition operators and some relaxation, which is the most critical part of whole algorithm, therefore it's parallelization will be described first.

3.1 Relaxation scheme

As smoother we use either damped Jacobi method or Gauss-Seidel method. Jacobi method is quite easy to parallelize as it doesn't have any dependence in calculation of new unknown values, each thread can be used to calculate any number of them. Gauss-Seidel on the other hand is inherently sequential and thus cannot be directly parallelized. Therefore it was needed to switch to its Red-Black version, which is easily parallelizable even though it has the disadvantage, that it can't be used for general matrices. This can be solved by multi-color coloring, but that wasn't implemented yet.

3.2 Transition operators

Simple Matrix-vector multiplication is used to convert quantities from one grid to another. So it can be easily parallelized over unknowns,

```
void Mult(const mat & A, const arr & x, arr & res)
{
 #pragma omp parallel for schedule(static)
 for (int r = 0; r < A.getRowsCount(); r++)
 {
  res[r] = 0;
  for (int i = 0; i < A.getRowSize(r); i++)
    res[r] += A.getRowValue(r,i) * x[A.getColIndex(r,i)];
 }
}
```

3.3 Coarsening

Most challenging part to paralelize is the coarsening algorithm. It needs to select many coarse unknowns at once but also ensure that there aren't strong dependencies between them. 2 standard strategies exist.

Grid decomposition Which divides the grid to smaller ones and each thread splits one of these sub-grids and afterwards boundaries are solved in some less parallel way. This version is however not really suitable for architectures with large number of threads (e.g. graphic cards), because then processing boundary points becomes the main computational part.

Noise adding Adds random values to score of each unknown, which creates local maxims, which can be chosen for coarse grid in parallel. However there are number of disadvantages to this system:

- Obtained results are random and will differ each time.
- More complicated structure for storing point scores must be created.
- Still difficult to paralelize effectively.

Because none of the method seems to be easily usable this part of the algorithm wasn't parallelized yet and will be dealt with later if it starts to hinder the computations.

4 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 several technologies, that allow programmer to create application for GPU, but most important are [6]:

- OpenGL - It is cross-platform graphical API so basic knowledge about computer graphics is needed and general problems have to be inconveniently masked as a graphical ones. This was the first way how graphics card could be used to solve general problems, but nowadays it is once again used only for graphics.

- CUDA - Is a technology from NVidia company designed specifically for general purpose computing on graphics cards, main disadvantage is that it only works with NVidia graphics cards. Advantages are that it is being rapidly developed and there exist a lot of example and documentation for it.

- OpenCL - Newest technology for general computation on graphics card, it is an industry standard and so it can be used for almost all new devices ranging from graphics cards to cell phones.

For the purpose of this article CUDA was used rather than OpenCL. However core parts of both these technologies are very similar, the main difference is only in the naming of the API functions.

4.1 GPU specifics

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.

4.2 GPU implementation

GPU implementation of the parallel algorithm was quite straightforward, only needed change was that if Red-Black Gauss-Seidel method is to be used, data must be reorder so that coalescing can occur during Red/Black phase (i.e. they must be reordered according to colors). Apart from that all parts of algorithm were easily parallelized by the manner that each unknown is handled by one thread and there is no communication between threads. This means:

- In matrix-vector multiplication each thread multiply one row of matrix with the given vector and write one value of final vector
- In Jacobi/Gauss-Seidel iteration each thread actualize one unknown
- In computation of residuum each thread computes one value of final vector (similar to matrix-vector multiplication).

One notable thing is that GPU isn't well suited for restriction operations like sum or min/max search. Algorithm however needs to compute L2 norm of residuum which is a sum type operation. To accomplish this efficient algorithm from [5] was used:

```
__global__ void Norm2Kernel(double *in, double * out, int N)
{
        __shared__ float sdata[CUDA_SUM_BLOCK_SIZE];
        const unsigned int tid = threadIdx.x;
        sdata[tid] = 0;

        for (int i = tid; i < N; i+=CUDA_SUM_BLOCK_SIZE)
        {
                sdata[tid] += in[i]*in[i];
        }
        __syncthreads();

        for( unsigned int s = blockDim.x/2 ; s > 0 ; s >>= 1 )
        {
                if( tid < s ) sdata[tid] += sdata[tid + s];
                __syncthreads();
        }

        if( tid == 0 ) out[0]= sdata[0];
}
```

and residuum was calculated only once each 10 smoother iterations.

All these operations are relatively undemanding on arithmetic computations, so their bottleneck is the memory bandwidth. Therefore coalescing became crucial for efficient implementation. Fortunately apart from red/black reordering it was quite easy to achieve it without too much added code complexity.

Also RAM/VRAM data transfers, don't create any issue because whole multigrid cycle is implemented on the gpu. So the problem is only once copied to VRAM and after the computation final result is copied back.

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 classical iterative solvers with both multigrid methods. It clearly shows that multigrid methods are faster by the order of magnitude and that geometric multigrid is faster than the more general algebraic one, but the difference isn't too big.

Second table (Tab 2) shows the same problem, this time it compares multigrid methods implemented on CPU and GPU.

Last table (Tab 3) shows again comparison of CPU and GPU version of multigrid but for a larger problem. Interesting fact, that should be noted here, is that GPU speed-up is worse than for the smaller problem. This is strange because GPU normally performs better for larger problems as the effect of GPU overhead becomes negligible. This issue should be looked into in the future.

	Time	Rel. Speed-up
Jacobi	1419 s	1
Gauss-Seidel	952 s	1.5
Geometric-Multigrid	87 s	16.3
Algebraic-Multigrid	113 s	12.6

Table 1: Comparison of different CPU solvers. For the 2D Laplace equation with 65 000 degrees of freedom

	Time	Speed-up
Geometric-multigrid	87 s	
GPU Geometric-multigrid	7 s	12.4
Algebraic-Multigrid	113 s	
GPU Algebraic-Multigrid	18 s	6.2

Table 2: Comparison of multigrid solvers on CPU and GPU. For the 2D Laplace equation with 65 000 degrees of freedom

Also because multicore architectures are quite common nowadays it would be good the compare GPU version also with version parallelized over OpenMP or similar technology. This was tested, however the obtained speed-up was quite insignificant (about 20% for 4 core CPU), so we suspect that there was some fundamental flaw and so we won't compare GPU implementation with this version until further testing.

	Time	Speed-up
Geometric-multigrid	1218 s	
GPU Geometric-multigrid	143 s	8.5
Algebraic-Multigrid	1538 s	
GPU Algebraic-Multigrid	276 s	5.6

Table 3: Comparison of multigrid solvers on CPU and GPU. For the 2D Laplace equation with 262 000 degrees of freedom

6 Summary

This article presented key principles of Algebraic multigrid and its parallelization as well as basics of GPU programming. The algebraic multigrid algorithm was successfully implemented and parallelized on GPU. It was proven that AMG is suitable for GPU implementation and it can be accelerated more than five times. In the future we would like to create also proper OpenMP implementation to compare it with the GPU version, test different coarsening strategies and use the AMG solver in our program for incompressible flow simulations.

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] V. Klement, *Diplomová práce: Implementace řešičů řídkých matic na GPU*, ČVUT-FJFI, 2011

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

[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

Dynamically Evolving Dislocations*

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 numerical simulation of dislocation dynamics. Dislocations are line defects in crystalline lattice causing the disturbance of the regularity of the crystallographic arrangement of atoms. From a mathematical point of view, the dislocations are defined as smooth closed or open planar curves which evolve in time. The motion itself is only two-dimensional and is driven by the equation for the mean curvature flow. We describe the evolving curves by parametric approach and the model is numerically solved it by means of semi-implicit finite differences and flowing finite volumes method. However, numerical experiments show this model exhibits unintended behaviour, since during the evolution, the grid points are accumulated in certain segments. We overcome this problem by adding the tangential velocity to the model, which does not affect the shape of the curve.

Keywords: dislocations, mean curvature flow, tangential redistribution

Abstrakt. Tento příspěvek se zabývá simulací dislokační dynamiky. Dislokace jsou čárové poruchy v krystalové mřížce způsobující nepravidelnost v uspořádání atomů. Z mamematického hlediska jsou dislokace popsány jako hladké uzavřené nebo otevřené planární křivky, které se vyvíjejí v čase. Jejich pohyb je popsán rovnicí pro pohyb křivek řízený jejich střední křivostí. Pohybující se křivky jsou definovány parametrickým popisem a model je numericky řešen semi-implicitní metodou založenou na metodě konečných diferencí nebo plovoucích konečných objemů. Numerické experimenty však ukazují, že tento model vykazuje defektní chování v podobě toho, že v průběhu časového vývoje se uzlové body křivky nahromadí v jistých segmentech. Tento nechtěný jev je řešen dodáním tečné složky rychlosti do modelu.

Klíčová slova: dislokace, pohyb křivek řízený střední křivostí, redistribuce

1 Introduction

Dislocations are line defects of the crystalline lattice. They acts in such a way that the crystallographic arrangement of atoms is disturbed along the dislocation line. Theoretical framework about the dislocations theory is extensivelly discussed in literature such as [1, 2]. From the mathematical point of view, the dislocations can be represented as a closed (inside the crystal) or open (ending on a surface of the crystal) curves, which can evolve in time and space. At a certain physical conditions, e.g. at low homologous temperatures, the dislocations can move only along so called slip planes, i.e. some crystallographic planes with the highest density of atoms.

*This work has been supported by the grant Two scales discrete-continuum approach to dislocation dynamics, project No. P108/12/1463 of the Grant Agency of the Czech Republic.

2 Curve Evolution in Plane

The dimensionless mathematical model of evolving dislocation curve Γ^t can be described by the equation for the mean curvature flow, which reads as

$$v = -k_\Gamma + F, \tag{1}$$

where v is the normal velocity, k_Γ is the mean curvature and F is the force term acting on the curve Γ^t in the normal direction. Our goal is to find a family $\{\Gamma^t : t \in [0, T_{max}]\}$ of closed or open curves in $\mathbb{R}^2$, whose normal velocity is proportional to the mean curvature, i.e. satisfying the equation for the mean curvage flow (1). Nowadays, there exist several approaches to treat the equation (1). Very popular methods come from the family of interface-tracking approaches, such as the phase-field method [5, 6] or the level set method [3, 4]. It is often reffered, that their main advantage is the ability to deal with the topological changes like merging or splitting with almost no difficulties. Hovewer, when considering to use such approaches for a dislocation dynamics problems, where is often required a longer time evolution, one might to experience some difficulties in the computational costs since in the case of a planar curve, it is required to solve the two-dimensional problem to obtain the curve, which is an one-dimensional object.

Very fast method for the time evolution of curves is provided by the parametric (sometimes referred as direct or Lagrangian) approach [7]. The curve Γ^t can be parametrized either by some fixed interval or directly by its length (so called arc-length parametrization). The parametric approach, hovewer, can not deal with the topological changes on its own, thus it requires development of separate algorithms to treat such changes.

In this contribution, we focus on the parametric approach. In this case, the planar curve Γ^t is given as the following set

$$\Gamma^t = \{\mathbf{X}(u,t) = (X_1(u,t), X_2 u, t) : u \in I_u\},$$

where the curve is described by the spatially and time dependent vector function called parametrization

$$\mathbf{X} : I_u \times I_t \to \mathbf{R}^2,$$

where $I_u = [0,1]$ is the fixed interval for the parameter u and $I_t = [0,T]$ is the time interval. The unit tangential vector $\mathbf{t}$ is defined as $\mathbf{t} = \partial_u \mathbf{X}/|\partial_u \mathbf{X}|$. The unit normal vector $\mathbf{n}$[1] is defined as $\mathbf{n} = \partial_u \mathbf{X}^\perp/|\partial_u \mathbf{X}|$, where the $\mathbf{X}^\perp$ is vector perpendicular to the $\mathbf{X}$, and hence the relation $\mathbf{t} \cdot \mathbf{n} = 0$ holds. The normal velocity v is defined as the time derivative of $\mathbf{X}$ projected into the normal direction

$$v = \partial_t \mathbf{X} \cdot \frac{\partial_u \mathbf{X}^\perp}{|\partial_u \mathbf{X}|}.$$

According to the Frenet formulae, one can determine the curvature k_Γ from the following relation

$$\frac{1}{|\partial_u \mathbf{X}|} \partial_u \mathbf{t} = -k_\Gamma \frac{\partial_u \mathbf{X}^\perp}{|\partial_u \mathbf{X}|}. \tag{2}$$

[1] in the case of closed curve, the outer normal vector is considered; in the case of open curve, there is a selected pre-defined direction

Differentiating the left-hand side of (2) and using the perpendicularity condition, one can derive the formula for the curvature as the following

$$-k_\Gamma = \frac{\partial_{uu}\mathbf{X}}{|\partial_u\mathbf{X}|^2} \cdot \frac{\partial_u\mathbf{X}^\perp}{|\partial_u\mathbf{X}|}.$$

To obtain the parametric equations, we can substitute the previous relations into the equation for the mean curvature flow (1) and multiply it by $\mathbf{n}$

$$\partial_t\mathbf{X} = \frac{\partial_{uu}\mathbf{X}}{|\partial_u\mathbf{X}|^2} + F(\mathbf{X},t)\frac{\partial_u\mathbf{X}^\perp}{|\partial_u\mathbf{X}|}. \tag{3}$$

The equation (3) is complemented with some initial condition

$$\mathbf{X}|_{t=0} = \mathbf{X}_{ini}$$

and appropriate boundary conditions. In the case of a closed curve, the periodic boundary condition is set

$$\mathbf{X}|_{u=0} = \mathbf{X}|_{u=1}.$$

For the open curves we choose the fixed ends boundary condition, i.e.

$$\mathbf{X}|_{u=0} = \mathbf{X}_0, \mathbf{X}|_{u=1} = \mathbf{X}_1.$$

3 Tangential Redistribution of the Grid Points

It is known when tracking a curve motion, the tangential terms do not affect its shape (see [11], Proposition 2.4) and hence it is sufficient for the analysis to take into the account only the terms in the normal direction to the curve. Hovewer, numerical experiments show that the parametric equations (3) are not apropriate for the numerical computation. Since the curve is discretized by a certain number of grid points, except the perfectly symmetric and uniform situations with constant curvature, like a shrinking circle, 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 employ some kind of tangential redistribution, i.e. to complement the equation (3) with a term standing for the tangential velocity

$$\partial_t\mathbf{X} = \frac{\partial_{uu}\mathbf{X}}{|\partial_u\mathbf{X}|^2} + \alpha\frac{\partial_u\mathbf{X}}{|\partial_u\mathbf{X}|} + F(\mathbf{X},t)\frac{\partial_u\mathbf{X}^\perp}{|\partial_u\mathbf{X}|}. \tag{4}$$

The term α, usually called redisribution coefficient, is a function of curvature (and hence position) and time, thus $\alpha = \alpha(k,t) = \alpha(\mathbf{X},t)$. 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 proposed by D. Ševčovič

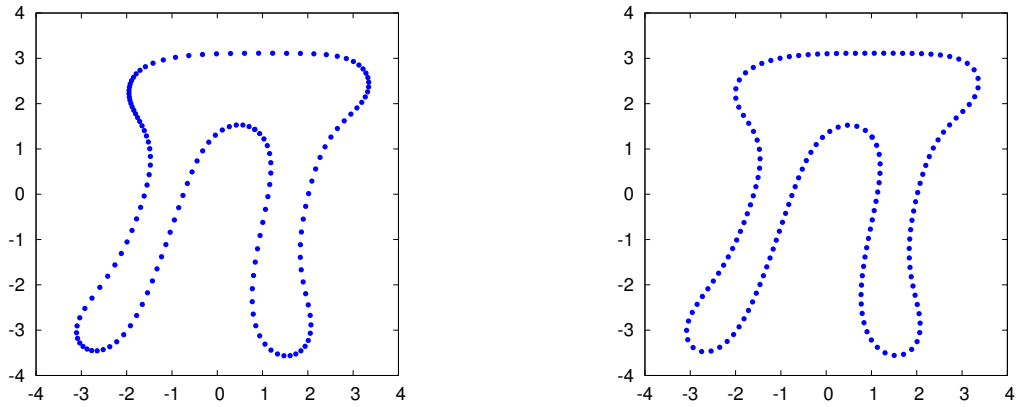


Figure 1: The impact of the tangential redistribution. On the left figure there is a case without the tangential velocity, the curve on the right figure was computed with the usage of the uniform redistribution.

and S. Yazaki in [12], in which one can also find a brief overview and a critical discussion of redistribution methods. The impact of the tangential redistribution is shown on the Figure 1.

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

$$\partial_s(\varphi(k)\alpha) = f - \frac{\varphi(k)}{\langle\varphi(k)\rangle}\langle f\rangle + \omega\left(\frac{L^t}{|\partial_u \mathbf{X}|}\langle\varphi(k)\rangle - \varphi(k)\right), \tag{5}$$

where ∂_s denotes the derivative with respect to the arc-length, i.e. $\partial_s \mathbf{X} = \partial_u \mathbf{X}/|\partial_u \mathbf{X}|$ and $\mathrm{d}s = |\partial_u \mathbf{X}|\mathrm{d}u$. The quantity L^t is the curve length in time t and ω is a given positive constant. The other factors in the problem (5) are as follows

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

To get the unique solution α of the equation (5), the following additional condition must be considered

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

The function $\varphi(k)$ plays important role because it is proposed to control the redistribution on the grid points. The special choice $\varphi(k) = 1$ produces the uniform redistribution for $\omega = 0$ and asymptotically uniform redistribution for $\omega > 0$. The function $\varphi = |k|$ was proposed for the crystalline curvature flow. Choosing $\varepsilon \in (0, 1)$, we obtain curvature adjusted redistribution [12].

4 Physical Model

Generally, there are several possibilities how to describe the motion of dislocations. We consider the model proposed by Kratochvíl and Sedláček [8], which enables to describe

the dislocation motion law by the mean curvature flow

$$Bv = Tk_\Gamma + F, \tag{6}$$

where the B denotes the drag coefficient equals to $B = 10^{-5}$ Pa·s, T denotes the line tension and F is the sum of all forces acting on the dislocation except the dislocation self-force, which is approximated by the mean curvature. The force term reads as the following

$$F = b(\tau_{app} + \tau_{wall} + \tau_{int} - \tau_{fr}), \tag{7}$$

where b is the magnitude of the Burger's vector $\vec{b}$ – vector which represents the magnitude and the direction of the lattice distorsion of dislocation in a crystal lattice. The particular force terms are caused by various stresses

- τ_{app} is the shear stress applied on the crystal,
- τ_{wall} is the stress from so called PSB channel, where the dislocation moves in,
- τ_{int} is the stress caused by mutual interaction between the dislocations,
- τ_{fr} is the stress caused by crystal lattice resistance, which slows down the movement of dislocation.

The value of τ_{app} is usually chosen in the range of $20 - 70$ MPa, the τ_{fr} is chosen as 5 MPa.

The quantity L is the dislocation line tension. This term is anisotropic and causes straightening of the dislocation curve. According to the [1], it can be approximated as

$$L \approx E^{(e)}(1 - 2\nu + 3\nu\cos^2\zeta),$$

where $E^{(e)}$ is the dislocation edge energy and ν is the Poisson's ratio. The quantity ζ is the angle between the Burger's vector and the segment of the dislocation line.

The motion of dislocation itself is considered within co called PSB (persistent slip band) channel [1, 2, 7]. Generally, it is a patterm consisting of ares with high densities of dislocations (channel walls) and low densities of dislocations (channel itself). This structure usually arises from cyclic loading of a crystal. The behavior of the channel stress field τ_{wall} can be described by the one-dimensional function in the Figure 2.

The problem of mutual interaction was theoreticaly solved by Devincre [9], Devincre's fomula (8) provides the 3D stress tensor field $\tau^A = \tau^A_{ij}$ at a location $\mathbf{x}$ generated by the dislocation half line from the grid point A to infinity

$$\begin{aligned}\tau^A_{ij} =& \frac{G}{4\pi}\frac{1}{R(U+R)}\left[(\vec{b}\times\vec{Y})_i t_j + (\vec{b}\times\vec{Y})_j t_i - \frac{1}{1-\nu}((\vec{b}\times\vec{t})_i Y_j + (\vec{b}\times\vec{t})_j Y_i)\right.\\ &\left. -\frac{\vec{b}\cdot(\vec{\varrho}\times\vec{t})}{1-\nu}\left[\delta_{ij} + t_i t_j + \frac{(\varrho_i t_j + \varrho_j t_i + U t_i t_j)(U+R)}{R^2} + \frac{\varrho_i\varrho_j(2+U/R)}{R(U+R)}\right]\right],\end{aligned} \tag{8}$$

where the meaning of the parameters is as follows

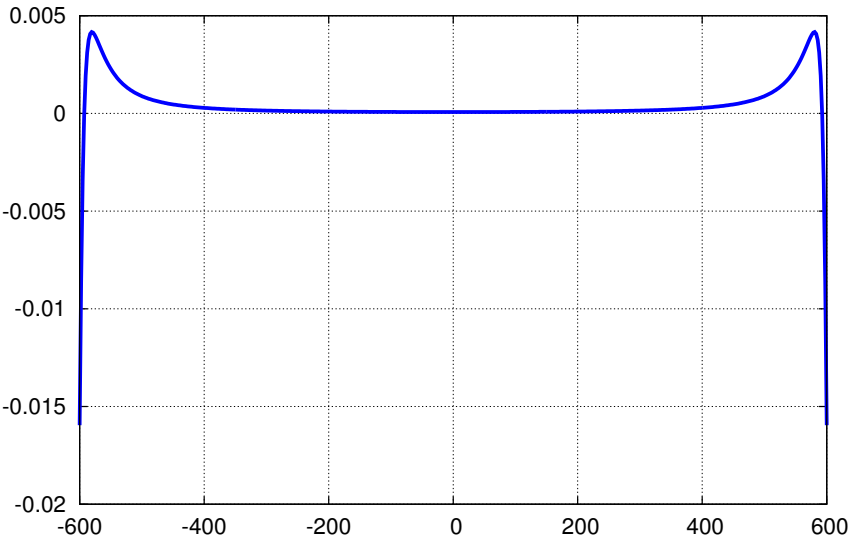


Figure 2: Wall force function, the x-axis is in nm, the y-axis is in N.

$\vec{t}$ tangential vector of the dislocation segment,
$\vec{R}$ $= (R_1, R_2, R_3)^T$ vector to the location x from A,
R $= \sqrt{R_1^2 + R_2^2 + R_3^2}$,
U $= \vec{R} \cdot \vec{t}$,
Y_i $= R_i + Rt_i$,
ϱ $= \vec{R} - U\vec{t}$ normal component of R to the dislocation segment,
G shear modulus.

The stress tensor generated by a straight dislocation segment AB is then given as a difference of tensors τ^A and τ^B, i.e.

$$\tau_{int} = \tau^A - \tau^B.$$

In this contribution, we consider the Burger's vector $\vec{b} = (b, 0, 0)^T$ parallel with x-axis and slip planes, where dislocation moves, parallel with xz-plane in mutual distance h. Generally, to compute the forces acting on the dislocation exposed to a stress field τ_{int} generated by other dislocations can be used so called Peach-Koehler formula [10], which reads as

$$\vec{F}_{int} = (\tau_{int}\vec{b}) \times \vec{t}. \tag{9}$$

Using the formula (9) greatly simplifies the problem, since for the considered model problem of parallel slip planes and Burger's vector parallel with the x-axis it is sufficient to compute only the component τ_{12} of the stress tensor τ_{int}.

5 Numerical Schemes

For the numerical computations we use the fully discrete semi-implicit numerical scheme based on finite differences method

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

where $\mathbf{X}_j^k \approx \mathbf{X}(jh, k\tau)$ for the spatial step h and the time step τ, $\mathcal{Q}(\mathbf{X}) = \sqrt{X_1^2 + X_2^2 + \varepsilon^2}$ serving as the regularization term since we want to avoid dividing by zero. The symbols $\mathbf{X}_{u,j}^k$ and $\mathbf{X}_{uu,j}^k$ denote the first and the second central differences. We also use the semi-implicit scheme based on flowing volumes method proposed by D. Ševčovič and S. Yazaki [12]

$$-a_j^{k+\frac{1}{2}} \tau \mathbf{X}_{j-1}^{k+1} + (1 + b_j^{k+\frac{1}{2}} \tau) \mathbf{X}_j^{k+1} - c_j^{k+\frac{1}{2}} \tau \mathbf{X}_{j+1}^{k+1} = \mathbf{X}_j^k + \tau F \frac{\mathbf{X}_{u,j}^{\perp,k}}{\mathcal{Q}(\mathbf{X}_{u,j}^k)},$$

where

$$a_j^{k+\frac{1}{2}} = \frac{2}{r_j^k + r_{j+1}^k} \left(\frac{1}{r_j^k} - \frac{\alpha_j^{k+1}}{2} \right), c_j^{k+\frac{1}{2}} = \frac{2}{r_j^k + r_{j+1}^k} \left(\frac{1}{r_j^k} + \frac{\alpha_j^{k+1}}{2} \right), b_j^{k+\frac{1}{2}} = a_i^{k+\frac{1}{2}} + c_i^{k+\frac{1}{2}}$$

for the quantity $r_j^k = |\mathbf{X}_j^k - \mathbf{X}_{j-1}^k|$ – line segment representing the control volume.

6 Computational Results

We present the results of the two numerical experiments , when we deal with the interaction of two dislocations on nearby parallel slip planes in the PSB channel. We suppose there are two initial dislocation lines with fixed points in the channel walls, driven by the forces (7) with opposite signs. Each dislocation is located in a different slip plane h apart.

In the first experiment on the Figure 3, the distance between the slip planes is $h = 65$ nm. The interaction force is attractive and speeds up the motion. When the dislocations overlap, the interaction force become repulsive. In this case of a relatively long distance h, the force generated by channel walls and applied stress is greater than the repulsive force and the dislocations continue to glide.

In the second experiment on the Figure 4, the distance between the slip planes is $h = 35$ nm. The interaction force also attracts the dislocations. However, since the slip planes distance is smaller, the interaction force is bigger and when overlapping, the repulsive force stops the movement at a certain position and the dislocations remain in steady state.

7 Conclusion

We have presented the mathematical model of evolving curves based on the parametric approach. The discussed disadvantage of this approach was treated by adding the tangential velocity to the model, which proved to be very useful technique for stabilizing the

algorithm. We have also introduced the physical model of evolving dislocations based on the equation for the mean curvature flow and described several force terms acting on the dislocations. The presented results of numerical simulations show the motion of dislocations in PSB channel and their mutual interaction.

References

[1] D. Hull and D Bacon. *Introduction to dislocations.* Butterworth-Heinemann (2001).

[2] T. Mura. *Micromechanics of Defects in Solids.* Kluwer Academic Publishers Group (Netherlands, 1987).

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

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

[5] M. Beneš. *Mathematical analysis of phase-field equations with numerically efficient coupling terms.* Interfaces and Free Boundaries, 3 (2001), 201–221.

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

[7] 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.

[8] J. Kratochvíl and R. Sedláček. *Statistical foundation of continuum dislocation plasticity.* Physical Review B, 77 (2008), p. 134102.

[9] B. Devincre. *Three dimensional stress field expression for straight dislocation segment.* Solid State Communications, 93 No. 11 (1995), p. 875.

[10] M. Peach and J. S. Koehler. *The forces exerted on dislocations and the stress fields produced by them.* Physical Review (1950).

[11] C. L. Epstein and M. Gage. *The curve shortening flow.* Wave motion: theory, modelling and computation (California, 1986, Berkeley).

[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.

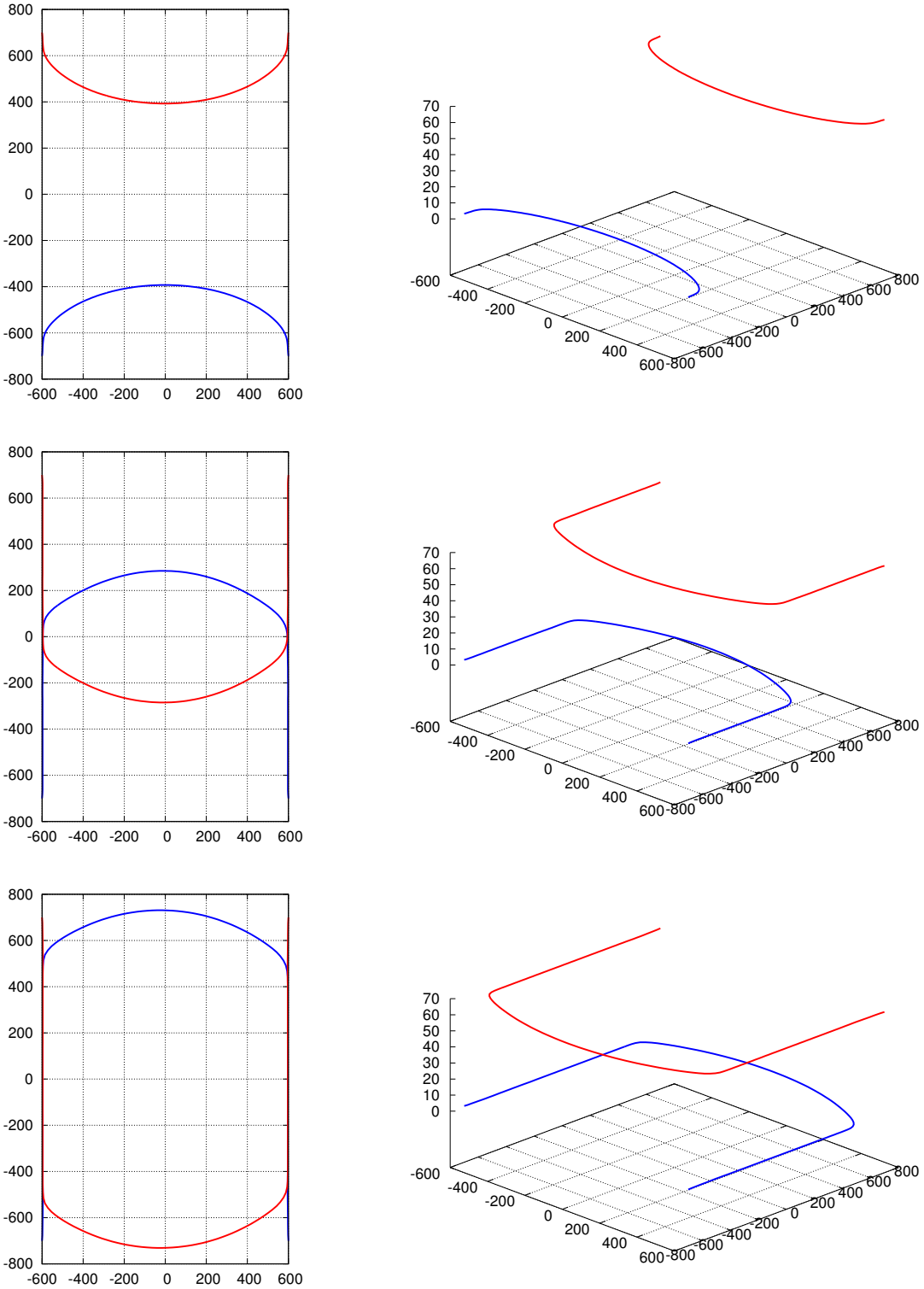


Figure 3: Time evolution of two dislocation curves with the distance $h = 65$ nm. During the passing, the dislocations slightly change their shape. All axes are in nm.

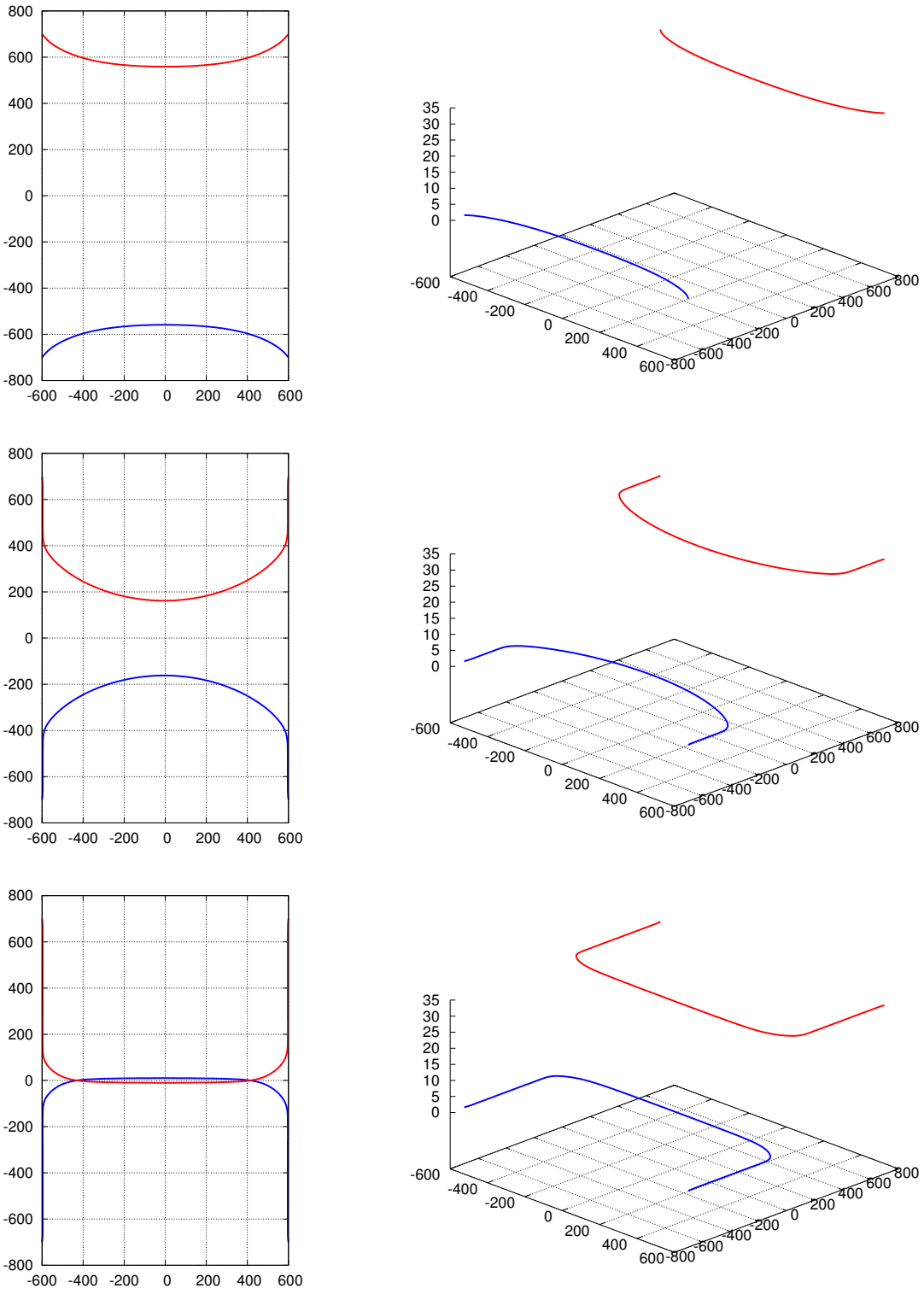


Figure 4: Time evolution of two dislocation curves with the distance $h = 35$ nm. At a certain position, the repulsive force is too high and the dislocations stop moving. All axes are in nm.

Modeling Financial Time Series: Multifracta Cascades and Rényi Entropy

Jan Korbel

1st 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. We show that a number of realistic financial time series can be well mimicked by multiplicative multifractal cascade processes. The key observation is that the multi-scale behavior in financial progressions fits well the multifractal cascade scaling paradigm. Connections with Kolmogorov's idea of multiplicative cascade of eddies in the well developed turbulence are briefly discussed. To put some flesh on a bare bones we compare volatility time series for S & P 500 stock index with a simulated multiplicative multifractal cascade processes. Qualitative agreement is surprisingly good. Salient issues, such as Codimension functions or Multifractal Diffusion analysis and its role in scaling identification are also discussed.
This article has been presented and is part of proceedings of the International symposium on complex systems held in Prague, 10.–13. September, 2013.

Keywords: Multiplicative cascades, Rényi entropy, Multifractal volatility

Abstrakt. Značná část reálných časových řad může být dobře popsána procesy založenými na multifraktálních kaskádách. Klíčové pozorování je, že více-škálové chování při vývoji časových řad se shoduje s koncepcí multifraktálních kaskád. V článku jsou také diskutovány spojitosti s původní Kolmogorovou myšlenkou multifraktálních kaskád jako sobě-podobných turbulenčních vírů. Pro ilustraci tohoto přístupu srovnáme časovou řadu volatility burzovního indexu Standard and Poor's 500 (S&P 500) s časovou řadou, která byla vytvořena jako multifraktální kaskáda. Kvalitativnní shoda těchto dvou řad je velmi dobrá. Další typické problémy, jako např. multifraktální kodimenze nebo metoda MF-DEA pro určení škálovacích exponentů jsou také diskutovány. Tento článek byl prezentován a je obsažen ve sborníku konference International symposium on complex systems held in Prague, 10.–13. September, 2013.

Klíčová slova: Multiplikativní kaskády, Rényiho entropie, Multifraktální volatilita

The Influence of an Interacting Substrate on Turing Instability Conditions*

Karolína Korvasová

1st year of PGS, email: korvasova@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. Diffusion-driven (or Turing) instability of the standard reaction-diffusion system is only achievable under the well-known and rather restrictive conditions on both the diffusion rates and the kinetic parameters. In this study we generalize the standard model by letting the reactants bind to a substrate and investigate the influence of such binding on the Turing parameter space. The idea that binding of the self-activator to a substrate may effectively reduce its diffusion rate and thus destabilize a steady state that would otherwise be stable was formulated in an article by Lengyel and Epstein [4], where the authors reduce the original system of three linear partial differential equations to a two-dimensional reaction-diffusion system under the assumption that the bound state evolves on a fast timescale. We, however, analyse the full system outside this limit. Our results obtained from the full model are in agreement with the results by Lengyel and Epstein [4] in the sense that Turing instability does not require the reactants to diffuse at different rates. We show that, unlike the reduced system, the full system allows relaxing the standard kinetic constraints on Turing instability, particularly two self-activators to generate a pattern.

Keywords: Diffusion-driven instability, interacting substrate

Abstrakt. Difuzí způsobené (neboli Turingovy) nestability standardního reakčně-difuzního systému lze docílit pouze za známých a poměrně restriktivních podmínek jak na difuzní konstanty, tak na parametry chemické kinetiky. V tomto příspěvku uvažujeme o něco obecnější model sestávající ze dvou chemických látek, které difundují a navzájem spolu reagují, přičemž jedna z nich se navíc navazuje na nepohyblivý substrát. Následně studujeme vliv rychlosti navazování na velikost Turingova prostoru. Myšlenku, že navazování na substrát může snížit efektivní rychlost difuze a tím destabilizovat jinak stabilní stacionární stav, zformulovali již Lengyel a Epstein [4] a toto tvrzení demonstrovali na systému dvou reakčně-difuzních rovnic, který získali asymptotickou redukcí původního systému tří rovnic za předpokladu, že kinetika navazování je výrazně rychlejší než ostatní děje. V tomto článku analyzujeme původní (neredukovaný) systém, což nám umožní popsat chování, které pomocí redukovaného systému nelze postihnout. Naše poznatky jsou v souladu s výsledky, které obdrželi Lengyel a Epstein [4] v tom smyslu, že pro vznik Turingovy nestability není potřeba, aby chemické látky difundovaly s rozdílnou rychlostí. Navíc dokážeme, že na rozdíl od redukovaného systému lze v tom neredukovaném rozvolnit i podmínky na kinetické parametry. Jako ilustraci dokládáme příklad systému obsahujícího dvě sebeaktivující látky, který pro vhodnou volbu parametrů vykazuje Turingovu nestabilitu.

*Results contained in this article were presented at the conference BIOMATH 2013 in Sofia, Bulgaria and the article is going to be submitted to The Bulletin of Mathematical Biology. We are grateful for being supported by the grant SGS12/198/OHK4/3T/14.

Klíčová slova: Nestabilita způsobená difuzí, navazování na substrát

Summary

Diffusion-driven instability is an interesting phenomenon that was first formulated by Alan Turing [8] in 1952. The idea that diffusion can destabilize a system and generate a stationary pattern, i.e. a spatially non-homogeneous stationary solution, seemed to be revolutionary and inspiring for many researchers. From the mathematical point of view, a typical setting for such a study consists of a system of two reaction-diffusion equations that describe the time evolution of concentrations of two chemicals that both participate in a chemical reaction[1] and diffuse. In order to distinguish instability caused by diffusion from other types of instabilities, we assume that the trivial steady state of the system of ordinary differential equations

$$\begin{aligned} u_t &= f_u u + f_v v, \\ v_t &= g_u u + g_v v, \end{aligned} \tag{R}$$

that describes the time evolution[2] of concentrations u and v of two chemicals that only react but do not diffuse, is asymptotically stable. Additionally, if the corresponding system of partial differential equations

$$\begin{aligned} u_t &= D_u \Delta_x u + f_u u + f_v v, \\ v_t &= D_v \Delta_x v + g_u u + g_v v, \end{aligned} \tag{RD}$$

that has been derived from (R) by adding diffusion terms, is unstable, the system (RD) is said to exhibit Turing instability or diffusion-driven instability (DDI) [7, 2, 4, 6, 5]. The set of parameter values that permit Turing instability (meaning that for a suitable choice of domain the system exhibits DDI) is often referred to as Turing parameter space.

In this paper we consider a generalization of the system (RD) where we let one of the chemicals bind to a substrate, for example to an extra-cellular matrix. We distinguish two states of the chemical that is allowed to bind: bound and unbound. Let us denote the concentration of the binding chemical in the unbound state by u and its concentration in the bound state by w. The concentration of the second chemical that is not allowed to bind is denoted by v. The corresponding system of reaction-diffusion equations reads

$$\begin{aligned} u_t &= D_u \Delta_x u + (f_u - h_u) u + f_v v - h_w w, \\ v_t &= D_v \Delta_x v + g_u u + g_v v, \\ w_t &= h_u u + h_w w. \end{aligned} \tag{RDB}$$

Note that the third equation in (RDB) governing the time evolution of w does not contain a diffusion term. This is an important fact that enables two chemicals with identical diffusion rates to generate a pattern (see Klika et al. [3] for further reference). On the other hand, Mincheva and Roussel [6] have shown (using a graph-theoretic method)

[1] In this paper we restrict ourselves to linear reaction kinetics that allows us to use simple algebraic tools for stability analysis.

[2] We denote by u_t, resp. v_t, the derivative of u, resp. v, with respect to t and by Δ_x the laplacian with respect to x.

that if all the equations of the system contain a diffusion term, then the diffusion rates must be different in order for a pattern to emerge.

The system (RDB) has already been studied by I. Lengyel and I. R. Epstein [4] who consider an asymptotic reduction of the full model to a system of two reaction-diffusion equations. We provide an additional discussion on consistency of this reduction with the asumption of asymptotic stability of the trivial steady state of the corresponding system (R) without diffusion terms.

Furthermore, we employ methods of linear stability analysis to derive necessary and sufficient conditions for DDI in the full system (RDB). We show that, as opposed to the standard system (RD) [7, 2] and the reduced system considered by Lengyel and Epstein [4], the full system allows DDI even if the parameters f_u and g_v are both positive. This is a significant relaxation of the constraints on chemical kinetics. We also confirm the results by Lengyel and Epstein in the sense that in the full system RDB equal diffusion coefficients do not preclude DDI. Moreover, if DDI occurs for a particular choice of parameters with $D_u = D_v$, it automatically occurs for the same kinetic parameters and diffusion of any magnitude, as long as the diffusion constants are identical. We remark that identical diffusion constants are in contradition with DDI in the standard system (RD), see [7, 2] for further details.

To illustrate the results and to show that no fine parameter-tuning is needed in order to find an example of a system that exhibits Turing instability and violates the conditions for DDI in standard reaction-diffusion equations without binding [7, 2], we perform a simple sensitivity analysis of a concrete system with $D_u = D_v$, $f_u > 0$ and $g_v > 0$. We also plot a few slices of the Turing parameter space that were obtained numerically.

To summarize, we have shown that binding of the reacting chemicals to a non-diffusing substrate can significantly relax the constraints that DDI imposes on the model parameters. In particular, a system with binding allows two chemicals that diffuse at the same rates as well as two self-activators to generate a pattern due to diffusion.

References

[1] C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers*. Springer-Verlag, New York, (1999).

[2] L. Edelstein-Keshet. *Mathematical Models in Biology*, volume 46 of *Classics In Applied Mathematics*. SIAM, (2005).

[3] V. Klika, R. E. Baker, D. Headon, and E. A. Gaffney. *The influence of receptor-mediated interactions on reaction-diffusion mechanisms of cellular self-organisation*. Bull Math Biol **74** (2012), 935–957.

[4] I. Lengyel and I. R. Epstein. A chemical approach to designing turing patterns in reaction-diffusion systems. volume 89, 3977–3979, (1992).

[5] H. Meinhardt. *Models of Biological Pattern Formation*. Academic Press, (1982).

[6] M. Mincheva and M. R. Roussel. *A graph-theoretic method for detecting potential turing bifurcations*. J Chem Phys **125** (2006).

[7] J. D. Murray. *Mathematical Biology.* Springer-Verlag, Berlin Heidelberg, (2002).

[8] A. M. Turing. *The chemical basis of morphogenesis.* Philosophical Transactions of the Royal Society of London **237** (1952), 37–72.

Segmentation of MRI Data by Means of Nonlinear Diffusion*

Radek Máca

4th year of PGS, email: radek.maca@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. The article focuses on the application of a segmentation algorithm based on the numerical solution of the Allen-Cahn non-linear diffusion partial differential equation. This equation is related to the motion of curves by mean curvature. It exhibits several suitable mathematical properties including stable solution profile. This allows the user to follow accurately the position of the segmentation curve by bringing it quickly to the vicinity of the segmented object and by approaching the details of the segmentation curve. The purpose of the article is to indicate how the algorithm parameters are set up and to show how the algorithm behaves when applied to the particular class of medical data. We describe in detail the algorithm parameters influencing the segmentation procedure, namely the force term allowing the segmentation curve to quickly move towards the segmented object, choice of the gradient control and the stopping criterion. The algorithm itself is easy to implement and its parallelization is possible. The left ventricle volume estimated by the segmentation of scanned slices is evaluated through the cardiac cycle. Consequently, the ejection fraction which serves as a medical information is evaluated. This approach allows the user to process cardiac cine MR images in an automated way and represents, therefore, an alternative to other commonly used methods. Based on the physical and mathematical background, the presented algorithm exhibits the stable behavior in the segmentation of MRI test data, it is computationally efficient and allows the user to perform various implementation improvements.

This article has been published in Kybernetika ([1]).

Keywords: cardiac MRI, co-volume method, image segmentation, level set method, PDE

Abstrakt. Tento článek se zabývá aplikací segmentačního algoritmu založeném na numerickém řešení Allenovy-Cahnovy parciální diferenciální rovnice. Pomocí této rovnice lze popsat pohyb křivek, který je závislý na jejich křivosti. Tato vlastnost dovoluje pohybovat segmentační křivkou tak, že popíše segmentovaný objekt. Hlavním obsahem této práce je popis výpočetních parametrů, jejich nastavení a vlastnosti algoritmu aplikovaného na segmentaci medicínských dat. Podrobně jsou popsány parametry ovlivňující průběh segmentace. Použitý algoritmus je aplikován na segmentaci levé srdeční komory ze série snímků obsahující celý srdeční cyklus. Díky tomu lze vyčíslit tzv. ejekční frakci. Tento přístup uživateli umožňuje zpracovat snímky z magnetické rezonance automaticky a může sloužit jako alternativa ke stávajícím segmentačním algoritmům.

Tento článek byl publikován v časopise Kybernetika ([1]).

*This work has been supported by the grant No. SGS11/161/OHK4/3T/14.

Klíčová slova: MRI srdce, metoda duálních objemů, segmentace obrazu, vrstevnicová metoda, PDR

References

[1] R. Chabiniok, R. Máca, M. Beneš and J. Tintěra. *Segmentation of MRI data by means of nonlinear diffusion.* In 'Kybernetika' volume 49, no. 2, (2013), 301–318

Distributed Data Processing in High-Energy Physics

Dzmitry Makatun

2nd year of PGS, email: makatun@rcf.rhic.bnl.gov
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Michal Šumbera, Nuclear Physics Institute, AS CR
Jérôme Lauret, STAR, Brookhaven National Laboratory, USA

Abstract. Processing data in distributed environment has found its application in many fields of science (Nuclear and Particle Physics (NPP), astronomy, biology to name only those). Efficiently transferring data between sites is an essential part of such processing. The implementation of caching strategies in data transfer software and tools, such as the Reasoner for Intelligent File Transfer (RIFT) being developed in the STAR collaboration, can significantly decrease network load and waiting time by reusing the knowledge of data provenance as well as data placed in transfer cache to further expand on the availability of sources for files and data-sets. Though, a great variety of caching algorithms is known, a study is needed to evaluate which one can deliver the best performance in data access considering the realistic demand patterns.

Records of access to the complete data-sets of NPP experiments were analyzed and used as input for computer simulations. Series of simulations were done in order to estimate the possible cache hits and cache hits per byte for known caching algorithms. The simulations were done for cache of different sizes within interval 0.001 - 90% of complete data-set and low-watermark within 0-90%. Records of data access were taken from several experiments and within different time intervals in order to validate the results. In this paper, we will discuss the different data caching strategies from canonical algorithms to hybrid cache strategies, present the results of our simulations for the diverse algorithms, debate and identify the choice for the best algorithm in the context of Physics Data analysis in NPP. While the results of those studies have been implemented in RIFT, they can also be used when setting up cache in any other computational work-flow (Cloud processing for example) or managing data storages with partial replicas of the entire data-set.

Keywords: data transfer, cache, optimization, algorithm

Abstrakt. Zpracování dat v distribuovaném prostředí nachází své uplatnění v mnoha oblastech vědy (např. v jaderné a částicové fyzice (NPP), astronomii, biologii). Efektivní přenos dat mezi sítěmi je nedílnou součástí takového zpracování. Implementace strategií kešování v softwaru pro přenos dat a v nástrojích, jako je např. Reasoner for Intelligent File Transfer (RIFT), který byl vyvinut v rámci experimentu STAR, může výrazně snížit zatížení sítě a čekací doby využitím znalostí o původu dat, stejně jako data v přenosové mezipaměti, k dalšímu rozšíření dostupnosti zdrojů souborů a dat. Přestože je známo velké množství různých kešovacích algoritmů, je nutné prozkoumat a vyhodnotit, který z nich může podávat nejlepší výkon v přístupu k datům při zvážení realistických modelů požadavků. Záznamy o přístup do kompletních datových sad experimentů v NPP byly analyzovány a použity jako vstup pro počítačové simulace. Řady simulací

byly provedeny za účelem odhadu možných zásahů keše a zásahů keše na bajt pro známé algoritmy. Simulace byly provedeny pro cache různých velikostí v intervalu 0,001-90 V tomto článku budeme diskutovat různé strategie kešování dat, od kanonických algoritmů po hybridní kešovací strategie, budeme prezentovat výsledky našich simulací pro různé algoritmy, rozebereme a určíme výběr nejlepšího algoritmu v souvislosti s fyzikální analýzou dat v NPP. Výsledky těchto studií byly začleněny do RIFT, mohou však být použity také pro nastavení cache v jakémkoli jiném výpočetní prostředí (např. zpracování v cloudu) nebo řízení datových úložišť s částečnými replikami celé sady dat.

Klíčová slova: přenos dat, mezipaměť, plánování, algoritmus

1 Introduction

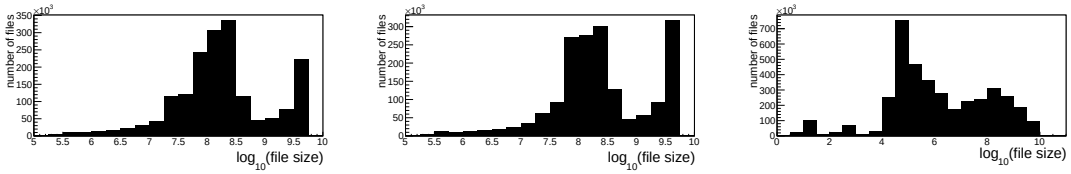
Efficient usage of available cache space is important for transferring and accessing data in computational grids. Though, a great variety of caching algorithms is known, a study is needed to evaluate which one can deliver the best performance in data access considering the realistic demand patterns.
Cache cleaning algorithms can be applied to keep in the cache of data-transfer tools files that may be reused. The size of those cache is small (several percent of the entire dataset) and the clean up has to take place regularly to make space for further transfers. Another task can, for example, be to delete a part of local data replica if no longer in use or requested. The problem posed by cache cleanup is to select and delete files which are the least likely to be used again. An investigation to find the most appropriate algorithm is required.

In this study, all the caching algorithm were implemented following the concept known as "water-marking". Water-marking is an approach where thresholds are set for the cache cleanup starts and stops. It considers the current disk space occupied by the data in cache and the high-mark and the low-mark for cache size are externally set up and can be adjusted as needed. When the used cache size exceeds the high-mark, the cache clean-up starts, and files are deleted until the used cache size gets below the low-mark. The time interval between clean-ups depends on combination of high/low marks, cache size and data-flow.Therefore with watermarking concept more computational demanding algorithms can be implemented as the cleanup may be independent of the data transfers.

2 Access patterns

Several data access patterns were extracted from log files of data management systems at sites of HEP/NPP experiments in order to simulate caching. Three different access patterns were used as input for our simulations:

STAR1: the pattern was extracted from Xrootd [8] logs taken from the STAR experiment's Tier-0 site of RHIC Computing Facility at Brookhaven National Laboratory (RCF@BNL), it consist of records made during a 3 months period (June-August 2012) of all datasets available in STAR.



a b c

Figure 1: Distribution of files by size for three datasetst: a - STAR1, b - STAR2, c - GOLIAS.

STAR2: the pattern was extracted from Xrootd [8] logs of Tier-0 site of STAR experiment (RCF@BNL), it consist of records made during a 7 months period (August 2012 - February 2013) under similar conditions.

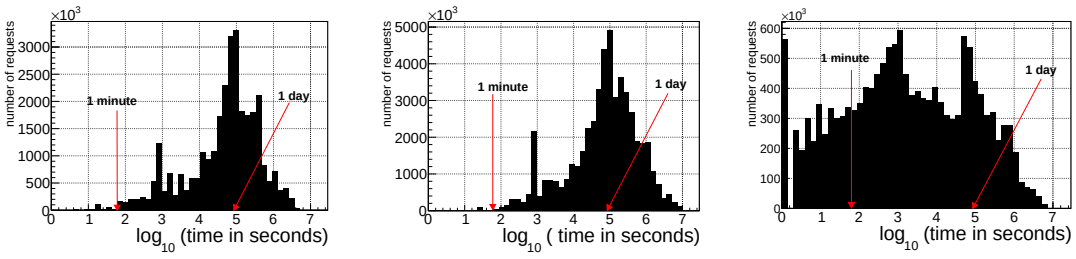
GOLIAS farm is a part of regional computing center for particle physics at the Institute of Physics (FZU) in Prague, and is part of a Tier-2 site for the CERN/ATLAS experiment. The facility also performs data processing for another experiment - AUGER, which makes less than 1% of the total requests. The pattern was extracted from DPM [9] logs for a 3 months period (November 2012 - February 2013).

The usage of access patterns corresponding to different time periods and experiments helps to verify the results of our simulations. As input of our simulations, we tried to focus on a few characteristic access patter. The key parameters we came up with for the three access patterns are given in Table 1. Both STAR access patterns have similar parameters. It is noteworthy to mention that the first one was taken right before the Quark Matter 2012 conference and the second one, right after. This is important as the access to data is intensified before a conference and without verification, it would be doubtful if our findings would be stable across time. The number of files requested only once during the period, is less than 10% in both cases.

The FZU/GOLIAS access pattern is taken from another experiment with different data-storage structure, DPM is used here within a Tier-2 data access context (user analysis). This access pattern is much less uniform and differs from the other two: the size of files is not explicitly limited and can reach 18 GB, the number of requests for a file varies from 1 up to 94260, with an average 5. 44% of files were requested only once.

When analyzing an access pattern one can subtract a set of unique filenames. It is a set of all files requested at least ones during the period of consideration. The following histograms at the Figure 1 represents the distribution of those unique files by size for each data-set. Here one can see that filesize distribution at GOLIAS is more dispersed than in STAR. Also, as it can be observed at the histograms, at STAR maximal filesize is limited to 5.3 GB (the files of larger size are splitted into several files). This fact explains the second peak at the histograms for STAR1 and STAR2 datasets. At GOLIAS there is no limitation for filesize, to peaks at the histogram can be explained with the presence of files with different types of data.

The timing characteristics of an access pattern can be pictured as a distribution of a time interval between two consequent requests for the same file. This histograms are given at Figure 2. In both STAR access patterns the distribution is close to log-normal with the peak time interval corresponds to 24 hours. This can be explained by the users behavior,



a b c

Figure 2: Distribution of time intervals between sequential requests for the same file. This distribution helps to understand the timing characteristics of access. Three different access patterns are presented: a - STAR1, b - STAR2, c - GOLIAS.

Table 1: Summary of three user access patterns used as input for simulations. The selection of two time sequence in STAR and one from a different experiment aims at verifying stability of our result and findings.

		STAR1	STAR2	GOLIAS
Time period	months	3	7	3
Number of requests	$\times 10^6$	33	52	21
Data transferred	PB	50	80	10
Maximal number of requests for one file	–	192	203	94260
Average number of requests per file	–	19	15	5
Number of unique files	$\times 10^6$	1.8	1.7	3.8
Total size of dataset	PB	1.45	2	1
Maximal file size	GB	5.3	5.3	18
Average file size	GB	0.8	1	0.3

one can imagine a situation when a scientist checks a result of computational job in the morning, edits the code and then resubmits the analysis on the same dataset, and the new output will be available only next working day. The GOLIAS access pattern is less regular. This can be explained with the large amount of jobs submitted automatically with different periods, and probably smaller average time of job running.

3 Cache simulation

Selection of cache policy depends on the user access pattern and the disk space available. The efficiency of caching can be estimated by two quantities, the cache hits $H(1)$ and cache hits per megabyte of data H_d(cache data hits) (2):

$$H = \frac{N_{cache}}{N_{req} - N_{set}} \tag{1}$$

$$H_d = \frac{S_{cache}}{S_{req} - S_{set}} \tag{2}$$

where N_{req} is the total number of requests, S_{req} -the total amount of transferred data in bytes, N_{set} -the number of unique filenames, S_{set} - the size of storage in bytes, N_{cache} - the number of files transferred from cache, S_{cache} - the amount of data transferred from cache in bytes.

By maximizing the cache hits H one reduces the number of files transferred from external sources and thus reduces the overall make-span due to transfer startup overhead for each file. By maximizing the cache data hits H_d one reduces the network load, since more data is reused from the local cache.

If the access pattern is completely random, the expected cache hit and cache data hits would be equal to *cache size/storage size*, so it can be useful to compare the actual cache performance to this estimation.

Altogether 27 different caching algorithms were simulated. But the majority of studied algorithms did not bring any improvements over the simplest one (FIFO). Only the algorithms that appeared to be the most efficient are discussed in this paper:

- **First-In-First-Out (FIFO)**: evicts files in the same order they entered the cache. Performance of this trivial algorithm provide a good comparison benchmark against more sophisticated ones which can demand significant computational resources.

◯ **Least-Recently-Used (LRU)**: evicts the set of files which were not used for the longest period of time.

● **Least-Frequently-Used (LFU)**: evicts the set of files which were requested less times since they entered the cache.

★ **Most Size (MS)**: evicts the set of files which have the largest size.

\+ **Adaptive Replacement Cache (ARC)**[5]: splits cached files into two lists: L1 - files with *access count* = 1, and L2 - files with *access count* > 1. LRU is then applied to both list, the self adjustable parameter $p = cache\ hits\ in\ L1/cache\ hits\ in\ L2$ defines the number of cached files in each list. The general idea is to invest more into the list which delivers more hits.

∗ **Least Value based on Caching Time (LVCT)**[4]: Deletes files with the smallest value of the Utility Function:

$$UtilityFunction = \frac{1}{CachingTime \times FileSize} \tag{3}$$

where **Caching Time** of a file F is total size of all files accessed after the last request for the file F.

▽ **Improved-Least Value based on Caching Time (ILVCT)**[3]: Deletes files with the smallest value of the Utility Function:

$$UtilityFunction = \frac{1}{NumberOfAccessedFiles \times CachingTime \times FileSize} \tag{4}$$

where **Caching Time** is the same as for LVCT and **Number Of Accessed Files** is a number of files requested after the last request for selected file.

4 Results

Three series of simulations with three access patterns were performed for each algorithm (90 simulations in total for each algorithm):

- 10 simulations with cache size 1-90 % of storage with fixed low-mark 75% and high-mark 95%. Those simulations aim to understand what would happen if we have large storage cache. Those cases are aligned with a DPM and Xrootd use where most (if not all) the dataset(s) are in the system.
- 10 simulations with cache size 1.2 - 0.0025% of storage with fixed low-mark 75% and high-mark 85%. We used those simulations to understand the behavior of cache cleanup if the cache size is by several orders less than the dataset size. This is in fact a most common case for transfer cache on data transfer nodes.
- 10 simulations with fixed cache size 10% of storage,fixed high-mark 95% and variable low mark within 0-90%. We performed those simulations to better understand the effect of data retention on cache (delete the least in hope of re-use).

In order to compare one algorithm against another an average improvement can be calculated in a following way:

$$Average\ improvement = \frac{\sum_{i=1}^{n} \frac{value2_i - value1_i}{value1_i}}{n}, \tag{5}$$

where n is the total amount of simulations with equal parameters for both algorithms, i is the number of the simulation,$value1$ - cache hits or cache data hits of a reference algorithm (FIFO), $value2$ - cache hits or cache data hits of a compared algorithm.

Table 2 contains the results of comparison of all the algorithms represented in this paper against FIFO. Results of simulation series 1 and 2 were used to calculate the average improvement (60 values for each algorithm). According our results, the LFU algorithm does not bring any improvement over FIFO due to its well known flaw - it accumulates files which were popular for a short period of time, and those files prevent newer ones from staying in cache. The ARC algorithm was developed as an improvement to LRU, and not surprisingly, it outperforms LRU by $\sim$5% in cache hits and $\sim$7% in cache data hits. Therefore, LFU and LRU algorithms could be excluded from the further analysis in our case studies.

The graphical detailed results of simulations for all 3 series are given at Figures 3-5. The performance of FIFO and 3 algorithms appeared to be the most efficient (MS, ACR and LVCT) is presented at the plots.

Difference between Tier-2 and Tier-0 access patterns leads to distinct cache performance. Only the data dedicated for the ongoing analysis is placed at the Tier-2 site, while at the Tier-0 site all the experimental data is stored. As a result – the access pattern at the Tier-2 site has stronger access locality. STAR1 and STAR2 access patterns correspond to Tier-0 site and GOLIAS to a Tier-2 site. Thus, any particular algorithm at the plots delivers higher cache hits and cache data hits for GOLIAS access pattern than for STAR1 and STAR2.

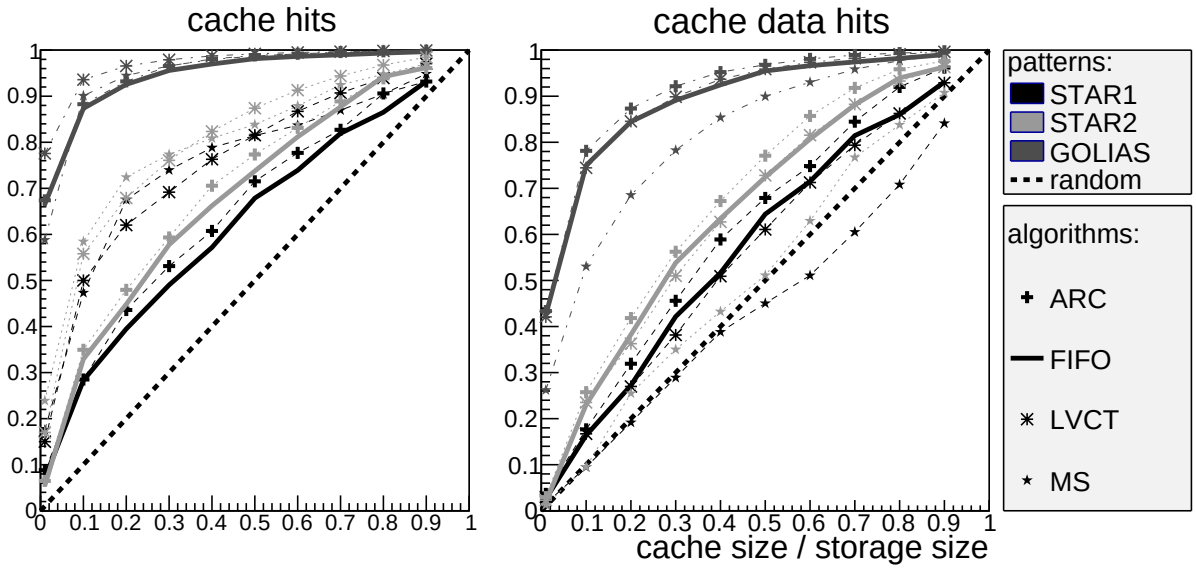


Figure 3: Results of simulation. Performance of algorithms for cache of larger size can be compared. For all of the simulations on this plot the following parameters were fixed: low mark = 0.75, high mark = 0.95

The behavior of algorithms is similar within each dataset that is, their respective performance ordering is the same. This observation implies that if one of the algorithm appears to be most efficient for one of the datasets it is also the most efficient for the other datasets. This statement is also true for the rest of simulated algorithms not present on our figure. Though the communities represented by the STAR and GOLIAS access patterns are somewhat similar, this result is slightly surprising as our case studies represent two time sequence within the same usage and totally uncorrelated experiments. It would be interesting to study those algorithms in a different experimental context (outside the HEP/NPP communities) but this study is outside the scope of our paper.

The MS algorithm has shown outstanding cache hits, but the lowest cache data hits. At the same time the LVCT has cache hits comparable to the MS while cache data hits are 2% improved over the FIFO. This algorithm could be an optimal when the cache hits is the target optimization parameter. The ARC algorithm has shown the highest cache data hits for studied access patterns.

The dependence of algorithms performance on low mark is presented at Figure 5. With higher low mark the number of clean-ups increases and that is why the difference between algorithms becomes more notable. Performance of efficient algorithms (FIFO, LRU, ARC and LVCT) increases steadily with the low mark. One should be careful when setting up a cache low mark at a particular site, since a higher low mark can increase cache performance significantly, but at the same time it can result in running cache clean-ups too often, consuming significant computational resources (and potentially increasing the chance to interfere with data transfers hence, degrading transfer performances if delete/writes/read overlap).

