DOKTORANDSKÉ DNY 2011

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

11. a 25. listopadu 2011

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2011
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 303, Vydání 1.

ISBN 978-80-01-04907-5

Seznam příspěvků

Předmluva

Již pošesté se letos scházejí doktorandi oboru Matematické inženýrství studijního programu Aplikace přírodních věd na workshopu Doktorandské dny, který se koná pravidelně na katedře matematiky FJFI. Cílem setkání je sledovat pokrok ve výzkumu jednotlivých doktorandů a umožnit jim získat přehled o tematice svých kolegů. Oběma účelům slouží jednak ústní prezentace na workshopu konaném ve dnech 11. a 25. listopadu 2011, jednak tento sborník, přinášející texty přednášek, resp. jejich abstrakty. Toto je oproti předcházejícím vydáním novinkou. Oborová rada se rozhodla umožnit doktorandům vystoupit na Doktorandských dnech s příspěvkem již publikovaným v odborném časopise nebo proceedings jiné konference. V takovém případě je v našem sborníku pouze abstrakt přednášky.

I letos workshop proběhne v několika paralelních sekcích dělených podle tematického zaměření přednášejících doktorandů. To pokrývá numerické a stochastické modely, jakož i matematické základy moderní teoretické informatiky a fyziky.

Za morální podporu děkujeme Katedře matematiky a Dopplerovu ústavu pro matematickou fyziku a aplikovanou matematiku při FJFI. Finančně je workshop podpořen Studentskou grantovou soutěží při ČVUT v rámci grantu SVK 17/11/F4.

Editoři

Local Hausdorff Distance Maps Utilized in Alzheimer's Disease Automatic Detection

Kamil Barbierik

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

advisor: Jaromír Kukal, Department of Software Engineering in Economy, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. An accurate and early diagnosis of the Alzheimer's disease (AD) is of fundamental importance for the patient's medical treatment. Single Photon Emission Computed Tomography (SPECT) images are commonly used by physicians to assist the diagnosis, rating them by visual evaluations. In this work I present an automated diagnosis tool based on local Hausdorff distance maps (LDMaps). The proposed algorithm, in the learning mode, generates two average LDMaps. The first aggregates the differences between healthy brain SPECT pictures and the etalon. The second one aggregates the dissimilarities between pictures of brains affected by AD and the same etalon. In the testing mode it compares LDMaps of patients' brain SPECT pictures against the averages from learning process and according to which average is the patient's brain LDMap more similar to, is classified as healthy or AD.

Keywords: Alzheimer's disease, SPECT, Hausdorff distance, local Hausdorff distance map, automatic detection

Abstrakt. Pre efektívnu liečbu pacienta s podozrením, že trpí Alzheimrovou chorobou, je nevyhnutná jej správna diagnóza stanovená v čo najskoršom štádiu tohto degeneratívneho ochorenia. V praxi v tejto veci pomáha lekárom napríklad tomografia SPECT, pomocou ktorej môžu pacientov mozog vizuálne preskúmať a na základe určitých znakov rozhodnúť, či u pacienta existujú náznaky Alzheimrovej choroby. V tejto práci uvádzam spôsob automatického rozpoznávania Alzheimrovej choroby na základe dát o mozgu získaných tomografiou SPECT. Rozpoznávanie funguje na metóde porovnávania množín založenej na Hausdorffovej vzdialenosti. V prvej fáze algoritmu v tzv. fáze učenia sú vygenerované dve priemerné lokálne rozdielové mapy mozgu. Prvá mapa obsahuje informácie o tom, ako sa v priemere líšia zdravé mozgy od zdravého etalonu. Druhá lokálna rozdielová mapa naopak agreguje rozdiely medzi mozgami postihnutými Alzheimrovou chorobou a už spomínaného zdravého etalonu. V druhej fáze rozpoznávania sa potom mapa lokálnych vzdialeností mozgu pacienta, ktorá vznikne porovnaním so zdravým etalonom, porovná s oboma priemernými lokálnymi rozdielovými mapami. Na základe vzdialenosti medzi nimi, ktorá vychádza znova z Hausdorffovej vzdialenosti, sa mozog klasifikuje ako zdravý, alebo naopak ako postihnutý Alzheimrovou chorobou.

Kľúčové slová: Alzheimrova choroba, SPECT, Hausdorffova vzdialenosť, mapa lokálnych Hausdorffových vzdialeností, automatická detekcia

1 Introduction

Alzheimer's disease (AD) is the most frequent type of dementia. This serious health problem affects middle-aged and elderly people. The elderly are the fastest growing part of the population, and increases in life expectancy will inevitably lead to a further increase in the prevalence of Alzheimer's disease. Currently no cure is available for AD but different strategies are under development that are expected to delay the disease, and possibly prevent or offset the onset of AD in early stages. Therefore it is very important to recognize individuals with high risk for developing AD as soon as possible. These people may particularly benefit from early therapeutic interventions. Currently, the diagnosis of AD is based primarily on clinical and neuropsychological assessments. There is evidence that medical imaging examinations like 3D SPECT have higher predictive value than the clinical measures in identifying the presence of a progressive neurodegenerative disease. However, analysis of brain images is not a simple task because patterns of brain degeneration are highly variable and complex.

Several attempts were made towards automatic recognition of AD symptoms using various medical imaging systems. Most attempts were based on analysis of specific brain segments which requires segmentation of the image into regions and afterwards analyzing these segments. These techniques rely heavily on manual or semi-automatic extraction of the structures of interest. Furthermore, they are limited by the fact that the brain atrophy usually involves many brain regions and different regions are affected at different stages of the disease. Therefore, current techniques are focusing on the use of the entire brain pattern. Disadvantage of this approach is that it is necessary to analyze much greater amount of data. This, consequently, leads to need of a great computation power, or a need of applying sophisticated reduction techniques on input data.

Image processing algorithm for recognizing the AD introduced in this paper is working with the entire brain pattern. It is very simple and its computation is fast enough and memory efficient. It is based on a comparison of 3D SPECT images of brain where the dissimilarity is measured using a local Hausdorff distance maps. The decision whether the subject of interest is suffering from AD is made according to degree of dissimilarity between its brain image and images of healthy and AD brains.

2 Hausdorff distance

Hausdorff distance is a very powerful tool for measuring dissimilarity between two sets. The set in this paper will represent 3D SPECT image of brain which is a 3D set of discrete points. Using Hausdorff distance, we can measure a degree of mismatch between two object shapes very precisely. Unlike feature based methods, Hausdorff distance is zero if and only if the shapes of objects are exactly the same and increases with growing dissimilarity. What is more, if we need to minimize the Hausdorff distance over the space of some transformation parameters, any transformation of object (rotation, affine transform...) can be taken into consideration. An advantage is also the possibility of independently using the directed distances that Hausdorff distance is composed of.

On the other hand, the major disadvantage is computation burden of discussed measure and the fact that it is extremely sensitive to outliers. The latter disadvantage is

very restrictive, because it is almost impossible to receive noise free data sets from any scanning apparatus. Therefore several modifications of Hausdorff distance were proposed (MHD [3], PartialHD [2], WindowedHD [4] and more), that yield much more satisfactory results when applied on noisy data sets. This paper focuses on the modification that is local i.e. windowed Hausdorff distance. The Hausdorff Distance and its modifications are very often used in a different object matching or image registration algorithms [6], [7] even in medicine [8]. Very often the utilization of HD may be found in the specific object matching problem - face recognition [9], [10].

2.1 Conventional Hausdorff distance

Hausdorff distance is a max-min distance defined by the following definition.

Definition: Let $\{M, \varrho\}$ be a metric space where M is a finite set of points. Let $A = \{\vec{a_1}, \ldots, \vec{a_p}\}$ and $B = \left\{\vec{b_1}, \ldots, \vec{b_q}\right\}$ be two subsets of M. We define Hausdorff distance $H(A, B)$ by:

$$H(A,B) := \max\left\{\max_{\vec{a}\in A}\min_{\vec{b}\in B}\varrho\left(\vec{a},\vec{b}\right), \max_{\vec{b}\in B}\min_{\vec{a}\in A}\varrho\left(\vec{a},\vec{b}\right)\right\} \tag{1}$$

where ϱ is defined as

$$\varrho(\vec{x},\vec{y}) = \sqrt{\sum_{k=1}^{r}(x_k - y_k)^2} \tag{2}$$

where $\vec{x}, \vec{y} \in \mathbb{R}^r$.

Note: The definition of Hausdorff distance can be derived by a series of steps naturally extending the distance function ϱ in the underlying metric space $\{M, \varrho\}$ as follows:

Let $\{M, \varrho\}$ be a metric space. Given $\vec{a} \in M$ and non-empty set $B \subset M$ we define a distance $dist\,(\vec{a}, B)$ between point $\vec{a}$ and the set B by:

$$dist\,(\vec{a}, B) := \min_{\vec{b}\in B}\varrho\left(\vec{a},\vec{b}\right) \tag{3}$$

Using the previous distance we define $h\,(A, B)$, the distance between A and B where $A, B \subset M$:

$$h\,(A,B) := \max_{\vec{a}\in A} dist\,(\vec{a}, B) \tag{4}$$

$h\,(A, B)$ is called the *directed Hausdorff distance.*
To obtain an undirected Hausdorff distance, which is a metric, it is necessary to combine the two directed Hausdorff distances $h\,(A, B)$ and $h\,(B, A)$:

$$H\,(A,B) := \max\{h\,(A,B), h\,(B,A)\} \tag{5}$$

2.2 Windowed Hausdorff distance

In January 2007 in a preprint and later published in [4] a new approach was proposed for determining dissimilarity between two sets using Hausdorff-like distances. While previous and other modifications of Hausdorff distances were global and except the classical Hausdorff distance and the MHD [3] required some input arguments, the windowed Hausdorff distance operates locally and does not require any input parameter. Furthermore, while the global ones produce only one number that expresses the dissimilarity between two sets, the windowed Hausdorff distance produce a dissimilarity map where local mismatches can be examined.

The definition involves three different directed Hausdorff distances, which supplies three possible cases of presence of set points in the window. It makes use of the distance to the frontier $Fr(W)$ of the window W. In this discrete case we consider that the frontier $Fr(W)$ is between the elements. For example the frontier of the ball $B(x,n)$ is the line between $B(x,n)$ and $B(x,n+1) \setminus B(x,n)$. The distance of a point $x \in B(x,n)$ to the frontier is equal to the distance to the elements just behind the frontier.

Definition: Let A, B be two bounded sets of $\mathbb{R}^r$. $H_w(A,B) = \max\{h_w(A,B), h_w(B,A)\}$ where:

- If $A \cap W \neq \emptyset \wedge B \cap W \neq \emptyset$

$$h_w(A,B) := \max_{\vec{a} \in A \cap W} \left[\min_{\vec{b} \in B \cap W} \varrho\left(\vec{a}, \vec{b}\right), \min_{\vec{w} \in Fr(W)} \varrho(\vec{a}, \vec{w}) \right] \tag{6}$$

- If $A \cap W \neq \emptyset \wedge B \cap W = \emptyset$

$$h_w(A,B) := \max_{\vec{a} \in A \cap W} \left[\min_{\vec{w} \in Fr(W)} \varrho(\vec{a}, \vec{w}) \right] \tag{7}$$

- If $A \cap W = \emptyset$

$$h_w(A,B) := 0 \tag{8}$$

The algorithm that computes the local Hausdorff distance at each non zero pixel of both images consist of a sliding window whose radius is growing at each point until reaches the optimal value. This value is then recorded into *local distance map* (LDMap) at the point of the center of the window. However, this algorithm is time consuming. For $m \times m$ images the computation complexity is $O(m^4)$. To save most of the time computation, it is possible to utilize transform distances of compared images. Fast algorithms have been developed for computing distance transforms of binary images. Taking the advantage of these algorithms, the computation complexity may be reduced to $O(m^2)$. In the following section, the LDMap is defined using the transform distance of compared images.

3 Local distance map

Definition: Let A and B be two non-empty finite sets of points of $\mathbb{R}^r$ and let $\vec{x} \in \mathbb{R}^r$, the local distance map $LDMap(\vec{x})$ is defined by:

$$LDMap(\vec{x}) = |\mathbb{I}_A(\vec{x}) - \mathbb{I}_B(\vec{x})| \max\{dist(\vec{x}, A), dist(\vec{x}, B)\} \tag{9}$$

where $\mathbb{I}_A(\vec{x})$ is equal to 1 if $\vec{x} \in A$ and 0 otherwise.

The maximum value in the LDMap is the Hausdorff distance $H(A, B)$. This value is present in the map at least once.
The notion of local dissimilarity is illustrated by following 2D image.

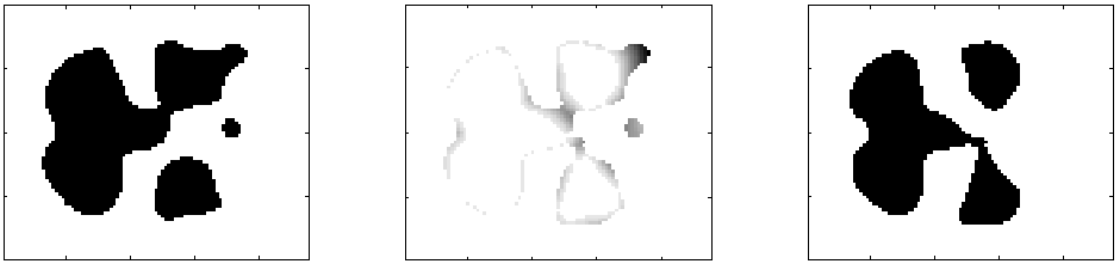


Figure 1: Slices of two brains at the same level (left, right) and slice of corresponding 2D LDMap (center)

The 3D image below shows the utilization of LDMaps for highlighting the most mismatched areas in the brain picture of the subject of interest compared to the etalon.

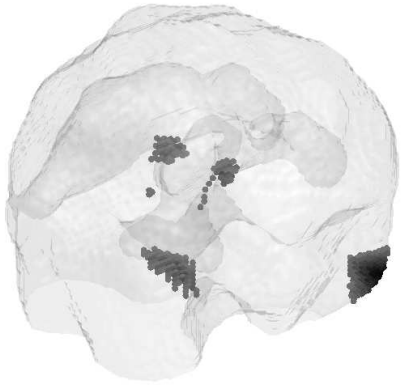


Figure 2: 3D LDMap utilized for highlighting the most significant differences from the healthy etalon

4 Proposed method for recognition of brains affected by Alzheimer's disease using LDMaps

The method utilizes the information included in LDMaps and is projected to keep the computation burden low:
Notation:

E	Binarized etalon
HLT	Set of binarized healthy brains images
ALZ	Set of binarized AD brains images
$HLT\_L(ALZ\_L)$	Set of healthy (AD) images chosen from $HLT(ALZ)$ for learning
$HLT\_T(ALZ\_T)$	Set of healthy (AD) images chosen from $HLT(ALZ)$ for testing
$HLTmap(ALZmap)$	LDMaps – results of comparison of $HLT\_L_i \in HLT\_L\,(ALZ\_L_i \in ALZ\_L)$ against E
$avgHLTmap(avgALZmap)$	Average LDMap generated from $HLTmap_i \in HLTmap\,(ALZmap_i \in ALZmap)$
$HLTmap\_t(ALZmap\_t)$	LDMaps – results of comparison of $HLT\_T_i \in HLT\_T\,(ALZ\_T_i \in ALZ\_T)$ against E
$ALZdist2HLTavg_i$	Distance of $ALZmap\_t_i$ to $avgHLTmap$
$HLTdist2HLTavg_i$	Distance of $HLTmap\_t_i$ to $avgHLTmap$
$HLTdist2ALZavg_i$	Distance of $HLTmap\_t_i$ to $avgALZmap$
$ALZdist2ALZavg_i$	Distance of $ALZmap\_t_i$ to $avgALZmap$

Algorithm:

1. Preprocess SPECT images of brains – binarization.

2. Divide data into learning sets (HLT_L, ALZ_L) and testing sets (HLT_T, ALZ_T).

3. Create LDMaps form learning data – healthy vs. etalon: $HLTmap$ and AD vs. etalon: $ALZmap$

4. Generate average LDMaps from learning data (healthy: $avgHLTmap$ and AD: $avgALZmap$)

5. Create LDMaps from Testing data - healthy vs. etalon: $HLTmap\_t$ and AD vs. etalon: $ALZmap\_t$

6. Compute the similarity of particular LDMaps $(HLTmap\_t, ALZmap\_t)$ with averages $(avgHLTmap, avgALZmap)$

7. Classify the image according the value and mark whether classification was correct

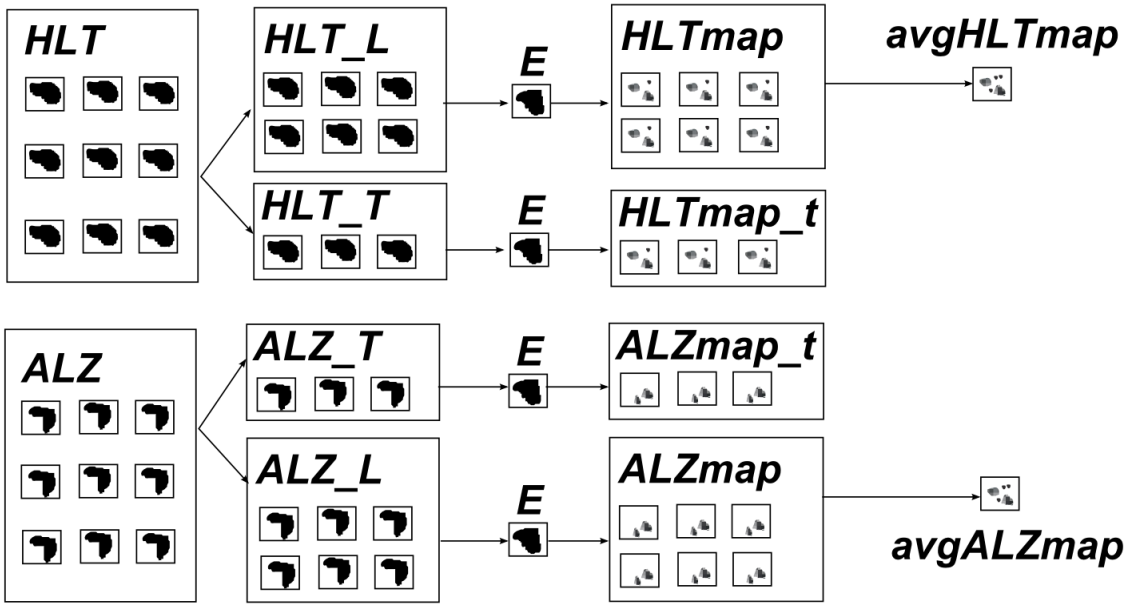


Figure 3: Schematic diagram of the algorithm (steps 2. - 5.): getting the average LDmaps from learning data and preparing LDmaps from testing sets

The figure (3) describes the first part of the algorithm where the set of healthy brain binarized pictures as well as set of binarized AD brains are divided into two parts. The larger parts are used for the learning process of the algorithm and the smaller parts are used for testing the method. At first, in the learning mode, LDMaps are created by comparing the elements of the learning sets with the healthy etalon. The aggregation of these LDMaps results in two average maps - healthy and AD average map. The healthy average map accumulates the differences of healthy brains against the healthy etalon. These differences are supposed to be not very marked and rather disseminated. On the other hand, the AD average map is expected to show more marked differences and rather localized. Except average maps, LDMaps of brain images from the testing sets are prepared. Every time the same healthy etalon is used.

The second part of the algorithm – classification – is captured on the figure below (4). In this part the testing healthy and the AD brain LDMaps are compared against the average maps. The dissimilarities are evaluated and accordingly the tested LDMaps are classified and marked weather the classification was correct or not.

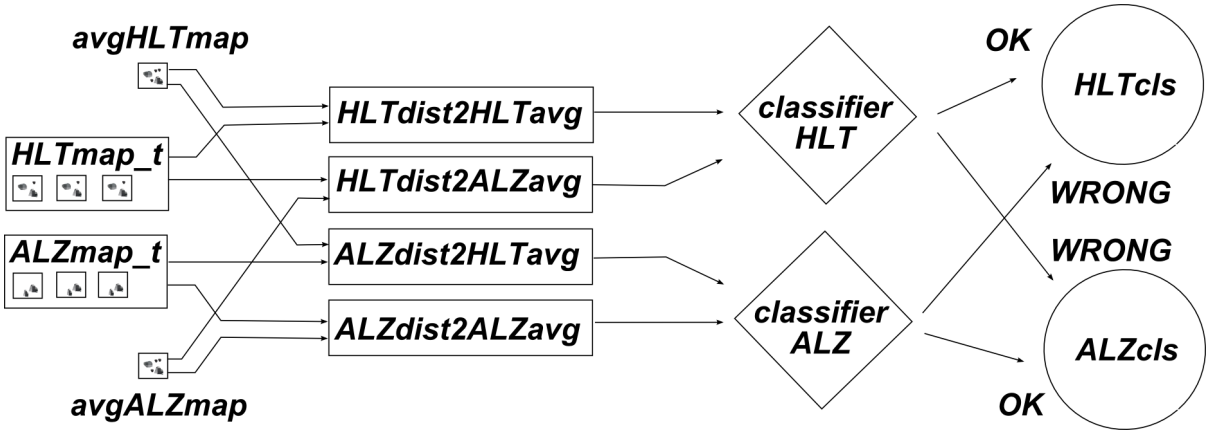


Figure 4: Schematic diagram of the algorithm (steps 6. and 7.): computing the distances between average LDMaps and LDMaps of testing samples, classifying the tested LDMaps and marking weather the classification was correct.

Further, each step of the proposed algorithm is explained into more detail:

1. In the first step every image $A \in HLT \cup ALZ$ is preprocessed. The values of images are linearly transformed to fit the range $\langle 0, 1\rangle$ and afterwards they are binarized using the user input threshold $t \in \langle 0, 1\rangle$. The output is a binary matrix B where

   - $B_{i,j} = 0$ where $A_{i,j} < t$ and
   - $B_{i,j} = 1$ where $A_{i,j} \geq t$.

2. Then, the HLT set as well as ALZ set are divided into two sets: a set for learning procedure (HLT_L, ALZ_L) and a set for testing purpose (HLT_T, ALZ_T). The ratio of HLT_L to HLT is the same as the ratio of ALZ_L to ALZ. Furthermore, the algorithm needs an etalon E which is also binarized using the same $t \in \langle 0, 1\rangle$ and constraints as in step 1. It can be computed as an average image of all healthy images of brains for example. Anyway, for needs of this paper, a healthy brain etalon was provided, which was checked and validated by professionals in the field.

3. The next two steps forms the learning process: $HLTmap(ALZmap)$ elements are created by comparison of each element of $HLT\_L(ALZ\_L)$ against E.

4. Average LDMaps $avgHLTmap$ and $avgALZmap$ are created:
   Let $HLT\_L_i(ALZ\_L_i)$ be the binarized image of healthy (AD) brain from the learning set $HLT\_L(ALZ\_L)$. Let $HLTmap_i(ALZmap_i)$ be the LDMap of comparison between $HLT\_L_i(ALZ\_L_i)$ and the etalon E. The average healthy and AD local distance maps are then computed as:

$$avgHLTmap = \frac{1}{\max_i (avgHLTmap_i)} \sum_{i=1}^{n} HLTmap_i \tag{10}$$

$$avgALZmap = \frac{1}{\max_i (avgALZmap_i)} \sum_{i=1}^{m} ALZmap_i \tag{11}$$

$avgHLTmap(avgALZmap)$ aggregates dissimilarities between healthy (AD) brains pictures and the etalon E. It contains information about how healthy (AD) patients' brains differ from the etalon.

5. The following step is testing. Images from $HLT\_T(ALZ\_T)$ are compared against the etalon E. The result is a set of LDMaps: $HLTmap\_t(ALZmap\_t)$.

6. LDMaps from step 5 are compared against $avgHLTmap$ and $avgALZmap$ and the dissimilarity between these images is computed. The directed Hausdorff distance (4) was chosen as a dissimilarity measure. The distance to both healthy and AD average LDMap is computed:

   - $HLTdist2HLTavg_i := h(HLTmap\_t_i, avgHLTmap)$
   - $HLTdist2ALZavg_i := h(HLTmap\_t_i, avgALZmap)$

   Similarly, the distances to both averages are computed for images from set $ALZmap\_t$:

   - $ALZdist2HLTavg_i := h(ALZmap\_t_i, avgHLTmap)$
   - $ALZdist2ALZavg_i := h(ALZmap\_t_i, avgALZmap)$

   LDMaps are gray 3D images. Before computing the directed Hausdorff distance (4) between two LDMaps it is necessary to binarize them. A user input threshold level may be used in this process. From the procedure how the average images are constructed is arising that all images from the learning set will be subsets of the corresponding average image after binarization:

   - $HLTmap_i \subset avgHLTmap$
   - $ALZmap_i \subset avgALZmap$

   Consequently, the following equations applies to directed distances:

   - $h(ALZmap_i, avgALZmap) = 0$ and also
   - $h(HLTmap_i, avgHLTmap) = 0$.

   While

   - $h(ALZmap\_t_i, avgHLTmap) \geq 0$ and
   - $h(HLTmap\_t_i, avgALZmap) \geq 0$.

   Thus, if $HLTmap(ALZmap)$ i.e. LDMaps from learning set will be passed into testing process instead of LDMaps from $HLTmap\_t(ALZmap\_t)$, they will always be classified correctly according to the following classification rules. It indicates the correct behavior of the algorithm.

7. The last step is the classification of images into two classes: $ALZcls$ or $HLTcls$; Marking of wrong and correct classification is done according the following rules: Set is classified as healthy ($HLTcls$) if the distance of $HLTmap\_t_i$ or $ALZmap\_t_i$ to $avgHLTmap$ is smaller than the distance to $avgALZmap$:

   - $ALZdist2HLTavg_i < ALZdist2ALZavg_i$ (wrong classification of test image to $HLTcls$)
   - $HLTdist2HLTavg_j < HLTdist2ALZavg_j$ (correct classification of test image to $HLTcls$)

   Set is classified as AD (into $ALZcls$) if the distance of $HLTmap\_t_i$ or $ALZmap\_t_i$ to $avgALZmap$ is smaller then the distance to $avgHLTmap$:

   - $ALZdist2HLTavg_i > ALZdist2ALZavg_i$ (correct classification of test image to $ALZcls$)
   - $HLTdist2HLTavg_j > HLTdist2ALZavg_j$ (wrong classification of test image to $ALZcls$)

   In case, when the LDMap of interest ($HLTmap\_t_i$ or $ALZmap\_t_i$) is within the same distance from the both of average LDMaps $avgHLTmap$ and $avgALZmap$, we cannot classify it to any class. It will be marked as wrong classification.

5 Experiment

Data for experiments consist of 55 3D SPECT images of brains marked by experts as brains affected by Alzheimer's disease and 91 3D SPECT images of healthy people brains. Furthermore, a healthy etalon is available. The external etalon of healthy brain image was obtained as an average over 2000 3D scans of healthy people, which were as well as healthy and AD samples normalized using Statistical Parametric Mapping (SPM5-Segment). The computations were accomplished using MATLAB software. Parameters were set as follows:

Threshold for image binarization $t = 0.35$; $|HLT\_L|/|HLT| = |ALZ\_L|/|ALZ| = 60\%$ ($|HLT\_L| = 55; |ALZ\_L| = 33; |HLT\_T| = 36; |ALZ\_T| = 22$)

5.1 Results

Experiment #	Wrong classification of healthy brains	Wrong classification of AD brains
1.	25%(**9**/36)	18%(**4**/22)
2.	14%(**5**/36)	14%(**3**/22)
3.	17%(**6**/36)	27%(**6**/22)
4.	28%(**10**/36)	9%(**2**/22)
5.	11%(**4**/36)	23%(**5**/22)

6 Conclusion

It was shown, that it is possible to automatically classify Alzheimer's disease quite successfully without any specific knowledge about the patient. What is more, using the local distance maps we are able to obtain specific local information while still retaining the advantage of global approach. Local information contained in local distance maps combined with appropriate display equipment may be very helpful for clinicians. It may help to focus the attention on the parts of a patient's brain which are most mismatched in comparison with healthy etalon. This may help to detect the developing Alzheimer's disease before any clinical symptoms appear.

In a future work one may also want to eliminate the binarization before every comparison to preserve as much information as possible and measure also the similarity in intensities. Furthermore, more sophisticated classification would be helpful, using for instance the support vector machine.

The algorithm proposed in this paper works with already spatially registered data sets, therefore the registration was not necessary in this case. However, HD and especially its modifications are suitable tools for object matching. Therefore, I plan to investigate the possibilities of using the local Hausdorff distance maps for object matching to be able to work with not registered brain images.

Acknowledgement: The support of grant OHK4-165/11 CTU in Prague is gratefully acknowledged. The author would also like to thank Helena Trojanová and Renáta Pichová from Clinique of Nuclear Medicine FNKV in Prague and Aleš Bartoš from Neurological clinique FNKV in Prague for providing the image data.

References

[1] A. M. J. Skulimowski. *Mathematical bases for the numerical evaluation of the Hausdorff distance*. Preprints of the 9th IMACS World Congress, Oslo, August 5-9, 1985; Vol. 5, pp.343-346.

[2] D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge. *Comparing images using the Hausdorff distance*. IEEE Transactions on pattern analysis and machine intelligence, vol. 15, no. 9, 1993, pp. 850-863.

[3] M. P. Dubuisson, A. K. Jain. *A modified Hausdorff distance for object matching*. In: Proc. 12th Internat. Conf. on Pattern Recognition, Jerusalem, Israel, October 1994, pp. 566-568.

[4] E. Baudrier, F. Nicolier, G. Millon, R. Su. *Binary-image comparison with local dissimilarity quantification*. Pattern Recognition, vol. 41, issue 5, May 2008, pp. 1461-1478.

[5] G. Fung, J. Stoeckel. *SVM feature selection for classification of Alzheimer's disease using spatial information*. Springer-Verlag London Limited 2006.

[6] Niu Li-pi, Jiang Xiu-hua, Zhang Wen-hui, Shi Dong-xin. *Image registration based on Hausdorff distance*. International Conference on Networking and Information Technology (ICNIT), 2010, pages: 252 – 256.

[7] Jian-xin Kang, Nai-ming Qi, Jian Hou *A Hybrid Method Combining Hausdorff Distance, Genetic Algorithm and Simulated Annealing Algorithm for Image Matching*. Second International Conference on Computer Modeling and Simulation, 2010, pages: 435 – 439.

[8] A. Fedorov, E. Billet, M. Prastawa, G. Gerig, A. Radmanesh, S. K. Warfield, R. Kikinis, N. Chrisochoides. *Evaluation of Brain MRI Alignment with the Robust Hausdorff Distance Measures*. Proceeding ISVC '08 Proceedings of the 4th International Symposium on Advances in Visual Computing, 2008.

[9] H. Dastmalch, J. Jafaryahya, R. Najafi, A. Daneshkhah. *Averaged Segmental Partial Hausdorff Distance for Robust Face Recognition*. Second International Conference on Intelligent Systems, Modelling and Simulation (ISMS), 2011, pages: 35 – 39.

[10] Huachun Tan, Yu-Jin Zhang *Computing Eigenface from Edge Images for Face Recognition Based on HD*. Fourth International Conference on Image and Graphics (ICIG), 2007, pages: 639 – 644.

Diskriminabilita afinních momentových invariantů

Petra Bednaříková

1. ročník PGS, email: petulado@centrum.cz
Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Jan Flusser, Ústav teorie informace a automatizace, AVČR

Abstract. Affine moment invariants are an important class of features used in image recognition. There is an infinite amount of these invariants, however only very few of them are independent. As part of our work we have applied several standard feature selection methods on a group of invariants and we were interested in finding patterns in selected invariants and based on this create some advice for the selection of invariants. The range of successful features was quite wide which indicated that the most successful features depends on the specific data. The first ten invariants, which contain eight independent invariants, have been on average significantly more successful than the other invariants. We have also noticed that generally invariants with lower order and weight have been more succesful probably because they are more robust to sampling error in the moment calculation.

Keywords: affine moment invariants, discriminability, image recognition, feature selection, independence

Abstrakt. Afinní momentové invarianty jsou důležitou třídou příznaků používaných v rozpoznávání obrazu, ale existuje jich nekonečné množství a je známo, že jen málo z nich je nezávislých. V rámci práce jsme na reálných datech z databáze listů aplikovali na skupinu invariantů vybrané standardní metody na výběr příznaků a snažili jsme se na základě výsledků nalézt nějaké zákonitosti nebo rady pro výběr příznaků. Vybraná škála příznaků byla poměrně široká, což poukazovalo na to, že nejlepší příznaky závisí na konkrétních datech. Jednou z vypozorovaných skutečností bylo, že relativně nejvyšší úspěšnost měla první desítka invariantů, o kterých je známo, že obsahují 8 nezávislých příznaků. Zároveň bylo viditelné, že úspěšnější v klasifikaci jsou invarianty s nižším řádem nebo vahou, zřejmě díky větší robustnosti vůči vzorkovací chybě při výpočtu momentů.

Klíčová slova: afinní momentové invarianty, diskriminabilita, rozpoznávání obrazu, výběr příznaků, nezávislost

Úvod

V typické klasifikační úloze v rozpoznávání obrazu se snažíme přiřadit objekt na obrázku do určité třídy na základě příznaků. Volba vhodných příznaků často záleží na charakteru konkrétní úlohy, ale jedním z důležitých požadavků na příznaky je invariantnost vůči transformacím obrázku.

Z tohoto důvodu hrají v rozpoznávání obrazu důležitou roli afinní momentové invarianty ([1]), které jsou invariantní vůči afinní transformaci prostorových souřadnic. Afinní

transformace není důležitá jen v dvourozměrných úlohách, ale také jako aproximace projektivní transformace při rozpoznávání fotografií trojrozměrné scény.

Pro každou klasifikační úlohu lze najít velké množství příznaků. Proto bývá v praxi často po definování příznaků dalším krokem snížení počtu příznaků používaných pro klasifikaci a to jak z důvodu výpočetního času, vyřazením nepodstatných částí dat, tak pro zlepšení úspěšnosti klasifikace (v některých případech může zvětšování počtu příznaků zvyšovat chybu klasifikace). Jelikož afinních momentových invariantů existuje teoreticky nekonečné množství, je výběr příznaků obzvláště důležitý. Mezi afinními momentovými invarianty se vyskytuje mnoho různých závislostí: od jednoduchých až po složité polynomiální závislosti. Jednoduché závislosti (nulové invarianty, lineární kombinace, násobky jiných invariantů) umíme teoreticky jednoduše najít. Polynomiální závislosti jsou známé jen některé a jejich hledání je náročné.

Cílem práce bylo prozkoumat na reálných datech, které momentové invarianty jsou vybírány standardními metodami na výběr příznaků ([2]) a také jak se tyto metody chovají vůči závislým invariantům. Dalším předmětem zkoumání bylo zjistit, jestli jsou vybírány invarianty vyšších řádů, které teoreticky mohou obsahovat více informace, ale kvůli numerické chybě jsou méně robustní, nebo jestli jsou úspěšnější invarianty nižších řádů, které ovšem často nesou podobnou informaci. Dá se předpokládat, že výběr vhodných invariantů závisí na konkrétní úloze a konkrétních datech. Chtěli jsme ale zjistit, jestli se i přesto dají vypozorovat nějaké obecnější zákonitosti pro předvýběr příznaků před spuštěním samotného výběrového algoritmu, které by vylepšily výsledek. Teoreticky je například žádoucí používat nezávislé příznaky, protože závislé příznaky nepřinášejí další užitečnou informaci a mohou úlohu komplikovat nebo dokonce snižovat kvalitu klasifikace.

V první části popisujeme afinní momentové invarianty a způsob jejich výpočtu. Ve druhé části se věnujeme použitým metodám pro výběr příznaků a v poslední části ukazujeme výsledky experimentů na reálných datech.

1 Afinní momentové invarianty

Afinní momentové invarianty jsou speciální polynomiální kombinace momentů, které mají tu vlastnost, že jsou invariantní vůči afinní transformaci prostorových souřadnic. Díky vlastnostem afinní transformace se nevyužívají pouze v úlohách, kde se přímo vyskytuje afinní deformace, ale často také jako náhrada invariantů vůči projektivní transformaci.

1.1 Afinní transformace

Afinní transformací myslíme jakoukoliv lineární transformaci prostorových souřadnic obrázku. **Afinní transformace** se dá vyjádřit následovně

$$\begin{aligned} x &= a_0 + a_1 x + a_2 y \\ y &= b_0 + b_1 x + b_2 y \end{aligned}$$

Jakobiánem afinní transformace je $J = a_1 b_2 - a_2 b_1$.

Afinní transformací je například posunutí, rotace, škálování nebo zrcadlení. Afinní transformace je významná v rozpoznávání obrazu zejména proto, že při zobrazování

scény ze vzdálenosti, která je velká v porovnání s velikostí objektů ve scéně, je afinní transformace dobrým přiblížením projektivní transformace. Projektivní (perspektivní) transformace je přesným modelem pro zobrazení rovinné scény, ovšem její nevýhodou je, že je nelineární a konstrukce invariantů vůči této transformaci je velmi složitá.

1.2 Afinní momentové invarianty

Konkrétní tvar afinních momentových invariantů se dá odvodit několika různými způsoby (teorií algebraických invariantů, teorií grafů, tenzorovou algebrou nebo řešením vhodných parciálních diferenciálních rovnic). Přehled těchto způsobů je popsán v [1].

Zde naznačíme jedno z možných odvození. Nejprve zaveďme označení pro **geometrický moment** m_{ik} a **centrální moment** μ_{ik}

$$\begin{aligned} m_{ik} &= \int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty} x^i y^k f(x,y)\,\mathrm{d}x\mathrm{d}y, \\ \mu_{ik} &= \int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty} (x - m_{10}/m_{00})^i (y - m_{01}/m_{00})^k f(x,y)\,\mathrm{d}x\mathrm{d}y. \end{aligned}$$

Dále nechť f je obrázek s dvěma body(x_1, y_1) a (x_2, y_2). Následující výraz označme jako **křížový produkt**

$$C_{12} = x_1 y_2 - x_2 y_1.$$

Pro počet bodů $r \geqq 2$ a sadu přirozených čísel n_{kj} budeme definovat výraz

$$I(f) = \int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty} \prod_{k,j=1}^{r} C_{kj}^{n_{kj}} \prod_{i=1}^{r} f(x_i, y_i)\,\mathrm{d}x_i\mathrm{d}y_i.$$

Afinní transformaci $I(f)$ lze vyjádřit jako

$$I(f)' = J^w |J|^r I(f),$$

kde $w = \sum_{k,j} n_{kj}$ se navývá **váha invariantu** a r se nazývá **stupeň invariantu**. Pokud normalizujeme $I(f)$ výrazem μ_{00}^{w+r}, získáme hledaný invariant vůči afinní transformaci (s nulovým posunem). Platí tedy

$$\left(\frac{I(f)}{\mu_{00}^{w+r}}\right)' = \left(\frac{I(f)}{\mu_{00}^{w+r}}\right).$$

V případě záporného J a liché váhy w je třeba do rovnosti přidat faktor -1.

Nejjednodušší invariant lze získat dosazením $r = 2$ a $n_{12} = 2$. Potom dostaneme

$$I(f) = \int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty} (x_1 y_2 - x_2 y_1)^2 f(x_1, y_1) f(x_2, y_2)\,\mathrm{d}x_1\mathrm{d}y_1\mathrm{d}x_2\mathrm{d}y_2 = 2\left(m_{20}m_{02} - m_{11}^2\right). \tag{1}$$

Nahrazením geometrických momentů centrálními momenty a normalizováním získáváme invariant vůči obecné afinní transformaci

$$I_1 = \left(\mu_{20}\mu_{02} - \mu_{11}^2\right)/\mu_{00}^4.$$

Maximální řád momentů vystupujících v invariantu se nazývá **řád invariantu** a platí, že je vždy menší nebo roven váze invariantu.

1.3 Nezávislost afinních momentových invariantů

Pokud chceme afinní invarianty používat jako příznaky v rozpoznávacích úlohách, je žádoucí používat nezávislé příznaky, protože závislé příznaky nepřinášejí další informaci navíc a úlohu komplikují nebo mohou i snižovat kvalitu rozpoznání. Existují následující druhy závislostí mezi afinními invarianty:

1. *Nulové invarianty.* Některé invarianty mohou být identicky rovné nule pro všechny obrázky
2. *Identické invarianty.*
3. *Násobky.* Některé invarianty mohou být násobky několika invariantů
4. *Lineární kombinace.* Některé invarianty mohou být lineární kombinací jiných invariantů
5. *Polynomiální závislost.* Invarianty jsou polynomiálně závislé pokud existuje konečný součet násobků invariantů, který je roven nule.

Invarianty, které mají některou závislost typu 1. - 4. se nazývají **reducibilní**. Pro ilustraci z celkových 2 533 942 752 invariantů s váhou menší nebo rovnou 12 (vygenerované na základě grafů) je 2 532 349 nulových, ze zbývajících je 1 575 126 rovných jinému invariantu, 2 105 je násobek jiných invariantů a 14 538 je lineární kombinací jiných invariantů. Zbývá tedy pouze 1 589 ireducibilních invariantů.

Nalezení ireducibilních invariantů je teoreticky poměrně jednoduché (i když výpočetně už poměrně náročné). Problematické je ovšem nalezení polynomiálních závislostí. Je známo, že z 1 589 ireducibilních invariantů do váhy 12 je možné, aby pouze 85 bylo nezávislých, z čehož plyne, že 1 504 invariantů je polynomiálně závislých. Nalezení všech polynomiálních závislostí je ovšem i pro nízké váhy mimo možnosti současných počítačů.

2 Metody pro výběr příznaků

Výběr vhodné podmnožiny příznaků je důležitá součást většiny úloh při rozpoznávání. Výběrem příznaků chceme dosáhnout maximální rozlišitelnosti různých tříd. Máme D příznaků a hledáme d ($d << D$) příznaků tak, abychom maximalizovali vhodné kritérium. Proces výběru příznaků se skládá z volby metody výběru, volby vhodné kriteriální funkce a určení počtu příznaků, které se mají vybrat.

Metody výběru se dělí na optimální a suboptimální. Optimální metody (úplné prohledávání nebo algoritmus větví a mezí) jsou velmi pomalé a vhodné pouze pro úlohy s nízkou dimenzí. My jsme používali suboptimální metody, které jsou kompromisem mezi

rychlostí vyhledávání a optimalitou řešení. Obě metody, které jsme použili, jsou založeny na sekvenčním vyhledávání, při kterém se přidávají nebo odebírají příznaky do nebo ze stávající množiny. Dopředný krok (přidání jednoho příznaku do stávající množiny) probíhá tak, že se z příznaků vybere ten, který má dohromady s danou množinou příznaků nejvyšší hodnotu kriteriální funkce. Zpětný krok probíhá tak, že je odebrán ten příznak z dané množiny, pro který má výsledná množina(po odebrání příznaku) nejvyšší hodnotu kriteriální funkce.

Sekvenční dopředný výběr (SFS) je jednodušší, ale často používaná metoda. Hledáme množinu příznaků o předem danné velikosti d. Algoritmus začíná od prázdné množiny příznaků. Opakují se dopředné kroky tak dlouho, dokud není dosaženo požadované velikosti množiny příznaků. Tato metoda zohledňuje závislosti mezi příznaky, ale její nevýhodou je, že může uváznout v lokálním maximu, protože nejde příznaky, které byly přidány, odebírat. Vylepšení této metody, které jsme také používali, je sekvenční plovoucí vyhledávání.

V případě **sekvenčního plovoucího vyhledávání (SFFS)** algoritmus opět začíná od prázdné množiny příznaků. Po každém dopředném kroku následují zpětné kroky tak dlouho, dokud jsou výsledné podmnožiny lepší, než ty, které byly předtím vyhodnoceny jako nejlepší na dané velikosti množiny. Hledáme – li množinu příznaků o velikosti d, pak algoritmus skončí po dosažení množiny o velikosti $d + n$ (n volitelné). Tato metoda je díky střídání sekvencí dopředných a zpětných kroků schopná nalézat dostatečně dobrá řešení a zároveň rychlost je postačující pro většinu praktických problémů.

V naší práci jsme využívali kriteriální funkci typu **wrapper**, která je vždy spojena s konkrétním klasifikátorem. Množinu obrázků je nutné rozdělit na trénovací množinu, která je použita k nastavení klasifikátoru, a testovací množinu. Hodnotou kriteriální funkce je potom úspěšnost klasifikace daným klasifikátorem na testovací množině. Pro spolehlivější výsledky je možné použít několik variant trénovacích a testovacích množin a jako hodnotu kriteriální funkce použít průměrnou úspěšnost přes těchto několik variant. V našem případě jsme použili metodu m-fold cross-validation (jako hodnotu m jsme použili 3), při které je množina náhodně rozdělena na m částí a v každém kroku je jedna z nich zvolena jako testovací (pro výpočet úspěšnosti) a zbývající jsou použity jako trénovací.

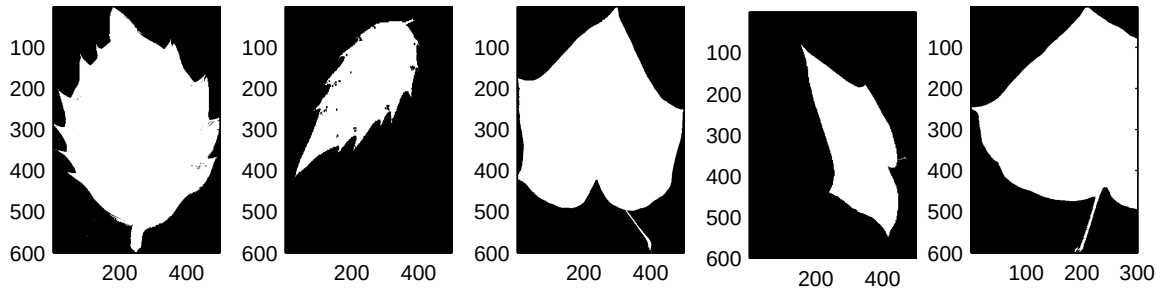
Jako klasifikátor jsme využívali metodu k **nejbližších sousedů (k-nn)**. Je to klasifikátor, který funguje na základě vzdáleností (v našem případě euklidovská) bodů v příznakovém prostoru. Algoritmus hledá nejbližší body daného obrázku v příznakovém prostoru a přiřadí ho do té třídy, která nejdříve obsahuje k bodů.

Kriteriální funkci typu wrapper jsme zvolili hlavně z toho důvodu, že závislosti mezi invarianty jsou polynomiální a dalo se očekávat, že metody fungující na základě korelace příznaků nebo předpokládající normální rozdělení příznaků tyto závislosti nedokáží zachytit.

3 Experimenty na reálných datech

3.1 Testovací obrázky

Obrázky, které jsme při testování používali, pochází z databáze listů stromů a keřů LEAF ([4]). Databáze vznikla naskenováním a binarizováním 800 skutečných listů od 90 druhů



Obrázek 1: Ukázka použitých obrázků. První a třetí obrázek jsou z původní databáze, druhý a čtvrtý jsou ukázky dodatečně vytvořených obrázků pomocí projektivní transformace. Poslední obrázek je ukázkou useknutého obrázku použitém v případě 3.

stromů a keřů. Počty vzorků jednoho daného stromu (tj. jedné třídy) jsou různé. Pro naši práci jsme zvolili těch 12 tříd, které obsahují alespoň 15 vzorků listů. I takovýto počet je pro účely rozpoznávání nedostatečný a tak jsme k existujícím reálným vzorkům dotvořili dodatečné umělé vzorky pozměněním původních obrázků. Z každého listu jsme vytvořili 4 transformované listy za pomoci projektivní transformace s náhodnými parametry. Parametry projektivní transformace byly ovšem nastaveny tak, aby byla velice blízká afinní transformaci.

Následující obrázek ukazuje příklad listu z databáze a příklad dodatečných obrázků, které jsme vytvořili projektivní transformací původního obrázku.

3.2 Výpočet afinních momentových invariantů

Pro popsanou sadu obrázků (původní i dodatečné) jsme vypočetli hodnoty všech 66 ireducibilních afinních momentových invariantů do řádu 4. Váhy všech invariantů jsou 2 až 19. Seznam těchto invariantů a kódy v MATLABU pro jejich výpočet jsme získali z přílohy knihy [1]. Byly odvozeny grafovou metodou popsanou v [1]. Metoda vyloučení reducibilních invariantů je také popsána v [1]. Tento soubor příznaků není nezávislý, protože v něm existují polynomiální závislosti mezi jednotlivými invarianty.

3.3 Výběr příznaků

Pro klasifikaci jsme použili C++ knihovnu FST3 vyvinutou v ÚTIA (popsána v [3]), která obsahuje implementaci nejpoužívanějších kriteriálních funkcí (k-nn wrapper, Mahalanobis, Bhattacharyya) a metod pro výběr příznaků (BIF, SFS, SFFS). V prvním kroku klasifikace jsme množinu všech obrázků rozdělili na dvě části. Jednu pro výběr příznaků (80% obrázků) a druhou pro závěrečné nezávislé otestování výsledného klasifikátoru (20% obrázků).

3.4 Výsledky

3.4.1 1. příklad

V prvním případě jsme pracovali s prvními 10 invarianty $I_1 \ldots . I_{10}$. V této množině invariantů je známá nezávislá podmnožina $(I_1, I_2, I_3, I_4, I_6, I_7, I_8, I_9)$. Velikost vybírané podmnožiny jsme zvolili 10 (tedy všechny) a úspěšnost konkrétního invariantu jsme určovali na základě pořadí ve výběru. Výběr příznaků jsme dělali na dvojicích tříd listů. Ze všech 12 tříd, které jsou popsány výše, jsme vytvořili všech 66 možných dvojic a na každé z nich jsme provedli výběr příznaků.

K výběru jsme použili metody SFS a SFFS. Všimli jsme si, že v některých případech při metodě SFS došlo k uvíznutí v lokálním maximu. Ovšem při porovnání obou metod na celém zkoumaném souboru dvojic se ukázalo, že obě metody dávají velmi podobné celkové výsledky.

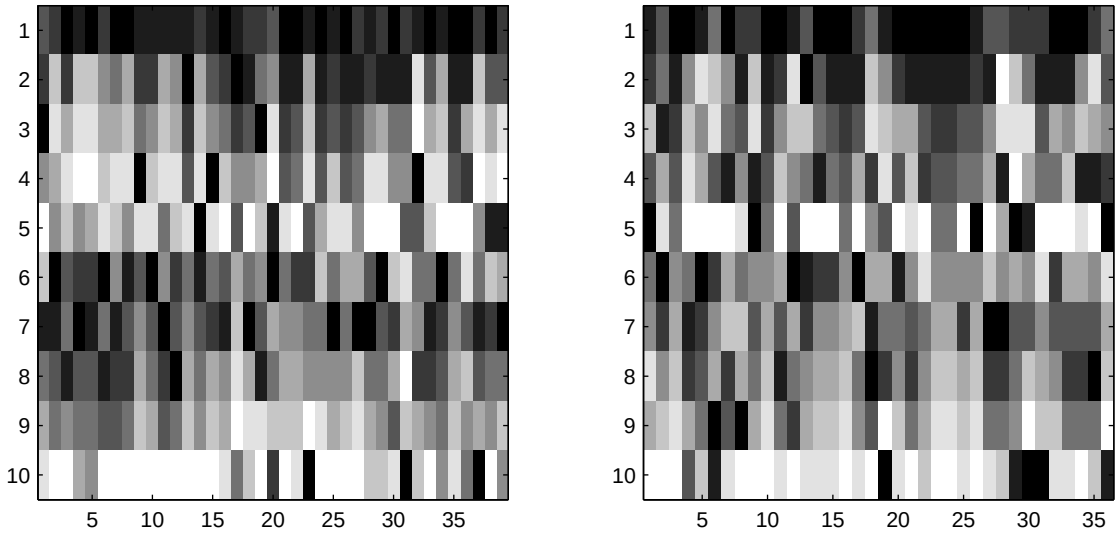
Výsledky výběru jsou zobrazeny na obrázku 2 vlevo. Řádky odpovídají invariantům (první řádek reprezentuje invariant I_1, atd.), sloupce jednotlivým dvojicím tříd. V každém řádku je barevně znázorněno pořadí, v kterém byl daný invariant vybrán při klasifikaci na dané dvojici tříd. Černá/bílá barva znamená, že daný příznak byl vybrán jako první/poslední. Na obrázku je vidět velká úspěšnost příznaku I_1, částečně také I_2, I_6 a I_7. Naopak I_5 a I_{10} jsou vybírány často jako poslední.

Jeden z možných důvodů je jejich závislost na ostatním příznacích ve skupině (první desítka invariantů bez příznaků I_5 a I_{10} jsou největší nezávislá podmnožina první desítky invariantů). Dalším důvodem může být lichá váha invariantů I_5 a I_{10}. Obrázky listů můžou být symetrické a na symetrických obrázcích jsou invarianty liché váhy rovny nule, čímž se ztrácí jejich diskriminační síla. K vyloučení tohoto důvodu jsme zkusili snížit symetričnost klasifikovaných obrázků useknutím části listu (příklad na obr. 1 vpravo). Výsledek této klasifikace je na obrázku 2 vpravo. Úspěšnost invariantů I_5 a I_{10} se na nesymetrických obrázcích ale nezlepšila. Z toho vyplývá, že lichá váha těchto příznaků není zřejmě důvodem jejich špatné diskriminační síly.

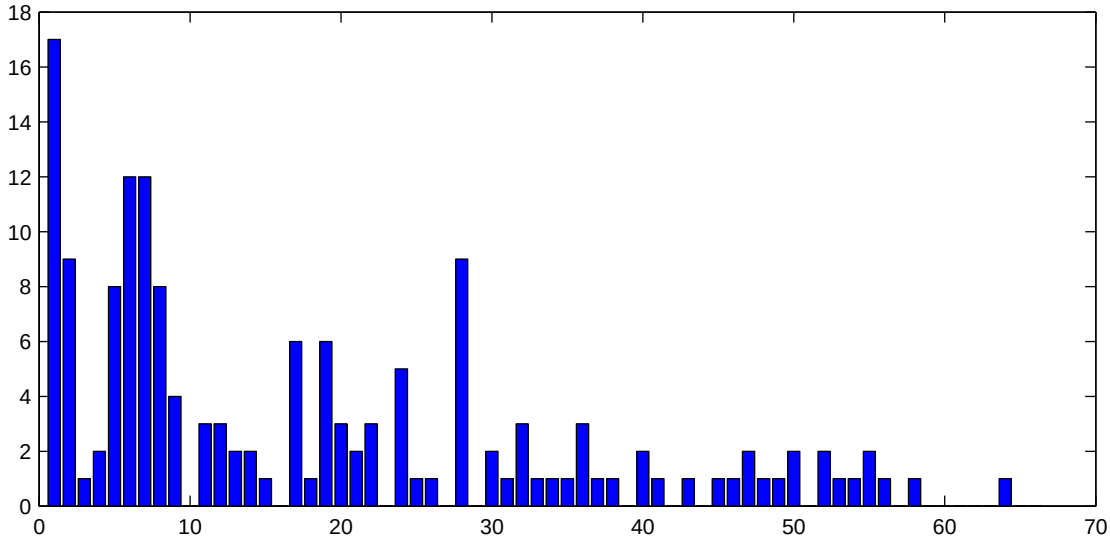
3.4.2 2.příklad

V dalším případě jsme testovali všech 66 ireducibilních invariantů do řádu 4. Závislosti mezi těmito invarianty nejsou teoreticky úplně prozkoumány. Přitom jen 9 z těchto 66 invariantů může být nezávislých. Velikost vybírané podmnožiny jsme zvolili 3. Výběr jsme jako v předchozím případě prováděli na každé z 66 vytvořených dvojic tříd.

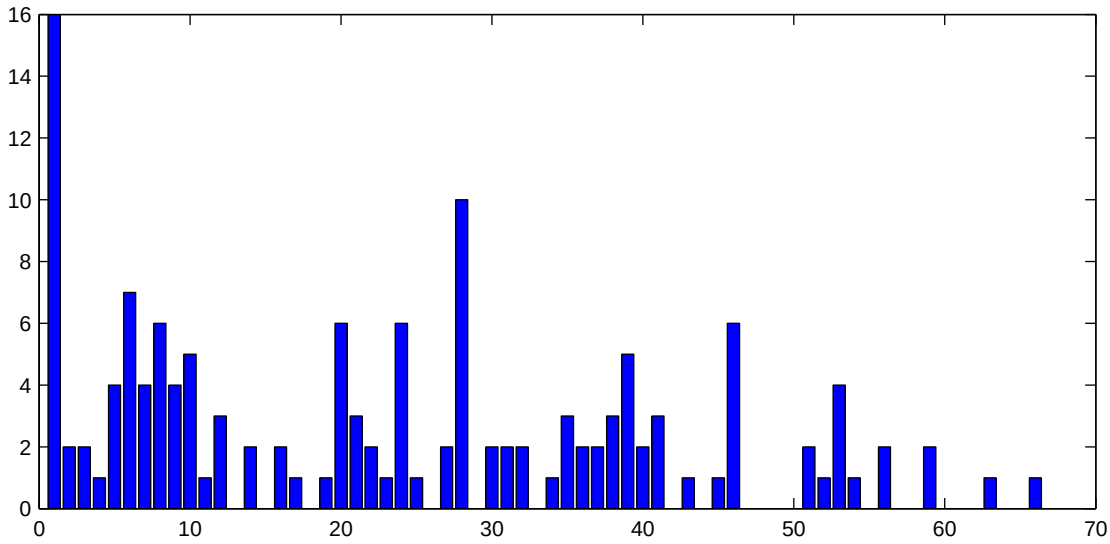
Výsledek metody SFFS pro obrázky celých listů je na obrázku 3. Výsledek metody SFFS pro useknuté listy je na obr. 4. V obou případech jsou nejvíce vybírány počáteční invarianty. Z invariantů vyšších řádů je výrazně úspěšný I_{28}. Možným důvodem je nízká váha invariantu I_{28} ve srovnání s okolními invarianty. V obou případech jsou upřednostňovány invarianty nižších řádů nebo s nízkou váhou. Důvodem by mohl být nižší řád momentů a menší počet sčítanců v těchto invariantech a tudíž jejich vetší robustnost a menší náchylnost k numerickým chybám (numerická chyba je menší v porovnání s informací obsaženou v příznaku).



Obrázek 2: Výsledky 1. případu (výběr na invariantech $I_1, ..., I_{10}$). Obrázek vlevo je výsledkem na první sadě obrázků (bez useknutí). Obrázek vpravo je výsledkem na useknutých obrázcích.



Obrázek 3: Výsledek 2. případu (výber 3 příznaků z 66 invariantů metodou SFFS) na původních (neuseknutých obrázcích). Sloupec odpovídá indexu invariantu a výška sloupce počtu výbrání.



Obrázek 4: Výsledek 2. případu (výber 3 příznaků z 66 invariantů metodou SFFS) na useknutých obrázcích. Sloupec odpovídá indexu invariantu a výška sloupce počtu výbrání.

Závěr

V naší práci jsme použili afinní momentové invarianty ke klasifikaci reálných obrázků (databáze listů doplněná dodatečnými transformovanými obrázky). K výběru příznaků jsme použili několik standardních metod a snažili jsme se na základě výsledků nalézt nějaké zákonitosti nebo rady pro výběr příznaků.

Zjišťovali jsme, jak si metody na výběr příznaků poradí se skutečností, že mezi 66 invarianty s kterými jsme pracovali, je možné vybrat maximálně devítičlennou nezávislou množinu, přičemž závislosti mezi invarianty jsou i nelineární. V prvním případě (výběr z 10 příznaků) bylo jasně vidět, že příznaky I_5 a I_{10}, o kterých víme, že jsou součástí polynomiální závislosti uvnitř první desítky příznaků, jsou při výběru velice neúspěšné. Je ale možné, že to může spíše souviset s nějakou jejich samostatnou vlastností (oproti nezávislosti, což je vlastnost skupiny příznaků), jelikož úspěšnost klasifikace na základě těchto příznaků při použití pouze jednoho příznaku byla výrazně horší než v případě jiných příznaků v první desítce invariantů.

Z teorie je zřejmé, že je výhodné klasifikaci dělat jen na nezávislých příznacích, ale díky existenci polynomiálních závislostí je nezávislá množina známá jen v první desítce invariantů (8 příznaků je nezávislých). Na datech, s kterými jsme pracovali, bylo nejčastěji maximální úspěšnosti dosaženo za pomocí dvou nebo tří příznaků a přidávání dalších příznaků již úspěšnost nezlepšovalo. Z tohoto důvodu v našem případě nehráli závislosti tak velkou roli (existuje mnoho nezávislých množin invariantů o velikosti dva nebo tři)

a proto byly v některých případech upřednostňovány příznaky i s vysokým pořadovým číslem, které se hodily na konkrétní dvojici tříd. Pro detailnější prozkoumání chování výběrových metod vůči závislostem mezi invarianty by bylo třeba použít data, pro jejichž klasifikaci je zapotřebí více příznaků.

Škála úspěšných příznaků je poměrně široká, ale není rovnoměrná, z čehož plyne, že konkrétní úspěšné příznaky závisejí na konkrétní dvojici tříd a není možné předem za pomoci teorie odhadnout, které příznaky budou pro daná data úspěšnější. V přítomnosti složitých závislostí mezi invarianty ovšem nelze plně spoléhat na spolehlivost výběru pomocí metod na výběr příznaků (zejména v případě metod pracujících na základě korelací nebo normálních rozdělení) a je nutné využívat teoreticky známých nezávislých příznaků společně s příznaky vybranými výběrovými metodami.

Zajímavým zjištěním práce bylo, že pro klasifikaci jsou upřednostňovány příznaky s nízkým pořadovým číslem, zejména z první desítky. Našlo se ale i velké množství úspěšných invariantů s vyšším pořadovým číslem. Většinou je spojovalo to, že mají v porovnání s okolními příznaky nižší váhu nebo řád. Důvodem je zřejmě to, že takovéto příznaky jsou díky nižším řádům momentů a menšímu počtu členů odolnější vůči numerickým chybám (z důvodu vzorkování) a míra informace v nich je oproti velikosti numerické chyby významnější.

Literatura

[1] J. Flusser, T. Suk, B. Zitová. Moments and Moment Invariants in Pattern Recognition. Wiley (Chichester, 2009)

[2] P. Somol, J. Novovičová, P. Pudil. *Moderní metody výběru příznaků ve statistickém rozpoznávání.* pre-print

[3] P. Somol, P. Vácha, S. Mikeš, J. Hora, P. Pudil, P. Žid, *Introduction of Feature Selection Toolbox 3 - The C++ Library for Subset Search, Data Modelling and Classification.* Tech. Report No. 2287, UTIA, (2010)

[4] Tree Leaf Database, Institute of Information Theory and Automation ASCR, Prague, Czech Republic, `http://zoi.utia.cas.cz/tree_leaves`

Different Measures of Reliability in Regression

Radim Demut

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

advisor: Martin Holeňa, Institute of Computer Science, AS CR

Abstract. This paper is concerned with the reliability of individual predictions in regression. It is well known that predictions by regression models have for different inputs different reliability, and if predictions by one regression model have a higher reliability than those by another in some part of the input space, they can nevertheless be less reliable in another part. Therefore, the reliability of individual predictions is a very important field of study. We describe conformal predictors and some methods for estimating the reliability of individual predictions such as sensitivity analysis or local modeling of prediction error. Finally, we carry out a simulation to compare the methods in an experiment.

Keywords: conformal predictors, sensitivity analysis, regression

Abstrakt. Tento článek se zabývá spolehlivostí pro jednotlivé odhady v regresi. Je známo, že predikce regresních modelů mají pro různé vstupy různou spolehlivost. Predikce jednoho modelu může být spolehlivější pro nějakou množinu vstupů než predikce jiného modelu, ale spolehlivost pro jinou množinu vstupů může být pro tento model nižší. V článku popíšeme konformní predikci a další metody odhadů spolehlivosti v regresi, jako například analýzu citlivosti nebo lokální model chyby odhadu. Nakonec provedeme simulaci, abychom jednotlivé metody srovnali i experimentálně.

Klíčová slova: konformní predikce, analýza citlivosti, regrese

Antimorphisms Generating $(-\beta)$-integers*

Daniel Dombek

2nd 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. This contribution is devoted to the study of positional numeration systems with negative base introduced by Ito and Sadahiro in 2009, called $(-\beta)$-expansions. We give an admissibility criterion for a more general case of $(-\beta)$-expansions and define the set of $(-\beta)$-integers, denoted by $\mathbb{Z}_{-\beta}$. We give a description of distances within $\mathbb{Z}_{-\beta}$ and show that this set can be coded by a biinfinite word over an infinite alphabet, which is a fixed point of an antimorphism.

Full version of this contribution, *Substitutions over infinite alphabet generating $(-\beta)$-integers*, was published in Electronic Proceedings in Theoretical Computer Science [2], and the proofs of given theorems can be found in paper [3] to appear in Theoretical Computer Science, and in preprint [1].

Keywords: numeration system, infinite word, antimorphism

Abstrakt. Tento příspěvek zkoumá poziční numerační systémy, které v roce 2009 definovali Ito a Sadahiro, tzv. $(-\beta)$-rozvoje. Formulujeme podmínku pro přípustnost řetězců cifer pro obecnější případ $(-\beta)$-rozvojů a definujeme množinu $(-\beta)$-celých čísel, značenou $\mathbb{Z}_{-\beta}$. Dále ukážeme jak spočítat mezery mezi po sobě jdoucími $(-\beta)$-celými čísly a jak kódovat množinu $\mathbb{Z}_{-\beta}$ pomocí oboustranně nekonečného slova nad nekonečnou abecedou, které je pevným bodem jistého antimorfismu.

Nezkrácená verze tohoto příspěvku, *Substitutions over infinite alphabet generating $(-\beta)$-integers*, byla publikována v Electronic Proceedings in Theoretical Computer Science [2] a důkazy uvedených tvrzení lze najít v článku [3] přijatém k publikaci v Theoretical Computer Science a v preprintu [1].

Klíčová slova: numerační systém, nekonečné slovo, antimorfismus

1 b-expansions

In 1957, Rényi introduced positional numeration system with positive real base $\beta > 1$, called β-expansions (see [8]). We can reformulate his definition in a more general way, such that it also covers an analogous numeration system with negative base, $(-\beta)$-expansions, introduced by Ito and Sadahiro in 2009 (see [6]).

*This work was supported by the Czech Science Foundation, grant GAČR 201/09/0584, by the grants MSM6840770039 and LC06002 of the Ministry of Education, Youth, and Sports of the Czech Republic, and by the grant of the Grant Agency of the Czech Technical University in Prague, grant No. SGS11/162/OHK4/3T/14.

Definition 1. *Let $b \in \mathbb{R}, |b| > 1$ be a base and consider $x \in [l, l+1)$, where $l \in \mathbb{R}$ is arbitrary fixed. We define the b-expansion of x as the digit string $d(x) = x_1x_2x_3\cdots$, with digits x_i given by*

$$x_i = \lfloor bT^{i-1}(x) - l \rfloor , \tag{1}$$

where $T(x)$ stands for the b-transformation

$$T : [l, l+1) \to [l, l+1) , \quad T(x) = bx - \lfloor bx - l \rfloor . \tag{2}$$

It holds that

$$x = \frac{x_1}{b} + \frac{x_2}{b^2} + \frac{x_3}{b^3} + \cdots$$

and we use the notation $d(x) = x_1x_2x_3\cdots$. Let us remark that the case $b = \beta > 1, l = 0$ coincides with Rényi definition of β-expansions and the case $b = -\beta < -1, l = -\frac{\beta}{\beta+1}$ corresponds to Ito and Sadahiro $(-\beta)$-expansions.

In the following we consider the negative base case $b = -\beta < -1$ with the condition $l \in (-1, 0]$. This choice of l guarantees the existence of $(-\beta)$-expansions of all real numbers. The set of digits used in $(-\beta)$-expansions depends on the choice of both β and l and can be calculated directly from (1) as

$$\mathcal{A}_{-\beta,l} = \left\{ \lfloor -l(\beta+1) - \beta \rfloor, \ldots, \lfloor -l(\beta+1) \rfloor \right\} . \tag{3}$$

Let $x \in \mathbb{R}$. Thanks to the fact that $0 \in [l, l+1)$, there exists an integer k, such that $\frac{x}{(-\beta)^k} \in [l, l+1)$ and $d\big(\frac{x}{(-\beta)^k}\big) = x_kx_{k-1}x_{k-2}\cdots$. The $(-\beta)$-expansion of x is then defined as

$$\langle x \rangle_{-\beta} = x_kx_{k-1}\cdots x_1x_0 \bullet x_{-1}x_{-2}\cdots .$$

Note that the similar procedure works also for all cases $b = \beta > 1$ with $l \in (-1, 0]$ and it gives unique expansions of all reals. However, the uniqueness of $\langle x \rangle_{-\beta}$ in the negative base case will still have to be discussed.

2 $(-\beta)$-admissibility

The so-called alternate order was used in the admissibility condition by Ito and Sadahiro and it will work also in the general case. Let us recall the definition. For the strings

$$u, v \in (\mathcal{A}_{-\beta,l})^{\mathbb{N}} , \quad u = u_1u_2u_3\cdots \quad \text{and} \quad v = v_1v_2v_3\cdots$$

we say that $u \prec_{alt} v$ (u is less than v in the alternate order) if $u_m(-1)^m < v_m(-1)^m$, where $m = \min\{k \in \mathbb{N} \mid u_k \neq v_k\}$. Note that standard ordering between reals in $[l, l+1)$ corresponds to the alternate order on their respective $(-\beta)$-expansions.

Definition 2. *An infinite string $x_1x_2x_3\cdots$ of integers is called $(-\beta)$-admissible (or just admissible), if there exists an $x \in [l, l+1)$ such that $x_1x_2x_3\cdots$ is its $(-\beta)$-expansion, i.e. $x_1x_2x_3\cdots = d(x)$.*

We give the criterion for $(-\beta)$-admissibility (proven in [3]) in a form similar to both Parry lexicographic condition (see [7]) and Ito-Sadahiro admissibility criterion (see [6]).

Theorem 3. ([3]) *An infinite string $x_1x_2x_3\cdots$ of integers is $(-\beta)$-admissible, if and only if*

$$l_1l_2l_3\cdots \preceq_{alt} x_ix_{i+1}x_{i+2}\cdots \prec_{alt} r_1r_2r_3\cdots, \qquad \text{for all } i\geq 1, \tag{4}$$

where $l_1l_2l_3\cdots = d(l)$ and $r_1r_2r_3\cdots = d^*(l+1) = \lim_{\epsilon\to 0+} d(l+1-\epsilon)$.

3 $(-\beta)$-integers

In the following, $(-\beta)$-admissibility will be used to define the set of $(-\beta)$-integers. However a further discussion concerning the uniqueness of $(-\beta)$-expansions is needed. In the following example we show, that non-invariance of the interval $[l,l+1)$ under division by $-\beta$ can cause problems.

Example 4. *Let β be the greater root of the polynomial x^2-2x-1, i.e. $\beta = 1+\sqrt{2}$, and let $[l,l+1) = \big[-\frac{\beta^9}{\beta^9+1}, \frac{1}{\beta^9+1}\big)$. Note that $\frac{1}{-\beta}[l,l+1)\notin[l,l+1)$.*

If we want to find the $(-\beta)$-expansion of number $x\notin[l,l+1)$, we have to find such $k\in\mathbb{N}$ that $\frac{x}{(-\beta)^k}\in[l,l+1)$, and then use $d\big(\frac{x}{(-\beta)^k}\big)$ to get the expansion of x. The problem is that, in general, different choices of the exponent k may give different $(-\beta)$-admissible digit strings which all represent the same number x.

Let us find possible $(-\beta)$-expansions of 1. It can be shown that by various choices of k we obtain five possible expansions:

$$1\bullet 0^\omega \quad = \quad 120\bullet 0^\omega \quad = \quad 13210\bullet 0^\omega \quad = \quad 1322210\bullet 0^\omega \quad = \quad 132222210\bullet 0^\omega\,.$$

However, only one of these digit strings possesses the property of being admissible even if we add the prefix 0 to its left end. For all $x\in\mathbb{R}$, $\beta>1$ and $l\in(-1,0]$, this unique digit string exists and we use it to define the unique $\langle x\rangle_{-\beta}$ for all real x.

Definition 5. *Let $\beta>1$, $l\in(-1,0]$. The set of $(-\beta)$-integers is defined as*

$$\mathbb{Z}_{-\beta} = \Big\{x=\sum_{i=0}^{k-1} a_i(-\beta)^i \,\Big|\, 0a_{k-1}a_{k-2}\cdots a_1a_00^\omega \text{ is admissible}\Big\}\,.$$

A phenomenon unseen in Rényi numeration arises, there are cases when the set of $(-\beta)$-integers is trivial, i.e. when $\mathbb{Z}_{-\beta}=\{0\}$. This happens if and only if both numbers $\frac{1}{\beta}$ and $-\frac{1}{\beta}$ are outside of the interval $[l,l+1)$. This can be reformulated as

$$\mathbb{Z}_{-\beta}=\{0\} \quad\Leftrightarrow\quad \beta<-\frac{1}{l} \text{ and } \beta\leq\frac{1}{l+1}\,.$$

Let us define a "value function" γ. Consider a finite digit string $x_{k-1}\cdots x_1x_0$, then $\gamma(x_{k-1},\cdots x_1x_0) = \sum_{i=0}^{k-1} x_i(-\beta)^i$. In order to describe distances between adjacent $(-\beta)$-integers, we will study ordering of finite digit strings in the alternate order. Denote by $\mathcal{S}(k)$ the set of infinite $(-\beta)$-admissible digit strings such that erasing a prefix of length k yields 0^ω, i.e. for $k\geq 0$, we have

$$\mathcal{S}(k) = \{a_{k-1}a_{k-2}\cdots a_00^\omega \mid a_{k-1}a_{k-2}\cdots a_00^\omega \text{ is } (-\beta)\text{-admissible}\}\,,$$

in particular $\mathcal{S}(0) = \{0^\omega\}$. For a fixed k, the set$\mathcal{S}(k)$ is finite. Denote by $\mathrm{Max}(k)$ the maximal element in $\mathcal{S}(k)$ with respect to the alternate order and by $\max(k)$ its prefix of length k, i.e. $\mathrm{Max}(k) = \max(k)0^\omega$. Similarly, we define $\mathrm{Min}(k)$ and $\min(k)$. Thus,

$$\mathrm{Min}(k) \preceq_{alt} r \preceq_{alt} \mathrm{Max}(k)\,, \qquad \text{for all digit strings } r \in \mathcal{S}(k).$$

Theorem 6. ([1]) *Let $x < y$ be two consecutive $(-\beta)$-integers. Then there exist a finite string w over the alphabet $\mathcal{A}_{-\beta,l}$, a non-negative integer $k \in \{0, 1, 2, \dots\}$ and a positive digit $d \in \mathcal{A}_{-\beta,l}\setminus\{0\}$ such that $w(d-1)\mathrm{Max}(k)$ and $wd\mathrm{Min}(k)$ are strongly $(-\beta)$-admissible strings and*

$$\begin{array}{lcll} x = \gamma(w(d-1)\max(k)) & < & y = \gamma(wd\min(k)) & \text{for } k \text{ even},\\ x = \gamma(wd\min(k)) & < & y = \gamma(w(d-1)\max(k)) & \text{for } k \text{ odd}. \end{array}$$

In particular, the distance $y-x$ between these $(-\beta)$-integers depends only on k and equals to

$$\Delta_k := \Big|(-\beta)^k + \gamma\big(\min(k)\big) - \gamma\big(\max(k)\big)\Big|\,. \tag{5}$$

4 Coding $\mathbb{Z}_{-\beta}$ by an infinite word

Let us now describe how we can code the set of $(-\beta)$-integers by an infinite word over the infinite alphabet $\mathbb{N}$.

Let $(z_n)_{n\in\mathbb{Z}}$ be a strictly increasing sequence satisfying

$$z_0 = 0 \quad \text{and} \quad \mathbb{Z}_{-\beta} = \{z_n \mid n \in \mathbb{Z}\}\,.$$

We define a bidirectional infinite word over an infinite alphabet $\mathbf{v}_{-\beta} \in \mathbb{N}^{\mathbb{Z}}$, which codes the set of $(-\beta)$-integers. According to Theorem 6, for any $n \in \mathbb{Z}$ there exist a unique $k \in \mathbb{N}$, a word w with prefix 0 and a letter d such that

$$z_{n+1} - z_n = \big|\gamma(w(d-1)\max(k)) - \gamma(wd\min(k))\big|\,.$$

We define the word $\mathbf{v}_{-\beta} = (v_i)_{i\in\mathbb{Z}}$ by $v_n = k$.

Theorem 7. ([1]) *Let $\mathbf{v}_{-\beta}$ be the word associated with $(-\beta)$-integers. There exists an antimorphism $\Phi : \mathbb{N}^* \to \mathbb{N}^*$ such that $\Psi = \Phi^2$ is a non-erasing non-identical morphism and $\Psi(\mathbf{v}_{-\beta}) = \mathbf{v}_{-\beta}$. Φ is always of the form*

$$\Phi(2l) = S_{2l}(2l+1)\widetilde{R_{2l}} \quad \text{and} \quad \Phi(2l+1) = R_{2l+1}(2l+2)\widetilde{S_{2l+1}}\,,$$

where $\widetilde{u}$ denotes the reversal of the word u and words R_j, S_j depend only on j and on $\min(k), \max(k)$ with $k \in \{j, j+1\}$.