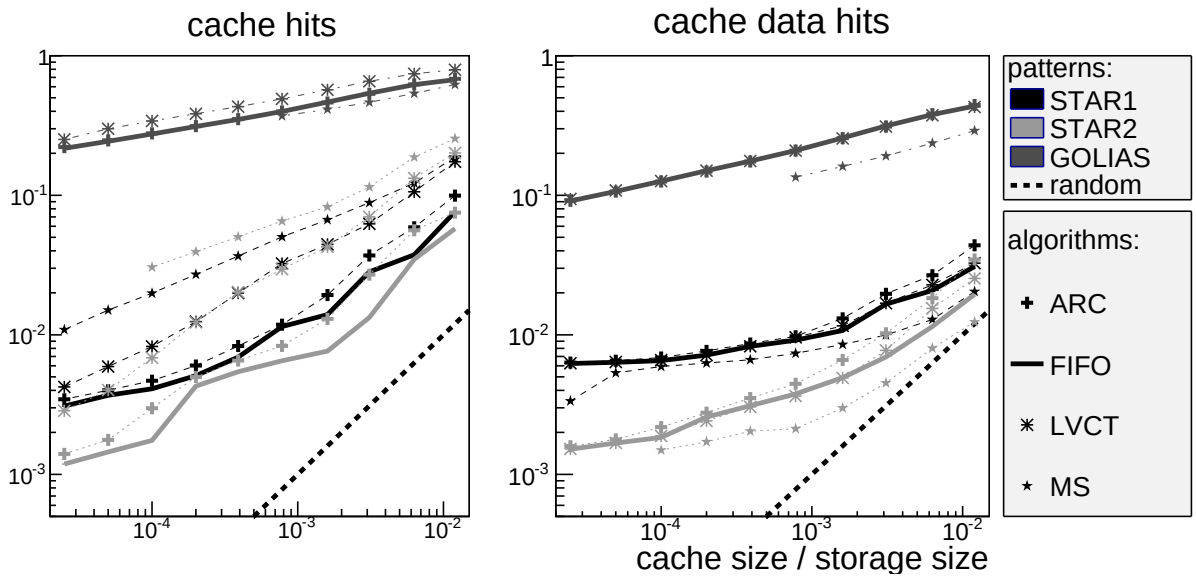


Figure 4: Results of simulation. Performance of algorithms for cache of smaller size can be compared. For all of the simulations on this plot the following parameters were fixed: low mark = 0.75, high mark = 0.85

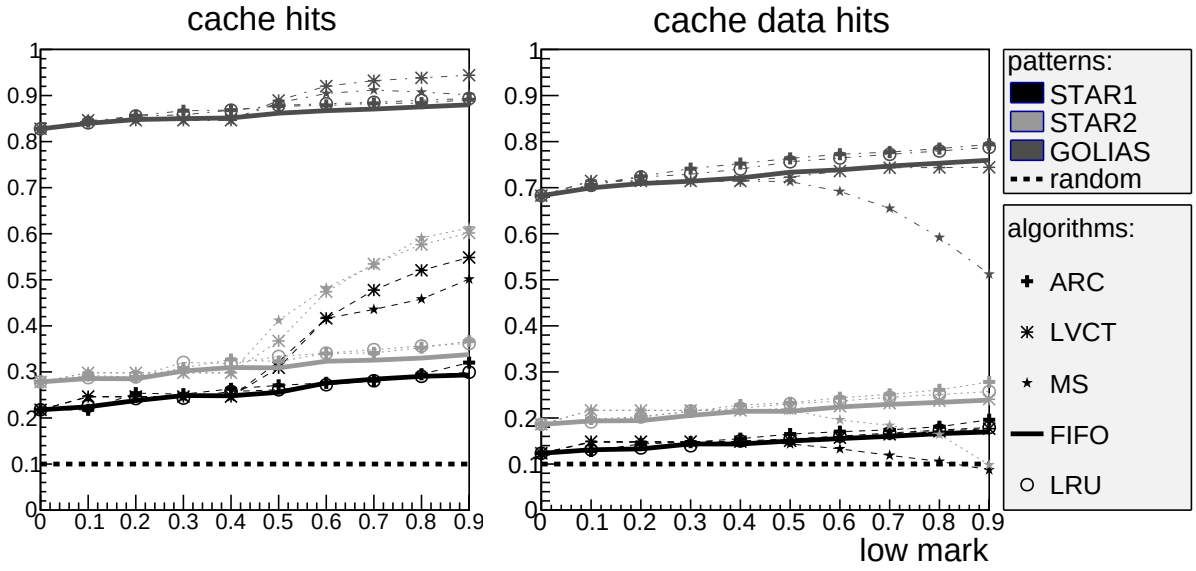


Figure 5: Results of simulation: dependence of cache performance on low mark. For all of the simulations on this plot the following parameters were fixed: cache size / storage size = 0.1, high mark = 0.95

Table 2: Average improvement of algorithms over FIFO.

Algorithm	cache hits	cache data hits
MS	**116** %	-20 %
LRU	8 %	5 %
LFU	-27 %	-18 %
ARC	13 %	**11** %
LVCT	**86** %	2 %
ILVCT	28 %	2 %

5 Conclusion

Performance of cache algorithms implemented with watermarking concept was simulated for a wide range of cache sizes and low marks. Three access patterns from Tier-0 and Tier-2 sites of 2 different experiments were used as input for simulations. Regardless of the cache size, Tier-level and specificity of experiment the LVCT and ARC appeared to be the most efficient caching algorithms for the communities we investigated. While we found the result surprising at first, we attribute this result to an access pattern which is intrinsically similar in nature. An extension of this work could be the investigation of this result in a different experiment context which is a work beyond our initial goal. LVCT and ARC could certainly be safely implemented in tools such as RIFT.

- If the goal is to minimize makespan due to a transfer startup overhead the LVCT algorithm should be selected.
- If the goal is to minimize the network load the ARC algorithm is an option.

Acknowledgments

References

References

[1] Makatun D, Lauret J and Sumbera M 2012 Distributed Data processing in high-energy physics *Proc. of PhD students workshop at FNSPN CVUT* (Prague) ISBN 978-80-01-05138-2

[2] Zerola M, Lauret J, Bartak R and Sumbera M 2012 One click dataset transfer: toward efficient coupling of distributed storage resources and CPUs *J. Phys.: Conf. Ser.* **368**

[3] Achara J P, Rathore A, Gupta V K and Kashyap A 2010 An improvement in LVCT cache replacement policy for data grid *Proc. of the 13th Int. Workshop on Advanced Computing and Analysis Techniques in Physics Research* (Jaipur) p 44

[4] Song Jiang and Xiaodong Zhang 2003 Efficient Distributed Disk Caching in Data Grid Management *Proc. of the IEEE Int. Conf. on Cluster Computing (CLUSTER'03)* (Hong Kong) pp 446-51

[5] Megiddo, Nimrod and Modha D S 2004 Outperforming LRU with an adaptive replacement cache algorithm *Computer* **37** 58-65

[6] Fast Data Transfer *Project web-site: http://monalisa.cern.ch/FDT/*

[7] High Performance Storage System *Project web-site: http://www.hpss-collaboration.org/*

[8] Xrootd *Project web-site: http://xrootd.slac.stanford.edu/*

[9] Disk Pool Manager *Project web-site:* https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm

Condensation in the Zero-range Processes

Michael Matějů

3rd year of PGS, email: matejmi8@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Lucie Fajfrová, Institute of Information Theory and Automation, AS CR

Abstract. This paper summarizes known facts about zero-range processes focused on condensation phenomenon. A brief summary on Markov semi-groups is given to develop satisfactory tools for building a Markov process on an infinite state space. Due to the known Markov generator, we are able to develop an appropriate Markov process and to find out its stationary measures. Defining a condensation phenomenon in probability sense we are able to describe phase transitions. Finally, we will give some insight into the dynamics of condensation.

Keywords: Condensation; zero range processes; interacting particle system

Abstrakt. Tento příspěvek shrnuje základní poznatky jevu kondenzace v zero-range procesech. Teorie Markovských semi-grup umožňuje vytvořit potřebné nástroje pro konstrukci Markovského procesu na nekonečném stavovém prostoru a určení stacionární míry. Definováním jevu kondenzace v pravděpodobnostním smyslu, jsme schopni popsat fázové přechody odpovídajícího Markovského procesu a podat krátký přehled dynamiky kondenzace.

Klíčová slova: Kondenzace; zero range proces; interacting particle system

1 Introduction

We are interested in the conservative interacting particle systems which are Markov processes with continuous time and discreet state space. The goal of the research is to study traffic phenomena, and in this article, we focus on known facts about a condensation phenomenon of the zero range process.

We will denote a state space as $\mathbf{S} = E^{\Lambda}$, where $\Lambda \subseteq \mathbb{Z}^d$, and E will be a local state space. For all $x \in \Lambda$, $\eta(x)$ will denote the number of particles occupying x, thus $\eta(x) \in E$ will be local state of the system on a position $x \in \Lambda$. Whole configuration of the system is denoted by $\boldsymbol{\eta} = (\eta(x))_{x\in\Lambda}$. In order to simply but correctly outline the process construction considered in section 2, we firstly consider the local state space to be of finite size, i.e. $E = \{0, 1, ..., \eta_{max}\}$, where $\eta_{max} \in \mathbb{N}$. And we assume product topology on $\mathbf{S}$, thus $(\mathbf{S}, \sigma(\mathbf{S}))$ is a compact measurable space with Borel σ-algebra.

2 Interacting particle systems

2.1 Construction of the process and dynamics

We will use a canonical construction of the process. Population Ω is the set of all right-continuous functions with left limits $\boldsymbol{\eta}(\cdot)$ on $[0,\infty)$ with values in $\mathbf{S}$, so called trajectories. A random variable is projective mapping

$$Y : \Omega \times \mathbb{R}_+ \mapsto \mathbf{S}; \quad Y(\boldsymbol{\eta}(\cdot), t) = \boldsymbol{\eta}(t)\,,$$

σ-algebra is a canonical one

$$\mathcal{F} = \bigotimes_{t\in[0,\infty)} \mathcal{B}(Y_t) = \sigma(Y_t, t \geq 0)\,,$$

and we define canonical filtration $(\mathcal{F}_t)_{t\in[0,\infty)}$ on $(\Omega, \mathcal{F})$, which is the system of σ-algebras $\{\mathcal{F}_t,\, t \geq 0\}$ such that $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ for all $0 \leq s \leq t$ and for which $Y(\boldsymbol{\eta}(\cdot), s)$ is $\mathcal{F}_s$-measurable.

To have defined Markov process on the state space $\mathbf{S}$, we need either the system of probability measures $\mathbb{P}_{\boldsymbol{\eta}}$ or transition probability $P_t(\boldsymbol{\eta}, A)$[1], where $\boldsymbol{\eta} \in \mathbf{S}$ and $A \in \sigma(\mathbf{S})$. Using the transition probabilities or the family of probability measures respectivly, we can define the remaining object by the following formula

$$P_t(\boldsymbol{\eta}, A) = \mathbb{P}_{\boldsymbol{\eta}}\{Y_t \in A\}\,, \tag{1}$$

where $Y_t = Y(\boldsymbol{\eta}(\cdot), t) = \boldsymbol{\eta}(t)$. It means, that the process Y which started in state $\boldsymbol{\eta}$ will be in time t in the set A.

Based on known probability measures $\mathbb{P}_{\boldsymbol{\eta}}$, we define Markov process[2] as a quadruple $(\Omega, (\mathcal{F}_t), Y, \mathbb{P}_{\boldsymbol{\eta}})$. Then the transition probabilities $\{P_t, t \geq 0\}$ defined by (1) form (Markov)[3] semi-group of one parameter bounded linear operators on $(\mathcal{C}_b(\mathbf{S};\mathbb{R}), ||\cdot||)$, where $(\mathcal{C}_b(\mathbf{S};\mathbb{R}), ||\cdot||)$ is Banach space of bounded continuous real functions on $\mathbf{S}$ with a supreme norm.

An infinitesimal operator of a semi-group P_t is defined by the formula:

$$\mathcal{L}f(\boldsymbol{\eta}) = \lim_{t\searrow 0+} \frac{\int_{\mathbf{S}} P_t(\boldsymbol{\eta}, \mathrm{d}\boldsymbol{\xi}) f(\boldsymbol{\xi}) - f(\boldsymbol{\eta})}{t} = \lim_{t\searrow 0+} \frac{P_t f(\boldsymbol{\eta}) - f(\boldsymbol{\eta})}{t}\,, \quad f \in D_{\mathcal{L}}\,, \eta \in \mathbf{S}\,.$$

Domain $D_{\mathcal{L}}$ of this operator are functions for which the limit exits, and $\mathcal{L} : D_{\mathcal{L}} \mapsto \mathcal{C}_b(\mathbf{S};\mathbb{R})$. For every $f \in D_{\mathcal{L}}$, function $P_t f$ is differentiable, hence we have got the *evolution equation*

$$\frac{\mathrm{d}(P_t f)}{\mathrm{d}t} = \mathcal{L}(P_t f) = P_t(\mathcal{L} f)\,, P_0 f = f\,, \tag{2}$$

for which function $P_t f$ is unique solution[4]. It is convenient to look on P_t as a "time-evolution" operator of observable f, which could be for example the number of particles occupying position x: $f(\boldsymbol{\eta}) = \eta(x)$.

[1] We assume homogeneous processes, and expect a homogeneous transition probability.

[2] Such quadruple must fulfil some assumptions imposed on its objects. Proper definition of the Markov process can be found in every monograph concerning Markov processes. For example see [3], chp.3.1.

[3] "Markov" designates that the semi-group is conservative, normal, positive contraction semi-group.

[4] See chp. 2.2 in [3] for more information concerning infinitesimal generator and uniqueness theorems.

We consider system without explosions, which means, that in any arbitrary bounded time-interval only finite number of jumps of particles could happen. If the length of the time-interval decreases to 0 we assume that only one jump happens. This one jump will be considered as an "infinitesimal transition".

Let us denote by $\mathcal{P}(\mathbf{S})$ the space, which is defined as the totality of finite, countable additive functions $\pi(\cdot)$ on $\sigma(\mathbf{S})$; especially it consists of all probability measures. Spaces $(\mathcal{C}_b(\mathbf{S};\mathbb{R}), ||\cdot||)$ and $\mathcal{P}(\mathbf{S})$ are naturally connected by the scalar product[5]

$$(f,\pi) = \int_{\mathbf{S}} f(\boldsymbol{\xi})\pi(\mathrm{d}\boldsymbol{\xi})\,.$$

Thus, it could be shown, that the space $(\mathcal{C}_b(\mathbf{S};\mathbb{R}), ||\cdot||)$ is regarded as a subspace of $\mathcal{P}^{\#}$, i.e. a dual-space to $\mathcal{P}$.

If we know the initial distribution $\pi \in \mathcal{P}(\mathbf{S})$ of the system, the time dependent distribution of the process denoted by $\pi P_t \in \mathcal{P}(\mathbf{S})$ is uniquely defined through the formula

$$\int_{\mathbf{S}} P_t f \mathrm{d}\pi = \int_{\mathbf{S}} f \mathrm{d}(\pi P_t)\,, \quad f \in \mathcal{C}_b(\mathbf{S};\mathbb{R})\,.$$

We are interested in the long-term behavior of the transition probability[6]. It requires solving equation (2) to find out behavior of the observable f in time. However, that problem can be transformed to the dual-space of measures, thus we are looking for the long-term behavior of the distribution of the particles on the lattice; i.e. a stationary distribution, such that

$$\pi P_t = \pi\,, \quad \forall t \geq 0\,.$$

It is worth noting, that by integrating equation (2) through $\mathbf{S}$ with respect to some stationary measure $\pi \in \mathcal{P}(\mathbf{S})$, the left-hand side of the equation vanishes and we obtain a new equation to be solved in the form

$$\int_{\mathbf{S}} \mathcal{L} f \mathrm{d}\pi = 0\,, \quad \forall f \in \mathcal{C}_0(\mathbf{S};\mathbb{R})\,, \tag{3}$$

where $\mathcal{C}_0(\mathbf{S};\mathbb{R}) \subset \mathcal{C}(\mathbf{S};\mathbb{R})$ denotes the set of cylindric functions.

2.2 Dynamics of particles

The infinitesimal operator (called Markov generator) for an interacting particle system on a lattice with the additional assumption that particles do not appear nor disappear is given by

$$\mathcal{L}f(\boldsymbol{\eta}) = \sum_{x,y\in\Lambda} c(x,y,\boldsymbol{\eta})[f(\boldsymbol{\eta}^{x\to y}) - f(\boldsymbol{\eta})]\,, \quad f \in \mathcal{C}_0(\mathbf{S};\mathbb{R}), \boldsymbol{\eta} \in \mathbf{S}\,,$$

where $c(x,y,\boldsymbol{\eta})$ denotes the infinitesimal transition rate with the meaning of the transition of one particle from a position $x \in \Lambda$ to a position $y \in \Lambda$ with the current configuration

[5] More precise is to say, that they are connected by duality pairing.

[6] Which describes the long-term behavior of the system

$\boldsymbol{\eta}$. The $c(x, y, \boldsymbol{\eta})$ is assumed[7] to be non-negative, continuous in $\boldsymbol{\eta}$, $c(\cdot, \cdot, \boldsymbol{\eta})$ is assumed to be irreducible[8], and the total rate of particles jumping to a position y is uniformly bounded; i.e. $\sup_{y\in\Lambda} \sum_{x\in\Lambda} \sup_{\boldsymbol{\eta}\in\mathbf{S}} c(x, y, \boldsymbol{\eta}) < \infty$. If Λ is a finite lattice, the infinitesimal transition rates can be expressed as

$$P_t(\boldsymbol{\eta}, \boldsymbol{\eta}^{x\to y}) = \mathbb{P}_{\boldsymbol{\eta}}(Y_t = \boldsymbol{\eta}^{x\to y}|\boldsymbol{\eta}) = c(x, y, \boldsymbol{\eta})t + o(t)\,, \text{for t} \searrow 0\,,$$

where $\boldsymbol{\eta}^{x\to y}$ is the configuration with one particle less on the position x and one particle more on the position y then the configuration $\boldsymbol{\eta}$ had.

The assumptions imposed on the infinitesimal transition rates $c(x, y, \boldsymbol{\eta})$ defining the infinitesimal operator $\mathcal{L}$ are reasonable due to the fact, that $\mathcal{L}$ generates strongly continuous Markov semi-group P_t on $\mathcal{C}_b(\mathbf{S}; \mathbb{R})$, which means, that P_t maps $\mathcal{C}_b(\mathbf{S}; \mathbb{R})$ to itself, being stochastically continuous. For $\mathbf{S}$ being a compact space[9], Hille-Yoshida[10] theorem gives us the unique solution to equation (2); i.e. functions $P_t f \in \mathcal{C}_b$, for all $f \in D_{\mathcal{L}}$.

2.3 Canonical & Grand-canonical measures

For the finite sized lattice Λ_L of size $L \in \mathbb{N}$, we have the corresponding state space $\mathbf{S}_L$. Considering the system is closed[11], the number of particles in state $\boldsymbol{\eta}$,

$$\sum\nolimits_L(\boldsymbol{\eta}) := \sum_{x\in\Lambda_L} \eta(x) \in \mathbb{N}\,,$$

is conserved in time for every $\boldsymbol{\eta} \in \mathbf{S}_L$, i.e. $\sum_L(Y(\boldsymbol{\eta}(\cdot), 0)) = \sum_L(Y(\boldsymbol{\eta}(\cdot), t))\,, \forall t > 0$. State space $\mathbf{S}_L$ is composed of non-communicating subsets

$$\mathbf{S}_{L,N} = \{\boldsymbol{\eta} \in \mathbf{S}_L | \sum\nolimits_L(\boldsymbol{\eta}) = N\}\,, \quad N \in \{0, 1, ..., L \cdot \eta_{max}\}\,.$$

This leads to non-uniqueness of the stationary measure on the whole state space. However, on each subset $\mathbf{S}_{L,N}$, the process is irreducible and have the unique stationary measure $\mu_{L,N}$, so-called *canonical*. All stationary canonical measures $\mu_{L,N}$ are exactly the extremal points of the set of stationary measures for a closed system[12],

$$I_e = \{\mu_{L,N} | N \in \{0, 1, ..., L \cdot \eta_{max}\}\}\,,$$

i.e. they are extremal points of convex hull of all stationary measures for the closed system, they are called *pure phases*. On contrary, for an open system, the number of particles is not conserved as they enter and leave the system and the process is irreducible on $\mathbf{S}_L$ and have one unique stationary measure; meaning $|I_e| = 1$.

[7] For the reasons behind these assumptions, see [7] Theorem I.3.9.

[8] So that a particle from arbitrary position can reach every position within a finite time.

[9] This concept was covered by Ligget, for the proof, see [7].

[10] Obtaining Markov process from infinitesimal generator is nicely covered by Kuo in [6], chp.10.9. with comparison to another approach using Kolmogorov equations and Itô theory.

[11] The particles in the system do not enter nor leave.

[12] For the more rigorous statement and its proof, see [7], proposition 1.1.8 .

For the closed system, we can define one measure for the whole state space $\mathbf{S}_L$ as a convex combination of the canonical measures[13]

$$\mu_\phi^L = Z(L,\phi)^{-1} \sum_{N=0}^{L\cdot\eta_{max}} \phi^N Z(L,N)\mu_{L,N}\,,$$

where $Z(L,\phi) = \sum_{N=0}^{L\cdot\eta_{max}} \phi^N Z(L,N)$. Measure μ_ϕ^L is so-called *grand-canonical* measure. It is easy to see that $\mu_{L,N}(\cdot) = \mu_\phi^L(\cdot|\sum_L(\cdot) = N)$. Because $\eta_{max} < \infty$, it is well defined for all $\phi \in [0,\infty)$.

For the system on infinite lattice, for instance consider $\Lambda = \mathbb{Z}^d$, the set of the extremal stationary measures become more complicated due to uncountable state space. Sometimes[14] there exists a one-parameter family of stationary measures μ_ρ for every "density"[15] $\rho \in [0,\eta_{max}]$, where these measures are the only extremal measures, i.e.

$$I_e = \{\mu_\rho | \rho \in [0,\eta_{max}]\}\,.$$

Parameter ρ could be comprehended in the similar way as in the finite system, i.e. as the density of particles in system ($\frac{N}{L}$ in the finite system).

We are interested, on the contrary, if for some ρ there exist more then one extremal stationary measure (pure phase), then we say, that the system exhibit *phase transition*, and if there is no extremal measure beyond a critical density[16], we say, that the system is in *condensation*.

3 Zero-range processes

3.1 Definition and construction

Now we consider state space with the number of particles not bounded on each position, i.e. the local state space $E = \mathbb{N}_0$ and a state space $\mathbf{S} = \mathbb{N}_0^\Lambda$.

We are interested first of all in zero-range processes. A zero range process is defined by its infinitesimal transition rates $c(x, x+y, \boldsymbol{\eta})$, which are dependent solely on the configuration at the position $x \in \Lambda$.[17] The transition means that the number of particles $\eta(x)$ on a position $x \in \Lambda$ decreases by one with the rate $g(\eta(x))$ and the leaving particle jumps to the position $x+y$ with probability $p(y)$ of finite range, i.e. $p(y) = 0\,, |y| > R \in \mathbb{N}$. The infinitesimal transition rates are given by

$$c(x, x+y, \boldsymbol{\eta}) = g(\eta(x)) \cdot p(y)\,.$$

Note that we assume only translation invariant probabilities on the lattice Λ.

[13] For more information, see any book concerning statistical physic. Brief summary could also be found in [5].

[14] Usually assuming certain monotonicity property, see [7], chp. II.2 and for example see [1].

[15] In the sense of the number of particles on the position.

[16] Proper definition of critical density will be given later.

[17] Not being concerned of the position, where the particle jumps in, but being concerned of the state on the position the particle jumps from, gives the model its name - it has zero range of scope.

Because of the state space not being compact, the construction of a Markov process corresponding to a given semi-group of operators, as it was outlined in section 2.1, is not true without any additional assumptions. Moreover, it is not obvious, if on such state space with given infinitesimal generator there exists a semi-group of bounded operators assigned to that generator.

However, it could be shown[18], that if we constrain the state space $\mathbf{S}$ on

$$\tilde{\mathbf{S}} = \{\boldsymbol{\eta} \in \mathbb{N}_0^{\Lambda}, ||\boldsymbol{\eta}||_\alpha < \infty\}\,,$$

where $||\boldsymbol{\eta}||_\alpha = \sum_{x\in\Lambda} \eta(x)\,\alpha(x)$ and $\alpha : \Lambda \mapsto (0,\infty)$ is some suitable function such that

$$\sum_{y\in\Lambda} \alpha(y) < \infty\,, \quad \sum_{y=-R}^{-R} p(y)\,\alpha(x+y) \le M\,\alpha(x)\,, \forall x \in \Lambda\,,$$

for some $M > 0$, and we assume that

$$\sup_{k\in\mathbb{N}} |g(k+1) - g(k)| =: \overline{g} < \infty\,, \quad g(k) > g(0) = 0\,, \forall k \in \mathbb{N}\,, \tag{4}$$

then infinitesimal generator $\mathcal{L}$:

$$\mathcal{L}f(\boldsymbol{\eta}) = \sum_{x,y\in\Lambda} c(x,y,\boldsymbol{\eta})[f(\boldsymbol{\eta}^{x\to y}) - f(\boldsymbol{\eta})]$$

defined for Lipschitz functions $f \in Lip(\tilde{\mathbf{S}};\mathbb{R})$, generates semi-group P_t of the operators on $Lip(\tilde{\mathbf{S}};\mathbb{R})$ with

$$|P_t f(\boldsymbol{\eta}) - P_t f(\boldsymbol{\zeta})| \le l_f \mathrm{e}^{\overline{g}(M+2)t} ||\boldsymbol{\eta} - \boldsymbol{\zeta}||_\alpha\,, \tag{5}$$

for all $\boldsymbol{\eta}, \boldsymbol{\zeta} \in \tilde{\mathbf{S}}$ and for all $f \in Lip(\tilde{\mathbf{S}};\mathbb{R})$, and where l_f is Lipschitz constant for f.

It is worth to note, that P_t is defined for $f \in Lip(\tilde{\mathbf{S}};\mathbb{R}) \subset C(\tilde{\mathbf{S}};\mathbb{R})$ and it is not strongly continuous on $C(\tilde{\mathbf{S}};\mathbb{R})$; however, the property (5) of the semi-group P_t assures that $\mathbb{P}_{\boldsymbol{\eta}}\{Y_t \in \tilde{\mathbf{S}}\} = 1$, thus being conservative and defining process $Y(\boldsymbol{\eta}(\cdot), t)$ for $\boldsymbol{\eta}(\cdot) \in \Omega$ and $t \ge 0$.

Since, in what follows, we will often consider only a finite lattice Λ, note that the zero range process is in this case a countable state space Markov process and as such it is well defined by rates only.

3.2 Stationary measures of zero range processes

Now we consider a finite lattice $\Lambda_L = (\mathbb{Z}/L\mathbb{Z})$ with the periodic boundary; i.e. particles jump from "the last position" to "the first position". It is known[19], that the zero range process on $\mathbf{S} = \mathbb{N}_0^{\Lambda_L}$ defined above has stationary product measures $\mu_\phi^L(\cdot) = \prod_{x\in\Lambda} \nu_\phi^1(x,\cdot)$, with one-point marginal

$$\nu_\phi^1(x,k) = \nu_\phi^1\{\eta(x) = k\} = \frac{1}{Z(\phi)} W(k)\phi^k\,, \tag{6}$$

[18] See Andjel [1], Theorem 1.4 .

[19] This basic result can be found for example in [1], [4] or [5].

with weight $W(k) = \prod_{i=1}^{k} \frac{1}{g(i)}$, fugacity $\phi \in [0, \infty)$, and $Z(\phi) = \sum_{k=0}^{\infty} W(k)\phi^k$ as a normalization constant. Fugacity is connected with the chemical potential α of the system via the formula $\mathrm{e}^{\alpha} = \phi$, thus being an "objective" variable. Then one-point marginals (6) define the grand-canonical product measure μ_{ϕ}^{L} and the corresponding canonical measures given by $\mu_{L,N}(\cdot) = \mu_{\phi}^{L}(\cdot | \sum_L(\cdot) = N)$ with arbitrary $\phi \in [0, \infty)$ as

$$\mu_{L,N}(\boldsymbol{\eta}) = \frac{1}{Z(L,N)} \prod_{x \in \Lambda_L} W(\eta(x))\, \delta(\textstyle\sum_L(\boldsymbol{\eta}), N)\,, \tag{7}$$

where $\delta(\cdot, \cdot)$ is Kronecker delta. For the quantitative description of condensation, we define the expected particle density per position

$$\rho = \mathbb{E}_{\mu_{\phi}^{L}} \eta(x) = \sum_k k \mu_{\phi}^{L}\{\eta(x) = k\} = \sum_k k \nu_{\phi}^{1}(k) =: R(\phi)\,,$$

which is position independent and which is a function of fugacity ϕ. We say, that the system is at critical density if the fugacity goes to a critical value ϕ_c, i.e.

$$\rho_c = \lim_{\phi \nearrow \phi_c} R(\phi)\,,$$

will denote the critical density. The measures μ_{ϕ}^{L} are well defined for all $\phi \in D_\phi \subset [0, \infty)$, which is determined by the radius of convergence of the series $Z(\phi)$. Often $D_\phi = [0, \phi_c)$, and the range of $R(\phi)$ is $D_\rho = R(D_\phi) = [0, \infty)$, i.e. the critical density ρ_c diverges[20]. We are more interested in the case when $D_\phi = [0, \phi_c]$, and $D_\rho = [0, \rho_c]$ for some $\rho_c < \infty$. In such case, the system exhibits condensation. It is achieved for slowly decaying tail of the rates $g(k)$ for large k, which introduces attraction between particles.

3.3 Generic model for condensation

Considering the rates of the following form[21] (8), we can obtain a borderline for rates for which condensation is observed: $\gamma < 1$ and $b > 0$ or $\gamma = 1$ and $b > 2$

$$g(k) = a + \frac{b}{k^{\gamma}}\,, a, \gamma > 0\,, b \in \mathbb{R}\,, \tag{8}$$

otherwise the condensation phenomenon does not appear. We will consider transition rates of the form (8), however, following theorems were proved for general rates, which are uniformly bounded away from zero and which are either uniformly bounded from above or there exists $\lim_{k\to\infty} g(k)$ in $(0, \infty]$.

One would expect, that for a large system size L, $N = \lfloor \rho L \rfloor \to \infty$ with a fixed particle density ρ, the canonical measure should be close to the grand-canonical measure in some sense. An important question arises, what happens with the grand-canonical measure if the limit of canonical measures is considered under a particle density $\rho > \rho_c$.

[20] For example in the case of non-decreasing rates $g(k)$; see [4].

[21] Introduced by Evans, see [4], studied further by Grosskinsky, see [5].

The convergence of canonical measures $\mu_{L,\lfloor\rho L\rfloor}$ to the grand-canonical product measure $\nu_{\Phi(\rho)}$ was proved[22] in the weak sense, and later also the convergence in the norm was proved[23] when we eliminate the most occupied position in the system, i.e. $\arg\max_{x\in\Lambda}\eta(x)$.

The former result is based on the following theorem considering only the n-point marginal and thus not consisting of the most occupied position.

Theorem 1. [24] *Let $\Phi(\rho)$ be defined by*

$$\Phi(\rho) = \begin{cases} R^{-1}(\rho), & \text{for } \rho < \rho_c \\ \phi_c, & \text{for } \rho \geq \rho_c \end{cases} .$$

Then the relative entropy of n-point marginals $\mu^n_{L,\lfloor\rho L\rfloor}$ and $\nu^n_{\Phi(\rho)}$ asymptotically vanishes, i.e.

$$\lim_{L\to\infty} H(\mu^n_{L,\lfloor\rho L\rfloor}|\nu^n_{\Phi(\rho)}) = 0\,,$$

for every $n \in \mathbb{N}$ and $\rho \in [0,\infty)$.

The following result concers the most occupied position in the case, when the chosen ρ is above the critical value ρ_c.

Theorem 2. [25] *Let $\nu^1_{\phi_c}(k)$ have a monotonic decreasing power law tail (write $\nu^1_{\phi_c}(k) \simeq k^{-b}$) with $b > 2$ and finite first moment ρ_c. Then for every $\rho > \rho_c$ the normalized maximum occupation number satisfies a weak law of large numbers, namely it converges in probability as*

$$\frac{1}{(\rho-\rho_c)L}\max_{x\in\Lambda_L}\eta(x) \xrightarrow{\mu_{L,\lfloor\rho L\rfloor}} 1\,, \quad \text{for } L\to\infty\,,$$

where $(\rho-\rho_c)L$ is the number of all excess particles in the system.

So far we know, that the typical configuration in the limit $L\to\infty$ has all positions except one, which is randomly chosen, distributed according to ν_{Φ_c}, and all excess particles are gathered on the one position forming the condensate.

3.4 Relaxation dynamics of ZRP

In this section we briefly describe results of simulations. We are interested in the relaxation time needed for the system to relax into the stationary state. Also, we are interested in dynamics of condensates; for the purpose of this work, we assume the rates $g(k)$ to be one of form (8), which are non-decreasing and allowing condensations. We start with the finite number of positions, $\Lambda_L = \{1,...,L\}$, with $N \in \mathbb{N}$ particles. By analyzing the normalization function $Z(\phi)$ and the particle density $R(\phi)$ near criticality[26], i.e. in the limit $\phi\to\phi_c$, we could find out the critical density

$$\rho_c = \begin{cases} \infty & \text{for } b \leq 2 \\ \frac{1}{b-2} & \text{for } b > 2 \end{cases} .$$

[22] For the proof see [5].
[23] For a proof see [2].
[24] Cite from [5], Theorem 5.2 .
[25] Cite from [5], Theorem 5.5 .
[26] By expanding proper hypergeometric function.

The initial configuration of the system is set to be "uniform", meaning all positions contain the same number of particles $\frac{N}{L} = \rho > \rho_c$ and the system contain $(\rho - \rho_c)L$ excess particles. Each position containing at least a α-fraction of the number of all excess particles, will be denoted as *cluster*. The coefficient α is from interval $(0, 1)$ and should be small enough.

The behavior of the system can be divided into 3 phases as could be seen from Figure 1:

1. Nucleation - particles are gathering on few positions, so-called *clusters*. This phase is very unstable (in the terms of time spent in this phase) and it is not physically interesting.

2. Coarsening - clusters exchange their particles and grow at the expense of the smaller ones, which finally leads to the saturation.

3. Saturation - when only one cluster survives with all $(\rho - \rho_c)L$ excess particles. This is the stationary distribution for the finite systems

For a further description of these phases some natural assumptions are needed (they arose from heuristic analysis of the process); the assumptions[27] of *separation of time scales* and *independence of excess particles in the bulk* express the average time a particle needs to move from one condensate to another one and that all positions except the clusters behave as a homogeneous medium, where excess particles move independently.

Based on these assumptions, it can be derived[28], that a typical condensate size grows with time according to

$$m(t) \sim t^{\beta}\,, \quad \text{where } \beta \in [\frac{1}{2}, 1]\,.$$

As simulations demonstrate, clusters exchange particles until only few of clusters survive. The saturation regime starts when only two clusters are surviving and exchanging particles. We would like to know the dynamics of exchanging. Let us describe it by the *master-equation* for condensate size m

$$\partial_t q(m,t) = -q(m,t)\left[\frac{1}{\frac{m}{M}} + \frac{1}{1-\frac{m}{M}}\right] + q(m-1,t)\frac{1}{1-\frac{m-1}{M}} + q(m+1,t)\frac{1}{\frac{m+1}{M}}\,, \qquad (9)$$

where $M = (\rho - \rho_c)L$ are all excess particles in the system, $q(m,t)$ is the probability to find m particles on one condensate and $M - m$ particles on the other one at time t.

For any initial condition the solution of (9) tends to the inverse binomial distribution $q^*(m) = 1/\binom{M}{m}$, with the two extreme occupation numbers $m = 0$ and $m = M$, which are most probable in the limit $L \to \infty$. Both with the probability $1/2$ of occurring.

4 Discussion

The goal of this paper was to lay some basic facts about zero range processes as the base for further research, to get familiar with the underlying Markov processes and to get some

[27] For a further explanation of these assumptions, see [5], chp. 6.2.1 .

[28] See [5].

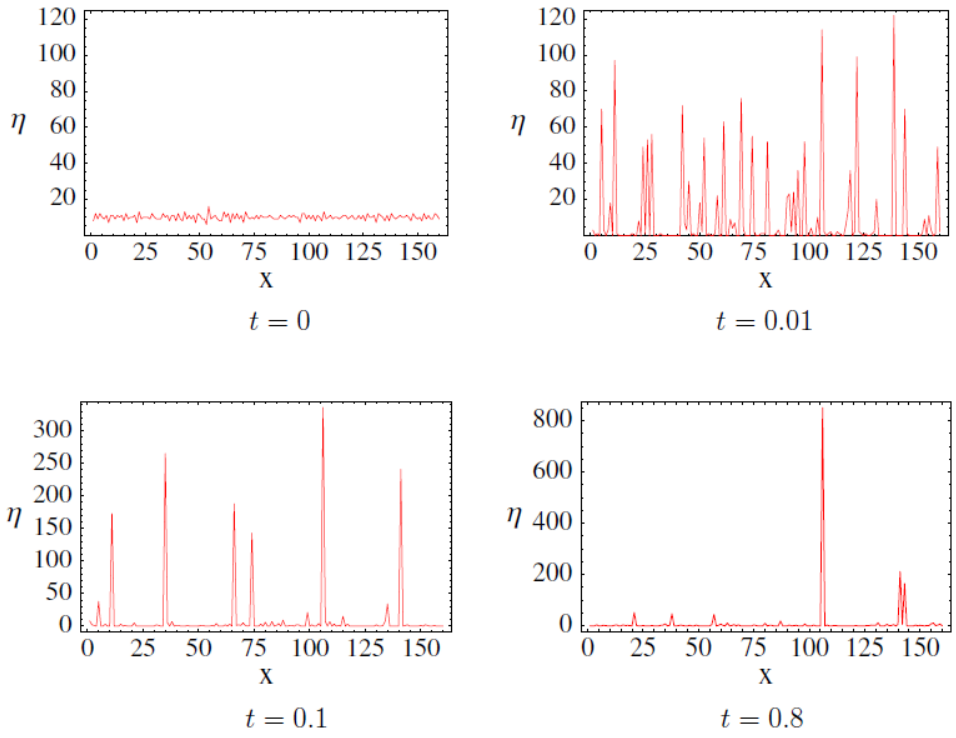


Figure 1: Relaxation dynamics in ZRP. This figure was taken from the [5].

insight into relaxation dynamics through simulations of zero range processes. This paper does not aim to summarizes the all know facts, which is not possible on a few pages. However, we restricted ourself on the periodic boundaries of lattice our motivation of this research is comparison between this simple model with results from traffic data.

5 Acknowledgement

In this place, I would like to thank my supervisor Mgr. Lucie Fajfrová, PhD. for suggesting this interesting topic, this beautiful and modern theory of Markov process, for advises and suggestions, but also for encouragement in the work and last but not least for her time and care with the topic and goals.

References

[1] E.D. Andjel. *Invariant measures for the zero range process. Ann. Probability*, **10** (1982), 525-547,.

[2] I. Armendáariz, S. Grosskinsky, M. Loulakis. *Zero-range condensation at criticality.* In 'Stoch. Proc. Appl.', **123(9)** (2013), 3466-3496 .

[3] E.B. Dynkin. *Markov Processes, vol I.* Springer-Verlag (1965)

[4] M.R. Evans. *Phase transitions in one-dimensional nonequilibrium systems.* In 'Braz. J. Phys.', **30** (2000), 42-57,

[5] S. Grosskinsky. *Disseration: Phase transition in nonequilibrium stochastic particle systems with local conservation laws.* Zentrum Mathematik Lehrstuhl fur Mathematische Physik (2004.)

[6] H.-H. Kuo. *Introduction to Stochastic Integration.* Springer (2006).

[7] T. M. Liggett. *Interacting Particle Systems.* Springer (1985)

Heuristic Time Complexity Analysis via Markov Chain*

Matej Mojzeš

3rd year of PGS, email: mojzemat@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. This paper introduces searching process as a means for analysis of discrete optimization heuristic algorithms. Performance of a searching process on a task with finite number of states is studied via Markov chain. Using this approach a thorough comparison of three different time complexity measures, which are introduced in the paper as well, is performed. According to the output of the measures on three different tasks the Q_∞ measure seems to provide most reasonable results for heuristic performance analysis.

Keywords: Integer optimization, heuristic algorithm, time complexity measure, Markov chain

Abstrakt. Táto práca predkladá prehľadávací proces ako prostriedok na analýzu heuristických algoritmov pre celočíselnú optimalizáciu. Výkonnosť prehľadávacieho procesu na úlohe s konečným počtom stavov je analyzovaná s využitím Markovových reťazcov. Týmto prístupom je vykonaná dôkladná analýza troch rôznych mier časovej náročnosti, ktoré su taktiež prezentované v práci. Na základe výstupov mier na troch rôznych úlohách je doporučovaná miera Q_∞, ktorá sa javí, že poskytuje najprimeranejšie výsledky pre analýzu výkonnosti heuristiky.

Kľúčové slová: Celočíselná optimalizácia, heuristika, miera časovej náročnosti, Markovov reťazec

1 Introduction

Researchers dealing with optimization problems and/or developing their own optimization heuristics are interested in time complexity measures since they can be used to determine the difficulty of distinct optimization problems and also to evaluate the suitability of given optimization heuristic for given task.

When dealing with discrete optimization task using a heuristic approach, the algorithm *searches* through the space of feasible solutions, or states, to find any of the goal states, which are optimal, or sub-optimal, solutions of the given problem. More formally, let $\mathbf{U}$ be a non-empty *set of states*. Let $\mathbf{G} \subset \mathbf{U}$ be a non-empty *set of goals*. Any state $\mathbf{x} \in \mathbf{G}$ is called a *solution* of the *searching task* $\langle \mathbf{U}, \mathbf{G} \rangle$. Let $N \in \mathbb{N}$ be the maximum number of searching steps. Any algorithm generating the sequence of $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathbf{U}^N$ is called a *searching process* (SP). The *number of searching steps* (time complexity of SP) is defined as $n = \min\{k \in \mathbb{N} \,|\, \mathsf{x}_k \in \mathbf{G}\}$, i.e. in the case of successful search. Should the search end with a failure we set $n = +\infty$.

*This paper has been supported by the grant OHK4-165/11 CTU in Prague

To make a relevant example, a typical discrete instance of *Optimization Problem* (OP) defined as the minimization of an objective function f: $\mathbf{D} \to \mathbb{R}$ where $\mathbf{D} = \{\mathbf{x} \in \mathbb{Z}^n \mid \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$ is an appropriate integer domain may be regarded as searching task $\langle \mathbf{U}, \mathbf{G} \rangle$ where $\mathbf{G} = \{\mathbf{x} \in \mathbf{U} \,|\, \mathrm{f}(\mathbf{x}) \leq \mathrm{f}_{\mathrm{opt}}\}$ with $\mathrm{f}_{\mathrm{opt}} = \min\{\mathrm{f}(\mathbf{x}) \,|\, \mathbf{x} \in \mathbf{U}\}$.

We may quite realistically suppose that the SP is produced by a stochastic algorithm and the complexity n is a stochastic variable with the domain of $\mathbf{D}_{\mathrm{SP}} = \{1, 2, \ldots, N, +\infty\}$ and densities $p_n \geq 0, \forall n \in \mathbf{D}_{\mathrm{SP}}$ satisfying $\sum p_n = 1$. The value of p_n for $n \leq N$ may be interpreted as the probability of finding the solution in n-th step of the SP. Moreover, we may define $p_{\mathrm{succ}} = \sum_{n=1}^{N} p_n$ as the probability of success and $p_\infty = 1 - p_{\mathrm{succ}}$ as the probability of failure in a single run of SP. In the following we will be studying SP with $p_{\mathrm{succ}} > 0$ only.

2 Traditional Approach to Time Complexity Measures

To study the behaviour of a SP we may use three widely used, e.g. by Yang and Deb [1], basic characteristics:

- p_{succ} as *reliability* of the SP,
- $\mathrm{E}\, n = p_{\mathrm{succ}}^{-1} \sum_{n=1}^{N} n\, p_n$ as *mean number of searching steps* in the case of successful search,
- $\sqrt{\mathrm{D}\, n} = p_{\mathrm{succ}}^{-1/2} \left(\sum_{n=1}^{N} (n - \mathrm{E}\, n)^2\, p_n\right)^{1/2}$ as *standard deviation of the searching step number* in the case of successful search.

To address the fundamental problem of measuring SP *time complexity* and thus performance, as long as the SP has reliability of $p_{\mathrm{succ}} = 1$, a very straightforward criterion of mean number of steps $\mathrm{E}\, n$ is frequently used. On the other hand, for $0 < p_{\mathrm{succ}} < 1$ we have to adjust the value of $\mathrm{E}\, n$ due to decreased reliability of the SP.

An example of adjusted time complexity evaluation is based on the Feoktistov criterion [2],

$$FEO = \frac{\mathrm{E}\, n}{p_{\mathrm{succ}}} \,. \tag{1}$$

Other authors use similar criterion as well, for example [6] defines time complexity measure $SP1$ as

$$SP1 = \frac{\mathrm{E}\, T_A^s}{p_{\mathrm{s}}} \,, \tag{2}$$

where $p_{\mathrm{s}} \in (0, 1]$ is probability of success and T_A^s number of evaluations for a run of an heuristic algorithm. Therefore $SP1 = FEO$.

Following the traditional approach we are able to search for optimal value of FEO,

$$FEO_{\mathrm{opt}} = \min\{FEO \,|\, N \in \mathbb{N}\} \,, \tag{3}$$

and more importantly to find the minimal number of steps that guarantee optimal quality, $N_{\mathrm{FEO,opt}}$,

$$N_{\mathrm{FEO,opt}} = \min\{N \in \mathbb{N} \,|\, FEO = FEO_{\mathrm{opt}}\} \,, \tag{4}$$

after which the SP should be terminated.

3 Extended Searching Process Measure

The recommendation of optimal running and even restarting strategy may be performed via analysis of the *Extended Searching Process* (XSP) time complexity analysis.

XSP is based on the idea of following trivial, but very practical habit – if the SP is successful in the first run, then the searching task is done. Otherwise, should the process end with a failure, we continue to repeat new runs until succeeding.

Let $k \in \mathbb{N}$ be the index of an SP run. Let $j = 1, 2, \ldots, N$ be the searching step index inside the individual SP run. The distinct runs of SP are supposed to be independent and, therefore, the XSP successfully terminates in the $n-$th step with probability

$$p_n^* = p_{N(k-1)+j}^* = (1 - p_{\text{succ}})^{k-1} p_j \ . \tag{5}$$

We can directly calculate

$$\mathrm{E}\, n^* = \sum_{n=1}^{\infty} n\, p_n^* = \tag{6}$$

$$= \sum_{k=1}^{\infty} (1 - p_{\text{succ}})^{k-1} \sum_{j=1}^{N} (N(k-1) + j) p_j = \qquad \frac{N\, p_{\text{succ}}\, (1 - p_{\text{succ}})}{p_{\text{succ}}^2} + \frac{p_{\text{succ}}\, \mathrm{E}\, n}{p_{\text{succ}}} \ . \tag{7}$$

Hence, the relationship between time complexities of the XSP and the original SP is

$$\mathrm{E}\, n^* = \mathrm{E}\, n + N \cdot \frac{1 - p_{\text{succ}}}{p_{\text{succ}}} \ . \tag{8}$$

Finally, this formula can be used directly to build up the quality criterion of XSP and, as has been noted, of SP time complexity, the Q_∞ *measure*:

$$Q_\infty = \mathrm{E}\, n + N \cdot (p_{\text{succ}}^{-1} - 1) \ . \tag{9}$$

Despite using different statistical reasoning similar approach was proposed by [6]:

$$SP2 = (\frac{1 - p_s}{p_s}) FE_{\max} + \mathrm{E}\, T_A^s \ , \tag{10}$$

where $FE_{\max}$ is the maximum number of function evaluations and therefore: $SP2 = Q_\infty$. It is obvious that the this criterion is quite similar to the Feoktistov's one. In fact, there is a very clear relation between the two. Starting from the inequality of $n \leq N$ we obtain $\mathrm{E}\, n \leq N$ and then $Q_\infty \geq \mathrm{E}\, n + \mathrm{E}\, n \cdot (p_{\text{succ}}^{-1} - 1) = \mathrm{E}\, n / p_{\text{succ}} = FEO$. We can conclude that Feoktistov's criterion is the lower bound of the novel criterion of SP time complexity.

Again, this way we are able to search for optimal value of Q_∞,

$$Q_{\infty,\text{opt}} = \min\{Q_\infty \,|\, N \in \mathbb{N}\} \ , \tag{11}$$

and more importantly to find the minimal number of steps that guarantee optimal quality, N_{opt},

$$N_{\text{Q,opt}} = \min\{N \in \mathbb{N} \,|\, Q_\infty = Q_{\infty,\text{opt}}\} \ , \tag{12}$$

after which the XSP should be terminated.

4 Random Shooting Envelope Measure

Another measure can be based on the rather trivial idea of random shooting heuristic algorithm. Let μ be number of states and ν be number of goal states. Then $p_{\mathrm{HIT}} = \nu/\mu$ is *probability of goal hitting* by a single random shot and number of evaluations has geometric distribution with probability density function (PDF) and cumulative distribution function (CDF) as follows:

$$p_n = p_{\mathrm{HIT}} \cdot (1 - p_{\mathrm{HIT}})^{n-1} \ , \tag{13}$$

$$F_n = 1 - (1 - p_{\mathrm{HIT}})^n \ . \tag{14}$$

Defining *time constant* as

$$T = \frac{1}{-\ln(1 - p_{\mathrm{HIT}})} > 0 \ , \tag{15}$$

we can reformulate CDF of *random shooting* as

$$F_n = 1 - \exp(-n/T) \ . \tag{16}$$

Random shooting is the only one heuristic which can be restarted without any change on CDF. That is why we have decided to compare random shooting CDF with CDF of given searching heuristics.

First, we define *upper bound* (envelope) of given CDF by condition:

$$\forall n \in \mathbb{N} : F_n \leq 1 - \exp(-n/T) \ . \tag{17}$$

After inequality rearrangement we obtain upper bound for time constant

$$\forall n \in \mathbb{N} : T \leq \frac{n}{-\ln(1 - F_n)} \tag{18}$$

which is the same as

$$T \leq \min\{\frac{n}{-\ln(1 - F_N)} \,|\, N \in \mathbb{N}\} \ . \tag{19}$$

This motivates us to define *random shooting time*:

$$T_{\mathrm{RS}} = \frac{N}{-\ln(1 - F_N)} \tag{20}$$

as third time complexity measure and to find its optimum value

$$T_{\mathrm{RS,opt}} = \min\{T_{\mathrm{RS}} \,|\, N \in \mathbb{N}\} \tag{21}$$

and corresponding optimum interruption time as

$$N_{\mathrm{RS,opt}} = \min\{N \in \mathbb{N} \,|\, T_{\mathrm{RS}} = T_{\mathrm{RS,opt}}\} \ . \tag{22}$$

5 Finite Time Horizon

Let N be number of searching steps in as single run, H be the number of independent serial runs, and M be constrain of total step number. In such case we have to minimize failure probability

$$p_{\text{fail}} = 1 - p_{\text{succ}} = (1 - F_N)^H \tag{23}$$

subject to $H \cdot N \leq M$. This constrained integer minimization task with two independent variables H, N can be converted to

$$\Phi = -\ln p_{\text{fail}} = \max \tag{24}$$

with $H = \lfloor M/N \rfloor$. Therefore, we obtain one-dimensional optimization task with unknown N as

$$\Phi = -\lfloor M/N \rfloor \cdot \ln(1 - F_N) = \max \ . \tag{25}$$

If M is large, we can approximate

$$\Phi \approx -\frac{M}{N} \ln(1 - F_N) = \frac{M}{T_{\text{RS}}} = \max \ . \tag{26}$$

Therefore, maximization of Φ is approximately minimization of T_{RS}, which is equivalent to random shooting envelope if $M \to \infty$, what can be written exactly as

$$T_{\text{RS}} = \lim_{M \to \infty} \frac{M}{\Phi} = \min \ . \tag{27}$$

There is also relationship between Φ criterion and Q_M measure as

$$Q_M = s_N^{k-1} \cdot s_j \approx s_N^k = (1 - F_N)^k = (1 - F_N)^{\lfloor M/N \rfloor} \ , \tag{28}$$

where s_j stands for failure probability $s_j = 1 - \sum_{k=1}^{j} p_k$ for $j = 0, 1, \ldots, N$.
Therefore

$$\lim_{M \to \infty} \frac{M}{-\ln Q_M} = T_{\text{RS}} = \min \ . \tag{29}$$

Finally, we recognize the equality of finite time horizon, random shooting envelope, and Q_M approaches which is a support argument for T_{RS} complexity measure and it suppresses individual examination of Φ and Q_M measures.

6 Searching Process Analysis via Hypothetical Searching Process

Having three time complexity measures FEO_{opt}, $Q_{\infty,\text{opt}}$ and $T_{\text{RS,opt}}$, we would like to compare their properties via hypothetical and real-world scenarios. First, we define three hypothetical searching processes with parameters $n_0 \in \mathbb{N}$ and $0 < p_{\text{succ}} \leq 1$. Their common characteristics are identical reliabilities p_{succ} and maximum running times n_0.

Degenerated Search (DES) has CDF

$$F_n = \begin{cases} 0, & n < n_0 \\ p_{\text{succ}}, & n \geq n_0 \ . \end{cases} \tag{30}$$

Trimmed Linear Searching Process (TLS) has CDF

$$F_n = \min(\frac{n \cdot p_{\text{succ}}}{n_0}, p_{\text{succ}}) \ . \tag{31}$$

Trimmed Random Shooting (TRS) has CDF

$$F_n = \min(1 - \exp(-n/T^*), p_{\text{succ}}) \tag{32}$$

where

$$T = \frac{n_0}{-\ln(1 - p_{\text{succ}})} \ . \tag{33}$$

If we terminate DES, TLS, TRS after $N = n_0$ steps, we obtain identical reliabilities p_{succ}, but different values of $\mathrm{E}\, n$:

$$\mathrm{E}\, n_{\text{DES}} = n_0 \ , \tag{34}$$

$$\mathrm{E}\, n_{\text{TLS}} = \frac{(n_0 + 1)}{2} \ , \tag{35}$$

$$\mathrm{E}\, n_{\text{TRS}} = \frac{(1 - p_{\text{succ}})^{1/n_0}}{1 - (1 - p_{\text{succ}})^{1/n_0}} \cdot \frac{1 - (1 - p_{\text{succ}}) \cdot (1 - p_{\text{succ}})^{1/n_0}}{p_{\text{succ}}} \ . \tag{36}$$

It is easy to prove $\mathrm{E}\, n_{\text{TRS}} \leq \mathrm{E}\, n_{\text{TLS}} \leq \mathrm{E}\, n_{\text{DES}}$ for $n_0 \geq 2$. Therefore, TRS is the fastest and DES is the slowest with the same reliabilities.

Applying three time complexity measures FEO_{opt}, $Q_{\infty,\text{opt}}$ and $T_{\text{RS,opt}}$, we can compare their decisive power. General results are collected in Tab. 1 except the case of FEO_{opt} for TRS which has to be investigated numerically. The dependency of measure values on reliability p_{succ} is demonstrated in Tab. 2 for $n_0 = 1000$.

Table 1: General comparison of measures

Process	FEO_{opt}	$Q_{\infty,\text{opt}}$	$T_{\text{RS,opt}}$	$N_{\text{FEO,opt}}$	$N_{\text{Q,opt}}$	$N_{\text{RS,opt}}$
DES	$\frac{n_0}{p_{\text{succ}}}$	$\frac{n_0}{p_{\text{succ}}}$	$\frac{n_0}{-\ln(1-p_{\text{succ}})}$	n_0	n_0	n_0
TLS	$\frac{n_0+1}{2p_{\text{succ}}}$	$\frac{n_0}{p_{\text{succ}}} - \frac{n_0-1}{2}$	$\frac{n_0}{-\ln(1-p_{\text{succ}})}$	n_0	n_0	n_0
TRS	numerically	$(1-(1-p_{\text{succ}})^{1/n_0})^{-1}$	$\frac{n_0}{-\ln(1-p_{\text{succ}})}$	numerically	1	1

For $n_0 \geq 2$ we observed:

$$FEO_{\text{opt,TRS}} < FEO_{\text{opt,TLS}} < FEO_{\text{opt,DES}} \ , \tag{37}$$

$$Q_{\infty,\text{opt,TRS}} < Q_{\infty,\text{opt,TLS}} < Q_{\infty,\text{opt,DES}} \ , \tag{38}$$

$$T_{\text{RS,opt,TRS}} = T_{\text{RS,opt,TLS}} = T_{\text{RS,opt,DES}} \ . \tag{39}$$

Therefore, $T_{\text{RS,opt}}$ measure of time complexity does not reflect the differences among DES, TLS, and TRS. Remaining measures $FEO_{\text{opt}} \leq Q_{\infty,\text{opt}}$ are of same importance, but we prefer $Q_{\infty,\text{opt}}$ for its relationship to XSP and also because $Q_{\infty,\text{opt}}$ advocates serial repetition of the searching process.

Table 2: Numerical comparison of measures for $n_0 = 1000$

p_{succ}	Process	DES	TLS	TRS
0.1	FEO_{opt}	10000.00	5005.00	4802.11
	$Q_{\infty,\text{opt}}$	10000.00	9500.50	9491.72
	$T_{\text{RS,opt}}$	9491.22	9491.22	9491.22
0.3	FEO_{opt}	3333.33	1668.33	1432.66
	$Q_{\infty,\text{opt}}$	3333.33	2833.83	2804.17
	$T_{\text{RS,opt}}$	2803.67	2803.67	2803.67
0.5	FEO_{opt}	2000.00	1001.00	743.53
	$Q_{\infty,\text{opt}}$	2000.00	1500.50	1443.20
	$T_{\text{RS,opt}}$	1442.70	1442.70	1442.70
0.9	FEO_{opt}	1111.11	556.11	229.43
	$Q_{\infty,\text{opt}}$	1111.11	611.61	434.79
	$T_{\text{RS,opt}}$	434.29	434.29	434.29

7 Markovian Simulation

The rather simple and straight-forward, but sometimes also very effective, heuristic we will use for experimental case study is the well-known *Shoot and Go* (SG) or *Iterated Local Search* (ILS) algorithm [4]. In our implementation the random solution is improved iteratively via local search in its neighbourhood. Effectiveness of this approach based on steepest descent is dependent mainly on the neighbourhood shape and size. General neighbourhood of $\mathbf{x} \in \mathbf{D}$ can be defined as $R_{p,\rho}(\mathbf{x}) = \{\mathbf{y} \in \mathbf{D} \,|\, ||\mathbf{x} - \mathbf{y}||_p \leq \rho\}$, where p is norm parameter and ρ is neighbourhood radius. In our experimental study we will apply Manhattan ($p = 1$) and Hamming norm ($p =$ `'HAMM'`) and a small neighbourhood size ($\rho = 1, 2$). Periodic extension of $\mathbf{D}$ is also possible but it is useful only in combination with Manhattan norm.

The resulting SP can be studied as a Markov chain [5] with a finite number of states as long as the $p_1, p_2, \ldots, p_N$ probabilities can be calculated for given N. The numerical study was performed for following three problems.

Weighted Sum Problem having objective function $\text{f}(\mathbf{x}) = \sum_{k=1}^{d} w_k x_k$, where $0 \leq x_k \leq R$, $R \in \mathbb{N}$, $w_k = a^{1/k}$, $a > 1$ is a relatively uncomplicated integer objective function with minimum at $\mathbf{0}$. Study for $d = 5$ and $R = 3$ was performed and results are collected in Tab. 3. We may clearly see that all criteria tend to prefer smaller neighbourhood size and the results suggest to proceed with searching in a relatively long runs ($Q_{\infty,\text{opt}}$ and $T_{\text{RS,opt}}$) – in other words, this criteria "trust" the heuristic.

Knapsack Problem having objective function $\text{f}(\mathbf{x}) = -\sum_{k=1}^{d} \pi_k x_k + \lambda \cdot \max(0, \sum_{k=1}^{d} w_k x_k - w^*)$, $0 \leq x_k \leq R$, $R \in \mathbb{N}$, $\pi_k, w_k, w^*, \lambda > 0$ with both weights $\mathbf{w}$ and item values $\boldsymbol{\pi}$ coming from geometrical sequence could be considered a more intricate integer objective function. Results of a study for $d = 5$ and $R = 3$ are collected in Tab. 4. As opposed to the first example, smaller neighbourhood is preferred only for the Manhattan norm. Smaller N advised by all criteria reflects increased difficulty of this problem as well.

Clerc's Zebra3 problem is a non-trivial binary optimization problem and part of discrete optimization benchmark problems [3]. Objective function value f is the sum of the result

Table 3: Weighted sum problem simulation results

Norm	Extension	ρ	FEO_{opt}	$Q_{\infty,\text{opt}}$	$T_{\text{RS,opt}}$	$N_{\text{FEO,opt}}$	$N_{\text{Q,opt}}$	$N_{\text{RS,opt}}$
Manhattan	None	1	25.3971	25.5177	6.5342	49	225	225
		2	40.6063	43.2484	25.2317	69	833	768
	Periodic	1	29.4557	29.9991	10.9	55	362	363
		2	60.1713	70.958	55.1548	85	1712	1538
Hamming	None	1	29.7584	30.9355	15.1241	53	500	501
		2	86.4279	116.3463	104.9813	92	3045	2716

Table 4: Knapsack problem results

Norm	Extension	ρ	FEO_{opt}	$Q_{\infty,\text{opt}}$	$T_{\text{RS,opt}}$	$N_{\text{FEO,opt}}$	$N_{\text{Q,opt}}$	$N_{\text{RS,opt}}$
Manhattan	None	1	159.74	261.12	258.44	30	21	21
		2	204.44	360.26	358.01	43	29	28
	Periodic	1	175.27	297.51	295.52	26	18	18
		2	221.15	399.94	398.26	44	29	29
Hamming	None	1	206.11	350.82	349.09	23	16	16
		2	189.38	337.07	334.77	66	48	48

of applying the function (40) to consecutive groups of three components each, if the rank of the group is even, or (41) otherwise. The maximum value is $d/3$, where d is dimension of the problem. We may use the value of $d/3 - f$ as modified objective function and search for its optimum, $\mathbf{0}$. Results of a study for $d = 12$ are collected in Tab. 5. In the case of Clerc's Zebra3 problem only more appropriate Hamming norm is presented, since the Manhattan one is practically the same (both with or without periodic extension). Results for $\rho = 1$ indicate the hardest problem of the three with minimal suggested N. Nevertheless, it may be of interest that by increasing the neighbourhood size we are able to significantly simplify the problem and advance with the search in much longer run.

$$\mathrm{f}_1(\mathbf{x}) = \begin{cases} 0.9 & |y| = 0 \\ 0.6 & |y| = 1 \\ 0.3 & |y| = 2 \\ 1.0 & |y| = 3 \end{cases} \qquad (40)$$

$$\mathrm{f}_2(\mathbf{x}) = \begin{cases} 0.9 & |y| = 3 \\ 0.6 & |y| = 2 \\ 0.3 & |y| = 1 \\ 1.0 & |y| = 0 \end{cases} \qquad (41)$$

Table 5: Clerc's Zebra3 problem results

Norm	Extension	ρ	FEO_{opt}	$Q_{\infty,\text{opt}}$	$T_{\text{RS,opt}}$	$N_{\text{FEO,opt}}$	$N_{\text{Q,opt}}$	$N_{\text{RS,opt}}$
Hamming	None	1	1597.17	2682.56	2681.71	10	7	7
		2	536.03	970.44	967.46	68	46	46

8 Conclusions

Even a very well designed heuristic should be terminated in the right moment and re-started in order to improve its chances of success. To accomplish this goal, the researcher should examine multiple runs of the heuristic. According to observed probabilities of

finding the optimum in distinct number of evaluations, the investigator should apply their preferred complexity measure and plan optimal termination and re-starting strategy accordingly. It may be useful also to get feedback on deliberation of the algorithm – instance with higher termination point can be regarded as more trusted to find the solution in one run. This way, one could also identify sophisticated heuristics in terms of their ultimate results and not based on their computational complexity.

In this paper, we have compared three different criteria that can examine performance of given heuristic algorithm on given problem. Using this criteria while studying the results of presented Markovian simulation and also performance of our own heuristics we suggest that the Q_∞ measure is worth using. While T_{RS} is a rather extreme criterion and FEO well-known and relevant criterion, we propose the Q_∞ measure for being more appropriately sensitive to performance of given heuristic on given problem and thus can provide important and suitable feedback when tuning algorithms.

References

[1] X.-S. Yang, S. Deb. *Cuckoo search via Lévy flights*, In: Proceedings of World Congress on Nature & Biologically Inspired Computing, IEEE Publications (2009), 210–214.

[2] V. Feoktistov. *Differential Evolution: In Search of Solutions*, Springer (2006).

[3] C. Hanning, Z. Yunlong, H. Kunyuan, H. Xiaoxian. *Hierarchical Swarm Model: A New Approach to Optimization*, Discrete Dynamics in Nature and Society, Vol.2010 (2010).

[4] H. Ramalhinho Dias Lourenço, O. C. Martin, T. Stutzle. *Iterated Local Search*, U. Pompeu Fabra Economics Working Paper, Vol.513 (2000).

[5] S. Meyn, R. L. Tweedie. *Markov Chains and Stochastic Stability*, Cambridge University Press (2009).

[6] A. Auger, N. Hansen. *Performance evaluation of an advanced local search evolutionary algorithm*, The 2005 IEEE Congress on Evolutionary Computation, Vol.2 (2005), 1777–1784.

A New Approach to Photo-Response Non-Uniformity Calculation*

Adam Novozámský†

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

advisor: Stanislav Saic, Institute of Information Theory and Automation, AS CR

Abstract. There are two essential tasks in image forensics. Integrity verification (genuineness analysis) of digital images and image ballistics. In image ballistics we address the problem of linking digital images under investigation to either a group of possible source imaging devices or to one particular source imaging device which has been used to capture these photos. The latter one is the main topic of this paper. Specifically, we develop a novel method to identify the source camera of a digital image by using its sensor unevenness caused by Photo-Response Non-Uniformity (PRNU).

Keywords: image forensics, PRNU, source camera identification

Abstrakt. Ve forenzní analýze obrazu existují dvě základní úlohy. Ověření integrity (autentičnosti) digitálního obrazu a obrazová balistika. V obrazové balistice řešíme problém nalezení typu záznamového zařízení, nebo konkrétního fotoaparátu, který byl použit k zachycení snímku. Tato druhá úloha je hlavním tématem našeho příspěvku. Konkrétně jsme vyvinuli novou metodu, jak identifikovat u digitálního obrazu zdrojové zařízení pomocí jeho senzorové nekonzistence způsobené PRNU.

Klíčová slova: forenzní analýza obrazu, PRNU, identifikace fotoaparátu

1 Introduction

Since image ballistics makes possible to differentiate between source cameras of the same make and model, it became especially useful in the forensic, law enforcement, insurance, and media industries. Insurance companies, for example, often need to know whether or not claim-substantiating photos were taken by the person looking for compensation. Law enforcement agencies are also tasked with finding the source camera when criminal activity is discovered in digital images (e.g. child pornography, etc).

Although in past researchers mainly were focused on data hiding and digital watermarking approach to carry out digital image integrity verification and image ballistics, today there is a relatively new approach called passive one which does not need embedding any secondary data into the image. So, in contrast to active methods, the passive approach does not need any prior information about the image being analyzed. There

*This work has been supported by the grants GAČR 13-28462S and VG20102013064
†Institute of Information Theory and Automation, AS CR

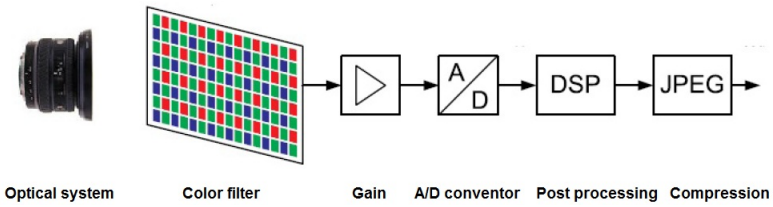


Figure 1: A typical digital camera system.

have been developed methods to detect image splicing, traces of non-consistencies in color filter array interpolation, traces of geometric transformations, cloning, computer graphics generated photos, JPEG compression inconsistencies, etc. Typically, pointed out methods are based on the fact that digital image editing brings specific detectable statistical changes into the image.

Our aim in this paper is to uncover some important drawbacks of existing source identification methods and analytically develop a novel way to identify particular source cameras by employing their sensor properties [1, ?]. Specifically, we will use the multiplicative nature of PRNU noise component present in digital images. Moreover, we also will deal with artifacts brought into the image by vignetting, JPEG, and embedded camera software. Effectiveness of proposed analytical concept will be experimentally measured and compared to state-of-the-art.

2 Fingerprints of Different Camera Components

A typical camera is consisted of several different components (see Fig. 1). As pointed out in [2], the core of every digital camera is the imaging sensor. The image sensor (typically, CCD or CMOS) is consisted on small elements called pixels that collect photons and covert them into voltages that are subsequently sampled to a digital signal in an A/D converter. Generally, before the light from the scene which is being photographed reaches the sensor it also passes through the camera lenses, an antialiasing (blurring) filter, and then through a color filter array (CFA).

The CFA is a mosaic of tiny color filters placed over the pixel of an image sensor to capture color information. Color filters are needed because typical consumer cameras only have one sensor which cannot separate color information. The color filters filter the light by wavelength range, such that the separate filtered intensities include information about the color of light. Most commonly, Bayer color filter is used. Here, each pixel captures intensity of one of the red, green, or blue color information. This output is further interpolated (demosaicked) using color interpolation algorithms to obtain all three basic color channels for each pixel.

The resulting signal is then further processed using color correction and white balance adjustment. Additional processing includes gamma correction to adjust for the linear response of the imaging sensor, noise reduction, and filtering operations to visually enhance the final image. Finally, the digital image might be compressed stored and stored in a specific image format like JPEG.

What is important in sense of forensic analyzes of digital images is that different

components of camera leave different kind of artifacts or fingerprints useful for integrity verification of photos or ballistics analysis. Typically, fingerprints left by CFA, post processing, and compression parts are in common for cameras of same make and model. In other words, assuming that we know their value and behavior for a particular camera make and model and based on the fact that digital image editing (e.g., photoshopping) change these values (fingerprints), they can be employed for verification of the originality of digital images .

On the other hand, each camera has its own unique sensor consisted on millions of pixels each of unique properties. Thus, if we are able to find such an information brought into image by the sensor and which will remain stable and present in all images captured by that sensor and cannot be fount in no image captured by any other sensor, then we can call it fingerprint of that sensor or camera. Such a fingerprint can be employed to link digital images to particular digital cameras which captured them.

2.1 Sensor as a Camera Fingerprint

Image sensors suffer from several fundamental and technology related imperfections resulting in their performance limitations and noise. As pointed out in [2], if we take a picture of an absolutely evenly lit scene, the resulting digital image will still exhibit small changes in intensity among individual pixels which is partly because of pattern noise, readout noise or shot noise.

While readout noise or shot noise are random components, the pattern noise is deterministic and remain approximately the same if multiple pictures of the same scene are taken. As a result, pattern noise can be the fingerprint of sensors which we are searching for.

Pattern Noise (PN) is consisted of two components called Fixed Pattern Noise (FPN) and photo response nonuniformity (PRNU). FPN is independent of pixel signal, it is an additive noise, and some high-end consumer cameras can suppress it. The FPN also depends on exposure and temperature.

PRNU is formed by varying pixel dimensions and inhomogeneities in silicon resulting in pixel output variations. It is a multiplicative noise. Moreover, it is not dependent on temperature and seems to be stable over time.

The values of PRNU noise increases with the signal level (it is more visible in pixels showing light scenes). In other words, in very dark areas PRNU noise is suppressed. Moreover, PRNU is not present in completely saturated areas of an image. Thus, such images should be ignored when searching for PRNU noise.

Despite the fact that there are not a lot of studies analyzing the PRNU noise in deeper details (probably due to physical limitations and no significant demand for it so far), it can be shown that it has dominant presence in the pattern noise component. This made possible Fridrich et al. [3, 1] to employ it in order to identify source cameras. In other words, PRNU noise is employed as the fingerprint of camera sensors.

3 Motivation

Generally, it can be claimed that state-of-the-art source identification methods are mostly based on methods proposed by Jessica Fridrich et al. (e.g., [3, 1]). There have been published some additional papers by other authors (e.g., [4, **?**]) aiming to improve accuracy of results. Generally, they bring modifications to the original paper of Jessica Fridrich et al. [3, 1] based on some theoretical or empirical findings. Nonetheless, the key concept of measuring sensor's fingerprint remained unchanged.

Nonetheless, having available a larger set of cameras of same and different models and a large set of ground-truth digital images captured by these devices, one can simply run an experiment to measure the effectiveness and fragileness of existing methods. By performing such an experiment, it is quite easy to notice that state-of-the-art source identification methods suffer of a number of essential non–perfections.

Below we discuss three important drawbacks specifically caused by optical zoom, JPEG, and embedded software in cameras.

3.0.1 Impact of optical zoom

When applying typical PRNU-based camera identification methods (e.g., [3, 1]) on digital images acquired by cameras having available rich optical zoom possibility then they typically fail. Let us demonstrate the problem with by carrying out a simple experiment using Fujifilm FinePix S100fs camera. The focal length of this camera can be changed from 28 mm to 400 mm. We captured 50 images of blue skye for each of the following focal lengths $Z \in \{28, 50, 100, 200, 400\}$ and used them to calculate camera sensor's fingerprint using the algorithm pointed out in [1]. In other words, 5 different fingerprints of the same camera have been obtained. Moreover, we took 5 images of a natural scene for each of mentioned focal lengths to carry a basic source identification task.

Figure 2 demonstrates results of 25 test images and 5 fingerprints. First image shown in Figure 2 demonstrates results of testing test images with sensor fingerprint of Fujifilm FinePix S100fs obtained by photos captured with focal length of 28 mm. Five test images captured by the same focal length exhibit high correlations (in other words, source camera has been found correctly). Nonetheless, all other test images captured by the same camera but differen focal lengths exhibited very low correlations (in other words, source camera has not been identified). Second image shown in Figure 2 shows result of testing test images with sensor fingerprint obtained by photos captured with focal length of 50 mm. Five test images captured by the same the focal length exhibit high correlation. Again, all other test images failed. Other images shown in Figure 2 uncovers the same problem under scenarios of using other focal lengths in Z.

We also carried out the same experiment with other cameras such as Nikon Coolpix L23, Canon PowerShot A495, Pentax Optio P80, etc. with very similar results. Apparently, this is a serious drawback as it is very difficult to create a stable fingerprint for a cameras having rich focal length. To cover all focal lengths, one should create one fingerprint per each available focal length, [5], which is very time consuming and almost impossible in real-life applications.

The question is why this problem happens? The reason behind this is, so called, vignetting which causes a change of PRNU values at different zoom levels. There are

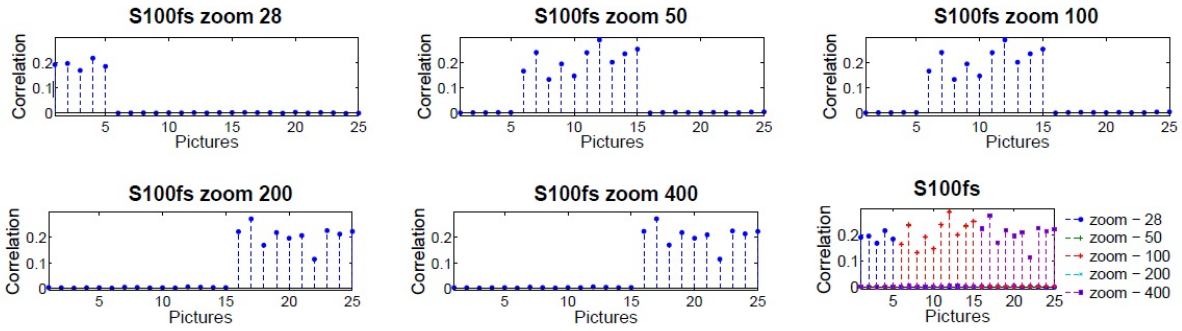


Figure 2: Problem of camera source identification caused by optical zoom. Fujifilm FinePix S100fs is a camera having different possibilities of focal lengths. Shown results demonstrate that correctness of source identification test is dependent on particular sensor reference images and corresponding focal length.

several types of vignetting such as mechanical, optical, natural or pixel vignetting [6]. Some types of vignetting can be completely covered by lens settings (using special filters), but most digital cameras use built-in image processing to compensate with vignetting when converting raw sensor data to standard image formats such as JPEG or TIFF. Typically, vignetting is stronger at the non-central parts of the photo.

3.0.2 Impact of embedded camera software

Assume we have 100 pieces of different iPhone 3 devices. Moreover, we have a digital image captured by one of these iPhones and our aim is to identify the particular source device. In other words, we need to have such a fingerprint of each device that distinguish it uniquely and eliminate features in common for these devices.

On the other hand, there is an embedded software in digital consumer cameras which perform operations like color filter array (CFA) interpolation, white balancing, gamma correction, color enhancement, interpolation (digital zoom), etc. Because of the fact that this embedded software is typically in common in cameras/smartphones of the same model, it brings into digital images of cameras of same model very similar changes. This is a serious problem which occurs in higher rate of false positives when having a higher number of source imaging devices of same model under investigation.

3.0.3 Impact of heavy JPEG

Let us assume the example with iPhone 3 mentioned before. Assume that this digital camera produces heavily compressed JPEG images. As it is known, highly JPEG compressed images exhibit blocking artifacts. Figure 3 provides a simple example of blocking artifact. Here, first 8 rows and 9 columns of the same photo compressed with different JPEG qualities is shown. As apparent, absolute difference between boundary pixels (pixels at 8th and 9th column) of (a) is 0. Same for (b) is 6. and for (c) is 14. These JPEG blocking artifacts is another change brought into the image by the embedded camera software and in common within the same model of cameras. In other words, this is an-

(a)

128	124	125	125	127	125	127	128	128
109	120	124	125	127	126	127	128	128
127	121	139	128	120	125	127	128	128
121	125	128	122	122	126	127	128	128
124	128	122	134	124	127	126	128	128
124	126	121	128	122	125	127	128	128
110	128	128	128	122	124	127	128	128
120	120	120	120	126	127	127	128	128

(b)

127	125	124	124	127	126	127	127	128
108	118	126	123	130	127	122	130	128
128	125	136	126	119	128	130	125	128
119	124	128	125	119	125	126	131	128
126	128	122	135	126	126	123	128	128
123	125	118	131	119	126	131	126	128
110	131	128	126	122	124	124	129	128
120	119	119	122	126	127	127	128	128

(c)

119	122	124	125	125	126	129	132	128
120	123	125	126	125	125	128	130	128
122	124	126	126	125	124	127	129	128
123	125	127	126	124	124	125	128	128
123	125	127	126	124	124	125	128	128
122	124	126	126	125	124	127	129	128
120	123	125	126	125	125	128	130	128
119	122	124	125	125	126	129	132	128

Figure 3: JPEG blocking artifact. (a) shows pixels of rows 1 to 8 and columns 1 to 9 of a RAW digital image. In (b) its JPEG 95% version is shown. In (c) JPEG 65% version of (a) is shown.

other source of false positive results when linking a photo to larger set of possible source cameras of same model. Moreover, this is a quite common problem occurred in real-life applications (for example, when inspecting Facebook photos or Youtube videos).

To understand why blocking artifacts occur, we need to understand how JPEG algorithm does work. Although JPEG file can be encoded in various ways, the most common algorithm is the following one.

Typically, the image is first converted from RGB to YCbCr, consisting of one luminance component (Y), and two chrominance components (Cb and Cr). Mostly, the resolution of the chroma components are reduced, usually by a factor of two. Then each component is split into adjacent blocks of 8×8 pixels. Block values are shifted from unsigned to signed integers. Each block of each of the Y, Cb, and Cr components undergoes a discrete cosine transform (DCT). Let $f(x, y)$ denote a pixel (x, y) of an 8×8 block. Its DCT is:

$$F(u,v) = \frac{1}{4} C(u) C(v) \sum_{x=0}^{7} \sum_{y=0}^{7} f(x,y) \cos \frac{(2x+1)u\pi}{16} \cos \frac{(2y+1)v\pi}{16}, \tag{1}$$

where

$$\begin{aligned} &u, v \in \{0 \cdots 7\}; \\ &C(u), C(v) = 1/\sqrt{2} \quad \text{for} \quad u, v = 0; \\ &C(u), C(v) = 1 \qquad \text{otherwise.} \end{aligned} \tag{2}$$

In the next step, all 64 $F(u, v)$ coefficients are quantized. This is done by simply dividing each component in the frequency domain by a constant for that component, and

then rounding to the nearest integer. More formally, the quantization step is given by a 64-element quantization table (QT):

$$F^{QT}(u,v) = \text{round}\left(\frac{F(u,v)}{QT(u,v)}\right), \quad u,v \in \{0\cdots 7\}$$

where $QT(u,v)$ defines the quantization step for each DCT frequency u and v. Commonly, there is one QT for Y and another, single QT for both Cb and Cr. In the final step, entropy coding is carried out. This part is, typically, performed by employing run-length encoding (RLE) and Huffman coding.

The JPEG decompression works in the opposite order: entropy decoding followed by de-quantization step, inverse discrete cosine transform, etc.

Now, it is apparent that it is the quantization step in conjunction with splitting the image into block 8×8 that bring into the decoded photo shown blocking artifacts.

4 Modeling and Extracting PRNU

Let us model the image acquisition process in the following way:

$$I_{i,j} = I^o_{i,j} + I^o_{i,j} \cdot \Gamma_{i,j} + \Upsilon_{i,j} \tag{3}$$

Here, $I_{i,j}$ denotes the image pixel at position (i,j) produced by the camera, $I^o_{i,j}$ denotes the noise-free image (perfect image of the scene), $\Gamma_{i,j}$ denotes PRNU noise and $\Upsilon_{i,j}$ stands for all additive or negligible noise components.

Following the approach proposed by [3, 1], the PRNU component is estimated in the following way. For a given camera, PRNU noise is estimated by averaging multiple images I_k, $k = 1, \cdots, N$ captured by this camera. This process is sped up by suppressing the scene content from the image prior to averaging. This is achieved by using a de-noising filter F and averaging the noise residuals I^d_k instead. In other words, PRNU of the camera C is computed by:

$$C_{PRNU} = \frac{1}{N}\sum_{k=1}^{N} I_k - I^d_k \tag{4}$$

Alternatively, maximum likelihood estimation (MLE) instead of simple averaging is employed.

In our work, we focus on multiplicative nature of PRNU component and analytically derive its estimation. Specifically, denoting the digital image captured by the camera by I, and the corresponding noise-free perfect image of the scene by I_0, then the fingerprint of the camera can be calculated in the following way.

Given Eq. 3, let us divide both sides of this equation by I^o and introduce a natural logarithm operator:

$$\frac{I_{i,j}}{I^o_{i,j}} = \frac{I^o_{i,j} + I^o_{i,j} \cdot \Gamma_{i,j} + \Upsilon_{i,j}}{I^o_{i,j}}$$

$$\ln(I_{i,j}) - \ln(I^o_{i,j}) = ln(1 + \Gamma_{i,j} + \frac{\Upsilon_{i,j}}{I^o_{i,j}}) \tag{5}$$

Having derived Eq. 5 and knowing that Taylor series expansions of the logarithmic function $\ln(1+x)$ is

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} \cdots$$

we can simply derive the following:

$$\ln(I_{i,j}) - \ln(I^o_{i,j}) = \Gamma_{i,j} + \frac{\Upsilon_{i,j}}{I^o_{i,j}} + \cdots$$

For the sake of simplicity, in the rest of this paper we omit pixel indexes (i,j) in our denotations. Now, having available N digital images captured by the same camera and considering the deterministic behavior of the PRNU noise component of its sensor, Γ_{sensor}, we can derive the following:

$$\frac{1}{N}\sum_{k=1}^{N} \ln(I_k) - \ln(I^o_k) = \Gamma_{sensor} + \frac{1}{N}\sum_{k=1}^{N} \frac{\Upsilon_k}{I^o_k} + \cdots$$

Assuming that Υ is a zero-mean noise component, we can conclude that

$$\lim_{n\to\infty} \frac{1}{N}\sum_{k=1}^{N} \frac{\Upsilon_k}{I^o_k} = 0$$

Ignoring higher order terms of Taylor expansion we can state that PRNU noise component of the sensor under analysis, Γ_{sensor}, can be estimated in the following way:

$$\Gamma_{sensor} = \frac{1}{N}\sum_{k=1}^{N} \ln(I_k) - \ln(I^o_k) \tag{6}$$

So, considering Γ_{sensor} as fingerprint of the camera's sensor based on PRNU noise, using Eq. 6 we can extract it from a set of image or even from one image. But, it is apparent that as $N \to \infty$ the more accurate estimate of Γ_{sensor} we get. As stated before, Eq. 6 we use the multiplication nature of PRNU component (recall that $ln(a) - ln(b) = ln(\frac{a}{b})$).

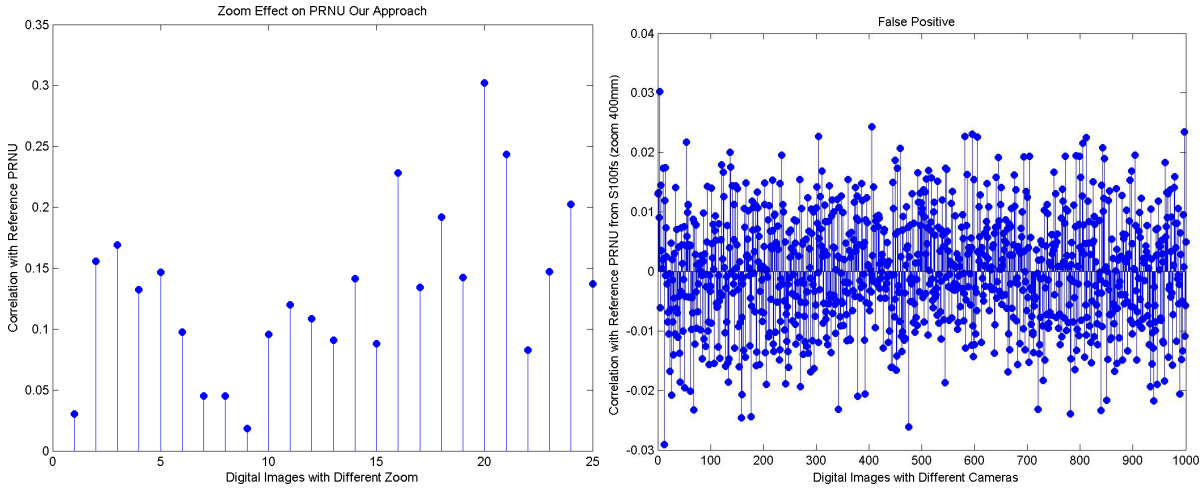
Now, using simple a correlation we can measure similarity of different fingerprints. For example, having available two different sensor fingerprints Γ_{s_1} and Γ_{s_1}, we measure their similarity by employing a normalized correlation:

$$corr(\Gamma_{s_1}, \Gamma_{s_2}) = \frac{(\Gamma_{s_1} - \overline{\Gamma_{s_1}}) \odot (\Gamma_{s_2} - \overline{\Gamma_{s_2}})}{(\|\Gamma_{s_1} - \overline{\Gamma_{s_1}}\|) \cdot (\|\Gamma_{s_2} - \overline{\Gamma_{s_2}}\|)} \tag{7}$$

where $\overline{X}$ denotes mean of the vector X, $\odot$ stands for dot product of vectors defined as $X \odot Y = \sum_{k=1}^{N} X(k)X(k)$ and $\|X\|$ denotes L_2 norm of X defined as $\|X\| = \sqrt{X \odot X}$.

It has been shown in [3] that a good way of approximating I^0 is by de-noising I and compute the residual of these two images:

$$I^0 \approx I - I^d \tag{8}$$



(a) Same images as in the Figure 2 (b) Images from others 10 cameras.

Figure 4: Problem of zooming camera by our approach.

Here, I^d denotes the de-noise image. While some studies were carried out about the specific choice and effectiveness of de-noising filters (e.g., [4]), our experiments uncovered that although a proper de-noising filter improves results of source identification, this part usually does not play the most critical part in receiving accurate results. It happens that in some cases (e.g., based on spatial distribution of the image) some filters work better and some a bit worse.

5 Experiments

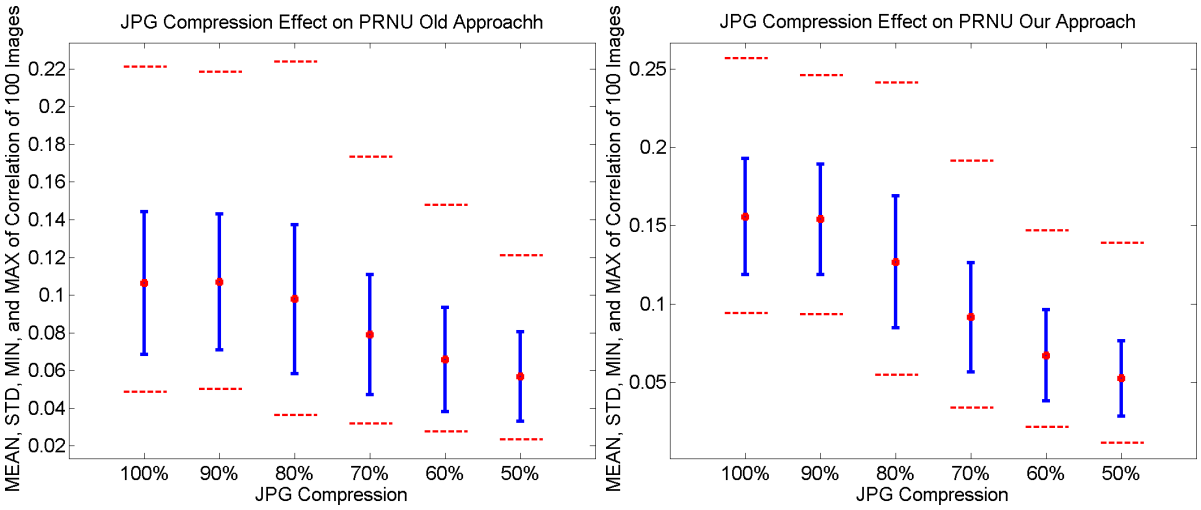
In this section we focus on testing the ultra-zoom camera Fujifilm FinePix S100fs for its possibility of manual zooming and wide range of zoom. While we got the similar results for other cameras.

5.1 Effect of optical zoom

We described in section 3.0.2, how strong influence has the optical zoom on the resulting PRNU. Therefore, we took the same 25 images as in Figure 2 and calculated their PRNU using our approach pointed out in section 4. Then we compared this PRNU with the camera sensor's fingerprint obtained by set of 50 photos captured with maximum focal length of 400 mm. As shown in Figure 4a almost all testing images captured by the different focal length exhibit higher correlation.

5.2 Effect of JPG compression

We captured 100 photos of different scene and store them with best quality of JPG compression (mark them as 100%). Then we resaved them with different JPG quality from 90% to 50%. The Figure 5 shows the results with state-of-the-art method, 5a, and our method 5b .



(a) Old Approach (b) Our Approach

Figure 5: Problem of camera source identification caused by jpg compression.

6 Conclusion

A new approach of counting Photo-Response Non-Uniformity Noise was developed. The standard method proposed by [3] assumes PRNU as the additive component of noise. We focused on multiplicative nature of PRNU component and analytically derived its estimation. In the experimental section, we show the resulting correlations for the jpg compression and zooming. Although, we need more tests with different camera settings for better understanding of influences on PRNU, we took the next step to more accurate identification of individual cameras in practice.

References

[1] 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.

[2] J. Lukáš, J. Fridrich, and M. Goljan, "Detecting digital image forgeries using sensor pattern noise," in *In Proceedings of the SPIE*. West, 2006, pp. 2006.

[3] 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, June 2006, pp. 205–214.

[4] 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.

[5] A. Novozámský, "Source Camera Identification Based on PRNU Invariant to Zoom", in *Doktorandské dny 2011 sborník workshopu doktorandů FJFI oboru Matematické inženýrství*, Praha, CZ, November 2011.

[6] D. B. Goldman and J.-H. Chen, "Vignette and exposure calibration and compensation," in *The 10th IEEE International Conference on Computer Vision*, Oct. 2005, pp. 899–906.

FPGA Based Data Acquisition System for COMPASS Experiment*

Josef Nový

1st 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 the present data acquisition system (DAQ) of the COMPASS experiment at CERN and presents development of a new DAQ. The new DAQ must preserve present data format and be able to communicate with FPGA cards. Parts of the new DAQ are based on state machines and they are implemented in C++ with usage of the QT framework, the DIM library, and the IPBus technology. Prototype of the system is prepared and communication through DIM between parts was tested. An implementation of the IPBus technology was prepared and tested. The new DAQ proved to be able to fulfill requirements. Full version of this contribution is available at http://arxiv.org/abs/1310.1308.

Keywords: data acquisition, FPGA, DIM

Abstrakt. Tento článek se věnuje současnému systému pro sběr dat experimentu COMPASS v CERN a popisuje dosavadní vývoj nového systému. Nový systém musí zachovat stávající formát dat a dále musí být schopný komunikovat s FPGA kartami. Návrh jednotlivých částí nového systému je založen na stavových automatech. Tyto části jsou realizovány v programovacím jazyce C++ s využitím knihoven QT, DIM a IPBus. Prototyp navrhnutého systému je připraven a části určené ke komunikaci skrze DIM a IPBus byly úspěšně otestovány. Díky testům bylo prokázáno, že nový systém dokáže splnit požadavky na něj kladené. Plná verze tohoto příspěvku je dostupná na adrese http://arxiv.org/abs/1310.1308.

Klíčová slova: sběr dat, FPGA, DIM

References

[1] 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/

[2] 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

*This work has been supported by grants LA08015 and SGS11/167/OHK4/3T/14

[3] 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

[4] T. Anticic, et al. (the ALICE collaboration): *ALICE DAQ and ECS User's Guide.* CERN, ALICE internal note, ALICE-INT-2005-015, 2005.

[5] M. Bodlák, V. Jarý, J. Nový: *Software for the new COMPASS data acquisition system.* In: COMPASS collaboration meeting, Geneva, Switzerland, 18 November 2011

[6] 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

[7] 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

[8] M. Bodlák: *COMPASS DAQ Database Architecture and Support Utilities* Prague, 2012, Master thesis, Czech Technical University in Prague

[9] J. Nový: *COMPASS DAQ - Basic Control System* Prague, 2012, Master thesis, Czech Technical University in Prague

[10] T. Anticic, et al. (ALICE DAQ Project): *ALICE DAQ and ECS User's Guide* CERN, EDMS 616039, January 2006.

Využití metod založených na jádrových funkcích v biomedicíně*

Jiří Palek

1. ročník PGS, email: jiri.palek@gmail.com
Katedra softwarového inženýrství
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Jaromír Kukal, Katedra softwarového inženýrství, Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. Kernel-based methods represent widely applicable branch of data mining algorithms. This paper deals with usage of kernel-based principal component analysis (PCA) in diagnostics of Alzheimer's disease from SPECT images. In general, these images are high-dimensional data which are not easy to classify. In order to solve this task, kernel based principal component analysis was used to reduce the dimensionality of the images, and quadratic discriminant analysis (QDA) was then used for classification.

Keywords: Alzheimer's disease, diagnostics, Kernel PCA, whitening, QDA, leave-one-out cross validation, classification, MATLAB, object oriented programming

Abstrakt. Metody založené na jádrových funkcích představují široce využitelný přístup k dolování znalostí z dat. V této práci je využita jádrová varianta analýzy hlavních komponent (PCA) k diagnostice Alzheimerovy choroby ze SPECT snímků. Tyto snímky představují vysokodimenzionální data, která se obecně obtížně klasifikují. Problém dimenze obrázků byl řešen rozdělením jejich analýzy na dvě části. V první bylo použito jádrové analýzy hlavních komponent ke snížení dimenze úlohy a v druhé části byla provedena klasifikace pomocí kvadratické diskriminační analýzy (QDA).

Klíčová slova: Alzheimerova choroba, diagnostika, jádrová PCA, whitening, QDA, leave-one-out křížová validace, klasifikace, MATLAB, objektově orientované programování

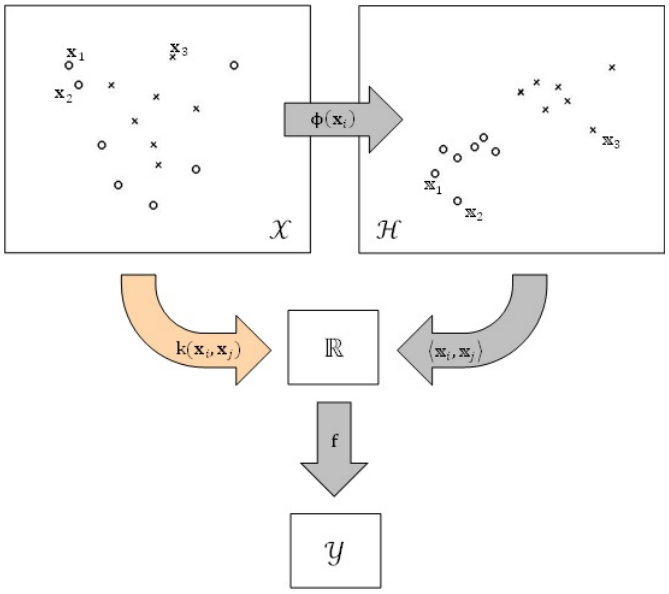
1 Úvod

Dolování dat je bouřlivě se rozvíjející vědecká disciplína, která stojí na pomezí tří oblastí; matematiky, informatiky a aplikačně zajímavé oblasti, jejíž data zpracovává.

K samotnému dolování dat existuje celá řada různých přístupů, které si velmi často berou inspiraci z matematiky. V tomto příspěvku je stěžejní přístup využívající teorii jádrových funkcí [3] [5]. Model jádrové analýzy hlavních komponent (PCA) je v článku využit k ilustraci přístupu, který jádrové metody k dolování dat využívají.

Výše uvedený matematický model je v praktické části práce použit pro analýzu 3D snímků mozků s cílem vytvořit binární klasifikátor Alzheimerovy choroby. Pro tento účel je jádrová PCA využita pro předzpracování dat pro kvadratickou diskriminační analýzu (QDA) [2].

*Tato práce byla podpořena grantem SGS11/165/OHK4/3T/14 ČVUT v Praze.



Obrázek 1: Využití jádrových funkcí k modelování

Samotné výpočty jsou realizovány v prostředí MATLAB v rámci vlastní objektově orientované implementace vycházející z [1].

2 Využité matematické modely

Praktickým cílem práce je analyzovat 3D snímky mozků a vytvořit binární klasifikátor rozlišující zdravé a nemocné lidi.

Vzhledem k povaze dat (viz část 4) byla analýza rozdělena do dvou částí; v první byla data transformována pomocí jádrové PCA, v druhé byla na tranformovaná data aplikována kvadratická diskriminační analýza.

2.1 Jádrový přístup k modelům

Mějme soubor pozorování a modelované vlastnosti $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, kde $\mathbf{x}_i \in \mathcal{X}$ jsou pozorování, $y_i \in \mathcal{Y}$ je modelovaná vlastnost a $n \in \mathbb{N}$. Klasickým přístupem je následné využití vztahů mezi pozorováními v prostoru $\mathcal{X}$ a skrze tyto vztahy modelovat požadovanou vlastnost $\mathcal{Y}$.

Myšlenka, kterou využívají jádrové funkce, je vložit mezi prostor $\mathcal{X}$ a $\mathcal{Y}$ další prostor; označme jej $\mathcal{H}$. Prostor $\mathcal{H}$ je zaveden jako Hilbertův prostor a obraz pozorování $\mathbf{x}_i$ v prostoru $\mathcal{H}$ dostaneme pomocí zobrazení Φ jako $\mathbb{x}_i = \Phi(\mathbf{x}_i)$. Takto získáme nový soubor pozorování $\{(\mathbb{x}_1, y_1), ..., (\mathbb{x}_n, y_n)\}$ v prostoru $\mathcal{H}$. Nyní, v prostoru $\mathcal{H}$, budeme skrze vzájemné vztahy mezi pozorování $\mathbb{x}_i$ modelovat prostor $\mathcal{Y}$. V kontextu teorie jádrových funkcí jsou vzájemné vztahy modelovány pomocí vzájemných vzdáleností vyjádřených pomocí skalárního součinu $\langle \mathbb{x}_i, \mathbb{x}_j \rangle$. Celý tento postup je znázorněn schematicky na obrázku 1 pomocí tmavé sekvence šipek.

Protože je prostor $\mathcal{H}$ volen buď jako vysokodimenzionální prostor, nebo dokonce spočetnědimenzionální prostor, je výše uvedený přístup technicky obtížně realizovatelný, v případě spočetnědimenzionálního prostoru dokonce nerealizovatelný. Tento zásadní nedostatek je odstraněn tím, že prostor $\mathcal{H}$ je konstruován tak, aby bylo možné skalární součin $\langle \mathbb{x}_i, \mathbb{x}_j \rangle$ počítat přímo z původních pozorování $\mathbf{x}_i$ pomocí tzv. jádrové funkce k jako $\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbb{x}_i, \mathbb{x}_j \rangle$. Tento postup je znázorněn na obrázku 1 pomocí světlé šipky. Vztahy mezi objekty můžeme shrnout následujícím způsobem

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbb{x}_i, \mathbb{x}_j \rangle = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle.$$

Přímým důsledkem předchozí formulace je skutečnost, že jednotlivé modely pro dolování dat založené na jádrových funkcích přebírají pozorování ve formě tzv. jádrové matice $\mathbb{K}$, která je definovaná následujícím způsobem

$$(\mathbb{K})_{ij} = \mathrm{k}(\mathbf{x}_i, \mathbf{x}_j), \forall i, j \in \{1, ..., n\}.$$

2.2 Příklady jádrových funkcí

Pro analýzu dat byla použita následující jádra

- nehomogenní polynomiální jádro s posunem $c = 1$ a parametrem d

$$\mathrm{k}(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^d,$$

- exponenciální jádro s parametrem σ

$$\mathrm{k}(\mathbf{x}, \mathbf{y}) = \exp(-\tfrac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}),$$

- gaussovské jádro s parametrem σ

$$\mathrm{k}(\mathbf{x}, \mathbf{y}) = \exp(-\tfrac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}).$$

2.3 Klasická PCA

Analýza hlavních komponent (PCA) [6] představuje účinnou techniku pro získávání struktur z vícedimenzionálních souborů dat. Z matematického hlediska se jedná o takovou ortogonální transformaci souřadného systému, která minimalizuje korelaci mezi proměnnými.

Nechť je dán soubor pozorování $\mathbb{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)', \mathbf{x}_i \in \mathcal{X} = \mathbb{R}^{p,1}$, který je vycentrovaný, tedy splňuje

$$\sum_{i=1}^{n} \mathbf{x}_i = \mathbb{O}_{p,1}, \tag{1}$$

potom hledání komponent představuje problém nalezení vlastních čísel λ a vlastních vektorů $\mathbf{v}$ kovarianční matice

$$\mathbb{C} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'. \tag{2}$$

Transformační matice $\mathbb{A}$ přechodu od původních souřadnic $\mathbb{X}$ k novým souřadnicím $\mathbb{Z}$ je pak dána jako $\mathbb{A} = (\mathbf{v}_1, ..., \mathbf{v}_p)$, kde pro vlastní čísla λ_i příslušející k vlastním vektorům $\mathbf{v}_i$ platí $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$.

Tento koncept lze samozřejmě dále modifikovat. Možné je na vstupu použít korelační matice místo matice kovarianční. Výstup je zase možné sférizovat pomocí následující transformační matice $\mathbb{W} = (\mathbf{v}_1/\sqrt{\lambda_1}, ..., \mathbf{v}_p/\sqrt{\lambda_p})$.

2.4 Jádrová PCA

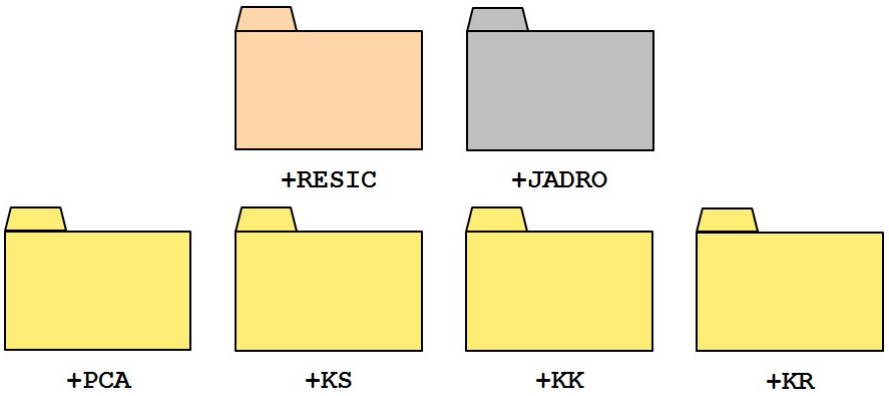
Základní myšlenka využití jádrových funkcí je uvedena v části 2.1. Rozšířit klasickou analýzu hlavních komponent tedy znamená provést ji v prostoru $\mathcal{H}$ [3] [5].

Lze ukázat [5], že provedení analýzy hlavních komponent v prostoru $\mathcal{H}$ odpovídá hledání vlastních čísel jádrové matice $\mathbb{K}$. Stručně řečeno, postup je stejný jako v předchozí kapitole, jen se místo matice (2) provádí s jádrovou maticí $\mathbb{K}$.

Před klasickou PCA se vstupní data centralizují pomocí vzorce (1). V prostoru $\mathcal{H}$ reprezentuje podobnou úpravu tzv. whitening matice $\mathbb{K}$, který je definován jako

$$\tilde{\mathbb{K}} = \mathbb{K} - \tfrac{1}{n}\mathbb{I}_{n,n}\mathbb{K} - \tfrac{1}{n}\mathbb{K}\mathbb{I}_{n,n} + \tfrac{1}{n^2}\mathbb{I}_{n,n}\mathbb{K}\mathbb{I}_{n,n}$$

Hledání vlastních čísel se potom provádí s maticí $\tilde{\mathbb{K}}$ místo matice $\mathbb{K}$.



Obrázek 2: Přehled implementovaných balíčků

2.5 Kvadratická diskriminační analýza

Kvadratická diskriminační analýza (QDA) [2] představuje model klasifikace s učením, který aproximuje data z tříd pomocí normálního rozdělení. Klasifikace nového pozorování je potom provedena tak, že se vypočítá pravděpodobnost příslušnosti ke všem třídám a pozorování se následně přisoudí do třídy s největší pravděpodobností příslušnosti.

Mějme N tříd C_i s rozděleními $f_i(\mathbf{x}), \mathbf{x} \in \mathcal{X} = \mathbb{R}^{p,1}, i \in \{1, ..., N\}$. Úkolem je vytvořit dekompozici prostoru $\mathcal{X}$ na N množin A_i tak, aby platilo

1. $\cup_{i=1}^{N} A_i = \mathcal{X}$,

2. $\mathbf{x} \in C_i \Leftrightarrow \mathbf{x} \in A_i$.

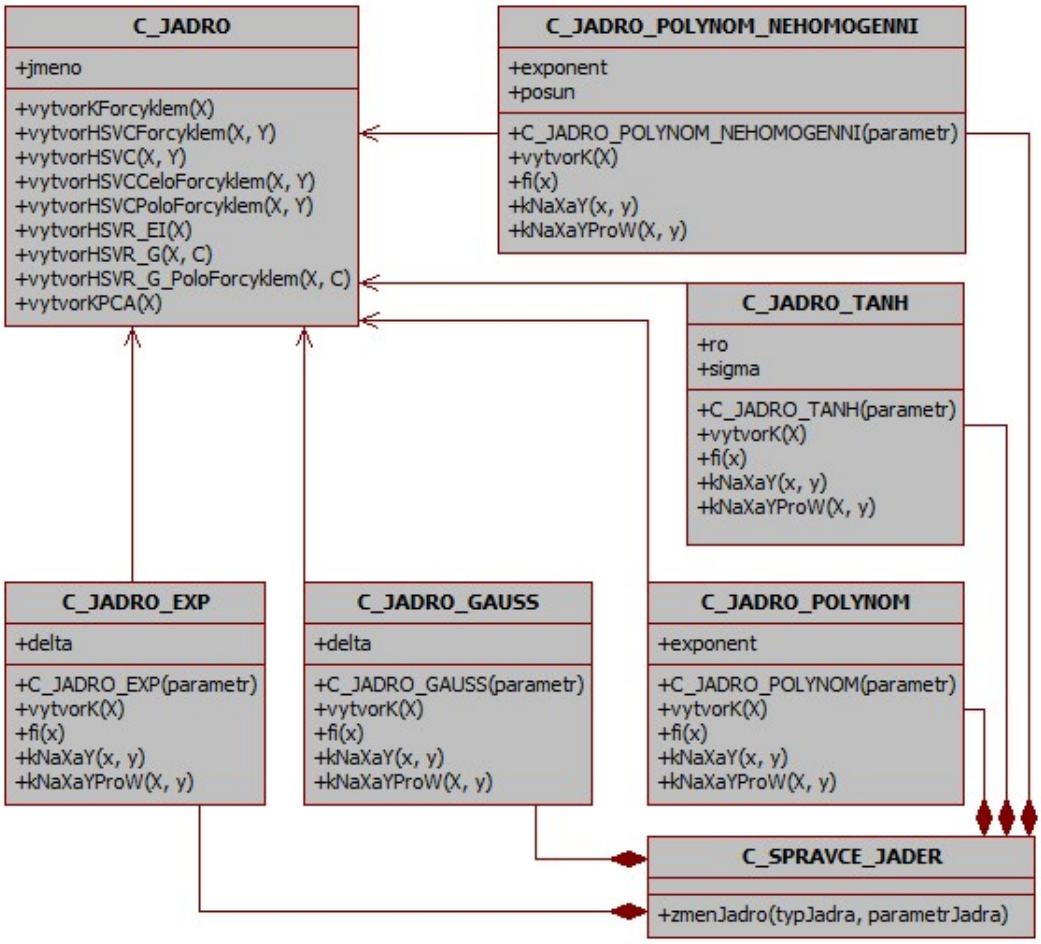
Nalezení optimálního řešení představuje nalezení minima funkcionálu

$$\mathbf{L} = \sum_{i=1}^{N} \int_{A_i} \sum_{j=1}^{N} \pi_j f_j(\mathbf{x}), \tag{3}$$

kde π_i je apriorní pravděpodobnost třídy C_i (například rovnoměrné rozdělení do tříd). V [2] je ukázáno, že klasifikační pravidlo 1. spolu s funkcionálem (3) lze převést na následující klasifikační pravidlo

$$\mathbf{x} \in C_t \Leftrightarrow \pi_t f_t(\mathbf{x}) > \pi_j f_j(\mathbf{x}), \forall j \neq t.$$

V QDA se používá normání rozdělení pravděpodobnosti $f_j \sim N(\mu_j, \Sigma_j)$.

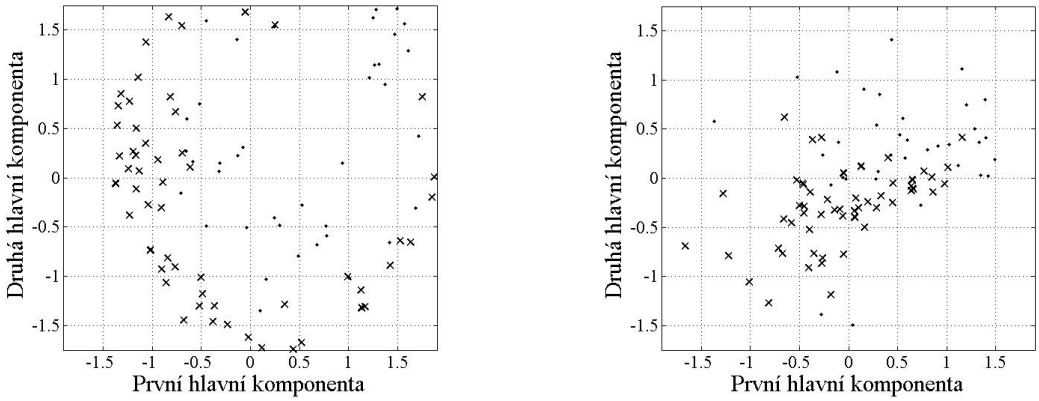


Obrázek 3: Schéma balíčku +JADRO

3 Implementace

Implementace se skládá s šesti balíčků. Jejich přehled je na obrázku 2. Struktura balíčků, popis jednotlivých metod a celkové použití je uvedeno v [1]. Oproti původní verzi z [1] však byl vylepšen balíček pracující s jádry. Jeho současná objektová struktura je uvedena na obrázku 3.

Současná struktura výrazně usnadňuje práci s jádrovými funkcemi díky třídě C_SPRAVCE_JADER, která zastřešuje a jednotně zastupuje chování jednotlivých jader. Díky uvedené koncepci balíčku je navíc možné velmi snadno rozšiřovat stávající portfolio jádrových funkcí.



Obrázek 4: Járová PCA: (vlevo) exponenciální jádro s parametrem $\sigma = 3900$, (vpravo) polynomiální jádro stupně tři

4 Analýza dat

Praktickým cílem bylo analyzovat 3D snímky mozků a vytvořit binární klasifikátor rozlišující zdravé lidi a lidi s Alzheimerovou chorobou [4].

4.1 Popis dat

Předmětem analýzy byly 3D SPECT snímky mozků lidí. Jednotlivé snímky jsou reprezentovány maticí o rozměrech 79x95x69. Snímky byly v rámci předzpracování normalizovány z hlediska intenzity.

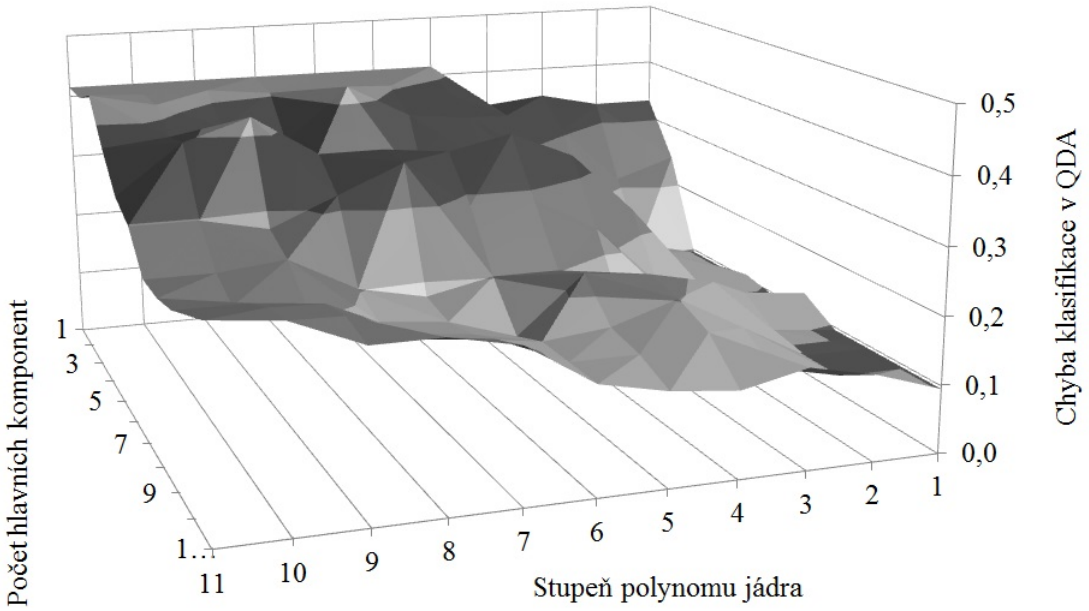
Data se skládala ze dvou skupin lidí. První část představovalo 56 snímků zdravých lidí. Zdravotní stav byl zde určován jednak na základě snímků jako takových, jednak na základě sady psychologických testů. Druhá část dat obsahovala 38 lidí s diagnózou Alzheimerovy choroby.

4.2 Postup analýzy

Jak plyne z popisu snímků, snímky jako data představují vysokodimenzionální objekty. Proto byla analýza rozdělena na dvě části.

Nejprve byla na celý soubor dat aplikovaná jádrová PCA a získány nové souřadnice snímků. K tomu účelu byly použity všechny jádrové funkce z části 2.2. Výpočty byly provedeny pro širokou škálu hodnot parametrů jednotlivých jader, aby bylo možné posoudit robustnost výsledků. Příklady výstupů pro dvě komponenty jsou uvedeny na obrázku 4, kde křížky reprezentují zdravé pacienty a tečky reprezentují nemocné pacienty.

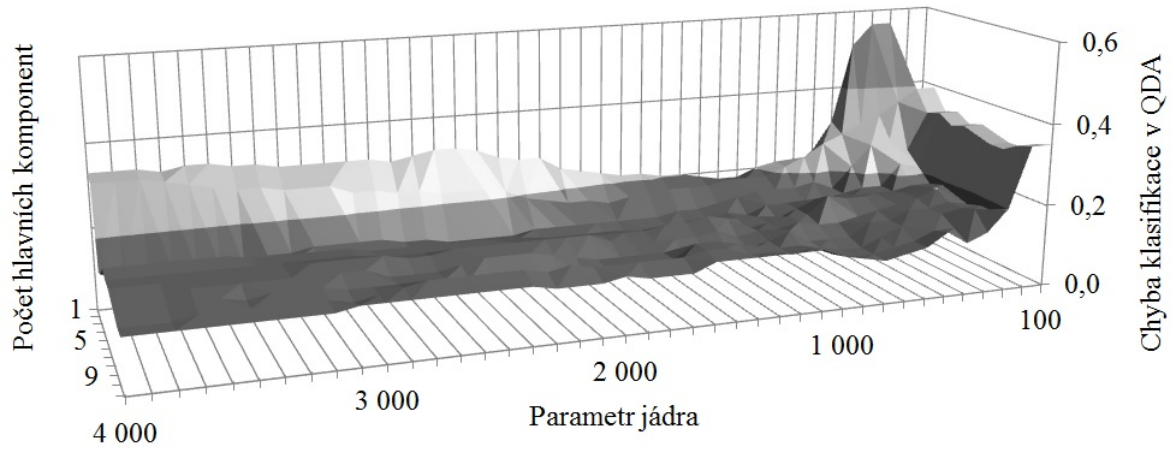
Druhým krokem bylo učení samotného klasifikačního modelu. Pro tyto účely byla použita kvadratická diskriminační analýza. QDA byla aplikována na data transformovaná všemi třemi použitými jádry. V této fázi bylo cílem zkoumat závislost výsledné chyby klasifikace na typu jádra, parametru jádra a počtu vybraných komponent. Pro měření chyby klasifikace byla použita validační metoda "leave-one-out". Výsledky jsou znázorněny na obrázcích 5, 6 a 7. Všechny tyto obrázky mají stejnou strukturu; na ose x je parametr použitého jádra, na ose y počet použitých komponent a osa z znázorňuje klasifikační chybu.



Obrázek 5: QDA s jádrovou PCA využívající polynomiální jádro

4.3 Výsledky

Kvadratická diskriminační analýza byla vybrána kvůli svému Bayesovskému základu a obecné použitelnosti. Dalším důvodem pro její volbu byla skutečnost, že z 2D a 3D výstupů analýzy hlavních komponent bylo usouzeno, že transformovaná data jsou kvadraticky separovatelná.



Obrázek 6: QDA s jádrovou PCA využívající exponenciální jádro

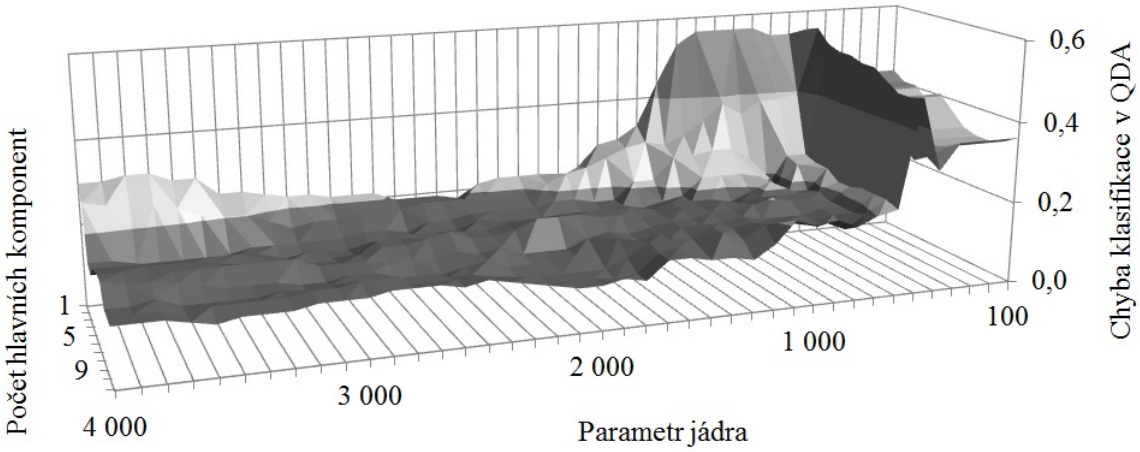
Všechny výsledky (viz obrázky 5, 6 a 7) jsou již při prvním přiblížení smysluplné, protože ukazují klesající chybu klasifikace s rostoucím počtem použitých komponent. Pokles chybovosti je nejvýraznější během prvních čtyř komponent. Navíc se u všech testovaných jader chybovost od sedmé komponenty ustaluje. Obecně lze tedy učinit závěr, že pro dosažení dobrých klasifikačních výsledků stačí přibližně osm dimenzí z jinak vysokodimenzionálního objektu.

Nejhorší výsledky s ohledem na volbu parametru vykazovalo polynomiální jádro (obrázek 5). Na druhou stranu jeho parametr je nejsnadněji interpretovatelný a proto je velmi snadné učinit závěry z provedené analýzy. Plyne z ní, že nejlepších výsledků bylo dosaženo pro polynom prvního a druhého stupně a že je potřeba vzít v úvahu alespoň tři komponenty. Potom je chybovost mezi 0,096 a 0,138 s průměrnou hodnotou 0,107.

Výsledek klasifikace s exponenciálním jádrem (obrázek 6) obsahuje dvě zajímavé oblasti. První z nich je oblast zahrnující volbu parametru $\sigma > 400$ a počet použitých komponent větší než tři. V této, z hlediska nastavení parametrů, rozsáhlé oblasti se chybovost stabilně pohybovala v rozmezí 0,096 až 0,191 s průměrnou hodnotou 0,143. Druhá zajímavá oblast je definovaná volbou parametru σ v rozmezí 600 až 1300. Zde bylo dosaženo chybovosti v rozmezí 0,181 až 0,213 s průměrnou hodnotou 0,191 již pro jednu použitou komponentu.

Výsledky pro QDA s Gaussovským jádrem (obrázek 7) jsou podobné výsledkům dosaženým v QDA s exponenciálním jádrem. První zajímavá oblast je pro volbu parametru $\sigma > 1100$ a počet komponent větší než tři. V této oblasti bylo dosaženo nejnižší chybovosti 0,096, nejvyšší 0,213 a průměrné hodnoty 0,149. V druhé oblasti bylo již pro jednu použitou komponentu dosaženo pro volbu parametru v rozmezí $600 < \sigma < 1300$ chybovosti mezi 0,181 a 0,191 s průměrnou chybovostí 0,189.

Obecně nejmenší chybovosti 0,096 bylo dosaženo s každým testovaným jádrem. Výhodou exponenciálního a Gaussovského jádra je, že dávají dobré výsledky pro širokou škálu parametrů. Navíc existuje oblast parametrů, pro kterou dávají dobré výsledky již



Obrázek 7: QDA s jádrovou PCA využívající Gaussovské jádro

pro jednu komponentu.

Literatura

[1] J. Palek. *Data Mining with MS SQL Server or with MATLAB?(Dolování dat v MS SQL Serveru nebo v MATLABu?)*. Master's Thesis, Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Prague, (2012).

[2] L. Pešková. *Discriminant Analysis (Diskriminační analýza)*. Bachelor's degree project,Bachelor's degree project, Palacky University in Olomouc, Faculty of Science, Olomouc, (2009).

[3] J. Shawe-Taylor, N. Cristianini. *Kernel Methods for Patter Analysis*. Cambridge University Press, (2004).

[4] J. Palek, J. Kukal, A. Bartoš, R. Píchová, H. Trojanová. *Kernel PCA in Alzheimer's Disease Diagnostics)*. 19th International Conference on Soft Computing, Brno, Czech Republic, June 26-28, 2013. MENDEL 2013 19th International Conference on Soft Computing, p. 349-352, (2013). R. Matoušek

[5] B. Schölkopf, A. Smola. *Learning with Kernels*. Cambridge: MIT Press, (2002).

[6] H. Řezanková, D. Húsek, V. Snášel. *Shluková analýza dat)*. Professional Publishing, Praha, (2009).

Numerical Simulation of Soil-Air Pressure*

Ondřej Pártl†

2nd 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 simulation of soil-air pressure. For the simulations, we employ a mathematical model that couples the continuity equation with the Darcy law; the problem obtained is solved numerically by means of the method of lines using the Galerkin finite element method and Runge-Kutta method. The results of our model are compared to experimental data measured in a wind tunnel. In the final part of the article, some interesting results obtained by simulating air flow in heterogeneous soil are presented.

Keywords: porous medium, soil-air pressure, Galerkin finite element method

Abstrakt. Tento příspěvek se zabývá simulací tlaku půdního vzduchu. K simulacím je používán matematický model, jenž spojuje rovnici kontinuity s Darcyho zákonem, přičemž vzniklá úloha je řešena numericky, a sice metodou přímek, s využitím Galerkinovy metody konečných prvků a Rungovy-Kuttovy metody. Výsledky našeho modelu jsou porovnávány s daty naměřenými ve větrném tunelu. V závěrečné části článku jsou rovněž prezentovány některé ze zajímavých výsledků získaných při simulaci proudění vzduchu heterogenní půdou.

Klíčová slova: porézní prostředí, tlak půdního vzduchu, Galerkinova metoda konečných prvků

1 Introduction

Flow of gases or liquids in porous medium is a part of a variety of complicated natural processes and, for this reason, it has been researched and simulated for years. In this paper, we deal with a seemingly simple phenomenon — we simulate only air flow in soil. This phenomenon proves, however, to be very complex and interesting as well.

The derivation of the mathematical model for the simulations is based on the ideas presented in [1] and [2]. We assume that the air flow occurs in dried soil (e.g., dried sand) which is represented by a bounded domain $\Omega \subset \mathbb{R}^2$, and it obeys the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = F, \tag{1}$$

*This work is partly supported by the project "Numerical Methods for Multiphase Flow and Transport in Subsurface Environmental Applications" number ME10009 of the Ministry of Education, Youth and Sports of the Czech Republic and "Advanced Supercomputing Methods for Implementation of Mathematical Models" number SGS11/161/OHK4/3T/14.

†The author would like to thank the following persons who kindly provided him experimental data: Radek Fučík, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague; Paul Schulte and Kate Smits, Center for Experimental Study of Subsurface Environmental Processes, Environmental Science and Engineering, Colorado School of Mines, Golden, Colorado, USA.

where ρ $[\frac{\mathrm{kg}}{\mathrm{m}^3}]$ is the air density, t [s] time and F $[\frac{\mathrm{kg}}{\mathrm{m}^3\cdot\mathrm{s}}]$ the source term of the air. The vector $\mathbf{u} = (u_1, u_2)^{\mathrm{T}}$ $[\frac{\mathrm{m}}{\mathrm{s}}]$ stands for the Darcy velocity

$$\mathbf{u} = -\frac{1}{\mu}\mathbf{K}\left(\nabla p - \rho\mathbf{g}\right) \tag{2}$$

of the air, where μ $[\frac{\mathrm{kg}}{\mathrm{m}\cdot\mathrm{s}}]$ is the dynamic viscosity, $\mathbf{K} = \begin{pmatrix} k_1 & k_2 \\ k_3 & k_4 \end{pmatrix}$ $[\mathrm{m}^2]$ the permeability tensor, p [Pa] pressure and $\mathbf{g} = (g_1, g_2)^{\mathrm{T}}$ $[\frac{\mathrm{m}}{\mathrm{s}^2}]$ the gravitational acceleration vector. Moreover, the pressure and density are assumed to be related by the ideal gas equation of state

$$\rho = p\frac{M}{RT}, \tag{3}$$

where M $[\frac{\mathrm{kg}}{\mathrm{mol}}]$ represents the molar weight of the air, R $[\frac{\mathrm{J}}{\mathrm{K}\cdot\mathrm{mol}}]$ the gas constant and T [K] the thermodynamic temperature.

It follows from (1)–(3) that the air flow in Ω is governed by the equation

$$\frac{\partial p}{\partial t} = -\nabla\cdot\left(-\frac{1}{\mu}p\mathbf{K}\nabla p + \frac{M}{RT\mu}p^2\mathbf{K}\mathbf{g}\right) + \frac{RT}{M}F \tag{4}$$

for the unknown pressure $p = p(x, y, t)$, where x, y are spatial variables. This problem is considered together with the initial condition

$$p(x, y, 0) = p_0(x, y), \qquad (x, y) \in \overline{\Omega}, \tag{5}$$

and the Dirichlet and Neumann boundary conditions

$$p|_{\Gamma_{\mathrm{Dir}}} = p_{\mathrm{Dir}}, \qquad -\frac{1}{\mu}p\mathbf{K}\nabla p|_{\Gamma_{\mathrm{Neu}}}\cdot\mathbf{n} = q_{\mathrm{Neu}}, \tag{6}$$

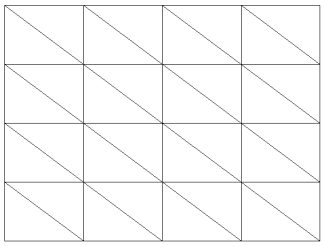
where $\Gamma_{\mathrm{Dir}}\cup\Gamma_{\mathrm{Neu}} = \partial\Omega$, $\Gamma_{\mathrm{Dir}}\cap\Gamma_{\mathrm{Neu}} = \emptyset$, and $\mathbf{n}$ denotes the unit outward normal to Γ_{Neu}.

2 Numerical Solution

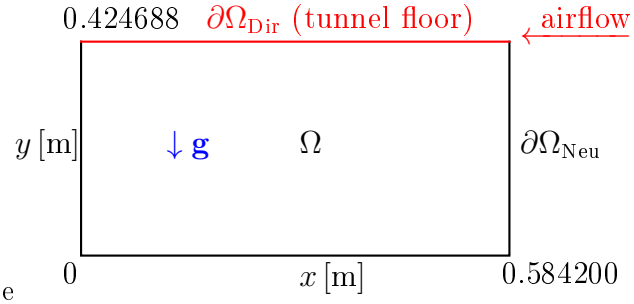
The problem (4)–(6) is solved numerically, by means of the method of lines; for the spatial discretization, the Galerkin finite element method is employed. The domain Ω is covered with the triangulation depicted in Figure 1a, and the linear Lagrange elements are used. Thus, the basis $\{\xi_j\}_{j=1}^N$ of the finite dimensional space consists of the functions which are linear on each triangle and take the value 1 at one node of the spatial mesh and vanish at the other nodes. The components of $\mathbf{K}$ are assumed to be constant on each triangle.

Hence, substituting the approximation $p = \sum_{i=1}^N p_i(t)\xi_i$, where N denotes the number of the mesh nodes, into the weak formulation, we get the following system of ordinary differential equations:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^N p_k(t)\int_\Omega \xi_i\xi_k\,\mathrm{d}x = &-\frac{1}{\mu}\sum_{i=1}^N\sum_{j=1}^N p_i(t)p_j(t)\int_\Omega \xi_i(\mathbf{K}\nabla\xi_j)\cdot(\nabla\xi_k)\,\mathrm{d}x - \int_{\partial\Gamma_{\mathrm{Neu}}} q_{\mathrm{Neu}}\mathbf{n}\cdot\,\mathrm{d}S \\ &- 2\frac{M}{RT\mu}\sum_{i=1}^N\sum_{j=1}^N p_i(t)p_j(t)\int_\Omega \xi_i\xi_k(\nabla\xi_j)\cdot(\mathbf{K}\mathbf{g})\,\mathrm{d}x\end{aligned} \tag{7}$$



(a) Spatial mesh used for the Galerkin finite element method.



(b) Description of Ω.

Figure 1: Description and triangulation of Ω.

for $k = 1, 2, 3, \ldots, N$.

Finally, the previous system is simplified by applying the method of lumped masses (see [5]) and solved by means of the Runge-Kutta-Merson method with the adaptive time step control (see [3]).

3 Simulation of Experimental Data

In order to verify our model, we simulated pressure distribution in the 12x9 sand tank mounted in the CESEP wind tunnel, and we compared the numerical results with experimental data obtained from P. Schulte.

The tank is block in shape. It is filled with Accusand #30/40, and 37 pressure ports are distributed across the north face of the tank. The complete tank is mounted in a wind tunnel so that the top side of the tank and sand are aligned with the floor of the wind tunnel, and the changes in pressure in the tank due to a moving stream of air in the tunnel are measured. This setup leads to the rectangular domain Ω depicted in Figure 1b.

The following boundary conditions are considered:

- At the top of the tank ($y = 0.424688\,\text{m}$), the Dirichlet boundary condition is considered (see Figure 1b); the values of pressure are obtained by the linear least squares minimization of the data measured by the five ports located most closely to the boundary.

- On the other three sides of Ω, the Neumann boundary condition is prescribed (see Figure 1b), specifically $q_{\text{Neu}} = 0$.

The initial condition is given by

$$p(x, y, 0) = p_{\text{ref}} - (0.424688 - y)\,\rho_{\text{air}}\,g_2,$$

where p_{ref} [Pa] denotes a reference pressure value and ρ_{air} [$\text{kg} \cdot \text{m}^{-3}$] the density of air.

The interior of the tank is considered to be homogeneous, i.e., the components of $\mathbf{K}$ are constant; the porosity of Accusand #30/40 was computed from the data in [4]. The values of all of the parameters in equations (4)–(6) are summarized in Table 1a. On each side of Ω, there are 21 mesh nodes (see Figure 1a).

parameter	value	unit
μ	$1.81 \cdot 10^{-5}$	$\frac{\mathrm{kg}}{\mathrm{m \cdot s}}$
k_1	$1.5219 \cdot 10^{-10}$	m^2
k_2	0	m^2
k_3	0	m^2
k_4	$1.5219 \cdot 10^{-10}$	m^2
M	0.02896	$\frac{\mathrm{kg}}{\mathrm{mol}}$
R	8.3144621	$\frac{\mathrm{J}}{\mathrm{K \cdot mol}}$
T	293.15	K
g_1	0	$\mathrm{m \cdot s}^{-2}$
g_2	-9.81	$\mathrm{m \cdot s}^{-2}$
F	0	$\frac{\mathrm{kg}}{\mathrm{m}^3 \cdot \mathrm{s}}$
p_{ref}	82000	Pa
ρ_{air}	1.2047	$\mathrm{kg \cdot m}^{-3}$

(a) Values used in Section 3.

parameter	value	unit
μ	$1.81 \cdot 10^{-5}$	$\frac{\mathrm{kg}}{\mathrm{m \cdot s}}$
M	0.02896	$\frac{\mathrm{kg}}{\mathrm{mol}}$
R	8.3144621	$\frac{\mathrm{J}}{\mathrm{K \cdot mol}}$
T	288.15	K
g_1	0	$\mathrm{m \cdot s}^{-2}$
g_2	-9.81	$\mathrm{m \cdot s}^{-2}$
F	0	$\frac{\mathrm{kg}}{\mathrm{m}^3 \cdot \mathrm{s}}$
p_{ref}	101325	Pa
ρ_{air}	1.2047	$\mathrm{kg \cdot m}^{-3}$

(b) Values used in Section 4.