As it turns out, in some cases (mostly when reference strings $l_1l_2l_3\cdots$ and $r_1r_2r_3\cdots$ are eventually periodic of a particular form) we can find a letter-to-letter projection to a finite alphabet $\Pi : \mathbb{N} \to \mathcal{B}$ with $\mathcal{B} \subset \mathbb{N}$, such that $\mathbf{u}_{-\beta} = \Pi\mathbf{v}_{-\beta}$ also encodes $\mathbb{Z}_{-\beta}$ and it is a fixed point of a an antimorphism $\varphi = \Pi \circ \Phi$ over the finite alphabet $\mathcal{B}$. Clearly, the square of φ is then a non-erasing morphism over $\mathcal{B}$ which fixes $\mathbf{u}_{-\beta}$.

Let us mention that $(-\beta)$-integers in the Ito-Sadahiro case $l = -\frac{\beta}{\beta+1}$ are also subject of [9]. For β with eventually periodic $d(l)$, Steiner finds a coding of $\mathbb{Z}_{-\beta}$ by a finite alphabet and shows, using only the properties of the $(-\beta)$-transformation, that the word is a fixed point of a non-trivial morphism. Our approach is of a combinatorial nature, follows a similar idea as in [1] and shows existence of an antimorphism for any base β.

References

[1] P. Ambrož, D. Dombek, Z. Masáková, E. Pelantová, *Numbers with integer expansion in the numeration system with negative base*, preprint (2011), 25pp. arXiv:0912.4597v3 [math.NT]

[2] D. Dombek, *Substitutions over infinite alphabet generating $(-\beta)$-integers*, Proceedings 8th International Conference Words 2011, EPTCS 63, 115–121 (2011).

[3] D. Dombek, Z. Masáková, E. Pelantová, *Number representation using generalized $(-\beta)$-transformation*, to appear in Theoret. Comput. Sci. (2011), 22pp. arXiv:1102.3079v1 [cs.DM]

[4] S. Fabre, *Substitutions et β-systèmes de numération*, Theoret. Comput. Sci. 137, 219–236 (1995).

[5] Ch. Frougny and A. C. Lai, *Negative bases and automata*, Discrete Mathematics and Theoretical Computer Science 13, 75–94 (2011).

[6] S. Ito and T. Sadahiro, *Beta-expansions with negative bases*, INTEGERS 9, 239–259 (2009).

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

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

[9] W. Steiner, *On the structure of $(-\beta)$-integers*, to appear in RAIRO Theor. Inf. Appl. (2011), 15pp. arXiv:1011.1755v1 [math.NT]

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

Experimental Signal Deconvolution in Acoustic Emission Identification Setup

Zuzana Farová

1st year of PGS, email: zuzana.farova@email.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. Identification of acoustic sources is a great deal and the most important problem emerging in biomedical applications and nondestructive testing. We present the basics of time reversal Acoustic Emission (AE) principle and point out the advantages of AE nondestructive testing for localization and classification of acoustic sources. This approach provides the tool for the defect localization by means of the reversal wave focusing in nonlinearity position in the material under consideration. Also, using time reversal (TR) acoustics to AE signal, we are able to perform so-called experimental deconvolution. That means we are able by means of TR acoustics to eliminate the influence of material properties and AE sensor characteristics in measured signal. Consequently, we obtain convolutional signal in the closest neighborhood of acoustic source, which gives us the pure image of the real characteristics of AE source after deconvolution process applied to the signals detected. We describe the mathematical background for backside deconvolution through the Green functions. Further, we derive basis for experimental deconvolution by means of time reversal acoustics and Fourier transform. Finally, we design the laboratory settings realizing experiments for AE signal deconvolution of the reversed signals measured by the AE sensor centering to the position of the defect. Our first experiment of experimental deconvolution was measured at the Institute of Thermomechanics, Academy of Science of the Czech Republic, on small aircraft component – a steering actuator bracket. This experiment confirmed our expectation.

The whole contribution will be published in **Proceedings of NDT in Progress 2011**, 10.–12.10.2011, Prague.

Keywords: TRA, Green function, experimental deconvolution, acoustic emission

Abstrakt. Určování zdroje akustické emise je velmi důležitou úlohou jak v biomedicinálních aplikacích tak i v nedestruktivním testování materiálů. V článku shrnujeme základy time reverzální akustiky (TRA) a výhody aplikace akustické emise při klasifikaci a lokalizaci defektů v materálech. Použití time reverzní akustiky pro hledání nelinearit v materiálu je účinným postupem pro lokalizaci defektů. Pomocí TRA jsme schopni též provést tzv. experimentální dekonvoluci, tedy jsme schopni pomocí TRA odstranit vliv materiálových vlastností na měřený signál. Tím dostáváme signál, jež je velmi podobný původnímu signálu ze zdroje akustické emise, a tudíž jsme schopni přesněji určit typ zdroje akustické emise. V práci popisujeme matematický základ pro dekonvoluci a vlastnosti Greenovy funkce. Dále jsme odvodili základy experimentální dekonvoluce pomocí TRA a Fourierovy transformace. Rovněž jsme navrhli experiment, kde jsme porovnali klasifikaci před a po experimentální dekonvoluci. Tento náš experiment byl proveden na Ústavu termomechaniky, Akademie věd ČR na malé součástce letadlového podvozku, která byla podrobena zátěžovému testu.

Celý článek byl uplikován v **Proceedings of NDT in Progress 2011**, 10.–12.10.2011, Prague.

Klíčová slova: TRA, Greenova funkce, experimentální dekonvoluce, akustická emise

References

[1] Aki K., Richards P. G., *Quantitative Seismology.* University Science Books, 2002.

[2] Chlada M., Prevorovsky Z., Blahacek M., Neural Network AE Source Location Apart From Structure Size and Material *Journal of Acoustic Emission*, **28**, 99–108, 2010.

[3] Fink M., Time-reversed Acoustics. *Rep. Prog. Phys.*, **63**, 1933-1995, 2000.

[4] Klibanov M.V., Timonov A., On the Mathematical Treatment of time reversal, Institute of physics publishing. *Inverse Problems*, **19**, 1299–1318, 2003.

Introduction to Total Least Trimmed Squares Estimation

Jiří Franc

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

advisor: Jan Ámos Víšek, Department of Macroeconomics and Econometrics, Faculty of Social Sciences, Institute of Economic Studies, Charles University in Prague

Abstract. In this paper we introduce the robustified version of total least squares, called total least trimmed squares. This method is the proper estimation in error-in-variables model, if outliers in datasets occur. We give different formulations of the estimator, obtain its breakdown point and show some properties. We summarize main information and recent developments in algorithms and computation. Small computational experiments on sets of benchmark instances show that the proposed algorithms performs well and gives reasonable estimation in sufferable computational time.

Keywords: robust regression analysis, error in variables mode, robustified total least squares, total least trimmed squares, mixed least trimmed squares - total least trimmed squares, breakdown point.

Abstrakt. V tomto článku představíme robustní verzi metody nejmenších totálních čtverců nazvanou totální nejmenší usekané čtverce. Tato metoda je vhodná zejména pro modely kde je jak závislá tak nezávislá proměnná měřena s náhodnou chybou a v datech se mohou vyskytovat odlehlá pozorování. Uvedeme různé formulace odhadu, spočteme jeho bod zlomu a ukážeme některé jeho vlastnosti. Dále shrneme nejnovější vývoj ve způsobu výpočtu totálních nejmenších usekaných čtverců a představíme vhodné algoritmy. Nakonec provedeme několik experimentů na vzorku testovacích dat s různými parametry, které nám ukáží, že odhad dává rozumná řešení v přípustném výpočetním čase.

Klíčová slova: robustí regresní analýza, metoda robustifikovaných totálních čtverců, metoda totálních nejmenších usekaných čtverců, bod zlomu, invariantnost odhadů

1 Introduction

The total least square (TLS) method is one of several linear parameter estimation method that has been proposed to solving a multivariate measurement error models. Let us consider the overdetermined set of linear equations

$$\mathbf{Y} \approx \mathbf{X}\beta^0,$$

where $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is vector of response (dependent) variable, $\mathbf{X} \in \mathbb{R}^{n\times p}$ matrix of predictors (independent variables), $\beta^0 \in \mathbb{R}^{p\times 1}$ unknown parameter vector and we have more

equations than unknowns, i.e. $n > p$. Our aim is to find such a linear model that explains the data set. The classical regression approach, such as ordinary least squares (OLS) technique, assumes that the matrix $\mathbf{X}$ (measurements of independent variables) is error free and hence, all errors are confined to the dependent variable. When this condition is broken, then not only OLS estimate of parameter β^0 is inconsistent. The assumption of error free measurements of response variable is frequently unrealistic, especially in econometrics or engineering applications were human, sampling or modeling errors occur. The TLS method is motivated by this asymmetry, where the dependent variable $\mathbf{Y}$ is corrected while the independent variables $\mathbf{X}$ not. The idea is to modify (correct) all data points in such a way that some norm of the correction is minimized subject to the constraint that that the corrected vectors satisfy a linear relation. Although the history of TLS technique is very long and the method is also known as orthogonal regression or errors-in-variables method, it became popular as lately as in last 20 years after the paper of Golub and Van Loan [2] and book of Van Huffel and Vandewalle [12]. The good paper that summarizes the recent development in TLS is [5]. According to these works we define the ordinary total least squares.

Given an overdetermined set of linear equations $\mathbf{Y} \approx \mathbf{X}\beta^0$. The total least squares problem seeks to

$$\hat{\beta}^{(TLS,n)} = \min_{\beta\in\mathbb{R}^p,[\varepsilon,\mathbf{\Theta}]\in\mathbb{R}^{n\times(p+1)}} \|[\varepsilon,\mathbf{\Theta}]\|_F \quad \text{subject to} \quad \mathbf{Y}+\varepsilon = (\mathbf{X}+\mathbf{\Theta})\beta. \tag{1}$$

$\hat{\beta}^{(TLS,n)}$ is called a TLS solution to the problem (1) and $[\varepsilon,\mathbf{\Theta}]$ is called the corresponding TLS correction.

The suitable norm used in previous definition of the TLS problem is called the Frobenius norm and for the matrix $\mathbf{X}$ is defined as follows

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{i=1}^{p} x_{ij}^2} = \sqrt{\text{trace}(\mathbf{X}^T\mathbf{X})} = \sqrt{\sum_{i=1}^{\min\{n,p\}} \sigma_i^2} = \sqrt{\sum_{i=1}^{\text{rank}(\mathbf{X})} \sigma_i^2},$$

where σ_i's are the singular values of the matrix $\mathbf{X}$.

Let us consider an n-element point set $P \in \mathbb{R}^{p+1}$, whose ith point is denoted by p_i, i.e.$p_i = (X_{i,1},\ldots,X_{i,p},Y_i)^T$. A model parameter vector β corresponds to a p-dimensional hyperplane, which we will denote by $\rho(\beta)$ or simply ρ. The residual $d_i(\rho)$ is defined to be the signed orthogonal distance from ρ to p_i. In this formulation the total least squares problem is equivalent to computing the hyperplane that minimizes the sum of the squared orthogonal distances from the data points p_i to the fitting hyperplane ρ. The normal vector of the hyperplane ρ is $\nu = \left[\beta^T,-1\right]^T$ and the formula for the orthogonal distance of point $p_i \in \mathbb{R}^{p+1}$ from the hyperplane ρ is

$$\frac{\left|\nu^T(A-p_i)\right|}{\|\nu\|},$$

where $A \in \rho$ is arbitrary point. Then we can formulate the total least square problem as

$$\begin{aligned}\hat{\beta}^{(TLS,n)} &= \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \frac{\left|\nu^T(A-p_i)\right|^2}{\|\nu\|^2} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \frac{\left|[\beta^T,-1]\begin{bmatrix} X_i \\ Y_i \end{bmatrix}\right|^2}{\|[\beta^T,-1]\|^2} \\ &= \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{1+\|\beta\|^2} \sum_{i=1}^{n} |Y_i - X_i\beta|^2 = \arg\min_{\beta\in\mathbb{R}^p} \frac{\|Y_i - X_i\beta\|}{\sqrt{1+\|\beta\|^2}}. \end{aligned} \quad (2)$$

It tell us that while the OLS minimizes a sum of squared residuals (vertical distances), TLS minimizes a sum of weighted squared residuals. The TLS weights the residuals by multiplying them with inverse of the corresponding error covariance matrix in order to derive a consistent estimate. In spite of the TLS theory connects the algebraic and numerical mathematics with statistics, there are not so many papers from the statistical point of view. The introduction into this field is for example in [10] and for further reading we can recommend [9] and [11].

In practice multivariate datasets contain often outliers, that is, data points that deviate from the usual assumptions or from the linear pattern formed by the majority of the data. If we apply classical statistical method such as TLS to these datasets we can obtain a misleading estimation. Our goal is to develop methods based on TLS technique that are robust against the possibility that one or several outliers may occur anywhere in the data. We focus on high-breakdown methods, which can deal with a substantial fraction of outliers in the data and is computationally efficient. We consider the idea of trimming bad observations proposed in least trimmed squares estimator (LTS) by Rousseeuw [6] and adapt it to TLS.

2 Total least trimmed squares

The total least trimmed squares method (TLTS) is based on the principle of robustification of OLS introduced by Rousseeuw [6] and developed by Rousseeuw and Leroy [7]. The estimator is based on the trimming of influential points, which is supposed to be outliers, and minimizes the sum of the h smallest squared orthogonal distances of data points p_i's from the pth dimensional fitting hyperplane $\rho(\beta)$.
The j-th orthogonal distances is denoted by $d_j(\beta)$ and defined by

$$d_j(\beta) = \frac{\left|Y_j - X_j^T\beta\right|}{\|[-1,\beta^T]\|}.$$

The TLTS estimation is defined by

$$\hat{\beta}^{(TLTS,n,h)} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{h} d^2_{(i)},$$

where h is an optional parameter called coverage and satisfying $\frac{n}{2} \le h \le n$, $d^2_{(i)}$ is the i-th least squared orthogonal distance, i.e. for any $\beta \in \mathbb{R}^p$

$$d^2_{(1)}(\beta) \le d^2_{(2)}(\beta) \le \ldots \le d^2_{(n)}(\beta).$$

Since TLTS has the infinite local sensitivity, we can define the total least weighted squares (TLWS) estimator. TLWS is based on the idea of an implicit weighting and multiply the squared orthogonal distances by a weights from $\langle 0, 1\rangle$. The definition and more details are in [1]. However, TLTS gives in overwhelming majority sufficiently good results as TLWS and for small datasets we have for TLTS in contrast to TLWS an useable exact algorithm.

If we consider the model with intercept or some columns of $\mathbf{X}$ are known exactly, the TLS solution does not give the accurate estimation. It is natural to require that the corresponding columns of the data matrix $\mathbf{X}$ be unperturbed since they are known exactly. The generalization of the TLTS approach is called mixed least trimmed squares - total least trimmed squares (LTS-TLTS) estimator and it minimizes the sum of the h smallest squared distances, which is given as the sum of both parts (orthogonal and vertical). To solve the mixed LTS-TLTS problem we us QR factorization. At first we solve the ordinary TLS problem of reduced dimension and after that we compute the components corresponding to non-perturbed variables. The detailed description of classical mixed LS-TLS is in [12] and the proper definition of its robustified version in [1].

3 Properties of total least trimmed squares

TLTS has some nice properties and may become the most popular robust estimation technique for error-in-variables models as LTS is for classical regression. The first important properties is that the solution of TLTS always exists. The TLTS estimator minimizes the objective function $\sum_{i=1}^{h} d_{(i)}^2$, where h is to be chosen between $n/2$ and n. This is equivalent to finding the subset of size h with the smallest TLS objective function. The TLTS estimate of the unknown parameter β^0 is then the TLTS estimate of that subset.

Since we want to have robust estimator, we have to show the resistance of TLTS against the occurrence of outliers. The most widely used measure of robustness is the finite sample breakdown value (breakdown point), which is a global measure of reliability and says when an estimator "still gives some relevant information". The breakdown value of a regression estimator T at a dataset $\mathbf{D}$ is the smallest fraction of outliers that can have an arbitrarily large effect on breakdown value. It is denoted by $\varepsilon_n^*(T, \mathbf{D})$ and the formally definition, due to [3] is following.

Let $\mathbf{D} = \{(X_{1,1}, \ldots, X_{1,p}, Y_1), \ldots, (X_{n,1}, \ldots, X_{n,p}, Y_n)\}$ be a sample of n data points, and let T be a regression estimator so that $\hat{\beta} = T(\mathbf{D})$. Consider all possible corrupted samples $\mathbf{D}'$ (submatrix of $\mathbf{D}$) that are obtained by replacing any m of the original data points by arbitrary values.
The breakdown value of the estimator T at the sample $\mathbf{D}$ is defined as

$$\varepsilon_n^*(T, \mathbf{D}) := \min\left\{\frac{m}{n}; \sup_{\mathbf{D}'} \|T(\mathbf{D}') - T(\mathbf{D})\| = \infty\right\}.$$

If $\sup_{\mathbf{D}'} \|T(\mathbf{D}') - T(\mathbf{D})\|$ is infinite, this means that m outliers can have an arbitrary large effect on T then the estimator "breaks down". So the breakdown value is the smallest

fraction of contamination that can cause the estimator T fails. Note that this breakdown value usually does not depend on given dataset $\mathbf{D}$ and depends only slightly on the sample size n. The limit of $\varepsilon_n^*(T)$ for $n \to \infty$ gives the asymptotic breakdown value $\varepsilon^*(T)$ which is the value that is usually called breakdown point of given estimator and denotes in percents. For robust estimators, the breakdown point should be more than 0 % and the highest breakdown point that can be achieved is 50%, this estimators are called high-breakdown.

If the data come from a continuous distribution, the breakdown point of the TLTS estimator for any integer h with $[(n+p+1)/2] \leq h \leq n$ is

$$\varepsilon_n^*(TLTS, h) = \frac{n-h+1}{n}.$$

The maximum breakdown value is reached for $h = [(n+p+1)/2]$, when the breakdown point $\varepsilon^*(TLTS, h) = 50\%$, whereas $h = n$ gives the TLS estimator with breakdown point equal to 0%. To prove the previous statement we have to find lower and upper bound. First we show that $\varepsilon_n^*(TLTS, h) \leq (n-h+1)/n$. From the definition of the breakdown point is obvious that we have to show, that we can always construct a corrupted sample $\mathbf{D}'$ with $(n-h+1)$ observations, such that the TLTS estimation breaks down. Let us define $\hat{\beta} = \hat{\beta}^{TLTS,n,h}(\mathbf{D})$, $\hat{\beta}' = \hat{\beta}^{TLTS,n,h}(\mathbf{D}')$, and take some $M > \beta$. Further define $M_X = \max_i \|X_i\|$. Now we set all $(n-h+1)$ corrupted observation equal to the point $(X, Y) = (X, M_x M^2 + K)$ for which $\|X\| = M_x$ and $K > 0$. These replaced observations satisfy

$$|X_i\beta| \leq \|X_i\|\|\beta\| < \|X\|M = M_x M < M_x M^2 + K = Y$$

and thus

$$d_i(\beta) = \frac{|Y_i - X_i^T\beta|}{\|[-1, \beta^T]\|} \geq \frac{|Y_i| - |X_i^T\beta|}{\sqrt{1+\|\beta^T\|^2}} \geq \frac{M_x M^2 + K - M_x M}{\sqrt{M^2+1}} = \frac{M_x M(M-1) + K}{\sqrt{M^2+1}}.$$

Since $(n-h+1) > n-h$ we obtain

$$\sum_{i=1}^{h} d_{(i)}^2 > \frac{M_x M(M-1) + K}{M^2+1}$$

and since we can choose K arbitrary large the minimum of the objective function of TLTS will not be reached for $\|\beta\| < M$. As $\|\beta'\| \geq M$ and letting M go to infinity causes the TLTS estimation to break down. The lower bound needs more complicated construction, but the proof is similar as in previous case and as for the LTS estimation (see [7] chapter 3). It has to be shown that if we replaced $n-h$ points in the original sample the TLTS estimation does not break down, i.e. $\|\beta - \beta'\|$ is bounded.

The upper bound we have to prove by another way than Rousseeuw and Leroy did in [7] for LTS. They showed that any regression equivariant estimator T satisfies

$$\varepsilon_n^*(T, \mathbf{D}) \leq \frac{\frac{n-p}{2}+1}{n}$$

at all samples $\mathbf{D}$. But unfortunately our TLTS estimator does not have this desirable property.

We can mention that any estimator $\hat{\beta}$ is

- regression equivariant if $\hat{\beta}(\mathbf{X}, \mathbf{Y} + \mathbf{X}\mathbf{v}) = \hat{\beta}(\mathbf{X}, \mathbf{Y}) + \mathbf{v}$, where $\mathbf{v}$ is any $p \times 1$ vector.
- scale equivariant if $\hat{\beta}(\mathbf{X}, c\mathbf{Y}) = c\hat{\beta}(\mathbf{X}, \mathbf{Y})$, where c is any scalar.
- matrix affine equivariant if $\hat{\beta}(\mathbf{X}\mathbf{A}, \mathbf{Y}) = \mathbf{A}^{-1}\hat{\beta}(\mathbf{X}, \mathbf{Y})$, where $\mathbf{A}$ is any $p \times p$ non-singular matrix.

It is easy to prove by finding simple univariate examples that both ordinary TLS and TLTS do not satisfy any mentioned equivariance. We proved it on computer for small datasets with 10 observations.

As well as most robust estimators, if we want to use TLTS, we have to make a compromise between robustness and efficiency. If we do not have any prior information about the occurrence of outliers, but we are sure that the data contains less than 25% of contamination, a good compromise between breakdown point and statistical efficiency is obtained by putting $h = floor(0.75 \cdot n)$, yielding an breakdown point of 25%.

The investigation of further properties such as consistency or asymptotic normality is task of future work.

4 Computation of total least trimmed squares

The computation of LTS estimates is not a straightforward task. The optional function of TLTS is continuous, non-convex, non-differentiable and has multiple local minima, whose number commonly rises with the number of observations and unknowns. It is obvious that the TLTS estimator coincides with the TLS estimator for the subset of h observations whose sum of squared orthogonal distances is minimal. Since the classical finite exhaustive algorithm has to compute the sum of squared orthogonal distances for all $\binom{n}{h}$ subsets, for large datasets in high dimensions it is practically not feasible to find the exact solution. In [1] we proposed the non-exhaustive exact algorithm based on a branch-and-bound (BAB) technique. This BAB algorithm slightly extends the set, for that we can compute the exact solution, but still the number of observations should be less than 60 and $p < 6$. For larger datasets with more observations and unknowns it is necessary to use some approximating algorithm. We developed the resampling algorithm for TLTS based on the idea of PROGRESS algorithm for LTS proposed by Rousseeuw and Leroy [7] and improved into FAST-LTS algorithm by Rousseeuw and Van Driessen in [8]. The algorithm is very simple and aims to find the h-subset which yields the smallest objective function. The algorithm usually finds a local minimum which is close to the global minimum, but not necessarily equal to that global minimum. A key element of the algorithm is the fact that starting from any h-subset, it is possible to construct another h-subset yielding a lower value of the objective function.

The simple description of this algorithm is following:

- suppose we have an h-subset H_{old} with corresponding TLS-estimate $\hat{\beta}_{old}^{(TLS,n)}$,
- compute the orthogonal distances d_i for all points from the hyperplane $\rho(\hat{\beta}_{old}^{(TLS,n)})$,
- set $H_{new} := \{h \text{ observations with smallest orthogonal distances } d_i\}$.

The TLS estimate $\hat{\beta}_{new}^{(TLS,n)}$, based on H_{new}, and its corresponding orthogonal distances then yield a value of the objective function that is smaller or equal to that of H_{old}. The idea of the algorithm is to construct many different initial h-subsets, apply previous steps until convergence, and keep the solution with the lowest value of the objective function. If we have small datasets and we construct the initial h-subset from a random (p+1)-subset more than thousand times, we usually obtain the global minimum. For large datasets we can not verify the optimality of the solution.

In spite of the algorithm gives reasonable estimations and is very fast, Hawkins and Olive [4] proved that elemental concentration algorithms are zero breakdown and that elemental basic resampling estimators are zero breakdown and inconsistent. For example the breakdown point of concentration algorithms that use K elemental starts is bounded above by K/n. For example if 100 starts are used and $n = 10000$, then the breakdown value is at most 1%. To cause a algorithm to break down, simply contaminate one observation in each starting elemental set. Since K elemental starts are used, at most K points need to be contaminated. Consequently, for small datasets we are using exact BAB algorithm, which gives robust estimation and for large datasets we tried to use as much as possible starting elemental sets.

5 Benchmark instances

In this small experimental study we use three widely used benchmark instances. All of them are taken from [7] and have been studied by many statisticians in robust literature, who in contrast to us assumed that independent variables were measured exactly and all random error is only in dependent variable.

The first dataset is Hertzsprung-Russell Diagram of the Star Cluster CYG OB1, which contains 47 stars in the direction of Cygnus, from C.Doom. The independent variable $\mathbf{X}$ is the logarithm of the effective temperature at the surface of a star and $\mathbf{Y}$ is the logarithm of its light intensity. The model with intercept is consider. Four stars, called giants, have low temperature with high light intensity and represent outliers. The second datasets is modified Wood Gravity Data. This is a real data set with five independent variables and intercept. It consists of 20 cases and some of them were replaced to contaminate the data by few outliers. The last dataset is called Hawkins, Bradu and Kass Data. This dataset is artificial and have been generatedfor illustrating the merits of a robust technique. It offers the advantage that at least the position of the good or bad leverage points is known. The Hawkins, Bradu and Kass data consists of 75 observations in four dimensions. The first ten observations form a group of identical bad leverage points, the next four points

are good leverage while the remaining are good data. In all three cases we supposed that there are mostly 20% of outliers in datasets and we set $h = 0,8 \cdot n$. We investigate, if our algorithms identify outliers and downweight them ("% success"). Further we observed the CPU computational time in seconds ("time"), the value of objective function (S) and the estimated parameter $\hat{\beta}$. All computations were performed in MATLAB on personal computer with 1826 Mhz processor. The resampling algorithm used 10000 initial p-subset estimates. Results are summarized in the table 1.

Data	Algorithm	n	p	S	CPU time	% success	$\hat{\beta}^{(LTS-TLTS,n,h=0,8\cdot n)}$
Stars	resampling	47	2	5.8739	8.9844	96	$(-21.54, 5.98)^T$
	BAB	47	2	4.9733	381.6406	100	$(-16.05, 4.75)^T$
Wood	resampling	20	6	0.0089	9.6563	90	$(0.63, 0.97, -3.87, -0.49, -0.87, 0.60)^T$
	BAB	20	6	0.0008	0.4219	100	$(0.36, 0.22, -0.12, -0.57, -0.42, 0.64)^T$
Hawkins	resampling	75	4	29.0801	17.5620	100	$(-0.42, 0.19, 0.16, -0.16)^T$

Table 1: Performance of the mixed LTS-TLTS algorithms on three small data sets.

6 Conclusion

In this paper we summarized existing knowledge on the robustified total least squares method, introduced TLTS estimator and its modification. We mentioned some properties of TLTS, investigate and explain its behavior and prove its positive breakdown point which depends on the subset size h to be chosen by the user. The choice of h is a option between efficiency and breakdown. We discussed the advantages and disadvantages of different algorithms to calculate an estimate. In the last section we showed some results and performance of mentioned estimators and algorithms and discussed them. Further work will focus primarily on prove properties of TLTS such as consistency and improvement of branch-and-bound algorithm.

All MATLAB source codes of all algorithms mentioned in this paper may be obtained on request without charge from the author.

References

[1] J. Franc. *Some computational aspects of robustified total least squares.* Stochastic and Physical Monitoring Systems proceedings 2011, Křižánky, Czech Republic (2011).

[2] G. Golub and C. Van Loan. *An analysis of the total least squares problem.* SIAM J. Numerical Analysis **17** (1980), 883–893.

[3] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, Inc., New York, (1986).

[4] D. M. Hawkins and D. J. Olive. *Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm.* Journal of the American Statistical Association **97** (2002), 136—-159.

[5] I. Markovsky and S. Van Huffel. *Overview of total least squares methods.* Signal Processing **87** (2007), 2283–2302.

[6] P. J. Rousseeuw. *Least median of squares regression.* Journal of the American Statistical Association (1984), 871—-880.

[7] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, Inc., New York, (1987).

[8] P. J. Rousseeuw and K. Van Driessen. *Computing lts regression for large data sets.* Data Mining and Knowledge Discovery (2006).

[9] S. Van Huffel. *Recent Advances In Total Least Squares Techniques and Errors-in-variables Modeling.* SIAM Proceeding Series, Philadelphia, (1997).

[10] S. Van Huffel. *Total least squares and errors-in-variables modeling: Bridging the gap between statistics, computational mathematics and engineering.* COMPSTAT 2004 -Symposium (2004).

[11] S. Van Huffel and P. Lemmerling. *Total Least Squares and Errors-in-variables Modeling: Analysis, Algorithms and Applications.* Kluwer Academic Publisher, Dordrecht, (2002).

[12] S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem: Computational Aspects and Analysis.* SIAM, Philadelphia, (1991).

Cramér–von Mises Type Estimators*

Jitka Hanousková

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

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

Abstract. This paper summarizes results published in Proceedings of the 17th European Young Statisticians Meeting [1] and Proceedings of SPMS 2011 [2]. This two articles deal with minimum distance estimates of unknown parameter θ_0. Two modification of well known Cramér–von Mises distance are defined and statistical properties of the corresponding estimators are investigated.

Keywords: minimum distance estimates, Cramér–von Mises distance, consistency, robustness

Abstrakt. Tento článek shrnuje výsledky publikované v Proceedings of the 17th European Young Statisticians Meeting [1] a Proceedings of SPMS 2011 [2]. Tyto články se zabývají odhady metodou s minimální vzdáleností neznámého parametru θ_0. Jsou definovány dvě modifikace známé Cramér–von Mises vzdálenosti a jsou zkoumány statistické vlastnosti jim odpovídajících odhadů.

Klíčová slova: odhady s minimální vzdáleností, Cramér–von Mises vzdálenost, konzistence, robustnost

1 Summary

We investigate minimum distance estimators based on two different modifications of Cramér–von Mises distance. We proposed this modifications due to improve robustness and consistency properties of Cramér–von Mises estimate. First modification is called generalized Cramér–von Mises distance (GCM) and is defined by replacing the second power in the original formula by general power.

$$d_{GCM}(F,G) = \int (F(x) - G(x))^{p/q} \mathrm{d}F(x), \text{ where } p \text{ is even, and } q \text{ is odd}.$$

Neither original CM distance nor GCM distance are symmetric and therefore there are two possibilities how to define minimum distance estimate based on this distance. We can search for minimum of (1) or (2)

$$d_{GCM}(F_n, F_\theta) = \int (F_n(x) - F_\theta(x))^{p/q} \mathrm{d}F_n(x) = \frac{1}{n}\sum_{i=1}^{n} (F_\theta(x_i) - F_n(x_i))^{p/q} \qquad (1)$$

$$d_{GCM}(F_\theta, F_n) = \int (F_\theta(x) - F_n(x))^{p/q} \mathrm{d}F_\theta(x)\,. \qquad (2)$$

*This work has been supported by the grant SGS 10/209/OHK4/2T/14

The integral form (2) could be simplified by straightforward calculating to the expression more appropriate for simulation purposes.

$$\int (F_\theta(x) - F_n(x))^{p/q} \mathrm{d}F_\theta(x) = \frac{q}{p+q} \sum_{i=1}^{n} \Big[\Big(F_\theta(x_i) - \frac{i-1}{n} \Big)^{\frac{p+q}{q}} - \Big(F_\theta(x_i) - \frac{i}{n} \Big)^{\frac{p+q}{q}} \Big] . \quad (3)$$

The second investigated modification is so called Kolmogorov–Cramér distance defined as distance between empirical distribution function F_n and theoretical distribution function F in the following way. Define a sequence $(d_i(F_n, F))_1^{2n}$ by

$$d_i(F_n, F) = (F_n(x_i) - F(x_i))^{p/q} \qquad \text{for } i = 1, \ldots, n\,, \quad (4)$$

$$d_{2n+1-i}(F_n, F) = (F_{n-}(x_i) - F_-(x_i))^{p/q} \qquad \text{for } i = 1, \ldots, n\,, \quad (5)$$

where $F_{n-}(x_i) = \lim_{x \to x_i -} F_n(x)$ and similarly $F_-(x_i) = \lim_{x \to x_i -} F(x)$, p is even, q is odd. Then we define KC distance

$$d_{KC}(F_n, F) = \frac{1}{m} \sum_{i=1}^{m} d_{(i)}(F_n, F)\,, \quad (6)$$

where $d_{(i)}(F_n, F)$ denotes ordered sequence of $(d_i(F_n, F))_1^{2n}$ and m is an integer less or equal to $2n$. The distance is called Kolmogorov–Cramér because for choice $m = 1$ it converts to p/q power of Kolmogorov distance. Obviously the corresponding MD estimator is the same as for standard Kolmogorov distance. And for $m = n$, $p = 2$, and $q = 1$ the KC distance becomes average of CM distance and CM distance with left limit instead of $F(x_i), F_n(x_i)$.

For the KC distance estimate we are able to prove consistency of the order $n^{-1/2}$ in the L_1–norm and in the expected L_1–norm under the assumption of finiteness of degree of variation of the family $\{F_\theta, \theta \in \Theta\}$. Moreover, the parameter m is allowed to depend on the sample size n, we denote the distance as KCp. Let $m = O(n^\beta), \beta \leq p/2q$. In this case we do not achieve the $n^{-1/2}$ consistency but only $n^{-\gamma}$ consistency, where $\gamma = 1/2 - \beta q/p$. Unfortunately we have not such theoretical results for GCM estimate, but extensive simulation study was produced to study its statistical properties experimentally and results are as follows.

As simulations show, the minimum distance estimators with CM and GCM distance posses consistency in the L_1–norm for all choices of parameters p, q with $p/q \in \langle 0, 2\rangle$ if the sample is non-contaminated and preserve some consistency even under contamination up to the proportion $\varepsilon = 0.15$. The character of their L_1–consistency is almost the same as for Kolmogorov estimator, which is proven to by L_1–consistent. But under increasing contamination proportion the Kolmogorov estimate loses its consistency in contrast to CM and GCM estimates. Further, the best choice of p/q was established. The heavier contamination is the smaller choice of p/q produces the best estimate. If the sample is non-contaminated, the choice p/q near to zero produces faulty estimator. Thus, the optimal could be the choice $p/q \approx 0.2$. For KC estimate we have proven L_1–consistency for non-contaminated sample. Under contamination KC estimate loses consistency in contrast to KCp which preserve some. Regarding robustness if m is chosen fixed for all n, the robustness of KC estimate depends on p/q only slightly. Situation differs if

the parameter m depends on the sample size n. There are two situations, if we choose the parameter β as large as possible for the given value of p/q then the best results are achieved for $p/q = 2$. On the other hand, if we fix the value of parameter β, then the smaller value of p/q the better results we gain.

To conclude, the GCM estimate is still more robust than the KC estimate, but not proven to be consistent. Thus the proposed KC estimate partly integrates the good properties of Kolmogorov and GCM estimates. These conclusions show us that it is worth to study the robustness and efficiency of GM and KC estimate theoretically, not only via simulations.

References

[1] J. Hanousková and V. Kůs. *Consistency and robustness of Cramér–von Mises type estimators*. Proceedings of 17th European Young Statisticians Meeting (2011), 99–103 .

[2] J. Hanousková and V. Kůs. *Generalized Cramér–von Mises distance estimators*. Proceedings of SPMS (2011), *Not published yet, accepted.*

MSAR BTF Model*

Michal Havlíček†

2nd 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. The Bidirectional Texture Function (BTF) is the recent most advanced representation of material surface visual properties. BTF specifies the changes of its visual appearance due to varying illumination and viewing conditions. Such a function might be represented by thousands of images of given material surface taken in different illumination and viewing conditions. Resulting BTF size, hundreds of gigabytes, excludes its direct rendering in graphical applications, accordingly some compression of these data is obviously necessary. This paper presents a novel probabilistic model based algorithm for realistic multispectral BTF texture modelling. This complex but efficient method combines several multispectral band limited spatial factors and corresponding range map to produce the required BTF texture. Proposed scheme enables very high BTF texture compression ratio and in addition may be used to reconstruct BTF space i.e. unmeasured parts of the BTF space.

Keywords: BTF, texture analysis, texture synthesis, data compression, virtual reality

Abstrakt. Obousměrná texturovací funkce je nejpokročilejší v současné době používaná reprezentace vizuálních vlastností povrchu materiálu. Popisuje změny jeho vzhledu v důsledku měnících se úhlů osvětlení a pohledu. Tato funkce může být reprezentována tisíci obrazy daného povrchu vzorku materiálu, které jsou pořízeny za různých světelných podmínek a pod různým úhlem. Výsledná velikost BTF, stovky gigabajtů, znemožňuje přímé využití v grafických aplikacích a je tedy třeba tato data nějakým způsobem komprimovat. Tento článek představuje nový pravděpodobnostní model umožňující realistické modelování multispektrálních BTF textur. Tato složitá ale účinná metoda kombinuje několik multispektrálních frekvenčních faktorů a odpovídající hloubkovou mapu výsledkem čehož je požadovaná BTF textura. Navržený postup umožňuje velmi vysokou úroveň komprese BTF textur a navíc může být využit k rekonstrukci BTF prostoru, tj. těch částí BTF, které nebyly původně naměřeny.

Klíčová slova: BTF, analýza textur, syntéza textur, komprese dat, virtuální realita

1 Introduction

Photo realism in virtual reality scenes cannot be realized without nature like colour textures covering visualised scene objects. These textures can be either smooth or rough. The rough ones have rugged surface and do not obey Lambertian law, their reflectance is both illumination and view angle dependent. Such textures might be represented by

*This research was supported by the grant GAČR 102/08/0593 and partially by the projects MŠMT 1M0572 DAR, GAČR 103/11/0335, CESNET 387/2010.

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

Bidirectional Texture Function (BTF) [3] which is 7-dimensional function describing texture appearance variations due to varying illumination and viewing angles. Generally, textures can be either digitised natural or artificial materials or images synthesised from an appropriate mathematical model.

The former simplistic option suffers among others with extreme memory requirements for storage of a large number of digitised cross sectioned slices through different material samples (apposite example can be found in [15]). This solution become unmanageable for rough textures which require to store thousands of different illumination and view angle samples for every texture so that even simple virtual reality scene featuring only several different textures requires to store tera bytes of texture data which is still far out of limits for any present day hardware. Several so called intelligent sampling methods (for example [4], [5] and many others) were proposed to reduce these extensive memory requirements. All these methods are based on some sort of original small texture sampling and the best of them produce very realistic textures. However they require to store thousands images for every combination of viewing and illumination angle of the original target texture sample and in addition often produce visible seams (except for method presented in [10]). Some of them are computationally demanding and they are not able to generate textures unseen by these algorithms as well.

While synthetic textures are more flexible and extremely compressed, because only several parameters have to be stored in contrast with gigabytes of original data [15]. They may be evaluated directly in procedural form and can be used to fill virtually infinite texture space without visible discontinuities. On the other hand, mathematical models can only approximate real measurements which results in visual quality compromise of some methods. Several multispectral modelling approaches were published for example [11], [1], [12], [13]. Modelling multispectral images requires three dimensional models but it is possible to approximate such model with a set of simpler two dimensional ones. Evidently this leads to certain loss of information (for example three dimensional Causal Autoregressive (CAR) model [7] versus two dimensional CAR model [8]).

Among such possible models the random fields are appropriate for texture modelling not only because they do not suffer with some problems of alternative options (see [6], [12] for details) but they provide relatively easy to implement and computational undemanding texture synthesis and sufficient flexibility to reproduce a large set of both natural and artificial textures. While the random field based models quite successfully represent high frequencies appeared in natural textures low frequencies are sometimes difficult for them. This slight drawback may be overcome by usage of a multiscale random field model. In this case the hierarchy of different resolutions of an input image provides a transition between pixel level features and region or global features and hence such a representation simplify modelling a large variety of possible textures. Each resolution component is both analysed and synthesised independently. Multiple resolution decomposition may be performed by means of Gaussian Laplacian pyramids, wavelet pyramids or subband pyramids. Because of its relative simplicity we decided to utilize Gaussian Laplacian pyramid decomposition for this task.

2 Smooth Texture Model

The overall roughness of a textured surface significantly influences a BTF texture appearance. Such a surface can be specified using its range map, which can be estimated by several existing approaches. The most accurate range map can be acquired by direct measurement of the observed surface using corresponding range cameras, however this method requires special hardware and measurement methodology [9]. Hence alternative approaches for range map estimation from surface images are more suitable. One of the most accurate approaches is the photometric stereo [9] which estimates surface range map from at least three images obtained for different position of illumination source and fixed camera position. This approach was utilized for range map estimation from textures used for experiments described below. Naturally it is enough to estimate range map once per material and reuse it whenever it is needed.

We propose a novel algorithm for efficient rough texture modelling which combines an estimated range map with synthetic multiscale smooth textures generated using Multispectral Simultaneous Autoregressive Model (MSAR) [1]. The material visual appearance during changes of viewing and illumination conditions can be simulated using the bump mapping [2] or displacement mapping technique [16]. The obvious advantage of this solution is the possibility to use hardware support of bump mapping and displacement mapping in up to date visualisation hardware. The overall appearance is guided by the corresponding underlying probabilistic model.

2.1 Spatial Factorization

An analysed texture is decomposed into multiple resolution factors using Laplacian pyramid and the intermediary Gaussian pyramid $y_{\bullet}^{''(k)}$ which is a sequence of images and each its element is a low pass down sampled version of its predecessor. The Gaussian pyramid for a reduction factor n is [8]:

$$y_r^{''(k)} = \downarrow_r^n (y_{\bullet,i}^{''(k-1)} \otimes w), \quad k = 1, 2, \dots \quad ,$$

where $\downarrow_r^n$ denotes down sampling with reduction factor n and $\otimes$ is the convolution operation. The Laplacian pyramid $y_r^{'(k)}$ contains band pass components and provides a good approximation to the Laplacian of the Gaussian kernel. It can be constructed by simple differencing single Gaussian pyramid layers:

$$y_r^{'(k)} = y_r^{''(k)} - \uparrow_r^n (y_{\bullet}^{''(k+1)}), \quad k = 0, 1, \dots \quad .$$

As previously noticed each resolution data are independently modelled by their dedicated MSAR model so that model parameters are estimated for each component independently in analysis step.

2.2 Multispectral Simultaneous Autoregressive Model

In the multispectral case random field models are defined as intensity levels on multiple two dimensional lattice planes (e.g. in case of widely used standard RGB colour model three such planes are considered). The value at each lattice location is taken to be a linear

combination of neighbouring ones and an additive noise component. For mathematical simplicity, all lattices are double toroidal (the same way as in case of Gaussian Markov Random Field model [9] for example). Let a location within an $M \times M$ two dimensional lattice be denoted by $s = (s_1, s_2)$, with $s_1, s_2 \in J$ and the set J is defined as $J = \{0, 1, \ldots, M-1\}$. The set of all lattice locations is then defined as $\Omega = \{s = (s_1, s_2) : s_1, s_2 \in J\}$. The value of an image observation at location s is denoted by $y(s)$. These random vectors are expected to have zero mean. Neighbour sets relating the dependence of image plane i on image plane j are defined as $N_{ij} = \{r = (k, l) : k, l \in \pm J\}$ with the associated neighbour coefficients $q_{ij} = \{q_{ij}(r) : r \in N_{ij}\}$. The set $\pm J = \{-(M-1), \ldots, -1, 0, 1, \ldots, M-1\}$. We also use shortened notation: $q = \{q_{ij};\, i, j \in \{1, \ldots, P\}\}$ and $r = \{r_i;\, i \in \{1, \ldots, P\}\})$. P equals number of image planes. Neighbour sets may be classified as symmetric or nonsymmetric. In particular, in the case of multispectral models, a symmetric neighbour set is defined as one for which $r \in N_{ij} \iff -r \in N_{ji}$. Our model is defined on symmetric neighbour set. Scale parameter ρ associated with the corresponding noise component of the model is defined for each particular lattice.

The Multispectral Simultaneous Autoregressive model (MSAR) [1] relates each lattice position $y_i(s)$ to its neighbouring pixels, both within and between image planes, according to the following equations:

$$y_i(s) = \sum_{j=1}^{P} \sum_{r \in N_{ij}} \theta_{ij}(r) y_j(s \oplus r) + \sqrt{\rho_i} w_i(s), \quad i = 1, \ldots, P\ , \tag{1}$$

where ρ_i is noise variance of image plane i, $w_i(s)$ are i.i.d. random variables with zero mean and unit variance and $\oplus$ denotes modulo M addition in each index. Virtually the MSAR model characterizes the spatial interactions between neighbouring pixels through the parameter vectors $\theta = (\theta_{ij};\, i = 1, \ldots, P;\, j = 1, \ldots, P)^T$ and $\rho = (\rho_i;\, i = 1, \ldots, P)^T$. Rewriting (1) in matrix form for the RGB colour model, i.e. $i \in \{r, g, b\}$, the MSAR model equations are then $B(\theta) y = w$ where

$$B(\theta) = \begin{pmatrix} B(\theta_{rr}) & B(\theta_{rg}) & B(\theta_{rb}) \\ B(\theta_{gr}) & B(\theta_{gg}) & B(\theta_{gb}) \\ B(\theta_{br}) & B(\theta_{bg}) & B(\theta_{bb}) \end{pmatrix},$$

$$y = (y_r(s),\ y_g(s),\ y_b(s))^T, \quad w = (\sqrt{\rho_r} w_r(s),\ \sqrt{\rho_g} w_g(s),\ \sqrt{\rho_b} w_b(s))^T$$

and both $y_i(s)$ and $w_i(s)$ are M^2-vectors of lexicographic ordered arrays $\{y_i(s)\}$ and $\{w_i(s)\}$, respectively. The transformation matrix $B(\theta)$ is composed of $M^2 \times M^2$ block circulant submatrices

$$B(\theta_{ij}) = \begin{pmatrix} B(\theta_{ij})_1 & B(\theta_{ij})_2 & \ldots & B(\theta_{ij})_M \\ B(\theta_{ij})_M & B(\theta_{ij})_1 & \ldots & B(\theta_{ij})_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ B(\theta_{ij})_2 & B(\theta_{ij})_3 & \ldots & B(\theta_{ij})_1 \end{pmatrix}$$

where each element $B(\theta_{ij})_p, \quad p \in \{1, \ldots, M\}$ is an $M \times M$ circulant matrix whose (m,n)-th element is given by:

$$b(\theta_{ij})_p(m,n) = \begin{cases} 1, & i=j,\ m=n, \\ -\theta_{ij}(k,l), & k=p-1,\ l=((n-m) \mod M),\ (k,l)\in N_{ij}, \\ 0, & otherwise. \end{cases}$$

Writing the image observations as $y = B(q)^{-1}w$, the image covariance matrix is obtained as $\Sigma_y = \varepsilon\{yy^T\} = \varepsilon\ \{\ B(q)^{-1}\ ww^T\ [B(q)^{-1}]^T\ \} = B(q)^{-1}\Sigma_w[B(q)^{-1}]^T$ where

$$\Sigma_w = \varepsilon\{ww^T\} = \begin{pmatrix} \rho_r I & 0 & 0 \\ 0 & \rho_g I & 0 \\ 0 & 0 & \rho_b I \end{pmatrix}.$$

2.3 Parameter Estimation

It is necessary to notice that the selection of an appropriate MSAR model support is important to obtain good results in modelling of a given random field. If the contextual neighbourhood is too small it can not capture all details of the random field. Contrariwise, inclusion of the unnecessary neighbours add to the computational burden and can potentially degrade the performance of the model as an additional source of noise [9].

A least squares (LS) estimate of the MSAR model parameters can be obtained by equating the observed pixel values of an image to the expected value of the model equations. As we prefer RGB colour model our task leads to three independent systems of M^2 equations:

$$y_i(s) = q_i(s)^T\theta_i, \quad s\in\Omega, \quad i\in\{r,g,b\},$$

with vectors θ_i and $q_i(s)$ formed as follows $\theta_i = (\theta_{ir}, \theta_{ig}, \theta_{ib})^T$ and $q_i(s) = (\ \{y_r(s\oplus t) : t\in N_{ir}\},\quad \{y_g(s\oplus t) : t\in N_{ig}\},\quad \{y_b(s\oplus t) : t\in N_{ib}\}\ )^T$. The LS solution $\hat{\theta}_i$ and $\hat{\rho}_i$ can be found then as

$$\hat{\theta}_i = \left(\sum_{s\in\Omega} q_i(s)q_i(s)^T\right)^{-1}\left(\sum_{s\in\Omega} q_i(s)y_i(s)\right),$$

$$\hat{\rho}_i = \frac{1}{M^2}\sum_{s\in\Omega}(y_i(s) - \hat{\theta}_i^{\ T}q_i(s))^2 \quad .$$

2.4 Texture Synthesis

The goal of texture synthesis in case of probabilistic model is to generate image of arbitrary size directly from the model parameters so that resulting texture has the same statistical properties as measured and analysed original. Several possibilities exist for a finite lattice MSAR synthesis. The most effective method uses the discrete fast Fourier transformation (DFT). The MSAR model equations (1) may be expressed in terms of the DFT of each image plane as

$$Y_i(t) = \sum_{j=1}^{P} \sum_{r \in N_{ij}} \theta_{ij}(r) Y_j(t) e^{\sqrt{-1}\omega_{rt}} + \sqrt{\rho_i} W_i(t), \quad i = 1, \ldots, P \tag{2}$$

where $Y_i(t)$ and $W_i(t)$ are the two-dimensional DFT coefficients of the image observation $\{y_i(s)\}$ and noise sequence $\{w_i(s)\}$, respectively, at discrete frequency index $t = (m, n)$ and $\omega_{rt} = \frac{2\pi(mk+nl)}{M}$ for $r = (k, l)$. For the RGB colour model equations (2) can be written in matrix form as

$$Y(t) \; = \; \Lambda(t)^{-1} \, \Sigma^{\frac{1}{2}} \, W(t), \quad t \in \Omega$$

where the vectors $Y(t)$ and $W(t)$ are formed this way:

$$Y(t) = (Y_r(t), \; Y_g(t), \; Y_b(t))^T, \quad W(t) = (W_r(t), \; W_g(t), \; W_b(t))^T,$$

and the matrices Σ and $\Lambda(t)$ are defined as:

$$\Sigma = \begin{pmatrix} \rho_r & 0 & 0 \\ 0 & \rho_g & 0 \\ 0 & 0 & \rho_b \end{pmatrix},$$

$$\Lambda(t) = \begin{pmatrix} \lambda_{rr}(t) & \lambda_{rg}(t) & \lambda_{rb}(t) \\ \lambda_{gr}(t) & \lambda_{gg}(t) & \lambda_{gb}(t) \\ \lambda_{br}(t) & \lambda_{bg}(t) & \lambda_{bb}(t) \end{pmatrix},$$

$$\lambda_{ij}(t) = \begin{cases} 1 - \sum_{r \in N_{ij}} \theta_{ij}(r) \; e^{\sqrt{-1}\omega_{rt}} & i = j \, , \\ - \sum_{r \in N_{ij}} \theta_{ij}(r) \; e^{\sqrt{-1}\omega_{rt}} & i \neq j \, . \end{cases}$$

Apparently, the MSAR model will be stable and valid if $\Lambda(t)$ is nonsingular matrix $\forall t \in \Omega$. Given the model parameters, a $M \times M$ MSAR image can be synthesized according to the following algorithm:

1) Generate the i.i.d. noise arrays $\{w_i(s)\}$ for each image plane using a pseudo random number generator.
2) Calculate the two-dimensional DFT of each noise array i.e. produce the transformed noise arrays $\{W_i(t)\}$).
3) For each discrete frequency index t compute $Y(t) = \Lambda(t)^{-1} \, \Sigma^{\frac{1}{2}} \, W(t)$.
4) Perform the two-dimensional inverse DFT of each frequency plane $\{Y_i(t)\}$, producing the synthesized image planes $\{y_i(s)\}$.
The resulting image planes will have zero mean thus it is necessary to add desired mean to each spectral plane after step 4. Fine resolution texture is obtained from the pyramid collapse procedure that is inversion process to the procedure described in section 2.1.

3 Results

We have tested the algorithm on colour BTF textures from the University of Bonn BTF measurements [15], namely on following materials: artificial leather, foil, glazed tails,

plastic floor and two different samples of wood. Each BTF material sample comprised in mentioned database is measured in 81 illumination and 81 viewing angles and each resulting image has resolution 800 × 800 pixels, so that 6561 such images had to be analysed for each material.

The open source project Blender[1] with special plugin for BTF support [14] was used to render the results i.e. the scene in virtual reality featuring three-dimensional object covered with synthesised BTF texture. Figure 1 demonstrates the result for one picked material, foil in this case, i.e. synthesised BTF texture combined with its range map in a displacement mapping filter of the rendering software mapped on bumpy board. Scene was rendered in several different illumination conditions with fixed view angle to demonstrate visual quality of synthesised BTF.

3.1 Implementation Details

The source code was written in C++ and compiled in several different environments (namely with g++ versions 3.4.4, 4.1.2, 4.3.2, 4.3.4 and 4.5.0)[2] and tested on many different systems including Microsoft's Windows XP operating system with cygwin[3] environment as well as Linux systems to prove stability and portability of the program. This implementation uses many parts of library developed at Pattern Recognition Department, Institute of Information Theory and Automation, ASCR[4], such as image reading and writing routines, memory management and XML format support.

4 Summary and Conclusion

Our testing results of the algorithm on available BTF data are encouraging. Some synthetic textures reproduce given measured texture images so that both natural and synthetic texture are almost visually indiscernible. The main benefit of this method is more realistic representation of texture colourfulness which is naturally apparent in case of very distinctively coloured textures. The multiscale approach is more robust and allows sometimes better results than the singlescale one due to capabilities of the model described above.

The proposed method allows huge compression ratio unattainable by alternative intelligent sampling approaches for transmission or storing texture data while it has still moderate computation complexity. It is necessary to mention that the complexity of analysis is not as important as the complexity of synthesis because the parameter estimation can be performed offline unlike the synthesis which should be as fast as possible. The method does not need any time consuming numerical optimisation like for example the usually employed Monte Carlo methods. The replacement of the bump mapping technique with the displacement mapping further significantly improve the visual quality of the results. The presented method is based on the mathematical model in contrast to intelligent sampling type of methods, and as such it can only approximate realism

[1] http://www.blender.org
[2] http://gcc.gnu.org
[3] http://www.cygwin.com
[4] http://www.utia.cas.cz

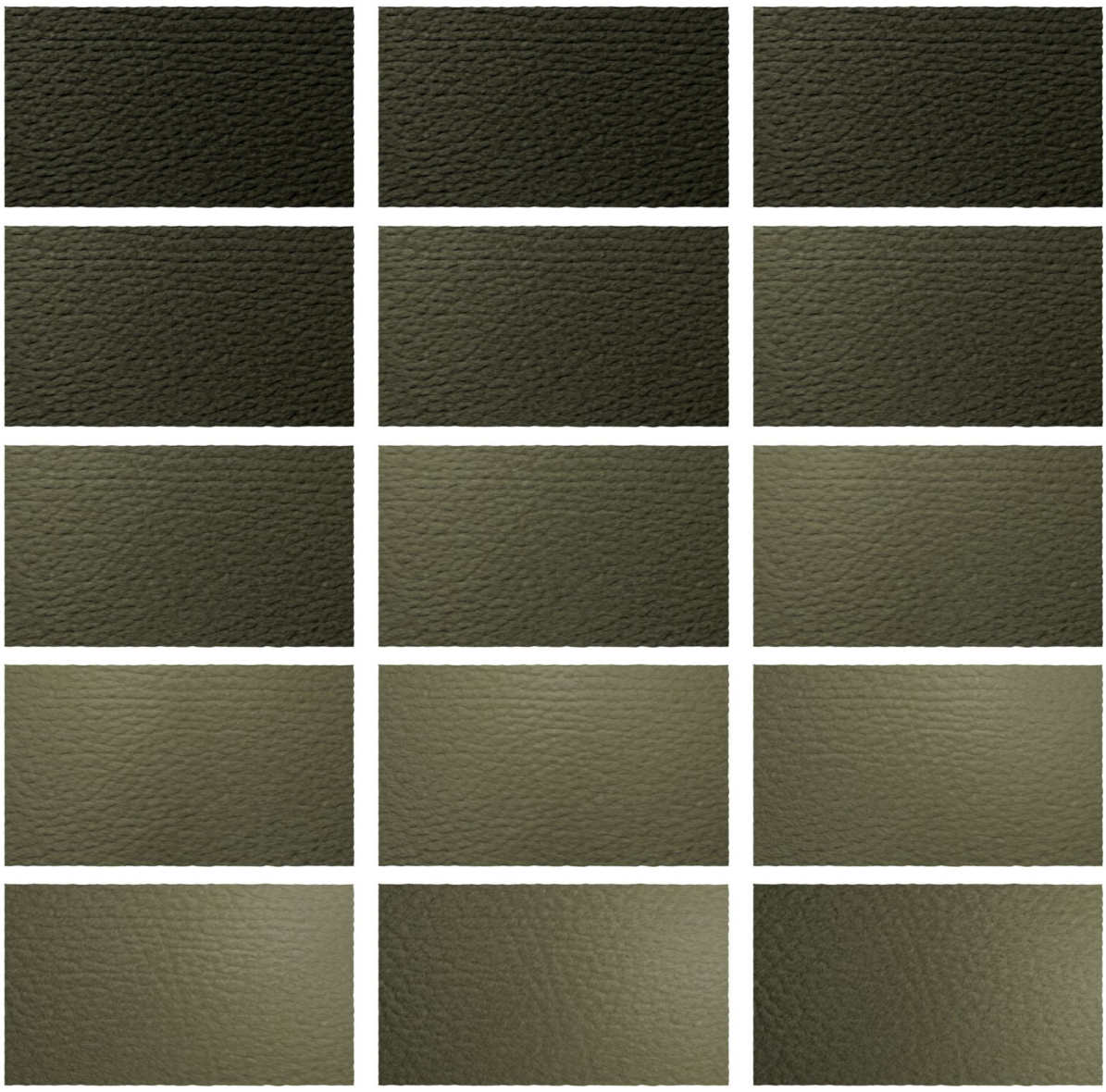


Figure 1: Resulting BTF texture of foil, synthesised texture combined with its range map mapped on bumpy board rendered with 15 different angles of illumination and fixed view angle.

of the original measurement. However it offers easy simulation of even non existing i.e. previously unmeasured BTF textures and fast seamless synthesis of texture of arbitrary size.

5 Future Work

This developed model might be further tested on different BTF measurements and compared with other random field based models such as already mentioned CAR or Gauss-Markov random field model [9]. Though the quality of the model was proven it would be interesting to find its limitation and study the influence of the size of the neighbourhood to overall performance for example. Naturally more interesting is possible extension of current implementation by means of parallel programming with use of OpenMP[5] library which is straightforward and would notably increase the model performance. It is also possible rewrite the source code so that program would perform all computations on a graphics processing unit.

References

[1] J. Bennett, A. Khotanzad. *Multispectral Random Field Models for Synthesis and Analysis of Color Images.* IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(3) (1998), 327–332.

[2] J. Blinn. *Simulation of Wrinkled Surfaces.* ACM SIGGRAPH Computer Graphics **12**(3) (1978), 286–292.

[3] K. Dana, S. Nayar, B van Ginneken, J. Koenderink. *Reflectance and Texture of Real-World Surfaces.* Proceedings of IEEE Conference Computer Vision and Pattern Recognition (1997), 151–157.

[4] J. De Bonet *Multiresolution sampling procedure for analysis and synthesis of textured images.* Proceedings of SIGGRAPH 97, ACM (1997), 361–368.

[5] W. Efros, A.A. Freeman. *Image quilting for texture synthesis and transfer.* SIGGRAPH 2001, Computer Graphics Proceedings, E. Fiume, Ed. ACM Press / ACM SIGGRAPH (2001), 341–346.

[6] M. Haindl. *Texture synthesis.* CWI Quarterly **4**(4) (1991), 305–331.

[7] M. Haindl, J. Filip, M. Arnold. *BTF Image Space Utmost Compression and Modelling Method.* Proceedings of 17th ICPR **3**, IEEE Computer Society Press (2004), 194–198.

[8] M. Haindl, J. Filip. *A Fast Probabilistic Bidirectional Texture Function Model.* Proceedings of ICIAR (lecture notes in computer science 3212) **2**, Springer-Verlag, Berlin Heidenberg (2004), 298–305.

[5] http://openmp.org

[9] M. Haindl, J. Filip. *Fast BTF Texture Modeling*. Proceedings of the 3rd International Workshop on Texture Analysis and Synthesis (2003), 47–52.

[10] M. Haindl, M. Hatka. *BTF Roller*. Texture 2005: Proceedings of the 4th International Workshop on Texture Analysis and Synthesis (2005), 89–94.

[11] M. Haindl, V. Havlíček. *Multiresolution colour texture synthesis*. Proceedings of the 7th International Workshop on Robotics in Alpe-Adria-Danube Region, K. Dobrovodský, Ed. Bratislava: ASCO Art (1998), 297–302, Berlin: Springer-Verlag (2000), 114–122.

[12] M. Haindl, V. Havlíček. *A multiresolution causal colour texture model*. Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, Springer-Verlag (2000), 114–122.

[13] M. Haindl, V. Havlíček. *A multiscale colour texture model*. Proceedings of the 16th International Conference on Pattern Recognition (2002), 255–258.

[14] M. Hatka *Vizualizace BTF textur v Blenderu*. Doktorandské dny 2009, sborník workshopu doktorandů FJFI oboru Matematické inženýrství, České vysoké učení technické v Praze (2009), 37–46.

[15] G. Müller, J. Meseth, M. Sattler, R. Sarlette, R. Klein. *Acquisition, Compression, and Synthesis of Bidirectional Texture Functions*. State of the art report, Eurographics (2004), 69–94.

[16] X. Wang, X. Tong, S. Lin, S. Hu, B. Guo, H.-Y. Shum. *View-dependent displacement mapping*. ACM SIGGRAPH 2002 **22**(3), ACM Press (2003), 334–339.

DAQ System for RelaxD Pixel Detector*

Martin Hejtmánek

1st year of PGS, email: hejtmank@fzu.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Václav Vrba, Institute of Physics, AS CR

Abstract. This paper concerns with the RelaxD readout interface for Medipix2 quad chips (four chips which form one unit). Particularly, the development of control software tools is discussed. The RelaxD interface is connected to the PC with the fast 1 GB ethernet link which currently provides readout speed 50 frames per second. Because of its high speed, many software issues related to the processing and storing of obtained data arise. In the last section some results of tests performed at the National Radiation Protection Institute in Prague are presented.

Keywords: pixel detector, Medipix2 quad, readout electronics

Abstrakt. Tento příspěvek se zabývá projektem RelaxD – vývoj vyčítacího rozhraní pro quady Medipix2 (čtyři spojené čipy tvořící jeden celek). Středem pozornosti v tomto článku je vývoj softwaru pro ovládání detektoru. Vzhledem k tomu, že RelaxD je navržen tak, aby umožňoval co nejrychlejší sběr dat pomocí rychlého 1 GB ethernetového spojení (současně 50 snímků za sekundu), je nutno vyřešit mnoho problémů s ukládáním a zpracováním velkého objemu dat. Na závěr je diskutováno jedno z uskutečněných měření ve Státním ústavu radiační ochrany v Praze.

Klíčová slova: pixelový detektor, Medipix2 quad, čtecí elektronika

1 Introduction

This article deals with the pixel detector RelaxD which can detect ionizing radiation and photons by converting ionization in a sensor into electrical signals. Such detectors are commonly used in particle physics and also for medical applications (e.g. X-ray imaging, X-ray computed tomography). The RelaxD (high-**RE**solution **L**arge **A**rea **X**-ray **D**etector) is a complete data readout system for the Medipix2 readout chip. It has been developed at Nikhef[1]. Each RelaxD module can serve one Medipix2 quad (i.e. four Medipix2 readout chips formed in a 2×2 grid).

There are several existing readout interfaces such as Muros2 (Nikhef) or USB interface (IEAP[2]) but the RelaxD is exceptional because of its readout speed. It uses 1 Gb Ethernet connection to the PC which enables the readout of the Medipix2 quad at 50 frames per second (current status, frame rate should be improved in future). This is crucial for many real applications such as high speed imaging of fast processes. For further technical overview on the RelaxD system see [4]. For further details on the Medipix2 readout chip see [5, 3, 2].

*This work has been done in cooperation with Nikhef, Amsterdam
[1]National Institute for Subatomic Physics, Amsterdam, The Netherlands
[2]Institute of Experimental and Applied Physics, Prague, Czech Republic

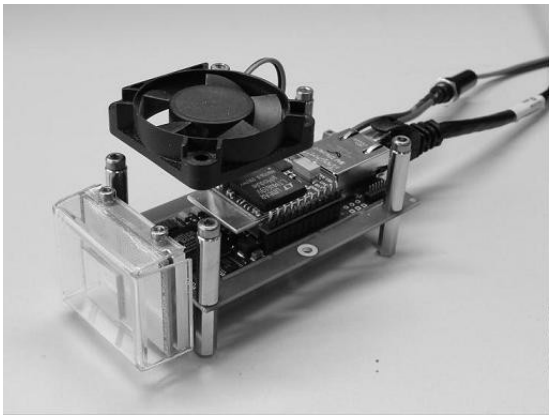


Figure 1: RelaxD module.

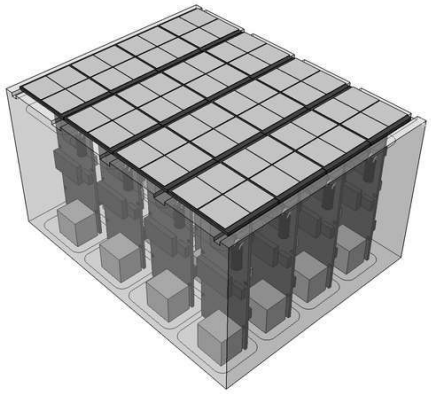


Figure 2: Modules tiled into grid.

1.1 RelaxD module physical layout

As mentioned above, the RelaxD module is designed to achieve the highest frame rate as possible. Since the total sensitive area of Medipix2 chip is relatively small (about 2 cm^2 per chip) it is expected that the modules will be tiled into a 2D grid as shown in figure 2. Therefore the RelaxD is built in a 'T' shape. There is the Medipix2 quad at the front side and readout electronics placed on a PCB (printable circuit board) which is connected to the perpendicular PCB of electronics board. The simplified scheme of the module layout is shown in figure 3.

The core of readout board forms Lattice LFSC15 Field Programmable Gate Array (FPGA) with embedded MICO32 32-bit microprocessor and built-in RAM memory. The FPGA is connected to the 2 blocks of EEPROM memory (electrically erasable programmable read-only memory). The first block serves as a storage for the FPGA controlling program written in the C programming language while the second block is used for storing user specific information such as settings and configurations. When the module is powered up, the FPGA code and user data are loaded into the microprocessor.

The connection to the PC is done via two links. The USB link is used for debugging purposes. The gigabit Ethernet link provides data transmission between the user PC and RelaxD module, i.e. it is used both for controlling of the device and reading out the measured data.

2 RelaxD software utilities

The main task of my stay at Nikhef was to develop and improve PC software applications for controlling RelaxD modules and for further processing of the obtained data. Some of applications were already functional, but since the readout speed of the module is relatively high, there remained still unsolved software challenges. All applications together form a standalone data acquisition (DAQ) system programmed in C++ programming language. For graphical user interfaces (GUI) the Qt framework from Nokia[3] was used. Since all software developing tools are cross-platform, user can use the applications both

[3] see `http://qt.nokia.com/`

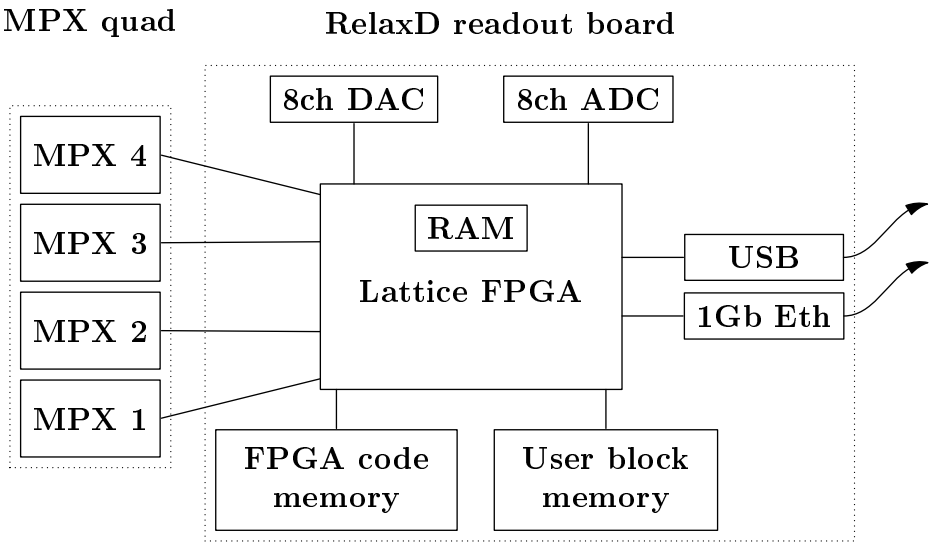


Figure 3: Simplified physical layout of the RelaxD module.

on Linux and Windows operating systems. However, in order not to duplicate effort in Medipix2 project, special features like calibration of the Medipix2 chips is done in already existing program Pixelman from IEAP, Prague.

The complete list of developed applications follows:

`MpxHwRelaxD`
: a library providing a class and an API (Application Programming Interface) for accessing and controlling a RelaxD module; the library is the basis for the other tools in this list; the library also complies to the definition of a so-called 'Pixelman hardware library', allowing the Pixelman program to control and read out multiple RelaxD modules.

`RelaxDaq`
: a program for fast read-out, frame data storage and frame display for up to 4 RelaxD modules, relying on the use of Pixelman for configuration of the modules before starting read-out.

`Convert`
: a command line tool to read the raw data files produced by RelaxDAQ, then decompress, decode and zerosuppress the frames and write out the frame data in an ASCII format more-or-less compatible with what Pixelman produces.

`DacView`
: a program to view, modify and store Medipix2 device DAC settings on the RelaxD module.

`SetId`
: a command line tool to change the IP-address of a RelaxD module.

In following subsections some of my solutions to the problems will be discussed.

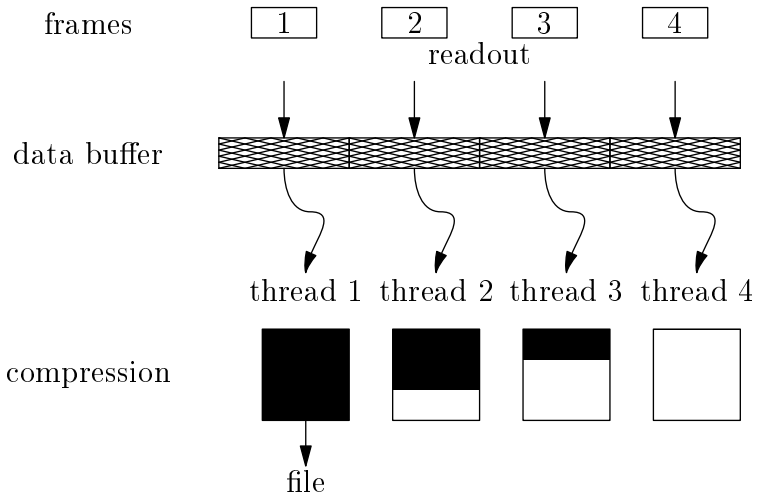


Figure 4: Scheme of data flow during storage of compressed data.

2.1 On-the-fly data compression

As already discussed above, one of the main advantages of the RelaxD is its readout speed. This causes serious problem with storing the obtained data. Normally, in not so fast systems the raw data (i.e. data from every 14-bit pixel register of Medipix2 quad) can be decoded and zero suppressed 'on-the-fly' (during measurement). This procedure becomes impossible in case of the RelaxD due to amount of data – system should be ready to take the information of another frame as soon as readout of previous frame is finished. Therefore the raw data in first approach were stored to computer hard drive 'as they are' without any further processing. This has 2 drawbacks – less comfort for the user, since he does not see the currently taken picture during the measurement and also problems with disk space, since the raw data produce very large data files which are decoded offline. The first drawback is quite easy to solve – every second one frame is decoded and displayed, so the user has some reduced but sufficient overview how his measurement is being performed. The second drawback is quite tricky, though.

As turned out, a compression of the raw data can be performed. Since the framerate is currently 50 frames per second and each frame consists of $512 \times 512 \cdot 14$ bits, the total amount of data to be stored per second is 448 MB. The total time of measurement can easily take minutes or even more, there are typically gigabytes of raw data which should be stored. When using pixel detectors in medical applications, commonly the frames are very sparse $\Rightarrow$ the standard LZ77 compression algorithm developed by Abraham Lempel and Jacob Ziv implemented in Qt framework is suitable to use.

The idea of data compression introduces one more interlink into the readout chain. The frame data are readout from the detector to the readout buffer in a computer RAM memory, then they are compressed and finally stored to a file. Since both compression and writing to disk consumes a significant amount of time, the benefits of multi-threaded programming should be employed. In the final solution, new thread which compresses the data is created for each readout frame. That means that the `RelaxDAQ` application does not wait until the compression of current frame is finished but it is continuously preparing the frames for writing into file. The application is checking if the current frame is already

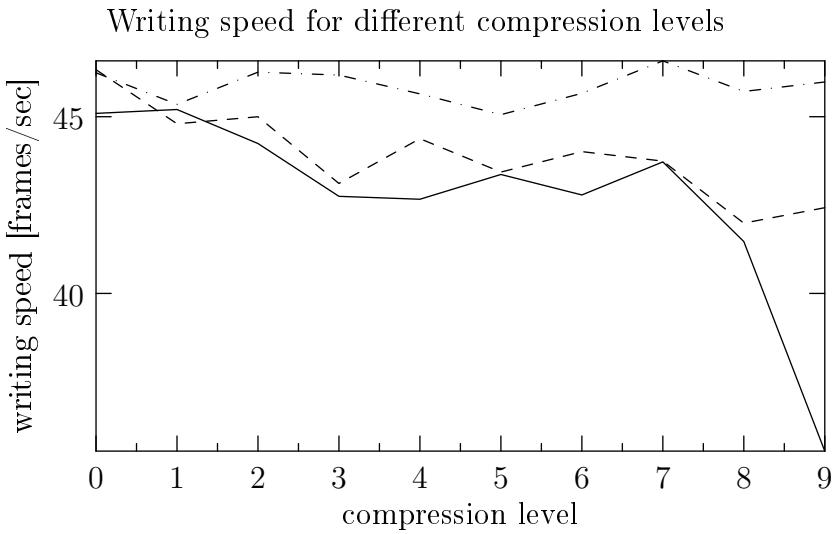


Figure 5: The writing-to-file speed when using different compression levels (0 means no compression, 9 the best but also the slowest compression). The solid line represents frames with 1/2 randomly generated non-zero pixels, the dashed line with 1/32, the dashed-dotted line with 1/512, respectively.

prepared and if it is, it writes the compressed frame immediately to file. This procedure allows to utilize time efficiently, since there is less idle time (when the `RelaxDAQ` does not write to file) because of data compression. However, to perform such solution succesfully, PC processor with multiple cores must be used. Scheme which ilustrates the creating of different compression threads is displayed in figure 4.

For testing purposes was used computer containing quad processor. The obtained writing speed results for different compression levels (quality of compression) and different occupancy of images (sparsity of images) can be seen in figure 5. The achieved file size using different levels of compression is displayed in figure 6. As can be seen, it does not make much sense to use higher compression levels.

2.2 Fast image preview application

In this subsection the application for fast RelaxD image preview will be introduced. As mentioned above, the RelaxD is designed to perform measurements consisting of many frames in sequence. Because of its readout speed, it is impossible to display each frame during the measurement. The data are stored in large files containing up to 1000 images. Processing of the frames is done offline, i.e. after the measurement. The application `EvtDisplay` was made for purposes of fast image browsing.