Table 1: Values of parameters.

The numerical results are shown in Figure 2. Since the numerical solution seems to steady in approximately one or two seconds, $t > 2$ is considered, and it is not indicated. Further, in Figure 3, the results are compared to the experimental data. Clearly, they do not agree with the experimental data. Although the experimental data attain the maximum values in the left bottom corner of Ω (the tank), the numerical solution exhibits different behaviour; the pressure reaches the maximum values in the left upper corner.

It is worth mention that the numerical solution does not seem to be affected by a significant change in the values of μ, T, $\mathbf{K}$ and $\mathbf{g}$ and by refinement of the spatial mesh. Similarly, the slight change in p_{Dir} affects the solution only near Γ_{Dir} (see Figure 4).

Finally, Figure 5 shows the numerical results obtained in case that the values of pressure are prescribed for $x = 0.068263\,\mathrm{m}$ (thus, the length of the tank is adjusted) as well. Now, the results are much closer to the experimental data.

4 Simulation of Pressure in Heterogeneous Soil

Further, we simulated pressure distribution in heterogeneous soil.

In this case, the domain $\Omega = (0.0, 1.0) \times (0.0, 1.0)$ (the units are [m]) depicted in Figure 6 is considered. On the upper side of Ω ($y = 1.0\,\mathrm{m}$), the Dirichlet boundary condition is prescribed; on the other three sides, the Neumann boundary condition $q_{\mathrm{Neu}} = 0$ is considered. The initial condition is given by

$$p(x, y, 0) = p_{\mathrm{ref}} + y\, \rho_{\mathrm{air}}\, g_2,$$

where p_{ref} denotes some reference pressure again.

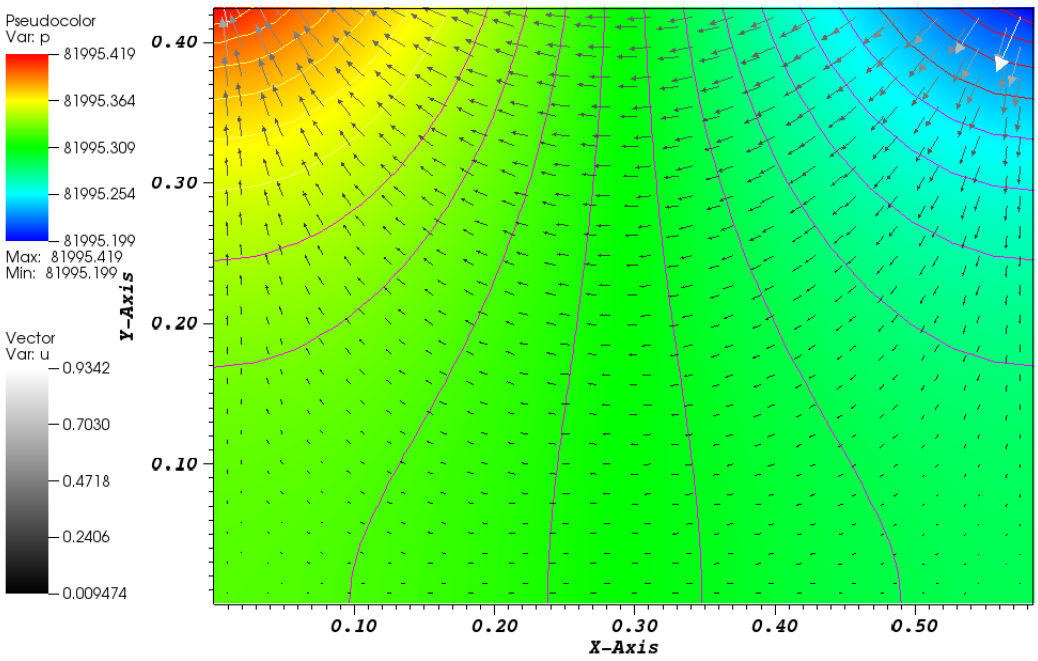


Figure 2: Pressure distribution in Ω. The arrows indicate the direction and magnitude of the pressure gradient at corresponding points. The lines are pressure isolines.

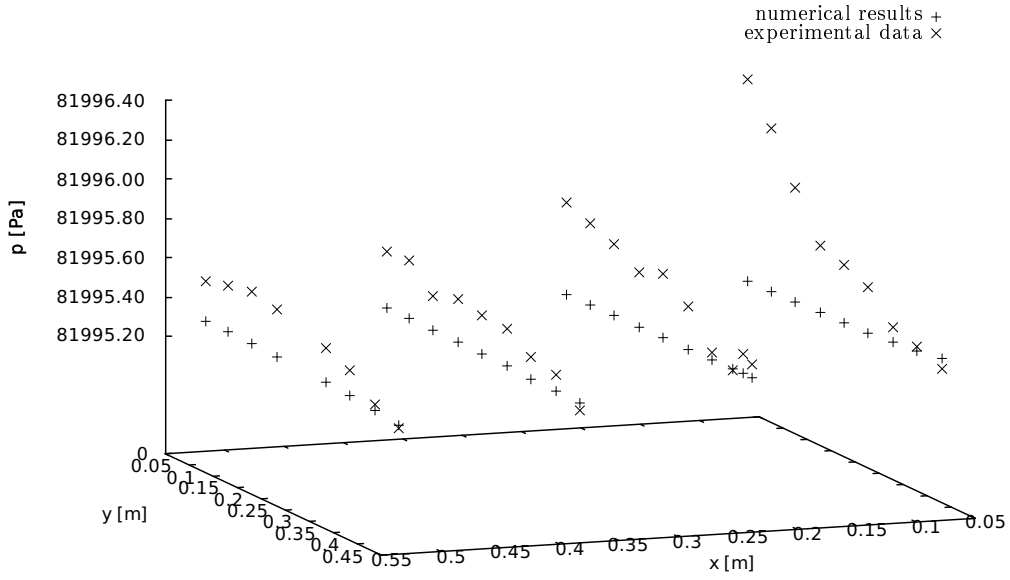


Figure 3: Comparison between the numerical results and experimental data.

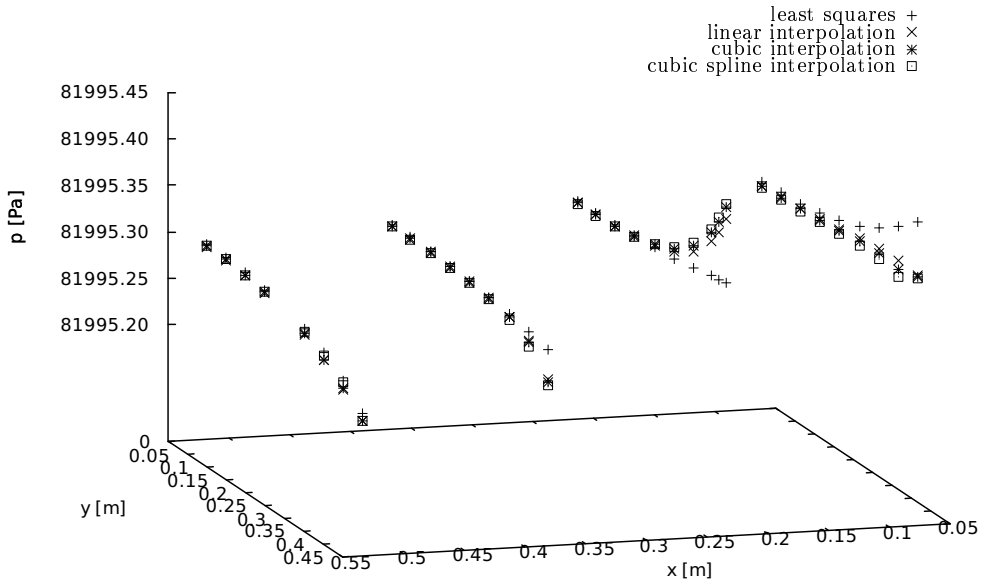


Figure 4: Comparison between the numerical results produced in cases that p_{Dir} is obtained by GNU Octave library functions for the linear least squares minimization, linear interpolation, cubic interpolation or cubic spline interpolation.

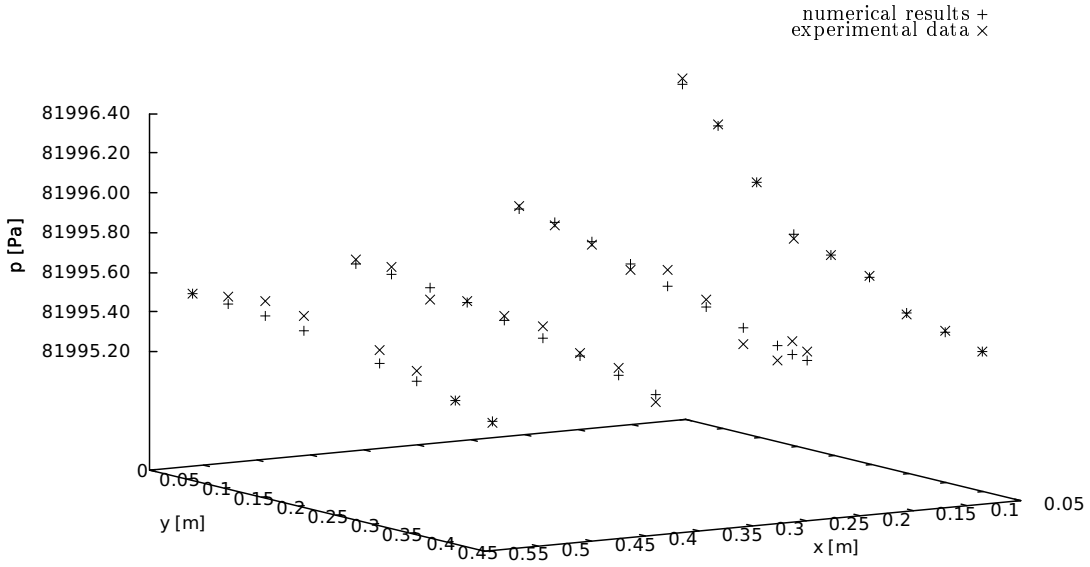


Figure 5: Comparison between the numerical results and experimental data. The values of pressure are prescribed for $x = 0.068263\,\mathrm{m}$ as well.

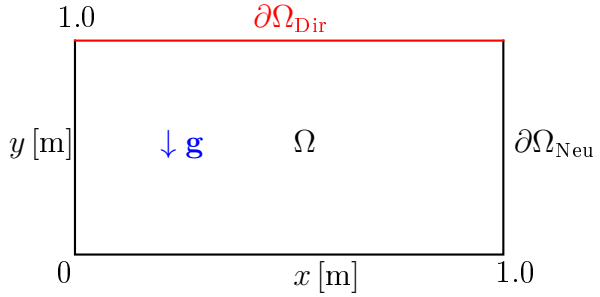


Figure 6: Description of Ω. The Dirichlet boundary condition is prescribed for $y = 1.0\,\text{m}$.

The components of the permeability tensor $\mathbf{K}$ are chosen as follows: $k_2 = 0$, $k_3 = 0$; and k_1, k_4 ($k_1 = k_4$) are spatial dependent; they will be specified later on. So will the boundary condition p_{Dir}. The values of the other parameters are summarized in Table 1b. On each side of Ω, there are 41 mesh nodes (see Figure 1a).

Several simulation were performed. In simulation 1, the domain Ω contained the spiral region of low permeability which is depicted in Figure 7a, and the constant boundary value $p_{\text{Dir}} = 151312.2677$ Pa was prescribed. The time evolution of pressure is shown in Figures 7b–7d. We can see how the pressure gradually rises in the interior of the spiral.

In simulation 2, the domain Ω contained several regions of low permeability depicted in Figure 8a, and the constant boundary value $p_{\text{Dir}} = 151312.2677$ Pa was prescribed. The time evolution of pressure is shown in Figures 8b–8d.

5 Conclusions

It has been shown in Section 3 that the numerical results do not agree with the experimental data. Nevertheless, it does not necessarily mean that the results or the model employed are wrong because the experimental data do not correspond to physical intuition whereas the numerical results do. This problem definitely requires further research.

The results presented in Section 4 illustrate the compressibility of soil-air.

References

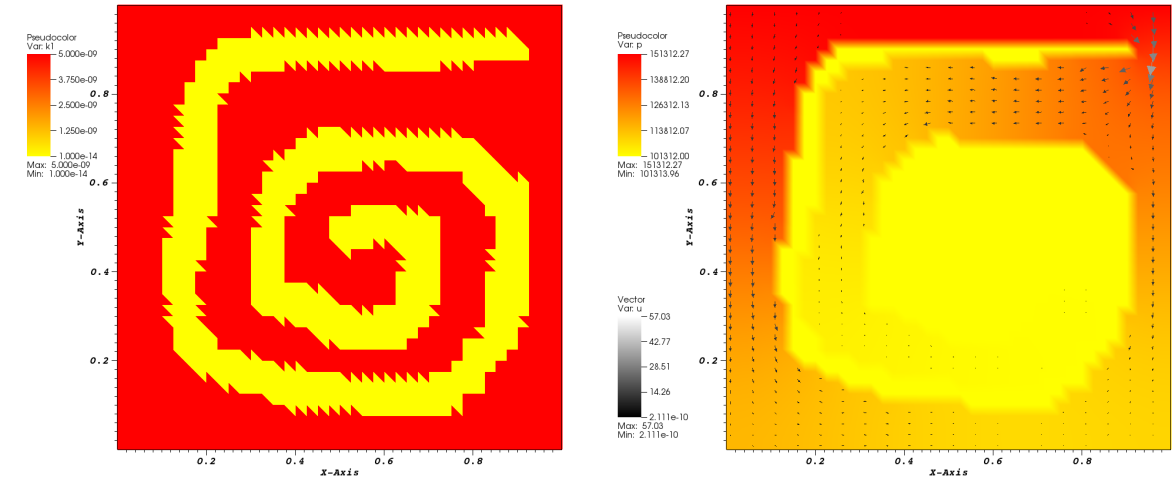
[1] J. Bear and A. Verruijt. *Modeling Groundwater Flow and Pollution.* Springer, (1987).

[2] M. Muskat. *The flow of compressible fluids through porous media and some problems in heat conduction.* Journal of Applied Physics **5** (1933).

[3] W. E. Schiesser. *The Numerical Method of Lines.* Academic, (1991).

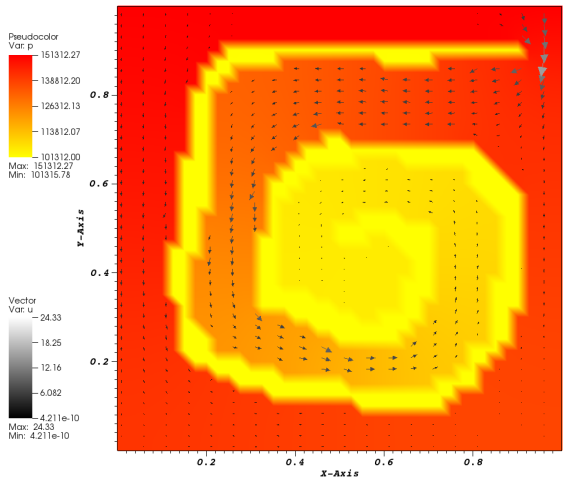
[4] M. H. Schroth, S. J. Ahearn, J. S. Selker, and J. D. Istok. *Characterization of miller-similar silica sands for laboratory hydrologic studies.* Soil Science Society of America Journal **60** (1996), 1331–1339.

[5] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems.* Springer, 1st edition, (1997).

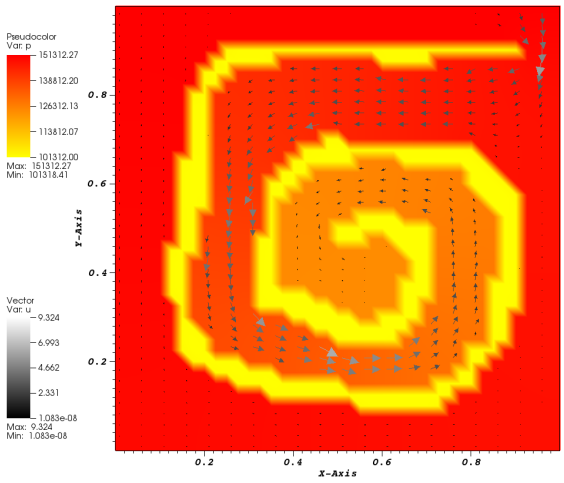


(a) Values of k_1.

(b) p at time $t = 0.01$ s.

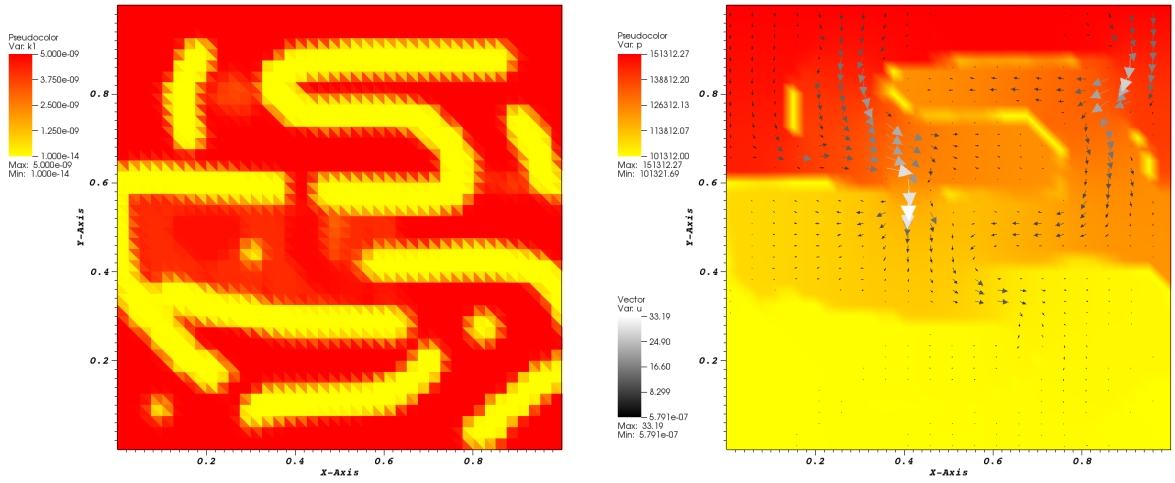


(c) p at time $t = 0.04$ s.



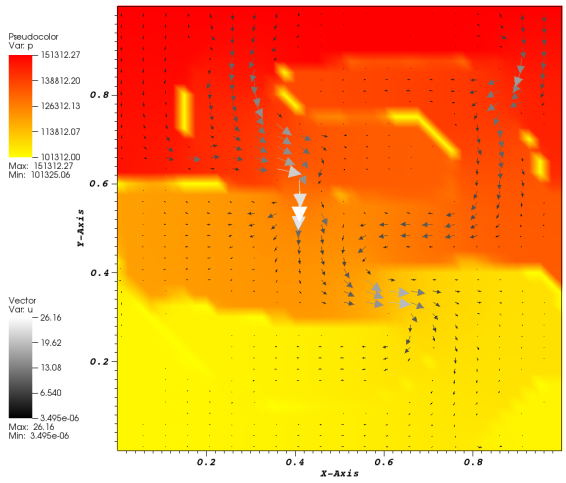
(d) p at time $t = 0.1$ s.

Figure 7: Simulation 1. Values of k_1 [m^2] and the time evolution of p [Pa]. The arrows indicate the direction and magnitude of the Darcy velocity $\mathbf{u}$ defined by (2).

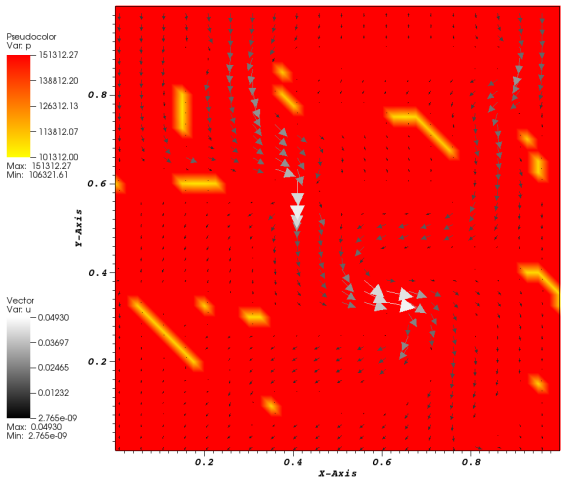


(a) Values of k_1.

(b) p at time $t = 0.01$ s.



(c) p at time $t = 0.02$ s.



(d) p at time $t = 0.5$ s.

Figure 8: Simulation 2. Values of k_1 [m^2] and the time evolution of p [Pa]. The arrows indicate the direction and magnitude of the Darcy velocity $\mathbf{u}$ defined by (2).

New Approach to Electricity Markets: Analytic Solution of ISO Problem

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:
Jiří Outrata, Institute of Information Theory and Automation, AS CR
Didier Aussel, Laboratoire PROMES, University of Perpignan

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. We present a natural condition for uniqueness of a solution to the ISO problem, and moreover we find an analytic formula for this solution. Such result is a key step towards a detailed analysis of the problem of a producer. 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. Našli jsme přirozenou podmínku pro jednoznačnost řešení ISO problému, a navíc i analytický vzorec pro toto řešení. Takový výsledek je klíčový pro následnou analýzu problému producenta. 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 [4]. 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 [1], or only a more specific setting with just two players may be considered [2].

In this article, we avoid the general problem of EPEC, and analyse the problem of the electricity market directly. We show that under a very natural assumptions the ISO problem possesses one solution on general, and moreover we find an analytic formula for such solution. Then, we may substitute this solution of lower level problem directly into the upper level problem, avoiding all these previously mentioned difficulties. However, such analysis is beyond the scope of this article. 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

$$\text{ISO(a,b)} \qquad \begin{array}{l} \min\limits_{q} \ \sum\limits_{i\in\mathcal{N}}(a_i q_i + b_i q_i^2) \\ \text{s.t.} \ \begin{cases} q_i \geq 0, \ \forall i \in \mathcal{N} \\ \sum\limits_{i\in\mathcal{N}} q_i = D \end{cases} \end{array}$$

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. [3]. In [1, 2] 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

$$\text{ISO(a,b)+(H)} \qquad \begin{array}{l} \min\limits_{q} \ \sum\limits_{i\in\mathcal{N}}(a_i q_i + b_i q_i^2) \\ \text{s.t.} \ \begin{cases} q_i \geq 0, \ \forall i \in \mathcal{N} \\ (a_i, b_i) = (a_j, b_j) \Rightarrow q_i = q_j, \forall i, j \in \mathcal{N} \\ \sum\limits_{i\in\mathcal{N}} q_i = D \end{cases} \end{array}$$

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 $\mathcal{N}^c(a,b) \subset \mathcal{N}$

$$\begin{aligned} \lambda^c(a,b) &= \min_{i\in\mathcal{N}, b_i=0} a_i \\ \mathcal{N}^c(a,b) &= \{i \in \mathcal{N} |\, a_i = \lambda^c(a,b), b_i = 0\} \\ D^c(a,b) &= \sum_{i\in\mathcal{N}_a(\lambda^c(a,b))} \frac{\lambda^c(a,b) - a_i}{2b_i} \end{aligned} \qquad (1)$$

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, cf. for instance (6). Then, $D^c(a,b)$ will be later identified with the overall amount of electricity produced by subcritical 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.

Remark 2.1. *Consider some $(a,b) \in \mathbb{R}^{2N}_+$, then we have $\lambda^c(a,b) > 0$ if and only if $a_i + b_i > 0$ for all $i \in \mathcal{N}$. In words, there is no producer offering electricity for free. This natural assumption will be useful afterwards.*

Next, we denote $\lambda_m(a) = \min_{i\in\mathcal{N}} a_i$ and 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.*

Proof. Consider $\lambda_m(a) < \lambda_1 < \lambda_2 < \lambda^c(a,b)$, since $\mathcal{N}_a(\lambda_1) \subset \mathcal{N}_a(\lambda_2)$ we have

$$F(a,b,\lambda_1) = \sum_{i\in\mathcal{N}_a(\lambda_1)} \frac{\lambda_1 - a_i}{2b_i} < \sum_{i\in\mathcal{N}_a(\lambda_1)} \frac{\lambda_2 - a_i}{2b_i} \leq \sum_{i\in\mathcal{N}_a(\lambda_2)} \frac{\lambda_2 - a_i}{2b_i} = F(a,b,\lambda_2).$$

□

The previous 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 owning 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). Then, in the forthcoming Corollary 2.6 we show 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).

Proof. The proof will be as follows. First, we find all solutions of the convex optimization problem ISO(a,b), i.e., we omit constraints (H) stemming from the equity property (H). Based on this solution set, we show that there exists a unique solution q of ISO(a,b) satisfying (H).

Since ISO(a,b) is a convex optimization problem, its solution set coincides with the solution set of the corresponding KKT system

$$\begin{cases} 0 = a_i + 2b_i q_i - \mu_i - \lambda \\ 0 \leq \mu_i \perp q_i \geq 0 \\ \sum_{i \in \mathcal{N}} q_i = D \end{cases} \tag{5}$$

where $\lambda \in \mathbb{R}$ and the first two equations are considered for all $i \in \mathcal{N}$. Let us first show that for the Lagrange multiplier λ we have

$$\lambda \in]0, \lambda^c(a,b)] \tag{6}$$

Indeed, assume for a contradiction that $\lambda \leq 0$ first. Since $D > 0$, there has to be some $j \in \mathcal{N}$ such that $q_j > 0$ and thus also $\mu_j = 0$. Then, however, $a_j + 2b_j q_j = \lambda \leq 0$ contradicts assumption $\lambda^c(a,b) > 0$, see Remark 2.1. Next, for any producer $i \in \mathcal{N}$ with linear bid, that is $b_i = 0$, the first equation of (5) gives $\lambda = a_i - \mu_i \leq a_i$ and so we have $\lambda \leq \lambda^c(a,b)$ by the definition of $\lambda^c(a,b)$.

Now, we show that

$$\{i \in \mathcal{N} | \mu_i = 0\} = \{i \in \mathcal{N} | a_i \leq \lambda\}.$$

Indeed, for all $i \in \mathcal{N}$ such that $\mu_i = 0$ we have

$$\lambda = a_i + 2b_i q_i \geq a_i \tag{7}$$

On the other hand, $\mu_i > 0$ implies $q_i = 0$ and thus also

$$\lambda = a_i - \mu_i < a_i.$$

Consequently, the last equation of (5) involves only such $i \in \mathcal{N}$ that $a_i \leq \lambda$. We may rewrite it as

$$\sum_{i \in \mathcal{N}; a_i < \lambda} q_i + \sum_{i \in \mathcal{N}; a_i = \lambda, b_i > 0} q_i + \sum_{i \in \mathcal{N}; a_i = \lambda, b_i = 0} q_i = D \tag{8}$$

Next, regarding (6) we observe that $a_i < \lambda$ implies $b_i > 0$ (we remark that λ will be at the end of the proof expressed as a function of a and b), and so we may substitute $q_i = \frac{\lambda - a_i}{2b_i}$ into the first sum using (7). Based on the same formula, we may omit the second sum since $a_i = \lambda$, $b_i > 0$ implies $q_i = 0$. To handle with the last sum, we observe that for each $i \in \mathcal{N}$ such that $a_i = \lambda$ and $b_i = 0$ we have $\lambda^c(a, b) \leq a_i$ since it is a linear bid. Next, using (6), we obtain

$$\lambda^c(a, b) \leq a_i = \lambda \leq \lambda^c(a, b) \tag{9}$$

for such i, and so $a_i = \lambda^c(a, b)$, or, in other words, $i \in \mathcal{N}^c(a, b)$. Now, if we treat all critical producers $i \in \mathcal{N}^c(a, b)$ together and use

$$Q^c(a, b) = \sum_{i \in \mathcal{N}^c(a,b)} q_i \geq 0$$

for their overall production, formula (8) reduces to

$$\sum_{i \in \mathcal{N}_a(\lambda)} \frac{\lambda - a_i}{2b_i} = D - Q^c(a, b) \tag{10}$$

We will solve this equation in a full generality in two steps. We begin with such solution of (10) that $\lambda < \lambda^c(a, b)$. This way we avoid such $i \in \mathcal{N}$ that $a_i = \lambda$ and $b_i = 0$ for the moment. For all $i \in \mathcal{N}^c(a, b)$ we have

$$\lambda = a_i + 2b_i q_i - \mu_i = a_i - \mu_i < \lambda^c(a, b) = a_i \tag{11}$$

implying $\mu_i > 0$ and thus also $q_i = 0$. Then, $Q^c(a, b) = 0$ and (10) reduces to $F(a, b, \lambda) = D$. Then, referring to Lemma 2.2 we deduce $D < D^c(a, b)$, and so using (3), we equivalently obtain $\lambda = \lambda(a, b, D)$. Altogether, all the statements in (4) are either valid or avoided provided $\lambda < \lambda^c(a, b)$. In this case, we did not consider equity property (H) assumption at all since constraints (H) are directly implied by the first equation of (4).

The second step is $\lambda \geq \lambda^c(a, b)$, but regarding (6) we have to deal with $\lambda = \lambda^c(a, b)$ only. For all $i \in \mathcal{N}_a(\lambda^c(a, b))$, the first formula in (4) derived in the previous paragraph is still valid. Thus, the overall production of this group of producers is given by

$$\sum_{i \in \mathcal{N}_a(\lambda^c(a,b))} \frac{\lambda^c(a, b) - a_i}{2b_i} = D^c(a, b)$$

Now, for $i \in \mathcal{N}^c(a,b)$ one immediately has $\lambda^c(a,b) = a_i - \mu_i = \lambda^c(a,b) - \mu_i$ and so $\mu_i = 0$. Then, we necessarily obtain $Q^c(a,b) = D - D^c(a,b)$ and thus also $D \geq D^c(a,b)$. Hence, we solved ISO(a,b) also for $\lambda = \lambda^c(a,b)$ omitting the additional assumption (H), but the solution with respect to production of critical producers $i \in \mathcal{N}^c(a,b)$ is not unique. It is unique only with respect to their overall production $Q^c(a,b)$. If $N^c(a,b) > 1$ then there are infinitely many ways how to divide $Q^c(a,b)$ among the participating producers $i \in \mathcal{N}^c(a,b)$. Then, it is the right time to tackle (H) resulting to the unique solution described by the second formula of (4). □

In general, $\lambda(a,b,D)$ is not a smooth function, but we may compute several directional derivatives easily. First, we introduce the notation. Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, then we denote the right directional derivative of $f(x_1,\ldots,x_n)$ with respect to x_i by

$$\partial^+_{x_i} f(x_1,\ldots,x_n) = \lim_{t\to 0^+} \frac{f(x_1,\ldots,x_i+t,\ldots,x_n) - f(x_1,\ldots,x_i,\ldots,x_n)}{t}$$

and analogously the left directional derivative $\partial^-_{x_i} f(x_1,\ldots,x_n)$. Since $\lambda(a,b,D)$ is a piecewise linear function in D for any $(a,b) \in \mathbb{R}^{2N}$, both directional derivatives with respect to D are well defined. Let us denote $m^\pm(a,b,D) := \partial^\pm_D \lambda(a,b,D)$

Lemma 2.4. *For fixed $(a,b) \in \mathbb{R}^{2N}$ and $D > 0$ we have*

$$\begin{aligned}
\frac{1}{m^-(a,b,D)} &= \sum_{i\in\mathcal{N}_a(\lambda(a,b,D))} \frac{1}{2b_i} && \text{if } D \leq D^c(a,b)\\
m^-(a,b,D) &= 0 && \text{if } D > D^c(a,b)\\
\frac{1}{m^+(a,b,D)} &= \frac{1}{m^-(a,b,D)} + \sum_{i\in\mathcal{N};a_i=\lambda(a,b,D)} \frac{1}{2b_i} && \text{if } D < D^c(a,b)\\
m^+(a,b,D) &= 0 && \text{if } D \geq D^c(a,b).
\end{aligned}$$

Proof. We separate (3) to three parts. For $D < D^c(a,b)$ we have $F(a,b,\lambda(a,b,D)) = D$, and so we may apply calculus of derivatives to composition of functions to obtain

$$\partial^\pm_D F(a,b,\lambda(a,b,D)) = \partial^\pm_\lambda F(a,b,\lambda(a,b,D))\, \partial^\pm_D \lambda(a,b,D) = 1$$

and so $\frac{1}{\partial^\pm_D \lambda(a,b,D)} = \partial^\pm_\lambda F(a,b,\lambda(a,b,D))$, which may be computed directly from (2). The indices of the participating producers are $\mathcal{N}_a(\lambda(a,b,D))$ in the case of $m^-(a,b,D)$, and $\{i \in \mathcal{N} | a_i \leq \lambda(a,b,D)\}$ in the case of $m^+(a,b,D)$, respectively. For $D > D^c(a,b)$ both $m^\pm(a,b,D) = 0$ since $\lambda(a,b,D)$ is constant with respect to D. Finally, we deduce the respective values at $D = D^c(a,b)$. □

Analogously, we derive the partial directional derivatives of $\lambda(a,b,D)$ with respect to the bid variables of player $i \in \mathcal{N}$.

Lemma 2.5. *For $D > 0$ and $(a,b) \in \mathbb{R}^{2N}_+$ such that $b_i > 0$, we have*

$$\begin{aligned}
\partial^\pm_{a_i} \lambda(a,b,D) &= \frac{m^\pm(a,b,D)}{2b_i} \\
\partial^\pm_{b_i} \lambda(a,b,D) &= \frac{\lambda(a,b,D) - a_i}{2b_i^2}\, m^\pm(a,b,D)
\end{aligned} \tag{12}$$

provided $a_i \geq 0$ *in the case* $\partial^+_{a_i}\lambda(a,b,D)$ *and* $\partial^\pm_{b_i}\lambda(a,b,D)$, *and* $a_i > 0$ *in the case of* $\partial^-_{a_i}\lambda(a,b,D)$.

Proof. For $D < D^c(a,b)$ we have $F(a,b,\lambda(a,b,D)) = D$. Based on partial derivaive calculus for composition of functions we immediately obtain

$$\partial^\pm_{a_i} F(a,b,\lambda(a,b,D)) + \partial^\pm_\lambda F(a,b,\lambda(a,b,D))\,\partial^\pm_{a_i}\lambda(a,b,D) = 0.$$

We note that $\partial^\pm_\lambda F(a,b,\lambda(a,b,D)) \neq 0$ because $\mathcal{N}_a(\lambda(a,b,D)) \neq \emptyset$ on the domain of F. Thanks to (2.4) we may conclude, using Lemma 2.4,

$$\partial^\pm_{a_i}\lambda(a,b,D) = -\frac{\partial^\pm_{a_i} F(a,b,\lambda(a,b,D))}{\partial^\pm_\lambda F(a,b,\lambda(a,b,D))} = -m^\pm(a,b,D)\,\partial^\pm_{a_i} F(a,b,\lambda(a,b,D)) = \frac{m^\pm(a,b,D)}{2b_i}$$

For $D > D^c(a,b)$ we have $\lambda(a,b,D) = \lambda^c(a,b)$, and so having $b_i > 0$ we see that $\lambda(a,b,D)$ is constant with respect to a_i. Thus we have $\partial^\pm_{a_i}\lambda(a,b,D) = 0$, which corresponds to our statement if we consider the appropriate equation in (2.4). Similarly, our statement complies with (2.4) also for $D = D^c(a,b)$. Finally, we note that the case of $\partial^\pm_{b_i}\lambda(a,b,D)$ is analogous. □

Since we know the formula for the unique minimizer of ISO(a,b)+(H), we may compute the overall cost $C(a,b,D)$ of production D defined as

$$C(a,b,D) = \sum_{i\in\mathcal{N}} a_i q_i(a,b,D) + b_i q_i(a,b,D)^2$$

.

Corollary 2.6. *Consider the setting of Theorem 2.3, than for* $C(a,b,D)$ *we have*

$$C(a,b,D) = \sum_{i\in\mathcal{N}_a(\lambda)} \frac{\lambda(a,b,D)^2 - a_i^2}{4b_i} \quad \text{if} \quad D < D^c(a,b) \tag{13}$$

$$C(a,b,D) = D\lambda^c(a,b) - \sum_{i\in\mathcal{N}_a(\lambda^c(a,b))} \frac{(\lambda^c(a,b) - a_i)^2}{4b_i} \quad \text{if} \quad D \leq D^c(a,b).$$

Moreover, it holds $\partial^-_D C(a,b,D) = \lambda(a,b,D)$.

Proof. For $D < D^c(a,b)$ we use $\lambda = \lambda(a,b,D)$ for brevity. With regards to (4), we restrict the sum in the definition of $C(a,b,D)$ to $i \in \mathcal{N}_a(\lambda)$ with $q_i(a,b,D) = \frac{\lambda - a_i}{2b_i}$ obtaining

$$C(a,b,D) = \sum_{i\in\mathcal{N}_a(\lambda)} a_i\frac{\lambda - a_i}{2b_i} + b_i\frac{(\lambda - a_i)^2}{4b_i^2} = \sum_{i\in\mathcal{N}_a(\lambda)} \frac{\lambda^2 - a_i^2}{4b_i}.$$

For $D \geq D^c(a,b)$ the way is analogous using formula (4) for $q_i(a,b,D)$ and splitting the sum between linear and non-linear bidders

$$C(a,b,D) = (D - D^c(a,b))\lambda^c(a,b) + \sum_{i\in\mathcal{N}_a(\lambda^c(a,b))} \frac{(\lambda^c(a,b))^2 - a_i^2}{4b_i}$$

where we moreover substitute $D^c(a,b) = \sum_{i\in\mathcal{N}_a(\lambda^c(a,b))} \frac{\lambda^c(a,b)-a_i}{2b_i}$ obtaining

$$C(a,b,D) = D\lambda^c(a,b) + \sum_{i\in\mathcal{N}_a(\lambda^c(a,b))} \frac{(\lambda^c(a,b))^2 - a_i^2 - 2(\lambda^c(a,b))^2 + 2\lambda^c(a,b)a_i}{4b_i}$$

directly giving the stated formula. We note that $C(a,b,D)$ computed using formulae (13) is continuous at $D = D^c(a,b)$. Finally, for the derivative $\partial_D^- C(a,b,D)$ at $D \in]0, D^c(a,b)]$ we have

$$\partial_D^- C(a,b,D) = \sum_{i\in\mathcal{N}_a(\lambda(a,b))} \frac{2\lambda(a,b,D)\, m^-(a,b,D)}{4b_i} = \lambda(a,b,D),$$

and the formula $\partial_D^- C(a,b,D) = \lambda^c(a,b) = \lambda(a,b,D)$ for $D > D^c(a,b)$ is immediate. □

3 Conclusion

In this article we found a new way how to treat the modelling of the electricity markets. We propose a natural assumptions called *equity property* stating that ISO does not make any difference among producers, and we show that under such assumption we may solve ISO problem analytically, see Theorem 2.3. We note that we obtain this result under a general setting newly including truly linear bids ($b_i = 0$) of producers here. Finally, we show that the central quantity $\lambda(a,b,D)$ defined by (3) is indeed a market marginal price, cf. Corollary 2.6. However, for us, all these results are mainly a workhorse to further analyse the problem of producers. This is however beyond the scope of this article as we already discussed.

References

[1] X. Hu & D. Ralph, *Using EPECs to Model Bilevel Games in Restructured Electricity Markets with Locational Prices*, Operations Research 55 (2007), 809-827.

[2] D. Aussel, M. Cervinka and M. Marechal, Day-ahead electricity market with production bounds, (2012), 24 pp.

[3] D. Aussel, R. Correa & P. Marechal *Spot electricity market with transmission losses*, J. Indust. Manag. Optim. (2012), 18 pages.

[4] 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

Prediction of the Homogeneous Droplet Nucleation by the Density Gradient Theory and PC-SAFT Equation of State

Barbora Planková

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

advisors:
Jan Hrubý, Department of Thermodynamics, Institute of Thermomechanics AS CR, v. v. i.

Václav Vinš, Department of Thermodynamics, Institute of Thermomechanics AS CR, v. v. i.

Abstract. We combined the density gradient theory (DGT) with the PC–SAFT and Peng-Robinson equations of state to model the homogeneous droplet nucleation and compared it to the classical nucleation theory (CNT) and experimental data. We also consider the effect of capillary waves on the surface tension. DGT predicts nucleation rates smaller than the CNT and slightly improves the temperature-dependent deviation of the predicted and experimental nucleation rates.

Keywords: Nucleation, density gradient theory, PC-SAFT, capillary waves

Abstrakt. Zkombinovali jsme teorii gradientu hustoty se stavovými rovnicemi PC–SAFT a Pengovou-Robinsonovou, abychom modelovali homogenní nukleaci kapek. Tyto výsledky jsme porovnali s klasickou nukleační teorií a experimentálními daty. Také jsme uvažovali efekt kapilárních vln na povrchové napětí. Gradientní teorie predikuje menší nukleační rychlosti než klasické a trochu vylepšuje odchylku teplotní závislosti teoretických a experimentálních nukleačních rychlostí.

Klíčová slova: Nukleace, theorie gradientu hustoty, PC-SAFT, kapilární vlny

1 Introduction

The classical nucleation theory (CNT) is widely used to model the homogeneous droplet nucleation. However due to the capillary approximation, even small molecular clusters are treated as macroscopic droplets. This flaw is at least partially overcome in the density gradient theory (DGT) [11, 2]. Unlike the CNT, this theory describes the surface tension varying with the size of the cluster. In this work, we compare both nucleation theories. We incorporate a physically based equation of state (EoS), the PC–SAFT [3, 5], into the DGT model and compare it with the classical cubic EoS, Peng–Robinson (PR). Another problem of DGT is that it ignores the effect of capillary waves (CW) [1]. We attempt to consider this effect.

The nucleation process is described by the nucleation rate J. In this work, a modified internally consistent (IC) [4] value is used,

$$J_{\mathrm{IC}} = \frac{\rho_{\mathrm{G}}\rho_{\mathrm{G}\infty}N_{\mathrm{A}}}{\rho_{\mathrm{L}\infty}}\sqrt{\frac{2\sigma_{\infty}}{\pi M}}\exp\left(-\frac{\Delta\Omega}{k_{\mathrm{B}}T}\right). \tag{1}$$

Here, subscript ∞ reffers to the saturated state, ρ_{G} and ρ_{L} are the densities of the bulk vapor and liquid of the system, N_{A} is the Avogadro constant, σ is surface tension, M is molecular mass, k_{B} is the Boltzmann constant, and $\Delta\Omega$ is the work of formation of the critical cluster.

The work of formation according to the DGT reads

$$\Delta\Omega(\rho) = \int_0^{\infty}\left[\Delta\omega_{\mathrm{hom}}(\rho) + \frac{c}{2}\left(\frac{\mathrm{d}\rho}{\mathrm{d}r}\right)^2\right]4\pi r^2\mathrm{d}r, \tag{2}$$

where $\Delta\omega_{\mathrm{hom}}$ can be found e.g. in [9], second term containing the density gradient and influence parameter c brings the inhomogeneity caused by the presence of the interface. Using Eq. (2), an Euler–Lagrange equation can be derived,

$$\frac{\mathrm{d}^2\rho}{\mathrm{d}r^2} + \frac{2}{r}\frac{\mathrm{d}\rho}{\mathrm{d}r} = \frac{1}{c}\Delta\mu(\rho), \tag{3}$$

where $\Delta\mu = \partial\Delta\omega_{\mathrm{hom}}/\partial\rho$. Including the boundary conditions $\rho(r\to\infty) = \rho_{\mathrm{G}}$, $d\rho/dr(0) = 0$, Eq. (3) defines a boundary value problem (BVP) that can be solved numerically.

The PC–SAFT EoS [3, 5] is based on the Statistical Associating Fluid Theory (SAFT) combining important interatomic and intermolecular forces, such as covalent bonding, hydrogen bonding, Coulombic forces and can be used for very different shapes of molecules. Due to the fact that the SAFT EoS works directly with the molecular structure of substances, it allows a more realistic modeling of fluids in the metastable region which is needed in the DGT model.

2 Numerical computations and results

The simply looking BVP defined by (3) has two difficulties: density profile near the vapor phase has a very sharp shape; its slope changes abruptly from the very steep decline to an almost constant profile. Second problem is that for large droplets density profile in the interior of the droplet changes only negligibly and is almost constant. This causes a significant cumulation of numerical errors. This work is based on results of [9], in which the shooting method was used instead of more sophisticated ones based on the finite difference schemes. The reason is that it was easier to develop a convergent routine algorithm in that way. To overcome many difficulties that arose during the solution process, several original numerical methods were developed.

Nucleation rates were computed using both nucleation theories, DGT and CNT, and two EoSs, the PC–SAFT and PR. Theoretical predictions were compared to experiments [10, 6, 7, 12, 13].

The left-hand side of Fig. 1 shows the nucleation rate J as a function of supersaturation S for four temperatures, both nucleation theories and both EoSs compared to

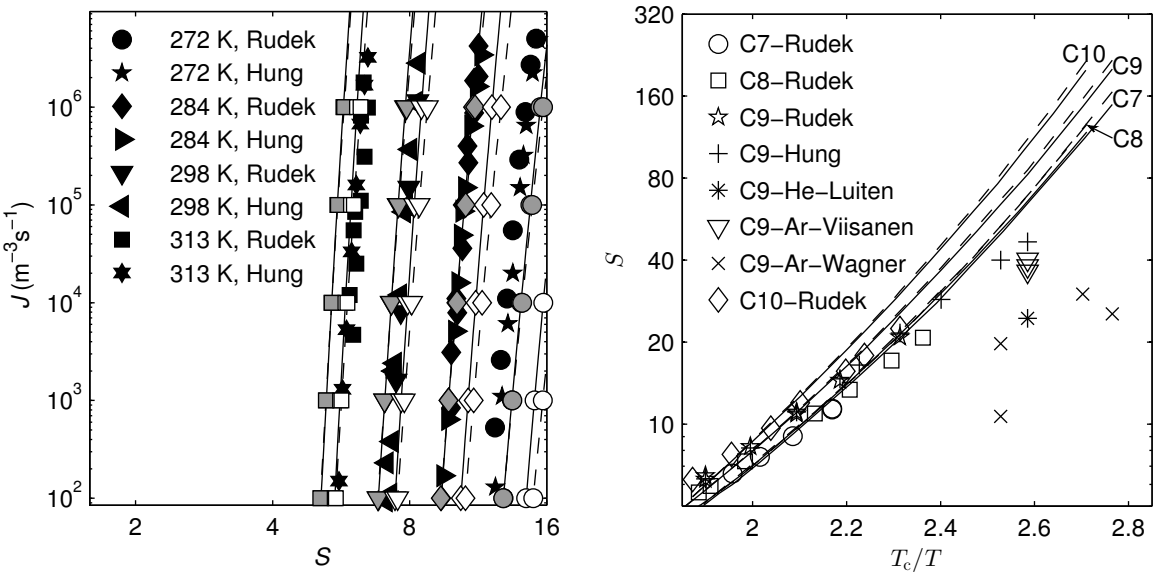


Figure 1: Left: Nucleation rates of n-nonane computed using the CNT (dashed line) and DGT (solid lines), PR EoS (white symbols) and PC–SAFT Eos (grey symbols) for temperatures 272 K, 284 K, 298 K, and 313 K. Right: Dependence of the supersaturation S on the inverse reduced temperature T/T_c. Lines correspond to values computed using the GT–IC (solid lines) and CNT–IC (dashed lines) for n-heptane (C7), n-octane (C8), n-nonane (C9), and n-decane (C10). As EoS was used only PC–SAFT.

the experimental nucleation rate data.The right-hand side of Fig. 1 shows supersaturations S as functions of temperature at a constant nucleation rate $J = 10^6\mathrm{m}^{-3}\mathrm{s}^{-1}$ (close to most experimental data range). These values were computed and compared for four substances: n-heptane, n-octane, n-nonane, n-decane. Supersaturations of experimental data were linearly interpolated to match the value of nucleation rate $J = 10^6\mathrm{m}^{-3}\mathrm{s}^{-1}$. Data [13] are far from this value causing a disagreement with others.

As aforementioned, despite the DGT does not contain a CW effect, the influence parameter c in Eqs. (2), (3) is determined using the experimental surface tension $\sigma_\infty \equiv \sigma_\mathrm{exp}$ that includes it. We attempt to avoid this inconsistency by using Meunier's mode-coupling theory [8], the surface tension without the CW effect can be expressed as

$$\sigma_\mathrm{non-cw} = \sigma_\mathrm{exp}\left(1 + \frac{3}{8\pi}\frac{T}{T_\mathrm{c}}\frac{1}{2.55^2\kappa}\right). \tag{4}$$

Here, T_c is the critical temperature and κ is the universal amplitude ratio determined by experiments and simulations to be $\kappa \cong 0.39$.

Figure 2 shows surface tension σ as a function of the pressure difference Δp, and nucleation rate J as a function of the supersaturation S for n-nonane computed using the DGT at $T = 313$ K. Two EoSs were used (PR, PC–SAFT), and both approaches are incorporated: influence parameter c is computed using σ_exp directly (grey symbols) and with removing the CW effect using (4) (white symbols). The effect for the nucleation rates is large which proves the importance of this procedure.

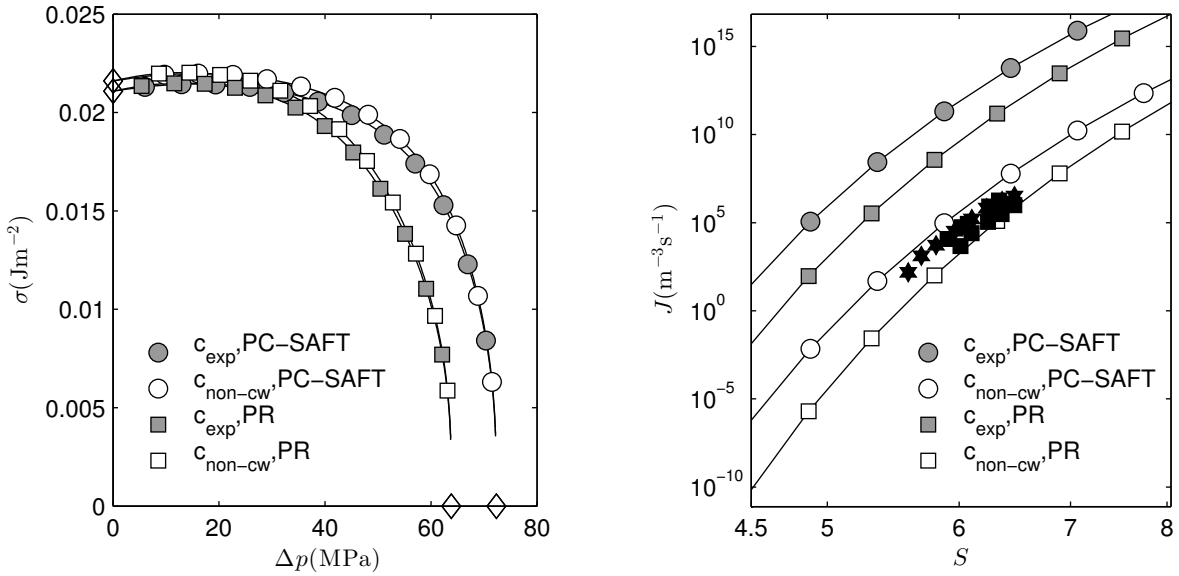


Figure 2: Surface tensions σ as a functions of Δp (left) and nucleation rates J as a functions of S (rights) computed using the DGT, PR EoS (squares) and PC–SAFT EoS (circles) for n-nonane at 313 K. Influence parameter computed using the experimental surface tension σ_{exp} (grey symbols) and using modified surface tension $\sigma_{\mathrm{non-cw}}$ are used.

3 Conclusions

Our computations show that the DGT predicts nucleation rates smaller than the CNT because the surface tension predicted by the DGT for the critical clusters is lower than for the planar phase interface. This effect is more pronounced at low temperatures and high supersaturations where the critical clusters are smaller. The more realistic PC-SAFT EoS predicts higher nucleation rates than the PR EoS. The influence of the capillary waves significantly lowers the predicted nucleation rates. This effect, however, requires further investigation. A large part of the temperature dependent deviation of theoretical predictions from experimental data still remains unexplained.

4 Acknowledgements

This work was presented at 19th International Conference on Nucleation and Atmospheric Aerosols 2013, 24-28 June 2013, Fort Collins, Colorado (USA) and published in the conference proceedings.

References

[1] F. P. Buff, R. A. Lovett, and J. F. H. Stillinger. *Interfacial density profile for fluids in the critical region.* Phys. Rev. Lett. **15** (1965), 621–623.

[2] J. W. Cahn and J. E. Hilliard. *Free energy of a nonuniform system. i. interfacial free energy.* J. Chem. Phys. **28** (1958), 258–267.

[3] W. Chapman, K. Gubbins, G. Jackson, and M. Radosz. *Saft: Equation-of-state solution model for associating fluids.* Fluid Phase Equilib. **52** (1989), 31–38.

[4] S. L. Girshick and C.-P. Chiu. *Kinetic nucleation theory: A new expression for the rate of homogeneous nucleation from an ideal supersaturated vapor.* J. Chem. Phys. **93** (1990), 1273.

[5] J. Gross and G. Sadowski. *Perturbed-chain saft: An equation of state based on a perturbation theory for chained molecules.* Ind. Eng. Chem. Res. **40** (2001), 1244–1260.

[6] C.-H. Hung, M. J. Krasnopoler, and J. L. Katz. *Condensation of a supersaturated vapor. viii. the homogeneous nucleation of n-nonane.* J. Chem. Phys. **90** (1989).

[7] C. Luijten. *Nucleation and Droplet Growth at High Pressure.* PhD thesis, Technische Universiteit Eindhoven, (1998).

[8] J. Meunier. *Liquid interfaces: role of the fluctuations and analysis of ellipsometry and reflectivity measurements.* J. Phys. **48** (1987), 1819.

[9] B. Planková. *Mathematical modeling of spherical phase interfaces in real fluids.* Master's thesis, CTU FNSPE, (2011).

[10] M. M. Rudek, J. A. Fisk, V. M. Chakarov, and J. L. Katz. *Condensation of a supersaturated vapor. xii. the homogeneous nucleation of the n-alkanes.* J. Chem. Phys. **105** (1996).

[11] J. D. van der Waals. *The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density.* Verhandel. Konink. Akad. Weten. **1** (1893).

[12] Y. Viisanen, P. E. Wagner, and R. Strey. *Measurement of the molecular content of binary nuclei. iv. use of the nucleation rate surfaces for the n-nonane-n-alcohol series.* J. Chem. Phys. **108** (1998), 4257.

[13] P. E. Wagner and R. Strey. *Measurements of homogeneous nucleation rates for n-nonane vapor using a two-piston expansion chamber.* J. Chem. Phys. **80** (1984), 5266.

Compositional Modeling in Porous Media Using Constant Volume Flash and Flux Computation without the Need for Phase Identification*

Ondřej Polívka

4th year of PGS, email: ondrej.polivka@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jiří Mikyška, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The paper deals with the numerical solution of a compositional model describing compressible two-phase flow of a mixture composed of several components in porous media with species transfer between the phases. The mathematical model is formulated by means of the extended Darcy's laws for all phases, components continuity equations, constitutive relations, and appropriate initial and boundary conditions. The splitting of components among the phases is described using a new formulation of the local thermodynamic equilibrium which uses volume, temperature, and moles as specification variables. The problem is solved numerically using a combination of the mixed-hybrid finite element method for the total flux discretization and the finite volume method for the discretization of transport equations. A new approach to numerical flux approximation is proposed, which does not require phase identification and determination of correspondence between the phases on adjacent elements. The time discretization is carried out by the backward Euler method. The resulting large system of nonlinear algebraic equations is solved by the Newton-Raphson iterative method. We provide seven examples of different complexity to show reliability and robustness of our approach.

This work was presented at **Interpore Conference 2013** in Prague (21.–24.5.2013) and the full article has been submitted to the **Journal of Computational Physics**.

Keywords: compositional simulation without phase identification, mixed-hybrid finite element method, finite volume method, phase-by-phase upwinding, constant-volume phase splitting, pressure computation

Abstrakt. Článek pojednává o numerickém modelování kompozičního modelu popisujícího stlačitelné dvoufázového proudění směsi složené z několika komponent v porézních prostředích s látkovou výměnou mezi fázemi. Matematický model je formulován pomocí rozšířeného Darcyho zákona, rovnic kontinuity pro složky směsi, konstitutivních vztahů a vhodných počátečních a okrajových podmínek. Rozdělení komponent mezi fázemi je popsáno pomocí nové formulace lokální termodynamické rovnováhy při zadaném objemu, teplotě a látkových množstvích jednotlivých komponent. Problém je řešen numericky za použití kombinace smíšené hybridní

*This work has been supported by the project P105/11/1507 "Development of Computational Models for Simulation of CO2 Sequestration" of the Czech Science Foundation, project KONTAKT II LH 12064 "Computational Methods in Thermodynamics of Hydrocarbon Mixtures" of the Ministry of Education, Youth and Sport of the Czech Republic, and project SGS11/161/OHK4/3T/14 "Advanced Supercomputing Methods for Implementation of Mathematical Models" of the Student Grant Agency of the Czech Technical University in Prague.

metody konečných prvků pro diskretizaci celkového toku a metody konečných objemů pro diskretizaci transportních rovnic. Je navržen nový přístup k aproximaci numerického toku, který nevyžaduje identifikaci fází ani určování odpovídajících si fází mezi sousedícími elementy. Časová diskretizace je provedena zpětnou Eulerovou metodou. Výsledná rozsáhlá soustava nelineárních algebraických rovnic je řešena Newtonovou-Raphsonovou iterační metodou. Pro znazornění stability a robustnosti našeho přístupu uvádíme sedm příkladů různého charakteru.

Tato práce byla prezentována na konferenci **Interpore 2013** v Praze (21.–24.5.2013) a celý článek je podán do časopisu **Journal of Computational Physics**.

Klíčová slova: kompoziční simulace bez fázové identifikace, smíšená hybridní metoda konečných prvků, metoda konečných objemů, upwind po fázích, fázový rozklad při konstantním objemu, výpočet tlaku

References

[1] F. Brezzi, M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, New York Inc. (1991).

[2] Z. Chen, G. Ma Y. Huan. *Computational Methods for Multiphase Flows in Porous Media.* SIAM, Philadelphia (2006).

[3] A. Firoozabadi. *Thermodynamics of Hydrocarbon Reservoirs.* McGraw-Hill, NY (1998).

[4] T. Jindrová, J. Mikyška. *Fast and Robust Algorithm for Calculation of Two-Phase Equilibria at Given Volume, Temperature, and Moles.* Fluid Phase Equilibria (2013), Vol. 353 (Sep 15, 2013), pp. 101–114.

[5] R. J. Leveque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge University Press, Cambridge (2002).

[6] J. Mikyška, A. Firoozabadi. *Implementation of higher-order methods for robust and efficient compositional simulation.* Journal of Computational Physics, 229 (2010), pp. 2898–2913.

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

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

[9] O. Polívka, J. Mikyška. *Numerical simulation of multicomponent compressible flow in porous medium.* Journal of Math-for-Industry Vol. 3 (2011C-7), (2011) pp. 53–60.

[10] A. Quarteroni, R. Sacco, F. Saleri. *Numerical Mathematics.* Springer-Verlag, New York (2000).

Feature Definition and Software Design for Java Source Code Classification Tool*

Michal Rost

3rd 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 utilizes the experience from the field of software engineering to formulate a set of features for identifying known structures in the source code of software project. Furthermore, a software tool for the analysis of project code is proposed.

Keywords: Java, software design, design patterns

Abstrakt. Tento článek využívá poznatků z oblasti softwaroveho inženýrství k sestavení sady příznaků pro rozpoznávání známých struktur ve zdrojovém kódu softwarového projektu. Dále předkládá návrh softwarového nástroje pro analýzu projektového kódu.

Klíčová slova: Java, návrh softwaru, návrhové vzory

1 Introduction

According to the majority of contemporary software development methodologies [2, 10], the design phase should precede the implementation stage during a software development process. In order to be easily maintainable and reusable, it is expected that developed software will be well designed, thus composed of interrelated objects, where each object is responsible for a particular task. Design patterns [11] are a well known standard for software design; however, once the implementation phase is over, it is difficult to determine whether software uses these patterns, or if they were implemented properly. Therefore, an effort is being made to create a tool for software quality assessment. In recent years, various approaches have been explored; [6] have performed statistical analysis of code smells [7] to suggest further refactoring techniques [5]. The authors of [4, 14] have focused directly on design patterns in order to determine which patterns are utilized in the examined source code. While [4] use a statistical approach based on previously defined predictors, [14] search for patterns using graph algorithms.

Our approach is to create a tool which will be able to detect well designed data types in a given project code and separate them from noise (poorly designed data types); we consider a data type to be well designed if it satisfies UML class stereotype [9] or represents single class design pattern. Second, we want our tool to perform an analysis of relationships among a project's data types, to detect multi-class design patterns. This

*This work has been supported by the grants SGS 11/167/OHK4/3T/14 and LA08015

paper focuses on the first phase, particularly on the definition of features for well designed types.

2 Classes and features

2.1 Classes

As mentioned above, we are trying to detect well designed data types with usage of eleven statistical classes (patterns) that are listed in Table 1.

Bean represents a storage type which holds attributes and provides access to them through setters and getters. DAO stands for data access object, which mediates access to a collection of data. The composite is a tree node of composite pattern. The constant is composed of constant or immutable objects and represents a configuration of a particular part of the application. A factory encapsulates methods for creation of new objects based on given parameters. A builder manages and sets up a newly created object. An adapter allows adaptation of an adaptee object from one interface to another. A proxy object substitutes another object of the same interface and allows changing of the implementation of some of its methods. A decorator adds additional properties to an object of the same interface. A worker combines or uses other objects in order to perform the main functions of a certain part of the application. An utility type manages static methods of a similar purpose in order to separate mechanical work from worker types.

Table 1: Recognized data type patterns

Name	*Satisfies*	*Represents*	*Responsible for*
Bean	type stereotype	crate pattern	data storage/access
Composite		composite pattern	data storage/access
Constant	utility stereotype		data storage/access
DAO	entity stereotype		data storage/access
Builder		builder pattern	object creation
Factory		factory method pattern	object creation
Adapter		adapter pattern	object manipulation
Decorator		decorator pattern	object manipulation
Proxy		proxy pattern	object manipulation
Worker	focus stereotype		object manipulation
Utility	utility stereotype		support

2.2 Features

Up to now, we have defined over forty different features; these features are divided into four major categories: *expression features*, *statement features*, *member features* and *relation features*. Expression and statement features are connected with expressions and statements in the project code, a typical expression feature is, for instance, a number of

instantiations within a definition of a data type weighted by total number of expressions in the same data type. A typical member feature is, for example, a number of public, non-static setters and getters in a selected data type weighted by total number of methods in the same data type. Relation features depict a relationship of a data type with its surroundings; this kind of feature is, for instance, a logical value which is set to true if the data type uses its direct parent type as an attribute. The explanation of selected features follows.

2.2.1 Feature fm#amr

Feature fm#amr is represented by (1), where n is a number of non-abstract, non-static, non-setter and non-getter methods of given data type, A is a number of non-static attributes, and a_i is a number of non-static attributes used in i-th method. Usage (invocation) of a setter or getter for a local attribute counts as usage of this particular attribute.

$$\frac{1}{n \cdot A} \cdot \sum_{i=1}^{n} a_i \tag{1}$$

This feature describes how much a given data type works with its own attributes. We have estimated that an fm#amr value may be close to number one for worker class data types.

2.2.2 Feature fm#mnew

Feature fm#mnew counts *factory methods* [11], thus public methods that contain instantiation (new expression) of a local (non-attribute) variable and return it as a result. Member methods that return invocation of factory methods also count as factory methods, this rule applies recursively. Member methods that return instance of same type as the type they are a member of do not count as factory methods for this feature. The resulting count of factory methods is weighted by the total number of member non-setter, non-getter and non-abstract methods in the corresponding data type.

This feature represents the share of factory methods in the total number of methods and might help to detect a factory class.

2.2.3 Feature fm#anew

Feature fm#anew represents the number of attributes instantied within constructors and non-static member methods, weighted by the total number of non-static attributes.

The feature tells how often the type's attributes are instantiated within its member methods and could be useful for builder class detection.

2.2.4 Feature fr2nsa

Value of fr2nsa equals to one if a given type is recursive; thus, holds a non-static attribute of same type. Otherwise the value is equal to zero. This feature could help to find composite, proxy or decorator types.

2.2.5 Feature fr2ia

Value of fr2ia equals to one if a given type has attributes of same type as its direct parent is; otherwise the value is equal to zero. Similarly to fr2nsa, this feature could help in finding composite, proxy or decorator types.

2.2.6 Feature fm#apc

Feature fm#apc represents number of public constant attributes of a given type weighted by the total number of type's attributes.

2.2.7 Feature fs#cyc

Value of fm#cyc holds number of cycle statement in a particular type weighted by the total number of statements in the same type.

2.2.8 Feature fm#mmou

Let M be a set of all public, non-abstract, non-setter and non-getter member methods of a particular data type. Then feature fm#mmou[1] is represented by (2), where n is the size of M; csa_i is the number of methods from M in that static method of i-th type is accessed; cpu_i is the number of methods from M in that parameter of same i-th non-trivial type is used; cau_i is the number of methods from M in that i-th attribute is accessed; cnv_i is the number of methods from M in that same i-th non-trivial type is instanced and used.

$$\frac{1}{n} \cdot \max\left\{\max_i csa_i, \max_i cpu_i, \max_i cau_i, \max_i cnv_i\right\} \tag{2}$$

We expect that this feature should be useful for separating DAO from other classes. Since DAO is utilized for querying a specific datasource, there always has to be an object, which mediates access to data. However, it is not known, if the object is passed to DAO's methods as a parameter, if it is an attribute of DAO, if it is singleton, utility type, or if it is created directly within DAO's methods. Therefore, fm#mmou chooses the most probable from all mentioned possibilities.

3 Tool design

3.1 Requirements

Functional and non-functional requirements were collected, before the design of the proposed tool was begun.

Functional requirements:

1. Manage features definitions and collect features data
2. Classify project code

[1] This feature has not been implemented yet.

3. Manage classification results

4. Provide posterior analysis for classified data

Ad 1. Collect features from the project's source code, store them in the corresponding objects and provide access to these objects.

Ad 2. Provide various statistical classifiers; enable their configuration and classification of the project's data types with the chosen classifier. Consequently, obtain results for all data types in the project and all the considered classes, where each result is represented by a probability that a given data type belongs to one particular class.

Ad 3. Manage the project's classification history, registered for all runs of all classifiers, because results from one uniformly configured classifier can vary over time.

Ad 4. Provide additional operations in order to measure the quality of classifiers, perform cross validations, or apply bilantion criteria to classification results.

Non-functional requirements:

1. Clean object design

2. OS independence

Ad 1. The application has to be separated into individual components that will provide their interfaces with the rest of the application. This allows easy interchangeability of component implementation, or simply adding a new implementation.

Ad 2. Since Java is an OS independent framework, the designed tool is also required to be OS independent in order to integrate it into Java IDEs in the future.

3.2 Components design

The four main components have been identified during the design phase: *collector*, *classifier*, *validator* and *launcher* (Figure 1).

The first component, the collector [13], is responsible for processing source codes and the collection of features data. A source code is parsed and an abstract syntax tree (AST) [1, 12] created as a result, consequently features are mined [13] from AST.

The second component, the classifier, is the core part of the whole application. Classifiers can be either simple or compound, where a compound classifier consists of two or more other classifiers and a balance criterion. The balance criterion is a judge among the sub classifiers and makes the final decision. The classifier is responsible for the identification of which category an observation belongs to.

The third component, the validator, is used for regression model validation, particularly a k-fold cross validation. This is a process of determination how results of a statistical analysis will affect independent data sets. During k-fold validation a project observation (a set of features of each data type) is split into k disjoint subsets, then k - 1 subsets are utilized to train a classifier and one subset is used for validation (testing). Cross validation is finished after all k subsets were used for validation.

The last component, the launcher, represents only the layer that performs top-level operations over the classifier, validator and collector components. It will allow the user to start classification or validation and configure their parameters.

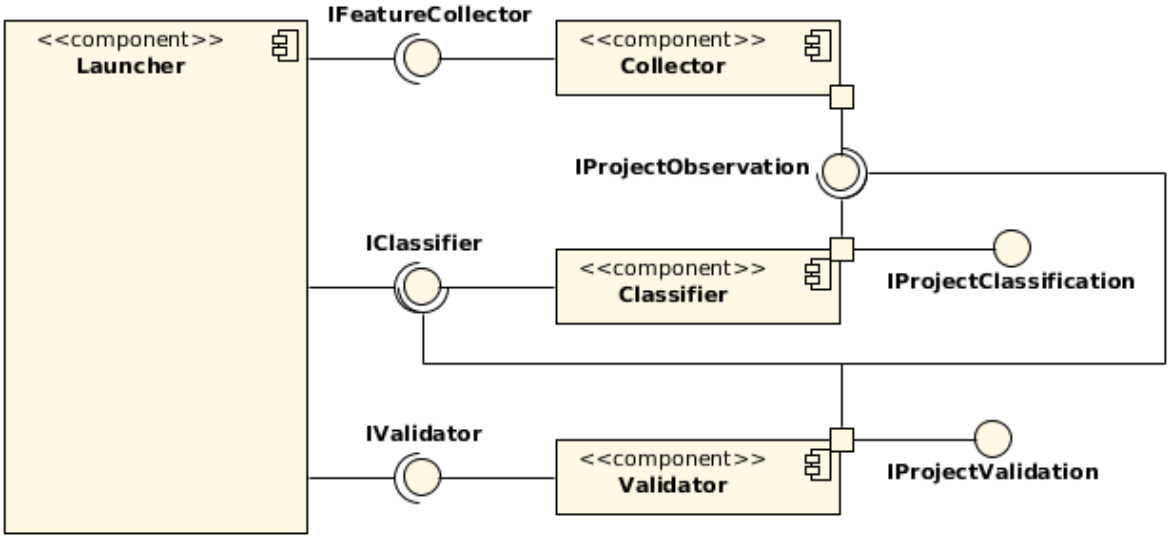


Figure 1: A component diagram of designed tool. Four main components have been identified: *collector*, *classifier*, *validator*, and *launcher*.

4 Training data

Four data sets were prepared in order to train our classifiers.

Mixed data set consists of java source files selected from different open source projects or design pattern tutorials. These files contain various implementations of all classes from Table 1; currently there are 175 types with at least fourteen representants of each class.

JaHoCa project (Java Home Cash) is a simple java application for monitoring personal incomes/expenses. There is a lack of design patterns in this project. Nevertheless, it is rich in beans and utility classes.

Andengine is an open source graphic 2D/3D engine for the android operating system. The project contains many workers, decorators, factories and utility classes.

JHotDraw is a simple java drawing/plotting tool, which is strongly based on design patterns, with many adapters, factories, composites or decorators.

5 Results

Until now, the launcher, collector and validator components of the tool have been implemented. Collected information about project's sources are being exported to the CSV format and passed to Matlab for subsequent analysis.

Table 2 shows results from analysis of the "Mixed data set" with *k-NN* (k Nearest Neighbours, k = 4 has proved to be optimal for this problem), *LDA* (Linear Discriminant

Table 2: Success rates of classifiers for "Mixed data set".

	k-NN	*LDA*	*SVM*
best submodel	0.87931	0.93103	0.82759
full model	0.82184	0.90805	0.77011

Table 3: Feature usage in 100 best sub-models.

Feature	*k-NN*	*LDA*	*SVM*
fr2nsa	100	100	100
fr2ia	100	100	100
fm#apc	98	100	96
fm#mnew	95	100	98
fs#cyc	91	100	100
fm#mase	91	93	99
fm#mpnov	82	100	100
fm#mpnop	86	100	95
fm#mpars	87	97	97
fm#anonp	77	98	98
fs#switch	79	100	93
fg2a	77	93	100
fs#sif	81	97	92
fm=mpard	65	100	100
fg#esr	62	100	97
fe#new	64	94	92
fm#mpnoo	49	100	98
fe#inv	55	100	91
fm#anew	45	100	100
fm#mpard	53	93	97

Feature	*k-NN*	*LDA*	*SVM*
fm#mps	50	100	92
fe#invm	53	92	95
fe#invo	54	97	88
fm#mpna	46	100	92
fs#elif	22	100	100
fm#mpnn	19	100	99
fm#anos	37	97	81
fm#mpngs	12	97	100
fe#newm	14	100	91
fr2ssa	8	97	97
fm=mase	23	100	78
fm#anonn	99	0	100
fm#mn	10	92	90
fm#mpara	8	95	88
fm#anpn	87	4	99
fm#apn	10	93	83
fm#amr	55	100	11
fm#anps	39	29	93
fe#casto	23	97	36
fm#mparcu	58	90	0

Analysis) and *SVM* (Support Vector Machines) classifiers [3]. Due to the large number of features, sub-models have been utilized [8] and FSA heuristic has been used [8] for finding the best sub-model; the heuristic has been applied ten times for each classifier and ten best results from each run have been recorded.

Table 3 summarizes usage of features in 100 best sub-models found by the heuristic. Features fr2nsa and fr2ia participated in all chosen sub-models, they successfuly separated recursive types from others. Frequent usage of fm#apc can be caused by fact that public attributes do not appear often in other classes than constant. Feature fm#mnew proved to be good separator of factory class, on the other hand fm#amr has to be improved.

6 Conclusion

The paper has focused on the problem of software quality measurement, and presented our solution to detect well designed and implemented data types, based on a newly defined set of features. Eight selected features were briefly explained; consequently, the tool for feature collection and statistical classification over the source code was proposed. In the future, we will continue to improve features and reduce their number. Moreover, we will implement classifiers into the proposed tool. Last but not least, we will focus on defining and collecting object relation features.

References

[1] A. V. Aho, M. S. Lam, R. Sethi and J. D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley, (2006), 2nd edition.

[2] J. Arlow and I. Neustadt. *UML 2 and the Unified Process: Practical Object-Oriented Analysis and Design*. Addison-Wesley, (2009), 2nd edition.

[3] R. O. Duda. *Pattern Classification*. John Wiley, (2001), 2nd edition.

[4] 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.

[5] M. Fowler. *Refactoring: Improving the Design of Existing Code*. Addison-Wesley, (1999).

[6] 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.

[7] 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.

[8] M. Mojzeš. *Quality of Fractographic Sub-Models via Cross-Validation*. In 'Proceedings Doktorandské dny 2012', Czech Technical University, (2012), 167–176.

[9] OMG. *Unified Modeling Language 2.5 specification*. Object Management Group, (28 September 2013).

[10] C. E. Otero. *Software Engineering Design: Theory and Practice*. CRC Press, (2012).

[11] R. Pecinovský. *Návrhové vzory*. Computer Press, (2007).

[12] J. Smolka. *Refactoring tool for Java programs*. Master's thesis, Czech Technical University, (2010).

[13] J. Smolka. *Feature collection for source code classification and pattern recognition.* In 'Proceedings Doktorandské dny 2013', Czech Technical University, (2013).

[14] N. Tsantalis, A. Chatzigeorgiou, G. Stephanides, and S. T. Halkidis. *Design Pattern Detection Using Similarity Scoring.* In 'Transactions on Software Engineering', IEEE Computer Society, volume 32, issue 11, (2006), 896–909.

Fock-Cadabra Approach to Stress-energy Tensor*

Josef Schmidt

3rd 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

Abstract. Cadabra system is used for implementation Fock method for finding conserved Lorentz covariant stress-energy complexes. The method is developed for linear second-derivative field equations and afterwords generalised for equations of motion containing non-derivative terms. The results for linearized vacuum Einstein field equations (with particular gauge) and Fierz-Pauli action are presented.

Keywords: stress-energy tensor, linearised gravity, conserved quantities, Fierz-Pauli action

Abstrakt. Systém Cadabra je použit pro implementaci Fockovy metody na hledání zachovávajících se Lorentz kovariantních komplexů energie-hybnosti. Nejprve je vypracována metoda pro rovnice pole obsahující lineárně druhé derivace a poté je zobecněna pro pohybové rovnice zahrnující nederivované členy. Prezentovány jsou výsledky pro linearizové Einsteinovy rovnice ve vakuu (v konkrétní kalibraci) a pro Fierz-Pauliho akci.

Klíčová slova: tenzor energie-hybnosti, linearizovaná gravitace, zachovávající se veličiny, Fierz-Pauliho akce

1 Introduction

The article is organised as follows. At the beginning the Fock method for finding conserved stress energy tensors for massless[1] equation of motion for rank two covariant tensor fields is presented. The important step in simplifying computations is to restrict ourselves only to Lorentz covariant expressions which is no serious limitation for obtained results. The method is then generalised for equations of motion containing non-derivative terms which is the case of Fierz-Pauli action of massive gravity. The crucial part plays the usage of symbolic tensor manipulation software Cadabra. Finally, the resulting tensors for field theory of linearized gravity and of massive gravity are presented.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS13/217/OHK4/3T/14

[1]Massless in the meaning that equation of motion lacks the non-derivative terms of field variable.

2 Fock formulation

We would like to find stress-energy complex[2] T^{ij} in the form of $T^{ij} = t^{ijabcrst}h_{ab,c}h_{rs,t}$, i.e. quadratic in the first derivatives of field, with $t^{ijabcmno}$ being constant coefficients symmetric in (a,b) and (m,n) and invariant with respect to the swap of triples (a,b,c) and (r,s,t), that is conserved in the sense of equation $T^{ij}_{,j} = 0$, whenever equations of motion of the form

$$P^A = p^{Amnop}h_{mn,op} = 0 \tag{1}$$

are satisfied (A being arbitrary multiindex). We can write our demands as a single condition using Langrange multipliers as

$$T^{ij}_{,j} = \lambda^i_A P^A, \tag{2}$$

i.e. it is required that divergence of stress-energy tensor is a linear combination of field equations. Coefficients λ^i_A can vary over the spacetime, so they are generally functions of spacetime point – $\lambda^i_A(x)$. As $T^{ij}_{,j}$ has the form $u^{iabcmnop}h_{ab,c}h_{mn,op}$, with $u^\bullet$-s given simply as $u^{iabcmnop} = 2t^{ipabcmno}$, the Lagrange multipliers will be of the following structure: $\lambda^i_A = L^{iabc}_A h_{ab,c}$.

Writing master equation (2) in terms of coefficients $u^\bullet$ and $L^\bullet$ we have

$$(u^{iabcmnop} - L^{iabc}_A p^{Amnop})h_{ab,c}h_{mn,op} = 0. \tag{3}$$

This has to be satisfied for every field h_{ab}, therefore terms $h_{ab,c}h_{mn,op}$ can be considered as linearly independent (taking into account symmetry in indices (a,b), (m,n) and (o,p)) and thus symmetrized coefficients has to be zero identically

$$u^{i(ab)c(mn)(op)} - L^{i(ab)c}_A p^{A(mn)(op)} = 0. \tag{4}$$

Our task is to eliminate Lagrange multipliers L^{iabc}_A using known coefficients p^{Amnop} to find constants $t^{ijabcmno}$ (easily recovered from $u^{iabcmnop}$).

3 Covariant formulation

We can take a great advantage when we consider only Lorentz covariant expressions. This step has the consequence of greatly reducing the number of unknowns in the master equation (2). Let's take a look at the different types of terms and their covariant contributions to the master equation.

3.1 General form of stress energy tensor

We would like to seek the most general form of the second rank Lorentz covariant tensor constructed from quadratic first derivatives of metric. So we need to find all different (with respect to index symmetries) contractions to the term $h_{ab,c}h_{de,f}$ to produce tensor of rank two. Raising and lowering indices is permitted via Minkowski metric. The resulting terms are listed in table 1.

[2] As it need not to transform as a tensor with respect to arbitrary coordinate change.

$\mathbf{B_i}$	$\mathbf{A_i}$	**Term**	**Abbr.**	$\mathbf{B_i}$	$\mathbf{A_i}$	**Term**	**Abbr.**
B_{12}	A_1	$h_{ik,a}h^{ab}{}_{,b}$		B_{10}	A_{11}	$h_{ka,}{}^{a}h_{b}{}^{b}{}_{,i}$	$h_{ka,}{}^{a}h_{,i}$
B_7	A_2	$h_{ik,a}h^{b}{}_{b,}{}^{a}$	$h_{ik,a}h^{,a}$	B_{19}	A_{12}	$h_{ia,b}h^{ab}{}_{,k}$	
B_{15}	A_3	$h_{ia,}{}^{a}h_{kb,}{}^{b}$		B_{18}	A_{13}	$h_{ka,b}h^{ab}{}_{,i}$	
B_{17}	A_4	$h_{ia,b}h_{k}{}^{a,b}$		B_6	A_{14}	$h^{a}{}_{a,i}h^{b}{}_{b,k}$	$h_{,i}h_{,k}$
B_{20}	A_5	$h_{ia,b}h_{k}{}^{b,a}$		B_{16}	A_{15}	$h_{ab,i}h^{ab}{}_{,k}$	
B_{13}	A_6	$h_{ia,k}h^{ab}{}_{,b}$		B_2	A_{16}	$\eta_{ik}h^{a}{}_{a,b}h^{bc}{}_{,c}$	$\eta_{ik}h_{,b}h^{bc}{}_{,c}$
B_{14}	A_7	$h_{ka,i}h^{ab}{}_{,b}$		B_3	A_{17}	$\eta_{ik}h_{ab,}{}^{a}h^{bc}{}_{,c}$	
B_8	A_8	$h_{ia,k}h_{b}{}^{b,a}$	$h_{ia,k}h^{,a}$	B_1	A_{18}	$\eta_{ik}h^{a}{}_{a,b}h^{c}{}_{c,}{}^{b}$	$\eta_{ik}h_{,b}h^{,b}$
B_9	A_9	$h_{ka,i}h_{b}{}^{b,a}$	$h_{ka,i}h^{,a}$	$B_{5/4}$	A_{19}	$\eta_{ik}h_{ab,c}h^{ab,c}$	
B_{11}	A_{10}	$h_{ia,}{}^{a}h_{b}{}^{b}{}_{,k}$	$h_{ia,}{}^{a}h_{,k}$	$B_{4/5}$	A_{20}	$\eta_{ik}h_{ab,c}h^{bc,a}$	

Table 1: List of possible contractions in stress-energy tensor. A_i-s denote coefficients of linear combination used in this paper, B_i-s are coefficients used in [1].

Therefore, the most general form of Lorentz covariant stress-energy tensor quadratic in first derivatives of metric is of twenty parameters as follows

$$\begin{aligned}T_{ik} = {} & A_1 h_{ik,a}h^{ab}{}_{,b} + A_2 h_{ik,a}h^{,a} + A_3 h_{ia,}{}^{a}h_{kb,}{}^{b} + A_4 h_{ia,b}h_{k}{}^{a,b} + A_5 h_{ia,b}h_{k}{}^{b,a} + A_6 h_{ia,k}h^{ab}{}_{,b} + \\ & A_7 h_{ka,i}h^{ab}{}_{,b} + A_8 h_{ia,k}h^{,a} + A_9 h_{ka,i}h^{,a} + A_{10} h_{ia,}{}^{a}h_{,k} + A_{11} h_{ka,}{}^{a}h_{,i} + A_{12} h_{ia,b}h^{ab}{}_{,k} + \\ & A_{13} h_{ka,b}h^{ab}{}_{,i} + A_{14} h_{,i}h_{,k} + A_{15} h_{ab,i}h^{ab}{}_{,k} + A_{16}\eta_{ik}h_{,b}h^{bc}{}_{,c} + A_{17}\eta_{ik}h_{ab,}{}^{a}h^{bc}{}_{,c} + \\ & A_{18}\eta_{ik}h_{,b}h^{,b} + A_{19}\eta_{ik}h_{ab,c}h^{ab,c} + A_{20}\eta_{ik}h_{ab,c}h^{bc,a}. \end{aligned} \tag{5}$$

We will denote term standing at coefficient A_α as $\mathcal{A}_{\alpha\, ik}$ (or, where the indices are not important, as $\mathcal{A}_\alpha$), then the tensor and its divergence can be written in the form

$$T_{ik} = \sum_{\alpha=1}^{20} A_\alpha \mathcal{A}_{\alpha\, ik}, \qquad T_{ik,}{}^{k} = \sum_{\alpha=1}^{20} A_\alpha \mathcal{A}_{\alpha\, ik}{}^{,k}. \tag{6}$$

3.2 Two indices equation of motion $P_{ab} = 0$

We will restrict ourselves here to the symmetric equations P_{ab} containing linearly field h_{ab} with second derivatives[3]. The contribution to the master equation has the form $\lambda_i^{rsqab} h_{rs,q} P_{ab}$. The demand of Lorentz covariance leaves us with six possibilities listed below in table 2, naturally we consider P_{ab} to be Lorentz tensor as well and we take into account all index symmetries. These terms will be denoted as $\mathcal{L}_{\alpha\, i}$ and corresponding Lagrange multipliers λ_α.

3.3 One index equation $P_a = 0$

We consider only equations linearly consisting of first derivatives of the field. Contributions to the master equation is $\mu_i^{mnopa} h_{mn,op} P_a$ leading to six covariant terms, labeled $\mathcal{U}_{\alpha\, i}$ with corresponding Lagrange multipliers μ_α, see table 3.

[3]E.g. it is the case of linearised gravity as the Einstein or the Ricci tensor contains linearly the second derivatives of metric perturbation.

Term	$\lambda_{\mathbf{i}}^{\mathbf{rsqab}}$	Explicitly
$\mathcal{L}_1$	$\lambda_1\eta^{rs}\eta^{qb}\delta_i^a$	$h_b{}^{b,a}P_{ia}$
$\mathcal{L}_2$	$\lambda_2\eta^{rs}\eta^{sq}\delta_i^a$	$h^{ab}{}_{,b}P_{ia}$
$\mathcal{L}_3$	$\lambda_3\delta_i^r\eta^{sq}\eta^{ab}$	$h_{ib}{}^{,b}P_a^a$
$\mathcal{L}_4$	$\lambda_4\eta^{rs}\delta_i^q\eta^{ab}$	$h_b{}^b{}_{,i}P_a^a$
$\mathcal{L}_5$	$\lambda_5\delta_i^r\eta^{sa}\eta^{qb}$	$h_i{}^{a,b}P_{ab}$
$\mathcal{L}_6$	$\lambda_6\eta^{ra}\eta^{sb}\delta_i^r$	$h^{ab}{}_{,i}P_{ab}$

Table 2: The list of possible covariant terms for equation $P_{ab} = 0$.

Term	$\mu_{\mathbf{i}}^{\mathbf{mnopa}}$	Explicitly
$\mathcal{U}_1$	$\mu_1\eta^{mn}\eta^{op}\delta_i^a$	$h_a{}^a{}_{,b}{}^bP_i$
$\mathcal{U}_2$	$\mu_2\eta^{mo}\eta^{np}\delta_i^a$	$h_{ab}{}^{,ab}P_i$
$\mathcal{U}_3$	$\mu_3\delta_i^m\eta^{no}\eta^{pa}$	$h_{ia}{}^{,ab}P_b$
$\mathcal{U}_4$	$\mu_4\delta_i^m\eta^{na}\eta^{op}$	$h_{ia,b}{}^bP^a$
$\mathcal{U}_5$	$\mu_5\eta^{mn}\delta_i^o\eta^{pa}$	$h_a{}^a{}_{,i}{}^bP_b$
$\mathcal{U}_6$	$\mu_6\eta^{mp}\eta^{na}\delta_i^o$	$h_{ab,i}{}^aP^b$

Table 3: The list of possible covariant terms for equation $P_a = 0$.

3.4 Scalar equation $P = 0$

Our linearity condition essentially restricts us to the only possible choices: $P = h_a{}^a$ or $P = h_a{}^a{}_{,b}{}^b$. Nevertheless in the master equation we have $\kappa_i^{qrs}h_{rs,q}P$ leading to two covariant terms, named $\mathcal{K}_{\alpha\,i}$ with multipliers κ_α, shown in table 4.

Term	$\kappa_{\mathbf{i}}^{\mathbf{qrs}}$	Explicitly
$\mathcal{K}_1$	$\kappa_1\delta_i^r\eta^{sq}$	$h_{ia}{}^{,a}P$
$\mathcal{K}_2$	$\kappa_2\eta^{rs}\delta_i^q$	$h_a{}^a{}_{,i}P$

Table 4: The list of possible covariant terms for equation $P = 0$.

3.5 Covariant form of master equation

The most general form of master equation is the following

$$\sum_{\alpha=1}^{20} A_\alpha\mathcal{A}_{\alpha\,ik}{}^{,k} - \sum_{\beta=1}^{6}\lambda_\beta\mathcal{L}_{\beta\,i} - \sum_{\beta=1}^{6}\mu_\beta\mathcal{U}_{\beta\,i} - \sum_{\beta=1}^{2}\kappa_\beta\mathcal{K}_{\beta\,i} = 0, \tag{7}$$

however there can be fewer terms depending on the type(s) of an equation(s) of motion used. Now we have equation for unknowns A_α, λ_β, μ_β and κ_β which has to hold for every field h_{ij}. Essentially we rewrite it in the form of (3) and because of the linear independence of the field terms $h_{ab,c}h_{mn,op}$[4] the linear equations for unknown variables are extracted. This extraction is done by Cadabra software as described in section 5.

4 Generalization of motion equation

In previous sections we considered equations of motion containing solely and linearly second derivatives of the field.

Now we will modify the procedure allowing non-differentiated field to be present linearly in equations of motion as in for example $\partial_a\partial^a h_{rs} + m^2 h_{rs} = 0$.

[4]But now appearing only as Lorentz covariant terms, hence at greatly reduced numbers!

At first, let's focus on the stress-energy tensor. If our demand is the tensor to be consisted of quadratic terms containing at most first derivatives of metric tensor then we should include also terms of the form $h_{ab}h_{cd}$ and $h_{ab}h_{cd,e}$. We can omit the latter one, because it is not possible to contract it to form second rank tensor. On the contrary the term free of derivatives produces four more Lorentz covariant terms, see table 5. The generalised tensor will be of the form

$$T_{ik} = \sum_{\alpha=1}^{20} A_\alpha \mathcal{A}_{\alpha\, ik} + \sum_{\beta=1}^{4} C_\beta \mathcal{C}_{\beta\, ik},$$

where we denoted terms corresponding to coefficient C_β as $\mathcal{C}_{\beta\, ik}$.

$\mathbf{C_i}$	**Term**	**Abbr.**
C_1	$h_{ik}h^a_a$	$h_{ik}h$
C_2	$h_{ia}h^a_k$	
C_3	$\eta_{ik}h^a_a h^b_b$	$\eta_{ik}h^2$
C_4	$\eta_{ik}h_{ab}h^{ab}$	

Table 5: Contractions of non-differentiated terms.

Term	$\nu_{\mathbf{i}}^{\mathbf{mna}}$	**Term**
$\mathcal{V}_1$	$\nu_1\eta^{mn}\delta^a_i$	$h_a{}^a P_i$
$\mathcal{V}_2$	$\nu_2\delta^m_i\eta^{na}$	$h_i{}^a P_a$

Table 6: The list of additional covariant terms for equation $P_a = 0$.

Why didn't we consider these $\mathcal{C}_\beta$-terms in the previous section? Because of the Fock procedure, they would vanish anyway – the equations of motion consist of only the second derivatives and choosing whatever form of Lagrange multipliers $\lambda^\bullet$ will never produce terms $h_{ab}h_{cd,e}$ occuring in $T^{ik}_{,k}$.

Let's have a look at the equation of motion. Now it contains also non-differentiated terms, so we need to modify relation (1) into

$$P^A = p_2^{Amnop}h_{mn,op} + p_0^{Amn}h_{mn} = 0. \tag{8}$$

What is then the form of Lagrange multipliers in the case of our new equation of motion and new stress-energy tensor? To answer this look at the terms occurring in the master equation (2). On the left side there are terms of type $h_{rs,t}h_{mn}$ and $h_{rs,t}h_{mn,op}$. On the right side we have from the equation of motion terms h_{mn} and $h_{mn,op}$. Consequently, the only needed and the only possible choice is to consider $\lambda^i_A = \lambda^{irst}_A h_{rs,t}$. The different choices wouldn't find pairing partners on the left side of master equation and would be condemned to vanish.

In the segment of Lagrange multipliers we need to reconsider only equations of the type $P_a = 0$ where in the case of presence of non-differentiated terms we get additional contribution to the master equation – $\nu_i^{mna}h_{mn}P_a$ leading to two more covariant terms, labeled $\mathcal{V}_\alpha$ with multipliers ν_α, see table 6. In the master equation there appears additional term $\sum_{\alpha=1}^{2}\nu_\alpha\mathcal{V}_{\alpha\, i}$.

5 Cadabra

With Cadabra software it is extremely easy to obtain equations for coefficients A_i (and C_i, λ_i,...). As was already said, it is needed to extract coefficients standing at the distinct

covariant terms. This is rather tedious task doing by hand because of the different naming of dummy indices, symmetry of tensor h, raising and lowering indices and last but not least the overwhelming number of terms (even though massively reduced by Lorentz covariance). Cadabra is asked to convert each term into its canonical form by the following set of commands:

```
@distribute!(%):                @prodsort!(%):
@eliminate_metric!(%):          @canonicalise!(%):
@eliminate_kr!(%):              @rename_dummies!(%);
```

The concrete canonical appearance of every term depends on the internal working of Cadabra algorithms and the way of storing tensorial structures. Grouping the canonicalized terms and collecting their coefficients is done with the command

```
@factor_in!(%){ ... list of coefficients to collect ... };
```

In order to satisfy master equation, it is necessary each collected group of coefficients to vanish, hence we get the set of linear equation which are fairly easy to solve (by hand or by arbitrary symbolic manipulation software such as Mathematica).

6 Some results

In this section a few examples of obtained results are presented. We begin with strongly conserved complex, continue with complex of linearized gravity in arbitrary gauge and also in particular gauge and end with conserved tensor for Fierz-Pauli action.

6.1 Strong conservation $T^{ik}_{\ ,k} = 0$

If we impose condition of vanishing divergence for arbitrary (gravitational) field, we obtain one parameter family with all constants A_i vanishing except for $\alpha = A_7 = -A_{13} = -2A_{17} = 2A_{20}$ and the resulting tensor (which is not symmetrical) is

$$T_{ik} = \alpha \left(h_{ka,i} h^{ab}{}_{,b} - h_{ka,b} h^{ab}{}_{,i} - \frac{1}{2} \eta_{ik} h_{ab,}{}^{a} h^{bc}{}_{,c} + \frac{1}{2} \eta_{ik} h_{ab,c} h^{bc,a} \right). \tag{9}$$

6.2 Linearised vacuum Einstein equations $T^{ik}_{\ ,k} = \lambda^{irs} R_{rs}$

Now we allow the divergence to be linear combination of linearized vacuum Einstein field equations. The resulting tensor depends on four parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and the dependence of A_i-s on α_j-s can be seen below.

$$\begin{aligned}
\alpha_1 = A_1 = -A_3 = A_4 = A_{10} = -A_{12}, \qquad & \alpha_2 = A_2, \\
\alpha_3 = A_7 = -2A_{17}, \qquad & \alpha_4 = A_9 = A_{15} = -2A_{19}, \\
0 = A_5 = A_6, \qquad & \alpha_1 + \alpha_2 - \alpha_4 = -A_{11} = A_{14} = A_{16}, \\
-\alpha_1 - \alpha_2 = A_8, \qquad & -\alpha_3 - 2\alpha_4 = A_{13}, \\
-\alpha_1 - \alpha_2 + \frac{1}{2}\alpha_4 = A_{18}, \qquad & \frac{1}{2}\alpha_3 + \alpha_4 = A_{20}.
\end{aligned} \tag{10}$$

And the explicit expression is

$$\begin{aligned}T_{ik} = &\alpha_1 \left(h_{ik,a}h^{ab}{}_{,b} - h_{ia,}{}^{a}h_{kb,}{}^{b} + h_{ia,b}h_{k}{}^{a,b} - h_{ia,k}h^{,a} + h_{ia,}{}^{a}h_{,k} - h_{ka,}{}^{a}h_{,i} - h_{ia,b}h^{ab}{}_{,k} + \right.\\ &\left. h_{,i}h_{,k} + \eta_{ik}h_{,b}h^{bc}{}_{,c} - \eta_{ik}h_{,b}h^{,b}\right) + \\ &\alpha_2 \left(h_{ik,a}h^{,a} - h_{ia,k}h^{,a} - h_{ka,}{}^{a}h_{,i} + h_{,i}h_{,k} + \eta_{ik}h_{,b}h^{bc}{}_{,c} - \eta_{ik}h_{,b}h^{,b}\right) + \\ &\alpha_3 \left(h_{ka,i}h^{ab}{}_{,b} - h_{ka,b}h^{ab}{}_{,i} - \frac{1}{2}\eta_{ik}h_{ab,}{}^{a}h^{bc}{}_{,c} + \frac{1}{2}\eta_{ik}h_{ab,c}h^{bc,a}\right) + \\ &\alpha_4 \left(h_{ka,i}h^{,a} + h_{ka,}{}^{a}h_{,i} - 2h_{ka,b}h^{ab}{}_{,i} - h_{,i}h_{,k} + h_{ab,i}h^{ab}{}_{,k} - \eta_{ik}h_{,b}h^{bc}{}_{,c} + \right.\\ &\left. \frac{1}{2}\eta_{ik}h_{,b}h^{,b} - \frac{1}{2}\eta_{ik}h_{ab,c}h^{ab,c} + \eta_{ik}h_{ab,c}h^{bc,a}\right). \end{aligned} \tag{11}$$

Now we want to find symmetric tensors – this condition will impose some restrictions on coefficients α_i. So the demand is $T_{ik} = T_{ki}$ or $T_{[ik]} = 0$. Omitting terms in T_{ik} which are already symmetric itself, we are left with

$$\begin{aligned}\tilde{T}_{ik} = &\alpha_1 \left(-h_{ia,k}h^{,a} + h_{ia,}{}^{a}h_{,k} - h_{ka,}{}^{a}h_{,i} - h_{ia,b}h^{ab}{}_{,k}\right)\\ &\alpha_2 \left(-h_{ia,k}h^{,a} - h_{ka,}{}^{a}h_{,i}\right)\\ &\alpha_3 \left(h_{ka,i}h^{ab}{}_{,b} - h_{ka,b}h^{ab}{}_{,i}\right)\\ &\alpha_4 \left(h_{ka,i}h^{,a} + h_{ka,}{}^{a}h_{,i} - 2h_{ka,b}h^{ab}{}_{,i}\right). \end{aligned} \tag{12}$$

For this leftover to be symmetric we obtain the conditions

$$-\alpha_1 - \alpha_2 + \alpha_4 = \alpha_1, \qquad -\alpha_1 - \alpha_2 = \alpha_4, \qquad -\alpha_1 = -\alpha_3 - 2\alpha_4, \tag{13}$$

with the one parameter solution (being merely a multiplicative constant) $\alpha_1 = 2\alpha$, $\alpha_2 = -3\alpha$, $\alpha_3 = 0$, $\alpha_4 = \alpha$.

To obtain correspondence with parametrization $(\beta_1, \beta_2, \beta_3, \beta_4)$ used in [1][5] we need the following linear transformation

$$\alpha_1 = -\beta_2, \qquad \alpha_2 = \beta_1 + 2\beta_3 + \beta_2, \qquad \alpha_3 = -2\beta_4, \qquad \alpha_4 = 2\beta_3. \tag{14}$$

We can also consider "complete" linearized Einstein equations $G_{rs} = 0$ and solve problem with defining equation $T^{ik}_{,k} = \lambda^{irs}G_{rs}$. The results won't change, because the following identity holds $\lambda^{irs}G_{rs} = \lambda'^{icd}R_{cd}$ with $\lambda'^{icd} = \lambda^{irs}\left(\delta^c_r\delta^d_s - \frac{1}{2}\eta_{rs}\eta^{cd}\right)$, i.e. the only change is the linear tranformation of Lagrange coefficients (it can be easily checked that the transformation is regular).

6.3 Linearized gravity with gauge condition $h^{ab}{}_{,b} = \frac{1}{2}h^{,a}$

In order to find conserved tensor for gravitational field satisfying gauge condition $h^{ab}_{,b} = \frac{1}{2}h^{,a}$ (which is gauge condition required in [2]) we can follow this procedure – impose condition (by converting all terms of form $h^{ab}_{,b}$ into $\frac{1}{2}h^{,a}$) on the general stress-energy

[5] In paper [1] parameters are labeled $\overset{i}{\alpha}$ instead of β_i – this relabelling is used to avoid confusion in notation used here.

tensor and general Einstein equations and plunge the resulting quantities into our F.-C.[6] machinery.

At first we get the following term equalities in stress-energy tensor

$$2\mathcal{A}_1 = \mathcal{A}_2, \quad 4\mathcal{A}_3 = 2\mathcal{A}_{10} = 2\mathcal{A}_{11} = \mathcal{A}_{14}, \quad 2\mathcal{A}_6 = \mathcal{A}_8, \quad 2\mathcal{A}_7 = \mathcal{A}_9, \quad 2\mathcal{A}_{16} = 4\mathcal{A}_{17} = \mathcal{A}_{18}. \tag{15}$$

Because of these equalities we simply make these redundant terms vanish via corresponding coefficients A_i, i.e. $A_1 = A_3 = A_{10} = A_{11} = A_6 = A_7 = A_{16} = A_{17} = 0$. Additionally we need to apply the gauge condition on divergence of stress-energy tensor once again – divergence produces terms of the type $h_{ac}{}^{,cb}$ which can be further converted into $\frac{1}{2}h_{,a}{}^{b}$.

Ricci tensor reduces simply into $2R_{ab} = -h_{ab,c}{}^{c} = 0$ and Ricci scalar into $2R = -h_{,c}{}^{c}$, hence the Einstein tensor is $2G_{ab} = -h_{ab,c}{}^{c} + \frac{1}{2}\eta_{ab}h_{,c}{}^{c}$. As a result of F.-C. procedure, we get five-parameter tensor

$$\begin{aligned} &\alpha_1 = A_2, && \alpha_2 = A_4 = -A_{12}, && \alpha_3 = A_9 = -\frac{1}{2}A_{13} = A_{20}, && \alpha_4 = A_{14}, \\ &\alpha_5 = A_{15} = -2A_{19}, && A_8 = -\alpha_1 - \frac{1}{2}\alpha_2, && A_{18} = -\frac{1}{4}\left(\alpha_1 + \alpha_3 + 2\alpha_4\right). \end{aligned} \tag{16}$$

Explicitly

$$\begin{aligned} T_{ik} = {} & \alpha_1\left(h_{ik,a}h^{,a} - h_{ia,k}h^{,a} - \frac{1}{4}\eta_{ik}h_{,b}h^{,b}\right) \\ & \alpha_2\left(h_{ia,b}h_{k}{}^{a,b} - h_{ia,b}h^{ab}{}_{,k} - \frac{1}{2}h_{ia,k}h^{,a}\right) \\ & \alpha_3\left(h_{ka,i}h^{,a} - 2h_{ka,b}h^{ab}{}_{,i} + \eta_{ik}h_{ab,c}h^{bc,a} - \frac{1}{4}\eta_{ik}h_{,b}h^{,b}\right) \\ & \alpha_4\left(h_{,i}h_{,k} - \frac{1}{2}\eta_{ik}h_{,b}h^{,b}\right) \\ & \alpha_5\left(h_{ab,i}h^{ab}{}_{,k} - \frac{1}{2}\eta_{ik}h_{ab,c}h^{ab,c}\right). \end{aligned} \tag{17}$$

Butcher's tensor is obtained after choosing $\alpha_1 = 0$, $\alpha_2 = 0$, $\alpha_3 = 0$, $\alpha_4 = -\frac{1}{8}$, $\alpha_5 = \frac{1}{4}$. In this case conditions of symmetry are as follows

$$-\frac{1}{2}\alpha_2 - \alpha_1 = \alpha_3, \qquad -\alpha_2 = -2\alpha_3, \tag{18}$$

with the result $\alpha_1 = 2\alpha$, $\alpha_2 = -2\alpha$, $\alpha_3 = \alpha$ and α_4, α_5 arbitrary.

6.4 Fierz-Pauli action

We will start with Fierz-Pauli action describing linearised massive gravity or massive spin 2 particle (see [3])

$$S_{FP} = \int -\frac{1}{2}\partial_k h_{ij}\partial^k h^{ij} + \partial_i h_{jk}\partial^j h^{ik} - \partial_i h^{ij}\partial_j h + \frac{1}{2}\partial_k h\partial^k h - \frac{1}{2}m^2\left(h_{ij}h^{ij} - h^2\right) d^4x. \tag{19}$$

[6] F.-C. a.k.a. Fock-Cadabra

Equations of motion are then obtained as variational derivative of action with respect to the field variables h_{ij}

$$\frac{\delta S}{\delta h^{ij}} = \partial_k{}^k h_{ij} - \partial_{ki} h_i^k - \partial_{kj} h_i^k + \eta_{ij}\partial_{k\sigma} h^{k\sigma} + \partial_{ij} h - \eta_{ij}\partial_k{}^k h - m^2\left(h_{ij} - \eta_{ij} h\right) = 0. \quad (20)$$

It can be easily shown that equations (20) are equivalent to the following set of equations

$$\left(\partial_k{}^k - m^2\right) h_{ij} = 0, \qquad \partial^i h_{ij} = 0, \qquad h = 0. \quad (21)$$

6.4.1 Results using equation (20)

Table showing nonvanishing coefficients follows.

$$\begin{aligned}
\alpha_1 =& A_7 = -2A_{17},\\
\alpha_2 =& A_9 = A_{11} = -A_{14} = A_{15} = -A_{16} = 2A_{18} = -2A_{19} = \frac{2}{m^2}A_{23} = -\frac{2}{m^2}A_{24},\\
A_{13} =& -\alpha_1 - 2\alpha_2,\\
A_{20} =& \frac{1}{2}\alpha_1 + \alpha_2.
\end{aligned} \quad (22)$$

Explicitly

$$\begin{aligned}
T_{ik} =& \alpha_1\left(h_{ka,i}h^{ab}{}_{,b} - h_{ka,b}h^{ab}{}_{,i} - \frac{1}{2}\eta_{ik}h_{ab,}{}^a h^{bc}{}_{,c} + \frac{1}{2}\eta_{ik}h_{ab,c}h^{bc,a}\right) +\\
&\alpha_2\left(h_{ka,i}h^{,a} + h_{ka,}{}^a h_{,i} - 2h_{ka,b}h^{ab}{}_{,i} - h_{,i}h_{,k} + h_{ab,i}h^{ab}{}_{,k} - \eta_{ik}h_{,b}h^{bc}{}_{,c} +\right.\\
&\left.\frac{1}{2}\eta_{ik}h_{,b}h^{,b} - \frac{1}{2}\eta_{ik}h_{ab,c}h^{ab,c} + \eta_{ik}h_{ab,c}h^{bc,a} + \frac{1}{2}m^2\eta_{ik}h^2 - \frac{1}{2}m^2\eta_{ik}h_{ab}h^{ab}\right).
\end{aligned} \quad (23)$$

This tensor cannot be made symmetric for any choice of parameters. However we can additionaly apply the second and the third equation from the set (21) (which are linearly independent of the original equation (20)) and get the tensor

$$\begin{aligned}
\tilde{T}_{ik} =& \alpha_1\left(-h_{ka,b}h^{ab}{}_{,i} + \frac{1}{2}\eta_{ik}h_{ab,c}h^{bc,a}\right) +\\
&\alpha_2\left(-2h_{ka,b}h^{ab}{}_{,i} + h_{ab,i}h^{ab}{}_{,k} - \frac{1}{2}\eta_{ik}h_{ab,c}h^{ab,c} + \eta_{ik}h_{ab,c}h^{bc,a} - \frac{1}{2}m^2\eta_{ik}h_{ab}h^{ab}\right),
\end{aligned} \quad (24)$$

which can be made symmetrical by the choice $\alpha = \alpha_2 = -\frac{1}{2}\alpha_1$ obtaining unique (up to a multiplicative constant) tensor

$$\bar{T}_{ik} = \alpha\left(h_{ab,i}h^{ab}{}_{,k} - \frac{1}{2}\eta_{ik}h_{ab,c}h^{ab,c} - \frac{1}{2}m^2\eta_{ik}h_{ab}h^{ab}\right). \quad (25)$$

6.4.2 Results using equations (21)

We use the same procedure as in subsection (6.3), i.e. at first equations $h^{ab}{}_{,b} = 0$ and $h = 0$ are applied on stress-energy tensor – only nonvanishing terms are then $\mathcal{A}_4$, $\mathcal{A}_5$,

$\mathcal{A}_{12}$, $\mathcal{A}_{13}$, $\mathcal{A}_{15}$, $\mathcal{A}_{19}$, $\mathcal{A}_{20}$, $\mathcal{C}_2$ and $\mathcal{C}_4$. Of course, the vanished terms can be arbitrarily add to the resulting tensor, since their divergence also vanishes, but if it is considered only on-shell situation then there is really no benefit of adding them.

The result of F.-C. procedure is a three-parameter tensor,

$$\begin{aligned} \alpha_1 &= A_{12} = -A_4 = -\frac{1}{m^2}C_2, \\ \alpha_2 &= A_{13} = -2A_{20}, \\ \alpha_3 &= A_{15} = -2A_{19} = -\frac{2}{m^2}C_4; \end{aligned} \tag{26}$$

explicitly

$$\begin{aligned} T_{ik} = &\alpha_1 \left(-h_{ia,b}h_k{}^{a,b} + h_{ia,b}h^{ab}{}_{,k} - m^2 h_{ia}h_k{}^a\right) + \\ &\alpha_2 \left(h_{ka,b}h^{ab}{}_{,i} - \frac{1}{2}\eta_{ik}h_{ab,c}h^{bc,a}\right) + \\ &\alpha_3 \left(h_{ab,i}h^{ab}{}_{,k} - \frac{1}{2}\eta_{ik}h_{ab,c}h^{ab,c} - \frac{1}{2}m^2\eta_{ik}h_{ab}h^{ab}\right). \end{aligned} \tag{27}$$

Condition of symmetry yields $\alpha_1 = \alpha_2$.

7 Conclusion

We presented a method for finding conserved Lorentz covariant stress-energy complexes for a certain class of equations of motion. The result for linearized gravity presented in [1] was reproduced. The requirement of particular gauge in [2] lead to a wider class of complexes, unlike the unique result obtained by specific procedure in [2]. Finally, the generalization of F.-C. method lead to computing of complexes for Fierz-Pauli action, one of the model of massive gravity.

References

[1] J. Bičák. *On the Question of the Uniqueness of the Energy-Momentum Complex in the Special and General Theory of Relativity.* Czech J. Phys. B **15** (1965) 81-94.

[2] L. M. Butcher, M. Hobson, A. Lasenby. *Localising the Energy and Momentum of Linear Gravity.* Phys. Rev. D **82**, 104040 (2010).

[3] K. Hinterbichler. *Theoretical aspects of massive gravity.* Rev. Mod. Phys. **84**, 671-710 (2012).

[4] K. Peeters. *Introducing Cadabra: a symbolic computer algebra system for field theory problems.* hep-th/0701238.

[5] K. Peeters. *Symbolic field theory with Cadabra.* Computeralgebra Rundbrief 41 (2007) 16.

Feature Collection for Source Code Classification and Pattern Recognition*

Josef Smolka

3rd year of PGS, email: smolkjos@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. The paper presents method of feature collection for the purpose of classification and pattern recognition in source codes of software projects. Design of a collector component is introduced, and an example implementation of a specific feature for recognition of Factory design pattern is given.

Keywords: Java, abstract syntax tree, XQuery, feature, classification

Abstrakt. Článek uvádí metodu pro sběr příznaků ze zdrojového kódu softwarových projektů pro účely klasifikace a rozpoznávání vzorů. Je prezentován návrh komponenty implementující samotný sběr dat a ukázka příznaku charakterizjícího návrhový vzor Továrna.

Klíčová slova: Java, strom abstraktní syntaxe, XQuery, příznaky, klasifikace

1 Introduction

The main domain of classification and recognition lays in an image processing, where pattern recognition of common objects on camera pictures is the most common task of this branch of machine learning. The idea is to emulate a process of a person, who looks at an unknown picture and instantly recognises and classifies various objects, other people and all the things that the person encounter earlier. Similar process takes places when an experienced software engineer looks at an unfamiliar source code. Such person orients itself in the code by way of identifying familiar structures or patterns, not by survey of functionality. This lead to a question, why not to apply principles from image processing to source code patterns recognition. Traditional approach to pattern recognition in a source code is by the means of graphs isomorphism and similarity scoring [1]. This is natural as the source code can be easily represented as a tree (abstract syntax tree) or as a graph (abstract semantic graph). This paper, on the other hand, deals with an application of more traditional classification methods like discriminant analysis, k nearest neighbours, naive bayes, neural networks and support vector machines, and specifically with the method of feature collection to support this methods.

*This paper was supported by grants SGS11/167/OHK4/3T/14 and LA08015.

2 Feature Space and Classes

Feature in machine learning is a measurable property which describes some quality of observed phenomena. Features used by the methods mentioned in the introduction have a numeric form and are typically used as a whole set called feature vector. Features in the feature vector should be independent and should discriminate the recognized patterns or classes from each other. Specifically in image recognition, features should ideally be invariant to translation and rotation, as a cat is a cat whether it is climbing a tree or laying on a grass.

Source code, and specifically source code of programs written in object oriented programming language, offers certainly many properties which can be measured: code length in the number of lines, class lengths, method lengths, number of methods, number of attributes, and so on. But these primitive features are not suited for recognizing such patterns as UML class stereotypes Focus, Auxiliary, Type, Utility, Entity, Boundary and Control [2]. A class (in the sense of data type) designated by stereotype Focus is meant to hold the core logic of the component or control flow of auxiliary classes. On the other hand, Auxiliary class takes a role of a supporting class for the fundamental core represented by the Focus classes and implements secondary logic or control flow. These classes are usually connected to Focus class by dependency relationship. Type classes represent domain objects and Utility classes are a special type of auxiliary classes that contains only static attributes and operations. Entity classes represent some, usually persistent, business or system information. Boundary class is a system boundary with its neighborhood, like user interface or system service. Finally, the Control class is an object used to model system or user workflow or some coordination in a system behaviour [2]. Another non-trivial patterns, which can be recognized in source code, are well-known design patterns like Factory, Proxy, Builder and others [3]. Completely different level represents recognition of enterprise integration patterns [4] in projects of large information systems.

To recognize such complex patterns, more sophisticated features have to be designed. For example, to support recognition of Factory design pattern, feature like this is required: a ratio of public methods that contain instantiation and return the result of this instantiation, where call of a member method (public or private) that returns a result of instantiation count as an instantiation, and where result of instantiation is not of the same type as covering class, and where returning object is not stored in covering class attribute [5]. To support recognition of Builder design pattern feature like this can help a lot: ratio of non-primitive non-static attributes, which are instantiated in within a member non-static method or constructor [5].

3 Method of Feature Collection

Features like the ones mentioned in the previous chapter can be collected on two levels. The first level is a textual representation of the source code. This approach leads in most cases to employment of some pattern matching mechanism like regular expressions. To implement feature collection from a textual representation of source code is very tedious and error-prone task as an implementer has to deal with all the syntactic sugar of the language. This is why the feature collector presented in the paper is implemented on

more abstract level, on the abstract syntax tree. Abstract syntax tree (AST) is a tree representation of the source code syntactic structure. The tree is called abstract, because it omits some details appearing in the real language syntax [6]. AST can be in many cases obtained as a result of a compilation process, where the source code is parsed and concrete syntax tree (CST) is created. Abstract syntax tree is then obtained by contextual analysis and information enrichment of the original tree. Again, pattern matching could be employed to implement feature collection in more elegant fashion then in the first case. Features are understood as mapping $F : A \to R$, where A is a tree and R is the set of real numbers. A set of collected data D_S for a feature space S can be defined as $D_S = \{F(A)|F \in S\}$. The F should be from an interval $< 0, 1 >$, but it is not a necessity as data are typically normalized before further use [7].

Presented feature collector is based on an idea that the AST of an object oriented code can be viewed upon as a hierarchical database of data types, attributes, methods, statements and expressions. It is then natural to think of some query language like SQL to query the database. Implementation of the first feature (factory) written in such pseudo query language could look like this:

```
define is-factory-method(method): boolean
  ∃ st ∈ method/statements | st/type = return-statement
  ∧ (
    ∃ expr ∈ st/expression | expr/type = class-instance-creation
    ∨ ∃ expr ∈ st/expression | expr/type = name
          ∧ not exists f from method/type/field | f/name = expr/name
          ∧ exists ei ∈ method/expression | ei = variable-assignment
              ∧ (ei/right-side = class-instance-creation
              ∨ ei/right-side = method-invocation
              ∧ is-factory-method(ei/right-side/method))
  )

select count(method) ∈ type/method | is-factory-method(method)
```

The query in the pseudo query language is composed of two parts:

- a definition of recursive boolean function that return boolean value true only if the passed method definition return result of an instantiation that is not stored in object's attribute,
- a query that is using the defined function to restrict the set of all methods to methods that could imply the presence of the factory class.

4 Collector Implementation

To implement collector as designed in the previous chapter several challenges have to be overcome. The first problem is parsing of source code and abstract syntax tree creation. This can be usually done by compiler of the language considered. To parse a Java code, parser from the Eclipse platform can be employed. The advantage is that the parser provides the tree in object form.

```
ASTParser parser = ASTParser.newParser(AST.JLS4);
parser.setKind(ASTParser.K_COMPILATION_UNIT);
parser.setSource(source.toCharArray());
CompilationUnit cu = (CompilationUnit) parser.createAST(null);
```

On the first line, parser object is created for parsing the Java source code according to Java Language Specification version 4 (JLS4). The second line states that the parser should expect whole compilation unit (whole class definition with package specification and import statements). Other possibilities are to parse only a statement or an expression. The fourth line is a creation of the AST, where CompilationUnit object is a root of the tree.

Now, the abstract syntax tree is available, but in a form which is own only to Java. Thus, implementing the query language directly on the Java AST would lead to a platform specific query language and platform specific feature definition. The intention is to have the feature definition platform independent and applicable to a whole family of relative languages. To achieve this goal, the AST has to be converted to a different form. The most used platform independent hierarchical structure is definitely XML, it is thus natural choice to convert the Java AST to XML form. But, XML just define the form not the content, so set of mapping rules has to be created to convert the AST to XML representation, specifying which nodes are mapped to which elements, attributes, and text values.

As was mentioned earlier, features are in fact mappings from tree to set of real numbers, so some mechanism of XML manipulation is required. In the world of XML technologies exist several possibilities when it comes to XML manipulation. Extensible Stylesheet Language Transformations (XSLT) is a language for transforming XML documents into various formats. XSLT is based on ideas of functional languages and text-based pattern matching languages [8]. XSLT could handle the required transformation of XML representation of AST, but it was not designed as a query language. XML Path Language (XPath) is query and computation language to easily select specific nodes in XML documents and carry out simple computations [9]. Despite being query language, it is too simple to cover the required functionality as was outlined in the example implementation of Factory feature in a pseudo query language. The finalist is thus XQuery, a functional programming language designed originally as a query language for XML databases. XQuery is in fact a superset of the XPath language. XPath in XQuery is used as an addressing mechanism of XML nodes, while XQuery provides additional features like the FLWOR construct [10]:

- F = FOR, specify a temporary variable, in which is stored currently processed node,
- L = LET, enable to specify additional variables during the query,
- W = WHERE, restriction of the queried set,
- O = ORDER BY, specify the sequence of result,
- R = RESULT, specify the form of a result.

5 Features implementation

Implementation of proposed features in XQuery is not such straightforward as in the example, because some details were omitted on purpose. First, some helper functions have to be defined. The find-method function is used to find a method in a type specified by name and number of arguments. Name and number of arguments might not be sufficient information to positively identify method definition because of methods overloading. In the case of overloaded methods, data types of all arguments would be required to identify positively. Identify arguments data types from a method invocation is quite a difficult task during static analysis of source code, thus this case is simplified in the function body.

```
declare function local:find-method($name as xs:string,
        $argnum as xs:integer,
        $type as element(type)) as element(method)* {

  let $methods := $type//method[./name/text() = $name]
  return
    if (count($methods) = 1)
    then $methods
    else
      if (count($methods) > 1)
      then($methods[count(./arguments/*) = $argnum])[0]
      else ()
};
```

Next function just verify that an instantiation expression is not using data type of covering class.

```
declare function local:is-proper-instantiation(
  $inst as element(expression)*) as xs:boolean {

  let $res :=
    for $ins in $inst
    return not($ins/variable-type/name/text()
      = $inst/ancestor::type/name/text())
  return true() = $res
};
```

That is all for helper functions and the main function, that verify whether the method could be a method of a factory, can be defined. The main problem is to identify all possibilities how could be a result of instantiation returned from the method.

```
declare function local:is-factory-method($m as element(method))
   as xs:boolean {
  let $field-names := $m/ancestor::type//field/name/text()
```

Function takes method definition as argument and returns boolean value indicating whether the method is a factory method. Local variable holding all names of attributes defined in the covering class is created.

```
let $var-declarations := $m//statement[
   @statement-type = 'variable-declaration'
   and count(./initializer//expression
     [@expression-type = 'class-instance-creation']) > 0
   and local:is-proper-instantiation(./initializer/expression
     [@expression-type = 'class-instance-creation'])
 ]/name
```

Variable containing names of all newly declared variables in the method body, which are also initialized by instantiation, is defined.

```
let $var-assignments := $m/body//expression[
  @expression-type = 'assignment'
  and count(./right-operand/expression
    [@expression-type = 'class-instance-creation']) > 0
  and local:is-proper-instantiation(./right-operand/expression
    [@expression-type = 'class-instance-creation'])
]/left-operand/name/text()
```

Another way of variable initialization is by the assignment, so all assignments, where on the right side is instantiation, are stored in another local variable.

```
let $return-statements := $m/body//statement[@statement-type = 'return']
let $rs-new := $return-statements[
  count(./expression
    [@expression-type = 'class-instance-creation']) > 0
  and local:is-proper-instantiation(./expression
    [@expression-type = 'class-instance-creation'])
  ]
let $rs-var := $return-statements[
  count(./name) = 1
  and (
    not(./name/text() = $field-names)
    and ./name/text() = $var-assignments
  )
  or ./name/text() = $var-declarations
]

let $rs-meth := $return-statements[
  let $is-inv := count(./expression
    [@expression-type = 'method-invocation']) = 1
```

```
    let $meth :=
      if ($is-inv and not(./expression
        [@expression-type = 'method-invocation']
        /name/text() = $m/name/text()))
      then local:find-method(./expression
        [@expression-type = 'method-invocation']/name/text(),
        count(./expression[@expression-type = 'method-invocation']
        /arguments/expression), $m/ancestor::type)
      else ()
    return
      if ($is-inv and count($meth) = 1)
        then local:is-factory-method($meth)
        else false()
  ]
```

The core of the function is composed of return statements analysis. Three types of return statements are detected:

1. Return statement where the returned expression is instantiation.

2. Return statement where the returned expression is a simple name. The name must not be a name of an attribute and there has to exist an assignment where the right side is an instantiation.

3. Return statement where the returned expression is a method invocation. The invoked method must comply to the same rules. This is ensured by recursive call of the is-factory-method function.

```
  return count($rs-new) > 0 or count($rs-var) > 0 or count($rs-meth)
};
```

The examined method is declared as a factory method if there is any of presented return statements.

6 Conclusion

The paper introduced an uncommon but effective method of feature collection for classification and pattern recognition in source code. Due to the conversion of abstract syntax tree to XML, the XQuery language could be employed as a query language and thus platform independent feature definition language. An extensive example of feature definition was given. The presented feature is important for classification of Factory design pattern. Besides the presented feature, over forty additional features have been proposed and implemented.

References

[1] Nikolaos, T., et al. *Design pattern detection using similarity scoring.* In 'Software Engineering', IEEE Transactions on 32.11 (2006), 896–909.

[2] Object Management Group. *OMG Unified Modeling Language (OMG UML), Superstructure.* Version 2.4.1 (2011).

[3] Fowler, M. *Patterns of enterprise application architecture.* Addison-Wesley Longman Publishing Co., Inc. (2002).

[4] Hohpe, G., Woolf, B. *Enterprise integration patterns: Designing, building, and deploying messaging solutions.* Addison-Wesley Professional (2004).

[5] Rost, M. *Feature definition and software design for Java source code classification tool.* In Doktorandske dny 2013, FJFI, CVUT (2013).

[6] Pfenning, F., Conal, E. *Higher-order abstract syntax.* In ACM SIGPLAN Notices. Vol. 23. No. 7. ACM (1988).

[7] Rost, M., Smolka, J., Mojzes, M., Virius, M. *Tool for Statistical Classification of Java Projects.* In: Software Development and Object Technologies 2013. Jihlava: VSPJ – The College of Polytechnics Jihlava (2013).

[8] Clark, J., W3C. *XSL Transformations (XSLT) Version 1.0* (1999).

[9] Clark, J., DeRose, S., W3C. *XML Path Language (XPath) Version 1.0* (1999).

[10] Boag, S., W3C, et al. *XQuery 1.0: An XML Query Language (Second Edition)* (2011).

Notes on Electro-Osmotic Drag Coefficient*

Lucie Strmisková†

4th 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, New Technologies Research Centre, University of West Bohemia

Abstract. This contribution deals with a transport of water in polymer membrane, which serves as an electrolyte in a hydrogen fuel cell. Special attention is paid to the electro-osmotic drag – phenomena, that has a significant influence on the humidification of whole membrane. The values of electro-osmotic drag coefficient obtained from different measurements are discussed. The value of this coefficient is obtained by simple model based on linear irreversible thermodynamic and it is compared with the experimental values.

Keywords: hydrogen fuel cell, water transport, electro-osmotic drag

Abstrakt. Tento příspěvek se zabývá transportem vody v polymerní membráně, která slouží jako elektrolyt ve vodíkovém palivovém článku. Velká pozornost je věnována elektro-osmotickému strhávání – jevu, který má podstatný vliv na zavodnění celé membrány. Hodnoty koeficientu elektro-osmotického strhávání získané z různých měření jsou diskutovány. Hodnota tohoto koeficientu je vypočítána pomocí jednoduchého modelu, který je založený na lineární nerovnovážné termodynamice, a tato hodnota je srovnána s hodnotami experimentálními.

Klíčová slova: vodíkový palivový článek, transport vody, elektroosmotické strhávání

1 Introduction

A fuel cell is defined as an electrochemical device, that converts the chemical energy of the fuel to the electrical energy. Unlike storage cells, fuel cells can produce electrical energy indefinitely, if we continuously feed them with a fuel and remove the reaction products. There are many types of fuel cells, but we will be interested only in hydrogen fuel cell with Nafion membrane as an electrolyte.

The basic operation of hydrogen fuel cell is quite simple. It is a reversed electrolysis of water. Hydrogen gas is driven to the anode, where it comes into contact with a platinum catalyst on the electrode surface. Then hydrogen ionizes to electron and proton.

$$2H_2 \rightarrow 4H^+ + 4e^- \tag{1}$$

The produced electrons pass through the external electrical circuit to the cathode due to an electrical potential gradient, creating thus required electrical current. Protons create

*This work has been supported by the grants CZ.1.07/2.3.00/20.0107 and SGS13/217/OHK4/3T/14.

†This work has been done in colaboration with Michal Pavelka and Petr Sedlák, NTC University of West Bohemia.

a bond with a water molecule from the membrane surface and in the form of hydronium ion H_3O^+ pass to the electrolyte and then they move to the cathode.

The cathode is fed by oxygen, usually in the form of air. Oxygen reacts with the electrons from the cathode and with the protons taken from the electrolyte and forms water.

$$O_2 + 4e^- + 4H^+ \rightarrow 2H_2O \tag{2}$$

The total oxidation-reduction reaction in the hydrogen fuel cell is thus

$$2H_2 + O_2 \rightarrow 2H_2O. \tag{3}$$

The reaction (1) is slightly endothermic, but the reaction (2) is highly exothermic so as a result, heat is produced within the cell.

Polymer membrane serves as an electrolyte and plays a vital role in fuel cells. It has to prevent mixing of reactant gases and provide good transport of protons from the anode to the cathode with as little resistance as possible. On the other hand, the resistance for the electron transport should be as high as possible. If electrons could pass through electrolyte, we will not gain the required electrical current.

The membrane also has to have high chemical and thermal stability and low production cost.

None of the currently developing materials satisfy all the requirements laid on the electrolyte. The most closed to the requirements and therefore the most common material used for the membrane is a material known under its commercial name Nafion, which was developed by DuPont company in the late 1960s. The biggest disadvantage of Nafion is its high price.

Nafion consists of a polytetrafluoroethylene backbone with the randomly attached perfluorinated side chains ending by a sulfonate acid group ($-SO_3H$). The structure of side chains varies for different types of Nafion and also for different membrane manufactures. The bonds between fluorine and carbon make Nafion very durable and chemical-resistant, they also provide high operating temperature.

Nafion is very good proton conductor, when it is sufficiently wet. So for good fuel cell operation, we need to keep membrane fully and uniformly humidified all the time. We will show in the next section, how difficult aim is it and which problems are necessary to overcome in order to ensure good Nafion humidification.

2 Role of water in Nafion

As we have said, it is well observed, that Nafion is a good proton conductor only if it is sufficiently wet. The conductivity of dry membrane is almost six orders of magnitude lower than the conductivity of fully humidified membrane. Insufficient water level inside the membrane does not lead only to the poor proton conductivity and thus to lower fuel cell performance, but dry membrane is also more prone to the pinhole formation and the degradation process is more fast or even membrane failure can occur.

On the other hand, if the level of water is too high, the excess water blocks the pores in gas diffusion or catalyst layers and the mass transport is limited, which leads to higher voltage losses[4]. Because of this, the design of catalyst layers has to ensure, that product

water is repelled from transport pores and that it is pulled to the membrane, where it increases the membrane conductivity.

The secret of a high proton conductivity of Nafion membrane is in its morphology, although the exact morphology of Nafion is not known, despite the fact, that it has been investigated extensively since the early 1970s.

The main difficulties are the facts, that the polytetrafluoroethylene chains has no uniform length, but their length is randomly distributed along the average length. Also the side sulphonated chains are not placed to exact place on the polytetrafluoroethylene backbone, but their placement is more or less random.

Despite this randomness, there are several generally accepted statements about Nafion membrane morphology. The most significant property is that the membrane is separated to into distinct hydrophobic and hydrophilic regions.

The bond between H^+ and SO_3^- is ionic and there is a strong mutation between the positive and negative ion of each molecule, therefore the side chains tend to cluster within Nafion. And because the polytetrafluoroethylene backbone is hydrophobic, while the suplphonated side chains are highly hydrophilic, these side chains clusters attract the water presented in membrane, so we have the structure composed from hydrated and dry regions. Protons inside these hydrated regions are able to move almost freely. For good proton membrane conductivity, these hydrated regions have to be as large as possible and there should be a connection between them.

The connecting path between these hydrated regions is really observed, when Nafion is sufficiently humidified, and protons move there almost like in fully aqeous environment. When membrane dries out, these channels are becoming narrower and the proton transfer is decelerated by the attractive forces of the surface of these channels.

There are four main causes of water transport inside the membrane: diffusion, electro-osmotic drag, pressure driven hydraulic permeation and capillary effect.

The pressure driven hydraulic permeation is negligible in comparison with drag and diffusion, if the operating temperature is under 70 C. But for higher temperatures, this factor can also highly affect water balance [2].

Water is created at the cathode by the oxygen reduction. Part of the generated water is removed by the air flow, but the rest diffuses to the anode due to the concentration gradient or differences in water activity.

Protons travel from the anode to the cathode, but isolated proton without electron cloud can exist freely in solutions only shortly, so when such proton meets a water molecule, it bounds to it forming thus hydronium ion H_3O^+. The higher ions $H_5O_2^+$ (Zundel ion) and $H_9O_4^+$ (Eigen ion) can be also created. These aggregates of water molecule and excess proton continue in the earlier proton direction to the cathode. This phenomena is called electro-osmotic drag.

There are two competing mechanisms of proton transfer in Nafion membrane: vehicular mechanism and Grothuss mechanism. The differences between both mechanisms are depicted in figure 1.

The vehicular mechanism is a diffusion of hydrated proton ($H^+(H_2O)_x$) due to gradient of electrochemical potential.

The Grotthuss mechanism is sometimes called as hopping. The produced proton sticks to the water molecule presented in the catalyst-membrane interface creating thus

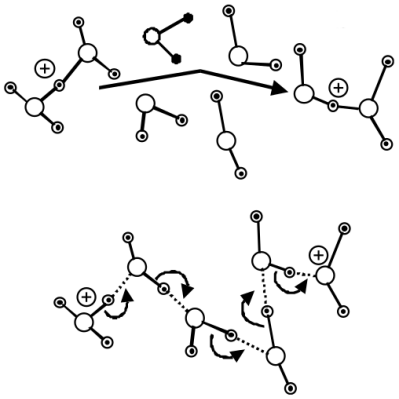


Figure 1: The mechanisms of proton transfer in Nafion membrane. Vehicle mechanism is in the top, Grothuss mechanism is in the bottom. [6]

the hydronium ion H_3O^+. When this ion is close to another water molecule, proton hops to it. Original ion turns again into water molecule and water molecule changes to hydronium ion. This way, proton hopping continues until it reaches cathode. Grotthuss mechanism is achieved through a local reorientation of water molecules and shuffling of hydrogen bonds and it was found, while trying to understand, why is the proton conductivity in water $5 - 8$ times higher than the conductivity of other cations.

At higher current densities, the produced protons thus do not allow water to reach the anode and although the cathode side of the membrane is flooded, the anode side can be completely dry. Humidifying of the anode is a solution, but it is no so easy, because excessive liquid water can block the pores and limited mass transport leads to significant voltage losses, so the level of hydration of the feed gases should be managed very carefully.

Maintaining the proper water level is not easy also because the membrane should stay optimally hydrated while varying power output. Another difficulty is the fact, that the reactants are not distributed homogeneously, some regions will be abundant to the fuel, some regions will be insufficiently supplied by the fuel. Therefore there will be large variations in local current. The distribution of the produced Joule heat will approximately correspond to the current distribution. The heat can form pinholes in the membrane and lead thus to the higher fuel crossover. It can also dry out the membrane and increase thus the membrane resistance.

The proper water management seems to be a key point in designing a fuel cell and many engineers and researchers are trying to find the ideal solution of it. The deep understanding of the electro-osmotic drag is necessary for solving the problem. Unfortunatelly the systematic experimental data on this phenomena are still missing.

It is generally known, that proton drags water molecules during its journey to the cathode. The average number of water molecules dragged by proton is called electro-osmotic drag coefficient. Its value is obtained from the experiments. The problem is, that different experimental techniques gives us significantly different values of this coefficient (between one and five water molecules per proton, see the figure 2)[2].

One need to know the electro-osmotic drag coefficient while modeling the transport

Table 1 – Comparison of the selected EOD coefficients in PEMFCs.