The `EvtDisplay` simply takes the data files from the `RelaxDAQ` and display its content on the PC screen. One can then easily find which frames are interesting and which frames should be processed further. The application also provides features like zooming of the image, printing it out into PDF, PNG or text file (just data formated in 3 columns). For better display functionality, additional Qt library Qwt[4] for technical GUI

[4]see `http://qwt.sourceforge.net/`

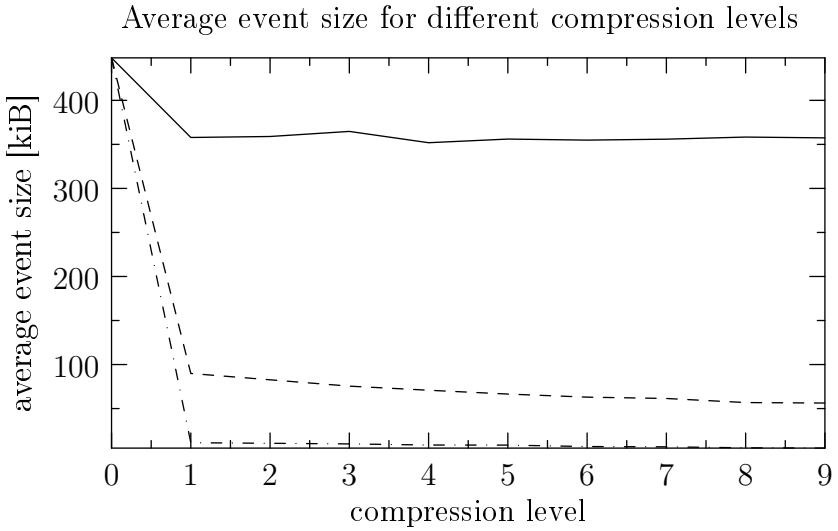


Figure 6: The average event size vs. different compression levels (0 means no compression, 9 the best but also the slowest compression). The solid line represents frames with 1/2 randomly generated non-zero pixels, the dashed line with 1/32, the dashed-dotted line with 1/512, respectively.

building was used. The main application window is displayed in figure 7.

3 Measurements with RelaxD

In the beginning of June, the RelaxD module was tested at the National Radiation Protection Institute in Prague[5]. In these measurements, molybdenum and tungsten X-ray tubes were used. As imaging objects several electronic chips and mammographic phantoms were used. Measurements were done in cooperation with Mária Čarná who describes it in more detail in her diploma thesis [1].

3.1 Imaging of electronic chips

In this measurements we were trying to display the composite structure of electronics chip. Various X-ray energies and intensities were used.

In figure 8 the image of dual inline memory module is shown. As can be seen, inside of the chip is displayed with high contrast. Both copper interconnects and soldered memory chips are clear and sharp. The used voltage of X-ray tube in this case was 40 kV, the current 20 mA and exposition time 1 s.

The second imaged chip is microcontroller MHB8748 from Tesla company (figure 9). The contrast is again very high, and at the top of the circle part on the right side can be even seen small 30 μm golden wire bonds. This image was taken with the voltage 50 kV, the current 10 mA, and exposition time 1 s.

[5]SÚRO – Státní ústav radiační ochrany, v.v.i.

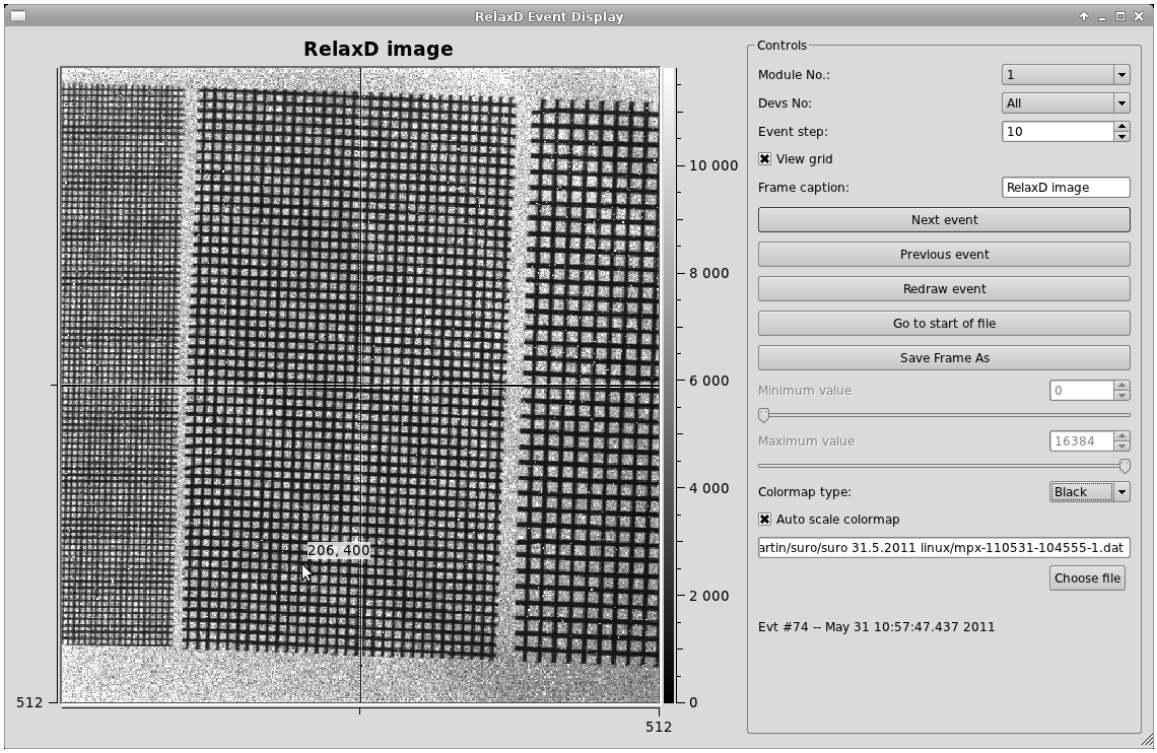


Figure 7: The main `EvtDisplay` application window. On the right are controlling buttons providing additional display features.

3.2 Mammographic phantom

Mammography is one of the most promising fields for application of the Medipix2 photon counting chip. Because the Medipix2 chip is very sensitive to radiation, the radiation doses needed for examination of patient could be significantly reduced.

Mammographic phantom provides the physical standard baseline for assuring the quality of the images produced by mammographic system. In this measurement, one of such phantoms was imaged (from company Kodak-Pathé). The results are displayed in figure 10. In the center of the figure there is an image of the whole phantom as provided in the phantom documentation (image taken with use of digital mammography). The rest are the images obtained in the measurements. Each image corresponds to some part of the phantom. The parameters of X-ray tube were in this case 25 kV, 30 mA, and 2 s.

4 Conclusion

The pixel detectors are very promising in many applications in particle physics as well as in medical imaging. In this work, we discussed control software development for the RelaxD readout interface for Medipix2 pixel detector, particularly some specific software issues cocerning its high readout speed. The PC software tools are necessary for performing measurements, their quality is crucial for the comfort of the user. With help of the developed software tools additional tests investigating the properties of Medipix2 sensors can be done, e.g. the qualities of so called 'edgeless sensors'. In section 3 we discussed one

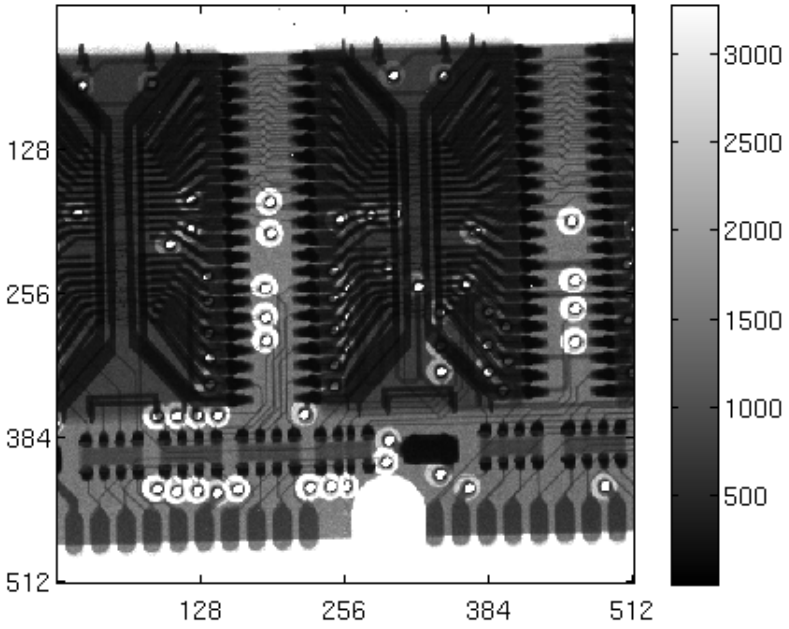


Figure 8: Dual inline memory module image taken by RelaxD detector.

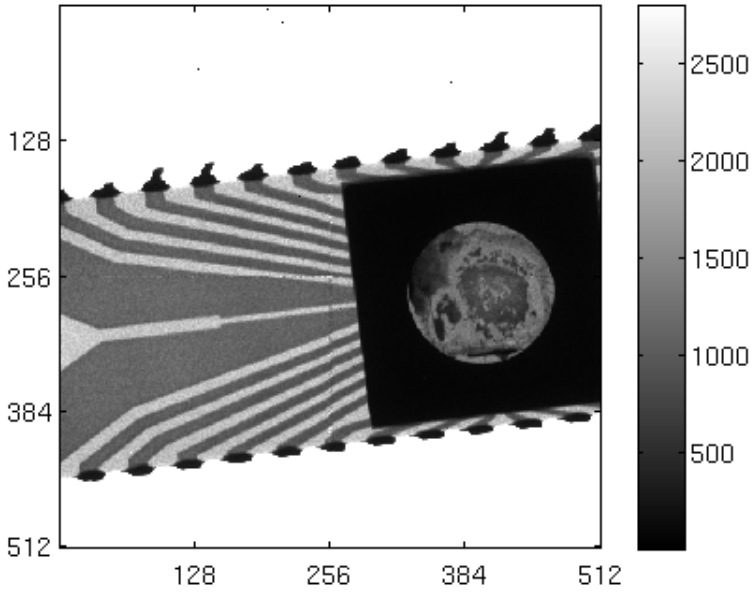


Figure 9: Microcontroller MHB8748 from Tesla company image taken by RelaxD detector.

of such tests, which was done at the National Radiation Protection Institute in Prague in June. These tests proved that the Medipix2 chip is very suitable for mammographic imaging – the obtained images of the testing mammographic phantom were nearly in the same or better quality than images taken by the digital mammograph (see figure 10).

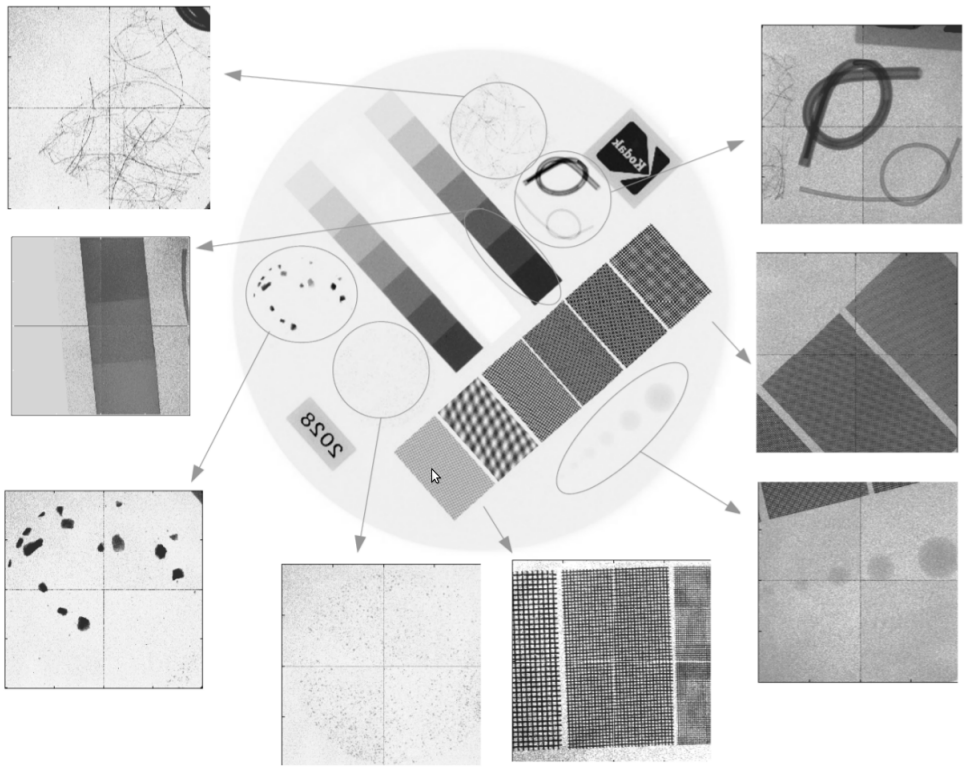


Figure 10: Comparison between the image of mammographic phantom taken with use of digital mammography (in the centre) and images of parts of phantom taken by RelaxD detector.

The developed software tools were introduced and presented at international Medipix2 collaboration meeting at CERN in March.

References

[1] M. Čarná. *Imaging Using Medipix2 Detector.* Diploma Thesis, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague (2011)

[2] X. Llopart. *MPIX2MXR20 Manual v2.3.* Medipix2 Collaboration, `http://medipix.web.cern.ch/MEDIPIX/Medipix2/PasswordProtected/Documents/MXR/Mpix2MXR20Documentv2.3.pdf`

[3] X. Llopart, M. Campbell, R. Dinapoli, D. San Segundo, and E. Pernigotti. *Medipix2: a 64-k Pixel Readout Chip With* $\mathbf{55} - \mu m$ *Square Elements Working in Single Photon Counting Mode.* Medipix2 Collaboration, `http://mcampbel.web.cern.ch/mcampbel/Papers/M7-3-Xavier-Llopart.pdf`

[4] J. Visser et al. *A Gigabit per second read-out system for Medipix Quads.* Nuclear Instruments and Methods in Physics Research A 633 (2011), 22–25

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

Phase-Field Modelling of Heteroepitaxial Growth

Hung Hoang Dieu

4th year of PGS, email: hoangdieu@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 phase-field method has appeared in the context of diffuse interfaces. It has been applied to the three major materials processes: solidification, solid-state phase transformation, and grain growth and coarsening. Very recently, a number of new phase-field models have been developed for modelling heteroepitaxial growth (see [2, 4, 5]). Accurate knowledge of morphological changes in heteroepitaxial thin films is crucial for governing the materials properties. We provide a computational tool based on the finite difference method for study of this phenomenon. Finally, we present our latest results.

Keywords: phase-field method, FDM, heteroepitaxy, ATG instability

Abstrakt. Metoda phase-field se objevila v souvislosti s difuzními rozhraními. Byla aplikována na tyto tři hlavní procesy: tuhnutí, fázový přechod a růst zrn. Řada modelů typu phase-field byla vyvinuta pro modelování heteroepitaxního růstu krystalů (viz [2, 4, 5]). Přesná znalost morfologických změn heteroepitaxního filmu v čase je rozhodující pro nastavení vlastností materiálů. Používáme zde metodu sítí pro simulaci tohoto jevu. Nakonec prezentujeme naše nejnovější výsledky.

Klíčová slova: metoda phase-field, metoda sítí, heteroepitaxe

Introduction

Crystallization is the process where solid crystals are formed from melt, solution, or vapour phase. There are two major stages involved in the crystallization process – *nucleation* and *crystal growth*. Nucleation is the stage where crystal forming units (atoms,

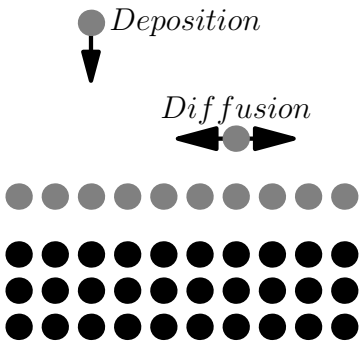
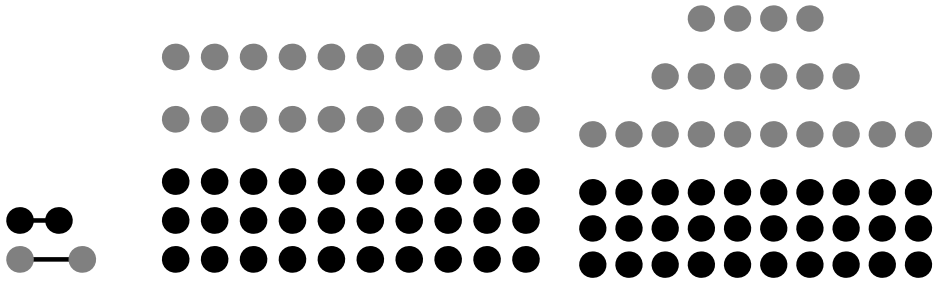


Figure 1: Atomistic view of the basic processes in epitaxy.



(a) High energy. (b) Low energy.

Figure 2: ATG instability.

ions or molecules) gather into clusters which are unstable until they reach a critical size. Stable clusters are called nuclei. After nuclei are created, crystal growth begins. It is the stage where new crystal forming units are incorporated into the crystal lattice. Seed crystals are used to bypass the nucleation stage; thus, the growth can start immediately.

In this contribution we deal with the growth of a thin film of single crystal material on a single crystal substrate so that the film has the same structure as the substrate. Such growth is called epitaxy. Here the substrate functions as a seed crystal. According to the theory of Burton, Cabrera, and Frank [1] atoms are first adsorbed to the crystalline surface where they are called adatoms. Then they diffuse freely along the surface. Finally they can detach from or attach to the crystal (see Fig. 1). The deposited film takes on a lattice structure and orientation identical to those of the substrate.

In general we distinguish two cases: homoepitaxy and heteroepitaxy. In homoepitaxy the film and substrate are made of the same material while in heteroepitaxy the film is made of a material different from the substrate. One example of heteroepitaxy is the growth of germanium film on a silicon substrate. The lattice parameter of the film differs from the substrate (less then 4% for Ge/Si). Hence strains are introduced into the heteroepitaxial film. Due to the effects of stress, the flat film surface is unstable to small perturbations. Heteroepitaxial films undergo a morphological instability, known as the Asaro-Tiller-Grinfeld (ATG) instability.

Figure 2 illustrates the physical mechanism of the ATG instability [3]. The surface tends to remain flat to get lowest surface free energy (Fig. 2a). But, if elastic energy is presented in the film, the corrugated surface has lower elastic energy than the flat one (Fig. 2b). The elastic energy is lowered by elastic deformation so that the film breaks into isolated islands (quantum dots). Each island then forms a quantum dot. Elastic energy is reduced as the surface area increases. Therefore, quantum dots are caused by the competition between surface and elastic energies. Here, Tte mass is transported by surface diffusion. The size of quantum dots is between several and hundreds of nanometers. Quantum dots have interesting electrical and optical properties. Quantum dots are widely used in optical and optoelectronic devices, quantum computation, or biology. Accurate knowledge of morphological changes in epitaxial thin films is crucial for governing the materials properties. Our goal is to provide a suitable computational tool for study of such phenomena.

Very recently, a number of new phase-field models have been developed for modelling heteroepitaxial growth (see [2]). [4, 5, 6] have developed a phase-field approach including

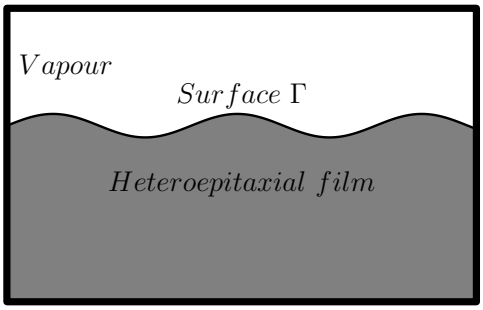


Figure 3: Model.

the stress as an active variable.

Model

We would like to describe the phase-field model of heteroepitaxial growth. We will follow [5] and [4]. Let us consider a system Ω consisting of two regions – a solid heteroepitaxial film $\Omega^f(t)$ and vapour phase $\Omega^v(t)$. The solid-vapour diffuse interface is denoted $\Gamma(t)$, which is a function of time t (see Fig. 3). We introduce a order parameter

$$\Phi(t,\mathbf{x}) \begin{cases} = 0 & \mathbf{x} \in \Omega^v \\ = 1 & \mathbf{x} \in \Omega^f \\ \in (0,1) & \mathbf{x} \in \Gamma \end{cases} .$$

Then the state of the entire system is represented by this order parameter. The total free energy of the system can be written as

$$F[\Phi, \{\epsilon_{ij}\}] = \int_V \left[f(\Phi, \{\epsilon_{ij}\}) + \frac{1}{2}\Gamma\xi^2(\nabla\Phi)^2 \right] dV, \tag{1}$$

where ϵ_{ij} is the strain tensor, V is the volume, ξ is the length parameter controlling the order of magnitude of the transition region described by the phase field. $\Gamma = 3\gamma/\xi$ is the energy density corresponding to the surface energy γ being distributed over a layer of widht $\approx \xi$. The first term represents the sum of the free energies of the film and vapour. The second term describes the interfacial energy.

The total free energy density is given by

$$f(\Phi, \{\epsilon_{ij}\}) = 2\Gamma g(\Phi) + f_{el}(\Phi, \{\epsilon_{ij}\}), \tag{2}$$

where $g(\Phi) = \Phi^2(1-\Phi)^2$ and f_{el} is the elastic energy density.

Here, the linear elastic theory is used. The stress tensor $\sigma_{ij}^{(v)}$ in the vapour is given by Hooke's law

$$\sigma_{ij}^{(v)} = 2\mu^{(v)}\epsilon_{ij} + \lambda^{(v)}\epsilon_{kk}\delta_{ij},$$

where einstein summation convention is implied.

Following [8] the stress tensor $\sigma_{ij}^{(f)}$ in the heteroepitaxial film is given by

$$\sigma_{ij}^{(f)} = 2\mu^{(f)}\epsilon_{ij} + \lambda^{(f)}\epsilon_{kk}\delta_{ij} - \epsilon^m\{\tfrac{1+\nu^{(f)}}{1-2\nu^{(f)}}\}\delta_{ij},$$

where $\mu^{(*)}, \lambda^{(*)}$ are Lamé constants, $\nu^{(*)}$ is Poisson's ratio, where $* \in \{f, v\}$. $\epsilon^m = \frac{a_f - a_s}{a_s}$ is the misfit strain, where a_f, a_s are lattice constants of epitaxial film or substrate. The strain tensor is given by $\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$, where u_i is the ith component of the displacement vector.

A straightforward ansatz for the elastic energy density is then

$$f_{el}(\Phi, \{\epsilon_{ij}\}) = h(\Phi)\{(\mu^{(f)} - \mu^{(v)})\epsilon_{ij}\epsilon_{ij} + \frac{\lambda^{(f)} - \lambda^{(v)}}{2}(\epsilon_{ii})^2 - \frac{1+\nu^{(f)}}{1-2\nu^{(f)}}(\epsilon^m)^2\}, \quad (3)$$

where $h(\Phi) = \Phi^2(3 - 2\Phi)$ may be interpreted as a "solid fraction" which must be equal to one in the solid and equal zero in the vapour.

The stress tensor in the system is determined from

$$0 = \frac{\partial}{\partial x_j}\{h(\Phi)\sigma_{ij}^{(e)} - [1 - h(\Phi)]\sigma_{ij}^{(v)}\}, \quad (4)$$

where $h(\Phi) = \Phi^2(3 - 2\Phi)$ is the weight function for the epitaxial layer.

Assuming relaxational dynamics, the equation of motion takes the form

$$\frac{\partial \Phi}{\partial t} = -R\frac{\delta F}{\delta \Phi}, \quad (5)$$

and the prefactor R should contain the mobility $1/k$. We choose $R = 1/(3k\rho_f\xi)$.

We arrive at

$$\begin{aligned}\partial_t\Phi = {} & A\Delta\Phi + \frac{B}{\xi^2}g'(\Phi) \\ & + \frac{C}{\xi}h'(\Phi)\{(\mu^{(f)} - \mu^{(v)})\epsilon_{ij}\epsilon_{ij} + \frac{\lambda^{(f)} - \lambda^{(v)}}{2}(\epsilon_{ii})^2 \\ & - \frac{1+\nu^{(f)}}{1-2\nu^{(f)}}(\epsilon^m)^2\},\end{aligned} \quad (6)$$

where A, B, C are constants, $g'(\Phi) = 2\Phi(1 - \Phi)(1 - 2\Phi)$, and $h'(\Phi) = 6\Phi(1 - \Phi)$.

Numerical scheme

We use an explicit scheme of the finite difference method to solve the free boundary problem of spiral crystal growth. The first step in the discretization is to divide the computational domain into a two-dimensional grid and then derivatives are replaced with equivalent finite differences.

We consider the computational domain S to be a rectangle $(0, L_1) \times (0, L_2)$ which is to be discretized. We partition the domain S using a grid of internal nodes $\omega_h = \{(ih_1, jh_2)|i = 1, ..., N_1 - 1, j = 1, ..., N_2 - 1\}$, where $h_1 = \frac{L_1}{N_1}, h_2 = \frac{L_2}{N_2}$ are the mesh sizes in S. We discretize the time interval using a mesh $[0, T] : T_\tau = \{k\tau|k = 0, ..., N_T\}$, where $\tau = \frac{T}{N_T}$ is a time step. Then we can consider a grid function $u : T_\tau \times \omega_h \to \mathbb{R}$ for which $u_{ij}^k = u(ih_1, jh_2, k\tau)$.

The time derivative is approximated by forward difference

$$\partial_t u_{ij}^k \approx \frac{u_{ij}^{k+1} - u_{ij}^k}{\tau},$$

and the space derivatives are approximated by second-order central differences:

$$\partial_x^2 u_{ij}^k \approx \frac{u_{i+1,j}^k - 2u_{ij}^k + u_{i-1,j}^k}{h_1^2},$$
$$\partial_y^2 u_{ij}^k \approx \frac{u_{i,j+1}^k - 2u_{ij}^k + u_{i,j-1}^k}{h_2^2}.$$

Then the Laplace operator in two dimensions is given by $\Delta_h u^k{}_{ij} \approx \partial_x^2 u_{ij}^k + \partial_y^2 u_{ij}^k$.

Finally we obtain this explicit scheme

$$\begin{aligned}\Phi_{ij}^{k+1} &= \Phi_{ij}^k \\ &+ \tau A \frac{\Phi_{i+1,j}^k + \Phi_{i,j+1}^k - 4\Phi_{ij}^k + \Phi_{i,j-1}^k + \Phi_{i-1,j}^k}{h^2} + \frac{\tau B}{\xi^2} g'(\Phi) \\ &+ \frac{\tau C}{\xi} h'(\Phi)\{(\mu^{(e)} - \mu^{(v)})\epsilon_{ij}\epsilon_{ij} + \frac{\lambda^{(e)} - \lambda^{(v)}}{2}(\epsilon_{ii})^2 \\ &- \frac{1 + \nu^{(e)}}{1 - 2\nu^{(e)}}(\epsilon^m)^2\},\end{aligned} \tag{7}$$

for $i = 1, ..., N_1 - 1, j = 1, ..., N_2 - 1, k = 0, ..., N_T$. That means we can obtain the values at time $k + 1$ from the corresponding ones at time k.

The boundary conditions are treated by mirroring the values in the inner nodes across the boundary.

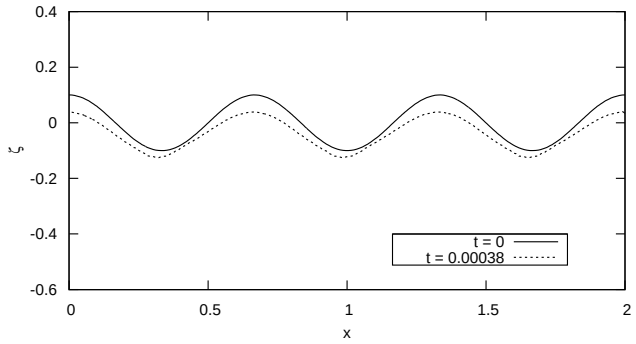
Numerical results

We implemented the model using the explicit scheme based on FDM for the phase-field equation (6). In numerical experiments, we used the rectangular domain $\Omega \equiv (0, 2)\times(0, 1)$ with the grid 200×100. The spatial step size in x-direction is set to $h_1 = 2/199$ and the spatial step size in y-direction is set to $h_2 = 1/99$. The other model parameters are as follows: $A = 0.005$, $B = -0.01$, $C = -0.00333$, $\xi = 0.015$, and time step $\tau = A{*}h_1{*}h_1/8$. The initial conditions for the phase-field variable are given by

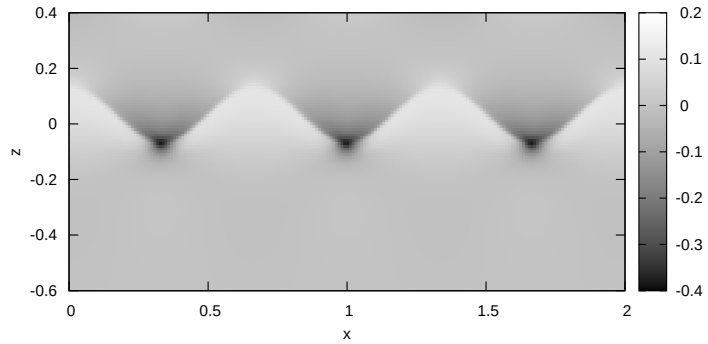
$$\Phi(x, y) = 0.5\left(tanh\left(\frac{1}{h_1}(0.1cos(3\pi x) - y)\right) + 1\right).$$

For the elastic problem, we used FreeFem++ [7] based on FEM. Material dimensionless parameters are taken as follows: $E^{(v)} = 1$, $\nu^{(v)} = 0$, $E^{(e)} = 1 \times 10^7$, $\nu^{(v)} = 0.278$, $\epsilon^m = 0.05$, $\mu = E/(2.0(1.0 + \nu))$, $\lambda = E\nu/((1.0 + \nu)(1.0 - 2.0 * \nu))$. Computations of stress field were very time consuming. Fig. 4b shows x-component of normal strain at t = 0.00038.

We observed the valleys of the surface profile deepen under stress. However, the tops deepen as well, even at higher speed (see Fig. 4a). Our study is at early stage and it is not obvious from the experiments whether it can lead to fracture or it evolves towards the planar interface. We also found that numerical noise avoid us to simulate the problem in longer time. Therefore, it is necessary to develop better numerical schemes suitable for modelling heteroepitaxial growth.



(a) Evolution of the interface. ζ is the interface position, given by its z coordinate.



(b) Normal strain in x-direction at $t = 0.00038$.

Figure 4: Numerical results.

References

[1] W. K. Burton, N. Cabrera, and F. C. Frank. *The growth of crystals and the equilibrium structure of their surfaces.* Phil. Trans. R. Soc. **243** (June 1951), 299–358.

[2] L. Q. Chen. *Phase-field models for microstructure evolution.* Annual Review of Materials Research **32** (2002), 113–140.

[3] D.J. and Srolovitz. *On the stability of surfaces of stressed solids.* Acta Metallurgica **37** (1989), 621 – 625.

[4] H. Emmerich. *The diffuse interface approach in materials science: thermodynamic concepts and applications of phase-field models.* Lecture notes in physics: Monographs. Springer-Verlag, (2003).

[5] K. Kassner, C. Misbah, J. Müller, J. Kappey, and P. Kohlert. *Phase-field modeling of stress-induced instabilities.* Phys. Rev. E **63** (Feb 2001), 036117.

[6] J. Müller and M. Grant. *Model of surface instabilities induced by stress.* Phys. Rev. Lett. **82** (Feb 1999), 1736–1739.

[7] O. Pironneau, F. Hecht, A. L. Hyaric, and J. Morice. Freefem++. http://www.freefem.org/ff++/.

[8] B. J. Spencer, P. W. Voorhees, and S. H. Davis. *Morphological instability in epitaxially strained dislocation-free solid films.* Phys. Rev. Lett. **67** (Dec 1991), 3696–3699.

Analysis of Microstructure of the Totally Asymmetric Simple Exclusion Process with Respect to Traffic Flow Modeling*

Pavel Hrabák

2nd year of PGS, email: hrabapav@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 Science and Physical Engineering, CTU in Prague

Abstract. This article serves as a summary of results published in articles [1] and [2], and several newest results which will be published in the Proceedings of the conference SPMS 2011 [3]. The main goal of the articles is to investigate the characteristics of the TASEP model, used for describing the microstructure of the traffic flow, i.e., the distance- and time- headway distribution. Using the matrix product Ansatz the exact formula for distance-headway distribution far from and close to the boundary has been derived. By means of this result the exact formula for the time-headway distribution is obtained.

Keywords: TASEP, distance-headway distribution, time-headway distribution.

Abstrakt. Tento článek slouží jako shrnutí výsledků publikovaných v článcích [1] a [2] a několika nejnovějších výsledků, které budou publikovány ve sborníku konference SPMS 2011 [3]. Cílem těchto článků je vyšetřit charakteristiky modelu TASEP, které jsou užívány k popisu mikrostruktury dopravního proudu, t.j. délkové a časové rozestupy. Užitím matrix product Ansatz je odvozen vzorec pro délkové rozestupy daleko od hranice i poblíž hranice. Na základě těchto výsledků je odvozen vztah pro časové rozestupy.

Klíčová slova: TASEP, délkový rozestup, časový rozestup.

1 Introduction

The totally asymmetric simple exclusion process (TASEP) is an interacting particle system defined on the one-dimensional lattice consisting of N cells. The particles are moving along the lattice in one direction by hopping to the neighboring cell according to following rules: New particle enters the system by hopping to the first cell with probability α, if the target cell is empty, a particle in the system hops to the neighboring cell with probability p, if the target cell is empty, the particle leaves the system by hopping out of the last cell with probability β.

Considering the system in stationary state (non-equilibrium) it has been proven that

*This work has been supported by the grant SGS10/209/OHK4/2T/14

the probability of finding the system in the configuration τ can be written as

$$P_N(\tau_1, \tau_2, \ldots, \tau_N) = Z_N^{-1} \langle W | \prod_{i=1}^{N} [\tau_i D + (1 - \tau_i) E] \, |V\rangle \,, \tag{1}$$

where E, D are square matrices and W, V vectors for which holds

$$DE = D + E \,, \quad \langle W|E = \frac{1}{\alpha} \langle W| \,, \quad D|V\rangle = \frac{1}{\beta} |V\rangle \,. \tag{2}$$

2 Results

Using the configuration distribution (1) and the rules (2) it has been shown in the article [2] that the distance-headway probability density, i.e., the probability of finding a gap of $k-1$ empty cells between two successive particles, is of the form

$$\wp(k; \alpha, \beta) = \begin{cases} \frac{1}{2^k} & \alpha \geq \frac{1}{2} \ \wedge \ \beta \geq \frac{1}{2} \,, \\ \alpha(1-\alpha)^{k-1} & \alpha < \frac{1}{2} \ \wedge \ \beta > \alpha \,, \\ (1-\beta)\beta^{k-1} & \beta < \frac{1}{2} \ \wedge \ \beta < \alpha \,, \\ \alpha\beta^{k-1} & \alpha + \beta = 1 \,, \end{cases} \tag{3}$$

when investigating the system far from the boundary. Concerning the headway distribution near the right boundary we obtain

$$\wp_0(k; \alpha, \beta) = \begin{cases} \dfrac{1}{\beta 2^k} \left(1 - \dfrac{k}{2} + \beta(k-1) \right) & \alpha \geq \frac{1}{2} \ \wedge \ \beta \geq \frac{1}{2} \,, \\ \dfrac{\alpha(1-\alpha)^k}{\beta} \left(\dfrac{\beta - \alpha}{1 - 2\alpha} \right) + \dfrac{\alpha^k (1-\alpha)}{\beta} \left(1 - \dfrac{\beta - \alpha}{1 - 2\alpha} \right) & \alpha < \frac{1}{2} \ \wedge \ \beta > \alpha \,, \\ (1-\beta)\beta^{k-1} & \beta < \frac{1}{2} \ \wedge \ \beta < \alpha \,, \\ \alpha\beta^{k-1} & \alpha + \beta = 1 \,. \end{cases} \tag{4}$$

Using the formula (3) the time-headway distribution, i.e., the distribution of time interval Δt between the passings of two successive particle through certain reference cell, can be derived. By means of the *random-sequential discrete-time update* the step-headway probability density far from the boundaries with the bulk density ϱ can be obtained as

$$\begin{aligned} \bar{f}_N(k) = {} & \frac{1}{\rho\sigma} \left[\left(1 - \frac{1}{N} \right)^{k-1} - \left(1 - \frac{\rho}{N} \right)^{k-1} \left(1 - \frac{\sigma}{N} \right)^{k-1} \right] + \\ & + \frac{\rho}{\sigma N} \left[\left(1 - \frac{\rho}{N} \right)^{k-1} - \left(1 - \frac{1}{N} \right)^{k-1} \right] + \frac{\sigma}{\rho N} \left(1 - \frac{\sigma}{N} \right)^{k-2} \left[1 - \left(1 - \frac{\rho}{N} \right)^{k-1} \right] \end{aligned} \tag{5}$$

for $k \geq 2$ and $f_N(1) = 0$, where$\sigma = 1 - \varrho$. To obtain the probability density $f_\varrho(t)$ for the time continuous dynamics, we will proceed as follows: The corresponding step-headway distribution function $F_N(t)$ will be calculated as

$$F_N(t) = \sum_{k=1}^{k(N)} f_N(k) \,, \tag{6}$$

where $\frac{k(N)}{N} \leq t < \frac{k(N)+1}{N}$. Using the large N limit we obtain the limiting distribution function in the form

$$F(t) = \frac{1}{\sigma}\left(1 - e^{-\rho t}\right) - \frac{\rho}{\sigma}\left(1 - e^{-t}\right) + \frac{1}{\rho}\left(1 - e^{-\sigma t}\right) - \frac{\sigma}{\rho}\left(1 - e^{-t}\right) + e^{-t}(1+t) - 1 \quad (7)$$

and the corresponding probability density function reads

$$f(t) = \frac{\rho}{\sigma}\left(e^{\sigma t} - 1\right)e^{-t} + \frac{\sigma}{\rho}\left(e^{\rho t} - 1\right)e^{-t} - te^{-t}\,. \quad (8)$$

3 Conclusion

Studying carefully the probability density function (8) we notice that the function is symmetrical against the exchange of ϱ and σ. This corresponds to the "particle-hole symmetry" of the system, i.e., the symmetry of particles moving in one direction and holes in the opposite direction. This symmetry is undesirable for traffic flow models for it does not appear in the real traffic.

The future goal of our research is to present such long-ranged interaction between particles, which breaks the symmetry of particles and holes. The dependance of the hopping probability λ on the distance-headway to the previous particle d is under investigation. Using the dependance

$$\lambda(d) = \frac{1 - p^{\min\{d, d_{\max}\}}}{1 - p^{d_{\max}}} \quad (9)$$

better agreement with the realistic traffic has been observed using numerical simulations. We aim to derive the headway distribution for the long-ranged model analytically and to extract the proper form of the dependance $\lambda(d)$ from the realistic traffic data.

References

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

[2] M. Krbálek, P. Hrabák. *Inter-particle gap distribution and spectral rigidity of totally asymmetric simple exclusion process with open boundaries.* Journal of Physics A: Mathematical and General **44** (2011), 175 203–175 224.

[3] P. Hrabák. *Time-headway distribution of Totally Asymmetric Exclusion Process with Nearest-Particle Interaction.* Proceedings of SPMS 2011 (2011), *not puplished yet, accepted.*

Digital Morphology of 3D Image in Dodecahedral Topology

Václav Hubata-Vacek

1st year of PGS, email: v.hubata@seznam.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. This paper deals with image segmentation and its use in recognizing Alzheimer's disease. Segmentation is performed by using the morphological method called watershed, in the image in the dodecahedral topology. The following describes the technique for converting the image into dodecahedral topology and the method for image filtering and edge detection. Finally, these methods are tested on the human brains, which are obtained by Single-Photon Emission Computed Tomography (SPECT).

Keywords: watershed, dodecahedral topology, Alzheimer's disease

Abstrakt. Článek se zabývá segmentací obrazu a jejího využití při rozpoznávání Alzheimerovy choroby. Segmentace se provádí pomocí metody rozvodí na obrazu v dodekaedrické topologii. Dále je popsán postup pro převedení obrazu do dodekaedrické topologie a metoda pro filtraci obrazu a detekci hran. V závěru jsou tyto metody testovány na snímcích mozků získaných pomocí jednofotonové emisní výpočetní tomografie (SPECT).

Klíčová slova: metoda rozvodí, dodekaedrická topologie, Alzheimerova choroba

1 Introduction

Alzheimer's disease, which is characterised by loss of neurons and their's synapses, is the most common form of dementia. This disease is still incurable by modern medicine, but it can be slowed down. Therefore time of recognition of diseased patient has highest priority.

SPECT is a technique using gamma rays to provide 3D imaging. Before the technique begins, an injection of radionuclide is introduced into the bloodstream of the patient. The result of this technique is the set of 2D slices of radionuclide distribution in the brain from which the final image is built.

This work is based on hypothesis that the brain scans of patients with Alzheimer's disease differ from the brain scans of healthy people. These changes are observed using watershed, which is the method for image segmentation based on morphological understanding the image and modeling of gradual flooding of virtual terrain. Moreover, watershed is implemented in dodecahedral topology to eliminate its sensitivity to the number of neighboring voxels. We calculate four main characteristics for each image: number of regions ($|r|$), volume of regions ($\sum r$), cardinalty of wateshed shapes ($\sum w$) and average volume of region ($\bar{r}$). The results of the measurement are summarized in the paper.

2 Dodecahedral Topology of 3D Images

Each 3D image is represented by a finite set of points in the computer. These points are generated by using three rectangular vectors of the same length in the cubic topology. Consequently, every voxel has 26 neighbors. Given that the neighbors vary in distance from the central voxel, this distribution of the voxels can cause a problem for the methods that work within the neighborhoods of the central voxel.

The points in the dodecahedral topology are generated by vectors of same length with mutual angles equal to $\frac{\pi}{3}$. A sample of three vectors (1) is used in foloving calculations. Based on this equation, the image can still be represented by the 3D matrix in the computer. Unlike the cubic topology, the image in dodecahedral topology consists of voxels in the shape of a rhombic dodecahedron. Every dodecahedral voxel has 12 neighbors, but distances from the central voxel to each neighboring voxel are equal.

$$\vec{a} = \begin{pmatrix} \cos\frac{1}{12}\pi \\ \sin\frac{1}{12}\pi \\ 0 \end{pmatrix}, \vec{b} = \begin{pmatrix} \cos\frac{5}{12}\pi \\ \sin\frac{5}{12}\pi \\ 0 \end{pmatrix}, \vec{c} = \begin{pmatrix} \frac{1}{\sqrt{3}}\cos\frac{\pi}{4} \\ \frac{1}{\sqrt{3}}\sin\frac{\pi}{4} \\ \frac{\sqrt{6}}{3} \end{pmatrix} \tag{1}$$

The conversion of 3D image (rectangular) input into the dodecahedral topology is performed via linear interpolation. Positions for neighboring voxels are show in Fig. 1.

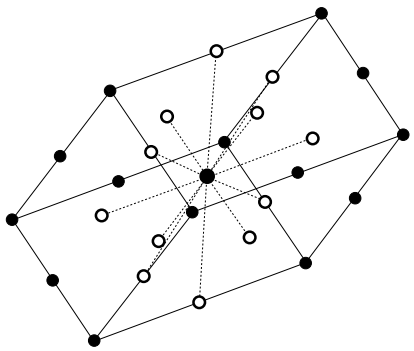


Figure 1: Neighbors in dodecahedral topology

Final representation for the computer in the form of a 3D matrix is shown in Fig. 2.

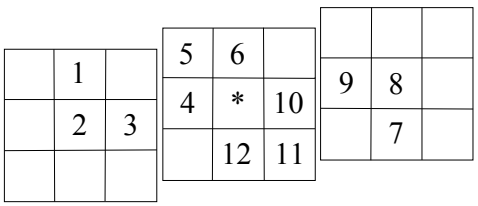


Figure 2: Neighbor representation in matrix

3 Edge Detection

We used linear filters to reduce the noise level and detect the edges. In this paper we discuss the best filter we tested: the DoG (Difference of Gaussian) filter. This method is based on the subtraction of two blurred images, where the images are blurred with a Gaussian kernel with a different parameters σ_1, σ_2 using the convolution.

$$F_1(x, y, z) = g_{\sigma_1}(x, y, z) * f(x, y, z) \tag{2}$$

$$F_2(x, y, z) = g_{\sigma_2}(x, y, z) * f(x, y, z) \tag{3}$$

We can calculate the convolution of the kernel, but we can only use the convolution once because the it is distributive. This operation provides the same results two times faster.

$$F_1(x, y, z) - F_2(x, y, z) = (g_{\sigma_1}(x, y, z) - g_{\sigma_2}(x, y, z)) * f(x, y, z) \tag{4}$$

4 Dodecahedral watershed

Watershed is a morphological method based on the modeling of a gradual flooding of virtual terrain. It starts in each local minimum to flood and build barriers (watershed shapes) in voxels where different domains join. This method gives us an image that is divided into regions.

The algorithm goes from the lowest level of grey to the highest level of grey. In each step, the Algorithm finds all coherent areas of a given level of grey and marks them depending on the following three conditions:

1. If it is touching just one area, then it will join to this area.
2. If it is touching more than one area, or it is touching only a barrier, then it will join to the barrier.
3. If it is touching neither area nor barrier, then it becomes a new area.

5 Matlab realization

There have been several documents written on the functions for loading, converting from cubic to dodecahedral topology, filtering, and rendering.

As mentioned before, we made the conversion from cubic to dodecahedral topology using linear interpolation. In this case, we used the Matlab function `interp3`. To use `interp3` correctly, we cannot rotate the cubic image; therefore, we rotated the future dodecahedral image in the opposite direction and placed the cubic image in the correct position in the space.

Filtering was performed via linear filters. To speed up the algorithm, we used the Fourier transform for convolution and the Matlab functions: `fftn`, `ifftn`, `fftshift`.

Since the cut in the image is composed of regular hexagons in dodecahedral topology, we could not use the standard Matlab function for renderig. Instead, we put together regular hexagons using function `fill` to make the final image.

6 Results

The main statistical properties for the DoG filter are similar to an arithmetic mean ($\bar{x}$), standard deviation (s), and coefficient of variation (γ) are collected in the Tab. 1 for groups of diseased and healthy patients. Our results provide a comparision between two sets of samples: ADT (set of testing samples of diseased brain scans), and CNT (set of testing samples of healthy brain scans).

Table 1: Properties for filter DoG ($\sigma_1 = 1.6$, $\sigma_2 = 3.6$)

set	stat.	$\lvert r \rvert$	$\sum r$	$\sum w$	$\bar{r}$
ADT	$\bar{x}$	3792	314280	224792	83.27
	s	265	10265	7984	6.78
	γ	0.0698	0.0327	0.0355	0.0814
CNT	$\bar{x}$	3439	335533	211654	98.03
	s	241	14934	6839	8.66
	γ	0.0700	0.0445	0.0323	0.0884

Before the interpolation from cubic to dodecaedric topology, the cubic image can be rotated. Sensitivity to the rotation is documented in the Tab. 2 for typical 3D scan. As shown, the rotation has only an insignificant effect on the final results.

Table 2: Rotation test for filter DoG ($\sigma_1 = 1.6$, $\sigma_2 = 3.6$)

stat.	$\lvert r \rvert$	$\sum r$	$\sum w$	$\bar{r}$
$\bar{x}$	4191	304756	238213	72.74
s	70	3261	1585	1.91
γ	0.0167	0.0107	0.0067	0.0262

The next property that we tested was shift resistance. The results for this testing are collected in the Tab. 3 for the typical 3D scan. In this case, the results are slightly better than for the rotation test, which means that the shift has only an insignificant effect as well.

Table 3: Shift test for filter DoG ($\sigma_1 = 1.6, \sigma_2 = 3.6$)

stat.	$\|r\|$	$\sum r$	$\sum w$	$\bar{r}$
$\bar{x}$	4184	304498	237714	72.78
s	59	2176	963	1.12
γ	0.0140	0.0071	0.0041	0.0154

The DoG filter was tested on the set of test samples (ADT and CNT). The results of this testing using two sample Students‘ t-test are shown in the Tab. 4 on the "training" line. In addition, the DoG filter was verified on the set of verification samples (ADV and CNV). The results for the verification samples are provided on the "verification" line. While the results for the verification samples are worse than the results for the test samples, they are still significant, with a probability of Type 1 error of about 1% for all characteristics.

Table 4: Results for filter DoG ($\sigma_1 = 1.6$, $\sigma_2 = 3.6$)

	p-value			
set	$\|r\|$	$\sum r$	$\sum w$	$\bar{r}$
training	0.0060	0.0016	0.0009	0.0005
verification	0.0121	0.0110	0.0096	0.0065

The difference between typical AD and CN patients is ilustrated in the Fig. 3. As seen, the watershed image of an patient with Alzheimer's disease is divided into more regions.

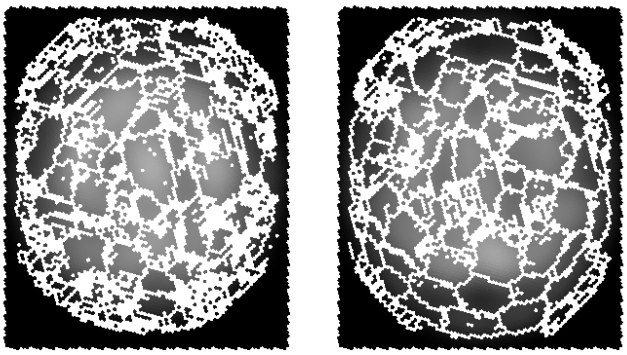


Figure 3: Images after segmentation: AD (left), CN (right)

7 Conclusion

Dodecahedral topology of 3D image is a useful structure for the digital diagnosis of Alzheimer's disease. Optimal parameter for DoG filter and watershed procedure were obtained. The watershed shape volume ($\sum w$) is the most robust measure related to rotation an shifting of the original image (patient) and second best in p-value for the classification of Alzheimer's diseased patient.

References

[1] M. Šebest. *Digitálna morfológia v hexagonálnej mriežke v Matlabe.* Bakalářská práce, FJFI, 2006.

[2] V. Cížek. *Diskrétní Fourierova transformace a její použití.* Diskrétní Fourierova transformace a její použití, SNTL 1981

[3] K. Haris, et al. *Hybrid Image Segmentation using Watersheds and Fast Region Merging*, IEEE Trans Image Processing, 7(12), 1684-1699, 1998.

[4] V. Grau, A. U. J. Mewes, M. Alcańiz *Improved Watershed Transform for Medical Image Segmentation Using Prior Information.* IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 23, NO. 4, April 2004

[5] G. Bertrand. *On topological watersheds*, Journal of Mathematical Imaging and Vision, Vol. 22, No. 2-3, pp. 217-230, 2005.

[6] M. Couprie and L. Najman and G. Bertrand. *Quasi-linear algorithms for the topological watershed*, Journal of Mathematical Imaging and Vision, Vol. 22, No. 2-3, pp. 231-249, 2005.

[7] K. Thangavel, R. Manavalan, I. Laurence Aroquiaraj *Removal of Speckle Noise from Ultrasound Medical Image basek on Special Filters.* ICGST-GVIP Journal, ISSN 1687-398X, June 2009.

[8] L. Najman and M. Couprie and G. Bertrand. *Watersheds, mosaics and the emergence paradigm*, Discrete Applied Mathematics, Vol. 147, No. 2-3, pp. 301-324, 2005.

[9] Y. H. Chai, L. Q. Gao, and S. Lu. *Wavelet-based Watershed for Image Segmentation Algorithm.* 6th World Congress on Intelligent Control and Automation, June 2006.

ATLAS DAQ-system for FE-I4

Zdenko Janoška

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

advisor: Václav Vrba, Institute of Physics, AS CR

Abstract. For the improvement of the present ATLAS Pixel detector, implementation of the new FE-I4 sensors is required. This implementation includes changes in the present DAQ-software as well as development of the hardware. The main aim is to make a system in which present FE-I3 modules and FE-I4 modules can co-exist.

Keywords: LHC, ATLAS, Pixel detector, Upgrade, DAQ, FE-I4

Abstrakt. Pre vylepšenie stávajúceho Pixelového detektoru v experimente ATLAS sa vyžaduje implementácia nového FE-I4 čipu. Implementácia zahŕňa zmeny nielen v stávajúcom DAQ-software ale súčasne je nutná aj úprava hardwaru. Hlavným cieľom je navrhnúť systém, kde dnešné FE-I3 moduly budú koexistovať s novými modulmi FE-I4.

Kľúčové slová: LHC, ATLAS, Pixel detektor, Rozšírenie, DAQ, FE-I4

1 Introduction

This article describes the data acquisition system of ATLAS Pixel detector and implementation of the new FE-I4 Si-detector to this system. It is an overview of the data acquisition chain from both sides - software and also hardware. The crucial point is the library which is responsible for proper work of the Pixel modules in the DAQ-system.

2 ATLAS Pixel Detector

The Pixel Detector is divided into three barrel layers in the center and three disks on both either sides for the forward direction. Due to its close distance to the beam pipe, the Pixel Detector faces the highest amount of particle flux, corresponding to the largest radiation damage and hit occupancies in ATLAS. Innermost barrel layer is the most occupied layer with occupancy approx. $5 \cdot 10^{-4}$ per $50 \times 400 \mu m^2$ area.
The Pixel Detector consists of amount of pixel modules including a single silicon sensor. The sensor is connected to 16 front-end FE-I3 chips using bump bonding. The FE chip is designed to digitise the charge signal received from the sensor pixels. The present version of FE chip (FE-I3) contains 2880 individual charge sensitive analogue circuits with a digital read-out cell. The chip is organised into 18 columns by 160 rows, so such that two columns are combined into pairs for the digital readout. As interface between FE-I3 chips and off-detector read-out system is Module Control Chip (MCC). The MCC is responsible for distributing commands to FE chips and collecting data from them.
For the future upgrade of ATLAS Pixel Detector, FE-I4 chip has been developed. The

FE-I4 integrated circuit contains readout circuitry for 26 880 hybrid pixels arranged in 80 columns on 250μm pitch by 336 rows on 50μm pitch. Many of the specifications and features of FE-I4 have been derived from the FE-I3 chip, but FE-I4 offers a lots of advantages:

- Much cheaper module manufacture ⇒ chip size as big as possible
- Greater fraction of the footprint devoted to pixel array ⇒ move the memory inside the array
- Lower power ⇒ don't move the hits around unless they are triggered
- Able to take higher hit rate ⇒ store the hits locally and distribute the trigger
- Still able to resolve the hits at higher rate ⇒ smaller pixels and faster recovery time
- No need for extra control chip ⇒ significant digital logic blocks on array periphery

The module with planar sensor consists of two front end chips and one sensor [1] and does not contain any MCC chip. Due to this a new protocol have been developed that is not fully compatible with the existing detector DAQ hardware. However, it is still compatible with the optical hardware used in the present detector. It is necessary to implement changes to DAQ software as well. Nevertheless, the command structure has been kept the same and "fast" commands (lower number of bits needed to transmit) are identical. The FE-I4 output is significantly different from the present detector mainly because of the requirement of capability of 160 Mb/s with the expected hit rate at 100 kHz trigger [1].

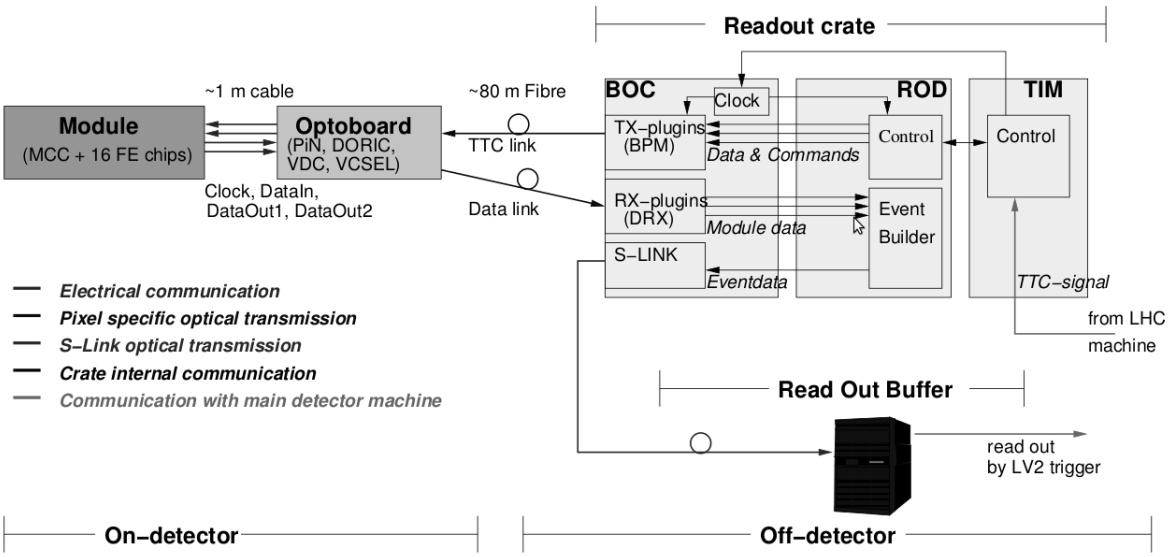


Figure 1: Readout chain of the pixel detector [5].

The figure 1 describes the readout chain of the pixel detector. The Read Out Driver (ROD) is a board designed to interface the detector specific readout components (optical interface, buffers) with the standard ATLAS DAQ chain consists of Timing, Triggering

and Control system - TTC Interface Module (TTC) and other components of Readout system. TIM is placed at ROD crate as well. Modules connections are grouped at PPo (PPo is not displayed at the figure 1). Each PPo can contain six or seven modules and these are connected to the ROD. The one ROD can be connected to 1, 2 or 4 PP0s, it depends on the speed and events count [2]. ROD is responsible for generating command bit-stream for the modules and for decoding incoming data from the sensors as well. From these data ROD creates event (from all modules) and transfers it in a general ATLAS format. Data from sensors are transferred by optical fibres to the off-detector interface (ROD) via the Back of Crate card (BOC) which is converting signals (optical to LVDS) and by S-LINK interface to the Read Out system (ROS). The all off-detector parts are located approx. 100 m away from the detector. The common signals other than trigger, timing, e.g. configurations for read-out chips are being transferred to the ROD via a VME interface form the DAQ-system. These data are transform in ROD card by DPS and FPGA chips to the form for modules. These chips are creating also histograms and some analysis and sending them to the rest of DAQ-system for online monitoring by VME interface or for further processing and archiving to the DAQ-system by S-LINK interface.

2.1 TurboDAQ set-up

For the testing purpose, known as Test-beam, we have TurboDAQ set-up, where ROD card is replaced by VME-crate, also known as Turbo Pixel Low Level card (TPLL) and interface between VME-crate and modules are made by Turbo Pixel Control Card (TPCC) according figure 2.

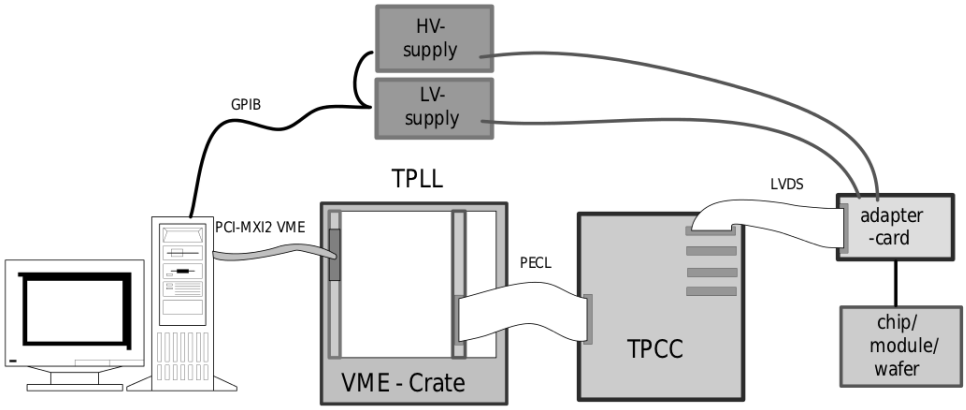


Figure 2: TurboDAQ test system [3].

For reading data from VME-crate we use standard PC connected to the TPLL by VME interface. TPCC is used for converting signals to specific format readable by modules consisting of the MCC and FE chips in number of maximum four modules.

3 DAQ-software

The data acquisition software (DAQ-software) is based on a library *PixLib*. The PixLib library is a complex collection of C++ applications and libraries used in a distributed

environment. This library provides the support of the hardware of the all Pixel Detector such as sub-library *PixBoc* for the BOC card, *RodPixController* for the ROD card, *PixFe* and *PixMcc* for modules, etc. as well as controlling, scanning, tuning capabilities, calibration, data-taking, making histograms, etc. This makes the Pixel DAQ-software different from the one than software used in other sub-detectors in the ATLAS. Pixel DAQ-software is making calibrations not just as standard taking-data and changing trigger set-up, but it makes scans, short sequences of triggers, with different detector conditions such as different thresholds, shaping time, pulse height, Opto-link, intensities or delays etc. These results are formed as histograms in the RODs and transferred to DAQ-system for monitoring. Differences between the local data-taking and calibration are marked in the figure 3 and 4. Each ROD can be used for data-taking mode as well as for calibration mode.

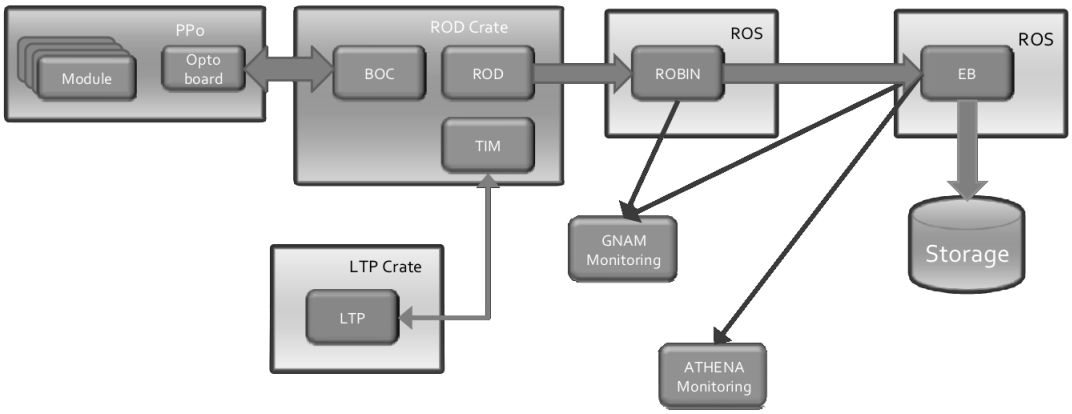


Figure 3: Local Data Taking in the Pixel Detector [2].

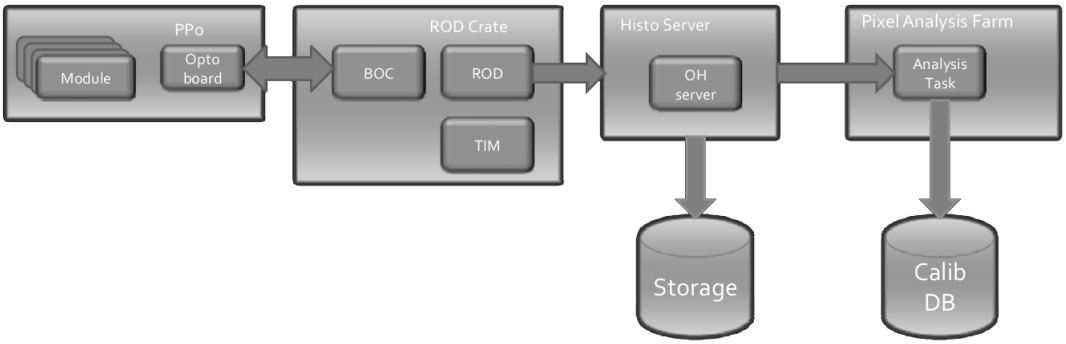


Figure 4: Calibration in the Pixel Detector [2].

Very important feature is also a sophisticated database system (DB) as well. In the configuration DB Module and Opto-link configurations are stored. That database is formed as Oracle server and in order to avoid too many processes in the DAQ-system access to this server at a same time, cache is created. In these caches basic configurations are stored. These are called DB Servers and can read configuration from Oracle server and distribute them to the DAQ-system as well as store the present configuration of set-up to the Oracle server [2]. As well Histogram Server as cache was created. The

Histogram Server is used to collect histograms from the RODs and distribute them to the processes running in the analysis farm.
The structure of DAQ-software is shown at figure 5. The software consists of an applications running on the different places in the system. Some are running on the Single Board Computer (SBC). This SBC is standard disk-less Intel based computer with VME interface for connection with ROD crate. The SBC is used for to controlling ROD crate, downloading configurations for FE chips and for the all modules from DB as well as for taking histograms saved in the ROD and sending them to the histogram server. SBC is also interacting with archiving system as well as with the rest of the DAQ-system. The communication between the SBC and the outside is based on the Gigabit Ethernet. The applications running in the SBC are called Action Servers. These are running as a couple of threads per one ROD. For each ROD only one application can be run. This is checked by Crate Broke application running in the SBC computer.
The DAQ-software includes applications that can perform different tasks and are running on different computers connected to each other via Inter Process Communication (IPC), so the IPC is used to address the crate controllers. This technology allows remote applications to call transparently across the network functions executed by different processors and to connect the ROD crate with the rest of the DAQ-system.

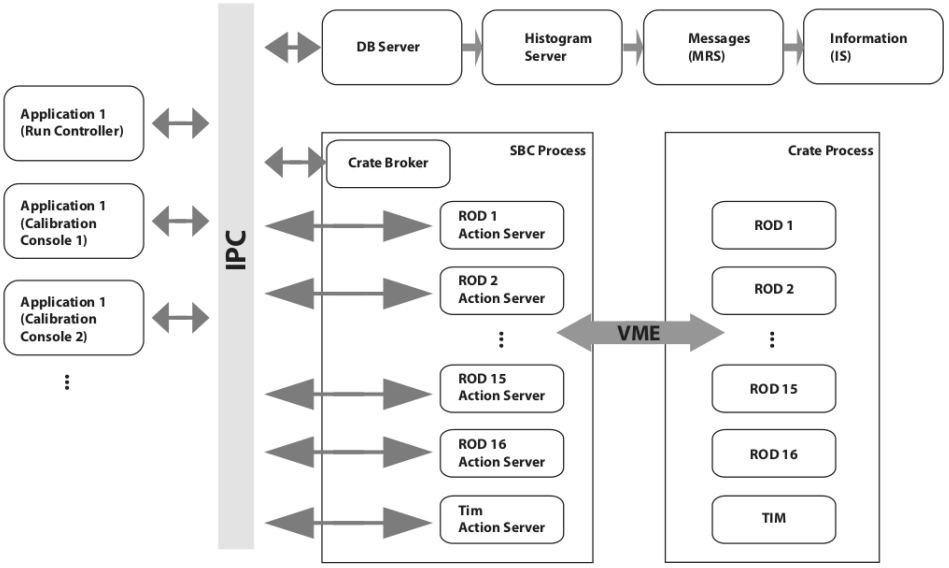


Figure 5: DAQ-software structure [1].

The whole system is created to be able to run different hardware on the same architecture. It means that the implementation of the new read-out electronics using the VME interface is limited only by changing software at the crate level such as drivers for BOC and ROD crates. When we implement new FE-I4 chip to the system, changes appear only at the hardware level and data and signals remain still compatible with the present Pixel detector.
Due to these aspects the implementation of the new FE-I4 chip is possible as additional layer to the present architecture of the Pixel Detector.

3.1 Detector partitions

The ATLAS detector consist of couple of detectors. Each detector can take data independently. Data-taking from one detector is called partition. In our case, Pixel detector is divided into three partitions: Layer 1 and 2 and Disks and each partition can run alone or with other partitions together. This division gives us an advantage of taking data with different (local) triggering - from Local Trigger Processor (LTP) or with trigger for all other partitions from Central Trigger Processor (CTP). Individual RODs can be removed from a partition and used for calibrations. Overview of partitions is in figure 6.

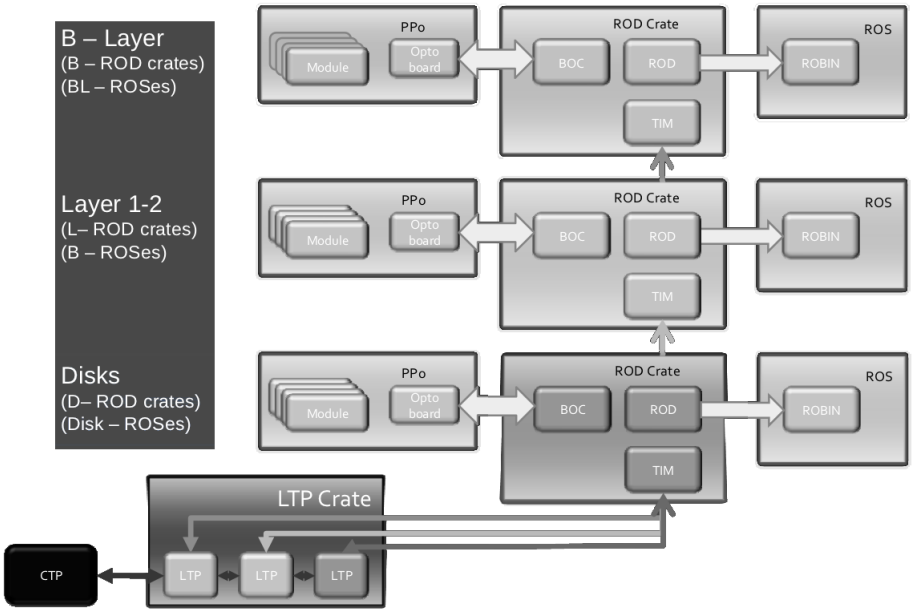


Figure 6: Pixel Detector Partitions [2].

DAQ-software is also divided to partitions for different tasks. TDAQ-system has TDAQ partition which is collections of processes running on different computers. These processes can be controlled by commands sent to the root controller for all processes or to the individual process. With the TDAQ partition it is possible to make data-taking from one or more detector partitions at the same time. On another hand it is possible to use the same tool for controlling processes such as calibrations, processes which are not taking data. TDAQ partition includes special partition called *PixelInfr* partition for control the Action and Broker processes and other servers. Always before a process such as calibration or data-taking can start, first we have to load correct configurations for current set-up from the DB server.

4 PixLib

The PixLib is the library which acts as software layer to interface the ROD with the other end user applications to access the Pixel modules. This library consist of couple of sub-libraries such as drivers for ROD, BOC, FE modules etc. as well as controlling applications as timing, scaning, tuning, calibration, data-taking, etc. A lot of them are

detector independent because of many actions which are made by ROD hardware itself. Many of these actions are running on SBC computer and connected with other GUI application on the different computer via a Ethernet connection.
In global, PixLib does not provide any specific task, it provides just access to control Pixel Modules which are hidden to the end users. The PixLib also provides an interface to the TIM and to the Detector Control System (DCS) which is computer controlling and supervising the detector and related services aiming at stable operation of the detector system such as the voltages, the temperatures etc. It also provides access to the database servers where calibration and histogram data are stored.

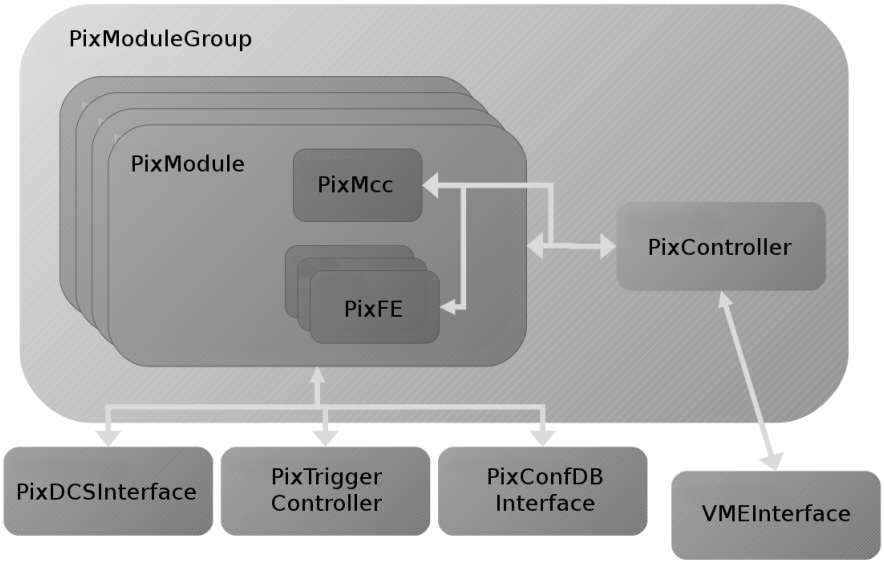


Figure 7: Structure of the PixLib library [4].

Figure 7 shows a basic structure of the *PixLib* library. The top level of the structure consist of *PixModuleGroup* class. This class corresponds with the set of the modules controlled by the same ROD. The structure of the *PixLib* library does not allow to control more than one ROD. It means that *PixLib* is just for one ROD and to access to the more RODs we need to run another application at the run controller. Exactly one Action Server in the SBC computer per one ROD card.
PixModuleGroup creates one or more instances of the following object depending on the configuration:

- PixModule - up to 6 or 7 instances according configuration
- PixController - only one instance

PixModuleGroup object also receives pointers to:

- VmeInterface
- PixTriggerController
- PixDCSInterface

- PixConfDBInterface

These are just main sub-classes listed. In point of fact there are plenty of the sub-classes. The *PixController* is the abstract class used mainly for communication with Modules and for access to the specific ROD implementation such as TIM controller.
The *PixModule* class is responsible for the most of the work because it contains the code for working with the module. This class will create instances of the following object:

- PixMcc - only one for FE-I3 or no instance for FE-I4
- PixFE - up to 16 for FE-I3 or up to two for FE-I4 according configuration

PixMcc and *PixFE* are representatives of the generation of the specific MCC and FE commands with communication with read-out chips, since the *PixModule* class provides the complex task for the all Module such as full module configuration, calibration loops, threshold scans, etc.
The *PixMcc* class contains the full image of Mcc registers. This register has to be loaded from configuration DB first and then copied to the ROD memory to be used for configuration. There are also methods to upgrade and save configurations to and from the database. PixMCC will also generate the bit stream corresponding to the command and pass it to PixModule for execution.
The *PixFE* class contains the whole image of all sensor and read-out chip settings such as FE DACs and pixel bits loaded also from the configuration DB in the same way as for MCC chip. The class can directly execute some commands or make bit-stream for execution by PixMcc.
We can divide commands to higher-level and low-level commands. Low level commands can be created by PixFE and be implemented only on the corresponding FE chip or higher-level commands that can create PixMcc, PixController or PixModuleGroup and be applied to all group of modules or FE chips.
I have to mention also another main classes as *PixTriggerController* which is used to interface trigger controller (TIM), *PixDCSInterface* which is used to interface DCS computer for setting voltages, temperatures etc. and *PixConfDBInterface* which is used to interface configuration databases.

4.1 Implementation of FE-I4 to PixLib

The class *PixLib* was developed originally for FE-I1, FE-I2 and FE-I3 read-out chips. Since the new FE-I4 chip has been developed, we need to adjust *PixLib*.
We started with DAQ-software version tagged as IBLDAQ-0-0-0 which is frozen and no code for FE-I4 is made. We already have instances as objects for FE-I4 derived from the USBPix application written for testing single FE-I4 chips. These already existing instances we have copied directly to PixLib without any changes.
Since FE-I3 chip is connected to MCC chip, software is full of hard-coded points with connection to these hardware constrains and these need to be eliminated. We need a code in witch MCC/FE-I3 and FE-I4 modules can co-exist. It means that both types of modules will be dealt within the same class. From another point of view, to keep the current code intact, and wherever things are different for FE-I4, implement them so that

new code shall be used for FE-I4-type modules as well. Just to be more specific, firmware changes (DSP software) are required for ROD and BOC card as well.
We have two option for fixing present daq-software:

- with a $dynamic\_cast < PixFeI4* > (\&fe)$
- using PixModule's $m\_feFlavour$ variable

First one can use C++ type-casting in order to find out what type of chip is connected. The second one is using the chip flavours, variables $m\_mccFlavour$ and $m\_feFlavour$ which are known to the PixModule. Example how to use the second variant with chip flavours is shown below:

```
// m_mccFlavour and m_feFlavour already read from DB
if(m_mccFlavour==PM_MCC_I2 && m_feFlavour==PM_FE_I2){
  m_mcc = new PixMccI2(dbServer, this, dom, tag, "MCC");
  conf.addConfig(&(m_mcc->config()));
  for (int i=0; i<16; i++) {
    m_fe.push_back(new PixFeI2(dbServer, this, dom, tag,"FE",i));
    conf.addConfig(&(m_fe.back()->config()));
  }
}
else if(m_mccFlavour==PM_NO_MCC && m_feFlavour==PM_FE_I4){
  m_mcc = 0;
  // to be seen if we have 1 or 2 FE-I4 per module
  for (int i=0; i<1; i++) {
    m_fe.push_back(new PixFeI4(dbServer, this, dom, tag,"FE",i));
    conf.addConfig(&(m_fe.back()->config()));
  }
}
else std::cerr << "Inconsistent or non-existing MCC/FE types for module "
               << m_name << std::endl;
```

For testing purpose we use DummyPixController, which is replacement for real Pixel Detector set-up. Working with this simulator makes process easier, because in major cases we are not able to set the all pixel set-up and debug software because of its enormous complexity. The DummyPixController needs just one PC, it is as host-PC and SBC computer at the same time.

5 Conclusion

Since the LHC brought the first results, there is effort for upgrade of particular components. One of theme is the replacement or addition of new detectors to the LHC experiment. In co-operation with an international group of scientists, we are working on the implementation of the new FE-I4 sensors to the present ATLAS Pixel detector. Despite quite robust architecture of the DAQ-software it is necessary to take DAQ-software as one cell and in aim of implementation of FE-I4 to be familiar with all this architecture.