Researchers	Measurement	PEM	T (°C)	EOD coefficient
Fuller et al. [64]	Concentrated cell	Nafion® 117	25	1.4 (vapor equilibrated) Decreased slowly as the membrane was dehydrated, falling sharply toward zero as the concentration of water approached zero.
Zawodzinski et al. [53]	Electro-osmotic drag cell	Nafion® 117 Recast Nafion membrane	30	Nafion® 117, $\lambda = 22$: 2.5–2.9 Nafion® 117, $\lambda = 11$: 0.9 Recast Nafion membrane: 2.9–3.4
Zawodzinski et al. [58]	Electro-osmotic drag cell	Nafion® 117 Dow XUS Membrane C		Nafion® 117, $\lambda = 22$: 2.0–2.9 Nafion® 117, $\lambda = 11$: 0.9 Dow Membrane: 1.4–2.0 Membrane C: 2.6–4.0
Zawodzinski et al. [59]	Electro-osmotic drag cell			1.0 (vapor equilibrated) 2.5 (liquid equilibrated) Independent of water content over $\lambda = 1.4$–14 (vapor equilibrated). Not significantly dependent on details of membrane microstructure.
Ren et al. [61]	DMFC analysis	Nafion® 117	60, 80	3.16 @ 80 °C, 2.82 bar BP 2.86 @ 60 °C, 2.82 bar BP
Ge et al. [67]	Water flux measurement	Nafion® 117	30, 50, 80	$\lambda < 4$, temperature had no influence on the EOD. $\lambda > 4$, the EOD coefficient increased with increasing temperature. 1.1 for the water activity = 1.0. Linearly increased from 1.8 to 2.7 at 15–85 °C (liquid equilibrated). Keep constant at 0.05–1.0 A/cm^2 current densities at 75 °C.
Ise et al. [65]	NMR	Nafion® 117	30–80	EOD coefficient increased with increasing temperature and water content ($\lambda = 11$–20). $\lambda = 13$, 1.7 @ 22 °C and 2.5 @79 °C
Yan et al. [7]	Water flux measurement	Nafion® 117	80	1.5–2.6
Takaichi et al. [62]	Inserted Pt potential probes	Nafion® 211	80	The ratio of EOD coefficient to BD coefficient was constant irrespective of the current density (0.2–200 mA/cm^{-2} at 20, 40, or 60% RH feed gases)
Ye et al. [2]	Hydrogen pumping cell	GORE-SELECT® membranes	80	1.07 (40 and 95% RH)
Husar et al. [8]	Water flux measurement	Nafion® 115	40, 60	0.25–0.4 (0.3–0.8 A/cm^2) at 40 °C 0.65–1.05 (0.3–1.0 A/cm^2) at 60 °C

Figure 2: The values of electro-osmotic drag coefficient. [2]

phenomena in the membrane. But such a scattering of experimental data have to make one desperate. There is also a big question, why is the range of data so wide.

In most of models, electro-osmotic drag coefficient is expected to be a constant. But it depends on state variables (temperature, pressure, thickness of the membrane). Different experimental methods show different values of drag coefficient even in same states. So it seems, that the choise of experimental technique has an influence on the measured data, although this is the situation, that should not occur in experimental physics.

But they are same common points in all measurements. It seems, for instance, that electro-osmotic drag coefficient linearly increases with the increasing temperature.

3 Thermodynamic constraint of electro-osmotic drag coefficient

In this section, the classical linear non-equilibrium thermodynamics will be used for determining the constraint of electro-osmotic drag coefficient.

The linear non-equilibrium thermodynamics describes the system, which is sufficiently close to the equilibrium [5]. There are no thermodynamic fluxes and forces, when the system is in the equilibrium. When the system is leaving the equilibrium, the forces and fluxes will smoothly grow. The Taylor expansion around the equilibrium can be done. There is a region, which is described by linear part of this expansion with high accuracy. This region is called linear region and thermodynamics valid in this region is called linear non-equilibrium thermodynamics.

In the linear region, the relation between thermodynamic fluxes (j_i) and thermodynamic forces (X_i) is simple.

$$j_i = \sum_{i=1}^{N} L_{ik} X_k, \tag{4}$$

where the coefficients L_{ik} are called phenomenological coefficients. The phenomenological coefficients are constants independent on thermodynamic forces X_k, but they are functions of state variables of a thermodynamic system.

Phenomenological thermodynamics has no tool for determining the values of these coefficients. Their values are generally determined experimentally. But thermodynamics put some constraints on their values. According to the second law of thermodynamics, the entropy production is not negative.

$$\sigma(S) = \sum_{i=1}^{N} j_i X_i = \sum_{i,k=1}^{N} L_{ik} X_i X_k \geq 0, \tag{5}$$

so the phenomenological coefficients have to fulfill Sylvester conditions.

Moreover, these coefficients are not independent, but they have to obey the Onsager relations of reciprocity [5].

$$L_{ik} = L_{ki} \tag{6}$$

The linear non-equilibrium thermodynamics will be used now for describing the transport of protons and water inside the Nafion membrane. The membrane is assumed to be a homogeneous space, where diffusivity and proton conductivity are constants, that depend only on state variables.

According to the linear non-equilibrium thermodynamics, the constitutive relations for the transport of protons and water inside the membrane are as follows.

$$j_{H^+} = L_{H^+H^+} X_{H^+} + L_{H^+w} X_w, \tag{7}$$

$$j_w = L_{wH^+} X_{H^+} + L_{ww} X_w, \tag{8}$$

where j_{H^+} is the flux of protons and j_w is the flux of water. The thermodynamic force X_w corresponds to the gradient of water chemical potential and X_{H^+} to the gradient of electrochemical potential. The other forces like temperature, pressure gradient or capillarity forces are considered to be negligibly small in comparison with the forces X_w and X_{H^+}.

The phenomenological coefficients L_{ij} have to satisfy Onsager relations of reciprocity, i.e.,

$$L_{H^+w} = L_{wH^+},$$

so there are only three independent coefficients. In order to gain the physical interpretation of these coefficients, the equations (7), (8) are rewritten into the following form

$$j_{H^+} = \sigma_{H^+} X_{H^+} + K j_w, \tag{9}$$

$$j_w = \sigma_w X_w + \xi j_{H^+}. \tag{10}$$

The coefficient $\xi = \left(\frac{j_w}{j_{H^+}}\right)_{X_w=0}$ is previously discussed electro-osmotic drag coefficient. The coefficient K is defined similarly as $K = \left(\frac{j_{H^+}}{j_w}\right)_{X_{H^+}=0}$.

Water cannot drag different protons than protons, that are present in the membrane, so the following inequality for the coefficient K must be valid.

$$K = \left(\frac{j_{H^+}}{j_w}\right)_{X_{H^+}=0} \leq \frac{c_{H^+}}{c_w}. \tag{11}$$

The relation between the coefficient ξ and K and the phenomenological coefficients follows from the original equations (7), (8). The force X_w in the equation (7) is substituted for the force X_w expressed from the equation (8). The same process is analogously done for the force X_{H^+} and the following set of the equations is thus gained.

$$j_{H^+} = \frac{L_{ww}L_{H^+H^+} - L_{wH^+}L_{H^+w}}{L_{ww}} X_{H^+} + \frac{L_{wH^+}}{L_{ww}} j_w, \tag{12}$$

$$j_w = \frac{L_{ww}L_{H^+H^+} - L_{wH^+}L_{H^+w}}{L_{H^+H^+}} X_w + \frac{L_{wH^+}}{L_{H^+H^+}} j_{H^+}. \tag{13}$$

The relation between coefficients ξ and K is obvious from these equations.

$$K = \frac{L_{wH^+}}{L_{ww}}, \tag{14}$$

$$\xi = \frac{L_{wH^+}}{L_{H^+H^+}}. \tag{15}$$

The coefficient of electro-osmotic drag can be rewritten using the coefficient K, which fulfills the inequality (11).

$$\xi = \frac{L_{H^+w}}{L_{H^+H^+}} = \frac{L_{ww}}{L_{H^+H^+}} \frac{L_{H^+w}}{L_{ww}} \leq \frac{L_{ww}}{L_{H^+H^+}} \frac{c_{H^+}}{c_w} \tag{16}$$

The Nernst-Einstein relations

$$\sigma_w = \frac{c_w D_w}{R}, \quad \sigma_{H^+} = \frac{c_{H^+} D_{H^+}}{R}, \tag{17}$$

where D_w, D_{H^+} is the diffusivity of water and protons respectively, are used for further arranging of the equations.

The value of the electro-osmotic drag coefficient is thus limited only with the ratio of diffusivity coefficients of water and protons.

$$\xi \leq \frac{D_w}{D_{H^+}} \tag{18}$$

The diffusivity coefficients for the examinated operating temperature and pressure can be found elsewhere in the literature. The diffusivity coefficients sugested by Choi et al. [1] will be put into the inequality just to have an estimation, how the inequality (19) limits the values of electro-osmotic drag coefficient.

$$\xi \leq \frac{D_w}{D_{H^+}} = \frac{2.26}{7.22} = 0.3 \tag{19}$$

This value is much lower than the experimental values. We have a strong suspicion, that the electro-osmotic drag is not so significant inside the membrane as it is generally thought. We think, that the surface and bulk properties of Nafion are different and that the electro-osmotic drag coefficient from experiments is so high, because it is high on the surface.

This idea is supported by the fact, that different measurement techniques give different values even at the same conditions. It seems, that each measurement technique slightly changes the surface of the membrane, changing thus also its properties and consequently the value of the electro-osmotic drag coefficient.

Moreover if the value of drag coefficient would be so high, the hydrogen fuel cell will be not able to work without external hydration of feed gases for long time. Benziger showed ([3]), that it is possible to run fuel cell without hydration without any significant changes of the membrane. So his experiments also support the idea, that the electro-osmotic drag should be smaller, than is measured.

References

[1] P. Choi, N.H. Jalani, R. Datta. *Thermodynamics and proton transport in Nafion. II. Proton diffusion mechanism and conductivity* Journal of the Electrochemical Society 152 (2005), 123–130

[2] W. Dai, H. Wang, X.-Z. Yuan, J. Martin, D. Yang, J. Qiao, J. Ma. *A review on water balance in the membrane electrode assembly of proton exchange membrane fuel cells* International Journal of Hydrogen Energy 34 (2009), 9461–9478

[3] Q. Duan, H. Wang, J. Benziger. *Transport of liquid water through Nafion membrane* Journal of Membrane Science 392–393 (2012), 88–94

[4] M. Eikerling, A.A. Kornyshev, A.R. Kucernak. *Water in polymer electrolyte fuel cells: Friend or foe?* Physics Today (2006), volume 59, issue 10, 38–44

[5] F. Maršík, I. Dvořák. *Lineární nerovnovážná termodynamika.* In 'Biotermodynamika (Praha, 1998)', Academia, 86–93.

[6] B. Pivovar. *An overview of electro-osmosis in fuel cell polymer electrolytes* Polymer 47 (2006), 4194–4202

SU(5) à la Witten*

Helena Šedivákova†

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

advisor: Michal Malinský, Institute of Particle and Nuclear Physics, FMP CU

Abstract. We argue that the Witten's loop mechanism for the right-handed Majorana neutrino mass generation identified originally in the $SO(10)$ grand unification context [8] can be successfully adopted to the class of the simplest flipped $SU(5)$ models [4, 5, 6]. In such a framework, the main drawback of the $SO(10)$ prototype, in particular, the generic tension among the gauge unification constraints and the absolute neutrino mass scale is alleviated and a simple yet potentially realistic and testable scenario emerges. Indeed, the perturbative baryon number violating processes such as proton decay are allowed in the flipped $SU(5)$ model, in particular, the partial proton decay widths are calculable (see [7] where also comparison with ordinary $SU(5)$ is given), and may be subjected to future experiments [1, 2, 3]. Firstly, the generic property of the simplest flipped $SU(5)$ models is $\Gamma(p \to K^+\overline{\nu}) = 0$. Next, the loop generation of the right-handed neutrino mass induces a tight correlation of the proton decay widths with the neutrino sector which results in predictions on the decay channels to neutral mesons. According to our analysis, $\Gamma(p \to \pi^0\mu^+)$ is bounded from above, whereas a lower bound on the sum $\Gamma(p \to \pi^0 e^+) + \Gamma(p \to \pi^0\mu^+)$ occurs.

Keywords: Witten's loop, flipped $SU(5)$, proton decay.

Abstrakt. Wittenův mechanismus pro generování majoranovských hmot pravotočivých neutrin byl původně implementován v teorii velké unifikace s kalibrační grupou $SO(10)$ (viz článek [8]), avšak v této práci ukazujeme, že jej lze použít i pro třídu modelů známých pod označením flipovaná $SU(5)$ [4, 5, 6]. Jednou z nevýhod použití Wittenovy smyčky v $SO(10)$ je nesrovnalost mezi absolutní škálou hmot neutrin a omezením plynoucím z podmínky unifikace – ukazuje se, že v tomto případě parametry modelu nelze nastavit tak, aby model odpovídal realitě. Tento problém mizí ve flipované $SU(5)$, neboť podmínky unifikace jsou zde mnohem slabší – jsou kladeny pouze na vazbové konstanty příslušné neabelovským kalibračním grupám, dostáváme tak realistický a potenciálně testovatelný model. Experimentální ověření modelů velké unifikace poskytují zejména procesy, kde se nezachovává baryonové číslo a které tedy ve standardním modelu nejsou dovoleny (neuvažujeme-li neporuchové efekty). Ve flipované $SU(5)$ lze velmi dobře vypočítat dílčí rozpadové šířky pro rozpad protonu (tyto výpočty a jejich srovnání se situací v běžných $SU(5)$ modelech jsou provedeny např. v publikaci [7]), které mohou být v budoucnu změřeny experimenty jako [1, 2, 3]. Typickým rysem flipované $SU(5)$ je absence rozpadu protonu na K^+ a antineutrino ($\Gamma(p \to K^+\overline{\nu}) = 0$). Náš model je navíc specifický silnou korelací mezi rozpadovými šířkami protonu a neutrinovým sektorem, která vzniká kvůli smyčkovému generování hmot pravotočivých neutrin a která umožňuje přesněji vypočítat rozpad

*The full paper written in cooperation with M. Malinský and Carolina Arbeláez Rodríguez is available at http://arxiv.org/abs/1309.6743. Publication in a peer-reviewed periodical is expected.

†The work of H.Š. is supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS13/217/OHK4/3T/14

protonu na neutrální mezony a nabité leptony. Naše analýza ukazuje, že lze nalézt horní odhad pro rozpadovou šířku $\Gamma(p \to \pi^0\mu^+)$, zatímco pro součet $\Gamma(p \to \pi^0 e^+) + \Gamma(p \to \pi^0\mu^+)$ existuje dolní odhad.

Klíčová slova: Wittenova smyčka, flipovaná $SU(5)$, rozpad protonu.

References

[1] K. Abe, T. Abe, H. Aihara, Y. Fukuda, Y. Hayato, et al. *Letter of Intent: The Hyper-Kamiokande Experiment — Detector Design and Physics Potential —.* (2011).

[2] T. Akiri et al. *The 2010 Interim Report of the Long-Baseline Neutrino Experiment Collaboration Physics Working Groups.* (2011).

[3] D. Autiero, J. Aysto, A. Badertscher, L. B. Bezrukov, J. Bouchez, et al. *Large underground, liquid based detectors for astro-particle physics in Europe: Scientific case and prospects.* JCAP **0711** (2007), 011.

[4] S. M. Barr. *A new symmetry breaking pattern for so(10) and proton decay.* Phys. Lett. **B112** (1982), 219.

[5] A. De Rújula, H. Georgi, and S. L. Glashow. *Model of neutrino-induced multilepton events.* Phys. Rev. D **17** (1978), 151–153.

[6] J. Derendinger, J. Kim, and D. Nanopoulos. *Anti-su(5).* Phys. Lett. **139B** (1984), 170–176.

[7] I. Dorsner and P. Fileviez Perez. *Distinguishing between SU(5) and flipped SU(5).* Phys.Lett. **B605** (2005), 391–398.

[8] E. Witten. *Neutrino masses in the minimal o(10) theory.* Phys. Lett. **B91** (1980), 81.

Paralelizace neuronové sítě s přepínacími jednotkami*

Vladimír Španihel

3. ročník PGS, email: vladimir.spanihel@seznam.cz
Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: František Hakl, Ústav informatiky, AV ČR

Abstract. This contribution is concerned with the possibility of making the learning process of neural network with switching units more parallel. Nvidia CUDA architecture will be used for paralelization of the learning process. This architecture serves for speedup of various computations due to huge number of CUDA cores which are available on GPU.

Keywords: CUDA, NNSU, MAGMA

Abstrakt. Tento příspěvek se zabývá možností paralelizace učícího procesu neuronové sítě s přepínacími jednotkami. Pro paralelizaci bude použita architektura Nvidia CUDA sloužící k urychlení nejrůznějších výpočtů díky velkému počtu výpočetních jader dostupných na grafických procesorech.

Klíčová slova: CUDA, NNSU, MAGMA

1 Úvod

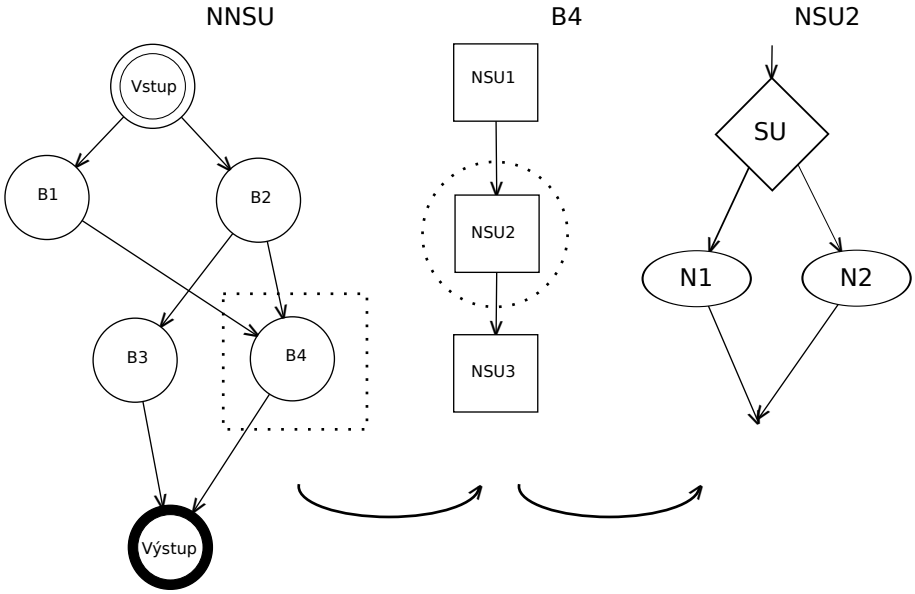
Neuronová síť s přepínacími jednotkami (NNSU) je nástroj určený k separaci dat či k aproximaci funkcí. Tento článek je sbírkou poznatků o této neuronové síti a o možnosti další optimalizace učícího procesu. Optimalizace bude spočívat v nahrazení sériového algoritmu řešení soustavy lineárních rovnic algoritmem paralelizovaným na GPU. K podpoře paralelizace na grafickém procesoru bude využita architektura Nvidia CUDA [9].

2 Neuronová síť s přepínacími jednotkami

Neuronová síť s přepínacími jednotkami je tvořena bloky zřetězených tzv. neuronů s přepínací jednotkou (NSU). Učení takovéto sítě probíhá na dvou úrovních. První úroveň učení v sobě skrývá optimalizaci architektury celé NNSU pomocí genetických algoritmů, zatímco na druhé úrovni dochází k učení jednotlivých NSU. Pro představu je na obrázku 1 vidět architektura NNSU.

Tedy neurony NNSU jsou složeny ze zřetězených NSU. NSU vždy obsahuje přepínací jednotku SU a tzv. výpočetní uzly. Uvnitř přepínací jednotky dochází k rozdělování vstupních vektorů do podmnožin, které přepínací jednotka posílá na konkrétní výpočetní uzel. Výpočetní uzel je vlastně obyčejný perceptron.

*Tato práce byla podpořena granty SGS11/167/OHK4/3T/14 a GA TA ČR TA01010490.



Obrázek 1: Levá část obrázku zobrazuje jednoduchou celou NNSU. Ve střední části je rozkreslen blok 4, který je složen ze tří neuronů s přepínací jednotkou. Pravá část vykresluje příklad, jak může vypadat struktura neuronu s přepínací jednotkou.

2.1 Genetické algoritmy v NNSU

Architektura NNSU lze jednoduše popsat jako acyklický graf. O výsledné uspořádání hran a uzlů se stará optimalizace pomocí genetických algoritmů [3]. Aby bylo vůbec možné použít genetickou optimalizaci bylo nutné najít vhodnou reprezentaci acyklického grafu. Nakonec byla zvolena kombinace dvou metod: PST (Program Symbol Trees [1]) a Readovy lineární kódy [10]. Roman Kalous ve své disertaci nazval tuto reprezentaci IP kód (Instruction-Parameter Code). Tato metoda dokáže serializovat uzly a hrany NNSU. Pomocí PST je reprezentována architektura ve tvaru stromu a Readovy kódy tuto strukturu převádějí na celočíselné řady. Více informací o IP kódu najdete v [3].

2.2 Učení NSU

Při učení NSU je nutné naučit jak přepínací jednotku tak jednotlivé výpočetní uzly. Přepínací jednotka se učí pomocí Forgyho metody (varianta k-means algoritmu) pro shlukovou analýzu. Učení výpočetní jednotky spočívá ve vyřešení soustavy lineárních rovnic. Vektor pravých stran je tvořen pouze prvky -1, 1. Matice soustavy je specifikována vstupními vektory. Řešení dané soustavy určí váhy pro vstupy do výpočetní jednotky. Jelikož většinou vstupy dané výpočetní jednotky nejsou lineární kombinací vah, je třeba provést odhad metodou nejmenších čtverců.

Zatímco genetické algoritmy pro určení architektury sítě NNSU jsou již implementovány paralelně, řešení soustav lineárních rovnic se stále spouští sériově. Naší snahou je přesunout řešení soustav lineárních rovnic na grafickou kartu s využitím architektury CUDA (Compute Unified Device Architecture).

3 CUDA

Nvidia CUDA je technologie umožňující vývojáři využít potenciál paralelní architektury grafických čipů od firmy Nvidia. Alternativní technologie pro paralelizaci výpočtů na libovolné vícejádrové architektuře (včetně GPU) je označována jako OpenCL. Pro naše účely byla vybrána CUDA z důvodů lepší dostupnosti studijních materiálů a přehlednějšího API.

Jednou z nevýhod použití architektury CUDA je, že není podporována grafickými čipy jiných výrobců. Díky této skutečnosti, však nemusí její rozhraní obsahovat množství přepínačů podle různých čipů, na kterých bude výsledná aplikace spouštěna, což vede k přehlednějšímu kódu.

3.1 Základní informace

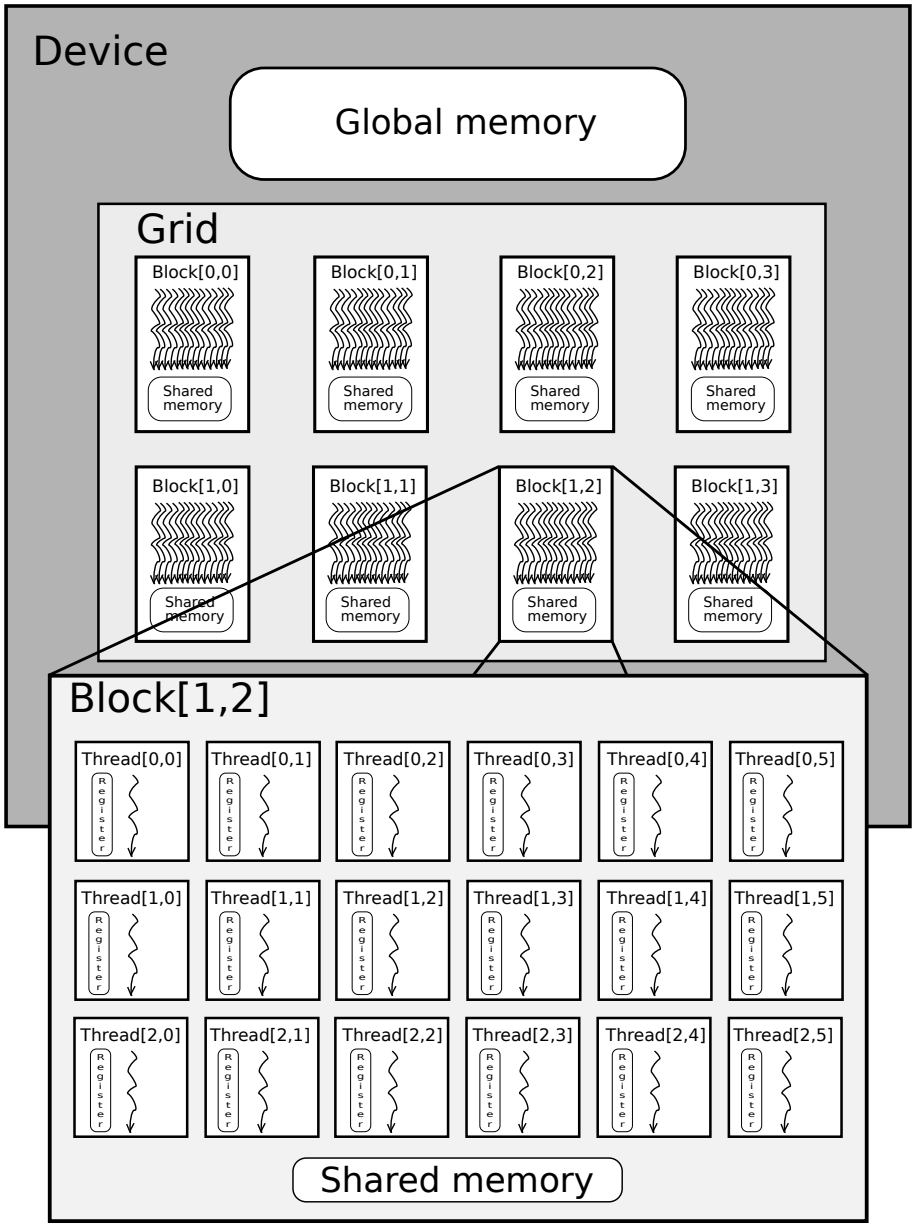
Programování grafických karet s využitím CUDA je velmi podobné psaní kódu v C nebo C++. Jde pouze o rozšíření syntaxe jazyka C, o čemž svědčí také způsob, jakým byl vytvořen překladač. Základní balík CUDA SDK obsahuje mimo jiné překladač nazvaný NVCC [6], který rozšiřuje běžně používaný překladač gcc.

Když programátor pročítá návody nebo dokumentaci setká se s pojmy jako *vlákno*, *warp*, *kernel* atd. Nyní si představíme nejčastěji používané pojmy při CUDA programování.

K rozlišení typu procesorů, na kterých má být prováděn kód, se používá označení *device* pro CPU resp. *host* pro GPU. Funkci volané z CPU, jejíž kód má být spuštěn na GPU, se říká kernel. Termín vlákno označuje část dat, zpracovávanou na jednom z CUDA jader. Vlákna je možné strukturovat do bloků a bloky do tzv. gridu, přičemž bloky i grid mohou být až 3 dimenzionální. Příklad takového rozdělení vláken znázorňuje obrázek 2.

Je třeba poznamenat, že GPU nemůže pracovat s daty uloženými v paměti CPU. Má-li program zpracovat určitá data, musí programátor zajistit jejich překopírování do paměti GPU před voláním daného kernelu. Kopírování mezi paměťmi je časově poměrně náročná operace. Další omezení spočívá v tom, že GPU neumožňuje spustit např. pouze jedno vlákno. Pro jednu instrukci se vždy spouští skupinky po 32 vláknech tzv. *warpy*, i když výsledek ze zbylých 31 není podstatný. Při implementaci algoritmů je třeba na toto myslet, aby byla karta vždy co nejvíce zatížená. Jen tak je možné vyvážit čas potřebný na zkopírování dat z paměti CPU do vnitřní paměti GPU a zpět.

Na obrázku 2 je také vidět některé typy pamětí, které má vývojář na grafické kartě k dispozici. Hlavní, největší a zároveň nejpomalejší pamětí GPU je tzv. globální paměť. Její velikost se dnes pohybuje v řádech gigabajtů a má do ní přístup libovolné běžící vlákno. Další typ paměti je paměť sdílená, jejíž velikost se pohybuje v řádu desítek kilobajtů. Tato je mnohem rychlejší než globální, ale přístup do ní je omezen pouze pro vlákna běžící v jednom bloku. Kromě globální a sdílené paměti ještě stojí za zmínku registry. Každé vlákno má svůj vlastní registr, do kterého jsou ukládány jeho lokální proměnné. Více o CUDA architektuře naleznete v [5].



Obrázek 2: Obrázek zobrazuje příklad, jak mohou být rozdělena vlákna.

3.2 Knihovny a vývojové prostředí

Firma Nvidia dodává v SDK spolu s překladačem NVCC také multiplatformní vývojové prostředí Cuda Parallel Nsight [8] a několik knihoven. Mimo jiné jde o knihovny Cublas [4] nebo Cusparse [7]. První jmenovaná poskytuje programátorovi základní operace z lineární algebry jako třeba maticové násobení a knihovna Cusparse je zaměřena na operace s řídkými maticemi.

Kromě knihoven dodávaných od firmy Nvidia existují užitečné knihovny od třetích stran. Pro naše účely jsou velice užitečné GPU implementace běžně používané knihovny LAPACK. Existují jak komerční tak i otevřené projekty. Příkladem komerční implementace funkcí z knihovny LAPACK může být CuLa Tools (*http://www.culatools.com*), naopak příklad otevřené knihovny je MAGMA (*http://icl.cs.utk.edu/magma*).

4 Integrace CUDA do projektu

Projekt NNSU je naprogramovaný objektově v jazyce C++ a jeho překlad je řízen systémem make souborů. Jedním z podúkolů bylo integrovat podporu výpočtů na GPU do stávajícího projektu. Vzhledem k tomu, že NNSU podporuje paralelní spouštění systémem Open MPI, obsahuje zdrojové soubory, které nemohou být přeloženy obyčejným gcc překladačem, ale musí být použit obalový překladač mpixx. Více informací o paralizaci NNSU pomocí Open MPI naleznete v [2]. Situace s překladem zdrojových souborů CUDA je podobná. Pro ty je nutné použít překladač NVCC.

Bylo tedy nutné najít odpovídající make soubory a doplnit do nich cesty k překladači NVCC, knihovnám a hlavičkovým souborům, což bylo vzhledem k chybějící dokumentaci značně obtížné.

4.1 Provedené práce

Nakonec se podařilo identifikovat všechny důležité make soubory a vhodný adresář pro umístění zdrojových souborů pro GPU implementaci učení výpočetních jednotek z NSU, o kterém byla řeč v kapitole 2.2.

Byl vytvořen testovací soubor `cuda_useful.cu` obsahující pouze kernel, který zatím jen ověřuje funkčnost spuštění kódu na GPU. Tento soubor zároveň obsahuje funkci, v níž se kopírují testovací data do a z paměti grafické karty a volá kernel.

V průběhu práce byl vytvořen menší samostatný projekt na vyzkoušení práce s otevřenou knihovnou MAGMA, o které byla zmínka v sekci 3.2.

5 Zbývající práce

Nyní je potřeba využít zkušenosti s knihovnou MAGMA z menšího projektu při paralelizaci učení NSU. V kódu bude zanesena kontrola, zda je k dispozici odpovídající grafická karta. V kladném případě budou překopírována nutná data do paměti GPU, spustí se kernel a po jeho doběhnutí budou data zkopírována zpátky pro další zpracovávání na CPU.

Navíc by bylo dobré, kdyby se podařilo najít mezní velikost učící úlohy, od které se vyplatí řešit tuto úlohu na grafické kartě, aby čas potřebný na přenos dat nepřesáhl dobu výpočtu na CPU.

Literatura

[1] F. GRUAU. *Neural Network Synthesis using Cellular Encoding and The Genetic Algorithm.* Doctor Thesis, Ecole Normale Supérieure de Lyon (1994).

[2] V. Jedek. *Paralelizace distribuovaného výpočtu geneticky optimalizovaných neuronových sítí.*
Master's degree thesis, CTU Prague, (2008).

[3] R. Kalous. *Optimization of Neural Networks Architectures.*
PhD thesis, CTU Prague, (2009).

[4] Nvidia. *CUBLAS LIBRARY v5.0.* (October 2012)

[5] Nvidia. *CUDA C Programming Guide.* [ONLINE] [cit. 2013-07-20]
URL: `http://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html`

[6] Nvidia. *CUDA COMPILER v5.0.* (October 2012)

[7] Nvidia. *CUSPARSE LIBRARY v5.0.* (October 2012)

[8] Nvidia. *NSIGHT ECLIPSE DG-06450-001_ v5.0.* (October 2012)

[9] Nvidia. *What is CUDA.* [ONLINE] [cit. 2013-07-20]
URL: `https://developer.nvidia.com/what-cuda`

[10] R. C. READ. *The coding of various kinds of unlabeled trees.* In: Graph theory and computing, Library of Congress Catalog Card Number: 74-187228, Academic Press, New York (1972), 153–182.

Spectrum of Jacobi Operators and Special Functions*

František Štampach

4th year of PGS, email: stampfra@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Pavel Šťovíček, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. A brief summary of results recently published in [2, 3] is presented in this article. Both papers are devoted to spectral analysis of operators with tridiagonal matrix representation. Results concerning spectral properties of operators under investigation are expressed with the aid of hypergeometric series and their q-analogues.

Keywords: Jacobi matrix, tridiagonal operator, characteristic function, special functions, hypergeometric function

Abstrakt. V tomto příspěku jsou shrnuty hlavní výsledky publikované letos v článcích [2, 3]. Obě práce jsou věnovány spektrální analýze operatorů s tridiagonální maticovou reprezentací. Spektrální vlastnosti studovaných operátorů jsou popsány pomocí hypergeometrických řad a jejich q-analogií.

Klíčová slova: Jacobiho matice, tridiagonální operátor, charakteristická funkce, speciální funkce, hypergeometrická funkce

1 Summary

We provide a review of some results taken from [2, 3]. The main contribution and the scope of the work is pointed out. However, we do not state theorems in their full and exact form and we omit many details, for the sake of simplicity. An interested reader is referred to papers [1, 2, 3].

In [2], we define a complex function $F_{\mathcal{J}}$, called characteristic function associated with a Jacobi matrix $\mathcal{J}$ of the form

$$\mathcal{J} = \begin{pmatrix} \lambda_1 & w_1 & & & \\ w_1 & \lambda_2 & w_2 & & \\ & w_2 & \lambda_3 & w_3 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1}$$

*This work has been partially supported from the following grants: Grant No. 201/09/0811 of the Czech Science Foundation, Grant No. LC06002 of the Ministry of Education of the Czech Republic, and Grant No. GA13-11058S of the Czech Science Foundation.

where sequences $\{\lambda_n\}_{n=1}^{\infty} \in \mathbb{C}$ and $\{w_n\}_{n=1}^{\infty} \in \mathbb{C} \setminus \{0\}$ satisfy following convergence condition:

$$\sum_{n=1}^{\infty} \left| \frac{w_n^2}{(\lambda_n - z)(\lambda_{n+1} - z)} \right| < \infty, \tag{2}$$

for at least one $z \in \mathbb{C}$. Function $F_{\mathcal{J}}$ is defined with the aid of function $\mathfrak{F}$ which has been introduced in [1], for the first time. Therein, algebraic and combinatorial properties of function $\mathfrak{F}$ has been studied in detail and several other results can be found in [2].

Function $\mathfrak{F}$ is of independent interest and, besides its importance concerning characteristic function of a Jacobi matrix, it is closely related also with orthogonal polynomials, continued fractions, solutions of bilateral second order difference equation, or hypergeometric functions and their q-analogues.

With matrix $\mathcal{J}$ one associates maximal domain operator $J_{\max}$ acting on $\ell^2(\mathbb{N})$. The main result of work [2] then states the part of the spectrum of $J_{\max}$ out of the set of finite accumulation points of diagonal sequence $\{\lambda_n\}_{n=1}^{\infty}$ coincides with the set of zeros of the characteristic function $F_{\mathcal{J}}$. With the only exception of the obstacle with limit points of the diagonal of matrix $\mathcal{J}$, this is familiar situation from (finite-dimensional) linear algebra. Further, we provide explicit formulas for corresponding eigenvectors, their ℓ^2-norm (in the real case), even the Green function, and the Weyl m-function, in particular.

As an application, we present several examples of concrete Jacobi matrices where general results can be further simplified, see the illustrating example at the end of this paper. Usually, the spectrum as well as eigenvectors are described in terms of of special functions and their zeros (Bessel, Ramanujan q-Airy). Special attention is paid on properties of zeros of the Bessel function $\nu \mapsto J_{\nu}(x)$ considered as a function of its order. In particular, a new asymptotic formula for these zeros has been found.

Realizing the occurrence of special function of hypergeometric type or their q-analogues, we present even more concrete examples of Jacobi matrices with solvable spectral problem in paper [3]. However, in this paper, we go even beyond the scope of the characteristic function indicated above. There appears operator for which convergence condition (2) is violated. Moreover, the spectral analysis of an operator with doubly-infinite tridiagonal matrix representation is involved. Further, we derive several asymptotic or summation formulas with special functions (Jackson q-Bessel function, confluent hypergeometric function ${}_1F_1$, and its q-version ${}_1\phi_1$).

To illustrate a typical result we recall the following example from [3].

Proposition 1. *For $0 < q < 1$, $\sigma \in \mathbb{R}$, and $\gamma > -1$, let $J = J(\sigma, \gamma)$ be the Jacobi matrix operator in $\ell^2(\mathbb{N})$ defined by (1) where*

$$w_n = \frac{1}{2} \sinh(\sigma) q^{(n-\gamma-1)/2} \sqrt{1 - q^{n+\gamma}}\,, \quad \lambda_n = q^{n-1}. \tag{3}$$

Then $z \neq 0$ is an eigenvalue of $J(\sigma, \gamma)$ if and only if

$$\big(\cosh^2(\sigma/2) z^{-1}; q\big)_{\infty}\, {}_1\phi_1\big(q^{-\gamma} \cosh^2(\sigma/2) z^{-1}; \cosh^2(\sigma/2) z^{-1}; q, -\sinh^2(\sigma/2) z^{-1}\big) = 0.$$

Moreover, if $z \neq 0$ solves this characteristic equation then the sequence $\{v_n\}_{n=1}^{\infty}$, with

$$\begin{aligned} v_n &= q^{-\frac{1}{2}\gamma n+\frac{1}{4}n(n-3)}\,\frac{\sinh^n(\sigma)\,(2z)^{-n}}{\sqrt{(q^{\gamma+n};q)_\infty}}\left(q^n\cosh^2\left(\frac{\sigma}{2}\right)z^{-1};q\right)_\infty \\ &\quad\times {}_1\phi_1\left(q^{-\gamma}\cosh^2\left(\frac{\sigma}{2}\right)z^{-1};q^n\cosh^2\left(\frac{\sigma}{2}\right)z^{-1};q,-q^n\sinh^2\left(\frac{\sigma}{2}\right)z^{-1}\right), \end{aligned} \tag{4}$$

is a corresponding eigenvector.

In the particular case $\gamma = 0$ the characteristic equation simplifies to the form

$$\left(\cosh^2(\sigma/2)z^{-1};q\right)_\infty\left(-\sinh^2(\sigma/2)z^{-1};q\right)_\infty = 0.$$

Hence in that case, apart from $z = 0$, one knows the point spectrum fully explicitly,

$$\operatorname{spec}(J(\sigma,0))\setminus\{0\} = \left\{q^k\cosh^2(\sigma/2);\ k=0,1,2,\ldots\right\}\cup\left\{-q^k\sinh^2(\sigma/2);\ k=0,1,2,\ldots\right\}.$$

This example is particularly interesting since it gives rise to a new class of orthogonal polynomials with interesting properties, as is non-uniqueness of the orthogonality measure, providing ranges of involved parameters are chosen conveniently. This is studied by the author at present.

References

[1] F. Štampach, P. Šťovíček: *On the eigenvalue problem for a particular class of finite Jacobi matrices*, Linear Alg. Appl. **434** (2011), 1336-1353.

[2] F. Štampach, P. Šťovíček: *The characteristic function for Jacobi matrices with applications*, Linear Alg. Appl. 438 (2013) 4130-4155.

[3] F. Štampach, P. Šťovíček: *Special functions and spectrum of Jacobi matrices*, Linear Alg. Appl. (2013) in press.

On Validation of Algorithms for Dynamic Medical Data Separation

Ondřej Tichý*

4th year of PGS, email: otichy@utia.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Václav Šmídl, Institute of Information Theory and Automation, AS CR

Abstract. The problem of dynamic medical image sequence separation is studied. We introduced the state of the art algorithms for medical sequence decomposition together with those that are proposed by us. The validation and the comparison of the algorithms are nontrivial and challenging task. We propose to use a synthetic data where a ground truth is available so it is possible to compute a significant statistics for comparison reason. Moreover, we proposed a comparison on 99 real data from renal scintigraphy where relative renal functions are automatically computed and compared with those obtained by physician.

Keywords: blind source separation, deconvolution, scintigraphy, medical image sequence

Abstrakt. Tento příspěvek se zabývá zpracováním dynamických dat získaných metodou nukleární medicíny, scintigrafie. State of the art algoritmy společně s těmi, které předkládáme my, jsou představeny a diskutovány. Validace a srovnání těchto algoritmů je netriviální úloha. Nejprve navrhujeme srovnání pomocí generovaných dat, kde jsou k dispozici zdrojová data, díky kterým je možno napočítat základní statistické ukazatele výsledků. Předkládáme i srovnání algoritmů pomocí 99 reálných studií ze scintigrafie ledvin. Na těchto studiích automaticky počítáme relativní renální funkci, která může být srovnána s výsledky získanými zkušeným lékařem.

Klíčová slova: slepá separace, dekonvoluce, scintigrafie, obrazová sekvence

1 Introduction

Medical data postprocessing and analysis is important step in diagnostic medical examination. In many imaging modalities such as scintigraphy, the activity of tissues can be observed only via observing of the particles coming from radiopharmaceutical applied to the body. It can be seen the activity during the time in the respective tissues or part of the body using the method; however, several issues must be considered. Since the scintigraphical camera observed the body from one direction, the resulting image pixel is a sum of all underlying tissues. As a result, we observe a superposition of all tissues in respective region of interest (ROI). The task of medical image processing is to reconstruct the original sources of signal, i.e., tissues and their time-activity curves (TACs).

The problem is called blind source separation (BSS) and it is well described in a literature. The current methods used in practice is typically based on manual or semi-manual

*Institute of Information Theory and Automation, Department of Adaptive Systems, AS CR

selection of ROIs of the examined tissues and subtraction of the background activity [13]. More automated models can be based on model of a factor analysis (FA), [8, 7]; however, the solution of the FA is ambiguous and biological meaningfulness is not guaranteed. Other approach is based on modeling of fluid flow using compartment models such as in [5]; however, this could be too strict for biological processes and suffers from artifacts and computation tractability. In recent years, we proposed a number of probabilistic models based on FA model and solved using Variational Bayes (VB) method, [15]. The models are based on modeling both, images and TACs. We proposed (i) a modeling of TACs as results of convolutions of common input function and restricted convolution kernels, [10], (ii) modeling a probability mask on images reflecting that activity do not cover the whole image but only relatively small area [9], and (iii) model combining the advantages from both forcoming model and using the automatic relevance determination (ARD), [1], as a general principle, [11].

This paper summarize mentioned methods and focus on theirs validation and comparison methodology. The issue with validation of models is in no ground truth, no golden standard. Even physician have very different results in scintigraphy on each patient [3] or using different methodology [4]. The synthetic data can be used as an indicator of feasibility but it never reflects the nature. Comparison with physician results can be done but with consideration that manual results suffers from inaccuracy. We propose a comparison on a data from renal scintigraphy where relative renal function is automatically computed.

2 Mathematical Models

We summarize the used mathematical models in our study. All selected methods provides automatic results so they are comparable without biased interpretation.

The objective is to analyze a sequence of n images obtained at time $t = 1, \ldots, n$ and stored in vectors $\mathbf{d}_t$ with pixels stacked columnwise. The number of pixels in each image is p, thus $\mathbf{d}_t \in \mathbf{R}^p$. The important assumption is that every observed image is a linear combination of r factor images, stored in vectors $\mathbf{a}_k \in \mathbf{R}^p$, $k = 1, \ldots, r$, using the same order of pixels as in $\mathbf{d}_t$. The dimensions of the problem are typically ordered as $r < n \ll p$. Each source image has its respective time-activity curve stored in vector $\mathbf{x}_k \in \mathbf{R}^n$, $k = 1, \ldots, r$, $\mathbf{x}_k = [x_{1,k}, \ldots, x_{n,k}]'$, $\mathbf{x}'$ denotes transpose of vector $\mathbf{x}$.

We propose probabilistic formulations of this problem using several probabilistic models. The models are solved using Variational Bayes approximation [15]. The Bayes rule is given as

$$p\left(\theta|D\right) = \frac{p\left(\theta, D\right)}{p\left(D\right)} = \frac{p\left(D|\theta\right) p\left(\theta\right)}{\int p\left(D|\theta\right) p\left(\theta\right) d\theta}, \tag{1}$$

where D are observed data and θ are parameters of $p(D|\theta)$ with prior knowledge $p(\theta)$. Approximation of the Bayes rule via VB approximation can be reached as

$$p(\theta_i) \propto \exp\left(\mathsf{E}_{p(\theta_{/i})}\left(\ln\left(p\left(\theta, D\right)\right)\right)\right), i = 1, \ldots, n \tag{2}$$

Here, $\theta_{/i}$ denotes the complement of θ_i in θ and $\mathsf{E}_{p(\theta)}(g(\theta))$ denotes expected value of function $g(\theta)$ with respect to distribution $p(\theta)$. Equation (2) forms a set of implicit

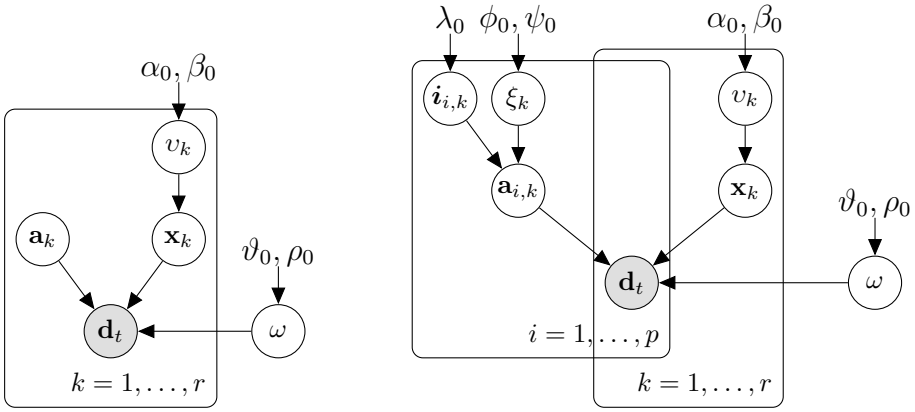


Figure 1: Hierarchical models of BSS+ (left) and FAROI (right).

equations which has to be solved iteratively.

2.1 Blind Source Separation Based on Factor Analysis

The described data sequence can be rewritten in terms of superposition, [7], as

$$\mathbf{d}_t = A\mathbf{x}_t, \tag{3}$$

where A is matrix of tissue images stored $\mathbf{a}_k$ as its columns. It is appropriate to set biologically motivated assumption such as (i) the observed data $\mathbf{d}_t$ are positive, (ii) the expected tissue images $\mathbf{a}_k$ and TACs $\mathbf{x}_k$ are also positive, (iii) the data $\mathbf{d}_t$ is strongly affected by a noise, and (iv) the number of relevant tissues, r, is unknown and should be estimated during the estimative procedure. These assumptions can be rewritten into the probabilistic model as:

$$f(\mathbf{d}_t|A, X, \omega) = \mathrm{tN}(A\mathbf{x}_t, \omega^{-1} I_p \otimes I_n), \tag{4}$$

$$f(\omega) = \mathrm{G}(\vartheta_0, \rho_0), \tag{5}$$

$$f(\mathbf{x}_k|\upsilon_k) = \mathrm{tN}(0_{n,1}, \upsilon_k^{-1} I_n), \tag{6}$$

$$f([\upsilon_1, \ldots, \upsilon_r]) = \prod_{k=1}^{r} \mathrm{G}(\alpha_{k,0}, \beta_{k,0}), \tag{7}$$

$$f(\mathbf{a}_k) = \mathrm{tN}(0_{p,1}, I_p), \tag{8}$$

where tN() denotes truncated normal distribution to positive values, G() denotes gamma distribution, I_p denotes identity matrix of the respective size, and symbol $\otimes$ denotes Kronecker product. The hierarchical model of this model is in Figure 1, left. The model will be denoted as the Blind Source Separation model with positivity constraints (BSS+).

2.2 Regions of Interest in Blind Source Separation

This model adopts the assumptions from section 2.1; however, it reflects the simple fact that tissues do not cover the whole scanned area but only a limited number of pixels.

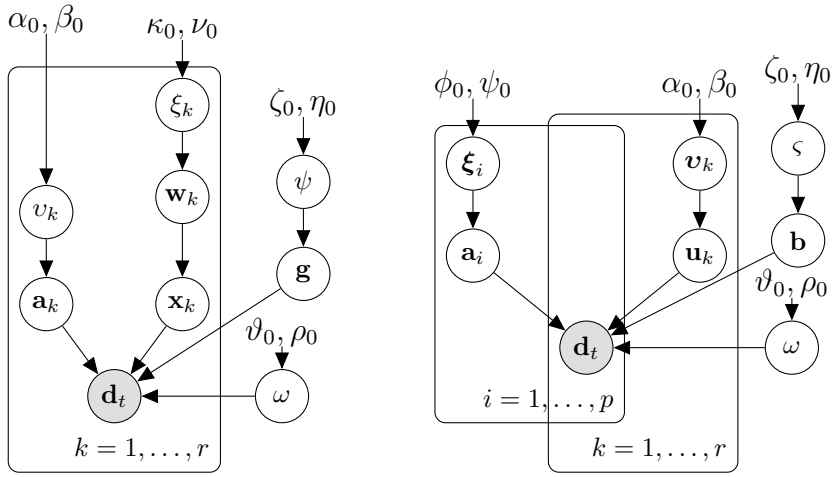


Figure 2: Hierarchical models of CFA (left) and S-BSS-DC (right).

Hence, we proposed a masking of each tissues image using indicator $\mathbf{i}$ of the same size as tissue image, [9]. This affects the model from section 2.1 as follows:

$$f(a_{i,k}|\mathbf{i}_{i,k}, \xi_k) = \mathrm{U}(0,1)^{\mathbf{i}_{i,k}} \mathrm{tN}(0, \xi_k^{-1})^{(1-\mathbf{i}_{i,k})}, \tag{9}$$

$$f(\xi_k) = \mathrm{G}(\phi_{k,0}, \psi_{k,0}), \tag{10}$$

$$f(\mathbf{i}_i, k) = \mathrm{tExp}(\lambda_{ik,0}, \langle 0, 1]), \tag{11}$$

where tExp() is truncated exponential distribution. The hierarchical model of this model is in Figure 1, right. The probabilistic masks $\mathbf{i}_k$ are estimated together with other parameters during the estimative procedure in VB method. This model will be denoted as the FAROI model (Factor Analysis with integrated ROI).

2.3 Convolution in Blind Source Separation

This model reflects the fact that each time-activity curve arise as a convolution of common input function and tissue-specific kernel, [6], such as

$$\mathbf{x}_k = \mathbf{b} * \mathbf{u}_k, \tag{12}$$

where $\mathbf{b} \in \mathbf{R}^{n \times 1}$ is input function, $\mathbf{u}_k \in \mathbf{R}^{n \times 1}$ is convolution kernel of the kth tissue, and $*$ denotes convolution. Both $\mathbf{b}$ and $\mathbf{u}_k$ are modeled as increases as vectors $\mathbf{g}$ and $\mathbf{w}_k$ respectively. This can be rewritten into the probabilistic model as [10, 12]:

$$f(\mathbf{w}_k|\xi_k) = \mathrm{tN}(M_{w_f}, \xi_k^{-1} I_n), \tag{13}$$

$$f(\xi_k) = \mathrm{G}(\kappa_{k,0}, \nu_{k,0}), \tag{14}$$

$$f(\mathbf{g}|\psi) = \mathrm{tN}(0_{n,1}, \psi^{-1} I_n), \tag{15}$$

$$f(\psi) = \mathrm{G}(\zeta_0, \eta_0), \tag{16}$$

where M_{w_f} is obtained in each iteration using clustering algorithm. The hierarchical model of this model is in Figure 2, left. This model will be denoted as the CFA model (Convolution with Factor Analysis).

2.4 Sparsity in Blind Source Separation and Deconvolution

Our latest model adopts ideas from the previous models from sections 2.1, 2.2, and 2.3. However, the assumptions of probabilistic masks, i.e. sparsity of tissue images, and of convolution are not so strict here. We use the Automatic Relevance Determination (ARD) principle, [1], to adopt the sparsity in both, tissue images and convolution kernels respectively. ARD principle is based on observation that variance of the redundant parameter tends to zero in VB solution.

The model can be written as [11]:

$$p(\mathbf{a}_i|\boldsymbol{\xi}_i) = \mathrm{tN}(\mathbf{0}_{1,r}, \mathrm{diag}(\boldsymbol{\xi}_i)^{-1}),\ i = 1, \ldots, p, \tag{17}$$

$$p(\boldsymbol{\xi}_i) = \prod_{k=1}^{r} \mathrm{G}(\phi_{ik,0}, \psi_{ik,0}), \tag{18}$$

$$p(\mathbf{b}|\varsigma) = \mathrm{tN}(0, \varsigma^{-1} I_n), \tag{19}$$

$$p(\varsigma) = \mathrm{G}(\zeta_0, \eta_0), \tag{20}$$

$$p(\mathbf{u}_k|\boldsymbol{\upsilon}_k) = \mathrm{tN}(0_{n,1}, \mathrm{diag}(\boldsymbol{\upsilon}_k)^{-1}), \tag{21}$$

$$p(\upsilon_{j,k}) = \mathrm{G}(\alpha_{jk,0}, \beta_{jk,0}),\ j = 1, \ldots, n, \tag{22}$$

where diag() denotes matrix with argument vector on its diagonal and zeros otherwise. The hierarchical model of this model is in Figure 2, right. This model will be denoted as the S-BSS-DC model (Sparsity in Blind Source Separation and Deconvolution).

2.5 CAM-CM algorithm

A complex compartment model for fMRI tumors imaging was described in [5] based on pharmacokinetic modeling using identifying representative pure pixels from each compartment in corners of cluster simplex. The algorithm is available online and is denoted as the CAM-CM algorithm.

3 Validation on Synthetic Data

Validation on synthetic data is widely used in cases when data with known ground truth are not available. This is the classical issue in the field of dynamic medical imaging including renal scintigraphy.

We propose synthetic data based on [5]. We adopt the image sources and generate our own TACs. It contains 3 image sources modeling the overlapping of all sources pairwise and shared overlap in the center. The size of images is 50×50 pixels; hence, $p = 2500$. The length of the generated sequence is 50 time steps; hence, $n = 50$. The image sources and theirs related TACs are in Figure 3, left.

We run each algorithm on this dataset. The number of expected tissues r is set to 3; hence, $r = 3$. The number of iteration is set to 100 which is reasonable for reach the convergence.

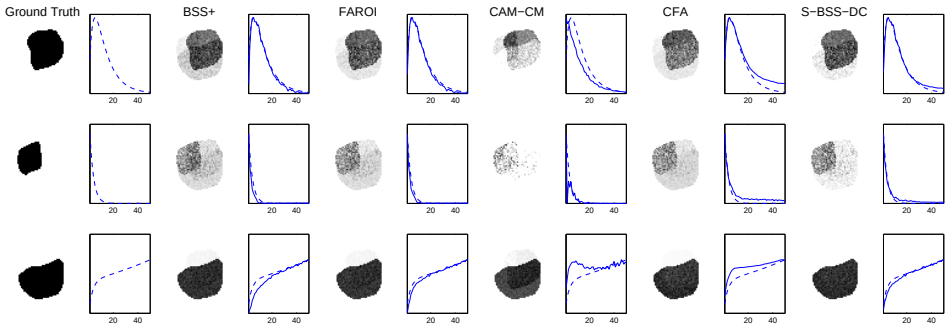


Figure 3: Results from the algorithms on synthetic dataset.

3.1 Results

The results from all algorithms are shown in Figure 3. The ground truth data are on the left and then results from algorithms (from left): BSS+, FAROI, CAM-CM, CFA, S-BSS-DC. The image sources are in the first column and the TACs are in the second. The dashed lines denotes ground truth and the full lines TACs estimated by algorithms. Note that the results are normed with respect to the activities of ground truth; hence, we study shapes, not amplitudes.

Since we have ground truth TACs, we can compute Mean square error (MSE), Mean absolute error (MAE), and Maximum error. The results is shown in Table 1. It can be seen that the computed statistics have significant explanatory value with S-BSS-DC algorithm being the best.

4 Validation on Real Data

Validation on real data is much more challenge then validation on synthetic data. Generally, we have no ground truth; hence, we can not compare results from algorithms with it. In renal scintigraphy, we have two main choices.

First, skilled operator can manually select regions contained each tissue and plot activities of the selected regions. Note that overlaps must be carefully considered. This task is extremely subjective and using of these types ground truths should be done with respect of this fact.

Second, diagnostic coefficients may be computed by a physician from the data. In renal scintigraphy, this task is very subjective too [3]. We are focused on computing of relative renal function (RRF) [2] which is a percentage of function of the left kidney and the right kidney. The RRF is estimated from the sum of activity in the left (L) and in the right (R) parenchyma during the uptake time. Then, $RRF_L = \frac{L}{L+R} \times 100\ \%$ and RRF_R can be computed analogically, both weighted by their time activity curves. Historically, the activity is taken only from the uptake time, the time when kidney accumulates activity only.

We propose a comparison on dataset [14] where RRF is computed by experienced physician. We select the sequences where both kidneys are present, i.e. 99 cases. The five

Mean Square Error					
	Algorithm				
Tissue no.	BSS+	FAROI	CAM-CM	CFA	S-BSS-DC
1	0.0061	0.0033	0.05	0.0135	0.0033
2	0.0047	0.0037	0.0205	0.0056	0.002
3	0.0455	0.0133	0.1420	0.0643	0.0095

Mean Absolute Error					
	Algorithm				
Tissue no.	BSS+	FAROI	CAM-CM	CFA	S-BSS-DC
1	0.0432	0.0416	0.1515	0.1017	0.0429
2	0.0321	0.0285	0.0363	0.0716	0.0374
3	0.1448	0.0737	0.2663	0.2208	0.0656

Maximum Error					
	Algorithm				
Tissue no.	BSS+	FAROI	CAM-CM	CFA	S-BSS-DC
1	0.4595	0.2827	0.7897	0.1684	0.2385
2	0.2651	0.2444	0.9516	0.1190	0.1589
3	0.5489	0.3569	0.8527	0.4362	0.2519

Table 1: Comparison of the algorithms on synthetic dataset is shown. The MSE, Mean error, and Maximum error are computed.

RRF estimation			
Algorithm	$<5\%$	$<10\%$	$\geqq 10\%$
BSS+	57.6%	78.8%	21.2%
FAROI	58.6%	83.8%	16.2%
CAM-CM	47.9%	63.8%	36.2%
CFA	59.6%	82.8%	17.2%
S-BSS-DC	68.7%	86.9%	13.1%

Table 2: Cumulative histogram of RRFs.

described algorithms will be compared via difference of their results of RRF computation from those provided by the experienced physician as a reference value. We will consider the automatic method that is closer to his results to be better [12].

4.1 Results

The results will be compared for BSS+, FAROI, CAM-CM, CFA, and S-BSS-DC algorithms. We use comparison over the cumulative histogram, see Table 2.

The results suggest the similar conclusion as results on synthetic data. The S-BSS-DC algorithm seems to outperform the other algorithms.

5 Conclusion

We study possibilities of comparison of algorithms for blind source separation of medical data sequence in this paper. We revise possible algorithms based on probabilistic modeling from base to more complex ones with additional assumptions. We discuss the way how to compare a performance of the algorithms. The synthetic data is proposed which provide a ground truth. It is possible to compute significant statistics using comparison of results with this ground truth. Comparison of the algorithms on real data from renal scintigraphy is more challenging task since no ground truth is available. We propose a comparison based on relative renal functions computation and comparison with those obtained from experienced physician.

We shown that the S-BSS-DC algorithm outperform other proposed algorithms in both synthetic and real data. In a future, we will prepare a comparison on directly manually selected tissue-images and related time-activity curves. It should prove the feasibility of algorithms in the best imaginable way.

References

[1] C. Bishop and M. Tipping. Variational relevance vector machines. In 'Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence', 46–53, (2000).

[2] M. Blaufox, M. Aurell, B. Bubeck, E. Fommei, A. Piepsz, C. Russell, A. Taylor, H. Thomsen, D. Volterrani, et al. *Report of the radionuclides in nephrourology com-*

mittee on renal clearance. Journal of nuclear medicine: official publication, Society of Nuclear Medicine **37** (1996), 1883.

[3] A. Brink, M. Šámal, and M. Mann. *The reproducibility of measurements of differential renal function in paediatric 99mtc-mag3 renography.* Nuclear medicine communications **33** (2012), 824–831.

[4] M. Caglar, G. Gedik, and E. Karabulut. *Differential renal function estimation by dynamic renal scintigraphy: influence of background definition and radiopharmaceutical.* Nuclear medicine communications **29** (2008), 1002.

[5] L. Chen, P. Choyke, T. Chan, C. Chi, G. Wang, and Y. Wang. *Tissue-specific compartmental analysis for dynamic contrast-enhanced mr imaging of complex tumors.* IEEE Transactions on Medical Imaging **30** (2011), 2044–2058.

[6] A. Kuruc, J. Caldicott, and S. Treves. *Improved Deconvolution Technique for the Calculation of Renal Retention Functions.* COMP. AND BIOMED. RES. **15** (1982), 46–56.

[7] J. Miskin. *Ensemble learning for independent component analysis.* PhD thesis, University of Cambridge, (2000).

[8] M. Šámal, C. Nimmon, K. Britton, and H. Bergmann. *Relative renal uptake and transit time measurements using functional factor images and fuzzy regions of interest.* European Journal of Nuclear Medicine and Molecular Imaging **25** (1997), 48–54.

[9] V. Šmídl and O. Tichý. Automatic Regions of Interest in Factor Analysis for Dynamic Medical Imaging. In '2012 IEEE International Symposium on Biomedical Imaging (ISBI)'. IEEE, (2012).

[10] V. Šmídl, O. Tichý, and M. Šámal. Factor Analysis of Scintigraphic Image Sequences with Integrated Convolution Model of Factor Curves. In 'Proceedings of the second international conference on Computational Bioscience'. IASTED, (2011).

[11] V. Šmídl and O. Tichý. *Sparsity in Bayesian Blind Source Separation and Deconvolution.* In 'Machine Learning and Knowledge Discovery in Databases', H. Blockeel et al., (ed.), volume 8189 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg (2013), 548–563.

[12] O. Tichý, V. Šmídl, and M. Šámal. Model-based Extraction of Input and Organ Functions in Dynamic Medical Imaging. In 'ECCOMAS Conference on Computational Vision and Medical Image Processing (VipImage 2013)'. Taylor and Francis, (2013). accepted.

[13] Y. Tomaru, T. Inoue, N. Oriuchi, K. Takahashi, and K. Endo. *Semi-automated renal region of interest selection method using the double-threshold technique: inter-operator variability in quantitating 99m tc-mag3 renal uptake.* European Journal of Nuclear Medicine and Molecular Imaging **25** (1997), 55–59.

[14] VFN Praha. Database of dynamic renal scintigraphy, (September 2013).

[15] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing.* Springer, (2006).

Monte Carlo Estimation of Correlation Dimension for EEG Analysis*

Lucie Tylová

2nd 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. The article presents selected options of the Monte Carlo algorithm application to study signals obtained by electroencephalographic examination (EEG). Using simulation algorithms considering the EEG signal within each measured channel as a chaotic system, there could be a faster and more efficient computation. Functionality of the Monte Carlo method is verified on the existing known systems. One of the goals is to find the appropriate characteristics and statistically significant indicators, applicable in the diagnosis of Alzheimer's disease.

Keywords: Monte Carlo method, chaotic system, EEG, Alzheimer's disease

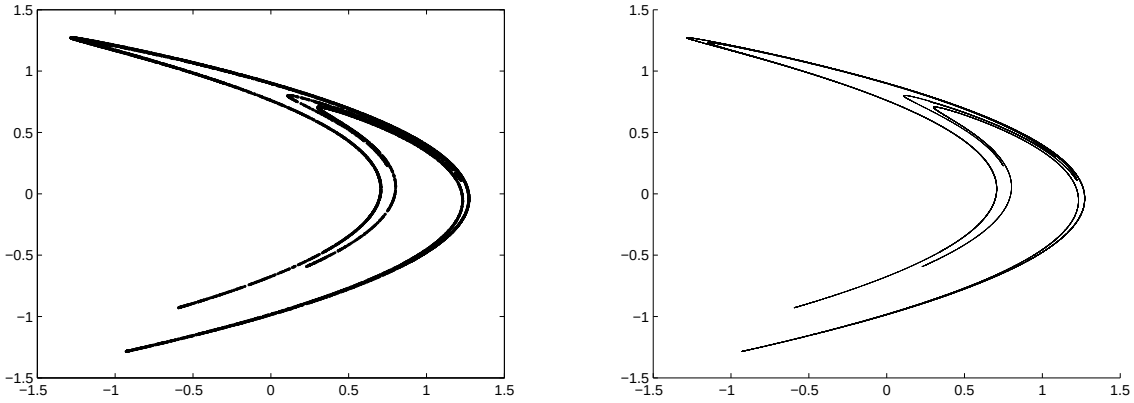
Abstrakt. Článek přináší vybrané možnosti aplikace algoritmů metody Monte Carlo při studiu signálů získaných elektroencefalografickým vyšetřením (EEG). Užitím simulačního algoritmu pohlížejícího na signál EEG v rámci jednotlivých měřených kanálů jako na chaotický systém, dochází ke zrychlení a zefektivnění výpočtu. Funkčnost metody Monte Carlo je ověřena na stávajících známých systémech. Jedním z cílů je nalézt vhodné charakteristiky a statisticky významné ukazatele, aplikovatelné v diagnostice Alzheimerovy choroby.

Klíčová slova: Monte Carlo, chaotický systém, EEG, Alzheimerova choroba

1 Introduction

There are many possibilities how to analyse EEG time series. Frequency analysis is the most popular methodology here. Another possibility is to analyse fractal properties of the time series. One of the possible characteristics of chaotic behaviour is called correlation dimension, which is based on calculations of correlation sum. In the case of EEG signal there are very large time series. Therefore, the time complexity of the correlation sum evaluation is unacceptable in real application. The novelty of this approach is in simultaneous and approximated calculations of correlation sums which makes the method applicable to real data.

*This work has been supported by the grant SGS11/165/OHK4/3T/14.



(a) Trajectory (b) Attractor

Figure 1: Hénon Discrete Dynamic System

2 Discrete Dynamic Systems

Let n $\in \mathbb{N}$ be system dimension. Let $\vec{x}_k \in \mathbb{R}^n$ be system state. Discrete dynamic system (DDS) can be driven by deterministic dynamics

$$\vec{x}_{k+1} = f(\vec{x}_k) \tag{1}$$

where f : $\mathbb{R}^n \to \mathbb{R}^n$ is continuous mapping.

Previous formula is useful for theoretical investigation and simulation of DDS. In the case of real data analysis, the state variable x_k cannot be directly observed. According to Whitney [6] and Takens [5], any state variable $y_k = x_{k,j}$ could be sampled and then the state space reconstruction calculated

$$\vec{\xi}_k = (y_{k,1}, ..., y_{k,D-1}) \tag{2}$$

where $D \in \mathbb{N}$ is embedding dimension of given DDS.