The work includes software as well as hardware and firmware development. The plan of the Pixel group is to have a functional set ready for testing at the end of this year, however, there are still a lot of problems with code debugging due to the request of proper environment.

References

[1] ATLAS IBL Community. *ATLAS Insertable B-Layer: Technical Design Report*. Internal Report ATLAS TDR 19 – CERN-LHCC-2010-013, (2010).

[2] ATLAS Pixel DAQ Group. *Training for DAQ shifters*. DAQ training V4, (2009).

[3] J. Große-Knetter. *Vertex Measurement at a Hadron Collider: The ATLAS Pixel Detector*. Bonn University, BONN-IR-2008-04, (2008).

[4] P. Morettini. *PixLib*. Presentation, SCT/Pixel ROD Software Workshop, Cambridge, (2002).

[5] T. Flick. *Studies on the Optical Readout for the ATLAS Pixel Detector: Systematical Studies on the Functions of the Back of Crate Card and the Timing of the Pixel Detector*. WUB-DIS 2006-05, (2006).

Towards a New Data Acquision Software for the COMPASS Experiment

Vladimír Jarý

3rd year of PGS, email: jaryvlad@kmlinux.fjfi.cvut.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. The COMPASS is a particle physics experiment with a fixed target situated on the SPS accelerator at the CERN laboratory in Geneva, Switzerland. The current data acquisition system of the experiment based on the ALICE DATE package suffers from the performance and stability problems. In this paper, several improvements of the current DAQ system are mentioned. Recently, the online database service of the experiment has been updated and operating system on all nodes participating in the data acquisition has been upgraded to a more recent version. Also, the requested remote control of the experiment has been implemented into the system. In parallel, a development of a brand new DAQ system based on a custom FPGA–based hardware has started. The proposal of the run control and monitoring software for this new platform is presented.

Keywords: data acquisition, COMPASS, remote control, distributed systems

Abstrakt. COMPASS je fyzikální experiment s pevným terčem umístěný na částicovém urychlovači SPS v laboratoři CERN v Ženevě, Švýcarsko. Současný systém pro sběr dat založený na softwarovém balíku ALICE DATE se potýká s problémy s výkonem a stabilitou. Tento článek popisuje několik řešení těchto problémů. V první řadě došlo nedávno ke kompletní výměně databázové služby. Zároveň proběhla migrace na novější verzi operačního systému na všech strojích účastnících se sběru dat a bylo nainstalováno a otestováno požadované vzdálené řízení experimentu. Paralelně k těmto aktivitám započal vývoj zbrusu nového systému pro sběr dat založeného na vlastním hardware. V článku je představen navržený řídící a dohledový software pro tuto hardwarovou platformu a jsou zmíněny první postřehy z implementace tohoto návrhu.

Klíčová slova: sběr dat, COMPASS, vzdálené řízení, distribuované systémy

1 Introduction

The *Common muon and proton apparatus for structure and spectroscopy* (COMPASS) is a particle physics experiment with a fixed target running on the *Super Proton Synchrotron* accelerator at CERN laboratory in Geneva, Switzerland [1]. The scientific program of the experiment was approved by the CERN Scientific Council in 1997, the data taking started after several years of construction and testing in 2002. The scientific program consists of the muon program, that includes the research of the transverse spin effects or the investigation of the muon polarization, and the hadron program, that covers the research of the Primakoff scattering or the exotic states. Recently, the extension of the

program of the experiment, known as COMPASS II, has been approved [2]. The extension includes the research of the generalized parton distribution function, the Drell-Yan effect, and the Primakoff scattering.

At first, the existing data acquisition system based on the ALICE DATE software package is presented. Then, several improvements of this system are described. These improvements include update of the online database service or implementation of the remote control of the experiment. Finally, a brand new data acquisition system that is currently under development is presented.

2 The DAQ system of the COMPASS experiment

A typical data acquisition system performs several tasks: it reads data produced by detector(s) (*readout*), assembles full events from fragments of data (*event building*), transfers data to a permanent storage (*data logging*), and provides control, configuration, and monitoring to human operators (*run control*).

The COMPASS data acquisition (DAQ) system strongly depends on the supercycle of the SPS accelerator that consists of the acceleration (12 s) and the extraction (4.8 s) period known also as a spill. The DAQ system must use the acceleration period to reduce the data rate to one third of the on–spill rate. Typical SPS spill contains 2×10^8 particles for the muon beam and 1×10^8 particles for the hadron beam. When a beam particle interacts with the COMPASS polarized target, secondary particles are produced and are later detected in a system of detectors that forms the COMPASS spectrometer. Detectors are used to track particle (various wire chambers), to identify particles (e.g. Ring Imaging Cherenkov counter, muon filters), and to measure deposited energy (hadronic and electromagnetic calorimeters).

Collection of data describing the flight of particle through the spectrometer is called the *event*. The majority of registered events does not correspond to any physically interesting phenomena. The purpose of the *trigger system* is to select interesting events in a high rate environment. The trigger decision is based on signals from fast detectors, e.g. hodoscopes. The trigger system greatly reduces the storage requirements and also the CPU power required to analyze data. Average event size is approximately 40 kB, the data collected per one spill can reach up to 18 GB. During the 2004 Run, the experiment collected almost $\frac{1}{2}$ PB of data. DAQ of the COMPASS experiment uses buffering and parallel processing to handle these data rates.

The COMPASS DAQ system consists of several layers. On the lowest layer, the *primary* (frontend) electronics lies. Its main task is to preamplify, discriminate, and digitize data from detectors. There are roughly 250000 channels; data streams from multiple channels are concentrated into the *concentrator modules* CATCH and GeSiCA. The readout of data is triggered by signals distributed by the *trigger control system* TCS. This system also distributes the event identification and the time reference. When the concentrator module receives the TCS signal, it performs detector readout and appends *subevent header* to data. Subevents are then transferred via optical bus S-Link to the *readout buffer* computers that form the following layer of the system. The readout buffers are standard servers equipped by the custom PCI cards called *spillbuffers*. Readout buffers receive subevents during spills and continuously transfer them to the *event builder* servers that

form the last layer of the DAQ. Thus, readout buffers use the SPS supercycle to reduce the data rate to one third of the on-spill rate. Connection between readout buffers and event builders is based on the TCP/IP standard. Event builders use information from the subevent headers to assemble full events. Full events are stored in the CERN permanent storage CASTOR; additionally, catalogue file with metainformation is prepared and stored in the Oracle database. Remaining CPU power of the event builders is dedicated for additional tasks such as event sampling or online data filtering.

DAQ software is based on the *DATE* (Data Acquisition and Test Environment) package that has been developed for the ALICE experiment. The DATE package is designed to perform DAQ tasks in a multiprocessor distributed environment. DATE was designed to be a very scalable system; at the ALICE experiment, it runs at two modes: proton–proton collisions and heavy ion collisions. The pp mode is characterized by a high interaction rate and small event sizes. On the contrary, the heavy ion mode is characterized by relatively small interaction rates and large events. The DATE performs the data acquisition at the ALICE experiment on hundreds of distributed nodes. On the other hand, it can be also used in a small experiments with just one node that performs all the tasks. At the COMPASS experiment, the DATE runs in the fixed-target mode, at ALICE in the collider mode. The performance of the package has been measured with the following results: 40 GB/s of the readout, 2.5 GB/s of the event building, and 1.25 GB/s of the storage [3]. The DATE requires each node to be x86–compatible machine powered by GNU/Linux operating system that supports the TCP/IP stack. From the functionality point of view, the DATE provides data flow control (*EDM*), run control (*dateControl*), interactive configuration (*editDB*), event sampling (*COOOL*), data logging (*infoLogger*), information reporting (*infoBrowser*, *MurphyTV*), and other tasks.

On the readout buffers, that are also known as the Local Data Concentrators in the DATE terminology, the process *recorder* runs. It off-loads the spillbuffer and passes the data to the recording device - in the case of the COMPASS experiment, the data is transferred to the event builders over the TCP/IP connection. On the event builders (also known as Global Data Collectors in the DATE terminology), the process *eventBuilder* runs. It receives subevents from readout buffers, uses subevent headers to assemble full events, passes the events to the next processing stage (e.g. online filter called Cinderella), and sends the events to the permanent storage. The transfer of subevents is initialized by the process *recorder* that runs on the LDCs. The *eventBuilder* process is also communicating with the *edm* (Event Distribution Manager) process that implements the load balancing. Event builders can send two types of messages to the EDM: nearly empty and nearly full. The EDM uses these messages to keep list of available event builders. Processes *edmClient* and *edmAgent* pass this list to the *recorder* process which selects an appropriate destination GDC for subevents. The destination GDC is selected using the round robin algorithm.

The current DAQ system suffers from a high dead time[1] caused by recent increases of the trigger rate and the beam intensity. Moreover, as the hardware gets older, the failure rate is also increasing. Thus it has been decided to propose and implement a new DAQ architecture. In the meantime, several improvements to the system have been

[1] DAQ dead time is a ratio between time when system is busy and cannot accept new events and total time.

implemented. At first, the online database service has been replaced.

3 Online database service

The DAQ system of the COMPASS experiment uses the *MySQL* database to store configuration, monitoring data, logbook, and software logs. As a consequence of increases in the trigger rates, the database service became overloaded and caused several crashes of the data acquisition during the 2009 Run, thus it has been decided to update it.

Original architecture consisted of two physical database servers *pccodb01* and *pccodb02* that were synchronized by the master–master replication. In this configuration, the server *pccodb01* acts as a replication master of the slave server *pccodb02*. At the same time, the server *pccodb02* also acts as a master of the slave server *pccodb01*. Clients connected to the database through the virtual address *pccodb00*. Normally, this address pointed to the *pccodb01* server. If the watchdog process detected a crash of the *pccodb01* server, it reset the virtual address to point to the remaining server *pccodb02*.

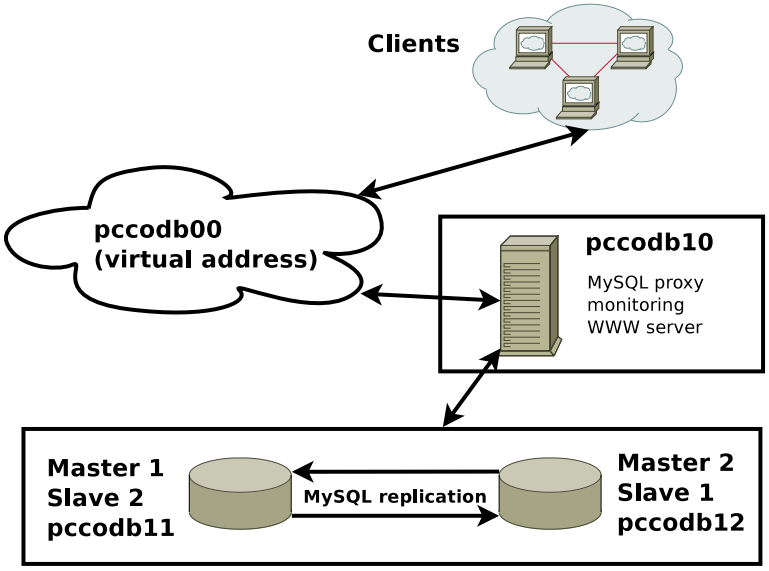


Figure 1: New database architecture

Before the start of the 2010 Run, the database service has been updated. In the updated architecture, the *pccodb01* and *pccodb02* servers are replaced by the servers *pccodb11* and *pccodb12* that are equipped by much more modern hardware. Also, the MySQL software and operating system have been upgraded to more recent versions. These new servers are also synchronized using the replication. Moreover, additional server *pccodb10* has been added. This server hosts the *MySQL Proxy* software that enables the (read only) load balancing. There is also monitoring service *Nagios* running on the proxy server. If some problem with a physical server is detected, the proxy is automatically reconfigured to forward all traffic to the unaffected server and e–mail message is sent to a database administrator. The web server is also running on the *pccodb10*. This service serves the web interface of the Nagios, the electronic logbook, and also the database management tool *phpMyAdmin*.

Server *pccodb11* is also replicated to the CERN IT center and from here into the computing centers of member organizations. This topology is known as a chain replication and can be regarded as a geographical backup. Moreover, regular backups are executed hourly (partial) and daily (full) by the system scheduler *cron*. During the replication process, the file with binary log is being created. This log can be used as an incremental backup. With the information contained in the daily, hourly, and the incremental backup, it is possible to reconstruct almost all data in case of the database failure.

In this configuration, the virtual address *pccodb00* still points to the proxy server. Since the virtual address is still the same, there was no need to reconfigure any clients during migration. During the 2010 Run, the database service ran stably and did not experience any crash. In case a higher performance is required, it is possible to easily add more replication slaves into the configuration and to enable the load balancing. More information about the new database service can be found in [6].

4 Remote control

The control room of the COMPASS experiment is placed directly in the experimental hall, just a few meters away from the spectrometer. The radiation levels increase with increasing beam intensity and safety limits might be exceeded in the future during the Drell-Yan program. Thus, the technical coordinator of the experiment has decided to install a remote control room.

Since the communication between nodes participating in the DAQ system is based on the TCP/IP protocols, the TCP/IP connection has been established between the experiment hall with detectors and the remote control room. Several applications are used by a shift crew to control and monitor the experiment and data taking.

The following equipment has been provided to power these applications: 8 HP workstations and 12 LCD screens. The operating system with the DATE package needed to be installed and configured on these workstations. The *Windows 7* has been installed on two workstations, *Scientific Linux CERN* (SLC) on others. As a Red Hat Enterprise Linux derivative, SLC contains the *Anaconda* system installer. Anaconda supports automated unattended installation using the kickstart technology. The *kickstart* is a text file that contains installation options such as a disk partitioning scheme, a network configuration, a package selection, or post-installation scripts. Using the CERN *Automatic Installations Management Facility*, the kickstart files are published in a network storage. During the installation, the Anaconda program downloads the appropriate kickstart file and performs the installation according to instructions stored in the file.

Moreover, it is not necessary to create a kickstart file for each computer, on the contrary, templates are supported. Computers participating in the DAQ can be divided into several groups: run control machines, event builders, readout buffers, gateways, file servers, or database servers. Each of this group is described by one kickstart template. During the installation, the template is parametrized by the IP address and the hostname of the machine. Additionally, the kickstart file can be used to quickly reinstall machine to original state in the case of a crash or a misconfiguration.

The run control workstations has been installed from the kickstart template. First workstation is running the human interface of the run control application, second is

displaying DAQ and detector errors, third one is running event sampling tool COOOL. Detector control system DCS is running on the fourth workstation, beam line and magnet monitoring program on the fifth, and IP cameras on the sixth. These 6 computers are part of the internal COMPASS network; the two remaining workstations are connected into the general purpose network (GPN) and are available for the shift crew.

A test run has been successfully started remotely, thus the remote control room is prepared for the future employment. Without the remote control room, the COMPASS collaboration would have to invest approximately 400000 EUR into the additional shielding of the spectrometer.

5 Research and Development of the new DAQ system

The development of the brand new data acquisition system has started. The main purpose is to increase the stability and decrease the dead time of the data acquisition. The new system is based on a custom FPGA[2] hardware that controls the data flow, the readout, and the event building [9]. Thus software is responsible only for the control and the monitoring. Moreover, the existing readout buffers and event builders can be turned into a dedicated filtering farm for COMPASS in the future.

At first, the possibility of using the DATE package with a new hardware has been evaluated [7]. It has been found out that the DATE is too complex software, moreover, it requires x86 compatible hardware [3]. Therefore, it has been decided to develop a new control and monitoring software. However, the DATE should be used as a source of inspiration during research and development of the new software.

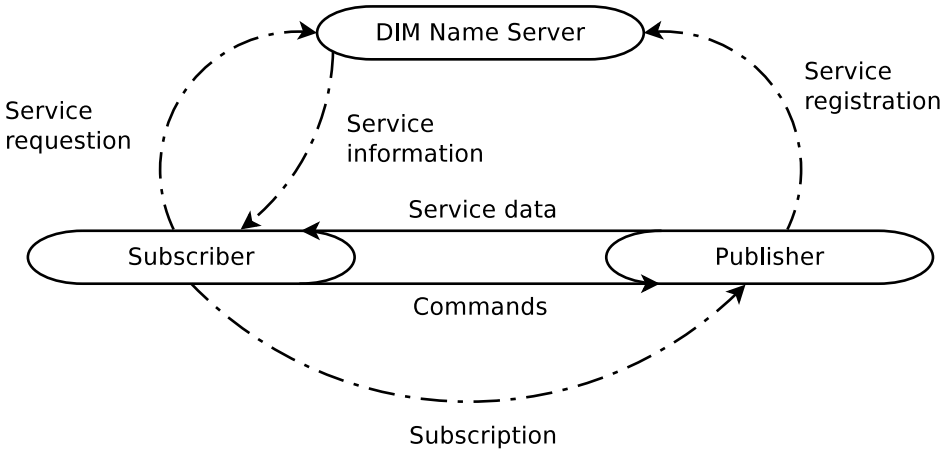


Figure 2: DIM name server

The new software should stay compatible with other parts of the experiment, especially with the Detector Control System. This implies that the new system should be built on top of the *DIM package* [4]. DIM (Distributed Information Management System) is a software library that provides asynchronous, one-to-many communication in a heterogeneous network environment. The library is based on the TCP/IP protocols, it

[2]Field Programmable Gate Array

extends the client–server paradigm with the concept of a name server. Each DIM service or command is identified by its unique name. When a server (publisher) publishes a service, it passes its name to the *DIM name server* (DNS) that registers it. When a client (subscriber) wishes to subscribe to a service, it also passes its name to the DNS which returns the location of the server that publishes the requested service. After that, normal TCP/IP connection is established between the subscriber and the publisher. The communication with the DNS is handled transparently by the library. Moreover, the library also handles the conversion of data between the host and the network encoding.

The library is written in the C language, however interfaces to the FORTRAN, C++, Java (using the Java Native Interface), and Python also exist. The performance of the C++ and the Java interface has been compared. In the test, the server publishes one information service and one command. The client sends command to the server, when the server receives this command, it updates its service. This forces client to fetch the updated information; when the client receives this information, it sends another command. This cycle is repeated million times and network usage and elapsed time is measured for different sizes of the message. The results have been measured on the 100 MBit/s network card. It has been found out, that the DIM performance scales well with the increasing size of the message. Moreover, the DIM is able to saturate the network, the overhead caused by the communication with the name server can be neglected for larger messages. As expected, the Java performance is lower than the performance of the C++ because of the JNI overhead. For smaller messages, the difference in performance is about 20%, however as the message size increases, the performance hit caused by the JNI diminishes. Unfortunately, the Java DIM interface is not complete, thus it has been decided to use the C++ version. Results of the performance test are discussed in more details in [8].

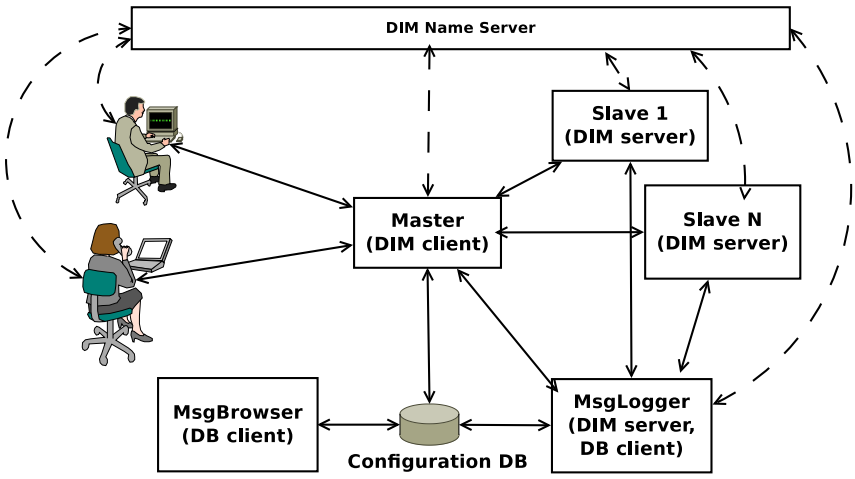


Figure 3: Nodes participating in the new DAQ architecture

The control and the monitoring software has been designed. Figure 3 shows the relations between actors participating in the system. The actors are running on distributed nodes, communication is implemented using the DIM library. The master node plays the key role. The master loads the global configuration from the online database. The MySQL database server has been selected because of the compatibility with the current DAQ system. Master distributes the configuration to all slaves; in this way, the slave do

not need a database access. The information stored in the database contains the list of the slave nodes. Each node is identified by its unique number and its hostname. Master connects to these nodes using the ssh client and wakes up the slave processes that start DIM servers. The master acts as a DIM client of these servers. The master can send DIM commands to the slaves. Typically, these commands are used to start or stop data taking. On the other hand, slave nodes publish some monitoring information, such as a buffer usage, data rates, or errors in data, as a DIM services. The master subscribes to these services and collects the monitoring data. The format of the message has been proposed. Each message starts with header with metainformation. The header is followed by the payload and the message is closed by the trailer with the check sum. The format is described in the table 1.

The format of the message			
Header			
1.	Data size	4 bytes	Total size of the message in 32b words = header size+payload size+trailer size
2.	Version	4 bytes	Version of the protocol
3.	Sender ID	4 bytes	Unique ID of the message's sender
4.	Message number	4 bytes	Number of the message
5.	Receiver ID	4 bytes	Unique ID of the message's receiver
6.	Message ID	4 bytes	ID of the message
7.-8.	Time	8 bytes	Time stamp
Payload			
9.	Body	(0-N) × 4 bytes	Body of the message (can be empty)
Trailer			
10.	Reserved	4 bytes	0x00000000
11.	Reserved	4 bytes	0x00000000
12.	Message number	4 bytes	Number of the message (the same as in the header)
13.	Check sum	4 bytes	Check sum of the message

Table 1: Message format

On the other hand, the master node is also containing the DIM server part. This part is receiving commands from the user interface application and publishes information about state of the system. At the same time, multiple user application can communicate with the master, however only one can control the system - the remaining users can only observe the behaviour of the system. The remote control is supported thanks to the DIM library. Communication protocol between the user interface applications and the master node uses the same message format. It has been decided to implement the user interface in the QT framework that is portable and contains rich class library that covers widget, database access, graphics, and platform independent data manipulation.

The master and all the slave nodes are also sending debug information to a *Message Logger* application. The Message Logger buffers these messages and periodically flushes them into the permanent storage, usually into the MySQL database. The debug information contains the time stamp, the identification of the node, the severity (notice, warning,

error, fatal error), and the actual description of the incident. The *Message Browser* is an application with a graphical user interface that will be used to display and to query the database with messages. The Message Logger and the Message Browser replace the InfoLogger and the InfoBrowser applications from the DATE package.

The master node, the user interface, and the Message Logger with the Message Browser will be running on a standard x86–compatible hardware powered by the Scientific Linux CERN operating system. Thus, it is possible to use some higher level libraries such as QT during the implementation of these applications. On the other hand, slave nodes will be running on a custom hardware (MICO32 softcore processor) under some Linux distribution for microcontrollers. The slave application will depend only on the DIM library which should be available also for the microcontroller Linux.

The work on implementation of the above described proposal has already started. The communication of the master node with slave nodes and user interface node has been tested on three nodes. Further tests on multiple nodes are scheduled into the nearest future. The code of the slave process needs to be ported to the microcontroller Linux. The goal is to have a fully functional prototype for the 2012 Run. During the year 2013, the shutdown of the entire accelerator complex is expected at CERN. This period should be used for final testing and installation so that the new data acquisition system is ready to be in operation in the 2014 Run. If proved to be successful, the system will also be deployed at the *PANDA* experiment at the *FAIR* facility at Darmstatd, Germany.

6 Conclusion and outlook

The existing data acquisition system of the COMPASS experiment has been described. The stability of the system decreases as the hardware gets older and the trigger rate increases with increasing beam intensity. Several interventions have been proposed and performed in order to improve the stability. The database service that has caused several crashes during the 2009 Run has been migrated to the new software and hardware. Since the migration, the database has not experienced any severe problem. In order to reduce the exposure of the shift crew to the radiation, a remote control room has been installed.

A new data acquisition system based on the custom hardware is being developed. The readout, the data flow control, and the event building is controlled by the hardware, the software is responsible for the run control and monitoring. The requirements has been analyzed and the proposal has been designed. The implementation of the proposal has started. It is projected to have the new system in a full operation for the 2014 Run.

7 Acknowledgement

This work has been supported by the MŠMT grants LA08015 and SGS 11/167.

References

[1] 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.

[2] Ch. Adolph et al. (the COMPASS collaboration). *COMPASS-II proposal.* CERN-SPSC-2010-014; SPSC-P-340 (May 2010)

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

[4] P. Charpentier, M. Dönszelmann, C. Gaspar. *DIM, a Portable, Light Weight Package for Information Publishing, Data Transfer and Inter-process Communication.* Available at: http://dim.web.cern.ch

[5] M. Bodlák, V. Jarý, T. Liška, F. Marek, J. Nový, M. Plajner. *Remote Control Room For COMPASS Experiment.* In: 37th Software Development, Ostrava: VŠB – Technická univerzita Ostrava, 2011, ISBN 978-80-248-2425-3. pp. 1–9.

[6] V. Jarý. *COMPASS Database Upgrade.* In: Doktorandské dny 2010, Praha: ČVUT, 2010, ISBN 978-80-01-04664-9. pp. 95–104.

[7] V. Jarý. *DATE evaluation.* In: COMPASS DAQ meeting, Geneva, Switzerland, 29 March 2011

[8] V. Jarý, T. Liška, M. Virius. *Developing a New DAQ Software For the COMPASS Experiment.* In: 37th Software Development, Ostrava: VŠB – Technická univerzita Ostrava, 2011, ISBN 978-80-248-2425-3. pp. 35–41.

[9] A. Mann, F. Goslich, I. Konorov, S. Paul. *An AdvancedTCA Based Data Concentrator and Event Building Architecture.* In 17th IEEE-NPSS Real-Time Conference 2010, Lisboa, Portugal, 24–28 May 2010

[10] 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

Discretization of Superintegrable Systems on a Plane

Zdeněk Kabát

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

advisor: Severin Pošta, Department of mathematics,
Faculty of Nuclear Science and Physical Engineering, CTU in Prague

Abstrakt. V článku konstruujeme diferenční analogie k tzv. Smorodinského-Winternitzovým superintegrabilním systémům v Eukleidovské rovině. Za použití metod umbrálního počtu získáváme diferenční rovnice pro zobecněný isotropní harmonický oscilátor na uniformní mřížce a taktéž jeho řešení ve tvaru mocninných řad. V případě kalibračně transformovaných Hamiltoniánů je řešením polynom, dobře definovaný v celé rovině.

Klíčová slova: diskrétní harmonický oscilátor, umbrální počet, uniformní mřížka

Abstract. We construct difference analogues of so called Smorodinsky-Winternitz superintegrable systems in the Euclidean plane. Using methods of umbral calculus, we obtain difference equations for generalized isotropic harmonic oscillator on the uniform lattice, and also its solution in the terms of power series. In the case of gauge-rotated Hamiltonian, the solution is a polynomial, well-defined in the whole plane.

Keywords: discrete harmonic oscillator, umbral calculus, uniform lattice

1 Introduction

Recent development in the field of theoretical and mathematical physics leads to an idea that existing models in quantum mechanics are only a continuous approximation of discrete space-time. Discretization has shown to be a convenient tool for quantum chromodynamics and renormalization theories [4], as well as for one of the candidates for a grand unification theory – loop quantum gravity [1]. This assumes an elementary length $l_P = \sqrt{\hbar k} = 10^{-33}$ cm which is referred to as Planck length.

There have been several attempts to create models for discrete quantum mechanics. An approach to the harmonic oscillator and hydrogen atom using special functions theory has been introduced in Atakishiev, Suslov [2] and Lorente [10]. Odake and Sasaki [12],[11] construct Hamiltonians as infinite-dimensional Jacobi matrices which can be understood as a discrete quantum mechanical system on a uniform grid or q-grid. An operator approach for discretization of harmonic oscillator has been used by Turbiner [17].

The problem with these methods is that one encounters issues with preserving Lorentz and Galilei invariance and symmetry algebras. This can be partially solved by using a mathematical tool called "umbral calculus", introduced by Roman [14] and Rota [15] in 1970's. An umbral approach for simple systems has been used in Dimakis [5] and later extended to two dimensions by Levi and Winternitz [9].

The aim of this article is to extend the application of umbral calculus and use a particular realization of difference operators that transfer some integrable systems to two-dimensional uniform grid. Thanks to the essence of the umbral theory, Lie symmetries are preserved and solutions are obtained by a simple substitution.

In Section II, we introduce the mathematics of umbral calculus and we show the particular difference operators to be used. In Section III, superintegrable systems are defined and two classes of harmonic oscillators on an Euclidean plane are described. Section IV is devoted to the own discretization and we find the solutions of the corresponding difference equations. Finally, in Section V some conclusions are drawn.

2 Discretization Method

Let $\mathbb{F}$ be a field of characteristic zero. We denote $\mathscr{P} = \mathbb{F}[x]$ a vector space of polynomials over $\mathbb{F}$ in variable x and $\mathscr{L}(\mathscr{P})$ a space of linear operators on $\mathscr{P}$. Addition and scalar multiplication are defined as usual.

Let $\mathscr{F}$ be an algebra of formal power series in variable t, i.e. the elements of $\mathscr{F}$ are in the form $\sum_{k=0}^{\infty} a_k t^k$. For $f(t) = \sum_{k=0}^{\infty} a_k t^k$ and $g(t) = \sum_{k=0}^{\infty} b_k t^k$ we define the algebraical operations as follows:

$$f(t) + g(t) = \sum_{k=0}^{\infty} (a_k + b_k) t^k,$$

$$f(t)g(t) = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} a_j b_{k-j} \right) t^k.$$

With these operations, $\mathscr{F}$ is an algebra with no zero divisors, and is called an *umbral algebra*. Moreover we define a formal derivative on $\mathscr{F}$ naturally as

$$f'(t) = \sum_{k=1}^{\infty} k a_k t^{k-1}.$$

There is a certain correspondence between formal power series and linear operators on $\mathscr{P}$. For each $f(t) \in \mathscr{F}$ we define an operator $U_f \in \mathscr{L}(\mathscr{P})$ as

$$f(t) \mapsto U_f = \sum_{k=0}^{\infty} a_k \partial_x^k,$$

where $\partial_x = \frac{\mathrm{d}}{\mathrm{d}x}$ is an operator of derivative with respect to x. The operator U_f is called a *delta operator* if and only if $a_0 = 0$ and $a_1 \neq 0$. For $\sigma \in \mathbb{F}$, we define a *shift operator* $U_f = T_\sigma \in \mathscr{L}(\mathscr{P})$ using power series $f(t) \in \mathscr{F}$ as

$$f(t) = \sum_{k=0}^{\infty} \frac{\sigma^k}{k!} t^k.$$

We can easily see that the action of T_σ on $\mathscr{P}$ is

$$T_\sigma p(x) = p(x + \sigma),$$

in other words, T_σ is indeed a shift in the variable x. Consequently, we can define an important subalgebra $\mathscr{A} \subset \mathscr{L}(\mathscr{P})$ in the following manner:

$$\mathscr{A} = \{S \in \mathscr{L}(\mathscr{P}) | \forall \sigma \in \mathbb{F}\ ST_\sigma = T_\sigma S\}\,.$$

We call the elements of $\mathscr{A}$ *shift-invariant operators*. There is one-to-one correspondence between $\mathscr{F}$ and $\mathscr{A}$.

Theorem 1. *The map $f(t) \mapsto U_f$ is an isomorphism between umbral algebra $\mathscr{F}$ and shift-invariant operators $\mathscr{A}$.*

Now we establish a connection between delta operators and certain polynomial sequences.

Theorem 2. *For each delta operator $Q \in \mathscr{A}$ there exists a unique associated sequence $p_n(x)$, where the degree of $p_n(x)$ is n, such that*

$$p_0(x) = 1, \quad p_n(0) = 0 \quad \textit{for} \quad n = 1, 2, \ldots$$
$$Qp_n(x) = np_{n-1}(x).$$

A simple example of an associated sequence is $p_n(x) = x^n$ for the delta operator $Q = \partial_x$. However, the previous theorem shows that similar sequences can be found for every delta operator.

Let $Q \in \mathscr{A}$ be a delta operator with an associated sequence $p_n(x)$. An operator $\theta \in \mathscr{L}(\mathscr{P})$ is called an *umbral shift* if for all $n \in \mathbb{Z}^+$ it holds $\theta p_n(x) = p_{n+1}(x)$. For the operator ∂_x an umbral shift is trivially $\theta = x$, that is a multiplication by x in $\mathscr{P}$. There is an important theorem that gives us a formula to find an umbral shift for any operator:

Theorem 3. *The umbral shift for a delta operator $Q \in \mathscr{A}$ has the form*

$$\theta = x\beta, \quad \textit{with} \quad \beta = (Q')^{-1},$$

where $Q' = Qx - xQ$ is so called Pincherle derivative *of the operator Q. The operator β is called a* conjugate operator *to Q. Moreover*

$$[Q, x\beta] = 1.$$

A Pincherle derivative is defined for every shift-invariant operator and is easy to compute even without the series expansion from $\mathscr{F}$. However, if $f(t)$ is an indicator (i.e. the defining series) for Q, it can be proved that the formal derivative $f'(t)$ is indeed an indicator of Q'.

Because $\theta = x\beta$ takes a polynomial $p_n(x)$ of a given delta operator to $p_{n+1}(x)$, it can be used to "generate" the complete associated sequence as

$$p_n(x) = \theta p_{n-1}(x) = \ldots = \theta^n p_0(x) = (x\beta)^n 1.$$

The discretization procedure is based on so called *umbral correspondence*. We use it in the particular form

$$\partial_x \longleftrightarrow Q, \qquad x \longleftrightarrow x\beta,$$

where Q is an arbitrary delta operator. This mapping, thanks to Theorem 3, preserves Heisenberg commutation relations, particularly it preserves Lie symmetries of a system.

Let

$$F(\partial_x, x)f(x) = 0$$

be a linear differential equation with a solution $f(x)$ that can be expanded into a power series around a nonsingular point x_0. Without loss of generality, we assume that $x_0 = 0$ and that the expansion is

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n.$$

Let Q be a delta operator with a conjugate operator β and an associated sequence $p_n(x)$. We make an operator substitution in the differential equation obtaining

$$F(Q, x\beta)\tilde{f}(x) \cdot 1 = 0.$$

Following the umbral correspondence, we can see that after substitution $x^n \longleftrightarrow p_n(x)$ in the solution $f(x)$ we can write down a solution of our new equation, that is

$$\tilde{f}(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} p_n(x).$$

This can be proved realizing that the pair $(Q, x\beta)$ acts on $p_n(x)$ in the same manner as (∂_x, x) acts on x^n.

A special case of delta operators, which is of our interest, is a case of *difference operators*. Using the formalism of shift operators, we can introduce three simple cases of delta operators, right, left and symmetric discrete derivatives. For the right discrete derivative we get

$$\Delta_+ = \frac{1}{\sigma}(T_\sigma - 1), \qquad p_n^+(x) = \prod_{i=0}^{n-1}(x - i\sigma),$$

for the left discrete derivative

$$\Delta_- = \frac{1}{\sigma}(1 - T_\sigma^{-1}), \qquad p_n^-(x) = \prod_{i=0}^{n-1}(x + i\sigma),$$

and finally for the symmetric one

$$\Delta_s = \frac{1}{2\sigma}(T_\sigma - T_\sigma), \qquad p_n^s(x) = x\prod_{i=1}^{n-1}[x + (n - 2i)\sigma].$$

For $\sigma \to 0$ the operators converge to a continuous derivative, whereas the associated sequences tend to the simple sequence x^n. A corresponding operator substitution in the differential equation leads to a difference equation which can be understood as a discrete analogue of the original system.

In the Hamiltonians and solutions of the considered systems, we need to substitute not only the positive powers of x, but also the negative ones. Therefore a following extension of the associated sequences will be convenient. Let $k \in \mathbb{Z}^-$, then

$$p_k(x) = (x\beta)^k \cdot 1 = \left[(x\beta)^{-1}\right]^{-k} \cdot 1 = \left(\beta^{-1}\frac{1}{x}\right)^{-k} \cdot 1.$$

For the difference operators $\beta = \Delta_+, \Delta_-, \Delta_s$, we get the following extensions:

$$p_k^+(x) = \frac{1}{(x+\sigma)(x+2\sigma)\dots(x-k\sigma)} = \frac{1}{\prod_{i=k}^{-1}(x-i\sigma)},$$

$$p_k^-(x) = \frac{1}{(x-\sigma)(x-2\sigma)\dots(x+k\sigma)} = \frac{1}{\prod_{i=k}^{-1}(x+i\sigma)},$$

$$p_k^s(x) = \frac{x}{[x+k\sigma][x+(k+2)\sigma]\dots[x-(k+2)\sigma][x-k\sigma]} = \frac{x}{\prod_{i=-k}^{0}[x-(k+2i)\sigma]}.$$

However, one has to be careful with the domain of these expressions: $p_k^+(x)$ has singularities in negative lattice points.

3 Smorodinsky-Winternitz Systems

In this section, we introduce a class of quantum-mechanical system that will be discretized using the methods of umbral calculus. Let P_i, Q_j be operators of canonical momenta and coordinates, $i, j = 1, \dots, n$. We say that a quantum mechanical system with n degrees of freedom described by the Hamiltonian

$$\mathcal{H} = \sum_{i=1}^{n} P_i^2 + V(Q_1, \dots, Q_n)$$

is *integrable* if it allows $n-1$ independent integrals of motion in involution, that is the operators $X_1, \dots, X_{n-1}$ such that

$$[\mathcal{H}, X_a] = 0, \qquad [X_a, X_b] = 0.$$

The system is called *superintegrable* if there exist further $1 \le k \le n-1$ operators $Y_1, \dots, Y_k$ commuting with the Hamiltonian.

Considerable attention is given to the superintegrable systems since 1920's, beginning with works of Jauch and Hill [8] (harmonic oscillator), Pauli [13], Fock [6] and Bargmann [3] (hydrogen atom). A complete classification of superintegrable systems on an Euclidean plane was provided by Winternitz and Smorodinsky [7], [18] in 1965, and it was later found that all these models are exactly solvable [16]. We are interested mainly in the following two oscillator systems:

I. **Generalized isotropic harmonic oscillator**

$$\mathcal{H}_I(x,y) = -\frac{1}{2}(\partial_x^2 + \partial_y^2) + \frac{\omega^2}{2}(x^2+y^2) + \frac{\alpha}{2x^2} + \frac{\beta}{2y^2} \tag{1}$$

with solution of eigenvalue problem in Cartesian coordinates

$$\psi_{n,m}(x,y) = x^p y^q L_n^{\left(p-\frac{1}{2}\right)}(\omega x^2) L_m^{\left(q-\frac{1}{2}\right)}(\omega y^2) e^{-\frac{\omega x^2}{2}} e^{-\frac{\omega y^2}{2}}, \qquad (2)$$
$$E_{n,m} = \omega(2n+2m+p+q+1),$$

where $\alpha = p(p-1) > -\frac{1}{8}$, $\beta = q(q-1) > -\frac{1}{8}$ are parameters and ω is frequency of the oscillator. Gauge-rotated Hamiltonian follows as

$$h_I = \frac{1}{2\omega}\psi_{0,0}^{-1}(\mathcal{H}_I - E_{0,0})\psi_{0,0}\Big|_{\substack{t=\omega x^2\\ u=\omega y^2}} = -t\partial_t^2 - u\partial_u^2 + t\partial_t + u\partial_u - (p+\tfrac{1}{2})\partial_t - (q+\tfrac{1}{2})\partial_u, \qquad (3)$$

having simple polynomial solution

$$\Xi_{n,m}(t,u) = L_n^{\left(p-\frac{1}{2}\right)}(t) L_m^{\left(q-\frac{1}{2}\right)}(u). \qquad (4)$$

This system is also separable (and can be solved) in polar coordinates.

II. **Generalized nonisotropic harmonic oscillator**

$$\mathcal{H}_{II}(x,y) = -\frac{1}{2}(\partial_x^2 + \partial_y^2) + 2\omega^2 x^2 + \frac{\omega^2}{2} y^2 + \frac{\beta}{2y^2}. \qquad (5)$$

The solution of the corresponding Schrödinger equation is given by

$$\psi_{n,m}(x,y) = y^q H_n(\sqrt{2\omega}x) L_m^{\left(q-\frac{1}{2}\right)}(\omega y^2)\, e^{-\omega x^2} e^{-\frac{\omega y^2}{2}}, \qquad (6)$$
$$E_{n,m} = \omega(2n+2m+q+\tfrac{3}{2}).$$

The parameter β is the same as in case I. After the gauge rotation, we get

$$h_{II} = \frac{1}{2\omega}\psi_{0,0}^{-1}(\mathcal{H}_{II} - E_{0,0})\psi_{0,0}\Big|_{\substack{t=\sqrt{2\omega}x\\ u=\omega y^2}} = -\frac{1}{2}\partial_t^2 + t\partial_t - u\partial_u^2 + u\partial_u - (q+\tfrac{1}{2})\partial_u. \qquad (7)$$

A polynomial solution of this equation is

$$\Xi_{n,m}(t,u) = H_n(t) L_m^{\left(q-\frac{1}{2}\right)}(u). \qquad (8)$$

This Hamiltonian also separates in parabolic coordinates.

There are also two classes of Coulomb-type systems in the Euclidean plane which are related to the generalized oscillators through so called coupling constant metamorphosis. Basically, their Hamiltonians in parabolic coordinates coincide with the systems I and II and, therefore, they can be discretized in similar manner.

4 Results

4.1 General Discretization

For the discretization of S.-W. systems, we need to perform an umbral correspondence in $\mathcal{E}_2$, i.e. in two coordinates. In variables x, y, the substitution is denoted

$$\begin{aligned} \partial_x &\longrightarrow \Delta_x & \qquad \partial_y &\longrightarrow \Delta_y \\ x &\longrightarrow x\beta_x & \qquad y &\longrightarrow y\beta_y \end{aligned}$$

where Δ_x and Δ_y are arbitrary difference operators in the corresponding variable and β_x, β_y their conjugates. Similar notation is used for different coordinates (after a substitution).

Both models of generalized oscillators in $\mathcal{E}_2$ are separable in Cartesian coordinates and their gauge-rotated partners are separable in the substituted variables. Consequently, the difference equation obtained by the umbral correspondence is also separable and the solutions can be written as products of two functions.

Let us start with an operator substitution in gauge-rotated Hamiltonian (3). The discrete version has the form

$$h_I^D = -t\beta_t\Delta_t^2 + t\beta_t\Delta_t - (p+\tfrac{1}{2})\Delta_t - u\beta_u\Delta_u^2 + u\beta_u\Delta_u - (q+\tfrac{1}{2})\Delta_u.$$

Since the eigenfunctions are polynomials, we can immediately discretize the solutions (4) of the corresponding Schrödinger equation:

$$\Xi_{n,m}^D(t,u) = \widehat{\Xi}_{n,m}\cdot 1 = L_n^{\left(p-\frac{1}{2}\right)}(t\beta_t)L_m^{\left(q-\frac{1}{2}\right)}(u\beta_u)\cdot 1 = \sum_{i=0}^{n} l_{i,n}^{\left(p-\frac{1}{2}\right)} p_i(t) \sum_{j=0}^{m} l_{j,m}^{\left(q-\frac{1}{2}\right)} p_j(u),$$

where $l_{i,n}^{\left(p-\frac{1}{2}\right)}$ is the i-th coefficient of Laguerre polynomial $L_n^{\left(p-\frac{1}{2}\right)}$ and $p_i(t)$ is the associated sequence for the delta operator Δ_t.

In case of the original Hamiltonian (1), the general discretization leads to the operator

$$\mathcal{H}_I^D = -\frac{1}{2}(\Delta_x^2+\Delta_y^2) + \frac{\omega^2}{2}[(x\beta_x)^2+(y\beta_y)^2] + \frac{\alpha}{2}(x\beta_x)^{-2} + \frac{\beta}{2}(y\beta_y)^{-2},$$

The eigenfunctions of this operator corresponding to the lowest eigenvalue can be written in the terms of power series in associated polynomials. We restrict ourselves to $p, q \in \mathbb{Z}$ which allows us to use the extended associated sequences. For the ground state, we get

$$\psi_{0,0}^D(x,y) = \sum_{k=0}^{\infty} \frac{(-\omega)^k}{2^k k!} p_{2k+p}(x) \sum_{l=0}^{\infty} \frac{(-\omega)^l}{2^l l!} p_{2l+q}(y),$$

and the excited states can be computed as

$$\psi_{n,m}^D(x,y) = L_n^{\left(p-\frac{1}{2}\right)}\left(\omega(x\beta_x)^2\right) L_m^{\left(q-\frac{1}{2}\right)}\left(\omega(y\beta_y)^2\right)\psi_{0,0}^D.$$

The expressions $x\beta_x$ and $y\beta_y$ in the expansion of Laguerre polynomials act as umbral shifts for the associated polynomials in the ground state $\psi_{0,0}^D$. The issue of convergence reduces to the convergence of the infinite series of discrete Gaussians in $\psi_{0,0}$.

Similar transfer to the lattice in case of the Hamiltonian (7) follows as

$$h_{II}^D = -\frac{1}{2}\Delta_t^2 + t\beta_t\Delta_t - u\beta_u\Delta_u^2 + u\beta_u\Delta_u - (q+\tfrac{1}{2})\Delta_u,$$

which is solved by

$$\Xi_{n,m}^D(t,u) = H_n(t\beta_t)L_m^{\left(q-\frac{1}{2}\right)}(u\beta_u)\cdot 1 = \sum_{i=0}^{n} h_{i,n}p_i(t)\sum_{j=0}^{m} l_{j,m}^{\left(p-\frac{1}{2}\right)}p_j(u).$$

The number $h_{i,n}$ is the i-th coefficient of the n-th Hermite polynomial $H_n(t)$, other notation as before.

Similarly, the original operator (5) takes the form

$$\mathcal{H}_{II}^D = -\frac{1}{2}(\Delta_x^2+\Delta_y^2) + 2\omega^2(x\beta_x)^2 + \frac{\omega^2}{2}(y\beta_y)^2 + \frac{\beta}{2}(y\beta_y)^{-2}.$$

The expression for the ground state is for $q \in \mathbb{Z}$

$$\psi_{0,0}^D(x,y) = \sum_{k=0}^{\infty}\frac{(-\omega)^k}{k!}p_{2k}(x)\sum_{l=0}^{\infty}\frac{(-\omega)^l}{2^l l!}p_{2l+q}(y),$$

For the excited states we get

$$\psi_{n,m}^D(x,y) = H_n(\sqrt{2\omega}x\beta_x)L_m^{\left(q-\frac{1}{2}\right)}\left(\omega(y\beta_y)^2\right)\psi_{0,0}(x,y).$$

The Hermite and Laguerre polynomials of the arguments $x\beta_x$ and $(y\beta_y)^2$ (up to constants) are the umbral shifts acting on the appropriate parts of the wave function $\psi_{0,0}^D$. The convergence is not affected by these terms.

4.2 Particular Discretization

In this paragraph we show the results for the particular difference operator mentioned in Section II. The solutions obtained by the umbral correspondence are well-defined on the lattice points $\sigma\mathbb{Z}$ (at least positive) and in the case of gauge-rotated Hamiltonian they converge everywhere. For brevity, the results will be demonstrated on the generalized isotropic harmonic oscillator.

With the right discrete derivative, we denote the spacings on the lattice as (σ_t, σ_u) or (σ_x, σ_y) (according to the coordinates used). Similarly, the shift operators are denoted T_{σ_t}, T_{σ_u} etc. The operator (3) is discretized as

$$h_+^D = \frac{1}{\sigma_t^2}\left[\left((\sigma_t+2)t+\sigma_t(p+\tfrac{1}{2})\right) - \left(t+\sigma_t(p+\tfrac{1}{2})\right)T_{\sigma_t} - t(\sigma_t+1)T_{\sigma_t}^{-1}\right] +$$
$$+\frac{1}{\sigma_u^2}\left[-\left(u+\sigma_u(q+\tfrac{1}{2})\right)T_{\sigma_u} + \left((\sigma_u+2)u+\sigma_u(q+\tfrac{1}{2})\right) - u(\sigma_u+1)T_{\sigma_u}^{-1}\right].$$

Written as a difference equation:

$$\frac{1}{\sigma_t^2}\left[-\left(t+\sigma(p+\tfrac{1}{2})\right)\Xi(t+\sigma_t,u)-t(\sigma_t+1)\Xi(t-\sigma_t,u)+\left((\sigma_t+2)t+\sigma(p+\tfrac{1}{2})\right)\Xi(t,u)\right]+$$
$$+\frac{1}{\sigma_u^2}\left[u(\sigma_u+1)\Xi(t,u-\sigma_u)+\left((\sigma_u+2)u+\sigma_u(q+\tfrac{1}{2})\right)\Xi(t,u)-\left(u+\sigma_u(q+\tfrac{1}{2})\right)\Xi(t,u+\sigma_u)-\right]=$$
$$=e\Xi(t,u).$$

The solution is a polynomial and can be expressed as

$$\Xi_+^D(t,u)=\sum_{i=0}^{n} l_{i,n}^{(p-\frac{1}{2})}\prod_{r=0}^{i-1}(t-r\sigma_t)\sum_{j=0}^{m} l_{j,m}^{(q-\frac{1}{2})}\prod_{s=0}^{j-1}(u-s\sigma_u).$$

If we return to the original problem and use the right discrete derivative, the eigenvalue problem can be formulated by the following difference equation on a lattice:

$$\left[\frac{\alpha}{2(x+\sigma_x)(x+2\sigma_x)}-\frac{1}{2\sigma_x^2}\right]\psi(x+2\sigma_x,y)-\frac{1}{2\sigma_x^2}\psi(x,y)+\frac{1}{\sigma_x^2}\psi(x+\sigma_x,y)+\frac{\omega^2}{2}x(x-\sigma_x)\psi(x-2\sigma_x,y)+$$
$$+\left[\frac{\beta}{2(y+\sigma_y)(y+2\sigma_y)}-\frac{1}{2\sigma_y^2}\right]\psi(x,y+2\sigma_y)-\frac{1}{2\sigma_y^2}\psi(x,y)+\frac{1}{\sigma_y^2}\psi(x,y+\sigma_y)+\frac{\omega^2}{2}y(y-\sigma_y)\psi(x,y-2\sigma_y)=$$
$$=E\psi(x,y).$$

Using the extended associated sequence, the ground state for this eigenvalue problem can be written as

$$\psi_{0,0}^D(x,y)=\left[\sum_{k=0}^{\left[-\frac{p+1}{2}\right]}\frac{(-\omega)^k}{2^k k!}\cdot\frac{1}{\prod_{i=-1}^{2k+p}(x-i\sigma_x)}+\sum_{k=\left[-\frac{p+3}{2}\right]}^{\infty}\frac{(-\omega)^k}{2^k k!}\prod_{i=0}^{2k+p-1}(x-i\sigma_x)\right]\times$$
$$\times\left[\sum_{k=0}^{\left[-\frac{q+1}{2}\right]}\frac{(-\omega)^k}{2^k k!}\cdot\frac{1}{\prod_{i=-1}^{2k+q}(y-i\sigma_y)}+\sum_{k=\left[-\frac{q+3}{2}\right]}^{\infty}\frac{(-\omega)^k}{2^k k!}\prod_{i=0}^{2k+q-1}(y-i\sigma_y)\right]$$

where we use the blanket hypothesis that $\sum_{k=0}^{-c}=0$ for c positive. This function solves the difference equation on the lattice points $\{(i\sigma_x,j\sigma_y)\mid i,j=0,1,2,\ldots\}$, that is for the first quadrant in $\mathcal{E}_2$. In other points the series diverges.

The excited states can be obtained as

$$\psi_{n,m}^D(x,y)=\sum_{j=0}^{n} l_{j,n}^{(p-\frac{1}{2})}\omega^j\sum_{k=0}^{\infty}\frac{(-\omega)^k}{2^k k!}p_{2k+2j+p}^{+}(x)\times\sum_{i=0}^{m} l_{i,m}^{(q-\frac{1}{2})}\omega^i\sum_{l=0}^{\infty}\frac{(-\omega)^l}{2^l l!}p_{2l+2i+q}^{+}(y)$$

where $p_n^+(x)$ is the generalized associated sequence for the right discrete derivative. For $p,q\in\mathbb{Z}_0^+$ these polynomials can be easily substituted.

The results for the left and symmetric discrete derivatives would be obtained in similar fashion.

5 Conclusions

We have shown that certain two-dimensional quantum-mechanical superintegrable systems can be transferred to a uniform lattice by the means of umbral discretization. The difference equations for the generalized isotropic harmonic oscillator using a simple example of right discrete derivative have been found and the solutions have been obtained by substituting into the original ones. In the case of gauge-rotated Hamiltonian, the solution of the difference analogue is a well-defined polynomial in $\mathcal{E}_2$, however, for the original system, we need to restrict ourselves to the lattice points only.

The method of umbral discretization offers an infinite number of difference operators (that approximates the derivative in an arbitrary order) and therefore this procedure can be done with various operator replacements. Moreover, there is no restriction for the dimension of the system, nor for the coordinate system. Therefore the difference analogues of quantum mechanics can be formulated on non-square lattices as well, while still preserving the symmetries.

Acknowledgments

The author would like to thank S. Pošta, M. Havlíček, P. Winternitz and A. Turbiner for helpful discussions and CRM Université de Montréal for its hospitality and support. This work was partially supported by research grant SGS.

References

[1] A. Ashtekar. *Quantum geometry and gravity: Recent advances.* In 'General Relativity and Gravitation', number 28, Word Scientific (2001).

[2] N. M. Atakishiev and S. K. Suslov. *Difference analogs of the harmonic oscillator.* Theoret. Math. Phys. **85** (1991), 1055–1062.

[3] V. Bargmann. *Zur theorie des wasserstoffatoms.* Z. Phys. **99** (1936), 578.

[4] M. Creutz. *Quarks, Gluons and Lattices.* Cambridge University Press, Cambridge, (1983).

[5] A. Dimakis, F. Müller-Hoissen, and T. Striker. *Umbral calculus, discretization, and quantum mechanics on a lattice.* J. Phys. A: Math. Gen. **29** (1996), 6861–6876.

[6] V. Fock. *Zur theorie des wasserstoffattoms.* Z. Phys. **98** , 145.

[7] J. Fris, V. Mandrosov, Y. A. Smorodinsky, and P. Winternitz. *On higher symmetries in quantum mechanics.* Phys. Lett. **16** (1965), 354.

[8] J. Jauch and E. Hill. *On the problem of degeneracy in quantum mechanics.* Phys. Rev. **57** (1940), 641.

[9] D. Levi, P. Tempesta, and P. Winternitz. *Umbral calculus, difference equations and the discrete schrödinger equation*. J. Math. Phys. **45** (2004), 4077.

[10] M. Lorente. *Continuous vs. discrete models for the quantum harmonic oscillator nad the hydrogen atom*. Phys. Lett. A **285** (2001), 119–126.

[11] S. Odake and R. Sasaki. *Shape invariant potentials in discrete quantum mechanics*. J. Nonlin. Math. Phys. **12** (2005), 507–521.

[12] S. Odake and R. Sasaki. *Orthogonal polynomials from hermitian matrices*. J. Math. Phys. **49** (2008), 053053.

[13] W. Pauli. *Über das wasserstoffspektrum von standpunk der neuen quanten-mechanik*. Z. Phys. **36** (1926), 336–363.

[14] S. Roman. *The Umbral Calculus*. Academic Press, San Diego, (1984).

[15] G.-C. Rota. *Finite Operator Calculus*. Academic Press, San Diego, (1975).

[16] P. Tempesta, A. Turbiner, and P. Winternitz. *Exact solvability of superintegrable systems*. J. Math. Phys. **42** (2001), 4248.

[17] A. Turbiner. *Canonical discretization: I. discrete faces on (an)harmonic oscillator*. Int. J. Mod. Phys. A **16** (2001), 1579.

[18] P. Winternitz, Y. A. Smorodinsky, M. Uhlir, and I. Fris. *Symmetry groups in classical and quantum mechanics*. Sov. J. Nucl. Phys. **4** (1967), 444.

Analýza pravděpodobnostního rozdělení ve více-segmentovém buněčném termodynamickém modelu

Katarína Kittanová

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

školitel: Milan Krbálek, Katedra matematiky,
Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. This article is dealing with microscopic structure of multi-segment cellular thermal-like model. In the beginning the meaning of congestions as macroscopic phenomenon is reminded. The simulation procedure is described. Then the clearance probability density is analyzed inside segments and also at the edges. The function derived for the one-segment model is tried to be used to approximate the probability density also for multi-segment model.

Keywords: thermal-like traffic model, clearance distribution and traffic congestion

Abstrakt. Tenhle článek se zaobírá analýzou mikroskopické struktury více-segmentového buněčného termodynamického modelu. V úvodu je shrnut význam výskytu kongescí jako zástupce makroskopických jevů. Připomenuta je simulační procedura vedoucí k ustálenému stavy daného systému, který je následně zkoumán. Dále je pozornost věnovaná pravděpodobnostnímu rozdělení vzdáleností a to nejen uvnitř segmentu ale taky na jeho okrajích. Testuje se, jestli je funkce popisující uvedené rozdělení pravděpodobností pro jednodušší verzi modelu jenom s jedním segmentem vhodná i pro hraniční úseky.

Klíčová slova: termální dopravní model, pravděpodobnostní rozdělení vzdáleností a dopravní zácpa

1 Úvod

První pokus o matematický popis dopravního systému se datují k roku 1935 a vděčíme za něj doktorovi Michiganské university Bruceovi Greenshieldsovi. Od té doby ho následovalo mnoho vědců používajících různé přístupy. Požadavků kladených na dopravní modely je několik: analytická řešitelnost, nenáročná simulační procedura, mikroskopická a makroskopická struktura odpovídající reálným vzorkům, atd. Snahou je představit univerzální model postihující všechny fenomény vyskytující se v cestní dopravě. Zároveň by počet vstupních parametrů takového modelu neměl být neúměrně vysoký.

Našim cílem je nechat se inspirovat některými úspěšnými dopravními modely při modifikaci a tak postoupit o krůček dál v univerzifikaci. Jako základ slouží model založen na vlastnostech termodynamického plynu.

1.1 Termodynamický dopravní model

V krátkosti připomeneme východiskový model. Jedná se jednorozměrný modifikovaný Dysnův plyn, používaný pro analýzu mikroskopické struktury. Jednotlivé částice jsou identické s hmotností m. Uvažuje se soubor N částic umístněných na kružnici. Pozice částice je dána uhlovou souřadnicí φ_i, kde i udává pořadové číslo částice. Byla zvolena krátkodosahová varianta vzhledem k lepší korespondencí s chováním řidičů v cestní dopravě. Konkrétně částice v uvažovaném modelu interaguje pouze s nejbližší předchozí částicí. Míra vzájemné interakce je závislá pouze na vzdálenosti mezi částicemi r_i a značí se $V(r_i)$. Hamiltonián uvedeného souboru má pak tvar

$$H = \sum_{i=1}^{N} \frac{1}{2} m(v_i - \bar{v})^2 + \sum_{i=1}^{N} V(r_i),$$

přičemž v_i značí rychlost i-té částice a $\bar{v}$ průměrnou rychlost souboru. Teď ještě zbývá určit exaktní tvar odpudivého potenciálu $V(r_i)$. V starších verzích termodynamického modelu se pracovalo s logaritmickým tvarem $V(r_i) = \ln(r_i)$, pro který byl model relativně snadno analyticky řešitelný. Později se však ukázalo jako vhodnější popisovat vzájemnou interakci částic pomocí newtonovské odpudivé sily $F(r_i) \propto= -\frac{1}{r_i^2}$ a tedy používat potenciál

$$V(r_i) \propto= -\frac{1}{r_i}.$$

Průměrná vzdálenost částic $\langle r \rangle$ je normovaná na hodnotu 1, co znamená obvod kružnice, na které se částice pohybují roven N.
Z reálných dopravných měření a následné analýzy dat lze lehce vypozorovat měnící se mikroskopickou strukturu v závislosti na aktuální dopravní situaci. Ve zkoumaném modelu je vliv tohoto faktoru reprezentován teplotní lázní, ve které je soubor částic umístněn. Pohyb částic je pak ovlivněn její termodynamickou teplotou T. Z praktických důvodů je zavedena inverzní termodynamická teplota β definovaná vztahem

$$\beta = \frac{1}{kT},$$

kde k značí Boltzmannovu konstantu.
β představuje významný parametr celého modelu reprezentující míru dopravního stresu ovlivňujícího počínání řidiče.
Jednou z hlavních zkoumaných statistik bude pravděpodobnostní rozdělení vzdálenosti sousedních částic. To lze odvodit [1] z Hamiltoniánu daného systému a má tvar

$$P(r) = \Theta(r) A exp[-\frac{\beta}{r} - Br], \tag{1}$$

kde $\Theta(r)$ označuje Heavisidovu funkci, β představuje již zmíněnou inverzní termodynamickou teplotu a A a B jsou normalizační konstanty, získávané z dvou normalizačních rovnic:

$$\int_0^\infty A exp[-\frac{\beta}{r} - Br] \mathrm{d}r = 1,$$

$$\langle r \rangle = \int_0^\infty rAexp[-\frac{\beta}{r} - Br]\mathrm{d}r = 1.$$

První vztah musí splňovat každé pravděpodobnostní rozdělení, druhý zajišťuje normalizaci střední hodnoty r na 1. Hodnoty obou normalizačních konstant jsou závislé na hodnotě parametru β a lze je přibližně aproximovat funkcemi

$$B = \beta + \frac{3 - exp[sqrt\beta]}{2},$$

$$A^{-1} = 2\sqrt{\frac{\beta}{B}}K_{1(2\sqrt{\beta B})},$$

kde $K_1(x)$ označuje modifikovanou Besselovou funkci druhého druhu.
Získaný tvar pravděpodobnostního rozdělení vzdáleností odpovídá struktuře reálných dopravných dat, jak již bylo dokázáno. Tato funkce je vhodnou aproximací jak u vzorků v režimu volné dopravy, tak u těch, kde dochází ke zhuštění, rozdíl je v hodnotě parametru β. Přibližně platí, že režim volné dopravy koresponduje s nižšími hodnotami inverzní termodynamické teploty, zatímco při zácpách hodnota tohohle parametru stoupá. Pro extrémní případ $\beta = 0$ jde o Poissonovo rozdělení, které se používá pro nezávisle se pohybující částice.

2 Více-segmentový buněčný termodynamický modelu

V některých pracích [4] byla hodnota parametru β nastavena globálně, tj. všechny uvažované částice byly ve stejném dopravním režimu. Bylo již ukázáno [1], že volbou vhodné hodnoty inverzní termodynamické teploty β lze uspokojivě aproximovat reálná data získané pro různé dopravní režimy. Obecně lze říct, že volnému dopravnímu režimu odpovídají nižší hodnoty parametru β, zatímco pro synchronizovaný dopravní režim lze dosáhnout vyšší hodnoty, kolem $\beta = 3$. Extrémním případem je pak $\beta = 0$, kdy jde o Poissonovo rozdělení používáno pro zcela nezávisle se pohybující elementy.
V takhle definovaném modelu se nevyskytují pozorovatelné zácpy. Absence tohoto makroskopického dopravního fenoménu je způsobena globálním nastavením parametru β. Znamená to, že všechny uvažované elementy jsou ve stejném režimu, tedy přímo v případné kongesci. Navíc dochází k normalizaci vzdáleností mezi částicemi, takže je nelze využít k detekci zácpy.
Abychom uvedený makroskopický fenomén v termodynamickém modelu obsáhly je potřeba provést modifikaci. Konkrétně jde o změnu definice inverzní termodynamické teploty β z globální na lokální.
Pro naše účely zatím postačí rozdělit uvažovanou kružnici na dva segmenty s různými hodnotami parametru β. Cílem je přiblížit se modelování dopravní situace, kdy volnou dopravu komplikuje nějaký druh překážky. Tahle překážka je představována menším ze segmentů, který nazveme kritickou oblastí, a budou pro ní voleny vyšší hodnoty parametru označovány β_c. Pro hlavní segment reprezentující volnou dopravu bude odpovídající hodnota inverzní termodynamické teploty značena jako β_0. Jelikož uvažujeme stacionární překážku, segmenty jsou definovány pevně, pomocí uhlových souřadnic: volný segment od

$\varphi = 0$ po φ_c a kritická oblast od φ_c po $\varphi = 2\pi$. V následujících simulacích bylo zvoleno $\varphi_c = \frac{3}{2}\pi$.

Další modifikací představuje použití diskrétních vzdáleností v simulační proceduře. Jako inspirace sloužil slavný Nagelův-Schreckenbergův buněčný model, kterého velkou předností je jednoduchost.
Kružnice délky N je rozdělena na m stejně dlouhých buněk. Dále uvažujeme částice jako bezrozměrné elementy pohybující se skokově mezi buňkami. Takže je jednoznačně dáno, ve které buňce se částice aktuálně nachází a jestli je buňka obsazená částicí nebo nikoliv. V jedné buňce se může nacházet maximálně jedna částice.
Experimentálně bylo zjištěno, že při použití dostatečně podrobného dělení na buňky se tvar pravděpodobnostního rozdělení vzdáleností zachovává.

2.1 Simulační metoda

Pro numerické výpočty byla použita simulační metoda založená na Metropolisově algoritmu.
Nejdřív se vygeneruje počáteční rozmístění částic. Použít se může jak náhodné tak ekvidistantní rozložení. Pak se upravují pozice částic podle následujícího algoritmu:

- Pro aktuální rozmístění částic se vypočte hodnota potenciální energie podle vztahu

$$U = \sum_{l=1}^{N} \frac{1}{r_l}. \tag{2}$$

- Náhodně je zvolen index $l \in \{1, 2, ..., n\}$.

- Je vybrána aktuální hodnota inverzní termodynamické teploty vzhledem k umístnění l-té částice definovanému příslušným uhlem φ_l podle vztahu

$$\beta_l = \beta_0 \Theta(\varphi_l)\Theta(\varphi_c - \varphi_l) + \beta_c \Theta(\varphi_l - \varphi_c)\Theta(2\pi - \varphi_l)$$

- Je vygenerováno číslo δ rovnoměrně rozdělené na intervalu $(0, 1)$.

- Délka skoku je ještě diskretizována

$$\tilde{w} = \lceil \frac{m}{N}\delta \rceil.$$

- Je vypočtena nová předpokládaná pozice $l$té částice podle vztahu $\acute{x}_l = x_l + \tilde{w}$. V uvažovaném modelu není povoleno předbíhání, proto může být nová pozice $\acute{x}_l$ akceptována pouze v případě, že je nerovnost $\acute{x}_l < x_{l+1}$ splněna.

- Nová hodnota potenciální energie $\acute{U}$ je vypočtena pro konfiguraci, kde je poloha $l$té částice dána $\acute{x}_l$.

- Pokud je splněno $\acute{U} < U_0$, $l$tá částice zaujme polohu $\acute{x_l}$, jinak je potřeba vypočítat Boltzmannův faktor q

$$q = exp^{-\beta_l \Delta U} = exp^{-(\acute{U} - U_0)}$$

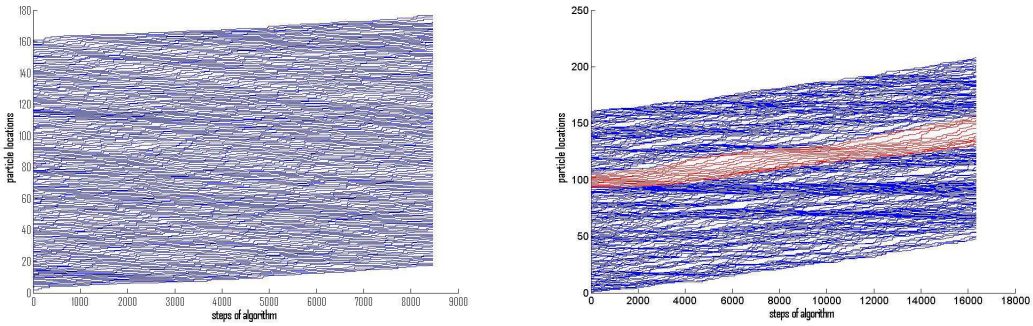
  Pak je vybráno náhodné číslo g rovnoměrně rozděleno na intervalu $(0, 1)$ a porovnáno s Boltzmannovým faktorem. Při splnění nerovnosti $q > g$ je skok přijat, jinak zůstává konfigurace nezměněna.

Tenhle postup provádí až do dosažení termální rovnováhy, která je charakterizována stabilní hodnotou potenciální energie U. Její hodnota nezávisí na počátečním rozmístění částic.

3 Výsledky simulací

3.1 Kongesce

Připomeňme si nyní výsledky získané v předchozí práci a sice evidenci kongescí u numerických simulací.
Dobrou metodou na jednoduchou detekci dat je vizuální zkoumání grafického znázornění trajektorií jednotlivých částic. Pro porovnání jsou vyobrazeny jak trajektorie pro více-segmentový model, tak pro model s globálně definovanou hodnotou inverzní termodynamické teploty β.

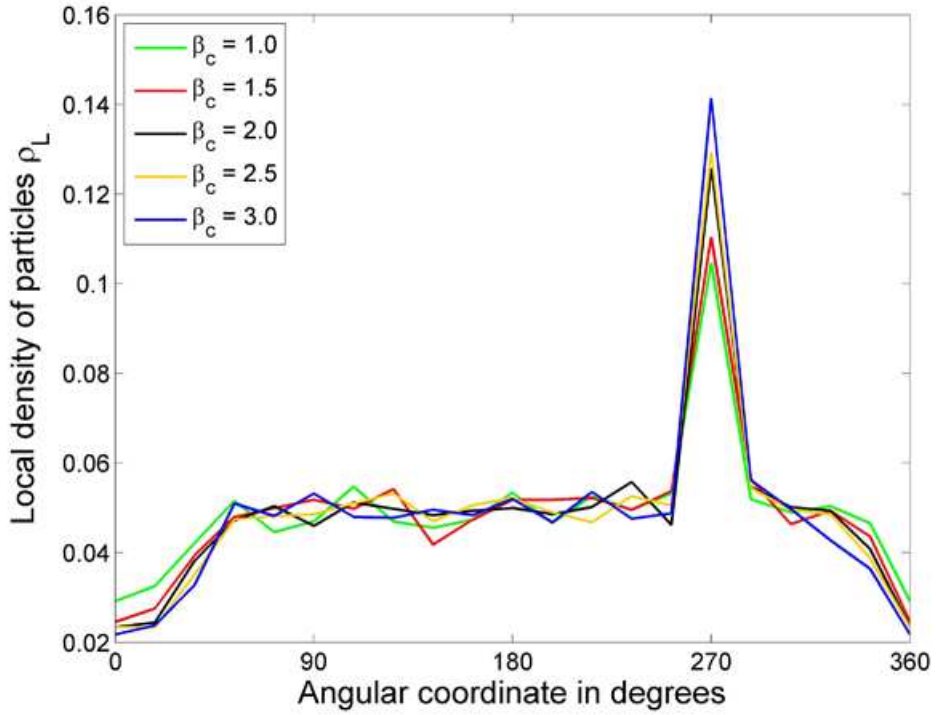


Porovnání trajektorií částic pro model s globální hodnotou β a více-segmentový model.

Z grafů (3.1) lze snadno odvodit, že ke kongescím dochází pouze u více-segmentového modelu. V grafické reprezentaci trajektorií pro tenhle model je detekováno několik výrazných zhuštění šířících se proti směru pohybu částic napříč celým souborem. Lze tedy dojít k závěru, že výše popsaná modifikace lokálního termodynamického plynu je dostačeným zobecněním vyvolávajícím vznik kongescí jako zástupce makroskopických fenoménů.

3.2 Rozmístění částic

Teď se zaměříme na podrobnou analýzu pozic částic ve stavu termální rovnováhy. Zkoumány jsou konečné konfigurace po ustálení hodnoty potenciální energie U. V následujících simulacích je použito $\beta_0 = 0,1$ a hodnoty β_c jsou voleny mezi 1 a 3.Simulace jsou pro

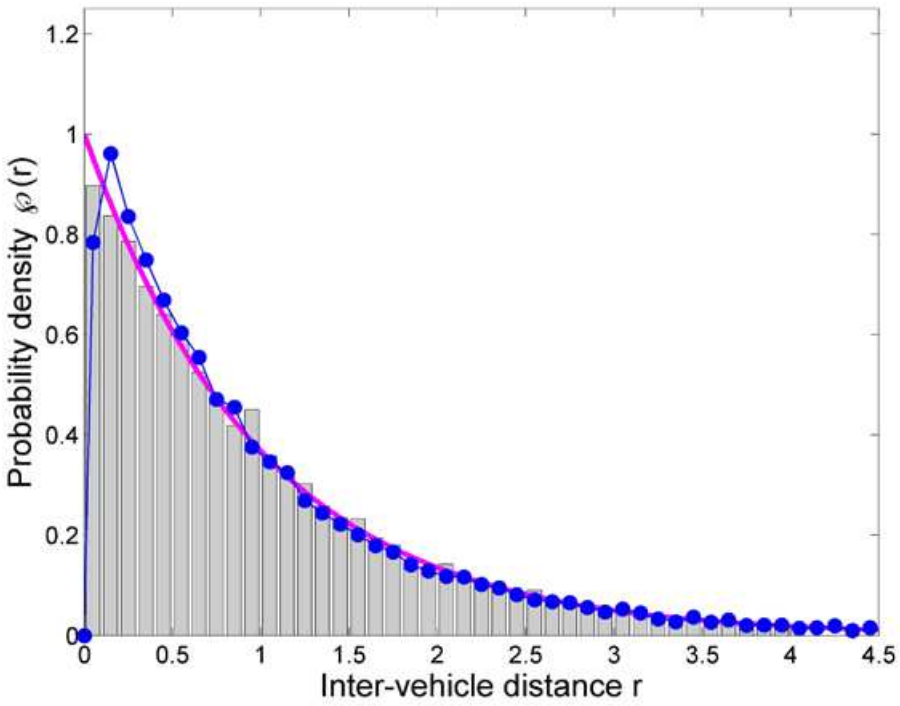


Obrázek 1: Graf hustotních profilů více-segmentového modelu

každý soubor parametrů několikrát opakovány, aby byl získán dostatek dat.
Než bude pozornost věnována pravděpodobnostním rozdělením vzdáleností, je zajímavé zaobírat se hustotním profilem zkoumaných dat. Pro model s globální hodnotou β je při dostatečně velkém objemu dat průměrná lokální hustota po ustálení termální rovnováhy ve všech částech stejná. Nyní se podíváme na hustotní profil více-segmentového systému.

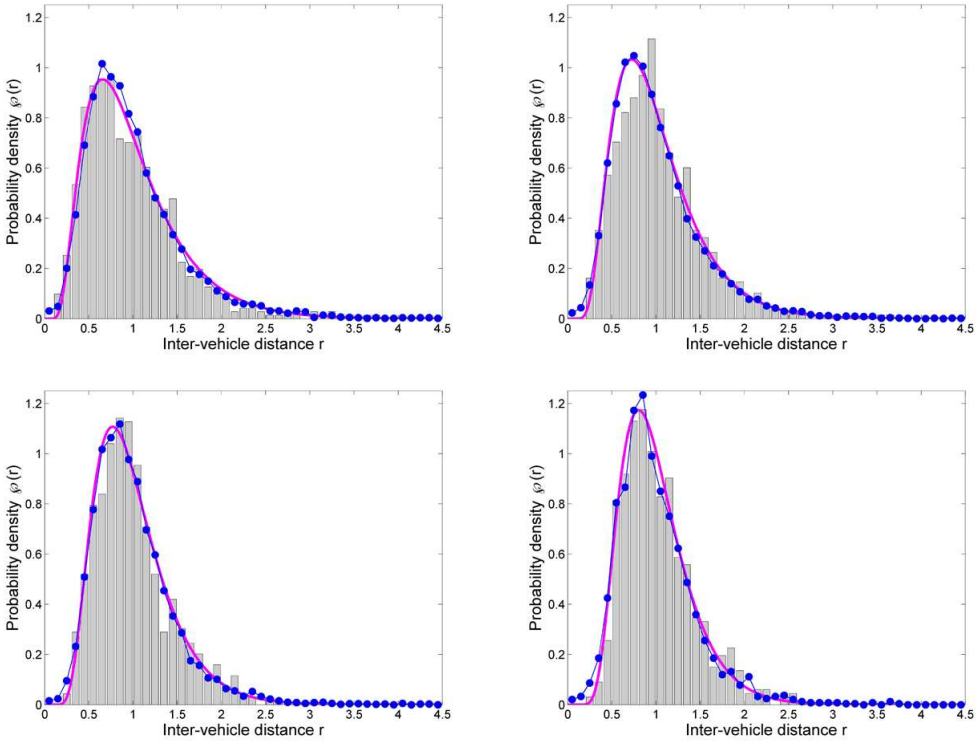
Na grafu (1) jsou patrné dva výrazné výkyvy v hustotních profilech. První představující významné zhuštění, kterému by v reálné dopravě odpovídala hustá zácpa, a druhý naopak podprůměrnou hustotu, ta může korespondovat se situací, kdy končí nějaké dopravní omezení. Vrcholy těchto výkyvů přesně odpovídají přelomům segmentů. Při nárůstu parametru β dochází ke kongesci a při poklesu naopak k ředění dopravy. Jelikož inverzní termodynamická teplota β představuje míru dopravního stresu a při její vyšší hodnotě předpokládáme snížení rychlosti, odpovídá uvedené pozorování předpokladu. Zajímavý je taky fakt, že v středních částech obou segmentů je hodnota průměrné lokální hustoty stejná. Kongesce představuje výraznější výkyv ale k zhuštění i rozpuštění dochází na kratším úseku. Další analýzou hustotních profilů lze odvodit, že velikost výkyvů roste s rozdílem hodnot β_0 a β_c.
A teď můžeme přistoupit ke zkoumaní pravděpodobnostních rozdělení vzdáleností sousedních elementů v různých částech uvažované kružnice.

Vzhledem k nastavení parametrů pro numerické simulace, je nejvíc dat k dispozici pro hodnotu inverzní termodynamické teploty $\beta_0 = 0,1$. Připomeňme, že se nepoužívají data z okrajových oblastí, aby se eliminoval vliv sousedního segmentu. Na příslušném



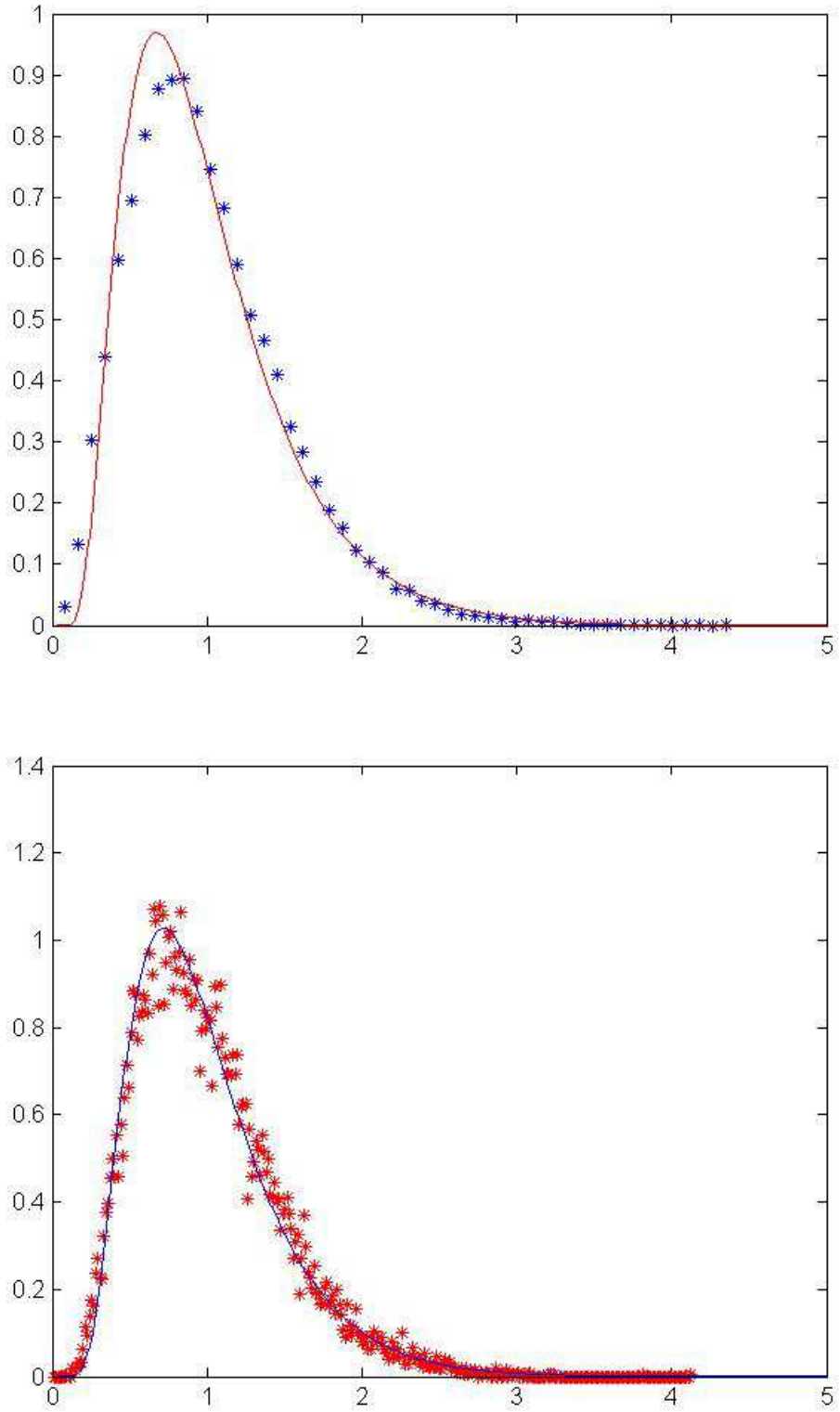
Obrázek 2: Graf pravděpodobnostního rozdělení vzdáleností pro $\beta_0 = 0,1$. Histogram reprezentuje výsledky numerických simulací, křivka analytickou aproximaci a body hodnoty získané z reálných dopravních dat.

grafu lze pozorovat nejlepší korespondenci dat získaných pomocí numerických simulací s analytickou předpovědí a taky s rozdělením pravděpodobnosti z reálných dopravných dat naměřených v odpovídajícím, volném dopravním režimu.
Důležité však je analyzovat i ostatní úseky. U kritické oblasti lze opět brát jenom střed segmentu a porovnat je s analytickými předpověďmi.



Grafy pravděpodobnostního rozdělení vzdáleností pro $\beta_c = 1; 1,5; 2; 2,5$. Histogramy reprezentují výsledky numerických simulací, křivky analytickou aproximaci a body hodnoty získané z reálných dopravních dat.

Výsledky pro kritické oblasti jsou zobrazeny na grafech (3.2). Vzhledem k menšímu počtu pro analýzu použitelných dat, nejsou uvedené histogramy tolik přesné. Stejně lze vypozorovat relativně uspokojivou korespondenci s očekávanou analytickou aproximací. Obecně tedy lze předpokládat, že i při rozšíření na více segmentů se uvnitř nich zachová očekávané pravděpodobnostní rozdělení vzdáleností.
Zbývá ještě zjistit, co se děje na okrajích segmentů. Za tímhle účelem byly vytvořeny speciální simulace s nastavením parametrů $\beta_0 = 0,1$ a $\beta_c = 2,5$ zaznamenávající právě situaci na přelomu segmentů. Průměrná vzdálenost naměřená v těchto úsecích se výrazně liší od globální průměrné vzdálenosti, proto je nutné získané data nejdřív normalizovat. Otázkou je, jetsli je analytická funkce (1) použitelná taky pro aproximaci dat okrajových oblastí.



Grafy pravděpodobnostního rozdělení vzdáleností pro přelomy segmentů. Hvězdičky reprezentují výsledky numerických simulací a křivky analytickou aproximaci.

Grafy (3.2) znázorňují pravděpodobnostní rozdělení vzdáleností na přelomu segmentů.

Ta se aproximuje funkcí (1), přičemž je optimální hodnota parametru β nalezena numericky. Opět lze konstatovat, že získané data relativně dobře korespondují s analytickou aproximací. Vypočtené hodnoty parametru β jsou $\beta = 1.0976$ pro první přelom segmentů (zvyšování inverzní termodynamické teploty) a $\beta = 1.4560$ pro druhý přelom segmentů (snižování inverzní termodynamické teploty).

4 Závěr

Výsledky numerických simulací naznačují, že funkce (1) by mohla být univerzální funkcí pro popis mikroskopické struktury. Vzhledem k zjištěním týkajícím se přelomu segmentů lze předpokládat, že parametr β se nemění skokově ale plynule. Dalším krokem by tedy mohla být změna způsobu definování hodnot parametru inverzní teploty v závislosti na uhlové souřadnici a to například pomocí spojité křivky. Jinou výzvou je navrhnout proces dynamického výpočtu aktuální hodnoty β na základě lokální hustoty případně number variance. Takhle modifikovaný model by mohl vést ke vzniku tzv. "phantom traffic jam".

Literatura

[1] M. Krbálek. *Equilibrum distributions in thermodynamical traffic gas.* J. Phys. A: Math. Theor **40** (2007).

[2] Nagel, K. and M. Scgreckenberg. *A cellular automaton model for freeway traffic.* Journal de Physique I, France, **2** (1992).

[3] Krbálek, M. and D. Helbing *Determination of interaction potentials in freeway traffic from steady-state statistics.* Physica A, **333** (2004).

[4] Krbálek, M. *Inter-vehicle gap statistics on signal-controlled crossroads.* J. Phys. A: Math. Theor **42** (2009).

Application of a Degenerate Diffusion Method in Medical Image Processing*

Radek Máca

2nd 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. This contribution is an extended abstract of the paper [6]. The paper deals with segmentation of image data using a partial differential equation of level-set type. The first part of this paper describes the level-set formulation and modification of the level-set equation. The evolution process are controlled by the segmented image data in such a way that the edges of objects can be found. The semi-implicit complementary-volume numerical scheme is used for solving the level-set equation. The final part of the paper describes algorithm parameters and their setting used for segmentation of the left heart ventricle in the cardiac MRI images.

Keywords: cardiac MRI, co-volume method, image segmentation, level set method, PDE

Abstrakt. Tento příspěvek je rozšířeným abstraktem článku [6]. Tématem tohoto článku je segmentace obrazových dat pomocí parciální diferenciální rovnice vrstevnicového typu. První část se zabývá vrstevnicovou formulací k odvození vrstevnicové rovnice. Pro nalezení objektů v daném obraze je třeba tuto rovnici modifikovat. Pro numerické řešení vrstevnicové rovnice je použita metoda duálních objemů. Druhá část se zabývá vhodným nastavením výpočetních parametrů k dosažení co nejlepších výsledků při segmentaci levé srdeční komory na snímcích získaných pomocí magnetické rezonance.

Klíčová slova: metoda duálních objemů, segmentace obrazu, vrstevnicová rovnice

1 Introduction

The presented work is motivated by the need of medical practice for evaluation of the dynamical images of the heart obtained by the magnetic resonance imaging (cardiac MRI). One of important purposes of cardiac MRI examination is an estimation of parameters reflecting current clinical state of patients. A typical example could be an accurate measurement of heart ventricle volume during the heart contraction showing the contractive ability of myocardium. Within this framework, it is necessary to find the inner contour of the ventricle in the MR images. We attempt to adapt and modify a segmentation model based on numerical solution of a partial differential equation of the level set type. The iterative algorithm is controlled by the segmented image data in such a way that the edges of the objects can be found. The level set equation is solved by the semi-implicit

*This work has been supported by the grant of the Ministry of Education of the Czech Republic MSM6840770010 "Applied Mathematics in Technical and Physical Sciences". Partial support of the project "Jindřich Nečas Center for Mathematical Modeling", No. LC06052.

complementary-volume numerical scheme. We describe parameters and their setting used for segmentation of the left heart ventricle from the cardiac MRI images.

2 Mathematical model

Segmentation of the left heart ventricle volume is an important part of the cardiac MRI data post-processing. Examination of the heart ventricle consists of several hundreds of MR images covering the entire left ventricle volume and recording complete cardiac-cycle interval with a given temporal resolution. The MRI images are segmented separately by our approach based on level set formulation for the motion of the curve $\Gamma_t \subset \Omega, \Omega \subset \mathbb{R}^2$ propagating in the normal direction with speed V. J. A. Sethian, belonging to the authors who first contributed to the level set methods, wrote on this topic a comprehensive work [9].

Main idea is to describe the motion of $\Gamma(t)$ by means of the zero level set of a function $u : [0,T] \times \Omega \to \mathbb{R}$ such that

$$\Gamma(t) = \{x \in \Omega \,|\, u(t,x) = 0\}\,. \tag{1}$$

We define the signed distance function (SDF) needed for our approach:

Definition 2.1. *Let Γ be a closed curve in $\mathbb{R}^2$ for which $\Gamma_{\text{in}} = \operatorname{int}\Gamma$ and $\Gamma_{\text{out}} = \operatorname{ext}\Gamma$ are defined and satisfies $\Gamma = \partial\Gamma_{\text{in}} = \partial\Gamma_{\text{out}}$, $\Gamma_{\text{in}} \cup \Gamma \cup \Gamma_{\text{out}} = \mathbb{R}^2$. We define the signed distance function (d_Γ) as*

$$d_\Gamma(x) = \begin{cases} \operatorname{dist}(x,\Gamma) & x \in \Gamma_{\text{out}}\,, \\ 0 & x \in \Gamma\,, \\ -\operatorname{dist}(x,\Gamma) & x \in \Gamma_{\text{in}}\,, \end{cases}$$

where $\operatorname{dist}(x,\Gamma) = \min\{|x-y| \;|\; y \in \Gamma\}$.

For a given initial closed simple curve Γ_0, we can define u_{ini} as follows

$$u_{\text{ini}}(x) = u(0,x) = d_{\Gamma_0}(x) \qquad \forall x \in \Omega\,. \tag{2}$$

The Hamilton-Jacobi equation for u implicitly describes the motion of $\Gamma(t)$ by (1) reads

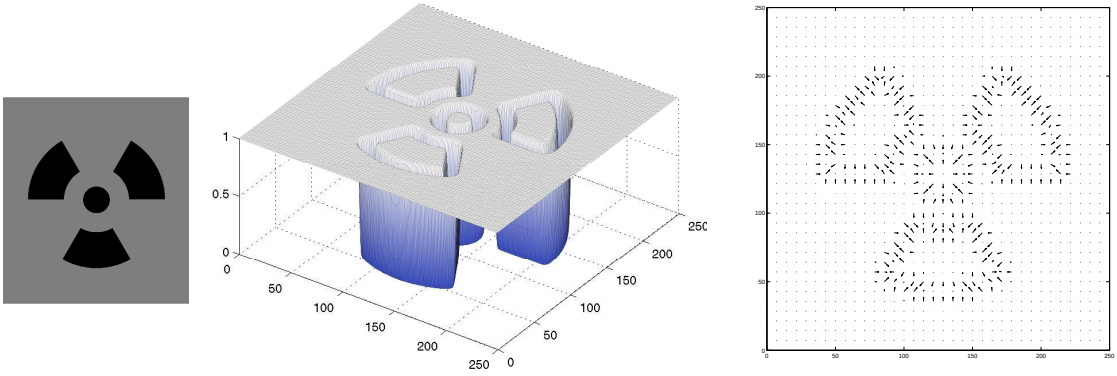
$$\frac{\partial u}{\partial t} + V|\nabla u| = 0\,. \tag{3}$$

The function $u(t,x)$ will be referred to as the segmentation function. Consider the following form of the normal velocity

$$V = -\kappa + F = -\nabla \cdot \frac{\nabla u}{|\nabla u|} + F\,, \tag{4}$$

where κ is the mean curvature of each level set defined as the divergence of its normal vector and F is an external force term. Substituting (4) to equation (3), we obtain the level set equation in the form

$$u_t = |\nabla u|\,\nabla \cdot \frac{\nabla u}{|\nabla u|} - |\nabla u|\,F\,,$$



(a) (b) (c)

Figure 1: The testing image (a) and the corresponding edge detector (b) together with the velocity field of the advection term (c).

where we denote $u_t := \partial u/\partial t$. Modification of the level set equation in the form

$$u_t = |\nabla u|_\varepsilon \nabla \cdot \frac{\nabla u}{|\nabla u|_\varepsilon} - |\nabla u|_\varepsilon F \,,$$

where $|\nabla u| \approx |\nabla u|_\varepsilon = \sqrt{\varepsilon^2 + |\nabla u|^2}$ denotes a regularization, can be used as a tool to prove existence of viscosity solution of the level set equation (see [3]). In this work ε is a computational parameter.

Detection of image object edges (boundaries) is a known task in image segmentation. Edges in the input image I^0 can be recognized by the magnitude of its spatial gradient ∇I^0. Application of the level set equation in this area requires an adaptation as follows

$$u_t = |\nabla u|_\varepsilon \nabla \cdot \left(g\left(\left|I^0 * \nabla G_\sigma\right|\right) \frac{\nabla u}{|\nabla u|_\varepsilon} \right) - g\left(\left|I^0 * \nabla G_\sigma\right|\right) |\nabla u|_\varepsilon F \,, \tag{5}$$

where $g : \mathbb{R}_0^+ \to \mathbb{R}^+$ is a non-increasing function for which $g(0) = 1$ and $g(s) \to 0$ for $s \to +\infty$. This function was first used by P. Perona and J. Malik ([8] in 1987) to modify the heat equation into a nonlinear diffusion equation which maintains edges in an image. Consequently, the function g is called the Perona-Malik function. We put $g(s) = 1/(1 + \lambda s^2)$ with $s \geq 0$. $G_\sigma \in \mathcal{C}^\infty(\mathbb{R}^2)$ is a smoothing kernel, e.g. the Gauss function with zero mean and variance σ^2

$$G_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|x|^2}{2\sigma^2}},$$

which is used in pre-smoothing (denoising) of image gradients by convolution

$$(I^0 * \nabla G_\sigma)(t) = \int_{\mathbb{R}^2} \bar{I}^0(t - \tau) \nabla G_\sigma(\tau) \, \mathrm{d}\tau \,,$$

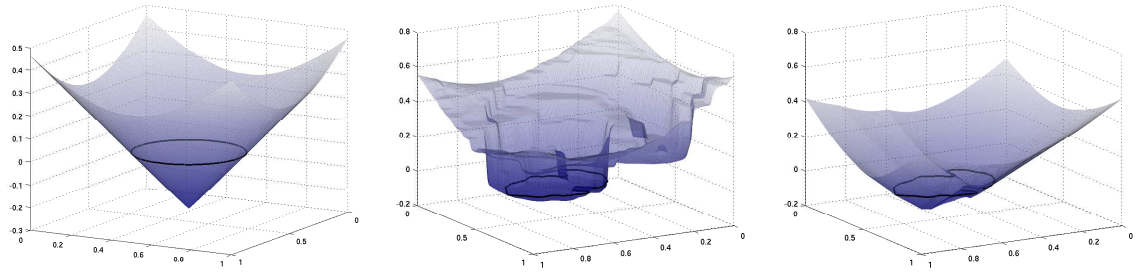


Figure 2: Example of segmentation function. Initial segmentation function u_0 (left), segmentation function u for $(t > 0)$ (middle), restored SDF (right)

where $\bar{I}^0$ is the extension of I^0 to $\mathbb{R}^2$ by mirroring, periodic prolongation or zero padding. Let us note that equation (5) can be rewritten into the advection-diffusion form

$$u_t = \underbrace{g^0 |\nabla u|_\varepsilon \nabla \cdot \left(\frac{\nabla u}{|\nabla u|_\varepsilon} \right)}_{(D)} + \underbrace{\nabla g^0 \cdot \nabla u}_{(A)} - \underbrace{g^0 |\nabla u|_\varepsilon F}_{(F)} . \tag{6}$$

For convenience, we use the abbreviation $g^0 = g(|I^0 * \nabla G_\sigma|)$. (D) in (6) denotes the diffusion term, (A) the advection term and (F) the external force term. The term g^0 is called the edge detector. For an example of an edge detector, see Fig. 1b. We can observe that value of the edge detector is approximately equal to zero close to the image edges. Here the evolution of the segmentation function slows down. On the contrary, in parts of the image with constant intensity the edge detector equals one. As we can see in Figure 1c, the advection term attracts the segmentation function to the image edges. We propose an advection parameter $\mathcal{A}$ to change the magnitude of the advection term. Finally, we obtain the final form of the modified level set equation, namely

$$u_t = g^0 |\nabla u|_\varepsilon \nabla \cdot \left(\frac{\nabla u}{|\nabla u|_\varepsilon} \right) + \mathcal{A} \nabla g^0 \cdot \nabla u - g^0 |\nabla u|_\varepsilon F . \tag{7}$$

2.1 Initial condition

A segmentation function $u(t, x)$ evolves from the initial guess (2). The initial curve Γ_0 has to be placed inside the segmented area (inside the left heart ventricle). To expand the initial curve, velocity (4) has to be positive. Positive value of V implies positive value of the external force F, rather $F > \kappa$. We use the signed distance function (SDF) for setting and restoring (redistancing) of the initial condition.

At the beginning of segmentation, i.e. for the first image, we have to place the initial curve Γ_0 into the left heart ventricle manually, e.g. as a circle. For a given Γ_0 we construct SDF d_{Γ_0} and set the initial condition as $u_{\text{ini}} = d_{\Gamma_0}$. The segmentation function u evolves from the initial guess (Fig. 2 left) according to (5). This evolution distorts the original shape of u_{ini} into $u(t, x)$ which fails to have unit gradient slopes (Fig. 2 middle). At the beginning of next image segmentation it is convenient to use the result of previous image

segmentation $\Gamma_t = \{x \in \mathbb{R}^2 \,|\, u(t,x) = 0\}$ and its signed distance function d_{Γ_t} as a new initial condition.

This is performed by means of the fast sweeping method introduced in [10]. This method is used to compute the viscosity solution of the eikonal equation

$$\begin{aligned} |\nabla u(x)| &= 1 \qquad x \in \Omega\,, \\ u(x) &= 0 \qquad x \in \Gamma \subset \Omega\,. \end{aligned}$$

Example of the restored signed distance function is shown in Figure 2 on the right.

3 Numerical scheme

A semi-implicit co-volume space discretization is used to solve (7) numerically. In [2], [4], [5], [7] a semi-implicit co-volume method discretizing (5) without the external force term is presented. First, we choose a uniform discrete time step τ. Then we replace time derivative in (7) by backward difference. Linear terms of the equation are considered at the current time level while the nonlinear terms (i.e. $|\nabla u|_\varepsilon$) are treated from the previous time level. In this way we obtain the following semi-implicit discretization

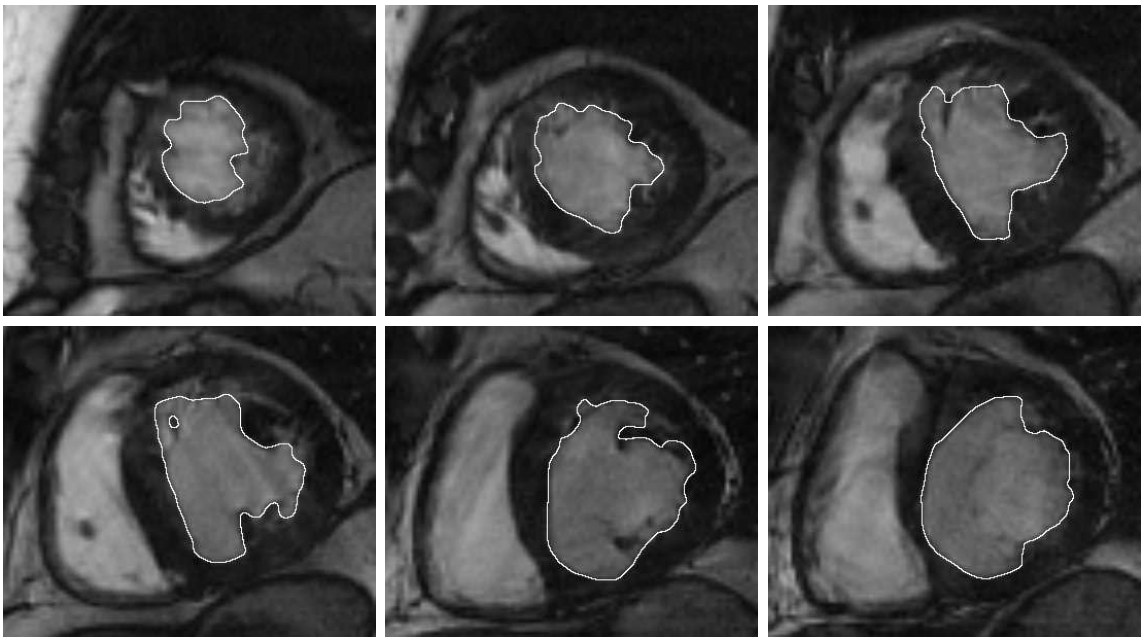
$$\frac{u^k - u^{k-1}}{\tau} = g^0 |\nabla u^{k-1}|_\varepsilon \nabla \cdot \left(\frac{\nabla u^k}{|\nabla u^{k-1}|_\varepsilon} \right) + \mathcal{A} \nabla g^0 \cdot \nabla u^k - g^0 |\nabla u^{k-1}|_\varepsilon F\,. \tag{8}$$

The derivation of numerical scheme using co-volume technique results in system of linear equation, which we solve by the SOR (Successive Over-Relaxation) iterative method.

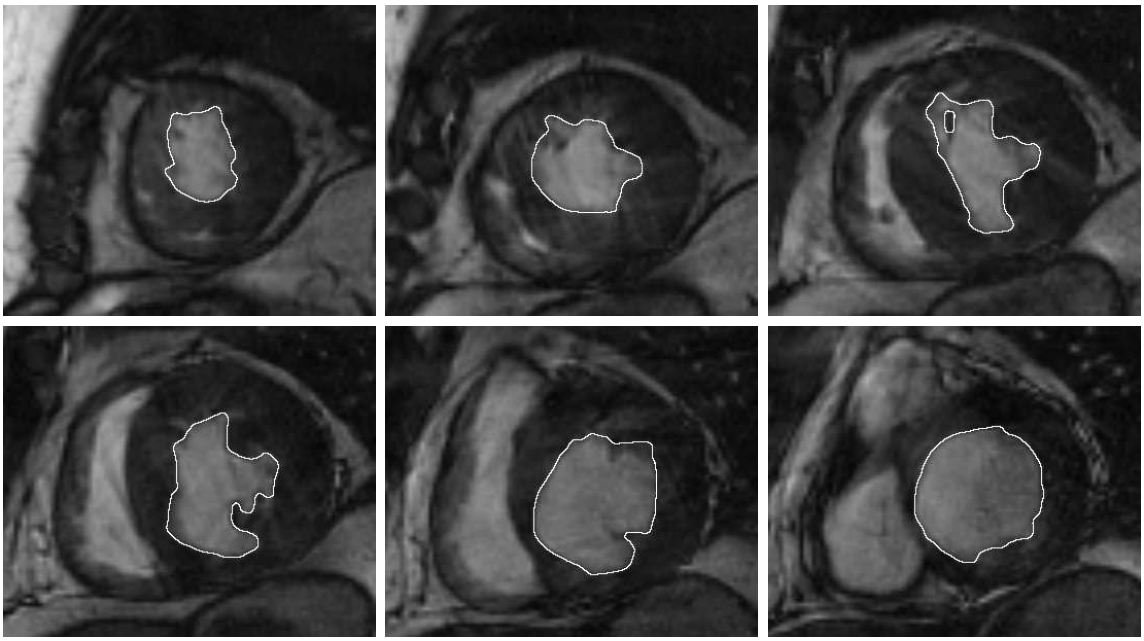
4 Results

To apply the scheme derived from (8), we have to specify correct values of the parameters of equation (7). The sensitivity of the edge detector depends on value of the parameter λ. Very low values of λ decrease the efficiency of edge detection. On the other hand, very high values of λ can cause detection of spurious edges (i.e. noise, blood flow artifacts, etc.). In our work we set $\lambda = 0.25$. The spatial discrete step is denoted by h and is given as $h = 1/(\max\{n_{x_1}, n_{x_2}\} - 1)$, the temporal discrete time step τ is given by $\tau = h/5$ and σ is set to the value $\sigma = 3h$.

We propose image dependent setting of the force parameter F and of the advection parameter $\mathcal{A}$. In the cardiac MR images obtained by means of the bright blood technique (see [1], chapter 4), the blood in the ventricle is lighter than the myocardium and the surrounding tissue. Also, blood in the ventricle has higher intensity than surrounding cardiac muscle. Using this information we can set a threshold I_{in} for picture elements certainly inside the ventricle and a threshold I_{out} for picture elements certainly in the myocardium and the surrounding tissue. These thresholds are set automatically using an algorithm based on minimum search on selected slices for a given initial condition. Using these thresholds we propose the threshold-dependent parameters F and $\mathcal{A}$ which are fundamental to obtain good segmentation results. On the other hand, the strong dependence of the algorithm on the thresholds I_{in} and I_{out} could be a hindrance. Indeed, wrong settings of these thresholds can cause incorrect segmentation results. Therefore it is



(a) Result of segmentation (end-diastole)



(b) Result of segmentation (end-systole)

Figure 3: Segmentation result for (a) end-diastole and (b) end-systole with parameters $h = 0.0028$, $\lambda = 0.25$, $\mathcal{A}_{\text{out}} = 2$, $F_{\text{out}} = -10$, $F_{\text{in}} = 50$.

important to apply robust automatic threshold selection. An example of the segmentation results can be seen in Fig. 3. The images are depicted in the end-diastolic phase (maximal volume of the ventricle) and in the end-systolic phase (minimum volume of the ventricle).

References

[1] Bogaert, J., Dymarkowski, S., Taylor, A. M.: Clinical cardiac MRI, *Springer-Verlag*, Berlin Heidelberg, (2005)

[2] Corsaro, S., Mikula, K., Sarti, A., Sgallari, F.: Semi-implicit co-volume method in 3D image segmentation, *SIAM Journal on Scientific Computing, Vol. 28*, No. 6 (2006), 2248–2265

[3] Evans, L. C., and Spruck, J.: Motion of level sets by mean curvature I, *J. Diff. Geom., Vol. 33*, (1991), 381–635

[4] Handlovičová, A., Mikula, K.: Stability and consistency of the semi-implicit co-volume scheme for regularized mean curvature flow equation in level set formulation, *Applications of Mathematics*, Vol. 53, No. 2 (2008), 105–129

[5] Handlovičová, A., Mikula, K., Sarti, A.: Numerical solution of parabolic equations related to level set formulation of mean curvature flow, *Computing and Visualization in Science , Vol.1, No. 3*, (1998), 179–182

[6] Máca R., Beneš M., Tintěra J., Application of a Degenerate Diffusion Method in Medical Image Processing, *Journal of Math-for-Industry*, accepted

[7] Mikula, K., Sarti, A., Sgallari, F.: Co-Volume Level Set Method in Subjective Surface Based Medical Image Segmentation, *Handbook of Biomedical Image Analysis, Springer US* (2005) 583–626

[8] Perona, P., Malk, J.: Scale space and edge detection using anisotropic diffusion, *Proc. IEEE Computer Society Workshop on Computer Vision* (1987)

[9] Sethian, J. A.: Level Set Methods, Cambridge University Press (1996)

[10] Zhao, H. K.: Fast sweeping method for eikonal equation, *Mathematics of Computation,* **74** (2005), 603–627

Rovnovážné fázové přechody při konstantním objemu, teplotě a chemickém složení systému

Kateřina Marková

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

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

Abstract. The traditional approach to equilibrium phase-splitting computation, based on the Gibbs energy minimization at given temerature, pressure and chemical composition of the mixture, cannot fully describe the state of a pure substance at the saturation pressure. This issue can be avoided by using the Helmholtz energy minimization instead of the traditional approach. The new approach allows to develop unified theory for both pure and multicomponent systems. It also simplifies the computation of phase-splitting in closed systems. This article briefly summarizes the thermodynamic formulation of the phase splitting at constant volume, temperature and moles. Two numerical methods are described and discussed, and finally, two examples solved with suggested algorithm are presented.

Keywords: equilibrium phase-splitting, VT-flash, Helmholtz energy, SSI method, Newton-Raphson method

Abstrakt. Tradiční metoda výpočtu rovnovážných fázových přechodů, založená na minimalizaci Gibbsovy energie při konstantním tlaku, teplotě a chemickém složení směsi, není schopná jednoznačně popsat stav čisté látky při saturačním tlaku. Tento problém zmizí, pokud je namísto tradičního přístupu minimalizována Helmholtzova volná energie. Nový přístup umožňuje formulaci společné teorie pro čisté látky a směsi. Další výhodou je zjednodušení popisu fázových přechodů v uzavřených systémech. Tato práce stručně popisuje termodynamickou formulaci fázových přechodů při konstantním objemu, teplotě a chemickém složení. Jsou popsány dva numerické modely a na závěr jsou ukázány dva příklady vyřešené navrhovaným algoritmem.

Klíčová slova: rovnovážné fázové přechody, VT-flash, Helmholtzova volná energie, metoda SSI, Newtonova-Raphsonova metoda

1 Úvod

Tradiční popis rovnovážných fázových přechodů (popsaný např. v [1]) využívá jako kritérium stability systému minimalizaci Gibbsovy energie, což je funkce tlaku, teploty a látkových množství jednotlivých komponent systému. Tzv. PT-rovnováha (*"Pressure-Temperature flash"*) je algoritmus, který při daných podmínkách určí množství a chemické složení jednotlivých fází v systému. Problém nastane v případě, kdy se pokusíme pomocí PT-rovnováhy popsat stav čisté látky při saturačním tlaku odpovídajícím dané teplotě. V tomto případě se může objem jednotlivých fází měnit při zachování konstantního saturačního tlaku. Proto je v této práci použit termodynamický model, který jako podmínku

stability využívá minimalizaci Helmholtzovy energie. Tato formulace vychází z [2]. Helmholtzova energie je závislá na objemu, teplotě a chemickém složení systému a proto je tzv. VT-rovnováha (*Volume-Temperature flash*) schopna rozlišit jednotlivé stavy nejen směsi o více komponentách, ale také čisté látky při saturačním tlaku.

V [2] je navržen algoritmus pro řešení VT-rovnováhy, založený na metodě SSI (*Successive Substitute Iteration*). V této práci je popsán také algoritmus, který využívá Newtonovu-Raphsonovu metodu spolu s metodou přímek. Navržený výsledný algoritmus využívá obě zmíněné metody.

2 VT-rovnováha

Systém v rovnovážném stavu nabývá nejmenší možné Helmholtzovy energie A při dané teplotě, celkovém objemu a látkovém množství. Nechť je systém tvořen k komponentami s látkovým množstvím $n_1, \ldots, n_k$. Označme termodynamickou teplotu T a objem celého systému V. Helmholtzova energie potom může být vyjádřena jako funkce $A = A(V, T, n_1, \ldots, n_k)$.

V případě systému, ve kterém je p různých fází a k různých komponent, označíme $n_{j,i}$ látkové množství j-té fáze ($j = 1, \ldots, p$) v i-té komponentě ($i = 1, \ldots, k$). Chemický potenciál i-té komponenty v j-té fázi značíme $\mu_{j,i}$. Nyní lze Helmholtzovu energii vyjádřit ve tvaru

$$A = \sum_{j=1}^{p} \sum_{i=1}^{k} \mu_{j,i} n_{j,i} - \sum_{j=1}^{p} P_j V_j, \tag{1}$$

kde V_j a P_j označují objem a tlak j-té fáze. Celkový objem a látkové množství jednotlivých komponent musí být zachovány a proto musí platit

$$\textstyle\sum_{j=1}^{p} V_j = V, \qquad \sum_{j=1}^{p} n_{j,i} = n_i, \tag{2}$$

kde V je objem celého systému a n_i je látkové množství i-té komponenty. Podmínku rovnováhy lze tedy vyjádřit jako požadavek na minimalitu Helmholtzovy energie vzhledem k proměnným $n_{j,i}$ a V_j takovým, že platí podmínka (2).

Chceme-li minimalizovat Helmholtzovu energii jako funkci objemu, teploty a látkového množství jednotlivých komponent systému, musíme v těchto proměnných vyjádřit i tlak P a chemický potenciál μ.

Pro výpočet hodnoty tlaku je v této práci použita Pengova-Robinsonova stavová rovnice [3]

$$P(V, T, n_1, \ldots, n_k) = RT \frac{\sum_{i=1}^{k} n_i}{V - B} - \frac{A}{V^2 + 2VB - B^2}, \tag{3}$$

kde A a B jsou závislé na vlastnostech materiálu komponenty. Funkce A a B lze vyjádřit pomocí materiálových konstant $a_{i,j}$ a b_i a látkových množství n_i, kde $i, j = 1, \ldots, k$ jako

$$A = \textstyle\sum_{i=1}^{k} \sum_{j=1}^{k} a_{i,j} n_i n_j, \qquad B = \sum_{i=1}^{k} b_i n_i. \tag{4}$$

Takto lze tlak v j-té fázi vyjádřit ve tvaru $P_j = P(V_j, T, n_{j,1}, \ldots, n_{j,k})$, $j = 1, \ldots, p$.

Pro vyjádření chemického potenciálu v závislosti na požadovaných proměnných využíváme tzv. koeficienty objemových funkcí jednotlivých komponent. Koeficient objemové funkce i-té komponenty $\Phi_i = \Phi_i(V, T, n_1, \ldots, n_k)$ zavádíme vztahem

$$RT \ln \left(\frac{V_2}{V_1} \frac{\Phi_i(V_2, T, n_1, \ldots, n_k)}{\Phi_i(V_1, T, n_1, \ldots, n_k)} \right) = \mu_i(V_1, T, n_1, \ldots, n_k) - \mu_i(V_2, T, n_1, \ldots, n_k). \quad (5)$$

Koeficienty objemových funkcí mohou být vyjádřeny jako funkce objemu, teploty látkových množství následujícím vztahem (odvození viz [2])

$$\ln \Phi_i(V, T, n_1, \ldots, n_k) = \int_V^\infty \left[\frac{1}{\tilde{V}} - \frac{1}{RT} \frac{\partial P}{\partial n_i}(\tilde{V}, T, n_1, \ldots, n_k) \right] d\tilde{V}. \quad (6)$$

Nyní je Helmholtzova energie vyjádřena jako funkce objemu, tlaku a látkových množství, a proto je možné přistoupit k její minimalizaci.

3 Dvoufázové systémy

Vzhledem k tomu, že naší hlavní motivací je popis fázových přechodů při sekvestraci CO_2, zajímáme se především o dvoufázové systémy, konkrétně systémy s kapalnou a plynnou fází. V následujícím textu bude tedy formulována teorie pro dvoufázové systémy.

Při popisu stavu systému nás především zajímá, zda se v systému vůbec vyskytuje druhá fáze a v případě, že ano, chceme popsat podíl jednotlivých komponent v obou fázích. Pro tento účel zavádíme funkci ΔA, což je rozdíl Helmholtzovy energie systému, který obsahuje jen jedinou fázi, a Helmholtzovy energie systému rozděleného do dvou fází vydělený celkovým objemem systému.

$$\Delta \tilde{A} = \frac{1}{V} \left(A(V, T, n) - A(V', T, n') - A(V'', T, n'') \right), \quad (7)$$

Místo minimalizace A tedy nyní minimalizujeme funkci ΔA. V případě, že je minimální hodnota ΔA nezáporná, je v systému přítomná jediná fáze. V opačném případě je třeba nalézt také rozložení komponent v obou fázích.

Zaveďme nyní značení pro dvoufázové systémy. Jednu fázi popisují proměnné V', n', $n'_1, \ldots$, n'_k ($n' = \sum_{j=1}^k n'_j$) a druhou fázi proměnné V'', n'', $n''_1, \ldots$, n''_k ($n'' = \sum_{j=1}^k n''_j$). Zachování objemu a látkových množství jednotlivých komponent (2) lze nyní zapsat jako podmínku

$$V = V' + V'', \qquad n_i = n'_i + n''_i, \; i = 1 \ldots k. \quad (8)$$

Dále lze formulaci problému zjednodušit zavedením nových proměnných

- saturací obou fází $S' = \frac{V'}{V}$, $S'' = \frac{V''}{V}$
- koncentrací $c = \frac{n}{V}$, $c' = \frac{n'}{V'}$, $c'' = \frac{n''}{V''}$
- molárních zlomků $z_i = \frac{n_i}{n}$, $x'_i = \frac{n'_i}{n'}$, $x''_i = \frac{n''_i}{n''}$, $i = 1 \ldots k$.

Při použití těchto proměnných přechází minimalizovaný problém do tvaru

$$\Delta A = A(1, T, cz_j) - A(1, T, c'x'_j) - A(1, T, c''x''_j), \tag{9}$$

s podmínkou (8) vyjádřenou ve tvaru

$$\begin{aligned} 1 = &\quad S' + S'', \\ cz_i = &\quad S'c'x'_i + S''c''x''_i,\ i = 1 \dots k. \end{aligned} \tag{10}$$

Aby byl systém v rovnováze, je nutné splnění rovností

$$\begin{aligned} P(1, T, c'x'_j) &= P(1, T, c''x''_j), \\ \ln \varphi_i(1, T, c'x'_j) - \ln c'x'_i &= \ln \varphi_i(1, T, c''x''_j) - \ln c''x''_i,\ i = 1 \dots k. \end{aligned} \tag{11}$$

4 Numerické metody

V této kapitole představíme dvě numerické metody, které byly v rámci této práce použity k řešení VT-rovnováhy. Výhody a nevýhody obou metod budou diskutovány a na závěr navrhneme algoritmus pro řešení tohoto typu problému.

4.1 SSI (Successive Substitute Iteration)

Předpokládejme, že známe n-té (iterační) řešení molárních zlomků x'_i, x''_i, $i = 1 \dots k$. Označme $K_i = \frac{x''_i}{x'_i}$. Dosazením $x''_i = K_i x'_i$ do rovnice popisující zachování látkového množství i-té komponenty v (10) a zavedením $\alpha = \frac{c''S''}{c}$, dostaneme vztahy

$$x'_i = \tfrac{z_i}{1+(K_i-1)\alpha}, \qquad x''_i = \tfrac{K_i z_i}{1+(K_i-1)\alpha}. \tag{12}$$

Protože $\sum_{i=1}^k x'_i = 1 = \sum_{i=1}^k x''_i$, dostaneme dosazením (12) do $\sum_{i=1}^k x'_i - x''_i = 0$ tzv. Rachfordovu-Riceovu rovnici

$$\sum_{i=1}^{k} \frac{(K_i - 1)z_i}{1 + (K_i - 1)\alpha} = 0. \tag{13}$$

Vyřešením Rachfordovy-Riceovy rovnice pro K_i získané v i-té iteraci dostaneme α. Poté stačí vyřešit soustavu 4 rovnic pro c', c'', S', S''

$$\begin{aligned} c'S' = &\quad c(1-\alpha), \\ c''S'' = &\quad c\alpha, \\ S' + S'' = &\quad 1, \\ P(\tfrac{1}{c'}, T, x'_1, \dots, x'_k) = &\quad P(\tfrac{1}{c''}, T, x''_1, \dots, x''_k). \end{aligned} \tag{14}$$

Řešení této soustavy je ekvivalentní řešení jediné rovnice pro neznámou $S'' \in (0, 1)$

$$P(\frac{1 - S''}{c(1-\alpha)}, T, x'_1, \dots, x'_k) = P(\frac{S''}{c\alpha}, T, x''_1, \dots, x''_k). \tag{15}$$

Tlak vyjadřujeme pomocí Pengovy-Robinsonovy stavové rovnice, což je kubická rovnice. Rovnice (15) je proto polynomiální rovnicí pátého řádu, ve všech námi prozkoumaných případech však existoval pouze jeden kořen v intervalu $(0, 1)$ a proto vede tento přístup k jednoznačnému řešení problému.

4.2 Newtonova-Raphsonova metoda

Newtonova-Raphsonova metoda je jednou z nejznámějších metod pro řešení rovnic ve tvaru $F(X) = 0$. Chceme-li použít tuto metodu pro VT-rovnováhu, jedná se o řešení vektorové rovnice, kde

$$X = \begin{pmatrix} c'x'_1 \\ \vdots \\ c'x'_k \\ c''x''_1 \\ \vdots \\ c''x''_k \\ S' \\ S'' \end{pmatrix} \qquad F(X) = \begin{pmatrix} S'c'x'_1 + S''c''x''_1 - cz_1 \\ \vdots \\ S'c'x'_k + S''c''x''_k - cz_k \\ P(1,T,c'x'_j) - P(1,T,c''x''_j) \\ \ln \frac{c'x'_1\varphi_1(1,T,c''x''_j)}{c''x''_1\varphi_1(1,T,c'x'_j)} \\ \vdots \\ \ln \frac{c'x'_k\varphi_k(1,T,c''x''_j)}{c''x''_k\varphi_k(1,T,c'x'_j)} \\ S' + S'' - 1 \end{pmatrix} \tag{16}$$

Newtonova-Raphsonova metoda je iterační metoda, která na základě znalosti n-tého řešení hledá $(n+1)$-ní řešení pomocí rovnosti

$$X^n - X^{n+1} = \left(\Delta F(X^n)\right)^{-1} F(X^n). \tag{17}$$

ΔF označuje gradient funkce F, což je matice se strukturou

$$\Delta F \sim \left(\begin{array}{ccccc|ccccc|cc} \blacksquare & & & & & \blacksquare & & & & & & \\ & \blacksquare & & & & & \blacksquare & & & & & \\ & & \ddots & & & & & \ddots & & & & \\ & & & \blacksquare & & & & & \blacksquare & & & \\ & & & & \blacksquare & & & & & \blacksquare & & \\ \hline \blacksquare & \blacksquare & \dots & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \dots & \blacksquare & \blacksquare & & \\ \blacksquare & \blacksquare & \dots & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \dots & \blacksquare & \blacksquare & & \\ \blacksquare & \blacksquare & & \blacksquare & \blacksquare & \blacksquare & \blacksquare & & \blacksquare & \blacksquare & & \\ \vdots & & \ddots & & \vdots & \vdots & & \ddots & & \vdots & & \\ \blacksquare & \blacksquare & & \blacksquare & \blacksquare & \blacksquare & \blacksquare & & \blacksquare & \blacksquare & & \\ \blacksquare & \blacksquare & \dots & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \dots & \blacksquare & \blacksquare & & \\ \hline & & & & & & & & & & \blacksquare & \blacksquare \end{array}\right) \tag{18}$$

4.3 Porovnání obou metod, navrhovaný algoritmus

Metoda SSI dosahuje nejvýše prvního řádu konvergence, je to velmi stabilní metoda. Newtonova-Raphsonova metoda dosahuje až druhého řádu konvergence, je však velmi citlivá na volbu počátečního odhadu řešení. Proto pro řešení VT-rovnováhy používáme v rámci této práce algoritmus, který nejdříve provede několik iterací metodou SSI, čímž se přiblíží přesnému řešení dostatečně na to, aby mohla být použita Newtonova-Raphsonova metoda, která vede rychleji k numerickému řešení s požadovanou přesností.

5 Numerické experimenty

V následujícím textu uvedeme dva experimenty, které simulují chování postupně stlačovaných směsí. Zmenšování celkového objemu systému vyjadřuje zvyšování celkové koncentrace látky. Z grafů je patrný přechod od dvoufázové směsi k jednofázové směsi v závislosti na celkové koncentraci. K výpočtům je použit algoritmus navržený v kapitole 4.3.

5.1 Směs methanu C_1 a n-penthanu nC_5

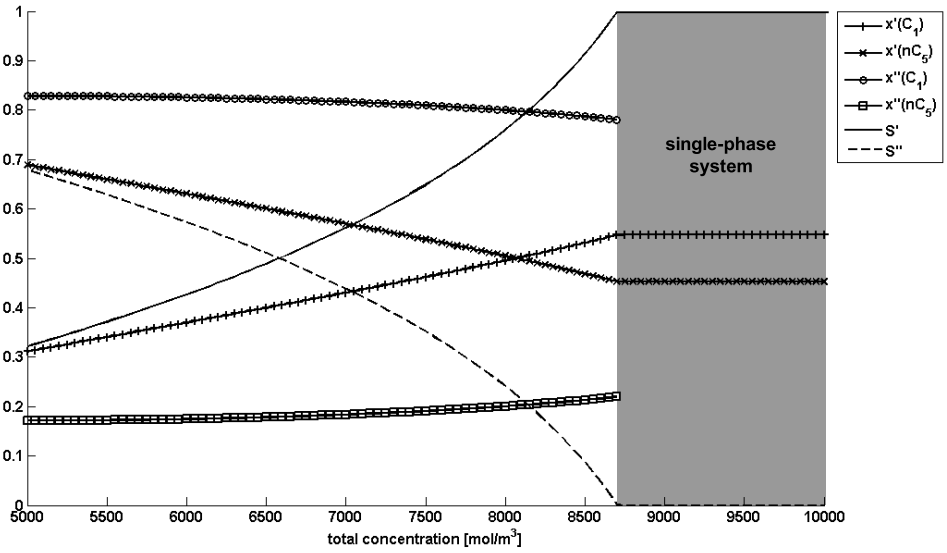
Nejdříve simulujeme kompresi binární směsi methanu C_1 a n-penthanu nC_5. Při značení z kapitoly 3 je komprese vyjádřena postupným nárůstem celkové koncentrace. Přesný popis parametrů výpočtu je následující:

- celková koncentrace roste z 5000 mol·m^3 na 10000 mol·m^3,
- termodynamická teplota je $T = 371$ K,
- složení směsi popisují molární zlomky $z_{C_1} = 0,547413$, $z_{nC_5} = 0,452587$.

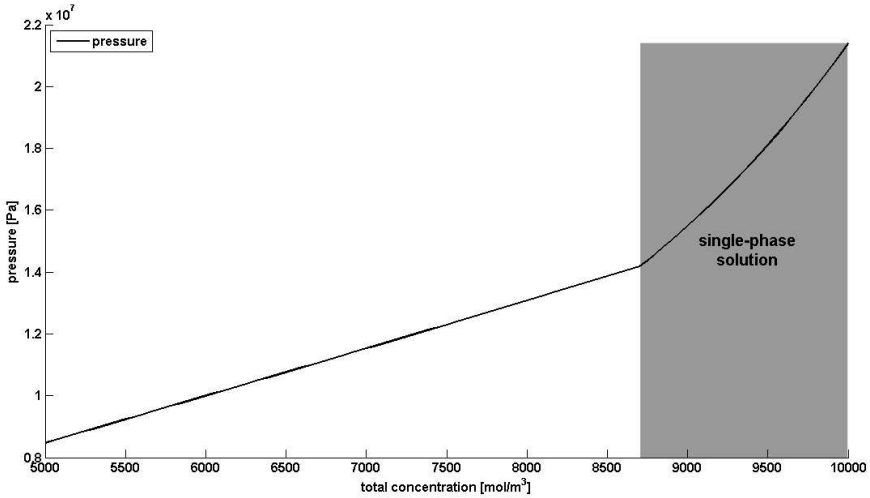
Na obr. 1 je vyobrazen průběh molárních zlomků metanu a n-penthanu v obou fázích. Z definice molárních fázových zlomků je nutné, aby se fázové molární zlomky obou komponent sčítaly na jedničku (tj. $x'_{C_1} + x'_{nC_5} = 1 = x''_{C_1} + x''_{nC_5}$). Z obrázku je vidět, že tato vlastnost je numerickým modelem zachována. V obr. 1 je vyobrazen také průběh saturací obou fází. Opět je z definice saturací nutné, aby $S' + S'' = 1$, což je opět velmi dobře vystiženo v grafu. Z průběhu saturací je zřejmé, že se zvyšováním celkové koncentrace roste saturace první fáze, až při koncentraci zhruba 8450 mol·m^{-3} dojde k přechodu původně dvoufázové směsi na směs jednofázovou.

Obr. 2 zachycuje průběh tlaku směsi při kompresi. Opět je možné vypozorvat místo přechodu z dvoufázové směsi na směs jednofázovou podle změny průběhu tlaku při zhruba 8450 mol·m^{-3}.

Obrázek 1: Molární zlomky C_1 a nC_5 a saturace obou fází v závislosti na celkové koncentraci pro výpočet popsaný v 5.1.



Obrázek 2: Vývoj tlaku uvnitř směsi C_1-nC_5 v závislosti na celkové koncentraci pro problém 5.1.



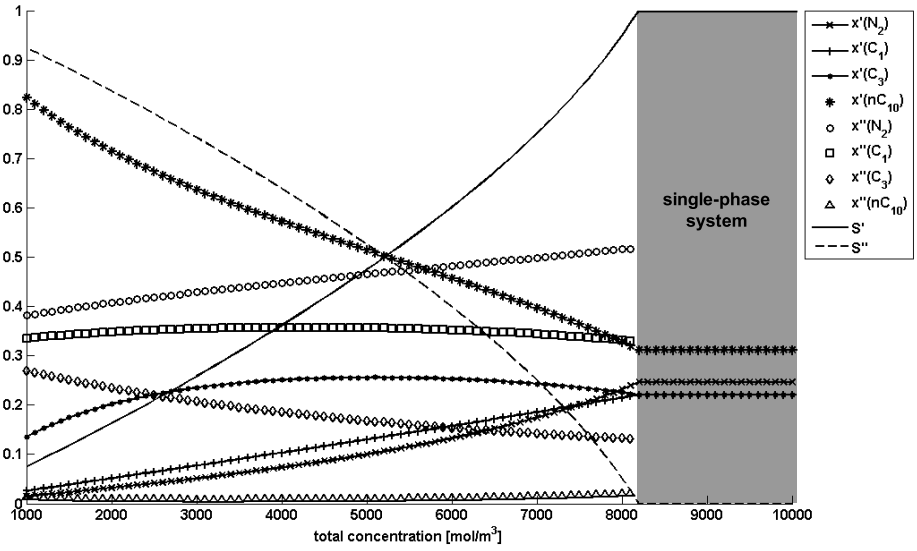
5.2 Směs dusíku N_2, methanu C_1, propanu C_3 a n-dekanu nC_{10}

Stejný algoritmus jako byl v kapitole 5.1 použit k řešení komprese binární směsi lze použít také k řešení VT-rovnováhy vícesložkových směsí. Jako příklad uvádíme výpočet VT-rovnováhy stlačované 4-složkové směsi tvořené dusíkem, methanem, propanem a n-dekanem. Parametry výpočtu jsou následující:

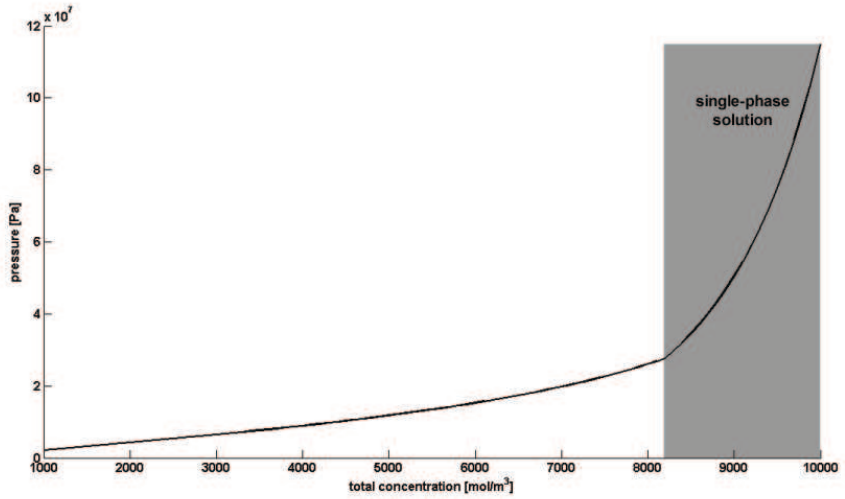
- celková koncentrace roste z 1000 mol·m^3 na 10000 mol·m^3,
- termodynamická teplota je $T = 393,15$ K,
- složení směsi je dáno molární zlomky $z_{N_2} = 0,2463$, $z_{C_1} = 0,2208$, $z_{C_3} = 0,2208$, $z_{nC_{10}} = 0,3121$.

Obr. 3 znázorňuje průběh fázových molárních zlomků všech 4 komponent. V tomto případě již není tak zřejmé zachování jednotkového součtu fázových zlomků komponent, nicméně při důkladném prozkoumání průběhu molárních zlomků je vidět, že opět platí $x'_{N_2} + x'_{C_1} + x'_{C_3} + x'_{nC_{10}} = 1 = x''_{N_2} + x''_{C_1} + x''_{C_3} + x''_{nC_{10}}$. Stejně tak je zachován jednotkový součet obou saturací S' a S''. Stejně jako v případě 5.1 převažuje s rostoucí celkovou koncentrací objem první fáze, jejíž saturace je označena S'. Při celkové koncentraci zhruba 8100 mol·m^{-3} dojde k přechodu dvoufázového systému na systém jednofázový. Tento přechod je patrný jednak z průběhu saturací a fázových molárních zlomků v grafu na obr. 3, při zmíněné koncentraci je však zřetelná i změna v průběhu celkového tlaku systému.

Obrázek 3: Molární zlomky komponent a saturace obou fází v systému N_2-C_1-C_3-nC_{10} popsaného v 5.2 v závislosti na celkové koncentraci.



Obrázek 4: Vývoj tlaku uvnitř systému N_2-C_1-C_3-nC_{10} popsaného v 5.2 v závislosti na celkové koncentraci.



6 Závěr

V této práci byl použit nový přístup k popisu termodynamiky rovnovážných fázových procesů, který je formulován při konstantním objemu namísto klasického pojetí, při kterém je udržován konstantní tlak směsi. Byly stručně diskutovány vlastnosti a důvody zavedení nového přístupu a navrženy dvě numerické metody pro řešení fázových přechodů v souladu s novou formulací.

V rámci této práce byl navržen algoritmus k výpočtu VT-rovnováhy. Tento algoritmus byl použit k výpočtu průběhu složení dvou vícesložkových systémů při jejich kompresi. Z výsledků je patrný postupný přechod od dvoufázové směsi k směsi jednofázové při narůstající celkové koncentraci. V obou experimentech systém přešel do fáze popsané proměnnými S', x'_i, c'. Z výsledků však není možné rozlišit o jakou fázi se jedná (zda bude při vysoké koncentraci směs v kapalném nebo plynném stavu).

V současnosti jsou kódy implementované v programu Matlab, do budoucna plánujeme jejich převedení do programovacího jazyka C++, aby bylo možné tento model fázových přechodů propojit např. s modelem vícefázového proudění.

Literatura

[1] Abbas Firoozabadi. *Thermodynamics of Hydrodynamic Reservoirs*. McGraw-Hill, 1999.

[2] Jiří Mikyška and Abbas Firoozabadi. A new thermodynamic function for phase-splitting at constant temperature, moles, and volume. *to appear in AIChE Journal*, 2011. Available on-line, DOI: 10.1002/aic.12387.

[3] Ding-Yu Peng and Donald B. Robinson. A new two-constant equation of state. *Industrial Engineering Chemistry: Fundamentals*, 15:59–64, 1976.

Heuristic Effectiveness Analysis on Knapsack Problem*

Matej Mojzeš

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

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

Abstract. Performance of a heuristic depends on the principle on which the algorithm is based on, on the setting of its parameters and also on the type of problem being solved. In the paper we are proposing a deterministic methodology to address the problem of the effectiveness evaluation of these algorithms. Furthermore, a demonstration of this methodology is presented. Its objective is to show how the proposed methods can be helpful for further development or application of heuristics.

Keywords: optimization, heuristics, performance

Abstrakt. Výkonnosť heuristiky záleží na princípe na ktorom je založená, na nastavení jej parametrov a taktiež aj na type problému, ktorý rieši. V našej práci navrhujeme deterministickú metodiku na riešenie problému vyhodnocovania efektívnosti týchto algoritmov. Ďalej predstavujeme demonštráciu, ktorá ukazuje ako môžu byť navrhnuté metódy užitočné pri vývoji a aplikácii heuristík.

Kľúčové slová: optimalizácia, heuristiky, výkonnosť

1 Introduction

As it follows from the "No Free Lunch" theorem [1], every heuristic algorithm, or even every distinct parameters setting of the same algorithm, performs differently on different problems. Moreover, there is not just one criterion for the evaluation of heuristic performance itself.

Thus, when when it comes to the development of a new algorithm or utilisation of an existing one to solve an *optimization problem* (OP), it is relevant to question *how to compare different heuristics and/or their parameter settings* when searching for the solution a specific problem.

A few attempts have been made [2] [3] to point out the difficulties one will deal with when analyzing and comparing heuristic algorithms performance, or to describe some of the "best practices" in this field, but a formalized framework is missing and we do believe that a contribution to this field could be made.

In our previous work [7] we have introduced a set of techniques we find useful for the analysis of the heuristic algorithms performance. These techniques can be divided into

*This paper has been supported by the grant OHK4-165/11 CTU in Prague

two main categories - *deterministic* and *stochastic*. Apart from the principles on which these methods are based, they are both delivering slightly different results. Using the deterministic methods, we arrive at a precise ranking of the heuristics performance, i.e. we know which has the best performance, which is the second one, etc. On the other hand, stochastic methods can generate a "cluster" of the best performing heuristics. Every heuristic in this cluster is better than the rest and heuristics in the cluster are all equally good by the means of statistical analysis.

However, in this paper we are focusing only on the deterministic methods and we are proposing an approach that enables also the deterministic methods to produce results similar to the latter category.

2 Methodology of comparison

Before detailed description of the proposed techniques, we begin with a more exact problem definition and the definition of heuristics performance measures.

An OP can be defined as minimization of the objective function f: $\mathbf{D} \rightarrow \mathbb{R}$ where

$$\mathbf{D} = \{\mathbf{x} \in \mathbf{X}^n \mid \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$$

is an appropriate domain. For purposes of this paper we are using the binary domain, $\mathbf{X}^n = \{0,1\}^n$, but it can be an integer or a real one as well. Let's suppose that we have an acceptable value of the objective function f^*. Then we can define a set of solutions, the goal set, as

$$\mathbf{G} = \{\mathbf{x} \in \mathbf{D} \mid \mathrm{f}(\mathbf{x}) \leq f^*\}$$

where

$$f^* \geq \min\{\, \mathrm{f}(\mathbf{x}) \mid \mathbf{x} \in \mathbf{D} \,\}.$$

We will suppose that we have a set of $H \in \mathbb{N}$ heuristics, each having $P_i \in \mathbb{N}$ parameters for $i = 1, 2, \ldots, H$. Then $p_{i,j} \in \mathbf{P}_{i,j}$ means that the setting of j-th parameter of i-th heuristic $p_{i,j}$ belongs to domain $\mathbf{P}_{i,j}$ which is a set of any distinct values specific for the given heuristic algorithm implementation, e.g. real numbers, logical values, or text strings. Domain $\mathbf{P}_{i,j}$ has the cardinality (number of elements) $C_{i,j} = \mathrm{card}(\mathbf{P}_{i,j})$. Actual parameter settings and their combinations may be based on recommendations (e.g. [6]), own experience, or by "random shooting" in the worst case.

2.1 Performance measures

Following the principles of Monte Carlo simulation we will run every *instance* of heuristic algorithm (meaning one specific heuristic approach with one specific setting of its parameters) for a specified, sufficiently large, number of times $q \in \mathbb{N}$ and then we will define the following estimates:

- *Reliability*, $REL = m/q$ where $m \in \mathbb{N}$ is the number of successful runs i.e. runs during which the algorithm found a solution from the goal set before exceeding the maximum allowed number of objective function evaluations (clearly $m \leq q$ and thus $REL \in [0,1]$);

- *Mean Number of Evaluations*, $MNE = \frac{1}{m}\sum_{i=1}^{m} NE_i$ where $NE_i \in \mathbb{N}$ is the number of evaluations of the objective function until the algorithm found a solution from the goal set;
- *Standard deviation of the Number of Evaluations*, $SNE = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(NE_i - MNE)^2}$.

It may happen, that reliability will be so low that it is impossible to evaluate SNE and for that reason it is necessary to discard such instances from further analysis. Also, since reliability is often a very sensitive attribute, we recommend further preliminary elimination of very unreliable instances, no matter how effective they are.

The number of evaluations is a positive integer stochastic variable which significantly varies in order. It is hardly possible to suppose that this variable has a distribution which is close to the Gaussian normal one. It is a good habit to analyze its natural or decadic logarithm instead of the original value to eliminate positive skewness [2]. So we introduce the logarithmic measures as follows:

- *Logarithm of Number of Evaluations*, $LNE_i = \ln NE_i$ for $i \in 1, 2, \ldots, m$;
- *Mean Logarithm of the Number of evaluations*, $MLN = \frac{1}{m}\sum_{i=1}^{m} LNE_i$;
- *Standard deviation of the Logarithm of Number of evaluations*, $SLN = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(LNE_i - MLN)^2}$.

Our task is to maximize REL and minimize MLN and SLN together. However, at this point it is obvious that REL, MLN and SLN are independent variables, and, typically, they are in contrast with each other (e.g. a very reliable heuristic has higher MLN than another, less reliable one). Therefore this is a typical example of a *multi-criteria decision analysis*.

We can start analysing weakly performing instances using the condition of *Pareto optimality* [4] – this way we can exclude heuristics which are worse than others in every measured characteristic. Nevertheless, in most cases this will not be of great impact and there will still be plenty of instances to choose from.

2.2 Proposed methodology

The *weighted approach*, which is discussed in this paper, is based on minimizing the general formula with positive weights $w_i \in \mathbb{R}$ for $i \in 1, 2, 3$ of included performance measures:

$$\mathrm{F} = w_1 \cdot MLN + w_2 \cdot SLN + w_3 \cdot LNR$$

where $LNR = -\ln REL$.

Weights can be determined using various techniques. Some of them are motivated by the traditional Feoktistov's criterion: $FEO = MNE/REL$ [5].

However, we are suggesting a criterion that works on a presumption that NE can be approximated by the *exponential distribution* with time constant T [7] and it is in the form of

$$F = MLN + \frac{C \cdot \sqrt{6}}{\pi} \cdot SLN + LNR$$

where C is the Euler's constant and just for illustration $\frac{C \cdot \sqrt{6}}{\pi} \cong 0.4501$.

Moreover, to be able to decide which instance of heuristic algorithm is efficient and which one is not, we are suggesting certain bounds of acceptance. We say that the instance k is efficient when the value of its criterion F_k satisfies the following inequality:

$$F_k \leq F_{\min} + \frac{\pi}{\sqrt{3q}} \cdot \Phi^{-1}(1-\alpha)$$

Here, $F_{\min}$ is the score of the best instance and Φ^{-1} is inverse of cumulative distribution function (cdf) of the normal distribution $\mathrm{N}(0,1)$.

Previous formula is based on pessimistic assumption that random variable F has a variance $\sigma_F^2 = \frac{\pi^2}{6q}$ which is q times smaller than variance in the case of random shooting. Now we can test the hypothesis $\mathrm{H}_0 : F_k = F_{\min}$ against $\mathrm{H}_1 : F_k \neq F_{\min}$ at the level of significance α.

Under this assumptions the variable $F_k - F_{\min}$ follows normal distribution $\mathrm{N}(0, 2\sigma_F^2)$ and after substitution:

$$F_k - F_{\min} \sim \mathrm{N}(0, \frac{\pi^2}{3q})$$

In the case of symmetric testing, the hypothesis is rejected when

$$|F_k - F_{\min}| > \frac{\pi}{\sqrt{3q}} \cdot \Phi^{-1}(1 - \frac{\alpha}{2})$$

However, we need to test whether $F_k < F_{\min}$ and thus reject the hypothesis if

$$F_k > F_{\min} + \frac{\pi}{\sqrt{3q}} \cdot \Phi^{-1}(1-\alpha)$$

and, for $\alpha = 0.05$, it is approximately

$$F_k > F_{\min} + \frac{2.9834}{\sqrt{q}}$$

3 Experimental results

3.1 Testing environment

As far as the testing problem is concerned we have chosen the 0-1 *Knapsack Problem* (KP) since it is a very well known one and for a reasonably small problem dimensions, i.e. the number of items, each having its own weight and value to put into the knapsack with limited maximum load, we can easily find an analytical solution using methods of binary programming and set the f^* value accordingly. Objective function is the total value of items in the knapsack, of course with a negative sign, incremented by a penalty for exceeding the maximal load of the knapsack, if appropriate. Both weights and values of items come from geometric sequences.

While on the other hand we have the heuristic algorithms to deal with the testing problem. The first algorithm we have used is the *Genetic Optimization* (GO) [6], and since its detailed description is out of scope of this paper, we will mention only the parameters and respective settings which are to be analysed:

- $N \in \{5, 50, 100\}$ – the size of population
- $T_{\text{sel}} \in \{0.1, 10, 100\}$ – temperature controlling the probability of selection
- $R \in \{0.001, 0.1\}$ – radius of mutation
- $rep \in \{1, 3, 5\}$ – number of generation repetitions
- $gdc \in \{0, 1\}$ – indication as to whether gradually decrease T_{sel} and R over time

Finally, our implementation of the *Fast Simulated Annealing* (FSA) [6], the second heuristic, has a following set of parameters and settings:

- $T_0 \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10\}$ – initial temperature
- $n_0 \in \{1, 10, 500, 1000\}$ – cooling delay
- $\alpha \in \{0.5, 1, 1.5, 2\}$ – cooling exponent

To summarize the number of instances, we get $NI = 3 \cdot 3 \cdot 2 \cdot 3 \cdot 2 + 7 \cdot 4 \cdot 4 = 108 + 112 = 220$. The first part of the sum is the product of individual distinct parameter values counts of GO and the second one of FSA, which performance we want to compare.

3.2 Results

To get a satisfactorily precise result, we set $q = 100$, and thus we analyzed $NI \cdot q = 22\,000$ of single runs of heuristic algorithms per problem. Table 1 gives us a rough idea of what the outcome of our experiment was.

Table 1: Overall performance results

	GO	**FSA**	**Total**
All	108	112	220
Reliable	81	112	193
Effective	5	0	5
The best one	1	0	1

Note: A *reliable* instance was the one having $REL \geq 0.2$.

Even from this, very high-level, statistic we can draw interesting conclusions about the two different heuristics: GO, despite being a bit unreliable, has been able to produce all the most effective instances. FSA on the other hand, seems to be at least a significantly more reliable heuristic.

Table 2: Basic performance characteristics

Characteristic	**GO**	**FSA**
Average reliability	0.62	0.93
Average MNE	519.9	2193.0
Average SNE	609.9	2803.1
$F_{\min}$	5.02	6.60
F_{avg}	6.17	7.62
$F_{\max}$	10.93	7.93

Table 2 supports the above mentioned conclusions and adds also another relevant observations, which are rather self-explanatory.

It is also worth mentioning that three instances of GO had $REL \leq 0.01$. This prevented us from calculating the F criterion value altogether.

Last, but not least, we are presenting a detailed overview of the effective instances parameter settings in table 3.

Table 3: Effective instances parameter settings

Heuristic	**Parameter**	**Value**	**Times present**	**Impact ratio**
GO	N	50	3	0.6
GO	N	5	2	0.4
GO	T_{sel}	10	3	0.6
GO	T_{sel}	0.1	2	0.4
GO	R	0.1	5	1.0
GO	rep	1	2	0.4
GO	rep	3	2	0.4
GO	rep	5	1	0.2
GO	gdc	1	5	1.0

A table similar to table 3 may be of significant use, since it does reveal the parameters and their settings which have the biggest impact on the performance of the heuristic. E.g. from the values in this specific one we can conclude that it may be a good idea to set $gdc = 1$ and $R = 1$, further search for the optimal setting of N and T_{sel} and stop tuning the rep as it will most probably not bring the desired effect.

For the sake of illustration, we are also including an analogous table, table 4, for the FSA heuristic, but of course, under the assumption that it would be the only available heuristic to deal with the given problem.

Table 4: Effective instances parameter settings - FSA only

Heuristic	**Parameter**	**Value**	**Times present**	**Impact ratio**
FSA	T_0	10	3	1.0
FSA	n_0	1000	2	0.7
FSA	n_0	50	1	0.3
FSA	α	1.5	2	0.7
FSA	α	2	1	0.3

4 Conclusions

Having the outcome of the previous section in mind, we can conclude that it is possible to analyse the performance of heuristic algorithms using the proposed deterministic approach.

In addition to that, we do believe that it is also relevant, and feasible as well, to introduce a deterministic methodology to address the problem of determining which instances of heuristic algorithms are performing equally well and to extract appropriate knowledge from such data.

As we have shown in the fore-mentioned case study, this approach can be used to better understand and interpret the behaviour of the heuristic and reveal the potential effect of settings tuning.

However, in the last part of the experimental section, we have analysed the effects of possible tuning of individual settings. Of course, the settings, and most importantly the effects they may have on the performance, are not completely independent – so it is here, where we see the most straight-forward potential for future work, to analyse effects of settings combinations.

References

[1] Wolpert D. H., Macready W. G., No Free Lunch Theorems for Optimization, *IEEE Transactions on Evolutionary Computation*, Vol. 1, No. 1, 1997

[2] McGeoch C. C., Experimental Analysis of Algorithms, *Handbook of Global Optimization Volume 2*, Kluwer Academic Publishers, 2002, pp. 489–513

[3] Battiti R., Machine Learning Methods for Parameter Tuning in Heuristics, *5th DIMACS Challenge Workshop: Experimental Methodology Day*, Rutgers University, 1996

[4] Van Veldhuizen D. A., Lamont G. B., Evolutionary Computation and Convergence to a Pareto Front, *Proceedings of the 3rd Annual Conference on Genetic Programming*, Stanford University, San Francisco CA, 1998, pp. 221–228

[5] Feoktistov V., *Differential Evolution: In Search of Solutions*, Springer, 2006

[6] Kvasnička V., Pospíchal J., Tiňo P., *Evolutionary Algorithms* (in Slovak), STU Bratislava, 2000

[7] Mojzeš M., Kukal J., Tran V. Q., Jablonský J., Performance Comparison of Heuristic Algorithms via Multi-Criteria Decision Analysis *Proceedings of Mendel 2011 Conference*, Brno University of Technology, Brno, 2011, pp. 244–251

Requirements Engineering*

Josef Myslín

2. ročník PGS, email: myslin@volny.cz
Katedra softwarového inženýrství v ekonomii
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Vojtěch Merunka, Katedra softwarového inženýrství v ekonomii,
Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. Information systems becomes more difficult and communication during their construction is still more difficult. But good communication and good requirements analysis is the condition for making good and usable information system. This paper tries to offer view on basic methods of business process modeling. This methods have one goal - make the communication between developers and their customers simplier. I also try to evaluace methods and outline their lability for different situation.

Keywords: Requeirement analysis, business modeling, IDEF, UML, BPMN, EPC

Abstrakt. Informační systémy se stávají stále složitějšími a komunikace při jejich vytváření je stále složitější. Přitom právě správná komunikace a správné nastavení požadavků umožňuje vytvořit dobrý a použitelný informační systém. Tento článek se snaží nabídnout pohled na základní metody modelování byznys procesů, přičemž tyto metody mají za cíl usnadnit komunikaci mezi vývojáři a zákazníky. Zároveň se pokouším o zhodnocení jednotlivých metod a nastínění jejich použitelnosti pro různé situace.

Klíčová slova: Správa požadavků, byznys modelování, IDEF, UML, BPMN, EPC

1 Úvod

Tvorba informačního systému je ve své nejvnitřnější podstatě světem dvou zcela odlišných světů. Prvním z nich je svět výpočetní techniky, tedy svět založený na hlubokém pochopení přírodních věd, na formální matematizaci problémů, na rigorózních důkazech. Informatici využívají modely a metody značně se lišící od běžného vnímání reality. Oproti tomu svět obchodu, svět byznysu, je zcela jiný. Je založen na tvrdém vnímání reality, často více na intuici než na exaktních metodách. Pokud mají lidé z tohoto světa vnímat modely, pak musí pochopit souvislost mezi nimi a tím, co jsou schopni vnímat v realitě. Jejich schopnost abstrakce je často nižší, navíc abstrakci často vnímají jen jako překážku reálné práce, přičemž právě reálná práce a její výsledky jsou generátory zisku, který je hlavním cílem podnikánijako takového.

Je pak logické, že v případě, že mají tyto dva zcela odlišné světy komunikovat, dochází ke značným rozporům, nedorozuměním a zmatkům. Důsledkem je pak fakt, že velká část (často až nadpoloviční většina) všech softwarových projektů končí naprostým fiaskem – tedy buď přímým zastavením projektu nebo, což je ještě horší, nasazením do provozu díla, které v žádném případě nesplňuje původní požadavky a které je často nikoliv pomocí,

*Tato práce byla podpořena grantem SGS2011

ale spíše příteží pro ty, kdož jsou nuceni takovéto systémy využívat [7]. Tato obrovská neefektivita způsobuje velké ztráty jak samotnému byznysu, tak také IT společnostem a pracovníkům v nich, neboť snižuje jejich důvěryhodnost a jejich možné další uplatnění a také vrhá negativní stín na moderní technologie. Ty však samy o sobě viníkem problémů nejsou a být nemohou. Na vině je špatná forma komunikace. Cílem mého odborného a vědeckého zájmu je hledat způsoby, jakými by bylo možné tyto bariéry odstranit nebo alespoň zmírnit, například formou vhodných transformací modelů [6), případně využitím analogií [4]. Důležitým aspektem pak je a zůstává modelování. A právě možnými přístupy modelování se zabývá tento článek.

2 Význam modelování byznys procesů v managementu

V předchozím odstavci jsme se zabývali modelováním byznys procesů pro potřeby SW inženýrství, nicméně tato oblast je jen jednou z těch, kde se modelování procesů využívá. Je nutno si uvědomit základní myšlenku:

Software je pouze nástroj, který má sloužit k efektivnímu řízení a zpracování informací. Podstatou je však proces.