Having a knowledge of system dimension n, Whitney's theorem could be applied and directly set $D = 2n + 1$. When fractal dimension of DDS attractor D_F is known, more optimistic estimate $D > 2D_F$ according to Takens' theorem could be obtained.

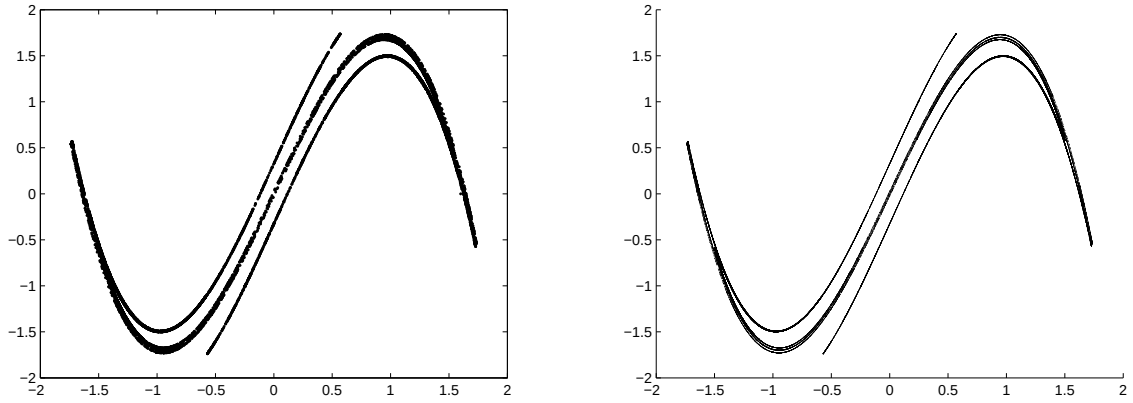
2.1 Hénon Map

Hénon system [1] is driven by formulas

$$\begin{aligned} x_{k+1,1} &= 1 - a{x^2}_{k,1} + bx_{k,2} \\ x_{k+1,2} &= x_{k,1} \end{aligned} \tag{3}$$

where usual parameters are a = 1.4 and b = 0.3.

Trajectory of Hénon DDS for $\vec{x}_0 = (0, 0.9)^T$ is depicted on Fig. 1. According to [4], fractal dimension of attractor is $D_F = 1.25827$.



(a) Trajectory (b) Attractor

Figure 2: Holmes Discrete Dynamic System

2.2 Holmes Cubic Map

Holmes system [2] is driven by formulas

$$\begin{aligned} x_{k+1,1} &= x_{k,2} \\ x_{k+1,2} &= -ax_{k,1} + bx_{k,2} - {x^3}_{k,2} \end{aligned} \tag{4}$$

where usual parameters are a = 0.2 and b = 2.77.

Trajectory of Holmes DDS for $\vec{x}_0 = (1.6, 0)^T$ is depicted on Fig. 2. According to [4], fractal dimension of attractor is $D_F = 1.26977$.

2.3 Lozi Map

Lozi system [3] is driven by formulas

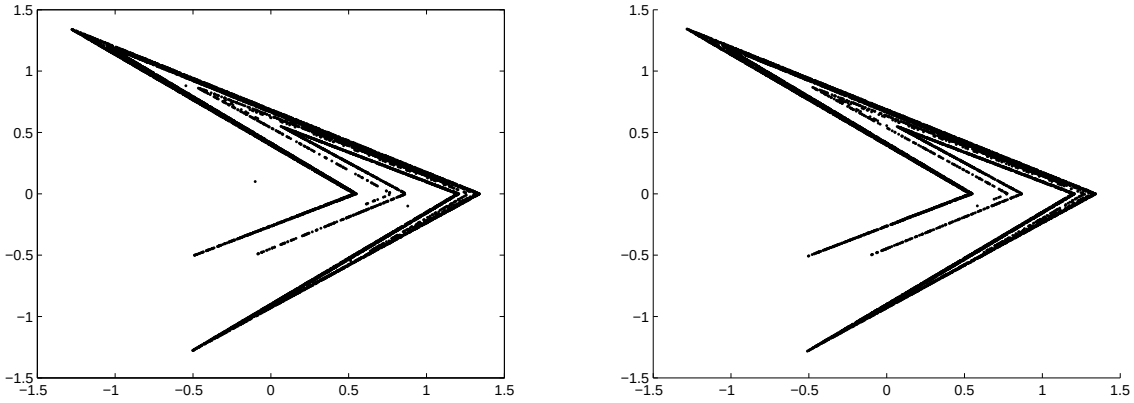
$$\begin{aligned} x_{k+1,1} &= 1 - a\,|x_{k,1}| + bx_{k,2} \\ x_{k+1,2} &= x_{k,1} \end{aligned} \tag{5}$$

where usual parameters are a = 1.7 and b = 0.5.

Trajectory of Lozi DDS for $\vec{x}_0 = (-0.1, 0.1)^T$ is depicted on Fig. 3. According to [4], fractal dimension of attractor is $D_F = 1.40419$.

2.4 Multichannel EEG Data

Electroencephalography (EEG) represents a basic electrophysiological method for examination of brain activity. The essential part of EEG registers spatio-temporal changes of brain biopotentials resulting from the continuous activity of excitatory membranes at synapses of columnarly arranged neural populations. The positive and negative charges create dipoles which are generally perpendicular to the surface of the cerebral. Sensing electrodes register the differences between particular areas.



(a) Trajectory (b) Attractor

Figure 3: Lozi Discrete Dynamic System

Most frequently used scheme of electrode placement is called 10-20, whose name corresponds to the ratio of the distances between particular electrodes. This diagram is shown in Fig. 4.

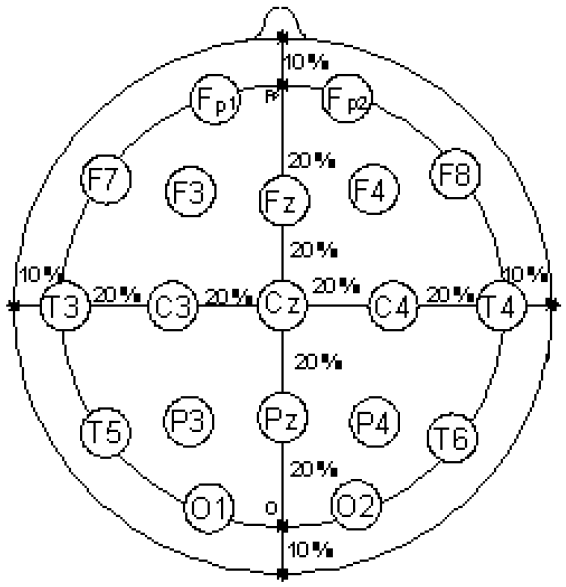


Figure 4: 10-20 scheme

3 Correlation Dimension

Correlation dimension D_2 is important characteristic of fractal structures. Their value lies between topological and Hansdorff dimensions according to inequalities

$$D_{\mathrm{T}} \leq D_2 \leq D_{\mathrm{H}} \tag{6}$$

Correlation dimension is determined by using correlation sum

$$\mathrm{C(r)} = \frac{2}{\mathrm{N(N-1)}} \sum_{\mathrm{j=1}}^{\mathrm{N}} \sum_{\mathrm{i=j+1}}^{\mathrm{N}} \Theta(\mathrm{r} - \mathrm{r_{i,j}}) \tag{7}$$

where r $>$ 0, Θ is Heaviside function, $r_{i,j} = \|\vec{x_i} - \vec{x_j}\|$ or $r_{i,j} = \|\vec{\xi_i} - \vec{\xi_j}\|$, respectively and N is a number of data points.

Correlation dimension is then defined as

$$D_2 = \lim_{r \to 0} \lim_{N \to \infty} \frac{d\mathrm{logC(r)}}{d\mathrm{logr}}. \tag{8}$$

For finite N, D_2 can be estimated via LSQ method using linearized model

$$\mathrm{logC(r)} = \mathrm{A} + \mathrm{D_2 logr}. \tag{9}$$

The main disadvantages of these approach are:

1. Unacceptable time complexity of C(r) calculations for large N,
2. unacceptable bias of D_2 estimate for small N.

Therefore, the original methodology of D_2 estimation is unacceptable in the case of EEG data analysis.

4 Numeric Experiments

4.1 Hénon Discrete Dynamic System

Monte Carlo approach was applied to Hénon DDS with $\vec{x}_0 = (0, 0.9)^T$. Time series of $\{x_{k,1}\}_{k=0}^{N}$ for $N = 10^6$ was used in state space reconstruction for $D = 5$ with $M = 10^4, 10^5, 10^6$ repetitions. Repeating approximations of C(r) are depicted on Fig. 5. Numeric estimates of capacity dimension are collected and compared with theoretical value ${D_2}^* = 1.220$ in Tab. 1.

M	ED_2	S	$z = \frac{\mathrm{E}D_2 - {D_2}^*}{S}$	p-value
10^4	1.2237	0.0404	0.9163	0.3617
10^5	1.2138	0.0140	-4.4075	2.6554×10^{-5}
10^6	1.2155	0.0050	-9.1118	9.5479×10^{-15}

Table 1: Correlation dimension for Hénon DDS

4.2 Holmes Discrete Dynamic System

Holmes DDS was approached with $\vec{x}_0 = (1.6, 0)^T$. Time series of $\{x_{k,1}\}_{k=0}^{N}$ for $N = 10^6$ was used in state space reconstruction for $D = 5$ with $M = 10^4, 10^5, 10^6$ repetitions too. Repeating approximations of C(r) are depicted on Fig. 6. Numeric estimates of capacity dimension are collected and compared with theoretical value ${D_2}^* = 1.260$ in Tab. 2.

M	ED_2	S	$z = \frac{\mathrm{E}D_2 - {D_2}^*}{S}$	p-value
10^4	1.2455	0.0770	-1.8834	0.0626
10^5	1.2536	0.0248	-2.5733	0.0116
10^6	1.2564	0.0076	-4.7804	6.0912×10^{-6}

Table 2: Correlation dimension for Holmes DDS

4.3 Lozi Discrete Dynamic System

Lozi DDS was approached with $\vec{x}_0 = (-0.1, 0.1)^T$. Time series of $\{x_{k,1}\}_{k=0}^{N}$ for $N = 10^6$ was used in state space reconstruction for $D = 5$ with $M = 10^4, 10^5, 10^6$ repetitions. Repeating approximations of C(r) are depicted on Fig. 7. Numeric estimates of capacity dimension are collected and compared with theoretical value ${D_2}^* = 1.384$ in Tab. 3.

M	ED_2	S	$z = \frac{\mathrm{E}D_2 - {D_2}^*}{S}$	p-value
10^4	1.3490	0.0636	-4.5001	2.9737×10^{-2}
10^5	1.3508	0.0212	-8.6321	1.4892×10^{-6}
10^6	1.3670	0.0101	-12.8742	7.9621×10^{-10}

Table 3: Correlation dimension for Lozi DDS

5 Case Study: Alzheimer's Disease Diagnosis

For the case study, a group of 165 patients has been used, from which 24 were affected by Alzheimer's disease (AD) and 139 were with control normal (CN). The data was recorded using the standard 10-20 scheme, thus values of 19 channels were obtained. The sampling frequency was 200 Hz and patients were measured for 5 minutes.

D_2 values of each channel for all AD and CN patients were studied. The model parameters were M = 10^6 repetitions, D = 19 for state space reconstruction, $p_min = 0.59$, $p_max = 0.6$, and minimum sample distance $\Delta = 200$.

Tab. 4 summarizes final results. The best $p_{\text{value}} = 0.0049$ was obtained for second channel and $p_{\text{value}} = 0.0306$ for sixth channel which correspond to frontal electrodes. The result confirms former research results that significant differences between AD and CN groups could be recognized in the case of frontal electrodes.

References

[1] M. Hénon. *A two-dimensional mapping with a strange attractor.* Communication in Mathematical Physics **50**, 69-77 (1976).

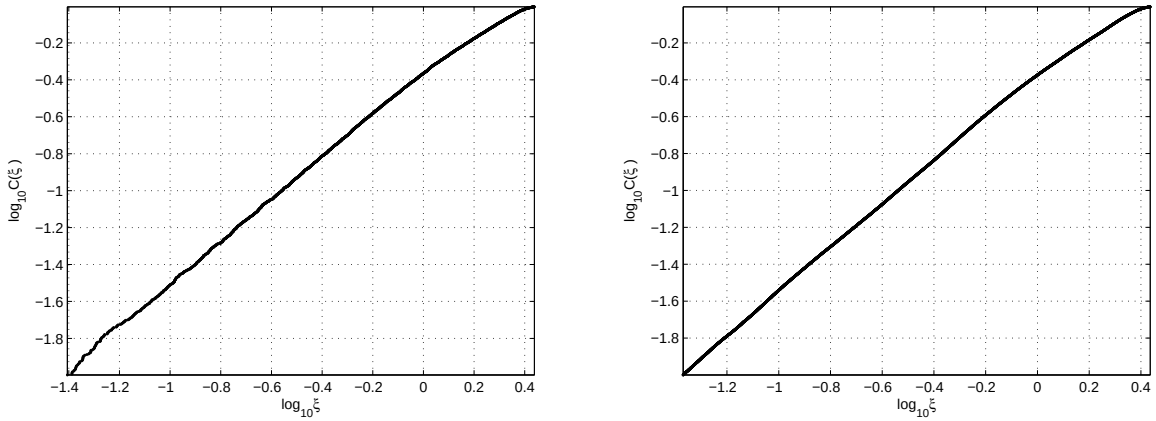
[2] P. J. Holmes. *A nonlinear oscillator with a strange attractor.* Philosophical Transactions of the Royal Society of London Series A **292**, 419-48 (1979).

[3] R. Lozi. *Un attracteur etrange? Du type attracteur de Hénon.* Journal de Physique **39**, 9-10 (1978).

[4] J. C. Sprott. *Chaos and Time-Series Analysis.* Oxford University Press (2003).

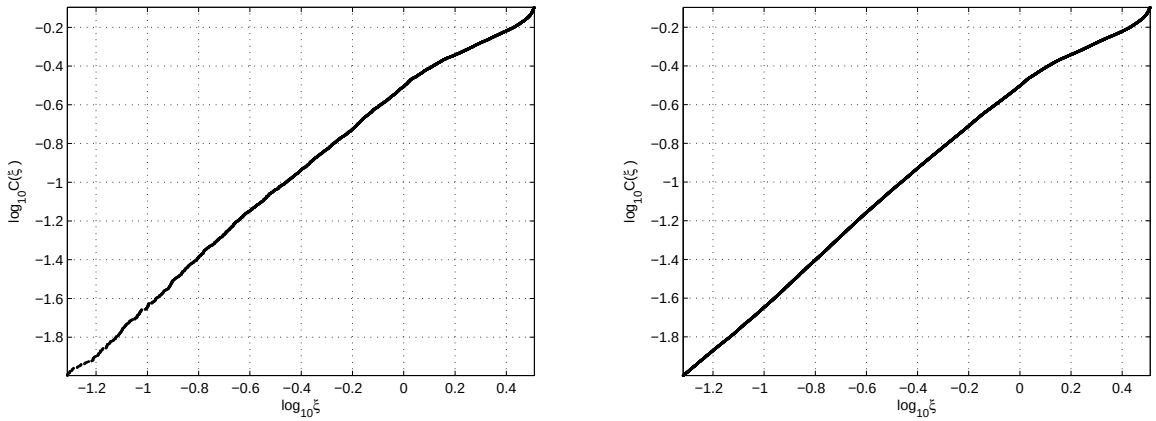
[5] F. Takens. *Detecting strange attractors in turbulence.* Dynamical systems and turbulence. Lecture Notes in Mathematics, No. 898, pp. 366-81. Springer (1981).

[6] H. Whitney. *Differentiable manifolds.* Annals of Mathematics **37**, 648-80 (1936).



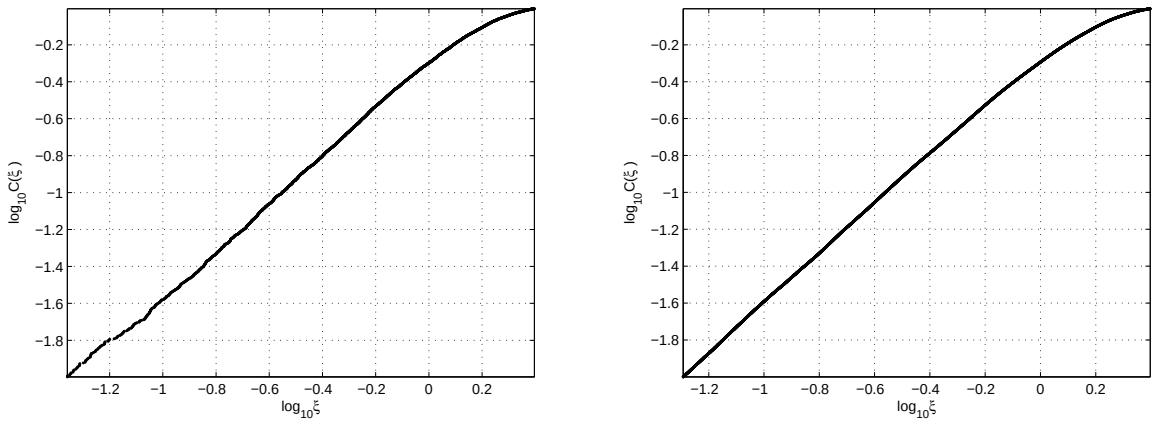
(a) Repetitions M = 10^4 (b) Repetitions M = 10^6

Figure 5: Hénon Discrete Dynamic System



(a) Repetitions M = 10^4 (b) Repetitions M = 10^6

Figure 6: Holmes Discrete Dynamic System



(a) Repetitions M = 10^4 (b) Repetitions M = 10^6

Figure 7: Lozi Discrete Dynamic System

Ch	D_2 CN		D_2 AD		p_{value}
	mean	std	mean	std	
1	1.4121	0.3922	1.2535	0.4540	0.0669
2	1.3829	0.3715	1.1492	0.4419	0.0049
3	1.4158	0.3698	1.4877	0.4557	0.3828
4	1.5757	0.3915	1.5150	0.4245	0.4752
5	1.6267	0.3742	1.4752	0.4439	0.0678
6	1.6090	0.3812	1.4227	0.4894	0.0306
7	1.4712	0.4206	1.3994	0.4745	0.4353
8	1.6114	0.4834	1.6329	0.5124	0.8365
9	1.7073	0.3872	1.6653	0.4539	0.6223
10	1.6895	0.4074	1.6035	0.4792	0.3381
11	1.6950	0.3990	1.6408	0.4389	0.5321
12	1.6654	0.5035	1.5828	0.5583	0.4519
13	1.6001	0.3848	1.6117	0.4267	0.8900
14	1.6459	0.3641	1.6027	0.4522	0.5942
15	1.6541	0.3456	1.5928	0.5132	0.4465
16	1.6450	0.3534	1.6354	0.4307	0.9019
17	1.6015	0.3911	1.6038	0.5378	0.9797
18	1.5791	0.3722	1.5076	0.4729	0.3913
19	1.5855	0.3560	1.5942	0.4405	0.9128

Table 4: Comparison of D_2 value of AD and CN patients

Arithmetical Aspects of a Number System with Negative Tribonacci Base*

Tomáš Vávra

2nd year of PGS, email: t.vavra@seznam.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Zuzana Masáková, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. We study arithmetical aspects of Ito-Sadahiro number systems with negative base. We present an effective algorithm for addition when the base is $-\gamma$ where $\gamma > 1$ is the tribonacci constant, the root of $x^3 - x^2 - x - 1$. In particular, we show that addition can be done by a finite state transducer. As a consequence of the structure of the transducer, we show that γ posseses the so-called finiteness property. Moreover, we determine the maximal number of fractional digits arising from addition of two $(-\gamma)$-integers.

Keywords: negative base, number system, tribonacci

Abstrakt. Příspěvek se zabývá aritmetickými vlastnostmi číselných soustav se záporným základem. Předvedeme efektivní sčítací algoritmus pro případ, že základem je $-\gamma$, kde $\gamma > 1$ je takzvaná tribonacciho konstanta, kořen $x^3 - x^2 - x - 1$. Přesněji řečeno, ukážeme, že sčítání může být provedeno konečným překladačem. Následně pak, jako důsledek struktury překladače, ukážeme, že γ má takzvanou vlastnost (-F). Navíc určíme počet zlomkových míst vznikajících při sčítání dvou $(-\gamma)$-celých čísel.

Klíčová slova: záporná báze, numerační systém, tribonacci

1 Introduction

Numeration systems with negative non-integer base received a non-negligible attention since the paper [5] of Ito and Sadahiro in 2009. Since then there have been written several papers concerning arithmetical aspects of such number systems with a Pisot base (see [6], [7], [1]).

It has been shown in [1] that the negative base number system posseses interesting properties when the base is taken to be root of

$$x^k - mx^{k-1} - \cdots - mx - n, \ m \geq n \geq 1 \quad \text{and } m = n \text{ for } k \text{ even}. \tag{1}$$

The most interesting of those properties is that the set of $(-\beta)$-integers coincides with the set

$$X(-\beta) = \Big\{ \sum_{i=0}^{n} a_i(-\beta)^i \mid a_i \in \{0, 1, \ldots, \lfloor \beta \rfloor\} \Big\},$$

*This work was supported by the Grant Agency of the Czech Technical University in Prague grant SGS11/162/OHK4/3T/14 and Czech Science Foundation grant 13-03538S.

i.e. the set of linear combinations of non-negative powers of $(-\beta)$ with coefficients in the canonical alphabet, even though the string $a_n a_{n-1} \dots a_0$ may be forbidden in the corresponding number system. An analogous result for positive based number systems comes from Ch. Frougny [3]. Another interesting result from [4] is that roots of (1) posses the so-called Property (F), namely that the set $\mathrm{Fin}(\beta)$ of numbers with finite expansion forms a ring.

As we will see, unlike in β-expansions, roots of (1) do not have Property (-F) with the exception of roots of $x^{2k+1} - mx^{2k} - \cdots - mx - m$. In particular the set $\mathrm{Fin}(-\beta)$ is not closed under addition. In this work we show that the tribonacci constant, i.e. the positive root of $x^3 - x^2 - x - 1$ has Property (-F). The proof is done by providing an algorithm for addition which is probably the first effective arithmetical algorithm for negative base number systems.

2 Preliminaries

The Ito-Sadahiro number system is a numeration system analogous to Rényi β-expansions which uses a negative base. Instead of defining the expansions of numbers from $[0, 1)$ first, the unit interval $[\ell, \ell + 1)$ with $\ell = \frac{-\beta}{\beta+1}$ was chosen. For $-\beta < -1$, any $x \in [\ell, \ell + 1)$ has a unique expansion of the form $d_{-\beta}(x) = x_1 x_2 x_3 \cdots$ defined by

$$x_i = \lfloor -\beta T_{-\beta}^{i-1}(x) - \ell \rfloor, \text{ where } T_{-\beta}(x) = -\beta x - \lfloor -\beta x - \ell \rfloor .$$

For any $x \in [\ell, \ell + 1)$ we obtain an infinite word from $\mathcal{A}^{\mathbb{N}} = \{0, 1, \dots, \lfloor \beta \rfloor\}^{\mathbb{N}}$.

Another analogous concept is the $(-\beta)$-admissibility, which characterizes all digit strings over $\mathcal{A}$ being the $(-\beta)$-expansion of some number. The lexicographic condition, similar to the one by Parry, was also proved in [5]. Ito and Sadahiro proved that a digit string $x_1 x_2 x_3 \cdots \in \mathcal{A}^{\mathbb{N}}$ is $(-\beta)$-admissible (or, if no confusion is possible, just admissible) if and only if it fulfills the lexicographic condition

$$d_{-\beta}(\ell) \preceq_{\mathrm{alt}} x_i x_{i+1} x_{i+2} \cdots \prec_{\mathrm{alt}} d^*_{-\beta}(\ell + 1) = \lim_{y \to l+1_-} d_{-\beta}(y) \text{ for all } i \geq 1 . \tag{2}$$

Here, the limit is taken over the product topology on $\mathcal{A}^{\mathbb{N}}$ and $\prec_{\mathrm{alt}}$ stands for alternate lexicographic ordering defined as follows:

$$u_1 u_2 \cdots \prec_{\mathrm{alt}} v_1 v_2 \cdots \;\Leftrightarrow\; (-1)^k (u_k - v_k) < 0 \text{ for } k \text{ smallest such that } u_k \neq v_k .$$

In analogy with β-numeration, the alternate ordering corresponds to the ordering on reals in $[\ell, \ell + 1)$, i.e. $x < y \;\Leftrightarrow\; d_{-\beta}(x) \prec_{\mathrm{alt}} d_{-\beta}(y)$.

The reference digit strings $d_{-\beta}(\ell)$ and $d^*_{-\beta}(\ell{+}1)$ play the same role for $(-\beta)$-expansions as Rényi expansions of unity for β-expansions. While $d_{-\beta}(\ell)$ is obtainable directly from the definition, the following rule (proved in [5]) is to be used for determining $d^*_{-\beta}(\ell + 1)$:

$$d^*_{-\beta}(\ell + 1) = \begin{cases} (0 l_1 \cdots l_{q-1}(l_q - 1))^\omega & \text{if } d_{-\beta}(\ell) = (l_1 l_2 \cdots l_q)^\omega \text{ for } q \text{ odd}, \\ 0 d_{-\beta}(\ell) & \text{otherwise.} \end{cases}$$

We can now recall the definition of $(-\beta)$-expansions for all reals.

Definition 1. *Let $-\beta < -1$, $x \in \mathbb{R}$. Let $k \in \mathbb{N}$ be minimal such that $\frac{x}{(-\beta)^k} \in (\ell, \ell+1)$ and $d_{-\beta}\left(\frac{x}{(-\beta)^k}\right) = x_1 x_2 x_3 \cdots$. Then the $(-\beta)$-expansion of x is defined as*

$$\langle x \rangle_{-\beta} = \begin{cases} x_1 \cdots x_{k-1} x_k \bullet x_{k+1} x_{k+2} \cdots & \text{if } k \geq 1, \\ 0 \bullet x_1 x_2 x_3 \cdots & \text{if } k = 0. \end{cases}$$

Similarly as in a positive base numeration, the set of $(-\beta)$-integers $\mathbb{Z}_{-\beta}$ can now be defined using the notion of $\langle x \rangle_{-\beta}$. Since the base is negative, we can now represent any real number without the need of a minus sign.

Definition 2. *Let $\beta > 1$. Then the sets of $(-\beta)$-integers and of numbers with finite $(-\beta)$-expansions are defined as*

$$\mathbb{Z}_{-\beta} = \{x \in \mathbb{R} \mid \langle x \rangle_{-\beta} = x_k \cdots x_1 x_0 \bullet 0^\omega\} \;=\; \bigcup_{i \geq 0} (-\beta)^i T_{-\beta}^{-i}(0)\,,$$

$$\mathrm{Fin}(-\beta) = \{x \in \mathbb{R} \mid \langle x \rangle_{-\beta} = x_k \cdots x_1 x_0 \bullet x_{-1} \ldots x_{-n} 0^\omega\} \;=\; \bigcup_{i \geq 0} (-\beta)^{-i} \mathbb{Z}_{-\beta}\,.$$

We say that β has Property (-F) *if* $\mathrm{Fin}(-\beta)$ *is a ring.*

3 Arithmetics in Ito-Sadahiro number systems

Let us recall that it has been shown in [1] that

$$\mathbb{Z}_{-\beta} = \Big\{ \sum_{i=0}^{n} a_i (-\beta)^i \mid a_i \in \{0, 1, \ldots, \lfloor \beta \rfloor\} \Big\}, \tag{3}$$

if and only if β is a root of

$$x^k - m x^{k-1} - \cdots - mx - n, \;\; m \geq n \geq 1 \quad \text{and } m = n \text{ for } k \text{ even}. \tag{4}$$

Such bases are promising candidates for having Property (-F). For it suffices to show that $x + 1 \in \mathrm{Fin}(-\beta)$ for any $x \in \mathrm{Fin}(-\beta)$, and that $-1 \in \mathrm{Fin}(-\beta)$. Because of the property (3), it means one has to show that any admissible string with $+1$ added to the position where maximal digit lies can be rewritten as a finite string over the alphabet $\{0, 1, \ldots, \lfloor \beta \rfloor\}$.

However, there are examples that this procedure is not possible for almost all roots of (4). We have

1. $$\langle m+1 \rangle_{-\beta} = 1m0 \bullet 0^{2k-3} 11m \big[0^{2k-3}110\big]^\omega$$
   for β root of $x^{2k} - mx^{2k-1} - \cdots - mx - m$;

2. $$\langle \beta + m + 1 \rangle = (m-n+1)(m-n+1)1(n+1)^\omega$$
   for β root of $x^3 - mx^2 - mx - n$, and

3.
$$\langle -\beta + m + 1\rangle_{-\beta} = 0 \bullet 0^{2k-1}(m-n+1)(m-n+1)0\big[0^{2k-3}1(n+1)n\big]^{\omega}$$
for β root of $x^{2k+1} - mx^{2k} - \cdots - mx - n,\ n < m, k \geq 2$.

In case of β being of odd degree and $n = m$, we have no counterexample. Later we will show that there is no such example for the root of $x^3 - x^2 - x - 1$, the tribonacci constant. We provide a transucer whose input is a digit-wise sum of $x, y \in \mathrm{Fin}(-\beta)$ and output is a representation of $x + y$ over $\{0, 1\}$.

Theorem 3. *Let $\gamma > 1$ be the root of $x^3 - x^2 - x - 1$. Then for any $x, y \in \mathrm{Fin}(-\beta)$ the computation of a representation of $x + y$ over the alphabet $\{0, 1\}$ can be done by a finite-state transducer.*

Proof. We define a tranducer $(S, s_0, \Sigma, \Lambda, T)$ where

- $S \subset \{\overline{2}, \overline{1}, 0, 1, 2, 3\}^3$ is the set of states;
- $s_0 = 000$ is the initial state;
- $\Sigma = \{0, 1, 2\} \cup \{0, 1, 2\}^2$ is the input alphabet;
- $\Lambda = \{0, 1\} \cup \{0, 1\}^2$ is the output alphabet;
- $T : S \times \Sigma \to \Lambda \times S$ is the transition function defined by the transitions in the list below.

The notation $s_1|a \to b|s_2$ means that the machine is in the state s_1 and reads symbol(s) a from the input tape, then it switches to the state s_2 and writes symbol(s) b onto the output tape. In fact, the machine reads the digit-wise sum of two numbers from the left side, looks only at four or five symbols wide window, and, if needed, adds a representation of zero. Then it moves the window to the right.

One can verify that the transition function defined bellow does not change the numerical value of the string since each image is obtained by adding or subtracting representation of zero, namely $0 = 11\overline{1}1$ which follows from the minimal polynomial for γ. Here $\overline{a}$ stands for $-a$. Moreover, with one exception, the transitions from any state are defined for any input symbol from Σ. The exception is the state 101 that cannot be escaped by reading symbol 0. However, the only path to the state 101 leads from 131 by reading 202 on the input (see Figure 1). Reading 2020 would mean that both x and y contain forbidden string 1010 that can be avoided (we have $d_{-\beta}(\ell) = 101^{\omega}$). Hence we assume that at least one summand does not contain 1010.

$000\|0 \to 0\|000$	$001\|2 \to 0\|012$	$003\|1 \to 1\|122$
$000\|1 \to 0\|001$	$002\|0 \to 1\|111$	$003\|2 \to 1\|123$
$000\|2 \to 0\|002$	$002\|1 \to 1\|112$	$00\overline{1}\|0 \to 0\|0\overline{1}0$
$001\|00 \to 00\|100$	$002\|21 \to 11\|131$	$00\overline{1}\|1 \to 0\|0\overline{1}1$
$001\|01 \to 11\|011$	$002\|22 \to 11\|132$	$00\overline{1}\|2 \to 0\|0\overline{1}2$
$001\|02 \to 11\|012$	$002\|20 \to 00\|220$	$011\|0 \to 0\|110$
$001\|1 \to 0\|011$	$003\|0 \to 1\|121$	$011\|1 \to 0\|111$

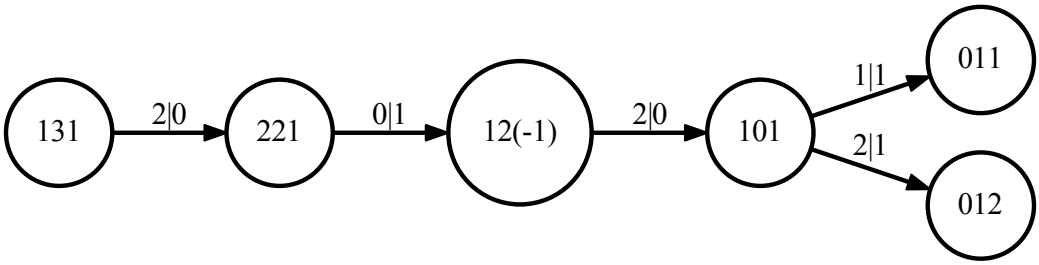


Figure 1: Transitions leading to the state 101

$011|2 \to 0|112$
$012|0 \to 0|120$
$012|1 \to 0|121$
$012|2 \to 0|122$
$013|00 \to 00|21\overline{1}$
$013|01 \to 11|120$
$013|02 \to 11|121$
$013|1 \to 0|131$
$013|2 \to 0|132$
$023|0 \to 0|230$
$023|1 \to 0|231$
$023|2 \to 0|232$
$0\overline{1}0|0 \to 1|0\overline{1}1$
$0\overline{1}0|1 \to 1|0\overline{1}2$
$0\overline{1}0|20 \to 00|111$
$0\overline{1}0|21 \to 00|112$
$0\overline{1}0|22 \to 11|023$
$0\overline{1}1|0 \to 1|001$
$0\overline{1}1|1 \to 1|002$
$0\overline{1}1|2 \to 1|003$
$0\overline{1}2|0 \to 1|011$
$0\overline{1}2|1 \to 1|012$
$0\overline{1}2|2 \to 1|013$
$0\overline{11}|00 \to 00|0\overline{1}1$
$0\overline{11}|01 \to 00|0\overline{1}2$
$0\overline{11}|02 \to 11|\overline{1}03$
$0\overline{11}|1 \to 0|\overline{11}1$
$0\overline{11}|2 \to 0|\overline{11}2$
$100|0 \to 1|000$
$100|1 \to 1|001$
$100|2 \to 1|002$
$101|1 \to 1|011$
$101|2 \to 1|012$
$10\overline{1}|0 \to 1|0\overline{1}0$
$10\overline{1}|1 \to 1|0\overline{1}1$
$10\overline{1}|2 \to 1|0\overline{1}2$
$110|0 \to 1|100$
$110|10 \to 00|100$
$110|11 \to 11|011$
$110|12 \to 11|012$
$110|2 \to 0|011$
$111|0 \to 1|110$
$111|1 \to 1|111$
$111|2 \to 1|112$
$112|0 \to 1|120$
$112|1 \to 1|121$
$112|2 \to 1|122$
$11\overline{1}|0 \to 0|00\overline{1}$
$11\overline{1}|1 \to 0|000$
$11\overline{1}|2 \to 0|001$
$11\overline{2}|0 \to 0|0\overline{11}$
$11\overline{2}|1 \to 0|0\overline{1}0$
$11\overline{2}|2 \to 0|0\overline{1}1$
$120|0 \to 0|11\overline{1}$
$120|1 \to 0|110$
$120|2 \to 0|111$
$121|00 \to 00|10\overline{1}$
$121|01 \to 00|100$
$121|02 \to 11|011$
$121|1 \to 0|120$
$121|2 \to 0|121$
$122|0 \to 1|220$
$122|10 \to 00|21\overline{1}$
$122|11 \to 11|120$
$122|12 \to 11|121$
$122|2 \to 0|131$
$123|0 \to 1|230$
$123|1 \to 1|231$
$123|2 \to 1|232$
$12\overline{1}|0 \to 0|10\overline{1}$
$12\overline{1}|1 \to 0|100$
$12\overline{1}|2 \to 0|101$
$131|0 \to 0|22\overline{1}$
$131|1 \to 0|220$
$131|2 \to 0|221$
$132|0 \to 0|23\overline{1}$
$132|1 \to 0|230$
$132|2 \to 0|231$
$21\overline{1}|0 \to 1|00\overline{1}$
$21\overline{1}|1 \to 1|000$
$21\overline{1}|2 \to 1|001$
$220|0 \to 1|11\overline{1}$
$220|1 \to 1|110$
$220|2 \to 1|111$
$221|0 \to 1|12\overline{1}$
$221|1 \to 1|120$
$221|2 \to 1|121$
$22\overline{1}|0 \to 1|10\overline{1}$
$22\overline{1}|1 \to 1|100$
$22\overline{1}|20 \to 00|100$
$22\overline{1}|21 \to 11|011$
$22\overline{1}|22 \to 11|012$
$230|10 \to 00|10\overline{1}$

$230|11 \to 00|100$
$230|12 \to 11|011$
$230|0 \to 1|21\overline{1}$
$230|2 \to 0|120$
$231|0 \to 1|22\overline{1}$
$231|1 \to 1|220$
$231|20 \to 00|21\overline{1}$
$231|21 \to 11|120$
$231|22 \to 11|121$
$232|0 \to 1|23\overline{1}$
$232|1 \to 1|230$
$232|2 \to 1|231$
$23\overline{1}|0 \to 0|11\overline{2}$
$23\overline{1}|1 \to 0|11\overline{1}$
$23\overline{1}|2 \to 0|110$
$\overline{1}03|0 \to 0|121$
$\overline{1}03|1 \to 0|122$
$\overline{1}03|2 \to 0|123$
$\overline{11}|0 \to 0|001$
$\overline{11}|1 \to 0|002$
$\overline{11}|2 \to 0|003$
$\overline{11}2|0 \to 0|011$
$\overline{11}2|1 \to 0|012$
$\overline{11}2|2 \to 0|013$

□

Remark 4. *It follows from the proof that algorithm can be extended to adding of more than two numbers with finite expansion by adding numbers consecutively. However, we always have to add an admissible string since the automaton may not accept the string* 2020.

Also, an extension to periodic expansions is possible. Since the digit-wise of two periodic representations is also periodic, the period of the result can be recognized by looking at the states in which the transducer is when the repetition of the period is being read.

The proof of Theorem 3 gives us two important consequences. The first is that γ posseses Property (-F). Closeness of $\mathrm{Fin}(-\beta)$ under addition can be seen from the subgraph of the transducer on Figure 2. It shows that when infinite repetition of zeros is on the input, the infinite repetition of zeros eventually appears also on the output. Although the representation obtained from the transducer may not be admissible, property (3) ensures that the expansion is also finite. Moreover, subtraction can be represented as addition since $\overline{1}\bullet = 11 \bullet 001$. This leads to the following theorem.

Theorem 5. *The tribonacci constant has Property (-F).*

Often observed property is the number of fractional points arising from addition of two $(-\beta)$-integers. We can determine this number again from Figure 2. One can see that when reading only zeros, the last nonzero digit is sent to the output at ninth position, i.e. six positions after the franctional point. For example, $112 \bullet 0^\omega = 100 \bullet 0110010^\omega$. Since the latter representation is admissible, this also shows that this bound can be reached.

Theorem 6. *Let β be the tribonacci constant. Then the number of fractional points arising from additon of two $(-\gamma)$-integers is at most 6. This bound is strict.*

References

[1] D. Dombek, Z. Masáková, T. Vávra, *Confluent Pisot bases in negative base number systems*, preprint (2013) 23pp.

[2] K. Dajani, M. de Vries, V. Komornik, P. Loreti. *Optimal expansions in non-integer bases.* Proc. Amer. Math. Soc. **140** (2012), 437–447.

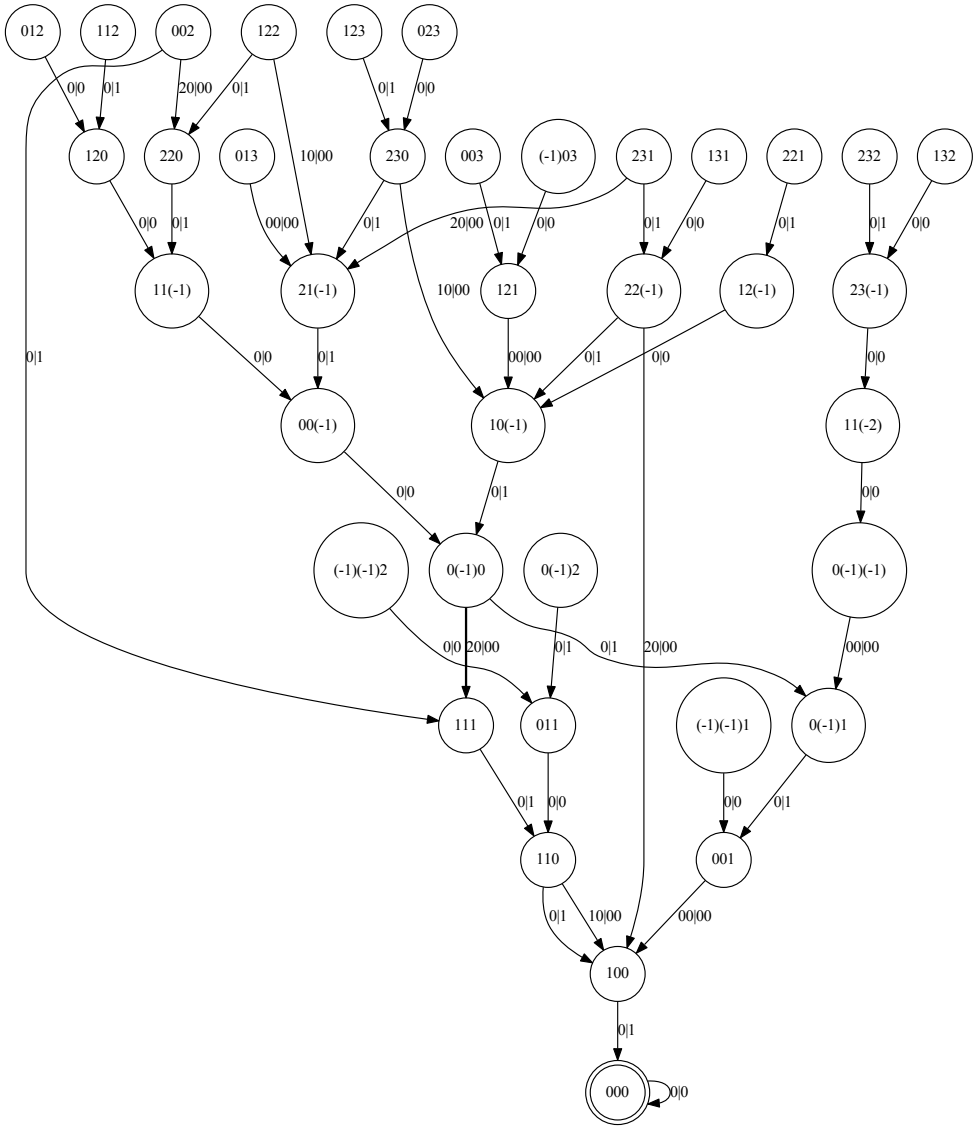


Figure 2: Subgraph induced by reading zeros at the input

[3] Ch. Frougny, *Confluent Linear Numeration Systems*, Theor. Comput. Sci. 106 (2) (1992), 183–219.

[4] Ch. Frougny, B. Solomyak, *Finite beta-expansions*, Ergodic Theory Dynam. Systems 12 (1992), no. 4, 713–723.

[5] S. Ito and T. Sadahiro, *Beta-expansions with negative bases*, INTEGERS 9 (2009), 239–259. doi:10.1515/INTEG.2009.023.

[6] Z. Masáková, E. Pelantová, T. Vávra, *Arithmetics in number systems with negative base*, Theor. Comp. Sci. 412 (2011), 835–845.

[7] Z. Masáková, T. Vávra, *Integers in number systems with positive and negative quadratic Pisot base*, preprint (2013), 20pp. arXiv:1302.4655

On the Generalized Geometry Origin of Noncommutative Gauge Theory*

Jan Vysoký

3rd 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

Abstract. We discuss noncommutative gauge theory from the generalized geometry point of view. We argue that the equivalence between the commutative and semiclassically noncommutative DBI actions is naturally encoded in the generalized geometry of D-branes.

Keywords: Generalized geometry, noncommutative gauge theory, generalized metric, effective action

Abstrakt. V článku se zaobíráme nekomutativní kalibrační teorií z pohledu zobecněné geometrie. Ukážeme, že ekvivalence komutativní a semiklasicky nekomutativní DBI akce je přirozeně zakódována v zobecněné geometrii D-bran.

Klíčová slova: Zobecněná geometrie, nekomutativní kalibrační teorie, zobecněná metrika, efektivní akce

1 Introduction

Generalized geometry [14, 13] recently appeared to be a powerful mathematical tool for the description of various aspects of string and field theories. Here we mention only few instances of its relevance that are more or less directly related to the present paper. Topological and non-topological Poisson sigma models are known to be intimately related to a lot of interesting differential, in particular generalized, geometry. For instance, the topological Poisson sigma models are of interest for the integration of Poisson manifolds (and Lie algebroids) [7] and are at the heart of deformation quantization [9]. Field equations of (topological) Poisson sigma models can be interpreted as Lie algebroid morphisms [4] and as such can further be generalized in terms of generalized (complex) geometry [20], [19]. Poisson sigma models can be twisted by a 3-form H-field [18] and also generalized to Dirac sigma models [19], where the graph defined by the corresponding (possibly twisted) Poisson structure is replaced by a more general Dirac structure. In turn, at least in some instances, D-branes can be related to Dirac structures [22], [2], or coisotropic submanifolds [8]. In [1], it has been observed that the current algebra of sigma models naturally involves structures of generalized geometry, such as the Dorfman bracket and Dirac structures.

*Excerpts from the paper published with Branislav Jurčo and Peter Schupp.

2 Generalized geometry

2.1 Fiberwise metric, generalized metric

In this section we recall some basic facts regarding generalized geometry, see, e.g., [13], [5]. Although most of the involved objects can be defined in a more general framework, we focus on a particular choice of vector bundle. Namely, let M be a smooth manifold and $E = TM \oplus T^*M$. A fiberwise metric $(\cdot,\cdot)$ on E is a $C^\infty(M)$-bilinear map $(\cdot,\cdot) : \Gamma(E) \times \Gamma(E) \to C^\infty(M)$, such that for each $p \in M$, $(\cdot,\cdot)_p : E_p \times E_p \to \mathbb{R}$ is a symmetric non-degenerate bilinear form. There exists a canonical fiberwise metric $\langle \cdot,\cdot \rangle$ on E, defined as

$$\langle V + \xi, W + \eta \rangle = i_V(\eta) + i_W(\xi), \tag{1}$$

for every $(V+\xi),(W+\eta) \in \Gamma(E)$. This fiberwise metric has signature (n,n), where n is a dimension of M. Hence, we denote by $O(n,n)$ the set of vector bundle automorphisms preserving this fiberwise metric. That is

$$O(n,n) = \{O \in \Gamma(Aut(E)) \mid (\forall e_1, e_2 \in \Gamma(E))\ (\langle Oe_1, Oe_2 \rangle = \langle e_1, e_2 \rangle)\}. \tag{2}$$

There are three important examples of $O(n,n)$ transformations, which we will use in the sequel. Let $B \in \Omega^2(M)$ be a 2-form on M. In this paper we will always denote the induced vector bundle morphism from TM to T^*M by the same letter, i.e., we define

$$B(V) = -i_V B = B(\cdot, V), \tag{3}$$

for all $V \in \mathfrak{X}(M)$. Correspondingly, the map e^B is given as

$$e^B(V + \xi) = V + \xi + B(V). \tag{4}$$

In the block matrix form

$$e^B \begin{pmatrix} V \\ \xi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ B & 1 \end{pmatrix} \begin{pmatrix} V \\ \xi \end{pmatrix}, \tag{5}$$

for all $(V+\xi) \in \Gamma(E)$. Similarly, let $\theta \in \Lambda^2\mathfrak{X}(M)$ be a bivector. The induced vector bundle morphism is again denoted by the same letter, that is

$$\theta(\xi) := -i_\xi \theta = \theta(\cdot, \xi), \tag{6}$$

for all $\xi \in \Omega^1(M)$. Correspondingly, we have e^θ

$$e^\theta(V + \xi) = V + \xi + \theta(\xi). \tag{7}$$

In the block matrix form

$$e^\theta \begin{pmatrix} V \\ \xi \end{pmatrix} = \begin{pmatrix} 1 & \theta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} V \\ \xi \end{pmatrix}, \tag{8}$$

for all $(V+\xi) \in \Gamma(E)$. Finally, let $N : TM \to TM$ be any invertible smooth vector bundle morphism over identity. We define the map O_N as

$$O_N(V + \xi) := N(V) + N^{-T}(\xi), \tag{9}$$

where $N^{-T} : T^*M \to T^*M$ denotes the map transpose to N^{-1}. In the block matrix form

$$O_N \begin{pmatrix} V \\ \xi \end{pmatrix} = \begin{pmatrix} N & 0 \\ 0 & N^{-T} \end{pmatrix} \begin{pmatrix} V \\ \xi \end{pmatrix}. \tag{10}$$

Any $O(n,n)$ transformation with the invertible upper-left block can be uniquely decomposed as a product of the form

$$e^{-B} O_N e^{-\theta}. \tag{11}$$

More explicitly, for $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ in $O(n,n)$, i.e., $A_{21}^T A_{11} + A_{11}^T A_{21} = 0$, $A_{12}^T A_{22} + A_{22}^T A_{12} = 0$ and $A_{21}^T A_{12} + A_{11}^T A_{22} = 1$, we find $N = A_{11}$, $\theta = -A_{11}^{-1} A_{12}$ and $B = -A_{21} A_{11}^{-1}$.

Let now $\tau : \Gamma(E) \to \Gamma(E)$ be a $C^\infty(M)$-linear map of sections, such that $\tau^2 = 1$. For $e_1, e_2 \in \Gamma(E)$, we put

$$(e_1, e_2)_\tau := \langle \tau(e_1), e_2 \rangle. \tag{12}$$

If such $(.,.)_\tau$ defines a positive definite fiberwise metric, we refer to it as a generalized metric on E. From now on, we will always assume that this is the case. Since $(\cdot,\cdot)_\tau$ is symmetric, τ is a symmetric map, that is,

$$\langle \tau(e_1), e_2 \rangle = \langle e_1, \tau(e_2) \rangle, \tag{13}$$

for all $e_1, e_2 \in \Gamma(E)$. Also, because $\tau^2 = 1$, it is orthogonal and thus $\tau \in O(n,n)$. Moreover, from $\tau^2 = 1$, we get two eigenbundles V_+ and V_-, corresponding to $+1$ and -1 eigenvalues of τ, respectively. Using the fact that $(\cdot,\cdot)_\tau$ is positive definite, we get that $\langle \cdot,\cdot \rangle$ is positive definite on $\Gamma(V_+)$ and negative definite on $\Gamma(V_-)$. Finally, we can observe that $V_+^\perp = V_-$ with respect to $\langle \cdot,\cdot \rangle$ and vice versa, and using the knowledge of the signature of $\langle \cdot,\cdot \rangle$, we get the direct sum decomposition

$$E = V_+ \oplus V_-. \tag{14}$$

Conversely, for any subbundle V of E of rank n, on which $\langle \cdot,\cdot \rangle$ is positive definite, set $\tau|_V := +1$ and $\tau|_{V^\perp} = -1$ to get a generalized metric on E.

From positive definiteness on V_+, we have $V_+ \cap TM = 0$ and $V_+ \cap T^*M = 0$, and the same for V_-. This means that V_+ and V_- can be viewed as graphs of invertible smooth vector bundle morphisms:

$$V_+ = \{V + A(V) \mid V \in TM\} \equiv \{A^{-1}(\xi) + \xi \mid \xi \in T^*M\}, \tag{15}$$

$$V_- = \{V + A'(V) \mid V \in TM\} \equiv \{A'^{-1}(\xi) + \xi \mid \xi \in T^*M\}, \tag{16}$$

where $A, A' : TM \to T^*M$, respectively. We can view A as covariant 2-tensor field on M, and write uniquely $A = g + B$, where g is a symmetric part of A and B a skew-symmetric part of A. From the positive definiteness of V_+ we get that g is a Riemannian metric on M, whereas B can be an arbitrary 2-form on M. Using the orthogonality of V_+ and V_-, we see that $A' = -g + B$. From this equivalent formulation, i.e. using g and B, we can uniquely reconstruct τ. This will give

$$\tau(V + \xi) = (g - Bg^{-1}B)(V) - g^{-1}B(V) + Bg^{-1}(\xi) + g^{-1}(\xi), \tag{17}$$

for all $(V+\xi) \in \Gamma(E)$. In the block matrix form,

$$\tau \begin{pmatrix} V \\ \xi \end{pmatrix} = \begin{pmatrix} -g^{-1}B & g^{-1} \\ g - Bg^{-1}B & Bg^{-1} \end{pmatrix} \begin{pmatrix} V \\ \xi \end{pmatrix}. \tag{18}$$

The corresponding fiberwise metric $(\cdot,\cdot)_\tau$ can then be written in the block matrix form

$$(V+\xi, W+\eta)_\tau = \begin{pmatrix} V \\ \xi \end{pmatrix}^T \begin{pmatrix} g - Bg^{-1}B & Bg^{-1} \\ -g^{-1}B & g^{-1} \end{pmatrix} \begin{pmatrix} W \\ \eta \end{pmatrix}. \tag{19}$$

The important observation is that the block matrix in formula (19) can be written as a product of simpler matrices. Namely,

$$\begin{pmatrix} g - Bg^{-1}B & Bg^{-1} \\ -g^{-1}B & g^{-1} \end{pmatrix} = \begin{pmatrix} 1 & B \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g & 0 \\ 0 & g^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -B & 1 \end{pmatrix}. \tag{20}$$

Note the important fact that the 2-form B *does not* have to be closed, and this will remain true throughout the whole paper. Nevertheless, we assume that B is globally defined, i.e. $H = dB$ globally.[1] We thus consider only the models with trivial H-flux. The case of the non-trivial H-flux will be discussed elsewhere.

There exists a natural action of the group $O(n,n)$ on the space of generalized metrics. For each $O \in O(n,n)$ and given τ define $\tau' = O^{-1}\tau O$. Clearly $\tau'^2 = 1$ and

$$\langle \tau'(e_1), e_2 \rangle = \langle \tau(O(e_1)), O(e_2) \rangle = (O(e_1), O(e_2))_\tau.$$

Hence $(\cdot,\cdot)_{\tau'}$ is again a generalized metric. We may use the notation $(\cdot,\cdot)_{\tau'} = O(\cdot,\cdot)_\tau$.

2.2 Factorizations of generalized metric, open-closed relations

Let us start with a (different) generalized metric $\mathbf{H}$, described by a Riemannian metric G and a 2-form Φ. Hence

$$\mathbf{H} = \begin{pmatrix} 1 & \Phi \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G & 0 \\ 0 & G^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\Phi & 0 \end{pmatrix}. \tag{21}$$

Let θ be a 2-vector field on M. The action of the $O(n,n)$ map $e^{-\theta}$ on the generalized metric $\mathbf{H}$ gives us a new generalized metric $\mathbf{G}$, which has the form

$$\mathbf{G} = \begin{pmatrix} 1 & 0 \\ \theta & 1 \end{pmatrix} \begin{pmatrix} 1 & \Phi \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G & 0 \\ 0 & G^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\Phi & 1 \end{pmatrix} \begin{pmatrix} 1 & -\theta \\ 0 & 1 \end{pmatrix}. \tag{22}$$

By the previous discussion, there exists a unique Riemannian metric g and a 2-form B, such that

$$\mathbf{G} = \begin{pmatrix} 1 & B \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g & 0 \\ 0 & g^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -B & 1 \end{pmatrix}. \tag{23}$$

Comparing the two expressions (22) and (23) of $\mathbf{G}$, we get the matrix equations

$$g - Bg^{-1}B = G - \Phi G^{-1}\Phi, \tag{24}$$

[1] More precisely, we assume that the corresponding integral cohomology class $[H]$ is trivial.

$$Bg^{-1} = \Phi G^{-1} - (G - \Phi G^{-1}\Phi)\theta, \tag{25}$$

which can be uniquely solved for G and Φ. Since $e^{-\theta}$ is invertible, we can proceed the other way around as well. We also know how the corresponding endomorphism $\tau_{\mathbf{H}}$ is changed by $e^{-\theta}$. Namely, we have

$$\tau_{\mathbf{G}} = e^{\theta}\tau_{\mathbf{H}}e^{-\theta}. \tag{26}$$

From that, we can easily find the relation between $+1$ eigenbundles:

$$V_{+}^{\mathbf{G}} = e^{\theta}V_{+}^{\mathbf{H}}. \tag{27}$$

Since

$$V_{+}^{\mathbf{G}} = \{\xi + (g+B)^{-1}(\xi) \mid \xi \in T^*M\},$$

and

$$V_{+}^{\mathbf{H}} = \{\xi + (G+\Phi)^{-1}(\xi) \mid \xi \in T^*M\},$$

we get using the above formula that

$$(g+B)^{-1} = \theta + (G+\Phi)^{-1}. \tag{28}$$

Formulae (24) and (25) are the symmetric and antisymmetric parts of (28). If θ is Poisson, (28) is the Seiberg-Witten formula[2] relating closed and open string backgrounds in the presence of a noncommutative structure represented by θ. In particular, for given g, B and θ, we can find a unique G and Φ, and conversely, for given G, Φ and θ, there exists a unique pair g and B.

For $\Phi = 0$ the open-closed relations can be given a slightly more geometric interpretation [2]. Consider the inverse $\mathbf{G}^{-1}$ of the generalized metric $\mathbf{G}$. If we exchange the tangent and cotangent bundles TM and T^*M, respectively, $\mathbf{G}^{-1}$ has the same properties as $\mathbf{G}$. Obviously, $\mathbf{G}^{-1}$ and $\mathbf{G}$ have identical graphs as well as ± 1-eigenbundles. The open-closed relations, for $\Phi = 0$, is a simple consequence of that.

2.3 Dorfman bracket, Dirac structures, D-branes

Here we briefly recall some relevant facts concerning the Dorfman bracket and Dirac structures, see, e.g., [11], [13], [5]. Our vector bundle $E = TM \oplus T^*M$ can be equipped with a structure of a Courant algebroid. The corresponding Courant bracket is the antisymmetrization of the Dorfman bracket:

$$[V+\xi, W+\eta]_D = [V,W] + \mathcal{L}_V(\eta) - i_W(d\xi), \tag{29}$$

for all $(V+\xi) \in \Gamma(E)$. The corresponding pairing is the canonical fiberwise metric (1).

A Dirac structure is a (smooth) subbundle L of E, which is maximally isotropic with respect to $\langle\cdot,\cdot\rangle$ and involutive under the Dorfman bracket (29).

Let θ be a rank-2 contravariant tensor field on M. As before, define a vector bundle morphism $\theta : T^*M \to TM$ by $\theta(\xi) = \theta(\cdot,\xi)$. Define a subbundle G_θ of E as its graph, that is

$$G_\theta = \{\xi + \theta(\xi) \mid \xi \in T^*M\}. \tag{30}$$

[2]For an earlier appearance of this type of formulae in the context of duality rotations see [12].

It is known that G_θ is a Dirac structure with respect to the Dorfman bracket, if and only if θ is a Poisson bivector. Similarly, let B be any rank-2 covariant tensor field on M. Define $B(V) = B(V, \cdot)$ and its graph G_B as

$$G_B = \{V + B(V) \mid V \in TM\}. \tag{31}$$

Again, one can show that G_B is a Dirac structure, if and only if B is a closed 2-form on M.

Furthermore, for any closed $B \in \Omega^2(M)$, one has

$$e^B[V + \xi, W + \eta]_D = [e^B(V + \xi), e^B(W + \eta)]_D, \tag{32}$$

and

$$\langle e^B(V + \xi), e^B(W + \eta)\rangle = \langle V + \xi, W + \eta\rangle, \tag{33}$$

for all $(V + \xi), (W + \eta) \in \Gamma(E)$. In the other words, e^B is an automorphism of the corresponding Courant algebroid. Note that (32) is no longer true for e^θ, where $\theta \in \Lambda^2\mathfrak{X}(M)$, but (33) holds.

Generally, a Dirac structure L provides a singular foliation of M by presympletic leaves, which is generated by its image $\rho(L)$ of the Dirac structure under the anchor map. We refer to [2] for arguments in favor of the identification "D-branes $\sim$ leaves of foliations defined by Dirac structures". In the case we will consider later, L will be given as a graph of a Poisson tensor θ and the corresponding foliation of M will be the foliation generated by Hamiltonian vector fields, i.e., by symplectic leaves of θ. Hence, in this case we will identify the symplectic leaves and D-branes.

3 Gauge field as an orthogonal transformation of the generalized metric

Let us start with a given Riemannian metric g and 2-form B. Further, let F be a 2-form (at this point an arbitrary one[3]). The gauge transformation defines new 2-form $B' = B + F$. To the pair (g, B) corresponds the generalized metric $\mathbf{G}$, see (23). The generalized metric $\mathbf{G}'$ corresponding to the pair $(g, B+F)$ has the following block matrix form:

$$\mathbf{G}' = \begin{pmatrix} 1 & F \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & B \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g & 0 \\ 0 & g^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -B & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -F & 1 \end{pmatrix}, \tag{34}$$

that is, $\mathbf{G}'$ is related to $\mathbf{G}$ by the $O(n, n)$ transform e^{-F}. As shown before, we can always get $\mathbf{G}$ by action of $O(n, n)$ transformation $e^{-\theta}$ on the generalized metric $\mathbf{H}$, where $\mathbf{H}$ is described by fields G and Φ, see (21).

One may ask, if there is a bivector θ' on M, such that we get $\mathbf{G}'$ by the action of $e^{-\theta'}$ on the generalized metric $\mathbf{H}'$, which is described by the same G as $\mathbf{H}$, but by gauged 2-form $\Phi' = \Phi + F'$ for some gauge field F'. This can be achieved under some assumptions,

[3]Later, when discussing DBI action, F will be closed and defined only on a submanifold of M supporting a D-brane. In which case, all expression involving F will make sense only when considered on the D-brane.

however, only up to a certain additional $O(n,n)$ action. In particular, there exists a vector bundle morphism $N : TM \to TM$, such that

$$\mathbf{G}' = \begin{pmatrix} 1 & 0 \\ \theta' & 1 \end{pmatrix} \begin{pmatrix} N^T & 0 \\ 0 & N^{-1} \end{pmatrix} \mathbf{H}' \begin{pmatrix} N & 0 \\ 0 & N^{-T} \end{pmatrix} \begin{pmatrix} 1 & -\theta' \\ 0 & 1 \end{pmatrix}, \tag{35}$$

where

$$\mathbf{H}' = \begin{pmatrix} 1 & \Phi' \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G & 0 \\ 0 & G^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\Phi' & 1 \end{pmatrix}.$$

Indeed, examine the block matrix decomposition:

$$\mathbf{G}' = \begin{pmatrix} 1 & F \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \theta & 1 \end{pmatrix} \begin{pmatrix} 1 & \Phi \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G & 0 \\ 0 & G^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\Phi & 1 \end{pmatrix} \begin{pmatrix} 1 & -\theta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -F & 1 \end{pmatrix}.$$

It suffices to consider the three rightmost matrices in the above expression. Since we want to modify Φ to $\Phi + F'$, we may proceed by inserting $1 = e^{-F'}e^{F'}$:

$$\begin{pmatrix} 1 & 0 \\ -\Phi & 1 \end{pmatrix} \begin{pmatrix} 1 & -\theta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -F & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -(\Phi + F') & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ F' & 1 \end{pmatrix} \begin{pmatrix} 1 & -\theta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -F & 1 \end{pmatrix}.$$

Now it is enough to note that the product of the last three matrices, can be uniquely decomposed into a product of a diagonal and an upper triangular block matrix—of course, only if we assume that $(1+\theta F)$ is invertible. For this, use the decomposition of $e^{-\theta}e^{-F} \in O(n,n)$ according to (11) as

$$e^{-\theta}e^{-F} = e^{-F'}O_N e^{-\theta'}, \tag{36}$$

with $F' \in \Omega^2(M), \theta' \in \Lambda^2\mathfrak{X}(M)$ and $N \in \Gamma(\mathrm{Aut}(TM))$. What we find are the following expression for θ', F' and N:

$$\theta' = (1 + \theta F)^{-1}\theta = \theta(1 + F\theta)^{-1}, \tag{37}$$

$$F' = F(1 + \theta F)^{-1} = (1 + F\theta)^{-1}F, \tag{38}$$

$$N = 1 + \theta F. \tag{39}$$

Comparing (34) and (35), we get the equalities

$$g - (B + F)g^{-1}(B + F) = N^T(G - (\Phi + F')G^{-1}(\Phi + F'))N \tag{40}$$

and

$$(B + F)g^{-1} = N^T(\Phi + F')G^{-1}N^{-T} - N^T(G - (\Phi + F')G^{-1}(\Phi + F'))N\theta'. \tag{41}$$

Taking the determinant of (40), we find that

$$\det(g - (B + F)g^{-1}(B + F)) = \det(N)^2 \cdot \det(G - (\Phi + F')G^{-1}(\Phi + F')). \tag{42}$$

This equality will play the central role when later discussing the DBI action.

Furthermore, following the same type of arguments leading to (28) we see that the equations (40) and (41) can equivalently be written as

$$(g + B + F)^{-1} = \theta' + (N^T(G + \Phi + F')N)^{-1}. \tag{43}$$

Finally, let us examine the objects F' and θ' using the tools described in subsection 2.3. We will concentrate on the case important for the discussion of the DBI action and noncommutative gauge theory. Therefore, in the rest of this section, we assume that θ is Poisson and F is closed. θ' is a bivector on M. For the graphs of θ and θ' we have

$$e^F G_\theta = G_{\theta'}. \tag{44}$$

Since e^F is an automorphism of Dorfman bracket, $G_{\theta'}$ has to be again a Dirac structure of E. Hence, θ' is a Poisson bivector. Similarly, one can see that

$$e^\theta G_F = G_{F'}. \tag{45}$$

4 Seiberg-Witten map

For an approach to the non-abelian case, using cohomological methods akin to the ones of Zumino's famous decent equations [23], see [6, 10]. Here we follow the approach of [15], [16], [17], where it was shown that the Seiberg-Witten field redefinition from the commutative to the non-commutative setting has its origin in a change of coordinates given by a map $\rho : M \to M$, such that $\rho^*(\theta') = \theta$.[4] This map can be derived using a generalization of Moser's lemma: Consider the family of Poisson bivectors

$$\theta_t = \theta(1 + tF\theta)^{-1} \tag{46}$$

parameterized by $t \in [0, 1]$. Of course, we have to presume that the formula is well-defined. To see that these θ_t are indeed Poisson for all t, simply observe that $G_{\theta_t} = e^{tF} G_\theta$ holds for the respective graphs.[5] Partial differentiation of (46) with respect to t leads to the differential equation

$$\partial_t \theta_t = -\theta_t F \theta_t.$$

For $F = dA$, this can be rewritten as

$$\partial_t \theta_t = -\mathcal{L}_{\theta_t(A)} \theta_t,$$

with a vector field $\theta_t(A) := \theta_t(\cdot, A)$, with initial condition $\theta_0 = \theta$. This differential equation can be integrated to a flow ϕ_t, such that $\phi_t^*(\theta_t) = \theta$. Thus $\rho = \phi_1$. Obviously, ρ explicitly depends on the choice of gauge potential A, hence we shall use the notation ρ_A. To avoid possible confusion, we will for a moment notationally distinguish between the tensor itself and its components in coordinates. Therefore we introduce the matrix $(\boldsymbol{\theta})^{ij} := \theta^{ij}$. Also, denote $J^i{}_k = \frac{\partial \rho^i}{\partial x^k}$. We have

$$\rho_A^*(\theta'^{kl}) = J^k{}_i J^l{}_j \theta^{ij}.$$

We thus get that

$$\det \rho_A^*(\boldsymbol{\theta'}) = J^2 \det \boldsymbol{\theta}. \tag{47}$$

[4] As said before, here we assume only topologically trivial $[H]$-flux. The interested reader may find some relevant discussion concerning nontrivial H and the related non-commutative gerbe in [3].