Jinými slovy lze říci, že základní věcí, kterou manažer řeší, je samotný proces a jeho provedení, případně problémy, které se vyskytnou. To, zda použije či nepoužije informační systém, je pouze o volbě konkrétního nástroje. Procesní a projektové řízení patří mezi významné pilíře dnešních teorií managementu. Jsou to účinné nstroje uskutečňování podnikatelských strategií a dosahování cílů – oceňované pro systémový přístup, který umožňuje využít matematické a statistické metody k optimalizaci procesů i projektů, a informační systémy k jejich řízení. Ačkoliv se použití procesního a projektového managementu věnuje řada publikací, manažerských stylů, metodik či technik, bývá jejich přijetí v podmínkách České republiky poněkud rozpačité. Procesní přístup je u nás často zužován na certifikaci systémů managementu kvality (dle ISO 9001) [10]. Problémem zde bývá fakt, že se jedná o formální certifikaci, která slouží k přístupu do různých výběrových řízení, ale samotné procesní řízení je opomíjeno a přináší jen zvýšenou byrokracii. Přitom pro řízení je velmi důležité znát a být schopen přesně popsat jednotlivé procesy ve firmě, neboť jen tak pracujeme se skutečnými daty. Pro zlepšení fungování firmy v tržním hospodářství je nutné neustále inovovat, v případě procesů tedy provádět reengineering.

Reengineering znamená zásadní přehodnocení a radikální rekonstrukci (redesign) podnikových procesů tak, aby mohlo být dosaženo dramatických zdokonalení z hlediska kritických měřítek výkonnosti, jako jsou náklady, kvalita, služby a rychlost [13].

Praktické provedení reengineeringu je specifickým projektem řízení změny v podniku. Projekt je zpravidla iniciovn vrcholovým vedením, jehož členové dále i přijímají role (přinejmenším roli sponzora) v realizačním týmu. Metodiky reengineeringových projektů jsou charakterově blízké projektům implementace informačních systémů. Klasické metodiky člení projektu do tří základních etap:

1. Přípravy projektu reengineeringu.
2. Rekonstrukce procesu.
3. Implementace změny.

Pakliže chceme provádět reengineering procesu, musíme dokonale pochopit to, jak funguje daný proces v současnosti a jaké jsou jeho slabiny. První nám zajistí kvalitní procesní modely, druhé pak simulace (automatická či manuální) daných procesů.

3 Možné přístupy

Máme-li se dále zabývat modelováním procesů, musíme se seznámit s alespoň některými zajímavými přístupy, které můžeme pro tento účel využít. Nástroje můžeme obecně rozlišit na [10].

1. Grafické jazyky a nástroje pro modelování.
2. Formální jazyky pro automatizované provádění procesů a workflow.
3. Formální jazyky pro B2B výměnu dat.

Vzhledem k tomu, že cílem není naprosto formální a matematické popsání procesů, ale taková forma modelování, která umožní lepší komunikaci mezi zadavatelem a vývojářem, případně nám jde o modelování procesu pro manažerské řízení, budeme volit nástroje z první skupiny – tedy nástroje grafické, umožňující vizuální zobtazení jednotlivých procesů. Zde máme k dispozici například následující široce využívané nástroje a metodiky [10].

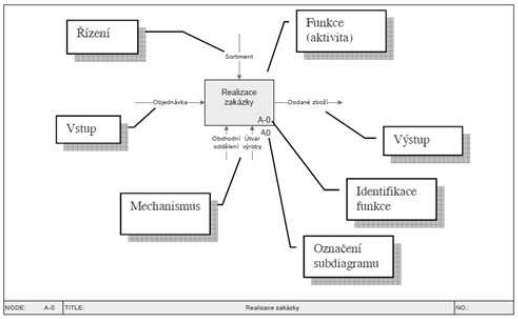
Metodika/nástroj	**Popis**
UML (Unified Modeling Language	Standard UML je rozsáhlou sadou nástrojů pro modelování informačních systémů a jejich chování. Je udržován konsorciem OMG. K modelování procesů jsou využitelné diagramy stavů a aktivit ze standardních profilů nebo rozšiřující profily jako např. (Eriksson, a další, 2000).
BPMN (Business Process Modeling Notation).	Dnes patrně nejrozšířenější notaci pro modelování procesů lze charakterizovat jako rozšíření diagramu aktivit UML. Standard spravuje konsorcium OMG.
EPC (Event-driven Process Chain).	EPC je notací používanou v metodice a nástrojích ARIS (Software AG), které patří k nejrozšířenějším a nejpropracovanějším nástrojům na trhu s BPM.
IDEF3 (The Integrated DEFinition).	Jde o jeden ze standardů vyvinutých v rámci US AirForce pro komplexní podporu modelování podnikové architektury.

4 IDEF

Metoda IDEF (Integration DEFinition), konkrétně IDEF0, poskytuje modelovací jazyk s danou syntaxí a sémantikou, umožňující vytvořit strukturovanou grafickou reprezentaci systému nebo organizace. Jejím použitím je možné sestavit konsistentní model tvořený popisem funkcí systému, jejich vzájemných vztahů a dat umožňujících tyto funkce integrovat. Metoda IDEF byla odvozena z graficky orientovaného jazyka SADT (Structured Analysis and Design Technique) na základě požadavků U.S. Air Force. Účelem bylo nalézt prostředek pro analýzu a komunikaci mezi lidmi zaměřenými na zvyšování produktivity výroby. Výsledkem bylo vytvoření celé řady technik, které byly v budoucnu ještě dále rozšířeny. Účelem použití IDEF0 je vytvoření modelu, který se sestává z hierarchicky uspořádané sady diagram a textů s přesně vytvořeným systémem vzájemných odkazů popisujícími funkce organizace či podniku. Primárními modelovacími komponentami jsou funkce a data/objekty, které vzájemně tyto funkce propojují. Konkrétně se jedná o následující syntaktické prvky (viz další obrázky:

1. Funkce (Function) popisující činnost transformující vstup na požadovaný výstup.
2. Vstupem (Input) jsou data nebo objekty, které budou funkcí transformovány na výstup.
3. Výstupem (Output) rozumíme data nebo objekty produkované funkcí.
4. Řízení (Control) je dáno pravidly potřebnými k vytvoření požadovaného výstupu.
5. Mechanismus (Mechanism) definuje prostředky nutné k realizaci funkce.

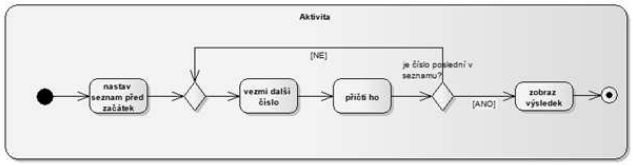
Z anglických názvů jednotlivých toků se pak hovoří o tzv. ICOM (Inputs, Controls, Outputs, Mechanisms), které je vyžadováno každou funkcí. Kromě toho, každá funkce nese své číselné označení (ID) a případně také označení diagramu, ve kterém je funkce pak dále rozpracována do svých dalších podfunkcí (obr. 2.4). Díky tomu je možné vytvářet hierarchii diagramů odpovídající dekompozici funkcí na své podfunkce (strukturovaný přístup). Vrchol této hierarchie je definován tzv.kontextovým diagramem označeným písmenem a číslem 0 (obr. 2.3). Při sestavování diagramů jsou dodržovány zásady jejich řazení ve směru diagonály a diagram by neměl mít méně než tři a více než šest funkcí. Platí zde také důležitá vlastnost těchto diagramů, kdy výstupy dané funkce mohou být vstupem, řízením či mechanismem jiných funkcí. Tímto způsobem jsou definovány vzájemné závislosti mezi funkcemi (převzato z [8]).



Obrázek 1: Ukázka diagramu IDEF [8]

5 UML

UML je zkratka pro Unified Modelling Language. Jedná se o obecně použitelný (unified) modelovací jazyk sloužící pro modelování softwarových systémů. Díky tomuto modelovacímu jazyku můžeme zvládnout téměř celý životní cyklus aplikace, od popisu požadavků, přes modelování statické struktury i dynamického chování aplikace. Bohužel, jedná se o jazyk zejména technicky zaměřený, tudíž nikoliv procesně. Jeho použití pro modelování byznys procesů se tak omezuje spíše na jednodušší a méně komplikované případy. Použitelný je zejména diagram aktivit [12]. Ukázku tohoto diagramu uvádím na obrázku.



Obrázek 2: Aktivitní diagram [12]

Použití tohoto diagramu je tedy možné, ale rozhodně není doporučené. Byznys modelování zkrátka není silnou stránkou UML, který byl primárně určen pro jiné účely. Pro takto snadný proces bude diagram aktivit postačující, u složitějších procesů brzy odhalíte nedostatky a chybějící notace pro zmíněné aspekty procesu. Pro modelování nesouvisející se SW inženýrstvím, určeným například pro tvorbu procesní struktury firmy či podkladů pro reengineering dokonce považuji tento modelovací jazyk za nevhodný.

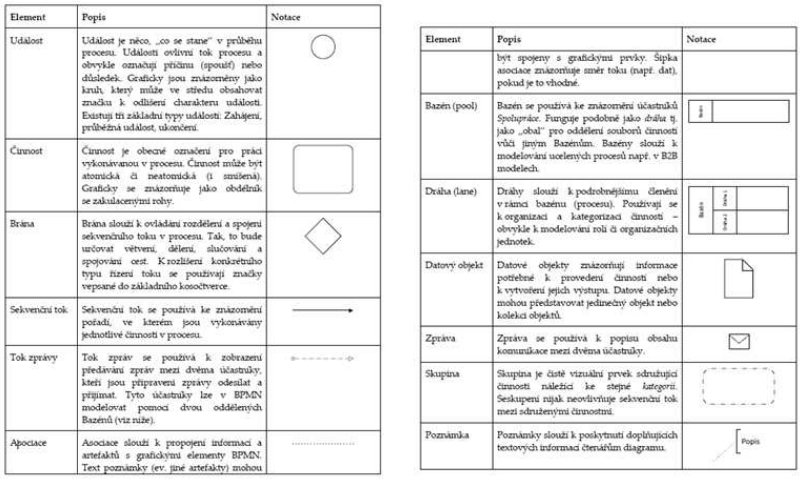
6 BPMN

BPMN vychází z diagramu aktivit UML. Jak jsem naznačil v předchozí kapitole, tyto diagramy jsou použitelné v omezené míře a nejsou přímo určeny pro modelování byznys procesů. Některé jednodušší procesy sice můžeme modelovat pouze pomocí samotného UML, nicméně pro složitější procesy v UML zkrátka postrádáme notaci pro zápis některých skutečností, zejména se pak jedná o hierarchii subprocesů, vazby mezi procesy, transakční zpracování či výjimky [10].

BPMN jde v tomto právě tím směrem, který nám v UML chyběl a přidává právě ty symboly a notace, které chyběly. Cílem BPMN je tedy standardizovaná notace, která bude srozumitelná jak byznys analytikům, tak také zákazníkům, kteří budou moci validovat jednotlivé procesy a jejich správnost a adekvátnost. Na následujících obrázcích pak můžete vidět základní elementy BPMN, ze kterých jsou dále tvořeny diagramy.

Velkou výhodou BPMN je fakt, že modely v něm vytvořené lze poměrně snadno uzpůsobit do podoby, ve kterém je možnost procesy virtuálně spustit a tím testovat, zda jsou skutečně použitelné a zda neobsahují cokoliv, co by mohlo způsobit nemožnost dokončení procesu (deadlock, uváznutí atd.). BPMN definuje, jak převádět jednotlivé elementy a sekvence těchto elementů do jazyka BPEL. Je tedy možné (manuálně) model procesu do jeho spustitelné podoby převést. Díky poměrné volnosti modelování v BPMN není možné vygenerovat BPEL automaticky, některé BPMS nástroje však tuto funkci nabízejí, a to za cenu určitých omezení při samotném modelování procesu [11].

BPMN definuje jediný diagram, tzv. Business Process Diagram (BPD). Ten je tvořen sítí grafických objektů, zejména aktivitami a zobrazením toku informací mezi nimi. Jednotlivé grafické objekty jsou od sebe dobře odlišené, což přispívá k přehlednosti diagramu. Jasně dány jsou tvary těchto objektů, které je třeba dodržovat, je ovšem možné volit pro ně vlastní barvy, například z odlišovacích účelů. V určitých případech lze použít v diagramu i vlastní grafický objekt, ten se však nesmí překrývat s žádným již existujícím a rovněž by neměl ovlivňovat samotný tok procesu, pouze jej upřesňovat, či poskytovat nějaké dodatečné informace [11].

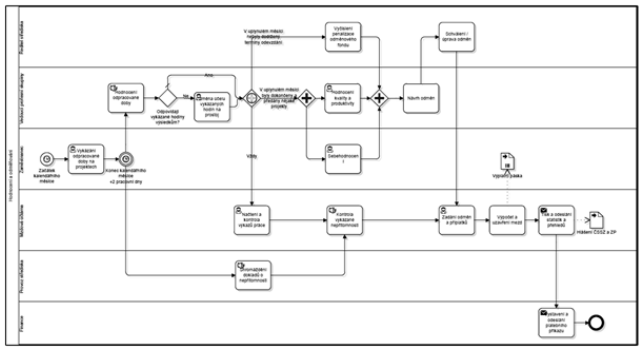


Obrázek 3: Grafické elementy v BPMN (zdroj: OMG, upraveno v [10])

Na následujících obrázcích jsou ukázány možnosti modelování konkrétních procesů pomocí BPMN. První model představuje model procesu Odměňování, druhý pak model procesu Nábor zaměstnanců.

7 ARIS (EPC)

Uvedl jsem metodiku IDEF. Tato metodika specifikuje funkce, které by vyvíjený systém měl obsahovat a plnit. Dále je však nutné definovat posloupnost aktivit, které k tomuto



Obrázek 4: Model procesu v BPMN [10]

cíli povedou a také definovat časovou, prostorovou návaznost a přiřazení rolím. Metoda EPC (Event-driven Process Chain) patří k jedné z nejrozšířenějším především proto, že se stala součástí systémů jako SAP R/3 (ERP/WFM) a ARIS (BPR). Podstata metody, jak vyplývá i z jejího názvu, spočívá v řetězení událostí a aktivit do posloupnosti realizující požadovaný cíl. Z obecného pohledu vykonávání procesu událost definuje vstupní podmínku (precondition) uskutečnění aktivity. Ukončení aktivity pak definuje další událost – výstupní podmínku (postcondition), na kterou mohou navazovat další aktivity. Z toho vyplývá, že každá aktivita je vymezena dvěmi událostmi a tak je i jednoznačně definován její začátek a konec. [8]

EPC diagramy obsahují tři základní prvky:

1. Aktivity - jednotlivé činnosti, které se v rámci procesu provádějí.
2. Události - jsou vstupními a výstupními podmínkami událostí - jinými slovy - daná událost vede k tomu, že je vykonána aktivita a aktivita může vést ke spuštění události.
3. Řídící spojky - umožňují větvit tok procesu.

Ukázka EPC diagramu popisujícího jednoduchý proces je opět na obrázku. Je zvykem kromě grafického odlišení aktivit, spojek a událostí také volit barevné odlišení, které ovšem není přímo popsáno v notaci EPC diagramů. Kromě standardního EPC diagramu existuje také rozšířený EPC diagram (extendedEPC, eEPC), který dále umožňuje modelovat podprocesy, organizační jednotky, podmínky atd. eEPC se tak stává kvalitním a rozhodně použitelným modelovacím nástrojem.

8 Zhodnocení

Uvedl jsem několik metodik, které je možno využívat pro modelování, případně simulaci byznys procesů. Pochopitelně, každý z nich má své uplatnění za daných podmínek. Žádná z metodik není samospásná a každá kromě použití ve správné situaci vyžaduje také znalosti a zkušenosti byznys analytika. V následující tabulce uvádím doporučené použití jednotlivých metodik.

Metodika/nástroj	Použití
UML (Unified Modeling Language	Aktivitní diagram lze využít pro menší, nekomplikované procesy, případně tam, kde procesy nebudou dále využívány pro komunikaci se zákazníky a nepředpokládá se jejich rozšiřování, při kterém by se mohly ukázat nedostatky této metodiky
BPMN (Business Process Modeling Notation).	Obecně použitelná metodika pro popis sekvencí aktivit a dalších aspektů procesu. Použitelný pro rozsáhlé projekty
EPC (Event-driven Process Chain).	Obecně použitelná metodika pro popis sekvencí aktivit a dalších aspektů procesu. Použitelný pro rozsáhlé projekty
IDEF3 (The Integrated DEFinition).	Použitelný pro menší i větší projekty, díky dekompozici je přehledný i u velkých projektů. Je použitelný spíše pro funkční náhled.

Z tohoto popisu tedy vyplývá, že seriozní modelování nebude příliš využívat aktivitní diagram z metodiky UML. Ostatní metodiky jsou použitelné a závisí na tom, jaký pohled chcete použít (funkční, sekvenční), případně jaké máte zkušenosti z praxe.

9 Závěr

V tomto krátkém textu jsem se snažil především o krátké seznámení se základními metodikami modelování procesů. Důvodem takového modelování je, jak bylo uvedeno, zejména pochopení firemních procesů zákazník, ať už pro potřeby vývoje nového SW či pro potřeby vytvoření procesní mapy pro řízení či reengineering. Výše uvedené metodiky umožňují obojí - je tedy možné přiblížit zákazníkovi procesy a zároveň mít k dispozici kvalitní procesní schéma, které umožní vytvořit produkt, který bude dále validován. Text není komplexním a vyčerpávajícím seznamem ani návodem, jak využívat jednotlivé procesní metodiky. U každé metodiky si musíme být vědomi nedostatků a toho, že žádná dnes využívaná procesní metodika není optimální z hlediska překonání sémantických bariér. Z tohoto důvodu bude dále vhodné vést výzkum a metodiky případně doplnit o další prvky.

Literatura

[1] Merunka V., Unal B., Myslín J.. *Formal techniques of data structure design in modern database systems*. In 'IFAFFE'10 Proceedings (Samsun, Turecko 2010)',

[2] Merunka V. *Objektové modelování*. Alfa nakladatelství s.r.o, Praha 2008.

[3] Molhanec M. *Krátká úvaha o normalizaci*. In 'Sborník konference Objekty 2009',

[4] Myslín J. *Využití analogií při výuce OOP*. In 'Sborník konference Informatika XXIII/2010)',

[5] Unal B., Myslin J. *Process Modeling and Business to IT Transformation.* In 'Sborník konference Doktorandské dny',

[6] Myslín J. *Transformace modelů RUP pro potřeby manažera.* In 'Sborník katedry informatiky VŠMIE 2/2009',

[7] Vondrák I. *Metody byznys modelování.* VŠB-TU Ostrava, 2004

[8] Vondrák I. *Úvod do softwarového inženýrství.* VŠB-TU Ostrava, 2002

[9] Mikulášek P. *Reengineering procesů v projektově orientované organizaci.* diplomová práce, ČZU, 2011.

[10] Vašíček P. *Úvod do BPMN.* http://bpm-sme.blogspot.com/2008/03/3-uvod-do-bpmn.html (22.8.2011)

[11] Wikipedia heslo *Diagram aktivit.* Wikipedia, (22.8.2011)

[12] Hammer M. Champy J. . *Reengineering - radik8lní proměna firmy, Manifest revoluce v podnikání.* Management Press, 2000

Source Camera Identification Based on PRNU Invariant to Zoom*

Adam Novozámský

2nd 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. In 2006 was proposed one of the most effective source camera identification technique based on uneven sensitivity of pixels. This unevenness causes unique noise for each sensor known as Photo-Response Non-Uniformity (PRNU). Methods based on PRNU are shown as powerful tool for image forensics. However, their accuracy can be reduced by the interpolation noise or image compression by the camera post-processing. In experiments with compact cameras, we found that the resulting PRNU is changing when we used the zoom in camera. The aim of this paper is analyze these changes and propose a methodology of the detection PRNU of compact cameras with zoom.

Keywords: image forensics, PRNU, source camera identification

Abstrakt. V roce 2006 byla navržena jedna z nejefektivnějších technik na identifikaci fotoaparátu založená na nerovnoměrné citlivosti pixelů. Tato nerovnoměrnost způsobuje šum, nazývaný Photo-Response Non-Uniformity (PRNU), který je specifický pro daný senzor. Metody založené na PRNU se ukazují jako silné nástroje pro forenzní analýzu obrazu. Ačkoli jejich přesnost může být snížena interpolačním šumem nebo kompresí probíhající ve fotoaparátech. Při experimentech s kompaktními fotoaparáty jsme zjistili, že se výsledné PRNU mění při použití zoomu. Cíl této práce je analyzovat tyto změny a navrhnout metodiku detekce PRNU u kompaktních fotoaparátů se zoomem.

Klíčová slova: forenzní analýza obrazu, PRNU, identifikace fotoaparátu

1 Introduction

Source Camera Identification is useful to determine if images were recorded using a specific camera, especially in the court for establishing the origin of images presented as evidence. In the past, many techniques have been described for this type of identification, the most effective was presented by Lukáš et al. [4]. This approach works very well in the basic settings without further editing photos. Castiglion [1] et al. examined some basic photo editing, but no publication has discovered a problem with different camera settings.
This paper deals with the zooming, which is one of the most frequently used camera

*This work has been supported by the grant PIZZARO, VG20102013064

settings. Cameras with adjustable zoom lenses are far more widespread than with fixed-parameter lenses. Therefore, we feel that this publication will be beneficial to use the method in practice.

2 PRNU-based Digital Camera Identification

These days the solid-state image sensors contain hundreds of thousands of pixels (CCD, CMOS). Individual pixels in the grid exhibit imperfection, which leads to *Pattern Noise*. It is a noise component superimposed on the image signal and is independent of it. The two main components of the *Patter Noise* are the *Fixed Pattern Noise* (**FPN**) and the *Photo-response Nonuniformity Noise* (**PRNU**). The first one is caused by *Dark Currents* and it is an additive noise, so most consumer cameras suppress FPN automatically. The second one is caused primarily by different sensitivity of pixels to light. It appears that this PRNU, formed during manufacturing, is unique for each image sensor and PRNUs from two images are correlated if they are coming from the same sensor (camera). If we have enough images with flat regions, we can estimate the PRNU and compare with the PRNU separated from the other images. The separation of PRNU is easy, it just take a picture and removes the noise. Then take the clean picture and subtract it from the original.
For a more accurate description of the method we recommend reading the publication [4], from which we base.

2.1 Our Implementation

We implemented the method proposed by [4] using the Matlab software. We tried both the wavelet [3] and median filter. Median may have a problem if you take pictures too rugged scene (e.g. Persian carpet), but in our test it has the same results as wavelet filter. Therefore, we chose the median for its faster calculations.

3 Experiment Results

As shown in Table 1, nine cameras were available for our experiments, while seven of them were different types. They are common-hundred-dollar compacts generally obtainable on the market, except S100fs, which is super-zoom SLR-like compact with electronic viewfinder (EVF) and was added to the experiment because of the possibility of saving in RAW format. All had a charge coupled device (CCD), but different size from 1/2.9 to the largest with 1/1.5 inch. In order to verify the identification method were chosen three same models L23 from one manufactures.
First, it was necessary to determine the change of PRNU by using zoom, the Section 3.2, and then to compare all the reference PRNU of cameras with their images, the Section 3.2.2. In conclusion we present a simplified method for comparing PRNU, and its results, the Section 3.3.

Table 1: Cameras used in experiments and their properties

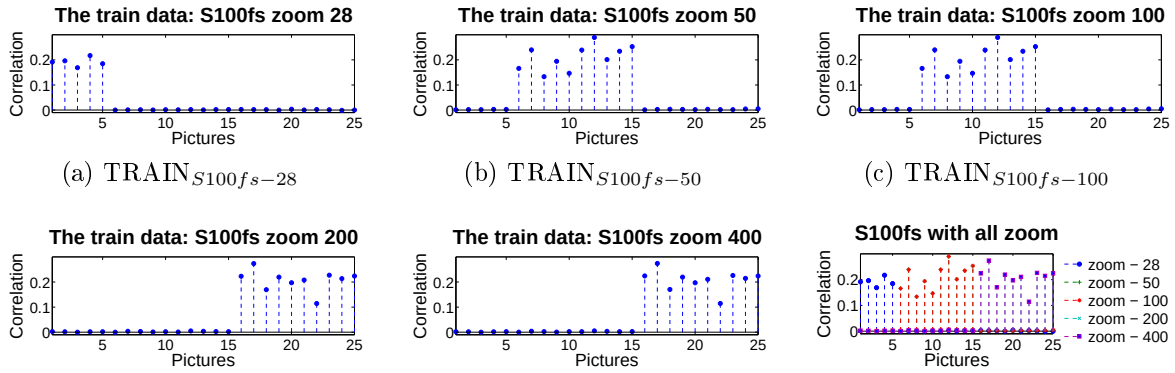
ID	Model	Sensor	Resolution	Opt. Zoom	Focal Length	JPEG	RAW
A495	Canon PowerShot A495	CCD 1/2.3-inch 4:3	3648x2736	3.3x	37-122 mm	x	-
S200	Casio EX S200	CCD 1/2.3-inch 4:3	4320x3240	4x	27-108 mm	x	-
S100fs	Fujifilm FinePix S100fs	CCD 1/1.5-inch 4:3	3840x2880	14.3x	28-400 mm	x	x
L23-1	Nikon Coolpix L23	CCD 1/2.9-inch 4:3	3648x2736	5x	28-140 mm	x	-
L23-2	Nikon Coolpix L23	CCD 1/2.9-inch 4:3	3648x2736	5x	28-140 mm	x	-
L23-3	Nikon Coolpix L23	CCD 1/2.9-inch 4:3	3648x2736	5x	28-140 mm	x	-
S3000	Nikon Coolpix S3000	CCD 1/2.3-inch 4:3	4000x3000	4x	27-108 mm	x	-
P80	Pentax Optio P80	CCD 1/2.3-inch 4:3	4000x3000	4x	27.5-110 mm	x	-
PL51	Samsung PL51	CCD 1/2.33-inch 4:3	3648x2736	3x	35-105 mm	x	-

3.1 Image Preprocessing

All images were captured with full resolution of sensor. First was necessary to square up with different size of images, because the native resolution was unequal between models. We had three possibilities how to solve this problem. **RESIZE** all images to a maximum size in the set, which is 4320x3240. Extract (**SUB**) from all images the sub-image with the same size, e.g. 1024x1024 originated at the beginning (0,0), or **CROP** the larger images to match the smaller images symmetrically from the center, in our case 3648x2736. Like a better method we had chosen the CROP, as [4], because the RESIZE brings additional interpolation noise and the SUB isn't indrawn to the center of the image.
For each camera model three sets of images were captured: the *Images for Reference Pattern* ($\mathbf{TRAIN}_{c-z}$), the *Images for Testing without Zoom* ($\mathbf{TESTnZ}_{c}$) and the *Images for Testing with Zoom* ($\mathbf{TEST}_{c-z}$), where $c \in C = \{A495, S200, ..., Pl51\}$, by ID in Table 1, denotes the using camera and z denotes the using zoom when acquiring images.
The images in TRAIN_{c-z} and TEST_{c-z} sets were taken on a tripod, with zoom, no flash, best JPEG compression quality, and other options were set to their default values. The TESTnZ_{c} were taken without tripod and no zoom (minimum focal length). Each TRAIN_{c-z} has 50 pictures of flat scene as was suggested in [4]. For the flat scene we used the white photographic paper. For all TRAIN_{c-z} was count the reference PRNU.

3.2 PRNU at Different Zoom

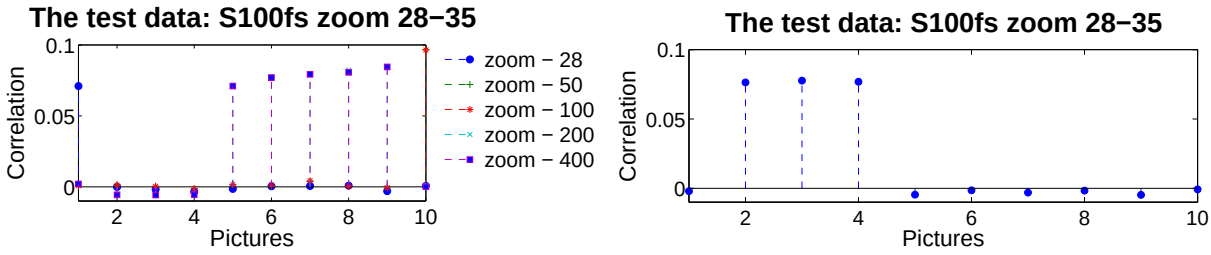
The EVF S100fs is very finely adjusting the zoom, because it is controlled mechanically by turning the lens. Other compacts do not have such fine adjustment of the zoom - mostly a few discrete steps of zoom which are controlled by electric motors. Therefore, we chose S100fs for testing if a sets of TRAIN_{c-z} and TEST_{c-z} give a high correlation for different settings of zoom.
The focal length of the S100fs lens is from 28 to 400 mm, so we took five $\mathrm{TRAIN}_{S100fs-z}$, where $z \in \{28, 50, 100, 200, 400\}$, each 50 images. We took with the same zoom also the $\mathrm{TEST}_{S100fs-z}$, each 5 images. Correlation values are plotted in Figure 1. On each graph corresponds to the first five values of correlations images with zoom 28, 5 more images with zoom 50, 5 more photos with zoom 100, etc. Although the TEST_{S100fs} composed of only 25 pictures, it is obvious that for various zoom, has the sensor a different PRNU. As is it seen in the Graph 1f, it would be sufficient for such a set of photos just three



(a) $\text{TRAIN}_{S100fs-28}$ (b) $\text{TRAIN}_{S100fs-50}$ (c) $\text{TRAIN}_{S100fs-100}$

(d) $\text{TRAIN}_{S100fs-200}$ (e) $\text{TRAIN}_{S100fs-400}$ (f) TRAIN_{S100fs} - all zooms

Figure 1: Scatter plot of the correlations between TEST_{S100fs} and the Image Reference Pattern of the camera with ID S100fs.



(a) $\text{TRAIN}_{S100fs-28,50,100,200,400}$ (b) $\text{TRAIN}_{S100fs-30}$

Figure 2: Correlations between images with zoom 28-35 and the Image Reference Pattern of the camera with ID S100fs.

$\text{TRAIN}_{S100fs-z}$, $z \in \{28, 100, 400\}$.

3.2.1 Entire Zoom Range

We took 5 external scenes so that we set the zoom to minimum (zoom = 28), and then we increased the zoom slightly for each image, until we reached the maximum zoom (zoom = 400). In this way we obtained the 50 photos of each scene, for a total of 250 photos for $\text{TEST}_{S100fs-all}$.
We found out after counting correlations that none of our TRAIN_{S100fs} makes good results in the zoom range 29-35, Figure 2a. So we took another set of zoom 30 - $\text{TRAIN}_{S100fs-30}$. The correlation show good response for this set, Figure 2b. Figure 3 presents the scatter plot of the correlations between images from $\text{TEST}_{S100fs-all}$ and the reference pattern of the $\text{TRAIN}_{S100fs-28,30,100,400}$. Although the zoom range is quite considerable, we covered it with only a few reference patterns. So we have assumed we can proceed similarly by the other cameras. We took more TRAIN sets for one camera, so we needed to know if some pictures from other eight cameras do not give False Positive. Analyzing Figure 4, it can be noted that all correlations with 1000 images (eight cameras, each 125) are an order of magnitude lower than the lowest correlations in Figure 3.

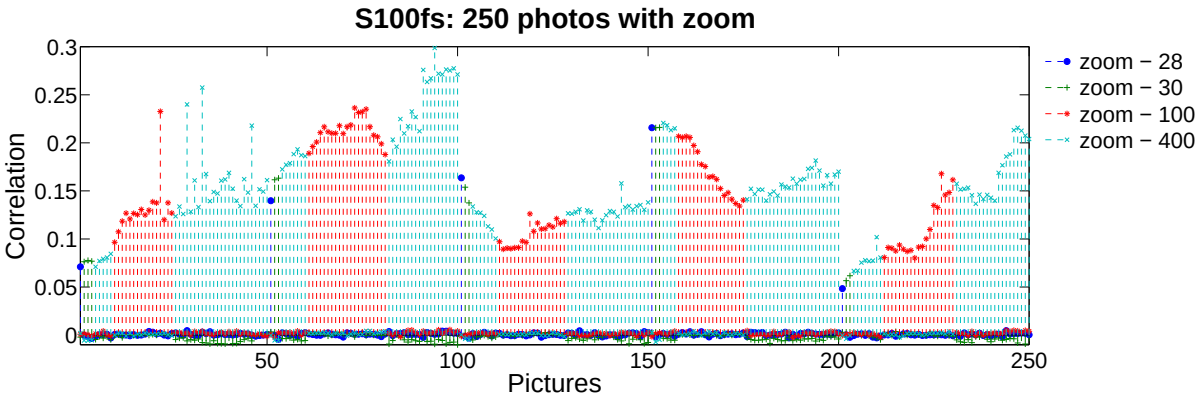


Figure 3: Correlations between 250 images of the camera with ID S100fs and its Image Reference Pattern.

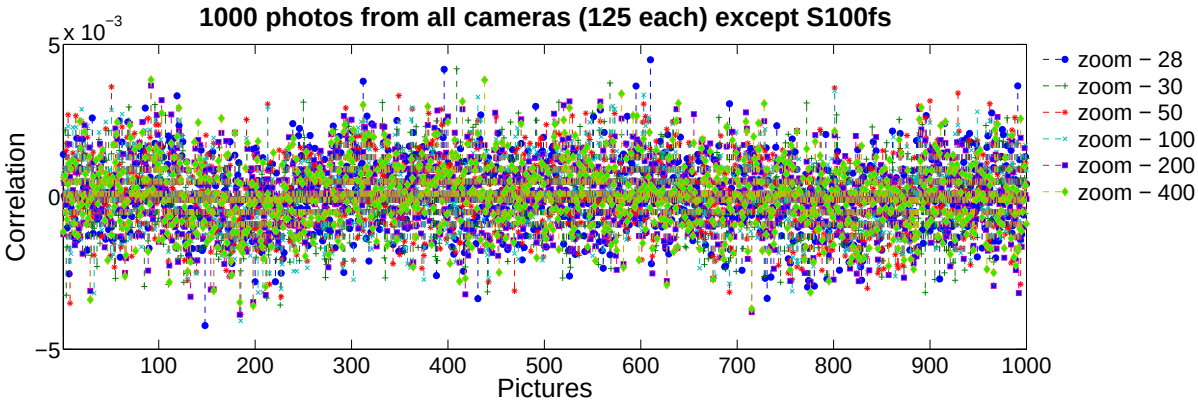


Figure 4: Scatter plot of the correlations between 1000 images form other cameras and the Image Reference Pattern of S100fs.

Table 2: Number of TRAIN for cameras in experiment

ID	A495	S200	L23-1	L23-2	L23-3	S3000	P80	PL51
Num. of TRAIN in Test	7	4	4	4	4	5	5	2
Min of TRAIN	1	4	4	4	4	5	4	1

3.2.2 Testing of Particular Cameras

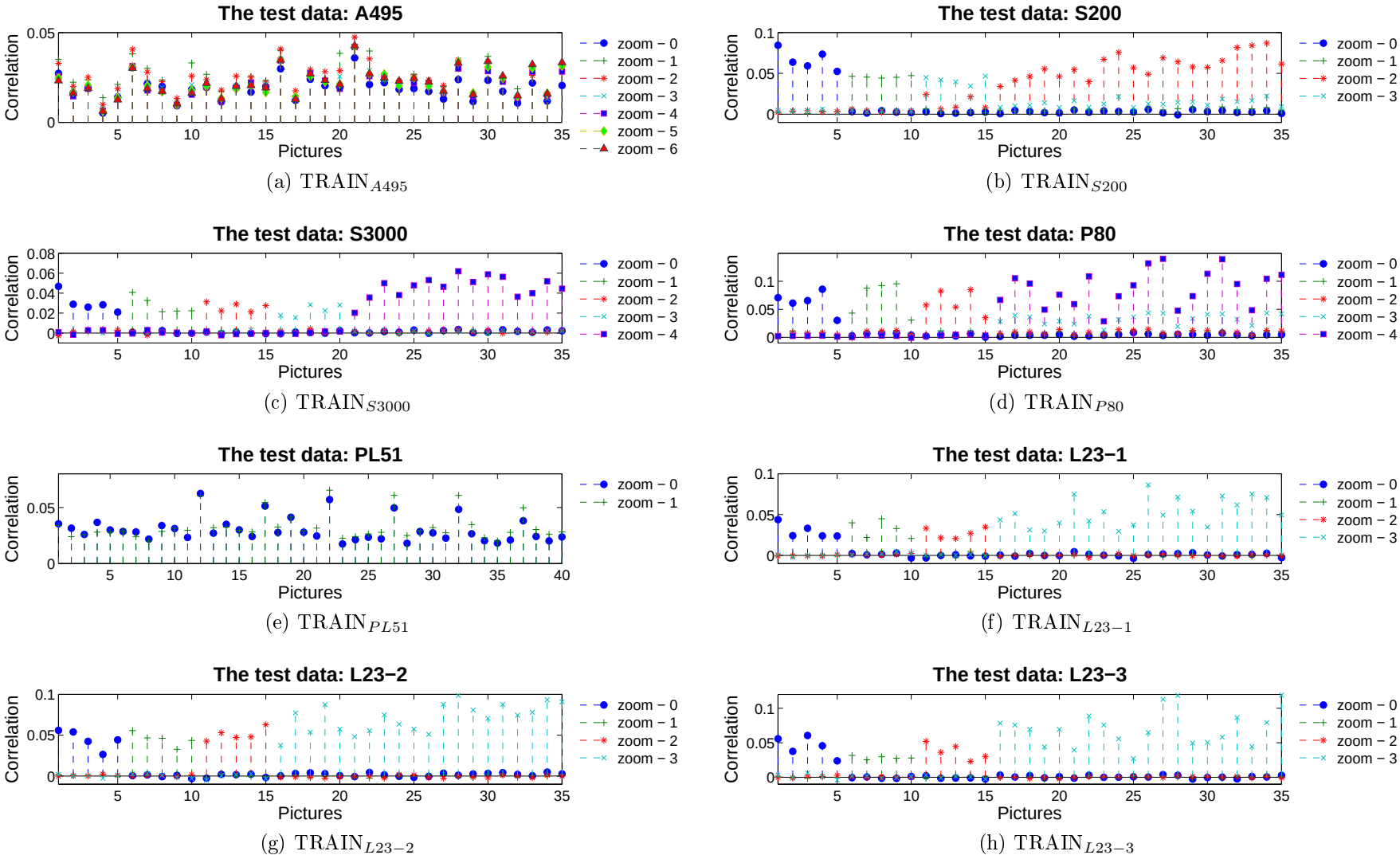
The classic low-cost compacts have only a few discrete steps of zoom, in our test no more than 10, but some just 7. Therefore, we decided to shoot a testing set consisting of 35 photos for each camera, TEST_{c-z}, where z equals 1 to 7, and where every 5 pictures is photographed by another zoom. For different models we took different number of TRAIN sets, see second row of Table 2. Results of testing of individual cameras can be seen on the graphs in Figure 5. As shown in Graph 5a and 5e some compacts do not need more than one TRAIN_c, but most need at least four. Figure 6 is a comparison of three compacts L23 among themselves. First 35 images are from L23-1, and another 35 from L23-2 and the last 35 from L23-3.

3.3 Experiment with Cropping

According to what we read in [4], PRNU should not depend on anything other than the properties of the actual sensor itself. We therefore believe that changes in PRNU using the zoom camera are caused by post-processing. To verify this, we took into RAW format 50 pictures one scene with varying zoom with S100fs. This camera has the option to save the RAW format as a only one in the test, if you like the RAF format (Fuji CCD-RAW Graphic File). The computed correlation of individual $\text{TRAIN}_{S100fs-z}$ is in Figure 7. Higher correlation has only $\text{TRAIN}_{S100fs-200}$ and $\text{TRAIN}_{S100fs-400}$, so sets shoot with more than a half zoom. We suppose it could be a correction of vignetting, which causes a change of PRNU at low zoom levels. We have many types of vignetting such as *Mechanical*, *Optical*, *Natural* or *Pixel Vignetting* [2]. Some types of vignetting can be completely cured by lens settings (special filters), but most digital cameras use built-in image processing to compensate vignetting when converting raw sensor data to standard image formats such as JPEG or TIFF.
Vignetting appears mainly at the edge of the image so we cropped all images in the $\text{TRAIN}_{S100fs-28,400}$ and TEST_{S100fs} relative to the center. Figures 8a, 8b, and 8c show the graphs for individual cuts - 1000x1000px, 500x500px, and 256x256px, and 8d shows correlation for 1125 images cropped to 256x256px from all cameras (nine cameras, each 125 according to Table 1). Figure 251-375 is shooting with S100fs. Although the tests of this camera fared well, not all other cameras given up good results in this cropping (256x256px), Figure 10. The Figure 9 shows the same graphs as 8d but for A495 and L23-3. For A495 we see a false positive for some images from other cameras. This method should therefore be used only in rare cases and we do not recommend it.

4 Conclusion

Source Camera Identification Based on PRNU was studied in this paper. The experimental results show, first of all, that the using zoom has a considerable influence on the resulting PRNU. Not by itself, as we think but used by the post-processing. This problem can be solved by comparing pictures with multiple TRAIN sets with different zoom, which is more difficult to lengthy, but gives good results. The second cropping method can be used as the first sighting. The research will be continued by testing various camera



(a) TRAIN_{A495} (b) TRAIN_{S200}

(c) TRAIN_{S3000} (d) TRAIN_{P80}

(e) TRAIN_{PL51} (f) TRAIN_{L23-1}

(g) TRAIN_{L23-2} (h) TRAIN_{L23-3}

Figure 5: Graphs of correlation between TRAIN_{c-z} and TEST_{c-z} for the same c and variable z

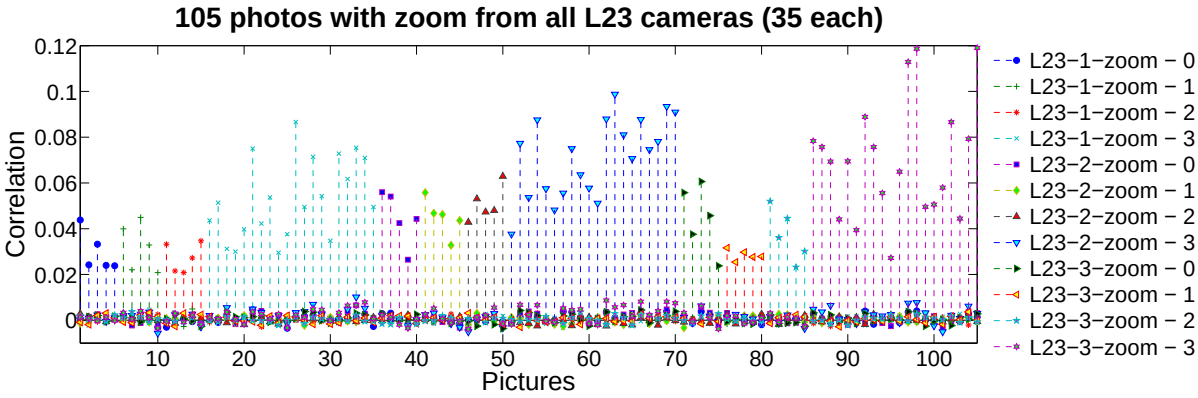


Figure 6: Comparison of three same models Nikon Coolpix L23.

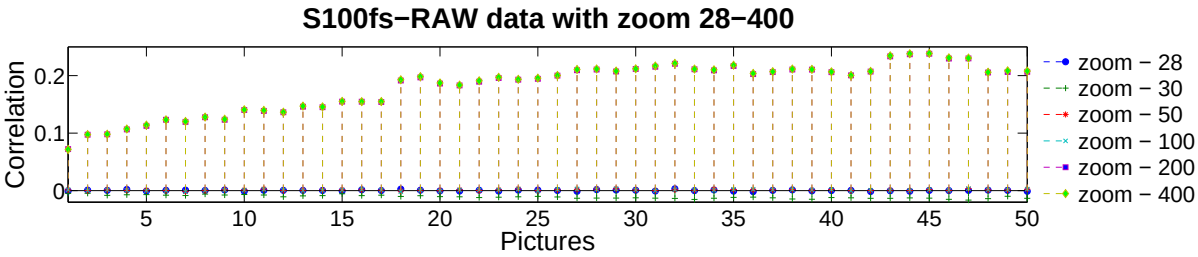


Figure 7: 50 RAW images and their correlation with $\mathrm{TRAIN}_{S100fs-z}$.

settings and shooting modes, such as flash, lower resolution, or macro.

5 Acknowledgment

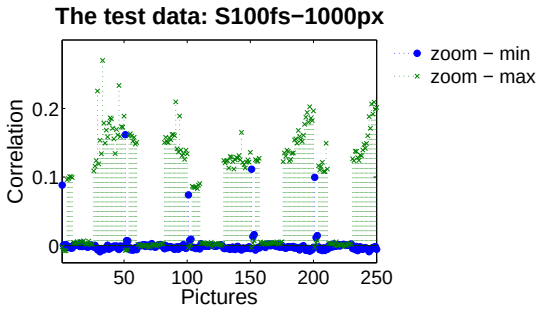
The author would like to thank his colleague Babak Mahdian and his advisor Stanislav Saic, for their valuable suggestions during the research phases.

A How to Read Graphs in This Paper

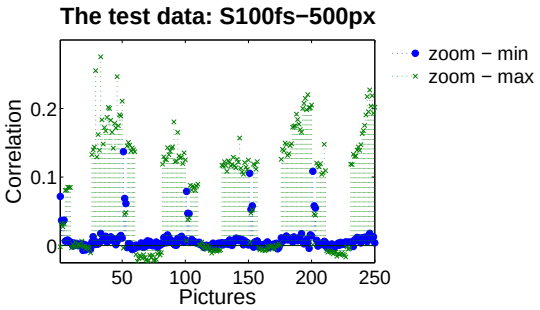
All graphs are scatter plots in this paper. The horizontal axis represents the individual pictures and the vertical axis indicates their correlation of PRNU with appropriate TRAIN sets. If there are more TRAIN sets in the graph, it contains a legend that tell us how the particular TRAIN sets are represent - what color and symbol.

References

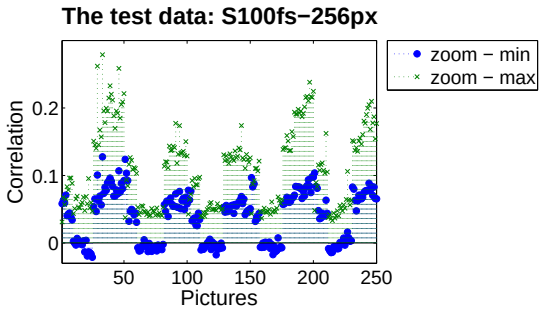
[1] A. Castiglione, G. Cattaneo, M. Cembalo, and U. Petrillo. *Source camera identification in real practice: A preliminary experimentation.* In 'Broadband, Wireless Computing, Communication and Applications (BWCCA), 2010 International Conference on', 417 –422, (nov. 2010).



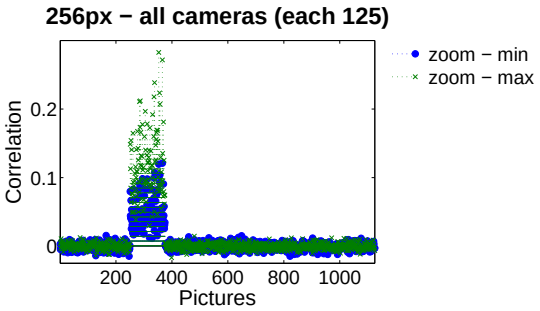
(a) $\text{TRAIN}_{S100fs-28,400}$ 1000x1000px



(b) $\text{TRAIN}_{S100fs-28,400}$ 500x500px

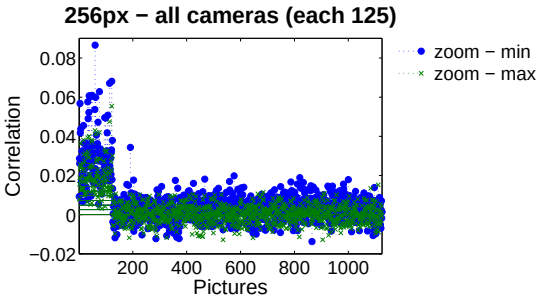


(c) $\text{TRAIN}_{S100fs-28,400}$ 256x256px

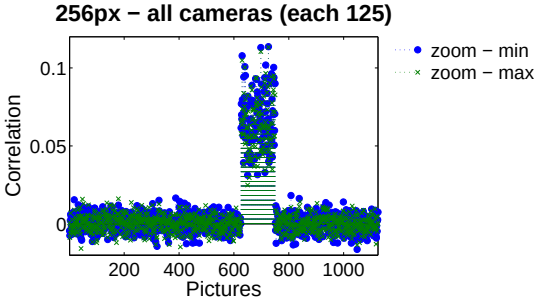


(d) $\text{TRAIN}_{S100fs-28,400}$ 256x256px with 1125 form all cameras

Figure 8: Changing PRNU when changing cropping (a), (b), (c), and False Positive for 256x256px (d).

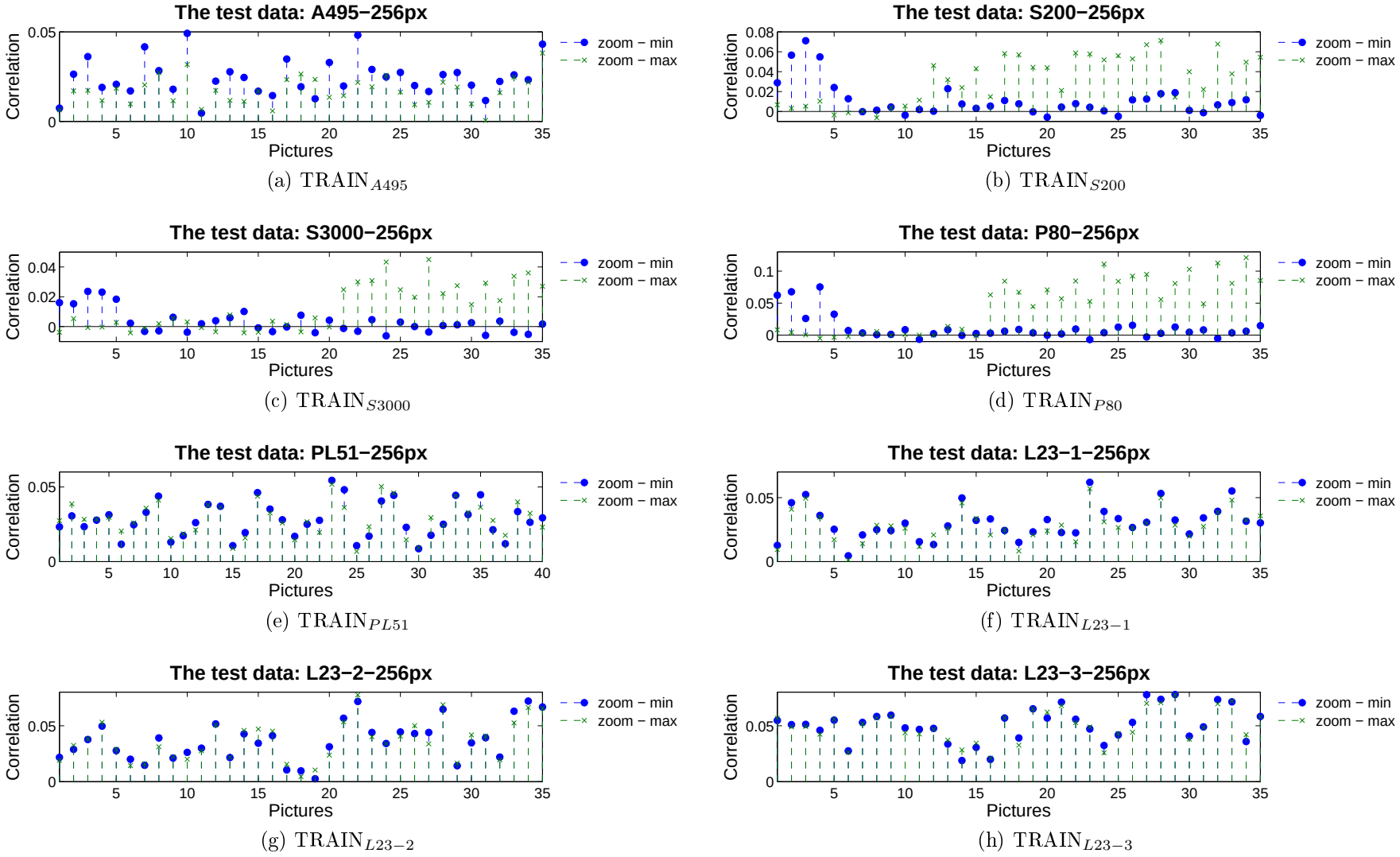


(a) $\text{TRAIN}_{A495-0,6}$ with 1125 images



(b) $\text{TRAIN}_{A495-0,3}$ with 1125 images

Figure 9: False Positive for A495 (a) and L23-3 (b) for 256x256px.



(a) TRAIN_{A495} (b) TRAIN_{S200}

(c) TRAIN_{S3000} (d) TRAIN_{P80}

(e) TRAIN_{PL51} (f) TRAIN_{L23-1}

(g) TRAIN_{L23-2} (h) TRAIN_{L23-3}

Figure 10: Grafy korelací PRNU TRAIN_{S100fs} množin s TEST_{S100fs} množinou

[2] D. Goldman and J.-H. Chen. *Vignette and exposure calibration and compensation.* In 'Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on', volume 1, 899 – 906 Vol. 1, (oct. 2005).

[3] F. Luisier, T. Blu, and M. Unser. *A new sure approach to image denoising: Interscale orthonormal wavelet thresholding.* Image Processing, IEEE Transactions on **16** (march 2007), 593 –606.

[4] J. Lukas, J. Fridrich, and M. Goljan. *Digital camera identification from sensor pattern noise.* Information Forensics and Security, IEEE Transactions on **1** (june 2006), 205 – 214.

Numerický model pro výpočet proudění směsi v porézním prostředí*

Ondřej Polívka

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

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

Abstract. The paper deals with the numerical modeling of compressible single-phase flow of a mixture composed of several components in a porous medium. The mathematical model is formulated by means of Darcy's law, components continuity equations, constitutive relations, and appropriate initial and boundary conditions. The problem is solved numerically using a combination of the mixed-hybrid finite element method for Darcy's law discretization and the finite volume method for the discretization of the transport equations. This approach provides exact local mass balance. The time discretization is carried out by the Euler method. The resulting large system of nonlinear algebraic equations is solved by the Newton-Raphson iterative method. The dimensions of obtained system of linear algebraic equations are significantly reduced so that they do not depend on the number of mixture components. The convergence of the numerical scheme is verified on a problem of methane injection into a homogeneous 2D reservoir filled with propane.

Keywords: mixed-hybrid finite element method, finite volume method, Newton-Raphson method, single-phase compressible multicomponent flow, miscible displacement

Abstrakt. Článek pojednává o numerickém modelování stlačitelného jednofázového proudění směsi o několika složkách v porézním prostředí. Matematický model je formulován pomocí Darcyho zákona, rovnic kontinuity pro složky směsi, konstitutivních vztahů a vhodných počátečních i okrajových podmínek. Úloha je řešena numericky kombinací smíšené hybridní metody konečných prvků použitou pro diskretizaci Darcyho zákona a metody konečných objemů pro diskretizaci transportních rovnic. Tento přístup poskytuje přesnou lokální bilanci hmoty. Časová diskretizace je provedena Eulerovou metodou. Výsledná soustava nelineárních algebraických rovnic je řešena Newtonovou-Raphsonovou iterační metodou. Rozměry získané soustavy lineárních algebraických rovnic jsou významně zredukovány tak, že nezávisí na počtu komponent směsi. Konvergence numerického schématu je ověřena na problému vtláčení metanu do homogenního 2D rezervoáru naplněného propanem.

Klíčová slova: smíšená hybridní metoda konečných prvků, metoda konečných objemů, Newtonova-Raphsonova metoda, jednofázové stlačitelné vícekomponentní proudění, mísitelné proudění

*Tato práce byla podpořena grantem "Mathematical modeling of multi-phase porous media flow", 201/08/P567, Grantové agentury České republiky a projektem "Numerical Methods for Multiphase Flow and Transport in Subsurface Environmental Applications", Kontakt ME10009, Ministerstva školství, mládeže a tělovýchovy České republiky

1 Úvod

Spolehlivá simulace transportu vícekomponentní směsi v podzemním porézním prostředí je důležitá při řešení řady problémů, jako je např. těžba ropy nebo sekvestrace CO_2. Klasické přístupy se zaměřují na plně implicitní nebo sekvenční metody [19, 5]. Plně implicitní přístup je stabilní, umožňuje volbu dlouhých časových kroků, ale vede k obrovským systémům lineárních algebraických rovnic, jejichž rozměry jsou úměrné počtu složek směsi. Při použití sekvenčních metod, jako je IMPEC (implicitní tlak, explicitní koncentrace) [12], se rovnice pro tlak získává sčítáním transportních rovnic [19, 5] nebo jinými metodami [11, 20, 1]. Tlak se pak počítá implicitně za použití koncentrací z předchozího časového kroku a molární zlomky se získávají explicitně. Tento postup umožňuje redukci rozměrů řešené soustavy, nicméně je často nutné volit velmi malé časové kroky.

V této práci se zabýváme numerickým modelováním stlačitelného jednofázového proudění směsi složené z několika komponent v porézním prostředí. Navrhujeme vlastní přístup postavený na kombinaci smíšené hybridní metody konečných prvků (MHFEM) a metody konečných objemů (FVM). Podobně jako u implicitních schémat vede i naše metoda k velkým systémům lineárních algebraických rovnic, ale jejich rozměry je v našem případě možné zredukovat tak, že nezávisí na počtu složek směsi. Výpočetní nároky jsou tedy srovnatelné se sekvenčními přístupy, u kterých je ale nutné sestavovat rovnici pro tlak. V našem případě je tlak získán přímo ze stavové rovnice.

2 Matematická formulace

Nechť $\Omega \subset \mathbb{R}^2$ je omezená oblast s porozitou ϕ [-] a $(0, \tau)$ je časový interval [s]. Uvažujme jednofázové stlačitelné proudění tekutiny o N_C složkách v oblasti při konstantní teplotě T [K]. Při zanedbání difúze je transport jednotlivých složek popsán následujícími rovnicemi [11]

$$\begin{aligned} \frac{\partial(\phi c_i)}{\partial t} + \nabla \cdot (c_i \mathbf{q}) &= f_i \,, \quad i = 1, \ldots, N_C \,, \\ c_i &= c_i(\mathbf{x}, t) \,, \quad \mathbf{x} \in \Omega \,, \ t \in (0, \tau) \,, \end{aligned} \tag{1}$$

kde neznámé veličiny c_i, $i = 1, \ldots, N_C$, jsou molární koncentrace komponent směsi [$\mathrm{mol\,m^{-3}}$]. Na pravé straně rovnice (1) stojí zdrojový člen f_i [$\mathrm{mol\,m^{-3}s^{-1}}$]. Darcyho rychlost $\mathbf{q}$ [$\mathrm{m\,s^{-1}}$] je dána Darcyho zákonem (viz [2])

$$\mathbf{q} = -\mu^{-1}\mathbf{K}(\nabla p - \varrho \mathbf{g}) \,, \tag{2}$$

kde $\mathbf{K} \in [L^\infty(\Omega)]^{2\times 2}$ je vlastní permeabilita [$\mathrm{m^2}$] (obecně symetrický stejnoměrně eliptický tenzor [15]), μ je viskozita [$\mathrm{kg\,m^{-1}s^{-1}}$], ∇p označuje gradient tlaku p [Pa], $\mathbf{g}$ je vektor gravitačního zrychlení [$\mathrm{m\,s^{-2}}$] a ϱ je hustota tekutiny [$\mathrm{kg\,m^{-3}}$]. Rovnice (1), (2) jsou svázány konstitutivními vztahy vyjadřujícími závislosti

$$p = p\,(c_1, \ldots, c_{N_C}, T) \,, \quad \mu = \mu\,(c_1, \ldots, c_{N_C}, T) \,, \quad \varrho = \varrho\,(c_1, \ldots, c_{N_C}) \,. \tag{3}$$

Tlak je předepsán Pengovou-Robinsonovou stavovou rovnicí (PR EOS) [17], viskozita je dána Lohrenzovou-Brayovou-Clarkovou (LBC) metodou [14]. Hustotu spočteme dle [10].

Počáteční a okrajové podmínky jsou následující

$$c_i(\mathbf{x},0) = c_i^0(\mathbf{x})\,, \quad \mathbf{x}\in\Omega\,,\ i=1,\dots,N_C\,, \tag{4a}$$

$$c_i(\mathbf{x},t) = c_i^D(\mathbf{x},t)\,, \quad \mathbf{x}\in\Gamma_c(t)\,,\ t\in(0,\tau)\,,\ i=1,\dots,N_C\,, \tag{4b}$$

$$p(\mathbf{x},t) = p^D(\mathbf{x},t)\,, \quad \mathbf{x}\in\Gamma_p\,,\ t\in(0,\tau)\ , \tag{4c}$$

$$\mathbf{q}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = q^N(\mathbf{x},t)\,, \quad \mathbf{x}\in\Gamma_q\,,\ t\in(0,\tau)\ , \tag{4d}$$

kde $\mathbf{n}$ je jednotkový vektor vnější normály k hranici $\partial\Omega$. Rovnice (4c), (4d) určují Dirichletovy a Neumannovy okrajové podmínky na částech hranice Γ_p, resp. Γ_q, přičemž platí $\Gamma_p\cup\Gamma_q=\partial\Omega$ a $\Gamma_p\cap\Gamma_q=\emptyset$. Okrajová podmínka (4b) pro molární koncentrace je také Dirichletova typu. Množina $\Gamma_c(t)$ značí vtokovou část hranice $\partial\Omega$ v čase t, tj.

$$\Gamma_c(t) = \{\mathbf{x}\in\partial\Omega\,|\,\mathbf{q}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x})<0\}\ .$$

3 Numerické řešení

Systém rovnic (1)–(4) je řešen numericky kombinací MHFEM aplikovanou na Darcyho zákon (2) a FVM aplikovanou na transportní rovnice (1). Časová diskretizace je provedena Eulerovou metodou a výsledné schéma získáno linearizací Newtonovou-Raphsonovou metodou (NRM).

Uvažujme 2D polygonální oblast Ω s hranicí $\partial\Omega$, která je rozdělena triangulací $\mathcal{T}_\Omega$ na trojúhelníky. Označme K prvek triangulace $\mathcal{T}_\Omega$ s plošným obsahem $|K|$, E je hrana trojúhelníku o délce $|E|$, N_K pak počet všech elementů triangulace a N_E počet hran trojúhelníkové sítě.

3.1 Diskretizace Darcyho zákona

Darcyho rychlost $\mathbf{q}$ lze aproximovat v Raviartově-Thomasově prostoru nejnižšího řádu (RT^0_K) nad elementem $K\in\mathcal{T}_\Omega$ jako

$$\mathbf{q} = \sum_{E\in\partial K} q_{K,E}\mathbf{w}_{K,E}\,, \tag{5}$$

kde koeficient $q_{K,E}$ vyjadřuje tok vektorové funkce $\mathbf{q}$ přes hranu E elementu K vzhledem k vnější normále a $\mathbf{w}_{K,E}$ po částech lineární bazickou funkci prostoru RT^0_K příslušející hraně E (viz [3, 4, 16]).

Vyjádřením gradientu tlaku z Darcyho zákona (2) získáme

$$\nabla p = -\mu\mathbf{K}^{-1}\mathbf{q} + \varrho\mathbf{g}\,. \tag{6}$$

Vynásobením (6) bazickou funkcí $\mathbf{w}_{K,E}$, integrací přes K, využitím vlastností prostoru RT^0_K, vztahu (5), Greenovy věty a věty o střední hodnotě odvodíme diskrétní tvar Darcyho zákona

$$q_{K,E} = \mu_K^{-1}\left(\alpha_E^K p_K - \sum_{E'\in\partial K}\beta_{E,E'}^K p_{K,E'} + \gamma_E^K\varrho_K\right)\,, \quad E\in\partial K. \tag{7}$$

V rovnici (7) jsou $\alpha_E^K, \beta_{E,E'}^K$ a γ_E^K koeficienty závisející na geometrii sítě a lokálních hodnotách permeability; p_K , $p_{K,E'}$ je průměrná hodnota tlaku na elementu K, resp. na hraně E' ; μ_K , ϱ_K značí střední hodnotu viskozity a hustoty na trojúhelníku K.

Spojitost toku a tlaku na hraně E mezi sousedícími elementy $K, K' \in \mathcal{T}_\Omega$ lze zapsat jako

$$q_{K,E} + q_{K',E} = 0\,, \tag{8}$$

$$p_{K,E} = p_{K',E} =: p_E\,. \tag{9}$$

Okrajové podmínky (4c), (4d) vyjádřené v diskrétním tvaru jsou

$$p_{K,E} = p^D(E)\,, \quad \forall E \subset \Gamma_p\,, \tag{10a}$$

$$q_{K,E} = q^N(E)\,, \quad \forall E \subset \Gamma_q\,, \tag{10b}$$

kde $p^D(E)$ je předepsaná hodnota tlaku p na hraně E a $q^N(E)$ předepsaný tok skrz hranu E.

Tok můžeme eliminovat dosazením $q_{K,E}$ ze vztahu (7) do rovnic (8) a (10b). Pro další odvození označme časově závislé veličiny v čase t_{n+1} horním indexem $n+1$. Pak rovnice (7)–(10) přejdou na následující soustavu N_E lineárních algebraických rovnic

$$F_E \equiv \begin{cases} \sum\limits_{K:E\in\partial K} \left(\mu_K^{n+1}\right)^{-1}\left(\alpha_E^K p_K^{n+1} - \sum\limits_{E'\in\partial K} \beta_{E,E'}^K p_{K,E'}^{n+1} + \gamma_E^K \varrho_K^{n+1}\right) = 0 & \forall E \not\subset \partial\Omega\,, \\ \left(\mu_K^{n+1}\right)^{-1}\left(\alpha_E^K p_K^{n+1} - \sum\limits_{E'\in\partial K} \beta_{E,E'}^K p_{K,E'}^{n+1} + \gamma_E^K \varrho_K^{n+1}\right) - q^N(E) = 0 & \forall E \subset \Gamma_q\,, \\ p_{K,E}^{n+1} - p^D(E) = 0 & \forall E \subset \Gamma_p\,. \end{cases} \tag{11}$$

Zde symbol $\sum\limits_{K:E\in\partial K}$ značí sčítání přes elementy obsahující hranu E. Podobný postup vedoucí ke smíšené hybridní formulaci lze nalézt v [15].

3.2 Aproximace transportních rovnic

Transportní rovnice (1) s počátečními a okrajovými podmínkami (4) jsou diskretizovány pomocí FVM [13]. Integrací (1) přes libovolný element $K \in \mathcal{T}_\Omega$ a použitím Greenovy věty dostaneme

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_K \phi(\mathbf{x}) c_i(\mathbf{x},t) + \int\limits_{\partial K} c_i(\mathbf{x},t)\mathbf{q}(\mathbf{x},t)\cdot\mathbf{n}_{\partial K}(\mathbf{x}) = \int\limits_K f_i(\mathbf{x})\,, \quad i = 1,\ldots,N_C\,. \tag{12}$$

Aplikováním věty o střední hodnotě a označením $\phi_K\,, c_i|_K\,, f_i|_K$ průměrných hodnot $\phi, c_i\,, f_i$ $(i = 1,\ldots,N_C)$ přes element K, přejde rovnice (12) na

$$\frac{\mathrm{d}(\phi_K c_i|_K)}{\mathrm{d}t}|K| + \sum_{E\in\partial K} \widetilde{c_i}|_E \underbrace{\int\limits_E \mathbf{q}\cdot\mathbf{n}_{K,E}}_{=q_{K,E}} = f_i|_K|K|\,, \tag{13}$$

kde $\widetilde{c_i}|_E$ představuje koncentraci c_i na hraně E. Integrál v (13) je roven toku skrz hranu E elementu K (složka $\mathbf{q}$ ve směru vnější normály k E).

Předpokládejme, že porozita nezávisí na čase. Časová derivace $c_i|_K$ v (13) je aproximována časovou diferencí s časovým krokem Δt_n. Při použití Eulerovy metody [13], máme pro každé n, všechny elementy $K \in \mathcal{T}_\Omega$ a komponenty $i = 1, \ldots, N_C$

$$F_{K,i} \equiv \phi_K |K| \frac{c_i|_K^{n+1} - c_i|_K^n}{\Delta t_n} + \sum_{E \in \partial K} \widetilde{c}_i|_E^n \, q_{K,E}^{n+1} \left(p_{K,E}^{n+1}, c_1|_K^{n+1}, \ldots, c_{N_C}|_K^{n+1} \right) - f_i|_K |K| = 0 \,, \tag{14}$$

kde $q_{K,E}^{n+1}$ je dáno vztahem (7). Hodnota $\widetilde{c}_i|_E^n$ je volena jako koncentrace ze sousedícího prvku v upwindovém směru, tj.

$$\widetilde{c}_i|_E^n = \begin{cases} c_i|_K^n & \text{pro} \quad q_{K,E}^{n+1} \geq 0 \,, \\ c_i|_{K'}^n & \text{pro} \quad q_{K,E}^{n+1} < 0 \ \wedge E \not\subset \partial\Omega : K \cap K' = E \,, \\ c_i^D|_E^n & \text{pro} \quad q_{K,E}^{n+1} < 0 \ \wedge E \subset \partial\Omega \,, \end{cases} \tag{15}$$

kde c_i^D značí koncentraci i-té komponenty na vtokové hranici. Poznamenejme, že schéma je téměř plně implicitní, jediný člen v (14), který je vyčíslen explicitně, je hodnota $\widetilde{c}_i|_E^n$.

Počáteční a okrajové podmínky (4a), (4b) v diskrétním tvaru můžeme psát jako

$$c_i|_K^0 = c_i^0(K) \,, \quad \forall K \in \mathcal{T}_\Omega \,, \ i = 1, \ldots, N_C \,, \tag{16a}$$

$$\widetilde{c}_i|_E^n = c_i^D(E, t_n) \,, \quad \forall E \subset \Gamma_c(t) \,, \ i = 1, \ldots, N_C \,, \ t_n < \tau \,. \tag{16b}$$

3.3 Propojení schémat z MHFEM a FVM

Označme F_E a $F_{K,i}$, pro hranu $E \in \{1, \ldots, N_E\}$, element $K \in \{1, \ldots, N_K\}$ a komponentu $i \in \{1, \ldots, N_C\}$, levé strany rovnic (11) a (14) s členem $q_{K,E}^{n+1}$ dosazeným ze vztahu (7). Průměrné hodnoty $p_K = p_K(c_1|_K, \ldots, c_{N_C}|_K)$, $\varrho_K = \varrho_K(c_1|_K, \ldots, c_{N_C}|_K)$ a $\mu_K = \mu_K(c_1|_K, \ldots, c_{N_C}|_K)$ jsou vyčísleny za použití konstitutivních vztahů (3). Soustava $N_E + N_K \times N_C$ rovnic

$$\mathbf{F} = [F_1, \ldots, F_{N_E}; F_{1,1}, \ldots, F_{1,N_C}, \ldots, F_{N_K,1}, \ldots, F_{N_K,N_C}]^T = \mathbf{0} \tag{17}$$

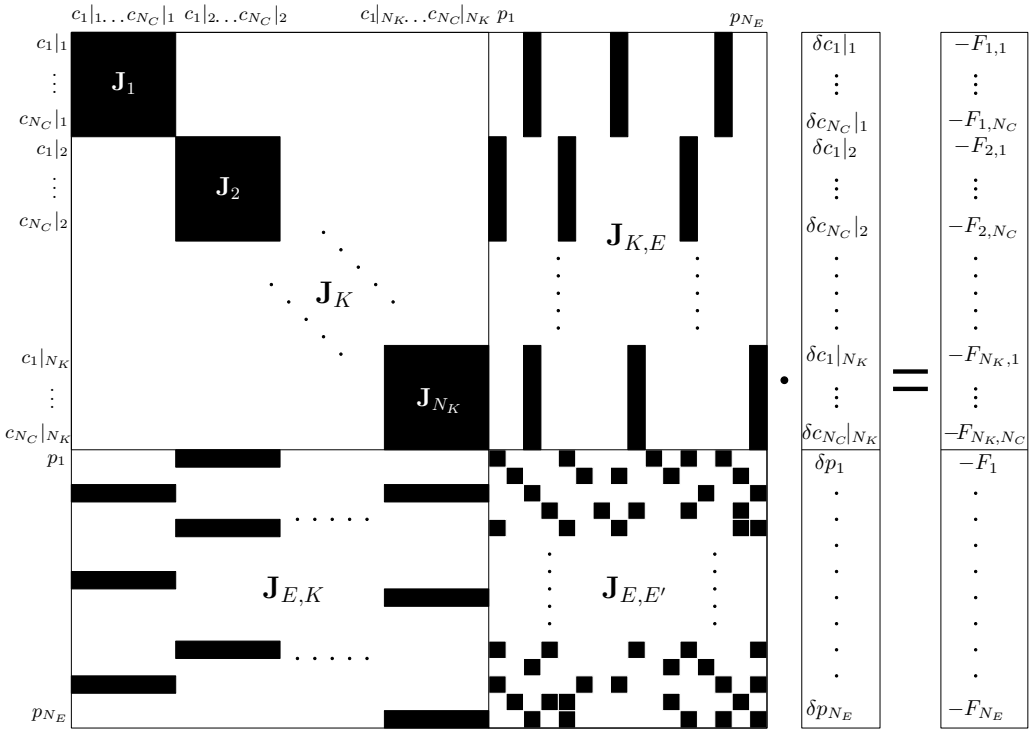
pro neznámé molární koncentrace $c_1|_K^{n+1}, \ldots, c_{N_C}|_K^{n+1}$, $K \in \{1, \ldots, N_K\}$ a tlaky na hranách p_E^{n+1}, $E \in \{1, \ldots, N_E\}$, je nelineární systém algebraických rovnic, který řešíme pomocí NRM. Výsledná soustava lineárních algebraických rovnic je zobrazena na obr. 1. Jacobiho matice je řídká, nesymetrická a vektor neznámých obsahuje korekce molárních koncentrací a tlaků na hranách. Nenulové černě vybarvené hodnoty na obr. 1 jsou dány parciálními derivacemi

$$(\mathbf{J}_K)_{i,j} = \frac{\partial F_{K,i}}{\partial c_j|_K^{n+1}} \,, \quad (\mathbf{J}_{K,E})_i = \frac{\partial F_{K,i}}{\partial p_{K,E}^{n+1}} \,, \quad (\mathbf{J}_{E,K})_j = \frac{\partial F_E}{\partial c_j|_K^{n+1}} \,, \quad J_{E,E'} = \frac{\partial F_E}{\partial p_{K,E'}^{n+1}} \,, \tag{18}$$

kde $J_{E,E'}$ je prvek matice $\mathbf{J}_{E,E'}$, $i, j = 1, \ldots, N_C$; $K = 1, \ldots, N_K$; $E, E' = 1, \ldots, N_E$. Derivace v (18) mohou být vyčísleny analyticky s využitím (3), (11) a (14).

Rozměry soustavy na obr. 1 lze zredukovat invertováním bloků $\mathbf{J}_K$ pro každé K (bloky jsou diagonálně dominantní pro dostatečně malé časové kroky) a eliminací vektorů $\mathbf{J}_{E,K}$ pro všechny E, K. Poté získáváme následující systém N_E rovnic pro N_E korekcí tlaků p_E

$$\sum_{K:E \in \partial K} \sum_{E' \in \partial K} \left(J_{E,E'} - \mathbf{J}_{E,K} \mathbf{J}_K^{-1} \mathbf{J}_{K,E'} \right) \delta p_{E'} = \sum_{K:E \in \partial K} \mathbf{J}_{E,K} \mathbf{J}_K^{-1} \mathbf{F}|_K - F_E \,, \tag{19}$$



Obrázek 1: Struktura soustavy lineárních algebraických rovnic v NRM.

kde $E = 1, \ldots, N_E$, $\mathbf{F}|_K = [F_{K,1}, \ldots, F_{K,N_C}]^T$. Jakmile máme spočteny δp_E, můžeme vyčíslit korekce koncentrací $\Delta\mathbf{c}|_K = [\delta c_1|_K, \ldots, \delta c_{N_C}|_K]^T$ na každém elementu K řešením

$$\Delta\mathbf{c}|_K = -\mathbf{J}_K^{-1}\left(\mathbf{F}|_K + \sum_{E' \in \partial K} \delta p_{E'}\, \mathbf{J}_{K,E'}\right), \tag{20}$$

pro $K = 1, \ldots, N_K$. Korekce δp_E, $\delta c_i|_K$ pro všechny hrany, složky a elementy spočtené v každé iteraci NRM jsou přičteny k hodnotám p_E^{n+1}, $c_i|_K^{n+1}$ z předchozí iterace. Iterační procedura končí při splnění podmínky

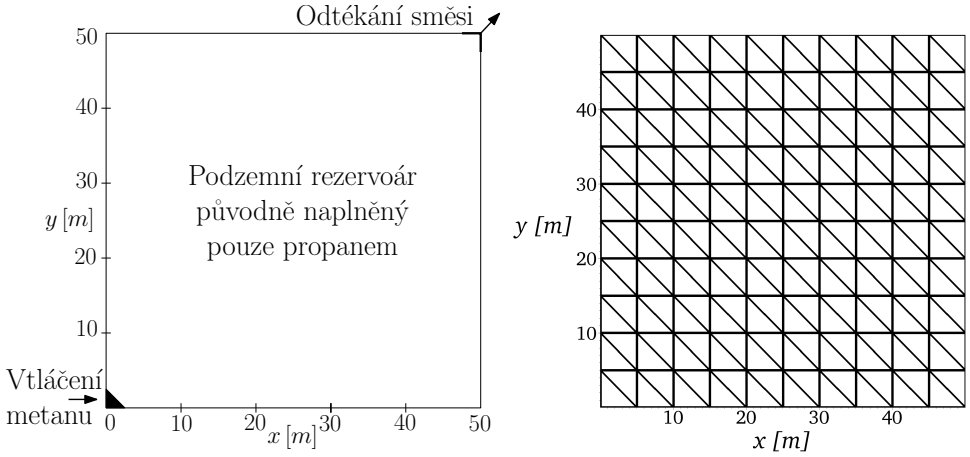
$$\left\| [\delta c_1|_1, \ldots, \delta c_{N_C}|_{N_K}, \delta p_1, \ldots, \delta p_{N_E}]^T \right\| < \varepsilon \tag{21}$$

pro zvolené $\varepsilon > 0$ (viz [18]).

4 Numerické výsledky

Uvažujme 2D čtvercovou oblast $50 \times 50\,\mathrm{m}^2$ reprezentující řez propanovým rezervoárem o porozitě $\phi = 0.2$ a izotropní permeabilitě $\mathbf{K} = k = 10^{-14}\,\mathrm{m}^2$ při počátečním tlaku $p = 5 \cdot 10^6\,\mathrm{Pa}$ a teplotě $T = 397\,\mathrm{K}$. V levém dolním rohu rezervoáru je vtláčen metan a v pravém horním rohu směs metanu a propanu odtéká (obr. 2). Hodnota vtláčení $f_1|_K$ je $2.643 \cdot 10^{-2}/|K|$ $\mathrm{mol\,m^{-3}s^{-1}}$, kde $|K|$ je obsah rohového trojúhelníku výpočetní sítě. Fyzikálně-chemické vlastnosti směsi jsou shrnuty v tab. 1. Při všech výpočtech jsou počáteční hodnoty voleny tak, aby ve směsi nedocházelo k fázovým přechodům. Hranice

oblasti je nepropustná kromě odtokového rohu, kde je udržován tlak $p = 5 \cdot 10^6\,\mathrm{Pa}$. Struktura výpočetní sítě o $2 \times m \times m$ elementech je zobrazena na obr. 2 (pro $m = 10$). Parametr ε z konvergenčního kritéria NRM (21) byl zvolen 10^{-6}. Soustava rovnic (19) byla řešena pomocí knihovny UMFPACK [6, 7, 8, 9].



Obrázek 2: Schéma simulovaného rezervoáru a struktura výpočetní sítě.

i (složka směsi)	p_{ci} [Pa]	T_{ci} [K]	V_{ci} [$\mathrm{m^3 mol^{-1}}$]
1 (CH_4)	$4.58373 \cdot 10^6$	$1.89743 \cdot 10^2$	$9.897054 \cdot 10^{-5}$
2 (C_3H_8)	$4.248 \cdot 10^6$	$3.6983 \cdot 10^2$	$2.000001 \cdot 10^{-4}$

i (složka směsi)	M_i [$\mathrm{kg\,mol^{-1}}$]	ω_i [-]	δ_{i1} [-]	δ_{i2} [-]
1 (CH_4)	$1.62077 \cdot 10^{-2}$	$1.14272 \cdot 10^{-2}$	0	0.0365
2 (C_3H_8)	$4.40962 \cdot 10^{-2}$	$1.53 \cdot 10^{-1}$	0.0365	0

Tabulka 1: Příslušné parametry PR EOS pro metan CH_4 a propan C_3H_8.

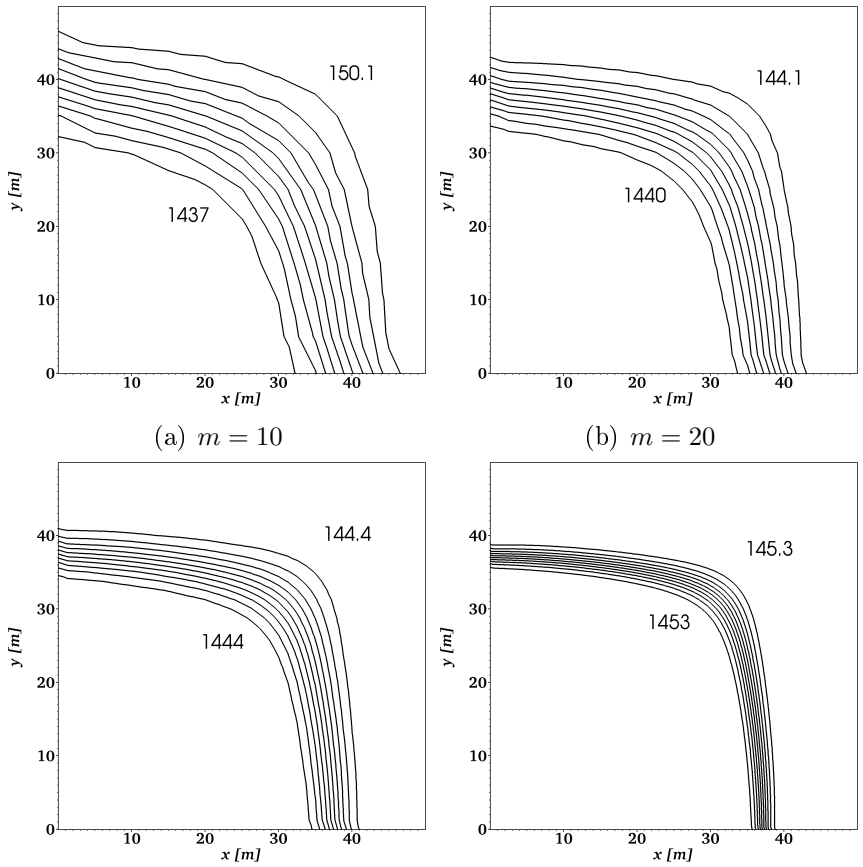
4.1 Konvergenční analýza

V této části ověříme konvergenci numerického schématu odvozeného v sekci 3 pomocí pseudoanalytického řešení – numerického řešení spočteného na nejjemnější síti $m = 160$ ($2 \times 160 \times 160$ prvků). Experimentální řád konvergence (EOC) budeme počítat mezi sítěmi $m = 10$, $m = 20$ a $m = 40$ s použitím L^1, L^2 a L^∞ konzistentních norem pro chyby E_m ve srovnání s řešením ze sítě $m = 160$. Chyba je vždy spočtena na nejjemnější síti projekcí řešení ze sítě hrubší a následnou lineární interpolací. Časový krok pro řešení s parametrem $m = 160$ je zvolen konstantní $\Delta t = 750\,\mathrm{s}$. Pro řešení na hrubších sítích je Δt čtyřikrát větší s každým zjemněním sítě ($\Delta t \sim m^{-2}$), tj. $\Delta t = 12000\,\mathrm{s}$ pro $m = 40$, $\Delta t = 48000\,\mathrm{s}$ pro $m = 20$, $\Delta t = 192000\,\mathrm{s}$ pro $m = 10$. EOC v normě $\|.\|_\nu$ spočteme jako

$$\mathrm{EOC}_\nu = \frac{\ln \|E_{m_1}\|_\nu - \ln \|E_{m_2}\|_\nu}{\ln m_2 - \ln m_1},$$

kde E_{m_1}, resp. E_{m_2} jsou chyby numerických řešení na sítích s parametry m_1, resp. m_2.

Konvergenční analýza je provedena na problému vtláčení metanu do horizontálního propanem naplněného rezervoáru (tj. s nulovou gravitací). EOC a chyby pro situaci v čase $\tau = 6 \cdot 10^6$ s jsou zahrnuty v tab. 2. Další tab. 3 obsahuje data z času $\tau = 2.4 \cdot 10^7$ s. Porovnání řešení na jednotlivých sítích v tomto čase je zobrazeno na obr. 3. Obdobně lze provést konvergenční analýzu na úloze s vertikální oblastí (s gravitací), která je zde ovšem pro nedostatek prostoru vynechána.



(a) $m = 10$ (b) $m = 20$

(c) $m = 40$ (d) $m = 160$

Obrázek 3: Koncentrace metanu c_1, $\tau = 2.4 \cdot 10^7$ s na různých sítích pro tab. 3. Izočáry jsou rozloženy rovnoměrně mezi dvěma zobrazenými hodnotami.

Grid (m)	$\|E_m\|_1$	EOC_1	$\|E_m\|_2$	EOC_2
10	$1.1025 \cdot 10^5$	0.6223	$6.5336 \cdot 10^3$	0.5086
20	$7.1621 \cdot 10^4$	0.8179	$4.5922 \cdot 10^3$	0.6814
40	$4.0627 \cdot 10^4$		$2.8635 \cdot 10^3$	

Grid (m)	$\|E_m\|_\infty$	EOC_∞
10	$1.1204 \cdot 10^3$	0.5077
20	$7.8804 \cdot 10^2$	0.5298
40	$5.4584 \cdot 10^2$	

Tabulka 2: Experimentální řády konvergence a chyby koncentrace c_1, $g = 0$, v čase $\tau = 6 \cdot 10^6$ s ve srovnání s numerickým řešením na síti $m = 160$ ($2 \times m \times m$ prvků) a časovým krokem $\Delta t = 750$ s. Na hrubších sítích je $\Delta t \sim m^{-2}$.

Grid (m)	$\|E_m\|_1$	EOC_1	$\|E_m\|_2$	EOC_2
10	$3.4079 \cdot 10^5$	0.6514	$1.109 \cdot 10^4$	0.5273
20	$2.1697 \cdot 10^5$	0.833	$7.6948 \cdot 10^3$	0.6814
40	$1.218 \cdot 10^5$		$4.7982 \cdot 10^3$	

Grid (m)	$\|E_m\|_\infty$	EOC_∞
10	$1.0485 \cdot 10^3$	0.584
20	$6.9948 \cdot 10^2$	0.4748
40	$5.0333 \cdot 10^2$	

Tabulka 3: Experimentální řády konvergence a chyby koncentrace c_1, $g = 0$, v čase $\tau = 2.4 \cdot 10^7$ s ve srovnání s numerickým řešením na síti $m = 160$ ($2 \times m \times m$ prvků) a časovým krokem $\Delta t = 750$ s. Na hrubších sítích je $\Delta t \sim m^{-2}$.

5 Závěr

V této práci jsme popsali numerické schéma pro řešení jednofázového stlačitelného proudění směsi v porézním prostředí založené na kombinaci MHFEM a FVM. Oproti tradičním přístupům je tlak spočten přímo ze stavové rovnice. Soustava nelineárních algebraických rovnic získaná kombinací MHFEM a FVM při použití Eulerovy metody je linearizována Newtonovou-Raphsonovou metodou. Rozměry získané soustavy lineárních algebraických rovnic závisí na počtu komponent směsi. Proto je navržena technika, která významně redukuje velikost soustavy tak, že již nezávisí na počtu složek směsi. Výpočetní složitost je tedy srovnatelná s klasickými sekvenčními přístupy. Výsledné schéma poskytuje přesnou lokální bilanci hmoty, která je důležitá zejména při řešení problémů v heterogenním prostředí. Numerický model jsme testovali na simulaci dvousložkové směsi – metan, propan. Konvergence numerického schématu byla ověřena vyčíslením experimentálních řádů konvergence na úloze vtláčení metanu do horizontálního propanem naplněného rezervoáru. V další práci bychom rádi vylepšili stávající model použitím MHFEM vyššího řádu a následně rozšířili o možnost simulace vícefázového proudění s přestupem komponent mezi fázemi.

Literatura

[1] G. Acs, S. Doleschall, E. Farkas. *General Purpose Compositional Model.* Society of Petroleum Engineers Journal, Vol.: 25, Issue: 4 (1985), 543–553.

[2] J. Bear, A. Verruijt. *Modeling Groundwater Flow and Pollution.* D. Reidel Publishing Company, Dordrecht, Holland (1987).

[3] F. Brezzi, M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, New York Inc. (1991).

[4] G. Chavent, J. E. Roberts. *A unified physical presentation of mixed, mixed-hybrid finite elements and standard finite difference approximations for the determination of velocities in waterflow problems.* Advances in Water Resources, 14(6) (1991).

[5] Z. Chen, G. Ma Y. Huan. *Computational Methods for Multiphase Flows in Porous Media.* SIAM, Philadelphia (2006).

[6] T. A. Davis. *A column pre-ordering strategy for the unsymmetric-pattern multifrontal method.* ACM Transactions on Mathematical Software, vol 30, no. 2 (2004), pp. 165–195.

[7] T. A. Davis. *Algorithm 832: UMFPACK, an unsymmetric-pattern multifrontal method.* ACM Transactions on Mathematical Software, vol 30, no. 2 (2004), pp. 196–199.

[8] T. A. Davis and I. S. Duff. *A combined unifrontal/multifrontal method for unsymmetric sparse matrices.* ACM Transactions on Mathematical Software, vol. 25, no. 1 (1999), pp. 1–19.

[9] T. A. Davis and I. S. Duff. *An unsymmetric-pattern multifrontal method for sparse LU factorization.* SIAM Journal on Matrix Analysis and Applications, vol 18, no. 1 (1997), pp. 140–158.

[10] E. O. Holzbecher. *Modeling Density-Driven Flow in Porous Media: Principles, Numerics, Software.* Springer-Verlag, Berlin (1998).

[11] H. Hoteit, A. Firoozabadi. *Multicomponent Fluid Flow by Discontinuous Galerkin and Mixed Methods in Unfractured and Fractured Media.* Water Resources Research (2005), 41, W11412, doi:10.1029/2005WR004339.

[12] P. S. Huyakorn, G. F. Pinder. *Computational Methods in Subsurface Flow.* Academic Press, Inc., New York (1983).

[13] R. J. Leveque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge University Press, Cambridge (2002).

[14] J. Lohrenz, B. G. Bray, C. R. Clark. *Calculating Viscosities of Reservoir Fluids From Their Compositions.* Journal of Petroleum Technology, Oct. (1964), 1171–1176.

[15] J. Maryška, M. Rozložník, M. Tůma. *Mixed-hybrid finite element approximation of the potential fluid flow problem.* Journal of Computational and Applied Mathematics, 63 (1995), 383–392.

[16] J. Mikyška, A. Firoozabadi. *Implementation of higher-order methods for robust and efficient compositional simulation.* Journal of Computational Physics, 229 (2010), 2898–2913.

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

[18] A. Quarteroni, R. Sacco, F. Saleri. *Numerical Mathematics.* Springer-Verlag, New York (2000).

[19] T. F. Russel, M. F. Wheeler. *Finite Element and Finite Difference Methods for Continuous Flows in Porous Media.* The Mathematics of Reservoir Simulation, Frontiers in Applied Mathematics, SIAM, Philadelphia (1983), 35–106.

[20] L. C. Young, R. E. Stephenson. *A Generalized Compositional Approach for Reservoir Simulation.* Society of Petroleum Engineers Journal, Vol.: 23, Issue: 5 (1983), 727–742.

Optická implementace kvantových procházek*

Václav Potoček

3. ročník PGS, email: vaclav.potocek@fjfi.cvut.cz
Katedra fyziky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Igor Jex, Katedra fyziky,
Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. Quantum walks represent an important research topic within the scope of quantum algorithms. Our workgroup has been focusing on the possibilities of a powerful compact interferometric layout which makes it possible to implement and observe quantum walks experimentally. The research is carried out in an international collaboration, in which all the theoretical results are accompanied by an actual experimental realisation. This extended abstract gives an introduction to the topic and summarises the main present results, emphasising the results and ongoing research questions covered during the last year.

Keywords: Quantum Walks, Optical Implementations, Statistical Physics

Abstrakt. Kvantové procházky jsou jednou z významných oblastí výzkumu v tématu kvantových algoritmů. Naše pracovní skupina se dlouhodobě zabývá studiem možností kompaktního interferometrického schématu, které umožňuje jednoduché kvantové procházky experimentálně implementovat a pozorovat. Výzkum probíhá v mezinárodní spolupráci, v jejímž rámci jsou všechny výsledky experimentálně realizovány. Tento rozšířený abstrakt poskytuje úvod do problematiky a shrnutí dosavadních výsledků s důrazem na nové teoretické výsledky a oblasti výzkumu pokryté po dobu posledního akademického roku.

Klíčová slova: Kvantové procházky, Optické implementace, Statistická fyzika

1 Úvod

Kvantové procházky se od původní myšlenky publikované v roce 1993 rozvinuly v široce diskutovanou oblast výzkumu mezi vědeckými komunitami zabývajícími se především kvantovými algoritmy a kvantovým zpracováním informace. Model kvantové procházky však našel významná uplatnění i v jiných, vzdálenějších oblastech, například v popisu systémů podléhajících náhodným vlivům okolí, v simulaci Andersonovy lokalizace či dokonce jako kandidát na popis překvapivých fyzikálně chemických vlastností některých významných biologických procesů.

S rozvojem významu kvantových procházek jako možného základního nástroje pro návrh kvantových algoritmů, stejně jako nástroje pro snadnou simulaci složitějších fyzikálních procesů, vyvstává otázka možností přímé či nepřímé experimentální realizace tohoto konceptu. Správná funkce kvantového algoritmu založeného na kvantovém procházení či přesnost simulace pomocí něj modelované jsou vitálně závislé na jednoduchosti, spolehlivosti a škálovatelnosti této realizace.

*Tato práce byla podpořena grantem SGS10/294/OHK4/3T/14

Výzkumná skupina kvantových procházek na Katedře fyziky FJFI se od roku 2008 účastní mezinárodní spolupráce, jejímž cílem bylo takový experiment navrhnout a realizovat, s důrazem na identifikaci, proměření a v nejlepším případě eliminaci možných zdrojů chyb, které výsledek degradují. Základní navržené schéma, úspěšně realizované v roce 2009, se vyznačuje vysokou flexibilitou i odolností vůči vlivům prostředí za zachování nízké časové i cenové náročnosti. Po tomto úspěchu se skupina zabývá rozšiřováním možností tohoto schématu nad rámec základní jednorozměrné kvantové procházky a dalším využitím, které se jeho prostřednictvím nabízejí.

Tento příspěvek je členěna následujícím způsobem. Sekce 2 poskytuje motivaci ke studiu kvantových procházek a stručný úvod do matematické definice tohoto modelu. Sekce 3 popisuje základní experiment, na kterém jsou založeny i další oblasti výzkumu v rámci spolupráce. Sekce 4 shrnuje nové možnosti, které původní experiment nabízí, a výsledky takto získané. Závěrečná sekce příspěvek shrnuje a poskytuje náhled na aktuální otevřené otázky podléhající probíhajícímu výzkumu.

2 Kvantové procházky

Popularita teoretického studia kvantových procházek těží z překvapivě širokého spektra aplikací klasických náhodných procházek, které sloužily jako předloha k hledání tohoto analogického principu podléhajícímu zákonům kvantové mechaniky. Za pár příkladů oblastí využití klasických náhodných procházek jmenujme statistickou fyziku (modelování difuze, termalizace, Brownova pohybu apod.), ekonomii, kombinatoriku a statistiku, zpracování obrazu a mnoho dalších. Náhodné procházky hrají také roli významného nástroje v algoritmizaci, kde se používají pro návrh nedeterministických algoritmů založených na vzorkování cest v datových strukturách, algoritmů pro odhad objemu či odhad globálních vlastností rozsáhlých grafů. Právě teoretická informatika, skřížená s nedávnou vlnou zájmu o výzkum kvantových počítačů a algoritmů pro ně, iniciovala myšlenku, zda kvantová verze náhodných procházek může nabídnout podobně univerzální uplatnění pro návrh kvantových algoritmů. Současné využití takto objevené teorie však, podobně jako v klasickém případě, široce přesáhlo původní očekávání.

Nejjednodušší model kvantové procházky získáme, hledáme-li analogii klasické náhodné procházky probíhající v diskrétních časových krocích po přímce, s pevnou délkou každého kroku. V klasickém schématu bychom polohu chodce popisovali celočíselnou veličinou podléhající Markovovu procesu, při kterém by se v každém kroku mohla zvýšit či snížit o jedničku. Pro jednoduchost můžeme dále předpokládat, že pravděpodobnosti obou možných přechodů (které budeme pracovně nazývat krokem doprava, resp. doleva), nezávisí na aktuální poloze a nemění se v čase. Pro pravděpodobnost výskytu náhodného chodce na poloze $x \in \mathbb{Z}$ v čase $t \in \mathbb{N}_0$, $P(x,t)$, tak získáváme rekurentní relaci

$$P(x,t) = p_+ P(x-1,t-1) + p_- P(x+1,t-1), \tag{1}$$

kde p_+ a p_- jsou nezáporné konstanty se součtem 1 a $t > 0$. Z možných počátečních podmínek se často předpokládá chodec lokalizovaný na jednom místě, díky translační invarianci v proměnné x postačí uvažovat $P(x,0) = \delta_{x,0}$, s čímž rovnice (1) vede na binomické rozdělení pravděpodobnosti v každém čase $t \geq 0$.

Při návrhu kvantového systému, který by představoval analogii takovéhoto procesu, vycházíme z předpokladu, že pravděpodobnostní rozdělení v každém okamžiku bude podloženo vlnovou funkcí definovanou na $\mathbb{Z}$. Diskrétní pohybová rovnice systému by pak měla mít tvar, dle kterého hodnota vlnové funkce v čase $t > 0$ a poloze x závisí pouze na hodnotách v čase $t-1$ a polohách $x \pm 1$, a být invariantní vůči posunu v poloze i čase. Kvantová mechanika dále vymezuje, že transformace vlnové funkce odpovídající přechodu z času $t-1$ do času t musí být v libovolném uzavřeném systému popsána unitárním operátorem.[1] Ukazuje se, že tato podmínka je natolik silná, že vylučuje jakoukoli dynamiku mimo triviálních případů (deterministický jednosměrný drift), dokud vlnová funkce je skalární.

Tuto významnou překážku snadno překonáme, jestliže v předpokladech umožníme, že částice má kromě polohy alespoň jeden další vnitřní stupeň volnosti. Pohybová rovnice, reprezentovaná předpisem časového vývoje během jednoho diskrétního kroku, pak může vždy být rozepsána jako složení dvou operátorů: rotace čistě v prostoru vnitřních stupňů volnosti následovaná posunutím, které na různé bázové vektory působí v odlišných směrech, v závislosti na tomto vnitřním stavu částice.

Konkrétněji, základní jednorozměrná kvantová procházka na přímce s diskrétními kroky o velikostech $+1$ a -1 je definována na stavovém prostoru daném tenzorovým součinem prostoru polohy $\ell^2(\mathbb{Z})$ a prostoru jednoho dvoustavového vnitřního stupně volnosti (dále jen „prostoru mince“), $\mathbb{C}^2$. Jestliže v prvním případě uvažujeme ortonormální bázi ketů označených celými čísly a ve druhém ortonormální bázi $\{|+\rangle, |-\rangle\}$, můžeme stavový prostor $\mathcal{H}$ také zapsat jako lineární obal tenzorového součinu těchto bází,

$$\mathcal{H} = \ell^2(\mathbb{Z}) \otimes \mathbb{C}^2 = \mathrm{Span}\{|n, d\rangle \mid n \in \mathbb{Z}, d \in \{+, -\}\}.$$

Časový vývoj okamžitého stavu $|\psi(t)\rangle \in \mathcal{H}$ je pak popsán rovnicí

$$|\psi(t+1)\rangle = U|\psi(t)\rangle, \tag{2}$$

kde propagátor U budeme uvažovat ve tvaru kompozice

$$U \ = S \cdot (I_{\ell^2(\mathbb{Z})} \otimes C), \tag{3}$$

operátor „hodu mincí“ C je libovolný unitární operátor na $\mathbb{C}^2$ a rovněž unitární operátor „kroku“ S definujeme předpisem akce na bázových stavech jako

$$S|n, \pm\rangle = |n \pm 1, \pm\rangle.$$

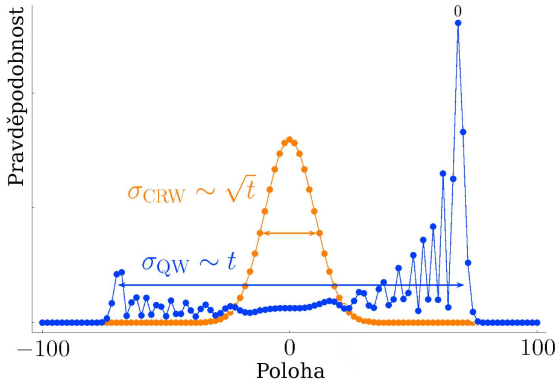
Tato definice vyjadřuje podmínku změny polohy o $+1$ nebo -1 v každém kroku, v závislosti na vnitřním stavu kvantové částice.

Při stanovené počáteční podmínce na tvar vlnové funkce $|\psi(0)\rangle$, platný v čase $t = 0$, pak snadno formálně dojdeme k obecnému řešení tvaru

$$|\psi(t)\rangle = U^t|\psi(0)\rangle.$$

V případě volby $|\psi(0)\rangle = \alpha|0, +\rangle + \beta|0, -\rangle$, vyjadřující chodce lokalizovaného v poloze 0, získané pravděpodobnostní rozdělení vykazuje zcela odlišné chování než binomické

[1] Poznamenejme, že časový vývoj kvantového uzavřeného systému je vždy deterministický. Kvantová procházka tedy nikdy nebude *náhodná*.



Obrázek 1: Typické pravděpodobnostní rozdělení pro náhodnou (světlá barva) a kvantovou (tmavá barva) procházku po 100 krocích. Asymetrie pravděpodobnostního rozdělení kvantové procházky je důsledkem silného vlivu počátečních podmínek.

rozdělení charakteristické pro klasické difuzivní systémy. Díky interferenci vlnové funkce dojde k jevu, že částice nebude setrvávat v blízkosti své počáteční polohy, ale postupně utvoří dvě vlny, jejichž čela se od této polohy vzdalují konstantní rychlostí (viz Obr. 1).

Důsledkem tohoto chování pravděpodobnostního rozdělení, typického pro kvantové procházky, je kvadraticky prudší růst variance v čase s okamžitými důsledky pro čas dosažení vrcholu v dané vzdálenosti, čas dosažení rovnoměrného rozdělení pravděpodobnosti na intervalu hodnot apod. Tyto vlastnosti jsou pak obzvláště důležité pro algoritmické využití procházky.

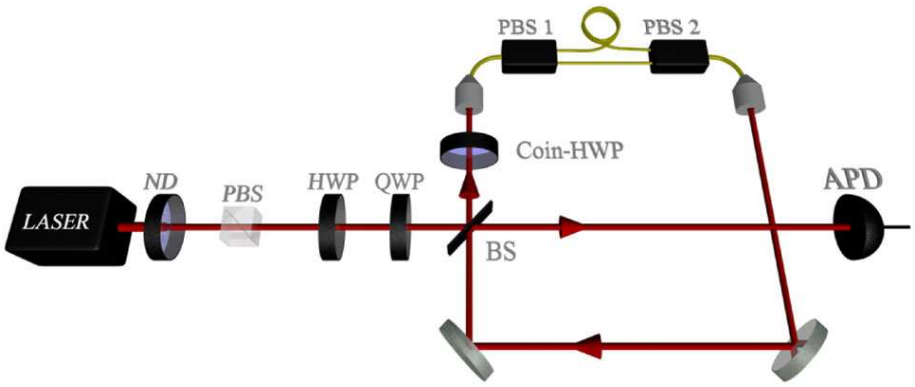
Podobným myšlenkovým postupem jako výše můžeme najít kvantové varianty náhodných procházek na složitějších podkladových prostorech než $\mathbb{Z}$, náhodných procházek s více možnými délkami kroku (či přestávkami) a procházek na spojitých prostředích, procházek spojitých v čase atd. Po zbytek tohoto příspěvku však zůstaneme u procházek diskrétních v čase na rovnoměrné mřížce.

3 Optická implementace lineární kvantové procházky

Základní zdroj [1] popisuje experiment, který byl navržen ve spolupráci pracovišť Heriot-Watt University (Edinburgh, Velká Británie), Max Planck Institute for the Science of Light (Erlangen, Německo) a FJFI ČVUT. Samotný experiment byl realizován v optických laboratořích Max Planck Institute. Schéma experimentu znázorňuje Obr. 2.

Základní myšlenkou kompaktní interferometrické realizace, kde počet optických elementů nebude růst s maximálním požadovaným počtem kroků náhodné procházky, je simulace časového vývoje převedením celé historie vlnové funkce na jednu společnou časovou osu. To je umožněno skutečností, že poloha i počet kroků vykonaných chodcem jsou celočíselné veličiny. Jestliže zvolíme dvě algebraicky nezávislé nezáporné reálné veličiny τ_1 a τ_2, pak celou množinu dvojic (čas, poloha) můžeme injektivně vnořit do jedné reálné osy pomocí zobrazení

$$i : \mathbb{Z} \times \mathbb{N}_0 : (x, t) \mapsto x\tau_1 + t\tau_2. \tag{4}$$



Obrázek 2: Schéma experimentálního rozložení (převzato z [1]). Podstatné zkratky: QWP – čtvrtvlnná destička, HWP – půlvlnná destička, BS – polopropustné zrcadlo, PBS – polarizující dělič svazku, APD – detektor (lavinová fotodioda).

Takto získaný čas budeme chtít interpretovat jako čas, kdy bude částice podléhající procházce naměřena na výstupu z celého zaměření. Jeden krok kvantové procházky, popsaný rovnicí (2), pak můžeme realizovat následujícím způsobem:

1. fyzikální částici podléhající procházce nejprve necháme projít prostředím, které způsobí změnu jejího vnitřního stavu,

2. necháme částici rozdělit svou dráhu v závislosti na vnitřním stavu,

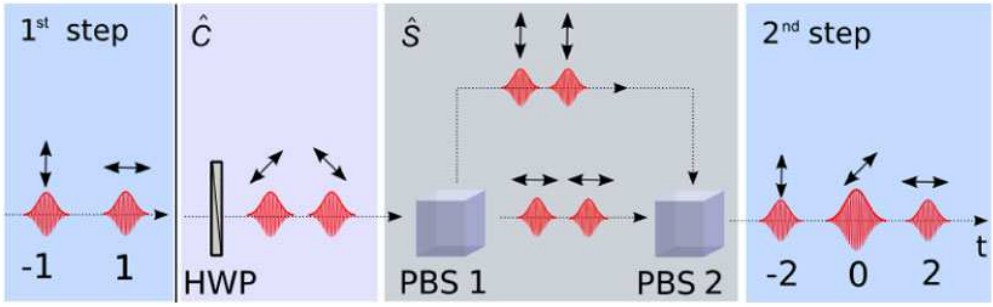
3. ve dvou větvích způsobíme časové zpoždění $\tau_2 + \tau_1$, resp. $\tau_2 - \tau_1$, a rekombinujeme je.[2]

Všechny tyto kroky můžeme snadno fyzikálně realizovat, konkrétní implementace závisí na druhu částice a zvolené vnitřní veličině, kterou použijeme.

V případě optické realizace je chodec reprezentován jednotlivým fotonem a jako vnitřní stav je použita polarizace. Jednotlivé kroky pak můžeme realizovat takto: Obecnou unitární transformaci je vždy možno sestavit ze čtvrtvlnné, půlvlnné a čtvrtvlnné destičky (v tomto pořadí) při vhodné volbě natočení optických os. K rozdělení dráhy fotonu v závislosti na jeho polarizaci ve zvolené bázi slouží polarizující dělič svazku, který jeden polarizační stav propouští a jeho ortogonální stav odráží. Totožný element můžeme použít pro rekombinaci, jestliže se polarizační stav po dobu separace nezmění. Posledním zbývajícím krokem je pouze zařazení vhodné zpožďovací linky mezi dvě takto separované trajektorie. Průchod dvou vlnových balíků (které samy jsou výsledkem podobného rozdělení v předchozím kroku) takovou sekvencí je znázorněn na Obr. 3.

Interferometr triviálně realizující t kroků procházky tímto způsobem by obsahoval t kopií této atomární sestavy zařazených za sebou. Po vyzáření jednoho fotonu do vstupního ramene a jeho průchodem celou sestavou by se jeho vlnový balík rozdělil na t pulzů oddělených časovým rozdílem $2\tau_1$, jejichž amplitudy přesně odpovídají hodnotám vlnové funkce $|\psi(t)\rangle$ v t možných polohách chodce, které tvoří její nosič.

[2]Typicky volíme $\tau_1 \ll \tau_2$. Výstupní časy odpovídající stejnému t, ale různému x, tak vystupují ve výrazně oddělených shlucích.



Obrázek 3: Transformace vnitřního stavu na půlvlnné destičce, rozdělení polarizací na polarizujícím děliči svazku a jejich rekombinace s časovým posunem odpovídajícím původnímu rozdílu obou pulzů. Zdroj: [1]

Schéma můžeme výrazně zjednodušit použitím zpětnovazební smyčky. Uzavřením interferometru sama do sebe dosáhneme toho, že sestavu realizující jeden krok procházky postačí implementovat jednou; stejný foton jí projde několikrát, než je z interferometru vypuštěn nebo jej náhodně opustí v případě pasivního uzavření smyčky polopropustným zrcadlem. Toto uzavírá popis schématu, které bylo přesně realizováno v práci [1].

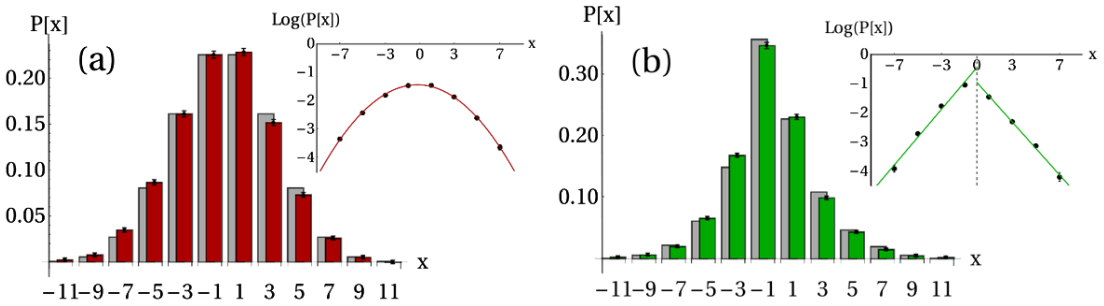
Experimentální výsledky potvrzují vysokou nadějnost výše popsané soustavy pro případnou realizaci ve větším měřítku. Zdroj [1] cituje pouze 5 kroků procházky pozorovaných s vysokou přesností, ve vyšších iteracích byl již signál příliš slabý vůči šumu. Nicméně analýza přítomných zdrojů rušení, výměna optických elementů za kvalitnější ekvivalenty a optimalizace polopropustného zrcadla, které realizuje vstup a výstup signálu z hlavní smyčky, snadno zvýšily tento počet na 28 [2]. Se zachováním součástek a implementací aktivního uzavírání, resp. otevírání smyčky elektronickou modulací optických parametrů prostředí by stejný interferometr byl schopen měřit až 100 kroků kvantové procházky. Tato úprava nebyla dosud realizována.

4 Další využití schématu

Původní schéma, zobrazené na Obr. 2, realizuje čistý unitární vývoj kvantové procházky s diskrétním časovým krokem na přímce. Myšlenka zpětnovazební interferometrickou realizací však nabízí i širší využití.

Po předvedení základní funkčnosti a životaschopnosti v [1] se tým realizující projekt zabýval studiem stability systému a vlivu prostředí na chyby měření. Díky realizaci maximální části optické dráhy, včetně celé zpožďovací linky, pomocí eliptického optického vlákna zachovávajícího polarizaci se tyto externí vlivy ukázaly jako zcela zanedbatelné. Ztráta koherence měla na viditelnost výsledné vlnové funkce neměřitelný vliv oproti ztrátám intenzity, daným disipací energie při odrazech a přechodech mezi volným prostředím a optickým vláknem. Tyto vlivy jsou nevyhnutelné, jde je však vyvážit naměřením většího množství experimentálních dat prodloužením doby konání experimentu. Popisovaný systém má tedy koherentní vlastnosti, jimž nemůže žádná jiná známá realizace konkurovat.

Vysoká nezávislost na chybách prostředí dává možnost do systému vnést plně řízený



Obrázek 4: Efekt silné řízené dekoherence na pravděpodobnostní rozdělení v případě (a) dynamické a (b) statické chyby. Porovnání teoretické předpovědi (sloupce v pozadí) s experimentálně naměřenými údaji (v popředí). Převzato z [2].

zdroj chyb a sledovat tak vliv dekoherence v kontrolovaných podmínkách. To je hlavním tématem navazující práce [2]. V ní využíváme původní schéma, do nějž je přidán elektro-optický modulátor, způsobující náhodné fázové změny, za část obvodu realizující kvantovou minci. Efektivně tak zkoumáme kvantovou procházku, jejíž mince je závislá na čase, či, v případě velmi rychlých změn, jejíž evoluční operátor (3) je upraven tak, že na bázové stavy s různou polohou působí různý operátor mince.

V práci [2] zkoumáme dva hlavní druhy chyb: modulaci vlnové funkce náhodnou fázovou chybou závislou na poloze po každém provedeném kroce, kde náhodná fáze je volena nezávisle v každém kroce a každé poloze, či tato fáze jakožto funkce polohy je zvolena náhodně na začátku vývoje, ale pro jednotlivé kroky kvantové procházky zůstává konstantní. V obou případech je možno experimentálně řídit rozptyl vnesené chyby. Pro extrémní případ rovnoměrného rozdělení fáze na celém intervalu jsou výsledky předpovědi i experimentu vyneseny na Obr. 4.

První situace (nazývaná dynamická chyba), která představuje ztrátu fázové informace pod vlivem silné dekoherence, vede k úplné ztrátě kvantového chování procházky. Její pravděpodobnostní rozdělení tak při jinak stejných podmínkách přechází v závislosti na síle rušení od průběhu typického pro kvantovou procházku k binomickému rozdělení a z kvantové procházky se stává klasická náhodná procházka.

Druhá situace, statická chyba, modeluje průchod chodce náhodným, ale stabilním prostředím. V takovém případě pro získání průměrných charakteristik přes možné konfigurace prostředí počítáme střední pravděpodobnostní rozdělení. Výsledky numerických simulací pak přesně ve shodě s experimentem předpovídají překvapivé chování pravděpodobnostního rozdělení polohy chodce, která zůstává výrazně lokalizovaná v blízkosti počáteční polohy. Práce [2] má prvenství v experimentálním předvedení tohoto jevu tzv. exponenciální lokalizace v kvantových procházkách.

Další významné zobecnění, které bude implementováno v následujícím běhu experimentu, je zvýšení dimenze procházky. Podobně jako procházku na přímce můžeme definovat procházku např. na mřížce. V kvantové procházce je pak nutno nahradit stavový prostor polohy částice prostorem $\ell^2(\mathbb{Z}^2)$ a prostor mince $\mathbb{C}^4$ v souvislosti s faktem, že chodec má v každém kroku 4 možnosti přechodu na sousední pole. V prezentované experimentální implementaci je snadné druhý rozměr (případně vyšší rozměry) zahrnout

zavedením dalšího časového rozestupu τ_3 a zobecněním rovnice (4) na

$$i : \mathbb{Z}^2 \times \mathbb{N}_0 : (x, y, t) \mapsto x\tau_1 + y\tau_2 + t\tau_3.$$

Problém nastává při potřebě čtyřrozměrného vnitřního stavového prostoru částice, jelikož není možno využít pouze polarizaci. Teoretický výzkum i první experimentální výsledky však nabízejí možnost využít polarizaci fotonu spolu s jeho orbitálním momentem hybnosti. Tento výzkum aktuálně probíhá a jeho úspěšné uzavření je příštím důležitým cílem projektu.

5 Závěr

Představili jsme základní úvod do problematiky kvantových procházek a skutečnou experimentální realizaci, na jejímž teoretickém podkladu se podílela Katedra fyziky FJFI. Experimentální schéma představuje realizaci kvantové procházky, která je výjimečná opakovaným používáním jedné sady optických elementů v konfiguraci zpětnovazebné smyčky. Díky tomu dosahuje experiment velmi kvalitních výsledků a dobré shody s idealizovanou teoretickou předpovědí. Jednoduchá úprava experimentu umožňuje plně kontrolovanou injekci šumu a pozorování jeho důsledků. Tyto výsledky jsou k dispozici s plnými detaily v publikacích [1, 2].

Nejbližší vyhlídky pro pokračování v projektu jsou realizace více než jednorozměrné kvantové procházky, tj. procházky po mřížce. Konečným cílem výzkumu je pak experimentální pozorování průběhu jednoduchého kvantového algoritmu založeného na kvantovém procházení.

Poděkování

Tento výzkum probíhal ve spolupráci s institucemi Heriot-Watt University (Edinburgh, Velká Británie) a Max Planck Institute for the Science of Light (Erlangen, Německo). Na české straně byl mimo autora příspěvku pracovní tým tvořen Dr. A. Gábrisem a Prof. I. Jexem, různé části projektu byly podporovány granty GAČR 202/08/H078, MSM 6840770039 a MŠMT LC06002. Autor příspěvku je od roku 2010 navrhovatelem grantu SGS10/294/OHK4/3T/14. Autoři tímto vyjadřují díky všem spolupracujícím organizacím i agenturám umožňujícím konání výzkumu a mezinárodní spolupráci svou podporou.

Literatura

[1] A. Schreiber, K. N. Cassemiro, V. Potoček, A. Gábris, P. J. Mosley, E. Anderson, I. Jex, and Ch. Silberhorn. *Photons Walking the Line: A Quantum Walk with Adjustable Coin Operations*. Phys. Rev. Lett. **104**, 050502 (2010).

[2] A. Schreiber, K. N. Cassemiro, V. Potoček, A. Gábris, I. Jex, and Ch. Silberhorn. *Decoherence and Disorder in Quantum Walks: From Ballistic Spread to Localization*. Phys. Rev. Lett. **106**, 180403 (2011).

Datové modelování*

Anna Rývová

1. ročník PGS, email: anna.ryvova@gmail.com
Katedra softwarového inženýrství v ekonomii
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Vojtěch Merunka, Katedra softwarového inženýrství v ekonomii,
Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. The paper deals with the possibilities and data modeling tools to better understanding of the modeled reality, their advantages and disadvantages. Flow Charts, UML diagrams and the method Business Object Relation Modeling (BORM) are methods that allow analysis and modeling of information system from different perspectives and serves not only to mutual communication between developers, but also to communicate with clients.

Keywords: Flow charts, UML, BORM method

Abstrakt. V příspěvku se zabývám možnostmi a nástroji datového modelování k lepšímu poznání modelované reality, jejich výhodami a nevýhodami. Vývojové diagramy, UML diagramy i metoda BORM jsou metody, které umožňují analýzu a modelování informačních systémů z různých hledisek a slouží nejen ke vzájemné komunikaci mezi vývojáři, ale i ke komunikaci s klienty.

Klíčová slova: Vývojové diagramy, UML, metoda BORM

1 Úvod

V literatuře můžeme najít celou řadu definic pojmu datové modelování. Uvádím některé z nich:

Datové modelování je základní součástí analýzy každého softwarového projektu. Správný návrh datové struktury může do značné míry ovlivnit bezporuchovost, udržovatelnost a rozšiřitelnost výsledné aplikace. [4]

Podle Thierry Bruneta je základním principem datového modelování centrální a standardizovaný návrh (schéma) databáze. Bez tohoto schématu nemůže existovat žádná robustní architektura a tomuto schématu musí rozumět všichni, kdo na projektu datové architektury pracují - obchodníci, technici, obchodní uživatelé, datoví architekti, analytici, návrháři databáhí, projektoví manažeři, vývojáři i databázoví administrátoři. [14]

Scott Ambler ve své knize The Object Primer definuje datové modelování jako "Data modeling is the act of exploring data-oriented structures. Like other modeling artifacts data models can be used for a variety of purposes, from high-level conceptual models to physical data models." [19]

Wikipedia přináší stručnější definici: "Data modeling in software engineering is the process of creating a data model by applying formal data model descriptions using data modeling techniques." [13]

*Tato práce byla podpořena grantem SGS2011

Podle Merunky je datové modelování specifická část softwarového inženýrství, která nemá za cíl tvorbu programů ani obsluhu databázových systémů. Není to ani programování, ani pouhé kreslení diagramů a psaní manažerské dokumentace. Při datovém modelování se používá jen vybraná část programovacích jazyků (zápis a manipulace s daty), nepoužívají se knihovny softwarových komponent, používají se jen návrhové vzory pro popis dat, formální aparát je nástrojem pro popis a manipulaci s daty (výroková logika, operace s množinami,...), používají se pouze diagramy popisující vlastnosti dat a vztahy mezi nimi. [8]

Graficky můžeme modely vyjádřit např. pomocí vývojových diagramů, UML diagramů, nebo grafické notace BPMN (Business Process Management Notation). K analýze můžeme využít např. metody BORM (Business Object Relationship Modelling) a Six Sigma. V této práci se budu podrobněji věnovat vývojovým diagramům, UML diagramům a metodě BORM a možnostem jejich využití při modelování reality.

2 Vývojové diagramy

Algoritmus je přesný postup, který vede k vyřešení určitého výsledku. Pokud programu dáme určitá data, vrátí nám výsledek, pokud mu tatáž data zadáme znovu, výsledek bude totožný s předchozím. [2]

Vývojový diagram je druh diagramu, který slouží k grafickému znázornění jednotlivých kroků algoritmu nebo obecného procesu. Vývojový diagram používá pro znázornění jednotlivých kroků algoritmu pomocí symbolů, které jsou navzájem propojeny pomocí orientovaných šipek. Symboly reprezentují jednotlivé procesy, šipky tok řízení. Vývojové diagramy standardně nezobrazují tok dat, ten je zobrazován pomocí data flow diagramů. Vývojové diagramy jsou často využívány v informatice během programování pro analýzu, návrh, dokumentaci nebo řízení procesu. [21]

Alan B. Sterneckert (2003) navrhl, že by vývojové diagramy mohly být tvořeny z nezávislého pohledu jiné skupiny uživatelů (např manažery, systémovými analytiky a úředníky) a díky tomuto návrhu existují čtyři hlavní typy vývojových diagramů:

1. **Document flowcharts** — ukazují řízení toků dokumentů v systému.

2. **Data flowcharts** — ukazují řízení toků dat v systému.

3. **System flowcharts** — ukazují řízení toků fyzické vrstvy nebo vrstvy zdrojů.

4. **Program flowchart** — ukazují řízení toků v programu v rámci systému. Každý z druhů diagramů se zaměřuje spíše na řízení, než na konkrétní tok. [16]

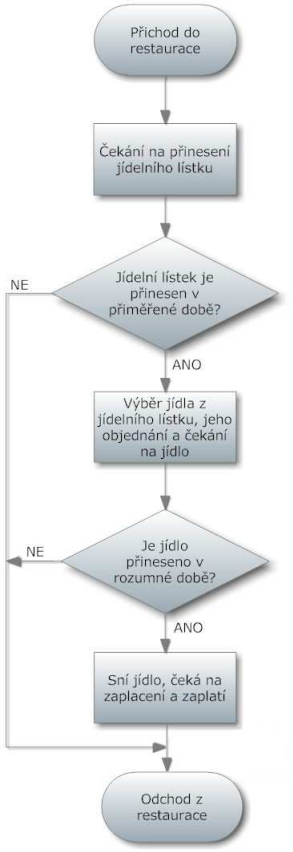
Tvorbu vývojových diagramů upravují ČSN 36 9030 "Značky vývojových diagramů pro systémy zpracování dat" z roku 1974 a ČSN ISO 5807 "Zpracování informací. Dokumentační symboly a konvence pro vývojové diagramy toku dat, programu a systému, síťové diagramy programu a diagramy zdrojů systému" z roku 1996.

Nevýhodou vývojových diagramů je pracnost a složitost konstrukce, větší diagramy se nevejdou na jednu stránku, což je činí méně přehlednými.

Postup při tvorbě vývojových diagramů: [6]

1. Co se stane nejdříve?
2. Co má následovat?
3. Co se děje rozhodne-li se ANO?
4. Co se děje rozhodne-li se NE?
5. Odkud přichází výrobek?
6. Kdo rozhoduje?

Pro tvorbu vývojových diagramů existuje celá řada softwarových nástrojů, např. MS Visio nebo SmartDraw, který umí kreslit nejen vývojové diagramy, ale i UML diagramy, myšlenkové mapy, organizační diagramy a celou řadu dalších diagramů.



Obrázek 1: Vývojový diagram objednání jídla v restauraci

3 UML

UML (Unified Modeling Language) je v softwarovém inženýrství grafický jazyk pro vizualizaci, specifikaci, navrhování a dokumentaci programových systémů. UML nabízí standardní způsob zápisu jak návrhů systému včetně konceptuálních prvků jako jsou business procesy a systémové funkce, tak konkrétních prvků jako jsou příkazy programovacího jazyka, databázová schémata a znovupoužitelné programové komponenty. UML podporuje objektově orientovaný přístup k analýze, návrhu a popisu programových systémů. UML neobsahuje způsob, jak se má používat, ani neobsahuje metodiku(y), jak analyzovat, specifikovat či navrhovat programové systémy. Standard UML definuje standardizační skupina Object Management Group (OMG). [20]

UML bylo přijato jako průmyslový standard ISO, v praxi se používá jak ke vzájemné komunikaci mezi vývojáři, tak ke komunikaci s klienty. Jazyk UML je sám definován pomocí modelu v UML – metamodel UML je zapsán pomocí diagramů tříd UML doplněných popisem sémantiky v přirozeném jazyce a formálním vyjádřením sémantiky v OCL.

UML pokrývá v podstatě celý vývojový cyklus informačního systému od sběru požadavků zákazníka až po nasazení. Analytik většinou není programátorem, nemusí být proto vhodné, aby navrhoval diagram tříd nebo API. Z tohoto důvodu je vhodné rozlišovat role analytika (definuje požadavky klienta) a architekta (navrhuje, jak požadavky realizovat). Tyto dvě role by měly být obsazeny dvěma lidmi.

Nevýhodou je, že v praxi jsou většinou modelované problémy velmi složité a podrobné zachycení všech detailů je z časových důvodů nemožné. Je proto nutné nalézt vhodný kompromis, určující míru detailů, které mají být zachyceny.

Elementy a vztahy mezi nimi:

1. Každý model je dokumentován sadou pohledů.
2. Model v UML je dokumentován sadou diagramů, dokumentujících určité rysy modelu.
3. Každý diagram je sestaven z jistých elementů a vztahů mezi nimi.
4. Obecně se v UML připouští, aby v každém modelu byl použít libovolný element.
5. Ne všechny kombinace jsou ale smysluplné, vždy určitá kombinace elementů a vztahů představuje superstrukturu diagramu jistého typu. Některé elementy a vztahy jsou použitelné obecně.

Standard ve verzi 2.0 se skládá z:

1. **UML 2.0 SuperStructure** - popis UML z hlediska uživatele (analytik/programátor) - popis jednotlivých diagramů.
2. **UML 2.0 Infrastructure** - metamodel specifikovaný pomocí Meta-Object Facility (MOF).

3. **UML 2.0 Object Constraint Language** (OCL) - jazyk pro specifikaci vstupních a výstupních podmínek, invariantů v diagramech.

4. **UML 2.0 Diagram Interchange** - popis XML struktur pro výměnu konkrétních modelů mezi jednotlivými modelovacími nástroji.

Nejznámější a nejpoužívanější částí standardu jsou diagramy, které můžeme dělit do několika skupin:

1. **strukturní diagramy**
   (a) diagram tříd
   (b) diagram komponent
   (c) diagram vnitřní struktury
   (d) diagram nasazení
   (e) diagram balíčků
   (f) diagram objektů

2. **diagramy chování**
   (a) diagram aktivit
   (b) diagram užití
   (c) stavový diagram

3. **diagramy interakce**
   (a) sekvenční diagram
   (b) diagram komunikace
   (c) diagram přehledu interakcí
   (d) diagram časování

Jakými diagramy začínáme?

1. **Funkční přístup** - diagram případů užití, model jednání.

2. **Datově orientovaný přístup** - diagram tříd.

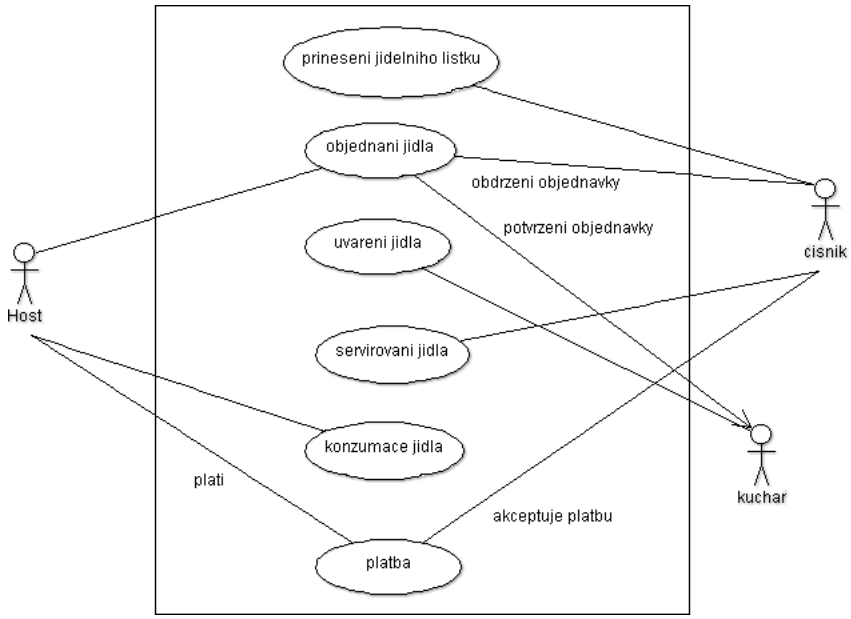
Pro tvorbu UML diagramů existuje celá řada softwarových nástrojů, např. Enterprise Architect, Rational Rose nebo free nástroje Umbrello UML Modeller a ArgoUML.

Podle Svačiny [18] programátor nebo softwarový architekt využívající UML pro modelování realizací případů užití musí dostatečně dobře znát jak model systému (aby mohl příslušné třídy/komponenty využívat opakovaně), tak cílové implementační prostředí (aby model měl hlavu a patu a byl implementovatelný).

Programátor nebo softwarový architekt musí vědět, jak danou funkčnost realizovat a v zásadě řeší především konkrétní detaily. Za tímto účelem musí vývojový tým sdílet společnou technickou vizi daného systému. Tato vize je realizována tzv. aplikačním

frameworkem. Framework definuje softwarovou architekturu požadoveného systému. V informačních systémech lze z pohledu architektury nalézt určité typické opakující se situace. Aplikační framework pak využije tzv. architektonické design patterns, ukazující typické řešení takovýchto situací. Příkladem architektonických design patterns mohou být systémové problémy, jako je autorizace, autentifikace, auditing, pooling a caching objektů, řízení transakcí apod., nebo typické situace z pohledu uživatele, např. práce s formulářem, průvodce, obsluha jednoduchého číselníku nebo práce se sestavami. Takovéto design patterns je vhodné vyhledat, popsat a opakovaně používat.

Klíčovou otázkou z pohledu modelování potom je, zda je třeba tyto konkrétní situace odpovídající nějakému design pattern znovu modelovat nebo "jen" odkázat na existující známé řešení. V takovéto situaci se jedná především o ekonomické rozhodnutí související s rozporem mezi náklady na projekt a touhou po co nejpodrobnější a nejpřesnější dokumentaci softwarového řešení. [18]



Obrázek 2: Diagram případů užití objednávání jídla v restauraci

4 BORM metoda

BORM (Business Object Relation Modeling) je objektově orientovaná metoda softwarového inženýrství třetí generace, která je velmi efektivní při vývoji znalostních systémů. Efektivity je dosahováno pomocí jednoduchých metod pro prezentaci všech aspektů relevatního modelu. Metoda má široké využití při modelování business procesů.

Metoda je vyvíjena od r. 1993 v rámci mezinárodního výzkumného projektu, od r. 1996 je vývoj podporován firmou Deloitte&Touche, kde je metoda prakticky využívána nejen při tvorbě softwaru, ale i k analýze požadavků a modelování business procesů. Od počátku byla orientována na podporu tvorby objektově orientovaných softwarových systémů založených na čistě objektově orientovaných programovacích jazycích a vývo-

jových prostředích (např. Smalltalk - VisualWorks, VisualAge...) a objektové databáze (Gemstone...). [8]

Metoda BORM a především její možnosti analýzy v počátečních fázích vývoje projektu byla prakticky použita například v projektech pro pražské zdravotnictví, Ústav pro státní informační systém ČR, elektroenergetiku, zemědělství, telekomunikace a plynárenství. Ve všech techto projektech se ukázalo, že BORM lze dobre využívat jako nástroj pro provádění business process reengineeringu. Výsledky takové analýzy také velmi dobře slouží pro podrobnou a úplnou specifikaci zadání softwarového projektu. [8]

Metoda BORM je podporována různými CASE nástroji, např. MetaEdit+ nebo CRAFT.CASE.

Fáze životního cyklu podle BORM [8, 17]

1. **Strategická analýza** - definice problému, stanovení jeho rozhraní, rozpoznání základních procesů odehrávajících se v systému a jeho okolí.
2. **Úvodní analýza** - rozpracování problému, mapování procesů v systému a vlastností základních objektů.
3. **Podrobná analýza** - detailní rozpracování analýzy jednotlivých objektů, vazeb mezi nimi a jejich životních cyklů. Toto je poslední analytická fáze, na jejímž konci by vše mělo být rozpoznáno.
4. **Úvodní návrh** - první fáze, ve které se snažíme upravit systém pro softwarovou implementaci.
5. **Podrobný návrh** - dochází k přeměně prvků existujícího modelu do podoby podřízené cílovému implementačnímu prostředí. Zohledňují se vlastnosti konkrétních programových jazyků, databází apod.
6. **Implementace** - vlastní vytváření požadovaného software programováním nebo generováním z CASE nástroje.

Ve fázi analýzy je nejčastěji vytvářen seznam funkcí, tabulka scénářů, architektura na business úrovni (vztah funkcí a scénářů) a ORD diagramy. Současně vznikají seznamy účastníků a datových toků. V této fázi se vlastně provádí kompletní OBA (Object Behavioral Analysis).

Ve fázi návrhu jsou vytvářeny diagramy tříd a komponent, tabulka podsystémů, tabulka balíčků a architektura na implementační úrovni (vztah podsystémů a balíčků).

Odlišnosti metody BORM oproti ostatním metodám podle V. Merunky: [8]

1. Většina metod je založena na analýze textového popisu zadání a odvozování objektů a jejich operací z podstatných jmen a sloves ve větách. UML poskytuje malou podporu pro identifikaci objektů ze zadání. U všech diagramů se předpokládá, že objekty a třídy jsou již rozpoznány.
2. BORM pro každou jednotlivou fázi životního cyklu využívá v diagramech omezenou sadu pojmů - předpokládá se, že během projektování dochází k postupným přeměnám objektů na jiné. Např. pojmy jako stav, přechod nebo asociace jsou používány jen během analýzy, pojmy jako agregace nebo dědičnost se používají jen ve fázi implementace.

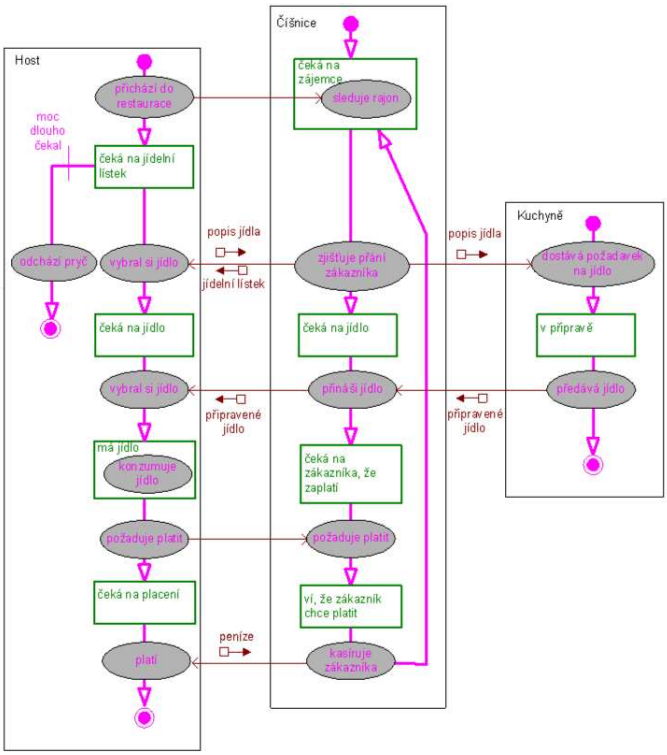
3. Nevyžaduje oddělování od sebe statických a dynamických pohledů na systém do různých typů diagramů s rozdílnou notací, je možno je v jednotlivých diagramech kombinovat.

Výhody metody BORM:

Metoda je založená na postupné transformaci modelu a v každé fázi se pracuje jen s určitou omezenou a konzistentní podnožinou BORM návrhu, což umožňuje její snadné osvojení analytiky, konzultanty i vývojáři. Je nadšeně přijímána programátory ve Smalltalku i v Javě stejně jako programátory objektových databází (Gemstone, ArtBase). BORM pracuje rovněž s hierarchií objektů (polymorfismus, is-a vztah, závislost objektů) [11].

Komplexnost metody BORM je i její nevýhodou, dnes nejrozšířenější software pro tuto metodu CRAFT.CASE je komerční a pracuje se s ním způsobem odlišným od většiny běžných CASE nástrojů.

Metoda BORM umožňuje v jednom grafu zachytit vývoj objektů účastnících se procesu, jejich stavy a akce, na kterých participují. Velké obdélníky jsou objekty účastnící se procesů, malé obdélníky stavy objektů, ovály představují aktivity objektů. Šipky mezi aktivitami představují komunikaci, která může obsahovat datový tok.



Obrázek 3: Diagram znázorňující vývoj objektů a vztahy mezi nim v rámci procesů objednávání jídla v restauraci

5 Závěr

Primárním cílem projektů informačních systémů je vytvoření požadovaného softwaru při dodržení kvality, kvantity, termínu a rozpočtu. Model jako takový není primárním cílem, ale jen podkladem sloužícím k vytvoření softwaru a ke vzájemné komunikaci mezi lidmi.

Informační systémy jsou často velmi komplexní, což klade vysoké nároky na analýzu. Proto je snaha o vylepšení úvodních fází vývoje softwaru. Modelovací nástroje musí sloužit nejen k vizualizaci kódu aplikace, ale musí co nejlépe umožňovat zachycení, analýzu a validaci požadavků uživatele.

Základem analýzy informačního systému je vytvoření správného datového modelu pro konkrétní aplikaci a databázový systém. CASE nástroje nám pak umožní z datového modelu vygenerovat SQL skripty pro vytvoření struktury databáze. CASE nástroje rovněž umožňují vytvoření diagramů a generování zdrojového kódu z modelu (resp. ze zdrojového kódu zpětné vytvoření modelu) a vygenerování dokumentace.

Všechny výše uvedené techniky modelování (vývojové diagramy, UML diagramy a metoda BORM) mají každá své výhody i nevýhody. Jejich vzájemnou kombinací můžeme lépe poznat modelovanou realitu ze všech možných hledisek a mají proto při vývoji informačního systému nezastupitelnou roli.

Literatura

[1] J. Arlow, I. Neustadt. *UML a unifikovaný proces vývoje aplikací. Průvodce analýzou a návrhem objektově orientovaného softwaru.* Computer Press, a. s. (2003) ISBN 80-7226-947-X.

[2] J. Chytil. *Vývojové diagramy - 1. díl.* Programujte.com (24. 7. 2005), http://programujte.com/clanek/2005080105-vyvojove-diagramy-1-dil/.

[3] P. Klobasa. *Na co se zapomíná v UML.* 4. května 2008, http://vyvojari.oxyonline.cz/uml-na-co-se-zapomina.

[4] R. P. KNOTT, V. MERUNKA, J. POLÁK. *The BORM Method: A Third Generation Object-Oriented Methodology.* In: Liu, L., Roussev B. (eds) Management of the Object-Oriented Development Process. pp. 337-360. IGI Publishing, [s.l.] (2006) ISBN 9781591406044, http://www.igi-global.com/viewtitlesample.aspx?id=25644.

[5] R. P. KNOTT, V. MERUNKA, J. POLÁK. *Principles of the Object-Oriented Paradigm - Basic Concepts, OO design methodologies, OOA & OOD.* In: Tutorial Book printed at EastEurOOP93 International Conference, Bratislava November 1993.

[6] R. Levay. *Vývojové diagramy.* ikvalita.cz, http://www.ikvalita.cz/tools.php?ID=25.

[7] V. Merunka, J. Polák. *Object-Oriented Analysis and Design - Teaching and Application Issues, in proceedings of ASU- Association of Simula Users & ACM.*(1993), Praha - Průhonice (září 1994), 22–36.

[8] V. Merunka, J. Polák. *BORM - Business Object Relation Modeling - Popis metody se zaměřením na úvodní fáze analýzy I.S.* konference TVORBA SOFTWARU '99, (Ostrava 26. - 28. 5. 1999), http://www.osu.cz/katedry/kip/aktuality/sbornik99/merunka2.html.

[9] V. Merunka, J. Polák, R. P. Knott, M. Buldra. *Object-Oriented Analysis and Design of IS of Agrarian Chamber Czech Republic in BORM.* In proceedings of 4Front Coordinators conference, Deloitte&Touche, Singapore (leden 1995).

[10] V. Merunka, J. Polák, J. Kofránek. *Úvod do metody BORM - Minikurz (Introduction into the BORM Method).* In the symposium of 5th Annual National Conference, "Objekty 2000", IEEE&ACM Czech Chapter, (Prague 2000), http://www.grada.cz/dokums_raw/usn/objekty2000.pdf.

[11] V. Merunka, J. Polák, L. Rivas. *BORM - Business Object Relation Modeling, in Proceedings of WOON.* In Fifth International Conference on Object-Oriented Programming, St. Petersburg (2001), http://www.grada.cz/dokums_raw/usn/woon2001.pdf.

[12] *OMG Unified Modeling LanguageTM (OMG UML).* Infrastructure (leden 2011), http://www.omg.org/spec/UML/2.4/Infrastructure/Beta2/PDF/.

[13] R. Pergl, Z. Struska. *Agilní modelování a metoda BORM.* (2008), http://www.google.cz/url?sa=t&source=web&cd=1&ved=0CCQQFjAA&url=http

[14] J. Polák, V. Merunka, A. Carda. *Umění systémového návrhu: objektově orientovaná tvorba informačních systémů pomocí původní metody BORM.* Grada, Prague (2003) ISBN 80-247-0424-2.

[15] J. Schmuller. *Myslíme v jazyku UML knihovna programátora.* Grada Publishing (2001) Praha ISBN 80-247-0029-8.

[16] A. B. Sterneckert. *Critical Incident Management* (2003).

[17] Z. Struska. *BORM Method and Complexity Estimation.* In Scientia Agriculturae Bohemica, 2008-1 Special, http://sab.czu.cz/cs/?r=4407&mp=download&sab=19&part=121.

[18] J. Svačina. *UML efektivně a prakticky.* In SYSTEMS INTEGRATION 2003, http://si.vse.cz/archive/proceedings/2003/uml-efektivne-a-prakticky.pdf.

[19] P. Šplíchal, R. Pergl, M. Pícka. *BORM Model Transformation.* In SYSTÉMOVÁ INTEGRACE 2/2011, http://www.cssi.cz/cssi/system/files/all/si-2011-02-07-Splichal-Pergl-Picka.pdf.

[20] *Unified Modeling Language.* Wikipedie, Otevřená encyklopedie, http://cs.wikipedia.org/wiki/Unified_Modeling_Language.

[21] *Vývojový diagram.* Wikipedie, Otevřená encyklopedie, http://cs.wikipedia.org/wiki/V

Simulace interakce bakteriálních kolonií*

Josef Smolka

2. ročník PGS, email: smolkjos@fjfi.cvut.cz
Katedra softwarového inženýrství v ekonomii
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Miroslav Virius, Katedra softwarového inženýrství v ekonomii,
Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. The paper introduces the problem of bacterial colony simulation and proposal of the discrete model for a simulation of mutually interacting bacterial bodies which is based on experimental observations. Special tool for simulations is designed and implemented in Java programming language.

Keywords: bacterial colony simulation, object-oriented database, Java, Groovy

Abstrakt. Příspěvek představuje problematiku simulace bakteriálních kolonií a návrh diskrétního modelu pro simulaci interakce dvou a více bakteriálních těles, který vychází z experimentálních pozorování. Pro potřeby simulace modelu je navržen a implementován speciální nástroj v programovacím jazyce Java.

Klíčová slova: simulace bakteriální kolonie, objektová databáze, Java, Groovy

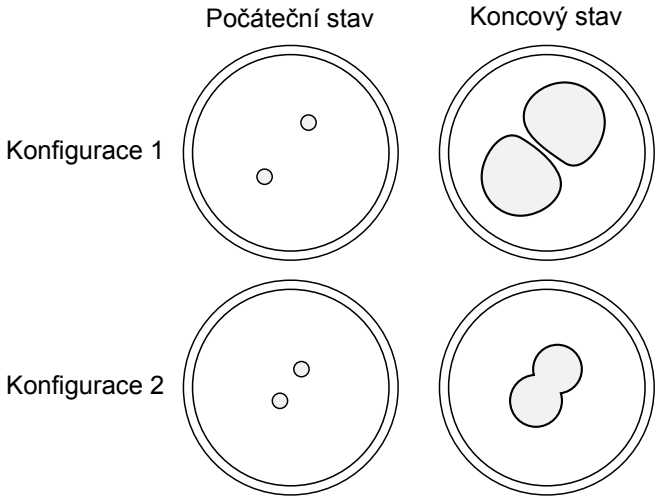
1 Úvod

Růst monoklonální bakteriální kolonie (konkrétně bakterie Serratia rubidaea) je ovlivněn různými faktory, jako například přísun živin nebo koncentrace a rozložení bakterií v prostoru. Experimenty ukazují, že vývoj kolonie může být značně ovlivněn i přítomností jiné kolonie stejného druhu, a to způsobem, který naznačuje rozvinutou komunikační schopnost bakterií. Jak tato komunikace probíhá? Jak se kolonie navzájem ovlivňují? Tyto otázky se snaží zodpovědět skupina mikrobiologů pomocí sady experimentů. Při svém výzkumu chtějí využít simulační nástroj, který by umožnil vzájemnou interakci bakteriálních kolonií simulovat. Tento článek popisuje návrh a realizaci tohoto nástroje.

1.1 Podoba experimentu

Experiment začíná inokulací bakterií na Petriho misku s agarem, který obsahuje potřebné živiny pro jejich kultivaci. Bakterie jsou umístěny v předem připraveném rozložení, podle kterého tvoří dvě nebo více kolonií. V rámci experimentu se pozorují interakce těles jednotlivých kolonií v závislosti na různých podmínkách: rozložení, přítomnost cizích těles, přítomnost různých látek, a jiné. Délka jednoho experimentu se pohybuje v řádech jednotek dnů.

*Tato práce byla podpořena grantem SGS 11/167



Obrázek 1: V případě konfigurace 1 jsou těla dvou bakteriálních kolonií natolik vzdálená, že koncentrace signálních látek, která zabraňuje dalšímu rozvoji bakterií, dosáhne kritické úrovně dříve, než se kolonie stačí spojit. Konfigurace 2 pak zobrazuje opačný případ.

2 Základní modely

Simulace množiny bakterií (tj. jedna nebo více kolonií) na misce s agarem je založena na modelování chování jednotlivých buněk v diskrétním prostoru a čase. Tento model vychází z experimentálních pozorování a ze všeobecně známých faktů. Základní myšlenky modelu vychází z několika jednoduchých spojitých modelů křivky růstu bakteriální populace.

Bakterie se množí metodou binárního dělení, tedy z jedné mateřské buňky vzniknou za příznivých podmínek dvě dceřiné [9]. Pokud budeme uvažovat pouze toto tvrzení, můžeme růst kolonie popsat obyčejnou diferenciální rovnicí 1, kde y je celkový počet bakterií v čase $t[hod]$ a $k_1[hod^{-}1]$ je koeficient růstu kolonie [3][9].

$$\frac{dy(t)}{dt} = k_1 y(t) \tag{1}$$

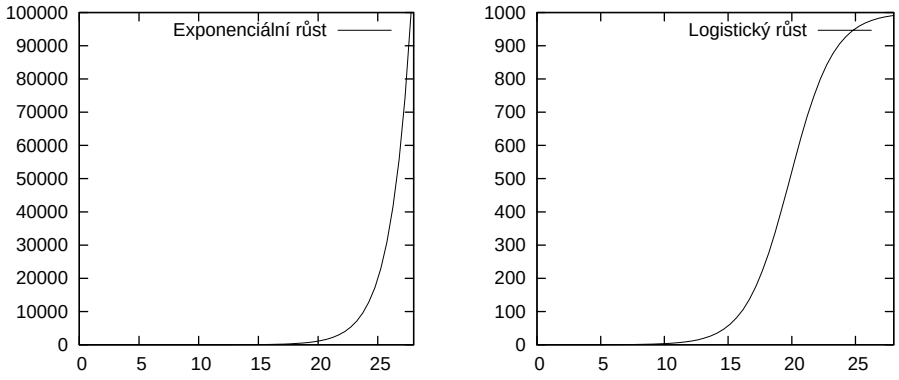
Rychlost růstu kolonie je tedy v tomto jednoduchém modelu přímo úměrná celkovému počtu bakterií a celková populace není nijak omezena.

Uvažujeme-li, že s přibývajícím počtem bakterií se úměrně zhoršují podmínky umožňující jejich další rozvoj (ubývající živiny, množící se odpadní látky), musí se tempo růstu těmto podmínkám přizpůsobit. Tento fakt zohledňuje rovnice 2.

$$\frac{dy(t)}{dt} = (k_1 t - k_2 y(t))y(t) \tag{2}$$

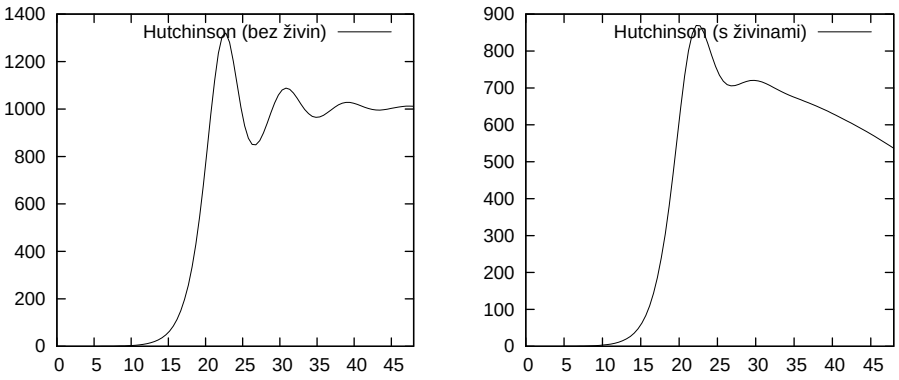
Z grafu 2, který ukazuje vývoj bakteriální populace v prvních dvaceti osmi hodinách, je vidět, že v případě druhého modelu má křivka růstu tvar logistické křivky a je tedy shora omezena. Tento model však obsahuje nepřesnost, jelikož rychlost růstu kolonie v čase t by měla být závislá na počtu bakterií v předchozí generaci. Tato skutečnost je zahrnuta v tzv. Hutchinsonově rovnici (viz rovnice 3), kde T_c je délka buněčného cyklu [5][3].

$$\frac{dy(t)}{dt} = (k_1 t - k_2 y(t - T_c))y(t) \tag{3}$$



Obrázek 2: Graf vlevo zobrazuje vývoj bakteriální populace modelovaný rovnicí 1, graf vpravo zobrazuje vývoj bakteriální populace modelovaný rovnicí 2. Osa X zobrazuje čas v hodinách, osa Y zobrazuje populaci v tis. bakterií. Rovnice byly řešeny ve volně dostupném programovém balíku Octave pomocí metody LSODE (autor Alan C. Hindmarsh). Výpočet byl proveden s $y_0 = 10$, $k_1 = 0.58$ a $k_2 = 0.00000058$.