[5] Let us note again that $e^{tF} G_\theta$ is a bona-fide Dirac structure even for non-invertible $(1 + tF\theta)$.

Let us assume for a moment that $\boldsymbol{\theta}$ is invertible. From (37) we see that so is $\rho_A^*\boldsymbol{\theta'}$. We immediately have that

$$J^{-2} = \det\left(\boldsymbol{\theta}(\rho_A^*\boldsymbol{\theta'})^{-1}\right). \tag{48}$$

For degenerate θ and hence also θ' the formula (48) still makes sense and we can argue as follows: Since the map ρ_A is infinitesimally generated by the vector field $\theta_t(A)$, and the kernels of all θ_t's are the same, we see that ρ_A only changes coordinates on the symplectic leaves (of θ). We can thus restrict ourselves to the non-degenerate case in order to carry out the computation of the Jacobian.

5 Noncommutative gauge theory and DBI action

In the previous sections we have described all ingredients needed for our discussion of noncommutativity of D-branes as a consequence of their generalized geometry. Namely, we have seen that the relations (24), (25), (40) and the (semiclassical) Seiberg-Witten have their root in generalized geometry. Actually, it is know for quite some time [17] that the equivalence of the commutative and (semiclassically) noncommutative DBI actions follows once one has established (24), (25), (40) and has understood the (semiclassical) Seiberg-Witten map as a (local) D-brane diffeomorphism. Nevertheless, according to our best knowledge, the direct relation to generalized geometry is new.

Assume that we have a D-brane D of dimension d, i.e, a submanifold of target space-time M equipped with a line bundle with a connection A and corresponding field strength F. Also, consider the restrictions (pullbacks) of the background fields (open and closed ones) to D. While describing the Seiberg-Witten map in the previous section, we have seen that it is quite natural to assume that there is a relation between the D-brane and the Poisson tensor θ.[6] Namely, assume that our D-brane is of a particular kind, i.e., one which comes as symplectic leaf of the Poisson structure θ.[7] As argued before, under this assumption, the Seiberg-Witten map is a D-brane diffeomorphism.

Before we turn to the discussion of the DBI action and its commutative and noncommutative description, we discuss the relation between the effective closed and open string coupling constants g_s and G_s, respectively [21]. These are related as

$$G_s = g_s\Big(\frac{\det(G+\Phi)}{\det(g+B)}\Big)^{1/2}.$$

A most intriguing relation is obtained from (**??**) and the relation (40), again using the above mentioned formula for the determinant of a sum of a symmetric and an antisymmetric matrix:

$$\frac{1}{g_s}\det{}^{1/2}(g+B+F) = \frac{1}{G_s}\det{}^{1/2}(1+\theta F)\det{}^{1/2}(G+\Phi+F'). \tag{49}$$

[6] Recall, in accordance with our above discussion of the open-closed relations, here we start from a given closed background (g, B), pick a θ and determine uniquely the open variables (G, Φ).

[7] It is straight-forward to modify everything to the case where the D-brane is a submanifold, such that the restriction of θ to it defines a regular Poisson structure, i.e. a Poisson structure having constant rank.

Integrating over the D-brane world-volume

$$\int d^dx \frac{1}{g_s} \det{}^{1/2}(g+B+F) = \int d^dx \frac{1}{G_s} \det{}^{1/2}(1+\theta F) \det{}^{1/2}(G+\Phi+F'), \tag{50}$$

recalling (48), and performing the change of coordinates according to the Seiberg-Witten map, we finally obtain a relation between the commutative and semiclassically noncommutative DBI actions

$$S^c_{DBI} := \int d^dx \frac{1}{g_s} \det{}^{1/2}(g+B+F) = \int d^dx \frac{1}{\hat{G}_s} \det{}^{1/2}\Big(\frac{\hat{\theta}}{\theta}\Big) \det{}^{1/2}(\hat{G}+\hat{\Phi}+\hat{F}') =: S^{nc}_{DBI}. \tag{51}$$

The hat $\hat{}$ has the following meaning: On matrix elements of θ it is defined as $\hat{\theta}^{ij} := \rho^*_A(\theta^{ij})$, and similarly for the other objects. As a result of this definition, $\hat{F}'$ is the semiclassically noncommutative field strength, which under the gauge transformation $\delta A = d\lambda$ transforms semiclassically noncommutatively, i.e.,

$$\delta \hat{F}'_{ij} = \{\hat{F}'_{ij}, \tilde{\lambda}\},$$

$$\tilde{\lambda} = \sum \frac{(\theta_t(A) + \partial_t)^n(\lambda)}{(n+1)!}|_{t=0}.$$

Here, the curly bracket is the Poisson bracket corresponding to the Poisson tensor θ and $\tilde{\lambda}$ is the (semiclassical) noncommutative gauge parameter.

References

[1] A. Alekseev and T. Strobl. *Current algebras and differential geometry.* JHEP **0503** (2005), 035.

[2] T. Asakawa, S. Sasa, and S. Watamura. *D-branes in Generalized Geometry and Dirac-Born-Infeld Action.* JHEP **1210** (2012), 064.

[3] P. Aschieri, I. Baković, B. Jurčo, and P. Schupp. *Noncommutative gerbes and deformation quantization.* J. Geom. Phys. **60** (2010), 1754–1761.

[4] M. Bojowald, A. Kotov, and T. Strobl. *Lie algebroid morphisms, Poisson sigma models, and off-shell closed gauge symmetries.* J.Geom.Phys. **54** (2005), 400–426.

[5] P. Bouwknegt. *Lectures on cohomology, T-duality, and generalized geometry.* Lect.Notes Phys. **807** (2010), 261–311.

[6] D. Brace, B. L. Cerchiai, A. F. Pasqua, U. Varadarajan, and B. Zumino. *A Cohomological Approach to the Non-Abelian Seiberg-Witten Map.* JHEP **0106** (2001), 047.

[7] A. S. Cattaneo. *On the Integration of Poisson Manifolds, Lie Algebroids, and Coisotropic Submanifolds.* Letters in Mathematical Physics **67** (January 2004), 33–48.

[8] A. S. Cattaneo and G. Felder. *Coisotropic submanifolds in Poisson geometry and branes in the Poisson sigma model.* Letters in Mathematical Physics **69** (July 2004), 157–175.

[9] A. S. Cattaneo and G. Felder. *A Path integral approach to the Kontsevich quantization formula.* Commun.Math.Phys. **212** (2000), 591–611.

[10] B. L. Cerchiai, A. F. Pasqua, and B. Zumino. *The Seiberg-Witten Map for Noncommutative Gauge Theories.* . Talk presented at Continuous Advances in QCD 2002/Arkadyfest.

[11] T. Courant. *Dirac manifolds.* Trans. Amer. Math. Soc. **319** (1990), 631–661.

[12] M. J. Duff and J. X. Lu. *Duality Rotations in Membrane Theory.* Nuclear Physics B **347** (1990), 394–419.

[13] M. Gualtieri. *Generalized complex geometry.* (2003).

[14] N. Hitchin. *Generalized Calabi-Yau manifolds.* Quart.J.Math.Oxford Ser. **54** (2003), 281–308.

[15] B. Jurčo and P. Schupp. *Noncommutative Yang-Mills from equivalence of star products.* Eur.Phys.J. **C14** (2000), 367–370.

[16] B. Jurčo, P. Schupp, and J. Wess. *Noncommutative gauge theory for Poisson manifolds.* Nucl.Phys. **B584** (2000), 784–794.

[17] B. Jurčo, P. Schupp, and J. Wess. *NonAbelian noncommutative gauge theory via noncommutative extra dimensions.* Nucl.Phys. **B604** (2001), 148–180.

[18] C. Klimčík and T. Strobl. *WZW - Poisson manifolds.* J.Geom.Phys. **43** (2002), 341–344.

[19] A. Kotov, P. Schaller, and T. Strobl. *Dirac sigma models.* Commun.Math.Phys. **260** (2005), 455–480.

[20] A. Kotov and T. Strobl. *Generalizing Geometry - Algebroids and Sigma Models.* In 'Handbook of pseudo-Riemannian geometry and supersymmetry (ed. by V. Cortes)', European Mathematical Society (2010).

[21] N. Seiberg and E. Witten. *String theory and noncommutative geometry.* JHEP **9909** (1999), 032.

[22] Ševera. P. *Letters to Alan Weinstein, 2.* http://sophia.dtp.fmph.uniba.sk/ severa/letters/ .

[23] B. Zumino. *Cohomology of Gauge Groups: Cocycles and Schwinger Terms.* Nucl.Phys. **B253** (1985), 477.

Implementation of the Finite Element Method for the Heat Equation*

Vítězslav Žabka

4th year of PGS, email: vitezslav.zabka@fjfi.cvut.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 describes a general implementation of the finite element method for the heat equation. The implementation allows for using domains of arbitrary dimensions and various types of finite elements. In addition, it is easily extensible to other partial differential equations. The main goal of this article is to analyze the finite element method from the implementational point of view while not ignoring its mathematical background.

Keywords: FEM, heat equation

Abstrakt. Tento příspěvek popisuje obecnou implementaci metody konečných prvků pro rovnici vedení tepla. Popisovaná implementace umožňuje použití domén libovolné dimenze a různých typů konečných prvků. Navíc je jednoduše rozšiřitelná i na jiné parciální diferenciální rovnice. Hlavním cílem tohoto článku je analyzovat metodu konečných prvků z implementačního pohledu a přitom zohlednit její matematickou stránku.

Klíčová slova: metoda konečných prvků, rovnice vedení tepla

1 Introduction

A long-term goal of our work is to create a GPU solver for the incompressible Navier–Stokes equations in 3D using unstructured meshes. Because we have experience with the finite element method for the Navier–Stokes equations in 2D, we intended to extend our original 2D implementation into 3D. However, this task proved to be difficult. The original implementation would have to be completely rewritten in order to include 3D computations. Thus, we decided to create a new implementation of the finite element method from scratch. Our requirements on the new implementation include its extensibility to other computational problems and its suitability for the adaptation to GPUs. Another option would be to use an existing finite element library, e.g., DUNE-FEM [4], DUNE-PDELab [5] or ViennaFEM [7]. Nevertheless, although such libraries are rather universal, they do not support computations on the GPU.

This article deals with the finite element method for the heat equation from two perspectives. First, it briefly summarizes the mathematical background of the method.

*This work has been supported by the grant No. SGS11/161/OHK4/3T/14 of the Student Grant Agency of the Czech Technical University in Prague and the project No. TA01020871 of the Technological Agency of the Czech Republic

And second, it attempts to describe a way of implementing the method generally, i.e., independently of the type of the finite elements and of the domain dimension. Moreover, the implementation should also, with slight modifications, be applicable to some other partial differential equations than the heat equation. The heat equation was chosen as an example for its simplicity.

2 FEM for the heat equation

Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz-continuous domain and $\partial\Omega$ its boundary. The heat equation for an unknown function $u = u(x,t)$ of the spatial coordinates $x \in \Omega$ and the time variable $t \in (0,T)$, supplemented with a Dirichlet boundary condition and an initial condition, takes the following form:

$$\frac{\partial u}{\partial t} = \Delta u \qquad \text{in } \Omega \times (0,T), \tag{1a}$$

$$u|_{\partial\Omega} = u_{\mathrm{Dir}} \qquad \text{on } \partial\Omega \times (0,T), \tag{1b}$$

$$u|_{t=0} = u_{\mathrm{ini}} \qquad \text{in } \Omega, \tag{1c}$$

where $u_{\mathrm{Dir}} = u_{\mathrm{Dir}}(x,t)$ and $u_{\mathrm{ini}} = u_{\mathrm{ini}}(x)$ are given functions and Δ denotes the Laplace operator. It is a second-order parabolic partial differential equation.

The first step of the finite element method consists of converting problem (1) into its corresponding weak formulation. This is accomplished by multiplying equation (1a) by a test function $v \in V$, where V is a suitable function space, and integrating over Ω:

$$\int_\Omega \frac{\partial u}{\partial t}\, v \,\mathrm{d}x = \int_\Omega \Delta u \, v \,\mathrm{d}x. \tag{2}$$

Here the usual choice of V is the set of all functions in the Sobolev space $W^{1,2}(\Omega)$ with zero trace on $\partial\Omega$. Applying Green's theorem on the right-hand side of (2) and using the fact that the trace of v is zero on $\partial\Omega$, the weak formulation of (1) is obtained:

$$\int_\Omega \frac{\partial u}{\partial t}\, v \,\mathrm{d}x = -\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x \quad \text{for all } v \in V. \tag{3}$$

The solution $u = u(x,t)$ might be thought of as a function $u : (0,T) \to W^{1,2}(\Omega)$ mapping $t \in (0,T)$ to $u(\cdot,t) \in W^{1,2}(\Omega)$. Additional assumptions on u are that it is differentiable with respect to t, that $\frac{\partial u}{\partial t}(\cdot,t) \in L^2(\Omega)$ for all $t \in (0,T)$ and that $u(\cdot,t)|_{\partial\Omega} = u_{\mathrm{Dir}}$ in the sense of traces for all $t \in (0,T)$.

2.1 Spatial discretization

In order to discretize (3) spatially by means of the finite element method, the function $u(\cdot,t) \in W^{1,2}(\Omega)$ is for all $t \in (0,T)$ decomposed as

$$u(\cdot,t) = u_0(\cdot,t) + u_{\mathrm{D}}, \tag{4}$$

where $u_0(\cdot,t) \in V$ and $u_{\mathrm{D}} \in W^{1,2}(\Omega)$ such that $u_{\mathrm{D}}|_{\partial\Omega} = u_{\mathrm{Dir}}$ in the sense of traces. Furthermore, we assume such u_{D} exists and is known. Substituting (4) into (3) gives:

$$\int_\Omega \frac{\partial u_0}{\partial t}\, v \,\mathrm{d}x = -\int_\Omega \nabla u_0 \cdot \nabla v \,\mathrm{d}x - \int_\Omega \nabla u_{\mathrm{D}} \cdot \nabla v \,\mathrm{d}x \quad \text{for all } v \in V. \tag{5}$$

Now let V_h be a finite dimensional subspace of V. The semi-discrete weak formulation of (1) is to find $u_h : (0,T) \to V_h$ satisfying

$$\int_\Omega \frac{\partial u_h}{\partial t}\, v_h \,\mathrm{d}x = -\int_\Omega \nabla u_h \cdot \nabla v_h \,\mathrm{d}x - \int_\Omega \nabla u_{\mathrm{D}} \cdot \nabla v_h \,\mathrm{d}x \quad \text{for all } v_h \in V_h. \tag{6}$$

Denoting by $\Phi = \{\varphi_1, \ldots, \varphi_N\}$ a basis for V_h, the function $u_h(\cdot,t)$ can be expressed for each $t \in (0,T)$ as a linear combination of the basis functions:

$$u_h(\cdot,t) = \sum_{j=1}^{N} u_j(t)\, \varphi_j, \tag{7}$$

where the coefficients u_j, $j = 1,\ldots,N$, are real functions of time. Plugging (7) into (6) and taking $v_h = \varphi_i$, $i = 1,\ldots,N$, leads to the following system of N ordinary differential equations:

$$\sum_{j=1}^{N} u_j'(t) \int_\Omega \varphi_j\, \varphi_i \,\mathrm{d}x = -\sum_{j=1}^{N} u_j(t) \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x - \int_\Omega \nabla u_{\mathrm{D}} \cdot \nabla\varphi_i \,\mathrm{d}x \tag{8}$$

with the initial condition $u_j(0)$, $j = 1,\ldots,N$, given by a projection of u_{ini} onto V_h. Equations (8) can be rewritten in a more compact form using matrices $\mathbf{M}$ (mass) and $\mathbf{S}$ (stiffness), whose elements are

$$M_{i,j} = \int_\Omega \varphi_i\, \varphi_j \,\mathrm{d}x \qquad \text{and} \qquad S_{i,j} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x, \tag{9}$$

and vector $\mathbf{f}$ composed of components

$$f_i = -\int_\Omega \nabla u_{\mathrm{D}} \cdot \nabla\varphi_i \,\mathrm{d}x. \tag{10}$$

Then system (8) becomes:

$$\mathbf{M}\mathbf{u}'(t) = -\mathbf{S}\mathbf{u}(t) + \mathbf{f} \tag{11}$$

with the initial condition $\mathbf{u}(0)$, where $\mathbf{u}(t)$ is a vector comprising components $u_j(t)$ and $\mathbf{u}'(t)$ its time derivative.

In the finite element method, the space V_h is chosen so that it contains continuous, piecewise polynomial functions on a triangulation of Ω. The support of the basis functions $\varphi_1, \ldots, \varphi_N$ usually consists of only several cells of the triangulation. Hence, the matrices $\mathbf{M}$ and $\mathbf{S}$ are sparse. Accordingly, a finite element is determined by a cell of the triangulation and by the restriction of the basis functions to the cell.

2.2 Time discretization

We employ the backward Euler method for the time discretization of (11). We introduce a time step $\tau > 0$ and the notation $\mathbf{u}^k$ for $\mathbf{u}(k\tau)$. The time derivative $\mathbf{u}'(t)$ is approximated by a difference quotient:

$$\mathbf{u}'(t) \approx \frac{\mathbf{u}^k - \mathbf{u}^{k-1}}{\tau}. \tag{12}$$

The time discretization of (11) is performed in an implicit manner, which results in the following system of linear equations for $\mathbf{u}^k$ at the time level $k = 1, 2, \ldots$:

$$(\mathbf{M} + \tau\mathbf{S})\,\mathbf{u}^k = \mathbf{M}\mathbf{u}^{k-1} + \tau\mathbf{f}. \tag{13}$$

Vector $\mathbf{u}^0$ represents the initial condition.

3 Implementation

The algorithm of the finite element method for the heat equation is divided into the following four basic steps:

1. triangulation of computational domain Ω,
2. evaluation of integrals in (9) and (10),
3. assembly of matrices $\mathbf{M}$ and $\mathbf{S}$ and of vector $\mathbf{f}$,
4. solution of linear system (13).

3.1 Triangulation of the computational domain

In the presented implementation, the computational domain Ω is triangulated prior to the start of the main program using an external application, e.g., Gmsh [6] or NETGEN [8]. The output of these applications is a conforming unstructured mesh representing the triangulation; i.e., the triangulation is the set of the mesh cells. The term conforming mesh means that neighboring mesh cells are required to meet face-to-face, edge-to-edge and vertex-to-vertex. In other words, if two mesh cells intersect, their intersection is always an entire face, edge or vertex of both of the cells.

We assume that Ω has piecewise linear boundary and so can be triangulated exactly. Thus, denoting the triangulation by $\mathcal{T}_h$,

$$\Omega = \bigcup_{K \in \mathcal{T}_h} K. \tag{14}$$

3.2 Evaluation of integrals

Given a mesh for the computational domain Ω and denoting by $\mathcal{T}_h$ the set of its cells, the integrals over Ω in (9) and (10) can be expressed by the sum of integrals over the mesh cells:

$$\int_\Omega \cdot \,\mathrm{d}x = \sum_{K \in \mathcal{T}_h} \int_K \cdot \,\mathrm{d}x. \tag{15}$$

A common way of computing such integrals over $K \in \mathcal{T}_h$ is to transform them to integrals over the corresponding reference element and evaluate them using quadratures.

Each cell $K \in \mathcal{T}_h$ together with the set $\Psi_K = \{\psi_1, \ldots, \psi_M\}$ of so-called local basis functions associated with the cell form a finite element. The local basis functions ψ_i are nonzero functions from K to $\mathbb{R}$ satisfying that for each $\psi_i \in \Psi_K$ there exists a basis function $\varphi_l \in \Phi$ such that $\varphi_l|_K \equiv \psi_i$. Moreover, if $\varphi_l|_K \not\equiv 0$ for some $\varphi_l \in \Phi$, then $\varphi_l|_K \in \Psi_K$. Hence, we never need the global basis functions $\varphi_l \in \Phi$ because the local basis functions $\psi_i \in \Psi_K$ hold all the necessary information. Note that the local basis functions depend on K although this dependance is not reflected in the notation.

3.2.1 Reference elements

A reference element is a specific finite element from which all finite elements of the corresponding type are derived by transformation. It is given by its geometric shape $\tilde{K}$ and the set of local basis functions $\tilde{\Psi} = \{\tilde{\psi}_1, \ldots, \tilde{\psi}_M\}$, where $\tilde{\psi}_i : \tilde{K} \to \mathbb{R}$ for all $\tilde{\psi}_i \in \tilde{\Psi}$.

The usual geometric shapes of reference elements are as follows:

- line segment $\tilde{K}_{\text{lin}} = \text{Conv}\,\{(0), (1)\} \subset \mathbb{R}$,
- triangle $\tilde{K}_{\text{tri}} = \text{Conv}\,\{(0,0), (1,0), (0,1)\} \subset \mathbb{R}^2$,
- quadrilateral $\tilde{K}_{\text{quad}} = \text{Conv}\,\{(0,0), (1,0), (1,1), (0,1)\} \subset \mathbb{R}^2$,
- tetrahedron $\tilde{K}_{\text{tetra}} = \text{Conv}\,\{(0,0,0), (1,0,0), (0,1,0), (0,0,1)\} \subset \mathbb{R}^3$,
- hexahedron $\tilde{K}_{\text{hex}} = \text{Conv}\,\{(0,0,0), (1,0,0), (1,1,0), (0,1,0), (0,0,1), (1,0,1), (1,1,1), (0,1,1)\} \subset \mathbb{R}^3$,

where $\text{Conv}\{\cdot\}$ is the convex hull of a set of points.

There are many ways to choose the reference basis functions; see, e.g., [3]. For example, the P1 reference element on the triangle $\tilde{K}_{\text{tri}}$ is given by the following reference basis functions:

$$\tilde{\psi}_1(\tilde{x}) = 1 - \tilde{x}_1 - \tilde{x}_2, \qquad \tilde{\psi}_2(\tilde{x}) = \tilde{x}_1, \qquad \tilde{\psi}_3(\tilde{x}) = \tilde{x}_2 \tag{16}$$

for all $\tilde{x} = (\tilde{x}_1, \tilde{x}_2) \in \tilde{K}_{\text{tri}} \subset \mathbb{R}^2$.

3.2.2 Transformation of the reference element

The transformation of the reference element to a finite element is based on a bijective mapping $g : \tilde{K} \to K$, i.e., from the geometric shape of the reference element to the geometric shape of the finite element. Obviously, the reference basis functions $\tilde{\psi}_i \in \tilde{\Psi}$ must be transformed as well. For the case of scalar basis functions, the relation between $\tilde{\psi}_i$ and the corresponding $\psi_i \in \Psi_K$ is:

$$\psi_i\big(g(\tilde{x})\big) = \tilde{\psi}_i(\tilde{x}) \qquad \text{for all } \tilde{x} \in \tilde{K}. \tag{17}$$

Considering the geometric shapes of the reference elements introduced in Section 3.2.1, the usual mappings g comprise affine, bilinear and trilinear transformations.

Affine transformation. Affine transformations are used to map $\tilde{K}_{\text{lin}}$ to an arbitrary line segment, $\tilde{K}_{\text{tri}}$ to an arbitrary triangle and $\tilde{K}_{\text{tetra}}$ to an arbitrary tetrahedron. The

general form of an affine transformation g is:

$$g(\tilde{x}) = L\tilde{x} + s, \tag{18}$$

where L is a matrix representing a linear transformation and s is a vector. The Jacobian matrix J_g of such transformation is equal to L:

$$J_g(\tilde{x}) = L. \tag{19}$$

When transforming $\tilde{K}_{\text{lin}}$ to an arbitrary line segment AB with endpoints A and B, the parameters L and s in (18) can be given as follows:

$$L = B - A, \qquad s = A. \tag{20}$$

Similarly, when transforming $\tilde{K}_{\text{tri}}$ to an arbitrary triangle ABC, the columns of L would be $B - A$ and $C - A$:

$$L = (B - A, C - A), \qquad s = A, \tag{21}$$

and in the same way for $\tilde{K}_{\text{tetra}}$ and an arbitrary tetrahedron $ABCD$:

$$L = (B - A, C - A, D - A), \qquad s = A. \tag{22}$$

Bilinear transformation. A bilinear transformation maps the reference quadrilateral shape $\tilde{K}_{\text{quad}}$ to an arbitrary convex quadrilateral:

$$g(\tilde{x}) = p_0 + \tilde{x}_1 p_1 + \tilde{x}_2 p_2 + \tilde{x}_1 \tilde{x}_2 p_{12}, \tag{23}$$

where $\tilde{x} = (\tilde{x}_1, \tilde{x}_2) \in \tilde{K}_{\text{quad}}$ and p_0, p_1, p_2 and p_{12} are column vectors. Its two-column Jacobian matrix is given by:

$$J_g(\tilde{x}) = (p_1 + \tilde{x}_2 p_{12}, p_2 + \tilde{x}_1 p_{12})\,. \tag{24}$$

To transform $\tilde{K}_{\text{quad}}$ to an arbitrary convex quadrilateral $ABCD$ (with vertices labeled in accordance with $\tilde{K}_{\text{quad}}$ as defined in Section 3.2.1), the parameters in (23) should be set as follows:

$$p_0 = A, \qquad p_1 = B - A, \qquad p_2 = D - A, \qquad p_{12} = A - B + C - D. \tag{25}$$

Trilinear transformation. Using a trilinear transformation, the reference hexahedral shape $\tilde{K}_{\text{hex}}$ is mapped to an arbitrary convex, quadrilaterally-faced hexahedron:

$$g(\tilde{x}) = p_0 + \tilde{x}_1 p_1 + \tilde{x}_2 p_2 + \tilde{x}_3 p_3 + \tilde{x}_1 \tilde{x}_2 p_{12} + \tilde{x}_2 \tilde{x}_3 p_{23} + \tilde{x}_3 \tilde{x}_1 p_{31} + \tilde{x}_1 \tilde{x}_2 \tilde{x}_3 p_{123}. \tag{26}$$

The corresponding Jacobian matrix has three columns:

$$\begin{aligned} J_g(\tilde{x}) = \big(&p_1 + \tilde{x}_2 p_{12} + \tilde{x}_3 p_{31} + \tilde{x}_2 \tilde{x}_3 p_{123}, \\ &p_2 + \tilde{x}_1 p_{12} + \tilde{x}_3 p_{23} + \tilde{x}_1 \tilde{x}_3 p_{123}, \\ &p_3 + \tilde{x}_2 p_{23} + \tilde{x}_1 p_{31} + \tilde{x}_1 \tilde{x}_2 p_{123}\big). \end{aligned} \tag{27}$$

An arbitrary convex, quadrilaterally-faced hexahedron $ABCDEFGH$ is obtained from $\tilde{K}_{\text{hex}}$ by g defined in (26) with the following parameters:

$$\begin{aligned} p_0 &= A, & p_1 &= B - A, & p_2 &= D - A, & p_3 &= E - A, \\ p_{12} &= C - D - B + A, & p_{23} &= H - E - D + A, & p_{31} &= F - B - E + A, \\ p_{123} &= G - F + E - H - C + D + B - A. \end{aligned} \tag{28}$$

Again, the vertices of the hexahedron should be labeled in accordance with $\tilde{K}_{\text{hex}}$ as defined in Section 3.2.1.

3.2.3 Transformation of integrals to the reference element

Let $g : \tilde{K} \to K$ be a bijective mapping from the reference element shape $\tilde{K}$ to $K \in \mathcal{T}_h$ as described in Section 3.2.2. Then, using integration by substitution and the fact that $g(\tilde{K}) = K$,

$$\int_K \psi_i(x)\, \psi_j(x)\, \mathrm{d}x = \int_{\tilde{K}} \psi_i\big(g(\tilde{x})\big)\, \psi_j\big(g(\tilde{x})\big)\, |\det J_g(\tilde{x})|\, \mathrm{d}\tilde{x}, \tag{29}$$

where $\psi_i, \psi_j \in \Psi_K$ and $J_g(\tilde{x})$ denotes the Jacobian matrix of g at point $\tilde{x} \in \tilde{K}$. Using (17) we can rewrite (29) as

$$\int_K \psi_i(x)\, \psi_j(x)\, \mathrm{d}x = \int_{\tilde{K}} \tilde{\psi}_i(\tilde{x})\, \tilde{\psi}_j(\tilde{x})\, |\det J_g(\tilde{x})|\, \mathrm{d}\tilde{x}. \tag{30}$$

In a similar manner, integrals involving gradients of the basis functions are transformed; for example:

$$\int_K \nabla\psi_i(x) \cdot \nabla\psi_j(x)\, \mathrm{d}x = \int_{\tilde{K}} \nabla\psi_i\big(g(\tilde{x})\big) \cdot \nabla\psi_j\big(g(\tilde{x})\big)\, |\det J_g(\tilde{x})|\, \mathrm{d}\tilde{x}. \tag{31}$$

Since $\nabla\psi_i(x)$ is equal to the transposed Jacobian matrix of $\psi_i(x)$ and the notation using Jacobian matrices is more general, we consider the following form of (31):

$$\int_K \nabla\psi_i(x) \cdot \nabla\psi_j(x)\, \mathrm{d}x = \int_{\tilde{K}} J_{\psi_i}\big(g(\tilde{x})\big) J_{\psi_j}\big(g(\tilde{x})\big)^T\, |\det J_g(\tilde{x})|\, \mathrm{d}\tilde{x}. \tag{32}$$

Differentiating the transformation formula of the basis functions (17) with respect to $\tilde{x}$ and writing the result in terms of the Jacobian matrices yields:

$$J_{\psi_i \circ g}(\tilde{x}) = J_{\tilde{\psi}_i}(\tilde{x}). \tag{33}$$

Application of the chain rule to the left-hand side of (33) leads to:

$$J_{\psi_i}\big(g(\tilde{x})\big) J_g(\tilde{x}) = J_{\tilde{\psi}_i}(\tilde{x}), \tag{34}$$

which is equivalent to

$$J_{\psi_i}\big(g(\tilde{x})\big) = J_{\tilde{\psi}_i}(\tilde{x}) J_g(\tilde{x})^{-1}. \tag{35}$$

It follows that

$$\int_K \nabla\psi_i(x)\cdot\nabla\psi_j(x)\,\mathrm{d}x = \int_{\tilde{K}} J_{\tilde{\psi}_i}(\tilde{x})J_g(\tilde{x})^{-1}\left(J_g(\tilde{x})^{-1}\right)^T J_{\tilde{\psi}_j}(\tilde{x})^T\,|\det J_g(\tilde{x})|\,\mathrm{d}\tilde{x}. \tag{36}$$

Owing to (10), we also need to compute integrals involving u_D which cannot be expressed as a linear combination of the basis functions:

$$\int_K \nabla u_\mathrm{D}(x)\cdot\nabla\psi_i(x)\,\mathrm{d}x = \int_{\tilde{K}} J_{u_\mathrm{D}}\big(g(\tilde{x})\big)\left(J_g(\tilde{x})^{-1}\right)^T J_{\tilde{\psi}_i}(\tilde{x})^T\,|\det J_g(\tilde{x})|\,\mathrm{d}\tilde{x}, \tag{37}$$

where $J_{u_\mathrm{D}}\big(g(\tilde{x})\big)$ is the Jacobian matrix of u_D with respect to x at point $g(\tilde{x})$.

Formulas (30), (36) and (37) cannot be used when transforming reference element shapes to mesh cells in a higher-dimensional space. This is the case of, e.g., surface meshes in 3D. Because the Jacobian matrix of g is not square, its determinant and inverse do not exist. However, $|\det J_g|$ can be replaced with a volume element [2]:

$$|\det J_g| \sim \sqrt{\det\left({J_g}^T J_g\right)} \tag{38}$$

and the inverse with a left inverse:

$${J_g}^{-1} \sim \left({J_g}^T J_g\right)^{-1} {J_g}^T. \tag{39}$$

3.2.4 Quadratures

The integrals on the right-hand side of (30), (36) and (37) are of the form

$$\int_{\tilde{K}} f(\tilde{x})\,\mathrm{d}\tilde{x}, \tag{40}$$

where f is some real function. Such integrals over a reference element shape $\tilde{K}$ are evaluated using quadratures; i.e., the integrals are approximated by a weighted sum of $f(x_i)$ at certain points $x_i \in \tilde{K}$:

$$\int_{\tilde{K}} f(\tilde{x})\,\mathrm{d}\tilde{x} \approx \sum_{i=1}^{m} w_i f(x_i). \tag{41}$$

The points $x_i \in \tilde{K}$ and weights $w_i \in \mathbb{R}$ are chosen so that the quadrature rule (41) yields exact results for polynomial functions f up to a certain degree. For examples of quadrature rules on the reference elements, see, e.g., [3].

3.3 Assembly of the finite element matrices and vector

The matrices $\mathbf{M}$ and $\mathbf{S}$ and the vector $\mathbf{f}$ are assembled from the contributions from each finite element. Given a finite element represented by a cell $K \in \mathcal{T}_h$ and a set $\Psi_K = \{\psi_1, \ldots, \psi_M\}$ of local basis functions, we can construct the so-called local matrices

$\mathbf{M}_K$ and $\mathbf{S}_K$ and the local vector $\mathbf{f}_K$. The matrices $\mathbf{M}_K$ and $\mathbf{S}_K$ are composed of elements $(M_K)_{i,j}$ and $(S_K)_{i,j}$ computed using (30), (36), (38) and (39) as:

$$(M_K)_{i,j} = \int_K \psi_i \, \psi_j = \int_{\tilde{K}} \tilde{\psi}_i \, \tilde{\psi}_j \sqrt{\det\left(J_g{}^T J_g\right)}, \tag{42}$$

$$(S_K)_{i,j} = \int_K \nabla\psi_i \cdot \nabla\psi_j = \int_{\tilde{K}} J_{\tilde{\psi}_i} \left(J_g{}^T J_g\right)^{-1} J_{\tilde{\psi}_j}{}^T \sqrt{\det\left(J_g{}^T J_g\right)} \tag{43}$$

for $i, j = 1, \ldots, M$. Similarly, the components $(f_K)_i$ of $\mathbf{f}_K$ are computed using (37), (38) and (39) as:

$$(f_K)_i = -\int_K \nabla u_\mathrm{D} \cdot \nabla\psi_i = -\int_{\tilde{K}} J_{u_\mathrm{D}} J_g \left(J_g{}^T J_g\right)^{-1} J_{\tilde{\psi}_i}{}^T \sqrt{\det\left(J_g{}^T J_g\right)} \tag{44}$$

for $i = 1, \ldots, M$, where J_{u_D} is understood as the Jacobian matrix of u_D with respect to x at point $g(\tilde{x})$.

The elements of $\mathbf{M}_K$ and $\mathbf{S}_K$ and the components of $\mathbf{f}_K$ are distributed to the global matrices $\mathbf{M}$ and $\mathbf{S}$ and the global vector $\mathbf{f}$. Let us recall that for each $\psi_i \in \Psi_K$ there exists $\varphi_l \in \Phi$ such that $\varphi_l|_K \equiv \psi_i$, as stated in Section 3.2. This statement defines a map from $\{1, \ldots, M\}$ to $\{1, \ldots, N\}$ mapping i to l. Denoting the map by γ_K, we can assert that $\varphi_{\gamma_K(i)}|_K \equiv \psi_i$ for each $\psi_i \in \Psi_K$. Consequently, each element $(M_K)_{i,j}$ of $\mathbf{M}_K$ is added to $M_{\gamma_K(i),\gamma_K(j)}$ of the global matrix $\mathbf{M}$, and in the same manner $\mathbf{S}$ and $\mathbf{f}$ are assembled.

To construct mapping γ_K, global information about the mesh is necessary. Typically, each basis function $\varphi_l \in \Phi$ is associated with a mesh entity, e.g., a vertex, an edge or a cell, and the global index of this entity within the mesh determines the index l of φ_l. On the element level, the local index i of $\psi_i \in \Psi_K$ is determined using the local index of the associated mesh entity within the cell K. Thus, the mapping γ_K is usually based on the local-to-global index mappings of the corresponding mesh entities.

3.4 Solution of the linear system

Equation (13) represents a system of N linear equations which can be written in the form

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{45}$$

where $\mathbf{x} = \mathbf{u}^k$ is the unknown vector, $\mathbf{A} = \mathbf{M} + \tau\mathbf{S}$ and $\mathbf{b} = \mathbf{M}\mathbf{u}^{k-1} + \tau\mathbf{f}$. The coefficient matrix $\mathbf{A}$ is symmetric positive definite and sparse. The system (13) can be solved by any method for the solution of linear systems, e.g., the conjugate gradient method.

4 Conclusion

We implemented the finite element method for the heat equation in C++. The implementation is based on the unstructured mesh library presented in our last year's article [1]. The mesh library enables the implementation to operate on various types of meshes,

e.g., triangular, quadrilateral, tetrahedral and hexahedral, in an arbitrary dimensional space. Furthermore, it was designed with its future adaptation to GPUs in mind. The implementation of the finite element method is general. It supports several types of finite elements, and it could be used for the numerical solution of various problems.

References

[1] V. Žabka and T. Oberhuber. Design of a General-Purpose Unstructured Mesh in C++. In 'Doktorandské dny 2012', P. Ambrož and Z. Masáková, (eds.), 299–306, (2012).

[2] P. Bastian et al. *The Distributed and Unified Numerics Environment (DUNE) Grid Interface HOWTO*, version 2.3-svn, (September 2013). Downloaded from `http://www.dune-project.org/doc/`.

[3] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* Studies in Mathematics and its Applications. Elsevier Science, (1978).

[4] A. Dedner, R. Klöfkorn, M. Nolte, and M. Ohlberger. *A Generic Interface for Parallel and Adaptive Scientific Computing: Abstraction Principles and the DUNE-FEM Module.* Computing **90** (2010), 165–196.

[5] DUNE team. *DUNE-PDELab Howto*, (March 2013). Downloaded from `http://www.dune-project.org/pdelab/`.

[6] C. Geuzaine and J.-F. Remacle. *Gmsh: A 3-D Finite Element Mesh Generator with Built-in Pre- and Post-Processing Facilities.* International Journal for Numerical Methods in Engineering **79** (2009), 1309–1331.

[7] K. Rupp. *ViennaFEM 1.0.0 User Manual.* Institute for Microelectronics and Institute for Analysis and Scientific Computing, TU Wien, (February 2012). Downloaded from `http://viennafem.sourceforge.net/`.

[8] J. Schöberl. *NETGEN User Manual*, (January 2009). Downloaded from `http://sourceforge.net/apps/mediawiki/netgen-mesher/`.

On Two Scale Approaches to the Frost Heave Modelling*

Alexandr Žák

2nd 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 two scale approaches to the mechanical manifestation modeling of freezing saturated soils. The first approach involves a macro-scale description of the problem. The mathematical model of two-dimensional two-phase system is designed. It comprises the modified heat equation involving the phase change of the pore water and the system of the Navier equations describing deformations of the body. Both equation types are coupled with the term which is related to the phase transition and which springs from the empirical considerations. Computational studies of the model for the control of the structural conditions within the mechanical heterogenious soil medium loaded by a concrete structure are presented. The second approach represents the pore-scale description of the problem. Several basic ideas regarding the local conditions of balance and mechanisms of the causes of the soil heaving inception under the thermal gradient are summed up. The preliminary simulations of the pore-scale freezing dynamics are shown.

Keywords: freezing, model, phase-transition, soil, heaving

Abstrakt. Tento příspěvek pojednává o dvojím přístupu k modelování mechnických projevů zamrzjících saturovaných zemin. První z přístupů popisuje v makro měřítku dvoudimenzionální termoelastický model dvoufázového systému. Model zahrnuje modifikovanou rovnici tepla popisující fázový přechod vody v pórech půdního materiálu a dále systém Navierových rovnic popisující deformaci tělesa. Oba typy rovnic jsou provázány členem, který je vztažený k fázovému přechodu a který vychází z empirických úvah. V příspěvku jsou ukázány počítačové studie tohoto modelu pro řízení strukturálních podmínek v mechanicky heterogenním půdním médiu zatíženém betonovou konstrukcí. Druhý přístup představuje problém na úrovni porézní struktury. Jsou shrnuty některé základní představy ohledně podmínek lokální rovnováhy a mechanismy příčin vzniku půdního vzdouvání při výskytu teplotního gradientu. Prezentovány jsou prvotní simulace dynamiky zamrzání na úrovní jednotlivých pórů.

Klíčová slova: zamrzání, model, fázová změna, půda, vzdouvání

*Partial support of the project of the "Numerical Methods for Multi-phase Flow and Transport in Subsurface Environmental Applications, project of Czech Ministry of Education, Youth and Sports Kontakt ME10009, 2010-2012" and of the project "Advanced Supercomputing Methods for Implementation of Mathematical Models, project of the Student Grant Agency of the Czech Technical University in Prague No. SGS11/161/OHK4/3T/14, 2011-13".

1 Introduction

During the temperature shifts of ground surface around 0°C, several qualitative property changes of upper soil layer can occur as a consequence of the phase change of water in pores. They include both mechanical and thermal property changes, and their range is substantial after a sufficient amount of the pore water, which is usually over 80 % of the soil porosity, is reached. Therefore, the saturated soil model is a convenient simplification for describing soil freezing problems. This model is used in our consideration as well.

One of the phenomena associated with the freezing of the high water content soil is the upward movement of the frozen ground. It is called the frost heave ([2],[6], [7]) and is caused by the formation of ice structures in the soil , which tend to grow as the freezing descends and which generate extra stresses affecting significantly the mechanical behavior of the soil.

One of our objectives is modeling the way how the freezing processes, including the frost heave, affect soil mechanical properties in macro-scale under various condition and heterogeneous properties.

2 Macro-scale model

Since the pore water interacts with the structure of porous medium, the water freezing conditions vary locally in the pore-scale, and water does not exhibit the phase transition all at once in a pore. For this reason taking approach from [8], it is convenient to define a function useful in describing the frozen state of soil:

$$\phi(T) = \begin{cases} 1 & : \quad T \geq T_\star \\ \dfrac{|T_\star|^b}{|T|^b} & : \quad T < T_\star \end{cases} , \tag{1}$$

where T is the soil temperature in °C, $T_\star$, $T_\star < 0$°C, is the freezing point depression (the freezing point of the pore water), η is the soil porosity, and b is a positive soil parameter. The function ϕ describes liquid water content in the pores, and its shape is supported with a number of experiments.

Heat balance reflecting the phase transition can be expressed by

$$C\frac{\partial}{\partial t}T + L\eta\frac{\partial}{\partial t}\phi(T) = \nabla\cdot(\lambda\nabla T) \ , \tag{2}$$

where C is the volumetric heat capacity of soil and λ is the effective thermal conductivity. They can be further related to the frozen water content as follows

$$C = C_f(1-\phi) + C_u\phi\,, \quad \lambda = \lambda_f^{1-\phi}\lambda_u^{\phi} \tag{3}$$

$$C_f = C_s(1-\eta) + C_i\eta\,, \quad C_u = C_s(1-\eta) + C_l\eta \tag{4}$$

$$\lambda_f = \lambda_s^{1-\eta}\lambda_i^{\eta}\,, \quad \lambda_u = \lambda_s^{1-\eta}\lambda_l^{\eta} \tag{5}$$

where subscripts f, u, s, i, and l denote heat capacity and thermal conductivity of frozen soil, unfrozen soil, solid particles, ice, and liquid water, respectively.

To cover the mechanical manifestations during the soil freezing, the soil is viewed as continuum, and the relation between the displacement vectors u, v and the temperature is proposed. Including the Navier equations in the model (2) and adding a linking term, the following governing system is applied

$$\begin{bmatrix} 0 \\ \rho\dfrac{\partial^2}{\partial t^2}u \\ \rho\dfrac{\partial^2}{\partial t^2}v \end{bmatrix} + \begin{bmatrix} \left(C + L\eta\dfrac{\mathrm{d}}{\mathrm{d}T}\phi\right)\dfrac{\partial}{\partial t}T \\ 0 \\ 0 \end{bmatrix} + \nabla\cdot\Gamma = 0 \tag{6}$$

where

$$\Gamma = \begin{bmatrix} -\lambda\dfrac{\partial}{\partial x}T & , & -\lambda\dfrac{\partial}{\partial y}T \\ -E\dfrac{(1-\nu)\dfrac{\partial}{\partial x}u + \nu\dfrac{\partial}{\partial y}v}{(1+\nu)(1-2\nu)} + \xi(T)\,, & & \dfrac{-E}{2(1+\nu)}\left(\dfrac{\partial}{\partial y}u + \dfrac{\partial}{\partial x}v\right) \\ \dfrac{-E}{2(1+\nu)}\left(\dfrac{\partial}{\partial y}u + \dfrac{\partial}{\partial x}v\right) & , & -E\dfrac{\nu\dfrac{\partial}{\partial x}u + (1-\nu)\dfrac{\partial}{\partial y}v}{(1+\nu)(1-2\nu)} + \xi(T) \end{bmatrix}, \tag{7}$$

$$\xi(T) = \chi\vartheta\left(T_\star - T\right)\,, \tag{8}$$

E is Young's modulus, ν is Poisson's ratio, ϑ stands for the Heaviside step function, and χ is the internal stress rate. The linking term ξ represents an intuitive switch function of internal stress between frozen and unfrozen soil material, and its more exact design will be objective of further development. It is supposed to be derived from the pore-scale considerations.

However, the occurrence of such a component can be justified by an analogy to the linear constitutive equation derivation process as follows. Let Σ stand for the deformation potential, which is given as the product of the mass density of undeformed body and the free energy density for a deformable body, i.e.

$$\Sigma(e_{ij}, T) = \rho_0 f\left(e_{ij}, T\right), \tag{9}$$

where e_{ij} denotes the strain tensor. Then, the stress tensor is expressed as

$$\sigma_{ij} = \frac{\rho}{\rho_0}\frac{\partial\Sigma}{\partial e_{ij}}\,. \tag{10}$$

To obtain a linear dependence on e_{ij}, Σ can be assumed to be written in the following form

$$\Sigma\left(e_{ij}, T\right) = \Sigma_0\left(T\right) + \Sigma_{ij}\left(T\right)e_{ij} + \frac{1}{2}\Sigma_{ijkl}\left(T\right)e_{ij}e_{kl}\,. \tag{11}$$

Analogously to the process of incorporation of the linear thermal expansivity term and springing from the inspiration in the abrupt volume change of pure materials during freezing, the linear coefficient in (11) can be expressed as

$$\Sigma_{ij}\left(T\right) = a_0^{ij} - \beta_{ij}\vartheta\left(T_\star - T\right) . \tag{12}$$

Assuming small deformations, the density ratio reads

$$\frac{\rho}{\rho_0} \doteq \frac{1}{1 + e_{kl}\delta_{kl}} \doteq 1 - e_{kl}\delta_{kl} , \tag{13}$$

and (10) transforms into

$$\sigma_{ij} = \left(1 - e_{kl}\delta_{kl}\right)\left(a_0^{ij} - \beta_{ij}\vartheta\left(T_\star - T\right) + \Sigma_{ijkl}e_{kl}\right) . \tag{14}$$

Dropping products of functions of T and e_{kl} and assuming no initial stress in undeformed unfrozen body, i.e. $a_0^{ij} = 0$, (14) gives

$$\sigma_{ij} = -\beta_{ij}\vartheta\left(T_\star - T\right) + \Sigma_{ijkl}e_{kl} . \tag{15}$$

Multiplying the previous equation by the inverse tensor Σ^{-1}_{ijkl}, it is possible to state the meaning to the coefficients. It is clear that the volumetric expansion coefficients of freezing and the elastic coefficients are

$$\left[\!\left[e_{ij}\right]\!\right]_{T_\star} = \beta_{ij}\Sigma^{-1}_{ijkl} = \alpha_{kl} , \qquad \left(\frac{\partial\sigma_{ij}}{\partial e_{kl}}\right)_T = \Sigma_{ijkl} , \tag{16}$$

respectively. Considering the material to be isotropic, the number of the independent coefficients in (15) decreases and the coefficient tensors reads

$$\Sigma_{ijkl} = \frac{E\nu}{(1+\nu)(1-2\nu)}\delta_{ij}\delta_{kl} + \frac{E}{2(1+\nu)}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) , \tag{17}$$

$$\alpha_{kl} = \alpha\delta_{kl} , \qquad \beta_{ij} = \frac{\alpha E}{3(1-2\nu)}\delta_{ij} = \chi\delta_{ij} . \tag{18}$$

When involved (15) with isotropic coefficients (17) and (18) in the dynamic balance equation, the Navier equations are obtained in the form as in (6).

The model given by (6) is used for preliminary quality studies of freezing ground situations, which involve investigations of the stress distribution through soil heterogeneities or assessments of the effect of the frost heave on building structures. Latter situation simulation is shown in Figure 4. It represents a cross-section of a simple concrete building constructed on freezing heterogenious ground (see Figure 1b) and shows mechanical processes within the structure when the soil heat leaks through the surface to the surroundings.

3 Pore-scale modeling

The soil mechanical property changes are a result of complex dynamic processes between the freezing pore water and the solid skeleton. Due to a force of attraction that water experiences in very close surroundings of a solid layer, a thin water film appears on the walls of the skeleton even within the frozen soil. When the solid skeleton is continuous under local pressure conditions, the film creates a continuous liquid net connected with unfrozen water reservoir below freezing soil. This enables water to flow through freezing zone until a discontinuity in the liquid net is reached. It occurs at a level, where the effective stress, σ_e, of the skeleton is fully supported by the stress produced by pore content reaction, σ_n. At the level, the solid particles are no more pressed horizontally to each other; the discontinuity appears, and cumulating water freezes and initiates an ice lens.

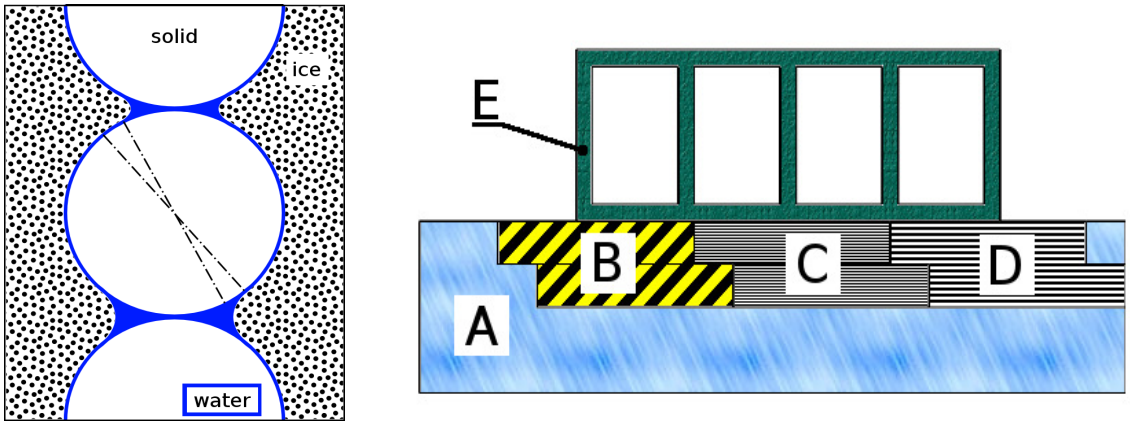
The basic force balance in the considered soil volume is expressed by the Terzaghi equation

$$P = \sigma_e + \sigma_n \,, \tag{19}$$

where P stands for the overburden pressure. As the phases are assumed to be continuous, σ_n is given by

$$\sigma_n = \zeta p_l + (1 - \zeta) p_i \,, \tag{20}$$

where ζ is the stress partition function ($0 \leq \zeta \leq 1$, $\zeta = 1$ when pores are filled only with water), p_i is the gage pressure of the pore ice, and p_l is the gage pressure of the pore water.



(a) (b)

Figure 1: (a) Scheme of the ice propagation under the vertical thermal gradient at the pore-scale level. Angles mark out the asymmetric areas. (b) Cross-section of soil ground with a structure on it. A, B, C, D stand for different soil types; E stands for concrete.

Stress σ_n is increasing through the freezing zone as the result of an asymmetric interactions of the film and the propagating pore ice. The ice propagation can be related to the temperature by the Clapeyron equation, which can be derived from free energy consideration (for more detail see [4]) in the following form:

$$\frac{p_l}{\rho_l} - \frac{p_i}{\rho_i} = \frac{lT}{T_a} \,, \tag{21}$$

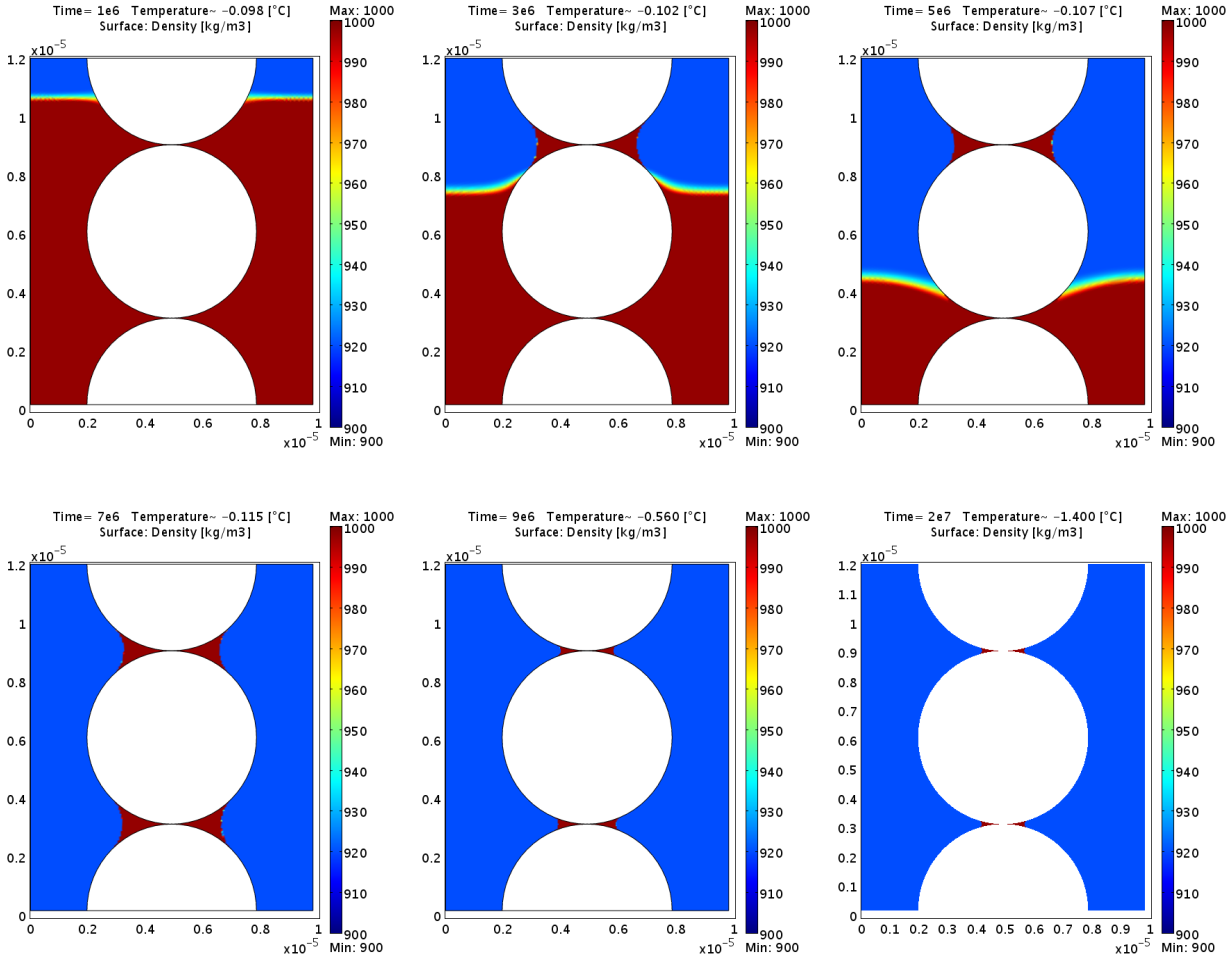


Figure 2: Pore-scale simulation of ice (grey) propagation under vertical thermal gradient through freezing water (dark) in pores around solid particles (white).

where l is the specific latent heat of freezing of water, ρ_i is the ice density, ρ_l is the water density, and T_a is the absolute temperature. At thermodynamic equilibrium, the ice pressure equals to the gage pressure of water, however, if the ice-water interface is curved, p_i and p_l differ. The difference is

$$p_l - p_i = \sigma_{il}\kappa\,, \tag{22}$$

where σ_{il} is the surface tension of an ice-water interface and κ is the mean curvature of the interface. From above equation it can be seen that the pressure conditions are determined by the interface curvature, i.e. by the geometry, and by the temperature.

When the balance of forces on an inner solid particle of a static column in the freezing zone (see Figure 1a) is considered, it follows from (21) and the temperature gradient that the areas of the film pressure action on the upper and lower hemisphere are not symmetric, and thus there is a downward component of force. The film pressure p_f is given by

$$p_f = p_i + \frac{2\sigma_{il}}{R+\tau}\,, \tag{23}$$

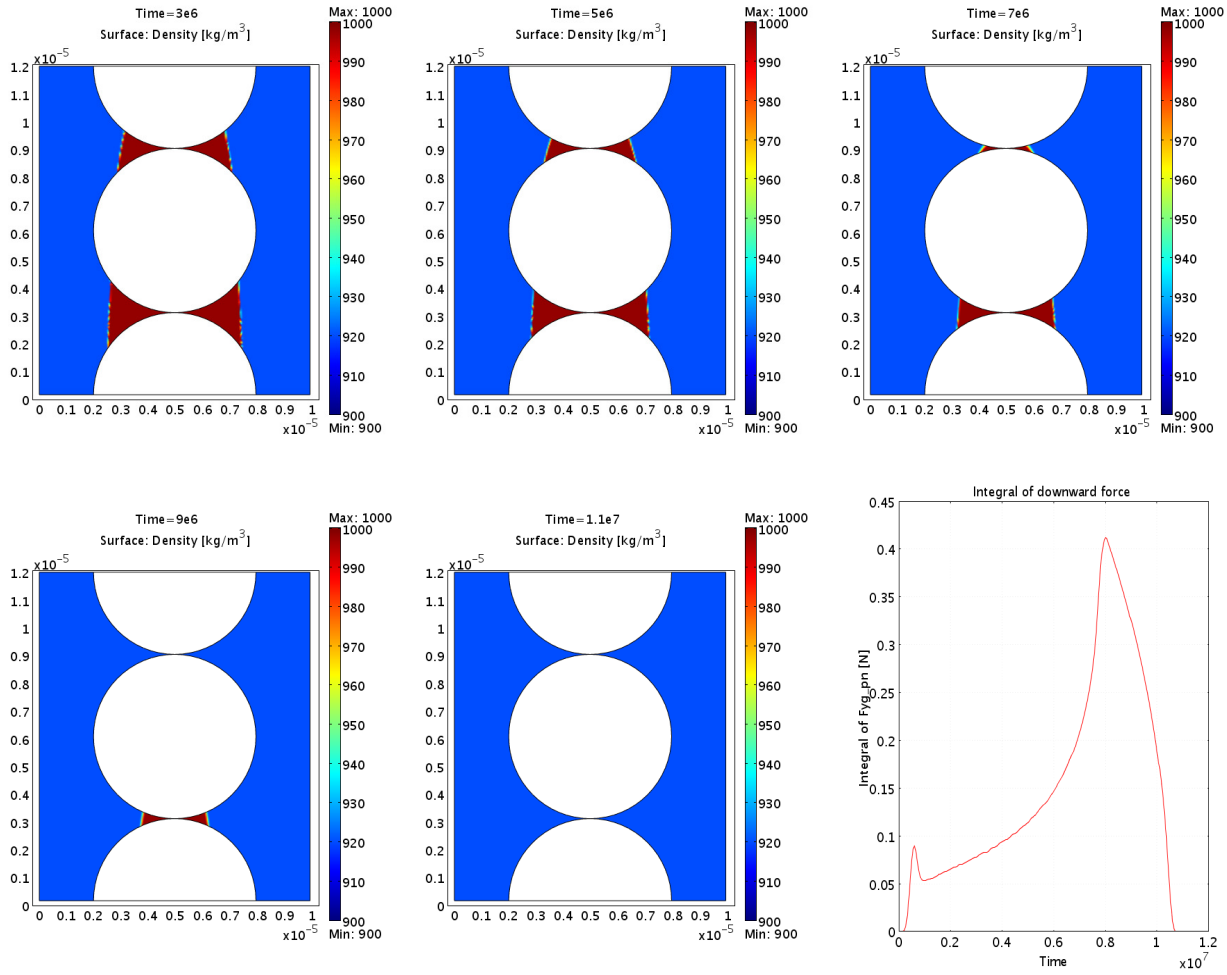


Figure 3: Pore-scale simulation of ice (grey) propagation under vertical thermal gradient through freezing water (dark) in pores around solid particles (white). Progress without curvature in menisci. Last figure illustrates the resultant surface force acting on the middle particle.

where R is the radius of the supposed particle and τ is the film thickness. In addition to this component, an another can by derived from a vertical gradient of p_f on the symmetric parts of the areas. If $R \gg \tau$, it is

$$\frac{\partial p_f}{\partial y} = \frac{\partial p_i}{\partial y} = \frac{-\rho_i l}{T_{a0}} \frac{\partial T}{\partial y}, \tag{24}$$

and the gradient is positive. Therefore, the component acts again downward.

The downward forces acting on every particle within the freezing zone represent the distributed force on the solid skeleton. This force is associated with an equal and oppositely distributed force on the pore content, which tend the ice body to move against the thermal gradient.

The process of new ice lens initiation is allowed when σ_n fully supports the load at some level, i.e. $\sigma_n = P$ or equivalently $\sigma_e = 0$. The pore water pressure at this level is

minimal. Therefore, the water flows from below to the level, where it cumulates, freezes, and makes the ice lens to grow. This remain until another lens is created below. This mechanism is described in more detail in [3], [5].

Several computational studies of the pore-scale ice propagation and its effect on the single particle have been performed and their result are in Figures 2 and 3.

4 Conclusions

The presented macro-scale model includes a basic heat and force balance and has been designed for the purpose of a preliminary study of structural changes in saturated soils caused by the phase transition of the water content due to alternations of climatic conditions. Although the model is based on the continuum approach and built on simplified relations, the produced simulations reflect adequately common empirical knowledge of the soil freezing and thawing process and the related mechanical manifestations. Further development will involve an application of more sophisticated and descriptive relations based on dynamical structure of freezing soil.

To fulfill this objective, the pore-scale structure modeling has been summed up, and numerical studies of the local balance conditions has begun.

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] S. Taber, "Frost Heaving", The Journal of Geology, Vol. 37, No. 5, pp. 428–461, Jul. - Aug. 1929.

[3] R. D. Miller, "Frost Heaving in Non-Colloidal Soils", in Proc. 3rd Int. Conference on Permafrost, pp. 707–713, 1978.

[4] 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.

[5] K. O'Neill and R. D. Miller, "Exploration of a Rigid Ice Model of Frost Heave", Water Resources Research, Vol. 21, No. 3, pp. 281–296, 1985.

[6] A. C. Fowler, "Secondary Frost Heave in Freezing Soils", SIAM J. APPL. Math., Vol. 49, No. 4, pp. 991–1008, 1989.

[7] 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.

[8] D. J. Nicolsky, V. E. Romanovsky, G. G. Panteleev: "Estimation of soil thermal properties using in-situ temperature measurements in the active layer and permafrost", Cold Regions Science and Technology, Vol. 55, pp. 120–129, 2009.

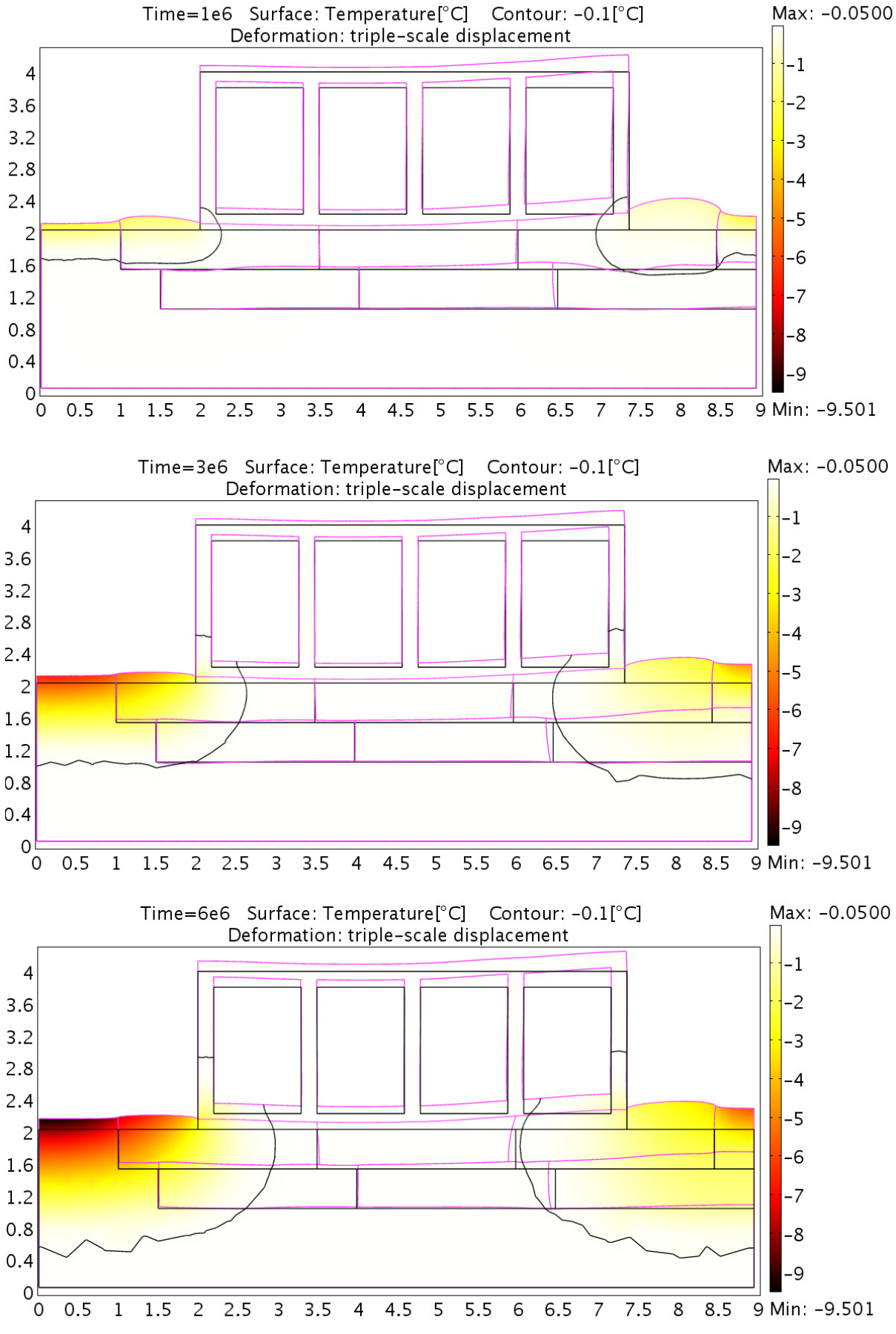


Figure 4: Strain evolution of the concrete construction during ground soil freezing