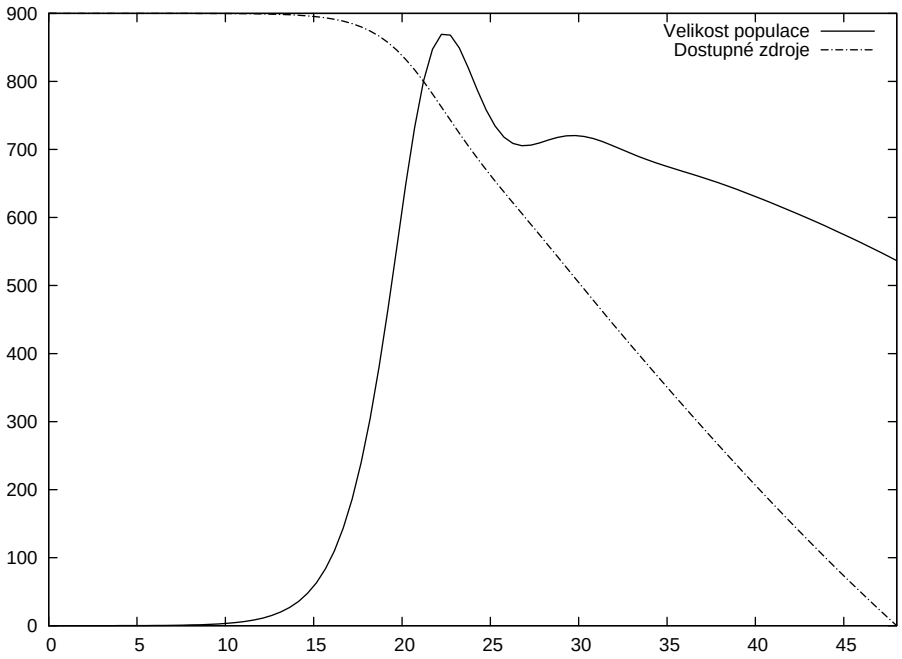
Na grafu 3 je vidět cyklické kolísání celkového počtu buněk, které se po nějaké době ustálí na podobné hodnotě jako v případě rovnice 2. Všechny dosud uvedené spojité modely uvažovaly pouze s vnitřními faktory, které ovlivňovaly tempo dělení a úmrtnosti buněk. Jedním z vnějších faktorů, které je určitě nutné do modelu zahrnout, je omezenost okolních zdrojů, především pak živin. Tuto skutečnost lze do modelu zahrnout přidáním další rovnice, popisující úbytek zdrojů v závislosti na velikosti populace (viz rovnice 4).



Obrázek 3: Graf vlevo zobrazuje vývoj bakteriální populace modelovaný Hutchinsonovou rovnicí, graf vpravo zobrazuje vývoj bakteriální populace modelovaný modifikovanou Hutchinsonovou rovnicí s přidaným omezením na množství živin. Osa X zobrazuje čas v hodinách, osa Y zobrazuje populaci v tis. bakterií. Rovnice byly řešeny ve volně dostupném programovém balíku Octave pomocí metody ODE78D (metoda pro řešení ODR se zpožděním).

$$\begin{aligned} \frac{dy(t)}{dt} &= (k_1 t - k_2 y(t - T_c) - k_3 \frac{y}{r}) y(t) \\ \frac{dr(t)}{dt} &= -k_r y(t) \end{aligned} \tag{4}$$

S ubývajícím počtem živin se zvyšuje počet buněk, které v daném čase nezískají dostatek živin pro další fungování a umírají. Závislost velikosti populace na počtu živin zobrazuje graf 4.



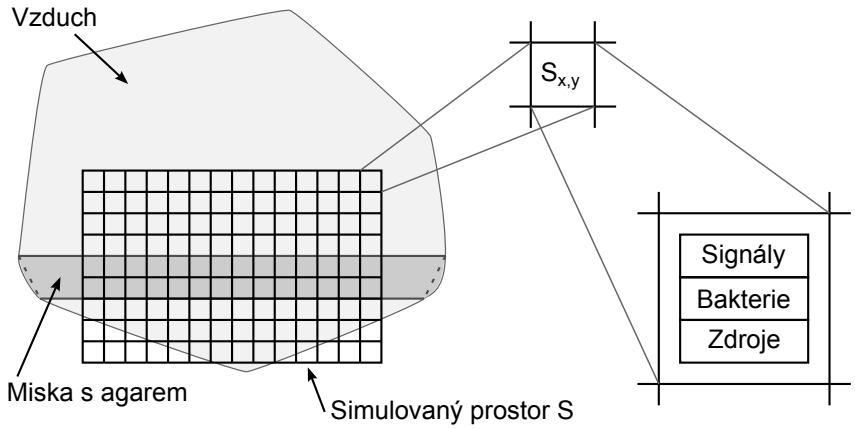
Obrázek 4: Graf zobrazující závislost velikosti populace na množství dostupných živin v prostředí. S klesajícím množstvím živin se zvyšuje úmrtnost buněk, která nakonec převládne nad tempem dělení buněk, a celková populace se začne snižovat.

3 Simulovaný model

Pro simulaci byl zvolen diskrétní model v prostoru a čase. Simulovaný prostor (tj. řez miskou s agarem a prostorem nad miskou) je pokryt pravidelnou čtvercovou sítí S. Každé políčko $S_{x,y} \in S$ je charakterizováno souborem vlastností (G, B, R), kde:

- $G = \{g_1 \dots g_n\}$ je množina signálních látek, jejichž koncentrace způsobují změny v metabolismu buněk,
- $B = \{b_1 \dots b_k\}$ je množina bakterií ve čtverci $S_{x,y}$, mohutnost množiny $|B|$ představuje koncentraci buněk v $S_{x,y}$,
- $R = \{r_1 \dots r_m\}$ je množina dostupných zdrojů v $S_{x,y}$.

Tím, že každé $S_{x,y}$ může obsahovat žádnou, jednu, nebo více bakterií, je dosaženo proměnlivé koncentrace buněk v prostoru. Pokud by model počítal pouze s žádnou, nebo jednou buňkou v $S_{x,y}$, neodpovídalo by to stavu, kdy u dospělé kolonie lze jednoznačně pozorovat hustější koncentraci buněk v jádře (způsobeno pravděpodobně inokulací buněk)[2]. Konkrétní bakterie b_i je charakterizována souborem vlastností (a, m, R, s), kde a je stáří bakterie, m je celková hmota bakterie, R je zásoba zdrojů a s je stav metabolismu. Samotná reprezentace kolonie je pouze první částí modelu. Druhou část tvoří evoluční



Obrázek 5: Reprezentace simulovaného prostoru v navrženém diskrétním modelu. Prostor je pokryt pravidelnou čtvercovou sítí, kde každý čtverec v síti je charakterizován souborem vlastností.

algoritmus, který dokáže pomocí několika procesů provést přechod kolonie ze stavu K_t do K_{t+1}, kde stav K_i je dán odpovídající konfigurací S v čase t.

3.1 Simulované procesy

Evoluční algoritmus implementovaný v simulátoru se skládá z procesů dvou druhů: procesy prvního druhu se týkají pouze samotných bakterií, procesy druhého druhu pak simulují změny v prostředí. Mezi první druh patří tyto:

- Příjem živin – Bakterie se snaží pozřít živiny v blízkém okolí, celkové množství přijatého zdroje R_i bakterií B_j v čase t je popsáno funkcí $R_{R_i}(m_j, h_j) = \frac{k_0 + k_1 m_j(t)}{k_2 h_j(t)}$, kde k_0 je základní přijaté množství, $k_1 m_j$ představuje živiny navíc, které dokáže pozřít mohutnější jedinec a $k_2 h_j$ je penalizace za vzdálenost bakterie od agaru, který živiny obsahuje.

- Příjem kyslíku – Pokud simulujeme aerobní bakterie, je příjem kyslíku bakterie B_j v čase t popsán funkcí 5, kde $C(S_{x,y})$ představuje okolí políčka $S_{x,y}$.

- Syntéza – Bakterie přeměňují přijaté živiny a kyslík na vlastní hmotu a vylučují signální látky.

- Údržba – Bakterie stárnou a udržují se.

- Reprodukce – Pokud jsou splněny všechny podmínky, může se bakterie pokusit rozdělit.

$$O(S_{x,y}) = (k_0 + k_1 m_j(t)) \left(1 - \frac{\sum_{c \in C(S_{x,y})} |B_c|}{8 \max_{c \in S} |B_c|} \right) \tag{5}$$

Úmrtnost buněk není v modelu zohledněna, protože v časovém rozsahu, ve kterém experimenty probíhají, je zanedbatelná. Mezi procesy druhého druhu patří difuze signálních látek ve vzduchu a v agaru, které následně ovlivňují metabolismus buněk a mohou například zapříčinit to, že se bakterie přestanou v oblasti zvýšené koncentrace dělit.

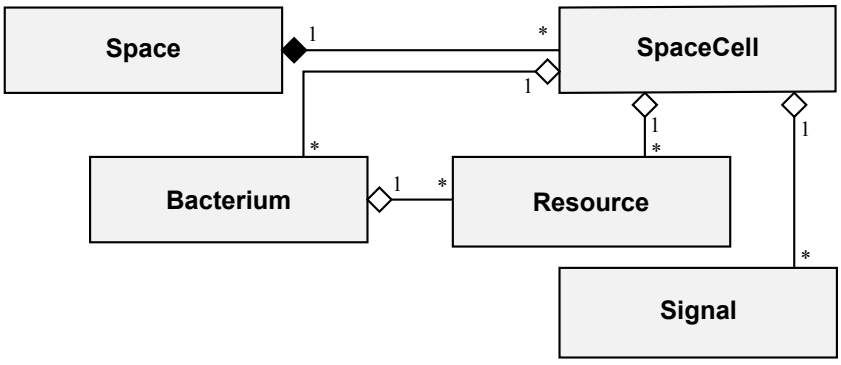
4 Návrh a implementace simulátoru

Základem simulátoru je jednoduchá objektová databáze uložená v operační paměti počítače, která obsahuje veškeré informace o simulovaném prostoru S. Schéma databáze zachycuje diagram tříd na obrázku 6. Databáze umožňuje práci v pseudotransakčním režimu v následujícím smyslu:

- Veškeré změny v databázi jsou zaznamenávány ve speciálním změnovém deníku, který je aplikován až při commitu.

- Čtení z databáze není tímto změnovým deníkem ovlivněn.

Toto chování se může zdát na první pohled podivné, ale odpovídá faktu, že všechny změny při přechodu mezi stavy se musí aplikovat najednou. Podstatnou částí databáze je prostorový index, který urychluje přístup k informacím na konkrétním políčku $S_{x,y}$. Tento index je implementován pomocí dynamické datové stromové struktury Quadtree.



Obrázek 6: Diagram tříd znázorňující datový model objektové databáze simulátoru.

Databáze poskytuje pro manipulaci s daty speciální dotazovací jazyk založený na dynamickém jazyku Groovy. Třída Space zde v podstatě zastupuje databázové schéma, přičemž každý experiment má svoje schéma. Práce s ním probíhá následovně:

```
// vytvoření nového schématu
d.create('space').name('myExperiment1')
// přístup k nově vytvořenému schématu
d.space['myExperiment1']
// přístup k předdefinovanému schématu
d.use('myExperiment1')
d.s
// smazání schématu a všech dat, které obsahuje
d.s.delete
```

Práce se samotnými daty je založená na schopnostech jazyka Groovy efektivně pracovat s kolekcemi objektů a vypadá například následovně:

```
// vytvoříme řadu políček v síti
for (x in 0 .. 10) d.s.create('cell').location(x, 0)
// do každého políčka vložíme jednu buňku s náhodně zvolenou hmotností
d.s.cells*.insert(d.s.create('bacterium').mass(d.rand))
// a nyní smažeme všechny bakterie, které mají hmotnost menší než 0.5
d.s.select('bacterium').filter{b -> b.mass < 0.5}*.delete
```

Co kdybychom například chtěli vybrat v určité oblasti bakterie starší než zadaná hranice?

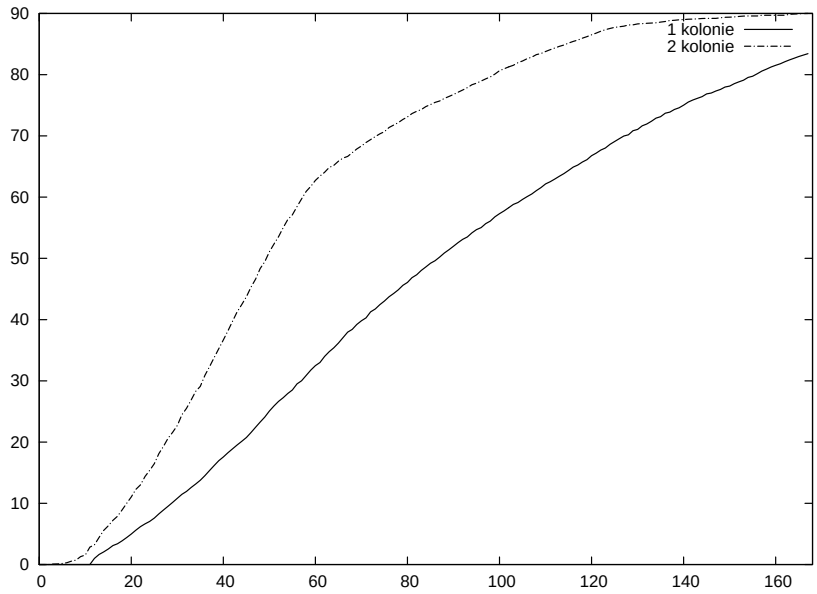
```
d.s.cells.findAll{c -> c.x <= 10}.collect{c -> c.bacteria}
         .flatten().findAll{b -> b.age > t}
// nebo jednodušeji
d.s.bacteria.findAll{b -> b.cell?.x < 10 && b.age > t}
```

Použití Groovy operátorů **.*, *?.* a funkcí *findAll* a *collect* umožňuje velice jednoduchou implementaci navrženého modelu v prostředí platformy Java.

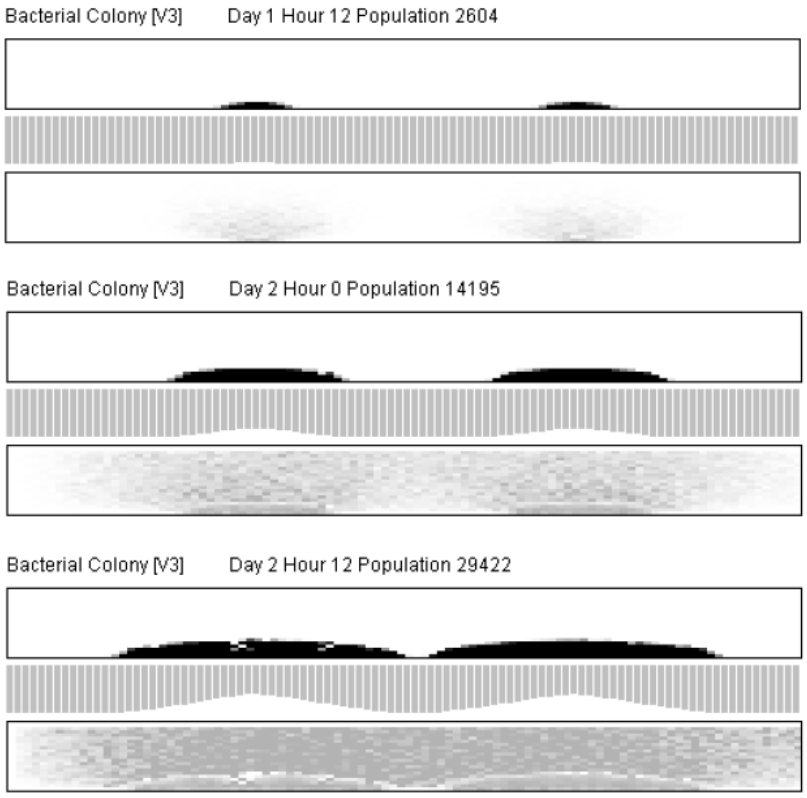
5 Výsledky

Implementovaný model prokazuje některé experimentálně pozorované vlastnosti. Na grafu 7 je vidět, že obě křivky vývoje populace ze začátku prokazují exponenciální charakter, což je zapříčiněno dostatkem místa i živin. V případě dvou kolonií je pozorovatelná výrazná změna charakteru křivky v momentě kontaktu bakteriálních těles. Difuze signální látky ve vzduchu v některých případech skutečně zapříčiňuje to, že se dvě kolonie zcela nespojí a zachovají si samostatnost, přestože jim nic jiného ve spojení nebrání. V současné době probíhá ladění a validace modelu.

Co se týče simulačního nástroje, jeho jádro bylo implementováno v programovacím jazyce Java. Samotný evoluční algoritmus byl však implementován v podobě pouhých dotazů do objektové databáze samotného nástroje. Obrázek 8 zobrazuje sérii tří snímků ze simulačního nástroje. Každý snímek obsahuje tři části: první část zobrazuje řez koloniemi, druhá část zásobu živin a poslední třetí část pak difuzi signální látky.



Obrázek 7: Graf vývoje celkové populace jedné, resp. dvou kolonií v průběhu sedmi dnů simulovaný navrženým diskrétním modelem. Osa X zobrazuje čas v hodinách, osa Y zobrazuje celkovou populaci v tisících bakterií.



Obrázek 8: Snímky obrazovky ze simulačního nástroje.

6 Závěr

Článek představil návrh diskrétního modelu pro simulaci interakce dvou a více bakteriálních kolonií, který vychází z několika základních spojitých modelů. Tyto modely byly v úvodu také stručně popsány. Model není zatím zcela hotov a probíhá jeho další zpřesňování.

Pro potřeby simulace byl navržen a implementován nástroj v jazyce Java. Pro zefektivnění implementačního procesu samotného modelu byla navržena jednoduchá objektová databáze, jejíž dotazovací jazyk umožnil rychlé prototypování modelu. Simulační nástroj je vytvářen v rámci výzkumu chování a interakce monoklonálních bakteriálních kolonií na Katedře filosofie a dějin přírodních věd Přírodovědecké fakulty UK.

Literatura

[1] R. Caol, M. Francisco-Fernández, E. J. Quinto. *A random effect multiplicative heteroscedastic model for bacterial growth.* In 'BMC Bioinformatics' (2010), 11:77

[2] J. Čepl, I. Pátková, A. Blahůšková, F. Cvrčková, A. Markoš. *Patterning of mutually interacting bacterial bodies: close contacts and airborne signals.* In: 'BMC Microbiology' (2010), 10:139

[3] R. Farana, L. Landryová, J. Lokosová, L. Smutný, A. Víteček, M. Vítečková, R. Wagnerová. *Programová podpora simulace dynamických systémů (sbírka řešených příkladů).* Ostrava, 1996, ISBN 80-02-01129-5

[4] M. Ginovart, D. López, J. Valls, M. Silbert. *Individual based simulations of bacterial growth on agar plates.* In: 'Physica A 305' (2002), 604–618

[5] G. E. Hutchinson. *Circular causal systems in ecology.* In: 'Annals of the New York Academy of Sciences' (1948), 221–246

[6] Z. Kutalika, M. Razaza, J. Baranyib. *Connection between stochastic and deterministic modeling of microbial growth.* In: 'Journal of Theoretical Biology 232' (2005), 285–299

[7] P. Melke, P. Sahlin, A. Levchenko, H. Jönsson. *A Cell-Based Model for Quorum Sensing in Heterogeneous Bacterial Colonies.* In: 'PLoS Computational Biology' Vol. 6, No. 6 (2010)

[8] C. Vlachosa, R. C. Patona, J. R. Saundersb, Q. H. Wuc. *A rule-based approach to the modeling of bacterial ecosystems.* In: 'BioSystems 84' (2005) 49–72

[9] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, K. van Riet. *Modeling of the Bacterial Growth Curve.* In: 'Applied and Environmental Microbiology' Vol. 56, No. 6 (1990), 1875–1881

[10] M. H. Zwietering, J. T. de Koos, B. E. Hasenack, J. C. de Wit, K. var Riet. *Modeling of Bacterial Growth as a Function of Temperature.* In: 'Applied and Environmental Microbiology' Vol. 57, No. 4 (1991), 1094–1101

Electronic Properties of Carbon Nanostructures

Jan Smotlacha

8th year of PGS, email: smota@centrum.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Richard Pincak, Institute of Experimental Physics,
Slovak Academy of Sciences

Abstract. The electronic structure of the carbon nanoparticles is investigated for the hyperboloidal geometries. The main object of our interest is the local density of states. We use a continuum gauge field-theory model for this purpose.

Keywords: graphene, carbon nanoparticles, disclination

Abstrakt. Elektronické vlastnosti uhlíkových nanoèástic jsou zkoumány pro geometrii hyperboloidu. Hlavním pøedmìtem výzkumu je lokální hustota stavù. Pro tento úèel používáme model kalibraèní invariance.

Klíčová slova: grafén, uhlíkové nanoèástice, krystalová porucha

1 Introduction

Nanostructured carbon materials are materials with a special geometrical structure of their molecules which we call carbon nanoparticles. This geometrical structure is accompanied by the topological defects in a hexagonal planar lattice called graphene.

There are many variously-shaped carbon nanostructures known. The most famous is fullerene, which has the structure of a soccer ball and can be approximated as a sphere. It is composed of 60 carbon atoms which create 20 hexagons and 12 pentagons. However other structures also exist, for example nanocones, nanotoroids, nanotubes, nanohorns etc. A wide variety of electronic properties of these structures have been studied. They suggest a potential use in nanoscale devices like quantum wires, transistors or molecular memory devices.

The electronic properties of these structures can be explored by solving the Dirac equation at a curved surface [4]. In most cases, the defects on this surface originate from the presence of the pentagons for the positive curvature and the heptagons for the negative curvature [1]. They cause breaking of the rotational symmetry of the wave function. It must be compensated by the addition of some local gauge fields. For the calculation of the local density of states, the hyperboloidal geometry is used.

2 Basic formalism

We introduce the Dirac equation in (2+1) dimensions. It has the form:

$$i\gamma^{\alpha}e_{\alpha}^{\mu}[\nabla_{\mu} - ia_{\mu} - iA_{\mu}]\psi = E\psi. \tag{1}$$

The wave function ψ, the so-called bispinor, is composed of two parts:

$$\psi = \begin{pmatrix} \psi_A \\ \psi_B \end{pmatrix}, \tag{2}$$

each corresponding to different sublattices of the curved graphene sheet [5].

Next, some additional gauge fields are introduced. Without them, the Hamiltonian in (1) would have the form:

$$H_0 = i\gamma^\alpha e_\alpha^\mu [\partial_\mu - iA_\mu] \tag{3}$$

and the corresponding wave function we denote ψ_0. Then:

$$\psi(r,\varphi) = \exp(i\Omega_1(r,\varphi)) \exp(i\Omega_2(r,\varphi)) \cdots \exp(i\Omega_n(r,\varphi))\psi_0(r,\varphi), \tag{4}$$

where $\Omega_i, i = 1, \ldots n$ are functions, theirs form follows from the boundary conditions. Because:

$$H_0\psi_0 = E\psi_0 \tag{5}$$

and, at the same time:

$$H\psi = E\psi, \tag{6}$$

it follows that:

$$H = \exp(i\Omega_1) \exp(i\Omega_2) \cdots \exp(i\Omega_n) H_0 \exp(i\Omega_n) \ldots \exp(-i\Omega_2) \exp(-i\Omega_1). \tag{7}$$

For our purpose, we introduce 2 additional gauge fields,a_μ and ω_μ, $\mu = \xi, \varphi$.

The gauge field a_μ arises from spin rotation invariance. The gauge field ω_μ comes from introducing the zweibein e_α which incorporates fermions on the curved 2D surface. It ensures the invariance of (1) for different choices of the frame and it satisfies

$$\partial_\mu e_\nu^\alpha - \Gamma_{\mu\nu}^\rho e_\rho^\alpha + (\omega_\mu)_\beta^\alpha e_\nu^\beta = 0, \tag{8}$$

where Γ_μ is the Levi-Civita connection coming from the metrics $g_{\mu\nu}$ (see below). Then ω_μ is called the spin connection. Next, the covariant derivative ∇_μ is defined as:

$$\nabla_\mu = \partial_\mu + \Omega_\mu, \tag{9}$$

where

$$\Omega_\mu = \frac{1}{8}\omega_\mu^{\alpha\beta}[\gamma_\alpha, \gamma_\beta] \tag{10}$$

denotes the spin connection in the spinor representation. The Dirac matrices γ_α can be replaced in two dimensions by the Pauli matrices σ_α:

$$\gamma_1 = -\sigma_2, \qquad \gamma_2 = \sigma_1. \tag{11}$$

A_μ is the vector potential arising from the external magnetic field.

The metric $g_{\mu\nu}$ of the 2D surface comes from following parametrisation using the two parameters ξ, φ:

$$(\xi,\varphi) \to \overrightarrow{R} = (x(\xi,\varphi), y(\xi,\varphi), z(\xi,\varphi)), \tag{12}$$

where

$$0 \leq \xi < \infty, \qquad 0 \leq \varphi < 2\pi. \tag{13}$$

The 4 components of the metric are defined as:

$$g_{\mu\nu} = \partial_\mu \overrightarrow{R} \partial_\nu \overrightarrow{R}. \tag{14}$$

The hyperboloid geometry which we use has, for both heptagons and pentagons, very similar but not identical parametrisation. We consider it in separate chapters. Generally, the non-diagonal components of the metric are:

$$g_{\xi\varphi} = g_{\varphi\xi} = 0. \tag{15}$$

For the zweibeins and the diagonal components of the metric the following relationships hold:

$$e^1_\xi = \sqrt{g_{\xi\xi}} \cos\varphi, \qquad e^1_\varphi = -\sqrt{g_{\varphi\varphi}} \sin\varphi, \tag{16}$$

$$e^2_\xi = \sqrt{g_{\xi\xi}} \sin\varphi, \qquad e^2_\varphi = \sqrt{g_{\varphi\varphi}} \cos\varphi, \tag{17}$$

and for the spin connection coefficients ω_μ:

$$de^1 = -\omega^{12} \wedge e^2, \qquad de^2 = -\omega^{21} \wedge e^1, \qquad \omega^{12} = -\omega^{21}, \tag{18}$$

so:

$$\omega^{12}_\varphi = -\omega^{21}_\varphi = 1 - \frac{\partial_\xi \sqrt{g_{\varphi\varphi}}}{\sqrt{g_{\xi\xi}}} = 2\omega, \tag{19}$$

$$\omega^{12}_\xi = \omega^{21}_\xi = 0. \tag{20}$$

Then the coefficients Ω_μ are:

$$\Omega_\xi = 0, \qquad \Omega_\varphi = i\omega\sigma_3. \tag{21}$$

If we write the wave function in the form

$$\begin{pmatrix} \psi_A \\ \psi_B \end{pmatrix} = \frac{1}{\sqrt[4]{g_{\varphi\varphi}}} \begin{pmatrix} \widetilde{u}(\xi) e^{i\varphi j} \\ \widetilde{v}(\xi) e^{i\varphi(j+1)} \end{pmatrix}, \qquad j = 0, \pm 1, ... \tag{22}$$

and substituting (22) into (1) we obtain

$$\partial_\xi \widetilde{u} - (j + 1/2 - a_\varphi + A_\varphi) \sqrt{\frac{g_{\xi\xi}}{g_{\varphi\varphi}}} \widetilde{u} = E \sqrt{g_{\xi\xi}} \widetilde{v}, \tag{23}$$

$$-\partial_\xi \widetilde{v} - (j + 1/2 - a_\varphi + A_\varphi) \sqrt{\frac{g_{\xi\xi}}{g_{\varphi\varphi}}} \widetilde{v} = E \sqrt{g_{\xi\xi}} \widetilde{u}. \tag{24}$$

It is possible to try to approximate the considered geometry by the metric of the cone [2]. However, this approach does not correspond to the real situation because of the point-like apex. Here we propose a method to avoid this problem.

3 Geometrical properties

3.1 Heptagonal defects

In the case of negative curvature and associated heptagonal defects, the parametrisation (12) for the case of the hyperboloid is:

$$(\xi, \varphi) \rightarrow (a\cosh\xi\cos\varphi, a\cosh\xi\sin\varphi, c\sinh\xi), \tag{25}$$

where a and c are some dimensionless parameters. The corresponding diagonal components of the metric are:

$$g_{\xi\xi} = a^2\sinh^2\xi + c^2\cosh^2\xi, \qquad g_{\varphi\varphi} = a^2\cosh^2\xi \tag{26}$$

and the nonzero spin connection term:

$$\omega_{\varphi}^{12} = 1 - \frac{a\sinh\xi}{\sqrt{g_{\xi\xi}}}. \tag{27}$$

The defect arises by the so-called cut and glue procedure - we cut a line in the graphene plane, add a 60° area and glue the resulting borders [2]. The geometrical properties of the new surface can be described with the help of the gauge potentials $\overrightarrow{W}_{\mu}^{(0)}$, which change the initial components of the metric (now denoted $g_{\mu\nu}^{(0)}$) [6]:

$$g_{\mu\nu}^{(0)} \rightarrow g_{\mu\nu} = \nabla_{\mu}\overrightarrow{R}_{(0)} \cdot \nabla_{\nu}\overrightarrow{R}_{(0)}, \tag{28}$$

where:

$$\nabla_{\mu}\overrightarrow{R}_{(0)} = \partial_{\mu}\overrightarrow{R}_{(0)} + [\overrightarrow{W}_{\mu}^{(0)}, \overrightarrow{R}_{(0)}]. \tag{29}$$

Then

$$g_{\mu\nu} = \partial_{\mu}\overrightarrow{R}_{(0)} \cdot \partial_{\nu}\overrightarrow{R}_{(0)} + \partial_{\mu}\overrightarrow{R}_{(0)}[\overrightarrow{W}_{\nu}^{(0)}, \overrightarrow{R}_{(0)}] +$$

$$\partial_{\nu}\overrightarrow{R}_{(0)}[\overrightarrow{W}_{\mu}^{(0)}, \overrightarrow{R}_{(0)}] + (\overrightarrow{W}_{\mu}^{(0)}\overrightarrow{W}_{\nu}^{(0)})\overrightarrow{R}_{(0)}^2 - (\overrightarrow{W}_{\mu}^{(0)}\overrightarrow{R}_{(0)})(\overrightarrow{W}_{\nu}^{(0)}\overrightarrow{R}_{(0)}) \tag{30}$$

and the components of the metric and the spin connection term will be changed such that:

$$g_{\xi\xi} = a^2\sinh^2\xi + c^2\cosh^2\xi, \qquad g_{\varphi\varphi} = a^2\alpha^2\cosh^2\xi, \tag{31}$$

$$\omega_{\varphi}^{12} = 1 - \frac{a\alpha\sinh\xi}{\sqrt{g_{\xi\xi}}}, \qquad \alpha = 1 + \nu, \tag{32}$$

where $\nu = N/6$ is called the Frank index and N is the number of heptagons in the defect. In this paper, we take $N = 1$. Let us stress that as the number of defects increases, the geometrical structure becomes more complicated and we have to take into account additional assumptions [2].

We can encircle the origin of the defect ($\xi = 0$) by a closed loop C_{ϵ} and integrate over it. The result is:

$$\oint_{C_{\epsilon}} d\overrightarrow{s} = 2\pi\nu. \tag{33}$$

No transformation of variables can change this value. If the values of the gauge field $\overrightarrow{W}_{\mu}^{(0)}$ are:

$$W_{\mu}^{(0)i=1,2} = 0, \qquad W_{\mu}^{(0)i=3} = W_{\mu}^{(0)}, \tag{34}$$

where:

$$W_{x}^{(0)} = -\nu y/r^2, \qquad W_{y}^{(0)} = \nu x/r^2, \qquad r = \sqrt{x^2+y^2}, \tag{35}$$

then:

$$\oint_{C_\epsilon} d\overrightarrow{s} = 2\pi\nu = \oint_{C_\epsilon} W_{\mu}^{(0)} dx^{\mu}, \tag{36}$$

so $\overrightarrow{W}_{\mu}^{(0)}$ serves as a vortex-like potential with a nonzero flux. This flux should be eliminated by the corresponding integral over the spin connection, so we must get:

$$\lim_{\epsilon\to 0} \oint_{C_\epsilon} \omega_{\varphi}^{12} d\varphi = -2\pi\nu. \tag{37}$$

Substituting (32) into the appropriate integral, the required result is readily obtained.

For our purpose, the gauge field $a_\varphi = N/4$. In the general case, a_φ depends on two constants N and M as $a_\varphi = N/4 + M/3$, where $M = -1, 0, 1$ for an even number of defects and $M = 0$ for an odd number of defects [2, 3, 5].

If the magnetic field is chosen in such a way that $\overrightarrow{A} = B(y, -x, 0)/2$, then:

$$A_\varphi = -\Phi \cosh^2\xi, \qquad A_\xi = 0, \tag{38}$$

where:

$$\Phi = \frac{1}{2} a^2 \Phi_0 B, \qquad \Phi_0 = \frac{e}{\hbar c}. \tag{39}$$

The geometric units are used, i.e. $e = \hbar = c = 1$.

3.2 Pentagonal defects

The case of the positive curvature is described in more detail in [3]. The parametrisation is changed into:

$$(\xi, \varphi) \to (a \sinh\xi \cos\varphi, a \sinh\xi \sin\varphi, c \cosh\xi), \tag{40}$$

and the diagonal components of the metric are:

$$g_{\xi\xi} = a^2 \cosh^2\xi + c^2 \sinh^2\xi, \qquad g_{\varphi\varphi} = a^2 \sinh^2\xi. \tag{41}$$

Introducing the gauge potentials $\overrightarrow{W}_{\mu}^{(0)}$ as for the heptagonal defects, the component $g_{\varphi\varphi}$ of the metric changes such that:

$$g_{\varphi\varphi} = a^2\alpha^2 \sinh^2\xi, \tag{42}$$

where $\alpha = 1 - \nu$. This means that in the cut and glue procedure, we cut a 60° area instead of inserting it. Then the nonzero spin connection term is:

$$\omega_{\varphi}^{12} = 1 - \frac{a\alpha \cosh\xi}{\sqrt{g_{\xi\xi}}}. \tag{43}$$

The values of the gauge field and the magnetic field are the same as in the previous case:

$$a_\varphi = N/4, \qquad \overrightarrow{A} = B(y, -x, 0)/2, \tag{44}$$

so that for the chosen parametrisation:

$$A_\varphi = -\Phi \sinh^2 \xi, \qquad A_\xi = 0, \tag{45}$$

where Φ is defined as in (39).

4 Solution of the Dirac equation

The solution of (23),(24) for heptagonal and pentagonal defects in the case of hyperboloidal geometry is introduced and the local density of states is calculated here. The linear elasticity theory [6] is used. For the numerical calculations of LDoS, the method described in [7] is exploited.

4.1 Heptagonal defects

The form of (23),(24) will be:

$$\partial_\xi \widetilde{u} - (\widetilde{j} - \widetilde{\Phi} \cosh^2 \xi)\sqrt{\tanh^2 \xi + \eta}\widetilde{u} = E\sqrt{g_{\xi\xi}}\widetilde{v}, \tag{46}$$

$$-\partial_\xi \widetilde{v} - (\widetilde{j} - \widetilde{\Phi} \cosh^2 \xi)\sqrt{\tanh^2 \xi + \eta}\widetilde{v} = E\sqrt{g_{\xi\xi}}\widetilde{u}, \tag{47}$$

where:

$$\widetilde{j} = (j + 1/2 - a_\varphi)/\alpha, \qquad \widetilde{\Phi} = \Phi/\alpha, \qquad \eta = c^2/a^2. \tag{48}$$

The parameter $\eta \ll 1$ is a dimensionless parameter which describes the elasticity properties of the initial graphene plane. Due to these properties, the defects can be interpreted as small perturbations in the graphene plane. In the case of finite elasticity, we can use an approximation $\eta \sim \sqrt{\nu\epsilon}$, where $\epsilon \leq 0.1$ [3]. In this way, the elasticity is described by a small parameter ϵ. Its value is usually taken between 0.01 and 0.1. If we perform some necessary corrections to the gauge field ω_μ, then as we take $\epsilon \to 0$ we obtain the metric of the cone.

Let us now suppose $E = 0$. This energy corresponds to the so-called zero-energy mode which is appropriate for the electron states at the Fermi level. Then, the solution of (46),(47) is:

$$\widetilde{u}_0(\xi) = C(\triangle(\xi) + k \sinh \xi)^{k\widetilde{j} - \frac{\eta\widetilde{\Phi}}{2k}} \left(\frac{\cosh \xi}{\triangle(\xi) + \sinh \xi}\right)^{\widetilde{j}} \exp\left(-\frac{\widetilde{\Phi}\triangle(\xi)\sinh \xi}{2}\right), \tag{49}$$

$$\widetilde{v}_0(\xi) = C'(\triangle(\xi) + k \sinh \xi)^{-k\widetilde{j} + \frac{\eta\widetilde{\Phi}}{2k}} \left(\frac{\cosh \xi}{\triangle(\xi) + \sinh \xi}\right)^{-\widetilde{j}} \exp\left(\frac{\widetilde{\Phi}\triangle(\xi)\sinh \xi}{2}\right), \tag{50}$$

where:

$$k = \sqrt{1 + \eta}, \qquad \triangle(\xi) = \sqrt{k^2 \cosh^2 \xi - 1}, \tag{51}$$

and C, C' are the normalisation constants.

For nonzero values of E we use a perturbation scheme described in [7]. Then, for a given ξ_0, the local density of states is defined as:

$$LDoS(E) = \widetilde{u}^2(E, \xi_0) + \widetilde{v}^2(E, \xi_0). \tag{52}$$

To evaluate the local density of states, we have to calculate the normalisation constants C,C'. They differ for different values of E. For unnormalised solutions $\widetilde{u}'(\xi), \widetilde{v}'(\xi)$ of (46),(47) and each value of energy:

$$1/C^2 = 1/C'^2 = \int_0^{\xi_{max}} (\widetilde{u}'(\xi)^2 + \widetilde{v}'(\xi)^2) d\xi. \tag{53}$$

Since for $\xi_{max} = \infty$ the integral diverges, some finite value of ξ_{max} in some interval which is of particular interest is needed. In this work, we take $\xi_{max} = 2.5$ and $\xi_{max} = 2$. It follows from the parametrisation (25) that for the given values of ξ, the corresponding distance is $r = a\cosh\xi$, which means that for $a = 1\ A$ we have $r_{max} = 6.13\ A$, or $r_{max} = 3.76\ A$. These values are of the same order as the size of the Brillouin zone which is formed by the single hexagons. Each atom in the hexagon lies at a distance $1.42\ A$ from its nearest neighbours. This is the main principle of the tight-binding approximation [8] in which we only account for the influence of the nearest neighbours.

4.2 Pentagonal defects

The form of (23),(24) is:

$$\partial_\xi \widetilde{u} - (\widetilde{j} - \widetilde{\Phi}\sinh^2\xi)\sqrt{\coth^2\xi + \eta}\widetilde{u} = E\sqrt{g_{\xi\xi}}\widetilde{v}, \tag{54}$$

$$-\partial_\xi \widetilde{v} - (\widetilde{j} - \widetilde{\Phi}\sinh^2\xi)\sqrt{\coth^2\xi + \eta}\widetilde{v} = E\sqrt{g_{\xi\xi}}\widetilde{u}. \tag{55}$$

In the case $E = 0$, the corresponding solution is:

$$\widetilde{u}_0(\xi) = C(\triangle(\xi) + k\cosh\xi)^{k\widetilde{j} + \frac{\eta\widetilde{\Phi}}{2k}} \left(\frac{\sinh\xi}{\triangle(\xi) + \cosh\xi}\right)^{\widetilde{j}} \exp\left(-\frac{\widetilde{\Phi}\triangle(\xi)\cosh\xi}{2}\right), \tag{56}$$

$$\widetilde{v}_0(\xi) = C'(\triangle(\xi) + k\cosh\xi)^{-k\widetilde{j} - \frac{\eta\widetilde{\Phi}}{2k}} \left(\frac{\sinh\xi}{\triangle(\xi) + \cosh\xi}\right)^{-\widetilde{j}} \exp\left(\frac{\widetilde{\Phi}\triangle(\xi)\cosh\xi}{2}\right), \tag{57}$$

where:

$$k = \sqrt{1+\eta}, \qquad \triangle(\xi) = \sqrt{k^2\sinh^2\xi + 1}. \tag{58}$$

To calculate the solution for nonzero values of E and the local density of states we use the same procedure as presented for the heptagonal defects.

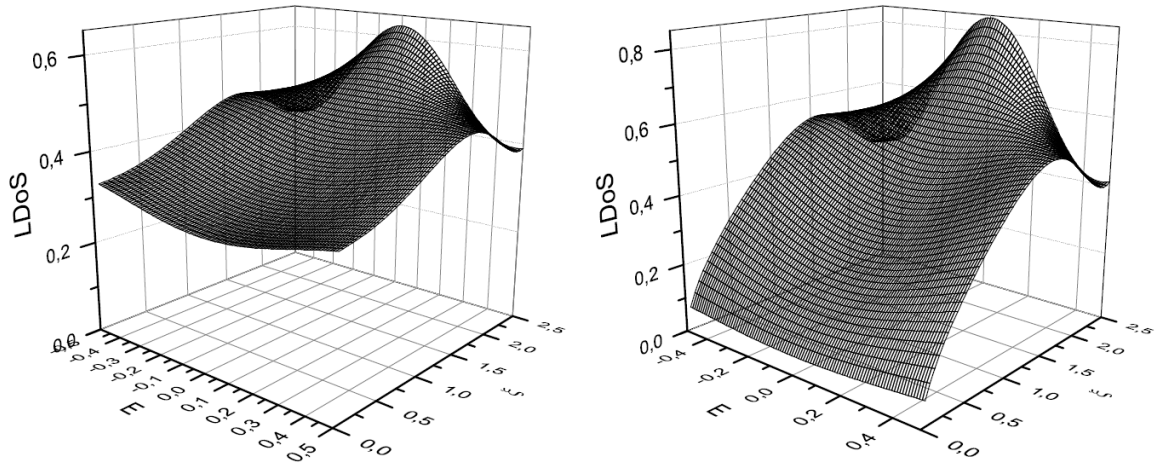


Figure 1: LDoS as a function of $E \in (-0.5, 0.5)$ and $\xi \in (0, 2.5)$ for 1-heptagon defects (left part) and 1-pentagon defects (right part) for $B = 0$; $\epsilon = 0.01$.

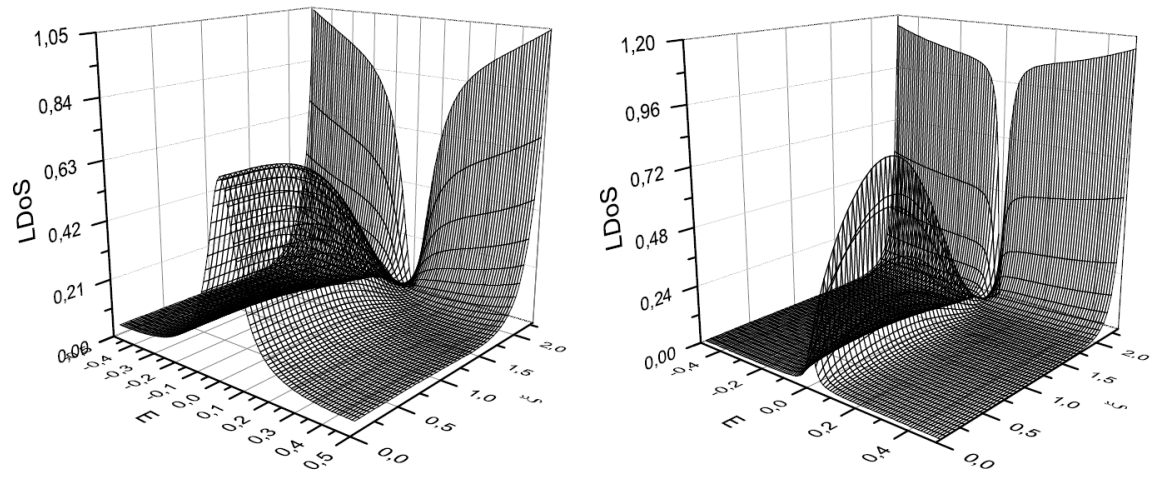


Figure 2: LDoS as a function of $E \in (-0.5, 0.5)$ and $\xi \in (0, 2.5)$ for 1-heptagon defects (left part) and 1-pentagon defects (right part) for $B = 0.5$; $\epsilon = 0.01$.

4.3 Local density of states

In Figs. 1 and 2, the LDoS as a function of energy E and the parameter ξ is presented for hyperboloidal surfaces with the defects formed by 1 polygon. In all of these figures, we set $j = 0$ in (48) and $\epsilon = 0.01$ in the expression for η. We can see the evidence that for increasing B or ξ_{max}, the LDoS is decreasing and the decrease is faster for the pentagonal defects. If we took larger ξ_{max}, the LDoS would go to zero with the exception of a small number of energies for which we would obtain plane waves. The larger values of ξ are, however, unphysical because of the limited interval of validity of the tight-binding approximation.

5 Conclusion

We have studied the electronic structure of disclinated graphene in the vicinity of heptagonal and pentagonal defects depending on the kind of a curvature (negative or positive). Hyperboloidal parametrisations (25),(40) were assumed after rejection of the conical metric. The continuum field-theory gauge model was used, in which the disclinations are incorporated using the vortex-like potential (34), (35) for the calculation of the components of the metric. The arising fictitious flux was compensated for by the gauge flux of spin connection field (19),(20). The potential (34), (35) also results in the dependence of the corresponding Dirac equation on the Frank index α which includes the number of defects. The defects are involved in (48) with the help of the parameter ϵ, which comes from the elasticity properties of the graphene.

Next, we incorporated a uniform magnetic field (38),(45) that can significantly influence the LDoS. These were calculated from the solution of the Dirac equation, which we obtained numerically with the help of the extension of an analytical solution for zero-energy modes (49),(50),(56),(57).

To conclude, the presented results have a large potential use for calculating the metallic properties of carbon nanohorns which have widespread application in electronic devices. Let us mention the significance of the zero-energy modes. Generally, they appear as a solution for disclinated graphene in the presence of a magnetic field [9] and they play a key role in explanations of anomalies, paramagnetism, high-temperature superconductance etc.

We have to stress that we assumed defects in which only 1 heptagon or 1 pentagon appeared. For a higher number of polygons in defects the calculation is more complicated, especially for heptagons, because in contrast to pentagonal defects problems with the geometrical interpretation occur. It will be useful to perform calculations for more complicated forms of defects in the future.

Acknowledgements

We sincerely thank Prof. V. A. Osipov for his helpful comments and advice. The work was supported by the Slovak Academy of Sciences in the framework of CEX NANOFLUID, by the Science and Technology Assistance Agency under Contract No. APVV 0509-07,

0171 10, VEGA Grant No. 2/0069/10 and by the Ministry of Education Agency for Structural Funds of EU in the frame of project 26220120021.

References

[1] C. Chuang, B.-Y. Jin: J. Chem. Inf. Model. **49**, 1664 (2009).

[2] P. E. Lammert, V. H. Crespi, Phys. Rev. B **69**, 035406 (2004).

[3] E.A. Kochetov, V.A. Osipov and R. Pincak, J. Phys.: Condens. Matter **22**, 395502 (2010).

[4] V.A. Osipov, E.A. Kochetov, M. Pudlak: JETP **96**, 140 (2003).

[5] D. V. Kolesnikov, V. A. Osipov, Eur. Phys. Journ. B **49**, 465 (2006).

[6] E. A. Kochetov, V. A. Osipov, J. Phys. A: Math. Gen. **32**, 1961, (1999).

[7] D. V. Kolesnikov, V. A. Osipov: JETP Letters **79**, 532 (2004).

[8] J. Gonzalez, F. Guinea, M. A. H. Vozmediano, Nucl. Phys. B **406**, 771 (1993).

[9] R. Jackiw, Phys. Rev. D **29**, 2375 (1984).

Circular D0L-systems, Their Critical Exponent and Factor Complexity

Štěpán Starosta*

4th year of PGS, email: stepan.starosta@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. We prove that a D0L-system is circular and non-pushy if and only if the critical exponent of its language is finite. We also explore the possible orders of growth of factor complexity of circular D0L-systems and include examples.

Keywords: circular D0L-system, critical exponent, factor complexity

Abstrakt. Ukazujeme, že D0L-systém je cirkulární a non-pushy právě tehdy, když kritický exponent jeho jazyka je konečný. Zkoumáme také možné řády růstu faktorové komplexity cirkulárních D0L-systémů a uvádíme příklady.

Klíčová slova: cirkulární D0L-systém, kritický exponent, faktorová komplexita

1 Introduction

Morphisms play an important role in combinatorics on words. A morphism is often used to define an infinite word – either as a fixed point of a morphism or as an image of another infinite word. Perhaps one of the most famous morphisms is the Thue-Morse morphism σ_{TM}, defined over the binary alphabet $\{0, 1\}$ as

$$\sigma_{TM} : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 10 \end{cases} .$$

Its fixed point

$$\mathbf{u}_{TM} = 011010011001011010\ldots$$

is the famous Thue-Morse sequence (see for instance [1] for various results).

Given a morphism σ it may not have only one fixed point. However, if σ is primitive, then all its fixed points have the same set of factors. In other words, the symbolic dynamical system associated to σ is the same (see for instance [6]). In this article, we are interested not only in primitive morphism but in a larger class: circular morphisms. Roughly speaking, a circular morphism is a morphism for which we can find preimages of

*This work has been supported by the Czech Science Foundation grant GAČR 201/09/0584, by the grants MSM6840770039 and LC06002 of the Ministry of Education, Youth, and Sports of the Czech Republic, and by the grant of the Grant Agency of the Czech Technical University in Prague grant No. SGS11/162/OHK4/3T/14.

factors which are long enough (see below for exact definition). Circular morphisms include injective primitive morphisms (see [12]). The notion of circularity has been explored for instance in [2] where an algorithm for deciding whether a language generated by a morphism avoids a certain factor is introduced, or in [7] where factor complexity of systems generated by a subclass of circular morphisms is given. Also, in [8], a method to enumerate bispecial factors for systems generated by non-pushy circular morphisms is derived. (See Definition 14 below for exact definition of non-pushy systems.)

Another notion from combinatorics on words is the critical exponent of a language. Given a language $\mathcal{L}$, its critical exponent $E_c(\mathcal{L})$ is defined as

$$E_c(\mathcal{L}) = \sup\{r \in \mathbb{Q} \mid w^r \text{ is a factor of } \mathbf{u} \text{ for some non-empty } w \in \mathcal{L}\},$$

i.e., it is the maximal (if attained) fractional power that occurs in the language $\mathcal{L}$. By critical exponent $E_c(\mathbf{u})$ of an infinite word $\mathbf{u}$ we mean the critical exponent of its set of factors.

The critical exponent of the Thue-Morse word is 2 and is attained for some factor. A more interesting example is the Fibonacci word, i.e., the fixed point of σ_F defined as

$$\sigma_F : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 1 \end{cases}.$$

Its fixed point has critical exponent $\frac{1}{2}(5+\sqrt{5})$ (see [10]) which is obviously not attained. As shown in [9], every real number greater than 1 is the critical exponent of some word. The notion is also connected to the famous Dejean's conjecture (see [4]). Let $n \geq 2$ be an integer. We define the repetition threshold $RT(n)$ as follows

$$RT(n) = \inf\{E_c(\mathbf{u}) \mid \mathbf{u} \text{ is an infinite word over } n\text{-letter alphabet}\}.$$

Dejean conjectured that

$$RT(n) = \begin{cases} 2 & \text{if } n = 2 \\ {}^{7}/_{4} & \text{if } n = 3 \\ {}^{7}/_{5} & \text{if } n = 4 \\ {}^{n}/_{n-1} & \text{otherwise} \end{cases}.$$

The conjecture is now proved, see [3] and [14].

Since in general a circular morphism can generate different languages, we use the notion of a D0L-system $S = (\mathcal{A}, \sigma, w)$ to specify the language. The language of the system S is the set of factors of words $\sigma^n(w)$ for all $n \in \mathbb{N}$ where $w \in \mathcal{A}^*$ is a non-empty word. The main result of this article is the following theorem.

Theorem 1. *Let $S = (\mathcal{A}, \sigma, w)$ be a D0L-system. Then the critical exponent of $\mathcal{L}(S)$ is finite if and only if S is circular and non-pushy.*

In the last section, we state that pushy circular D0L-systems have factor complexity in the class $\Theta(n^2)$, while non-pushy circular D0L-systems have their factor complexity in one of the following classes: $\Theta(n)$, $\Theta(n \log\log n)$, or $\Theta(n \log n)$. This result is based on the work done in [13] where factor complexities of fixed points of morphisms are explored.

2 Preliminaries

A finite set of symbols, usually called *letters*, is called an *alphabet* and denoted $\mathcal{A}$. A finite sequence $w = w_0w_1 \dots w_{n-1}$ of letters is said to be a *finite word*, its *length* $|w| = n$. The set of all finite words, including the *empty word*, denoted ε, and the operation of concatenation is the free monoid $\mathcal{A}^*$.

An infinite sequence $\mathbf{u} = (u_n)_{n=0}^{+\infty}$ of letters $u_n \in \mathcal{A}$ is an *infinite word*. A finite word $w = w_0w_1 \dots w_{n-1}$ is a *factor* of $\mathbf{u}$ if there exists an index $j \in \mathbb{N}$ such that $w_0w_1 \dots w_{n-1} = u_ju_{j+1}u_{j+n-1}$. The index j is called an *occurrence* of w in $\mathbf{u}$. The set of all factors of $\mathbf{u}$ is the *language of* $\mathbf{u}$ and is denoted by $\mathcal{L}(\mathbf{u})$.

Let z be a word (finite or infinite). If there exist words p, z', s such that $z = pz's$, then p is called a *prefix* of z and s a *suffix* of z.

Let $\mathbf{u}$ be an infinite word. If there are words p and w such that $\mathbf{u} = pwww\dots$, then $\mathbf{u}$ is said to be *eventually periodic*. If $p = \varepsilon$, then $\mathbf{u}$ is *purely periodic*. Otherwise, $\mathbf{u}$ is *aperiodic*.

We say that an infinite word is *recurrent* if any its factor occurs in it infinitely many times. It is *uniformly recurrent* if any its factor occurs with bounded gaps between successive occurrences.

The *k-power* of a finite word w is defined as $w^k = ww^{k-1}$ for $k > 0$ and $w^0 = \varepsilon$. If a finite non-empty word w can be factorized as $w = p^ke$ such that $k \geq 1$, e is a prefix of p, and $|p|$ is minimal, then w is $\frac{|w|}{|p|}$-*power* of p and we write $w = p^{\frac{|w|}{|p|}}$. For instance, we have $abbaab = (abba)^{\frac{3}{2}}$ and $starosta = (staro)^{\frac{8}{5}}$. The *critical exponent* of a language $\mathcal{L}$ is the number

$$E_c(\mathcal{L}) = \sup\{r \in \mathbb{Q} \mid w^r \in \mathcal{L} \text{ for some non-empty } w \in \mathcal{L}\}.$$

A *morphism* on a free monoid $\mathcal{A}^*$ is a mapping σ satisfying $\sigma(vw) = \sigma(v)\sigma(w)$ for all $v, w \in \mathcal{A}^*$. We say a morphism σ is *primitive* if for all $a \in A$ there exist an integer k such that $\sigma^k(a)$ contains all letters of $\mathcal{A}$. By *substitution* we understand a non-erasing morphism (for all $a \in \mathcal{A}$, $\sigma(a) \neq \varepsilon$) such that there exists a letter a and a non-empty word w such that $\sigma(a) = aw$. Note that we can find slightly different definitions of substitution in the literature. For instance, some authors define a substitution to be just a non-erasing morphism.

In what follows, we are interested in infinite words (more precisely their languages) defined as fixed points of a substitution. Given a substitution σ, one can construct its fixed point $\mathbf{u}$ as follows: $\mathbf{u} = \sigma^\omega(a) = aw\sigma(w)\sigma^2(w)\sigma^3(w)\dots$. Since there may be more fixed points, it is convenient to denote which fixed point are we working with. To specify the language generated by σ, we use the following definition of D0L-system which is a commonly used notion (see [15]).

Definition 2. A triplet $S = (\mathcal{A}, \sigma, w)$ is called D0L-system if $\mathcal{A}$ is an alphabet, σ a substitution on $\mathcal{A}$, and $w \in \mathcal{A}^*$ is a non-empty word. The language of the system $\mathcal{L}(S)$ is the set of all factors of words $\sigma^n(w)$ for all $n \in \mathbb{N}$.

We implicitly suppose that given a D0L-system $S = (\mathcal{A}, \sigma, w)$, it holds that $\mathcal{A} \subset \mathcal{L}(S)$. In other words, if the system does not contain all the letters of $\mathcal{A}$, we implicitly consider the restriction of $\mathcal{A}$ and σ to those letters it contains.

Remark 3. If σ is primitive, then it follows directly from the definition of primitivity that its associated D0L-system $(\mathcal{A}, \sigma, w)$ does not depend on the choice of w.

As already mentioned, circular systems include injective primitive morphisms.

Theorem 4 ([12]). *Any D0L-system $S = (\mathcal{A}, \sigma, w)$, with $a \in \mathcal{A}$ and σ injective and primitive, is circular.*

Given a language $\mathcal{L}$, we define its *factor complexity* as the mapping $\mathcal{C}(n)$ which associates to an integer n the number of factors of length n, i.e.,

$$\mathcal{C}(n) := \#\{w \in \mathcal{L} \mid |w| = n\}$$

for all $n \in \mathbb{N}$.

The fact that a language is given by a substitution is very important. Given a factor $w \in \mathcal{L}(S)$, there is a factor $v \in \mathcal{L}(S)$ such that w is a factor of $\sigma(v)$. If v is the shortest such factor, then it is called an *ancestor* (see [11]).

Example 5. Let φ_E be the substitution over $\{0, 1, 2\}$ defined as

$$\varphi_E : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 1012 \\ 2 \mapsto 1 \end{cases}.$$

We are interested in the system $S_E = (\{0, 1, 2\}, \varphi_E, 0)$. The fixed point $\mathbf{u}_E$ of φ_E begins with

$$\mathbf{u}_E = 011012101201101211012\ldots$$

and since φ_E is primitive, it is clear that $\mathcal{L}(S_E) = \mathcal{L}(\mathbf{u}_E)$.

The factor 01 has 2 ancestors: 0 and 1. The factor 011 has only 1 ancestor which is the factor 01.

As already mentioned, in a D0L-system every factor has at least one ancestor. To look for such factors means in fact to decompose a word w into images of letters by the substitution σ. The following notion (introduced in [2]) of synchronizing point indicates whether there is a common point for all such decompositions in a given factor.

Definition 6. Let $S = (\mathcal{A}, \sigma, w)$ be a D0L-system with σ injective and let $v \in \mathcal{L}(S)$. An ordered pair (v_1, v_2), where $v_1, v_2 \in \mathcal{L}(S)$, is called a *synchronizing point* point of v if $v = v_1 v_2$ and

$$\forall z_1, z_2 \in \mathcal{A}^*, \big(z_1 v z_2 \in \sigma(\mathcal{L}(S)) \Rightarrow z_1 v_1 \in \sigma(\mathcal{L}(S)) \text{ and } v_2 z_2 \in \sigma(\mathcal{L}(S))\big).$$

We denote this by $v = v_1|_s v_2$.

One can see that if a factor v has a unique ancestor and σ is injective, then v has a synchronizing point (provided v is long enough not to be central factor of an image of a letter). However, the notion of synchronizing point is more subtle and the converse is not true.

Example 7. Consider again the system S_E from Example 5. Its factor 101 can be decomposed as 1|01 or |101. On the other hand, the factor 0121 can be decomposed only as 012|1, i.e., it contains a common "bar" for all decompositions and thus a synchronizing point. For more examples see for instance [8].

Definition 8. A D0L-system $S = (\mathcal{A}, \sigma, w)$ is *circular* if σ is injective on $\mathcal{L}(S)$ and if there exists an integer $D \in \mathbb{N}$ such that any $v \in \mathcal{L}(S)$ longer than D has at least one synchronizing point. The integer D is called a *synchronizing delay.*

Example 9. Consider once more the system S_E from Example 5. Anytime the letter 2 occurs in a factor, we can place a synchronizing point after it. Since the substitution is primitive, its fixed point is uniformly recurrent. Therefore, the factors 2 occurs in a bounded distance and there exists synchronizing delay. Looking at its set of factors we can find the minimal synchronizing delay to be 4.

Example 10. The system $(\{0,1,2\}, \psi, 0)$ where ψ is defined as

$$\psi : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 11 \\ 2 \mapsto 02 \end{cases}$$

is not circular. One can see that the language of the system contains arbitrary powers of 1 and such factors have no synchronizing point, i.e., can be decomposed as $11|11|1\ldots$ or $1|11|1\ldots$.

The presence of arbitrary k-power of a factor in a language of a non-circular system are not a coincidence. The following result states this fact.

Theorem 11 ([11])**.** *If a D0L-system does not contain the k-power of any its factors for some $k > 0$, then it is circular.*

The following definitions and lemmas taken from [5] are used in the proof of Theorem 1.

Definition 12 ([5])**.** Let $S = (\mathcal{A}, \sigma, w)$ be a D0L-system. We say it is *strongly repetitive* if there exists a non-empty $v \in \mathcal{L}(S)$ such that $v^k \in \mathcal{L}(S)$ for all $k \in \mathbb{N}$.

Lemma 13 ([5])**.** *Given a D0L-system $S = (\mathcal{A}, \sigma, w)$, if $\mathcal{L}(S)$ contains a k-power for all $k \in \mathbb{N}$, then G is strongly repetitive.*

Definition 14 ([5])**.** Let $S = (\mathcal{A}, \sigma, w)$ be a D0L-system. A letter $b \in \mathcal{A}$ has *rank zero* if $\mathcal{L}((\mathcal{A}, \sigma, b))$ is finite.

S is *pushy* if for all $n \in \mathbb{N}$ there exits $v \in \mathcal{L}(S)$ of length n which is composed of letters having rank zero. Otherwise S is *non-pushy.*

Remark 15. One can see that if a system $S = (\mathcal{A}, \sigma, w)$ is pushy, then there exists a letter $a \in A$ such that $|\sigma(a)| = 1$.

As illustrated by the following example, the converse to the last remark is not true.

Example 16. Let σ_F be the Fibonacci substitution defined above. Since both $\{\sigma_F^n(0)\}$ and $\{\sigma_F^n(1)\}$ are infinite, the system $(\{0,1\}, \sigma_F, 0)$ is non-pushy.

Moreover, for pushy systems, the following lemma holds.

Lemma 17 ([5])**.** *If a D0L-system $S = (\mathcal{A}, \sigma, w)$ is pushy, then it is strongly repetitive.*

3 Proof of Theorem 1

Proof of Theorem 1. ($\Rightarrow$):
Circularity follows from Theorem 11. S being pushy is in contradiction with Lemma 17, thus, it is non-pushy.

($\Leftarrow$):
Suppose the critical exponent of $\mathcal{L}(S)$ is infinite. According to Lemma 13, there exists a non-empty factor $v \in \mathcal{L}(S)$ such that for all $n \in \mathbb{N}$, $v^n \in \mathcal{L}(S)$. Take the shortest factor v having such property. Since S is circular, there exists a finite synchronizing delay D. Take $N \in \mathbb{N}$ such that $|v^N| \geq D$. Then v^N contains a synchronizing point, i.e., $v^N = v_1|_s v_2$. It is clear that v^{N+1} contains at least two synchronizing points, i.e., $v^{N+1} = v_1|_s v_2 v = v v_1|_s v_2$. In general, v^{N+k} contains $k+1$ synchronizing points at fixed distances equal to $|v|$. Since σ is injective, it implies that there exists a unique $z \in \mathcal{L}(S)$ such that $v^{N+k} = p\sigma(z^k)s$ (for some factors p and s) and $z^k \in \mathcal{L}(S)$ for all $k \geq 0$. According to the choice of v, it is clear that $|\sigma(z)| = |z| = |v|$. Denote by $\mathcal{L}_1(z)$ the set of letters occurring in z. It is clear that $\sigma(\mathcal{L}_1(z)) = \mathcal{L}_1(v)$ and $\forall a \in \mathcal{L}_1(z)$ we have $|\sigma(a)| = 1$.

We can now repeat the process: take the factor z to play the role of factor v. Thus, we can find an infinite sequence of factors $z_0 = z, z_1, z_2 \ldots$ such that $\sigma(\mathcal{L}_1(z_{k+1})) = \mathcal{L}_1(z_k)$ and $|z_k| = |z|$ for all $k \geq 0$. Since $\mathcal{A}$ is finite, it is clear that there exists integers $m \neq \ell$ such that $\mathcal{L}_1(z_m) = \mathcal{L}_1(z_\ell)$. This implies that for all k the factor z_k is composed of letters of rank zero. This is a contradiction with S being non-pushy. □

The following claim illustrates that we cannot hope to have infinite critical exponent with a primitive substitution.

Claim 18. *Let σ be a primitive substitution and let $w \in A^*$ be a non-empty word. Then $S = (\mathcal{A}, \sigma, w)$ is non-pushy.*

Proof. Since $\mathcal{L}(S)$ is infinite, according to Remark 3, there are no letters of rank zero. Thus, the system is non-pushy. □

According to Lemma 17, a pushy system has always critical exponent infinite. The next example illustrates the last case of a D0L-system having infinite critical exponent: a system generated by a non-circular substitution.

Example 19. The system $(\{0,1,2\}, \psi, 0)$ where ψ is defined as

$$\psi : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 22 \\ 2 \mapsto 11 \end{cases}$$

is non-pushy and not circular. Its critical exponent is clearly infinite.

4 Factor complexity of circular D0L-systems

In this section, we use the results of [13] to explore possible orders of growth of factor complexities of circular D0L-systems.

Definition 20 ([13])**.** A substitution σ is *growing* if for all letters $a \in \mathcal{A}$ we have $|\sigma(a)| > 1$. Otherwise it is *non-growing*.

In the next definition, by the order of growth of a letter a we mean the order of growth of the sequence $(|\sigma^n(a)|)_{n=0}^{+\infty}$.

Definition 21 ([13])**.** A substitution σ is said to be *quasi-uniform* if for all letters have the same order of growth λ^n. It is said to be *polynomially divergent* if every letter a has its order of growth $n^{e_a}\lambda^n$ with $\lambda > 1$ and e_a non zero. Finally, it is *exponentially divergent* if there exist two letters a and b such that their order of growth is $n^{e_a}\lambda_a^n$ and $n^{e_b}\lambda_b^n$ with $1 < \lambda_a < \lambda_b$ and $\lambda_c > 1$ for all $c \in \mathcal{A}$.

Theorem 22 ([13])**.** *Let $\mathbf{u} = \sigma^\omega(a)$ be an aperiodic infinite word such that σ is a growing substitution. If σ is respectively quasi-uniform, polynomially divergent or exponentially divergent, the factor complexity of $\mathcal{L}(\mathbf{u})$ is respectively $\Theta(n)$, $\Theta(n \log\log n)$ or $\Theta(n \log n)$.*

Corollary 23. *Let $S = (\mathcal{A}, \sigma, w)$ be circular and non-pushy. Then its factor complexity $\mathcal{C}(n)$ is $\Theta(n)$, $\Theta(n \log\log n)$ or $\Theta(n \log n)$.*

Proof. Suppose σ is not growing. One can see that since S is non-pushy, we can find an integer $k > 0$ such that σ^k is growing and $\mathcal{L}(S) = \mathcal{L}(S')$ where $S' = (\mathcal{A}, \sigma^k, w)$. Thus, we may suppose σ is growing without loss of generality. For all $a \in \mathcal{A}$, the circularity of $S'' = (\mathcal{A}, \sigma, a)$ implies that there is no factor p such that $\mathcal{L}(p^\omega) = \mathcal{L}(S'')$ and thus the factor complexity of $\mathcal{L}(S'')$ is unbounded. (Since $\sigma^\omega(a)$ may not be well defined, we need this to replace the aperiodicity assumption in order to be able to apply Theorem 22.) To finish the proof it remains to see if σ is quasi-uniform, polynomially divergent, or exponentially divergent and apply Theorem 22. □

The systems generated by a primitive substitution have their factor complexity of order $\Theta(n)$. The next two examples illustrate the last two cases.

Example 24. Let Ψ be the following substitution

$$\Psi : \begin{cases} 0 \mapsto 012340 \\ 1 \mapsto 21112 \\ 2 \mapsto 12221 \\ 3 \mapsto 344443 \\ 4 \mapsto 433334 \end{cases}$$

is exponentially divergent ($|\Psi^n(1)| = |\Psi^n(2)| = 5^n$ and $|\Psi^n(3)| = |\Psi^n(4)| = 6^n$), the D0L-system $(\{0,1,2,3,4\}, \Psi, 0)$ is circular and non-pushy. Therefore, the factor complexity of this system is $\Theta(n \log n)$.

Example 25. Let Ψ be the following substitution

$$\Psi : \begin{cases} 0 \mapsto 0101010101 \\ 1 \mapsto 02220 \\ 2 \mapsto 20002 \end{cases}$$

is polynomially divergent ($|\Psi^n(1)| = |\Psi^n(2)| = 5^n$ and $|\Psi^n(0)| = (n+1)5^n$), the D0L-system $(\{0,1,2,3,4\}, \Psi, 0)$ is circular and non-pushy. Therefore, the factor complexity of this system is $\Theta(n \log\log n)$.

It remains to deal with push circular D0L-systems. The following theorem is rewritten in our terminology.

Theorem 26 ([13]). *Let* $\mathbf{u} = \sigma^\omega(a)$ *be an aperiodic infinite word such that* σ *is a non-growing substitution. If* $\mathcal{L}(\mathbf{u})$ *contains factors composed of letters of rank zero of arbitrary length, then its factor complexity satisfies*

$$c_1 n^2 \leq \mathcal{C}(n) \leq c_2 n^2,$$

with $c_1 > 0$.

The following corollary is an application of the last theorem to our setting.

Corollary 27. *Let* $S = (\mathcal{A}, \sigma, w)$ *be circular and pushy. Then its factor complexity satisfies* $\mathcal{C}(n) = \Theta(n^2)$.

Proof. Since S is pushy, there exists a letter $a \in \mathcal{A}$ such that $S' = (\mathcal{A}, \sigma, a)$ is pushy. According to Lemma 17, S' is strongly repetitive, i.e., there exists a factor $v \in \mathcal{L}(S')$ such that $v^k \in \mathcal{L}(S')$ for any $k > 0$. We can now use a similar argument as in the proof of Theorem 1 to see that v can be chosen to be composed of letters of rank zero. According to Remark 15, σ is non-growing.

Suppose that the factor complexity of S' is bounded. In other words, there exists a word p such that $\mathcal{L}(S') = \mathcal{L}(p^\omega)$. However, this is in contradiction with circularity of S' and $\mathcal{L}(S')$ is unbounded.

Thus, applying Theorem 26, one can see that the factor complexity of S' is quadratic. Finally, since for every letter $b \in \mathcal{A}$ the system $(\mathcal{A}, \sigma, b)$ has at most quadratic complexity (using also Corollary 23), it follows that the factor complexity $\mathcal{C}(n)$ of S is $\Theta(n^2)$. □

Let us illustrate this case by an example.

Example 28. Let σ_P be defined as

$$\sigma_P : \begin{cases} 0 \mapsto 001 \\ 1 \mapsto 1 \end{cases}$$

The system $S = (\{0,1\}, \sigma_P, 0)$ is circular and pushy. It is easy to see that its factor complexity is quadratic.

Acknowledgements

The author thanks to Karel Klouda and Julien Leroy for their fruitful remarks on the topic.

References

[1] J.-P. Allouche and J. Shallit. *The ubiquitous Prouhet-Thue-Morse sequence.* In 'Sequences and their applications, Proceedings of SETA'98', 1–16. Springer, (1999).

[2] J. Cassaigne. *An algorithm to test if a given circular hdol-language avoids a pattern.* In 'IFIP World Computer Congress'94', 459–464. Elsevier, (1994).

[3] J. D. Currie and N. Rampersad. *A proof of Dejean's conjecture.* preprint: http://arxiv.org/abs/0905.1129 (2009).

[4] F. Dejean. *Sur un théorème de Thue.* J. Comb . Theory A **13** (1972), 90–99.

[5] A. Ehrenfeucht and G. Rozenberg. *Repetition of subwords in D0L languages.* Inform. Comput. **53** (1983), 13–35.

[6] N. P. Fogg. *Substitutions in Arithmetics, Dynamics and Combinatorics.* Springer, 1st edition, (2002).

[7] A. E. Frid. *On uniform dol words.* In 'STACS'98, LNCS 1373', 544–554. Springer, (1998).

[8] K. Klouda. *Non-standard numerations systems and combinatorics on words.* PhD thesis, Czech Technical University in Prague, (2010).

[9] D. Krieger and J. Shallit. *Every real number greater than 1 is a critical exponent.* Theoret. Comput. Sci. **381** (2007), 177–182.

[10] F. Mignosi and G. Pirillo. *Repetitions in the Fibonacci in?nite word.* RAIRO Info. Theor. Appl. **26** (1992), 199–204.

[11] F. Mignosi and P. Séébold. *If a d0l language is k-power free the it is circular.* In 'ICALP '93: Proceedings of the 20th International Colloquim on Automata, Languages and Programming', 507–518, London, UK, (1993). Spriger-Verlag.

[12] B. Mossé. *Reconnaissabilité des substitutions et complexité des suites automatiques.* Bull. Soc. Math. Fr. **124** (1996), 329–346.

[13] J.-J. Pansiot. *Complexité des facteurs des mots infinis engendrés par morphismes itérés.* In '11th ICALP, Antwerpen', J. Paredaens, (ed.), volume 172 of *LNCS*, 380–389. Springer, (Jul 1984).

[14] M. Rao. *Last cases of Dejean's conjecture.* Theoret. Comput. Sci. **412** (2011), 3010–3018.

[15] G. Rozenberg and A. Salomaa. *The mathematical theory of L systems.* Academic Press, (1980).

Transportní jevy ve vodíkových palivových článcích*

Lucie Strmisková

2. ročník PGS, email: lucka.strmiskova@seznam.cz
Katedra fyziky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: František Maršík, Ústav termomechaniky, AV ČR, v.v.i.

Abstract. We explain the basic operation of the hydrogen fuel cell. Then we compute the voltage generated by fuel cell and its efficiency. We will present the major causes of voltage drop. Then we describe the most important transport phenomena in fuel cells. We are especially interested in water transport.

Keywords: hydrogen fuel cell, overvoltage, water transport

Abstrakt. V článku je vysvětlen princip fungování vodíkových palivových článků. Dále je vypočítáno napětí generované palivovým článkem a jeho účinnost. Jsou rozebrány nejčastější příčiny napěťových ztrát. Dále jsou popsány nejvýznamnější transportní jevy probíhající v palivovém článku, přičemž zvláštní pozornost je věnována pohybu vody.

Klíčová slova: vodíkový palivový článek, napěťové ztráty, transport vody

1 Úvod

Počet lidí na planetě neustále vzrůstá, stejně tak se zvyšuje i jejich životní úroveň. Vysoká životní úroveň spojená s možností cestování na velké vzdálenosti s sebou nese větší spotřebu elektrické energie a ropy. Tato enormní spotřeba vede k rychlému vyčerpávání zdrojů fosilních paliv. Obrovské množství aut také produkuje velké objemy výfukových plynů, které nenávratně ničí naše životní prostředí. V poslední době se proto usilovně hledá řešení, jak nahradit fosilní paliva jinými zdroji energie a jak napravit poškozené životní prostředí.

Jedním z možných řešení této situace je používání aut s vodíkovými palivovými články namísto aut s klasickými spalovacími motory.

Základní princip fungování vodíkového palivového článku je velmi jednoduchý. Jedná se o obrácenou elektrolýzu vody - vodík se slučuje s kyslíkem za vzniku elektrického proudu a vody. Protože voda je jediným vedlejším produktem tohoto procesu, jedná se o velmi ekologický způsob získávání elektrické energie.

Navíc vodíkové palivové články pracují velmi tiše a vyznačují se vysokou účinností. Z těchto důvodů je vodíkovým technologiím v současné době věnována značná pozornost. Ve vyspělých zemích se uvažuje o přechodu na vodíkovou ekonomiku, jež by znamenala podstatné snížení emisí skleníkových plynů a látek znečisťujících ovzduší. Pokud by byl

*Tato práce byla podpořena grantem číslo GD202/08/H072 financovaným Grantovou agenturou České republiky

vodík vyráběn z obnovitelných zdrojů, byly by škodlivé emise v podstatě nulové. Ale i při výrobě z fosilních paliv by došlo ke značnému zlepšení ovzduší ve velkých aglomeracích přesunutím exhalací do výroben vodíku.

Na následujících stránkách se proto budu věnovat vodíkovým palivovým článkům. Nejdříve popíšu princip fungování vodíkového palivového článku s polymerní elektrolytickou membránou, pak se budu zabývat jeho účinností. Nakonec popíšu nejdůležitější transportní jevy probíhající v membráně, přičemž zvláštní pozornost bude věnována pohybu vody.

2 Vodíkový palivový článek

První vodíkový palivový článek sestrojil William Grove už v roce 1839. Vznikající proud byl však velmi malý. Jednak proto, že přivedený plyn, tyčová elektroda a elektrolyt se stýkali na velmi malé ploše. Navíc velká vzdálenost mezi elektrodami způsobovala značné ohmické ztráty.

Aby se zvýšilo množství elektrického proudu, elektrody se dnes obvykle vyrábí ploché s tenkou vrstvičkou elektrolytu mezi nimi. Další věc, která limituje množství vznikajícího elektrického proudu, je pomalý průběh chemických reakcí na površích elektrod. Vodíkové palivové články totiž pracují při nižších teplotách, což právě vede k pomalému průběhu reakce. Rychlost reakce se běžně zvyšuje užitím katalyzátorů a zvýšením teploty, za které chemická reakce probíhá. U vodíkových palivových článků se užívá jako katalyzátor platina a i když se jedná o velmi drahý kov, v článku je ho použito tak malé množství, že ve srovnání s cenou zbylých komponent je jeho cena zanedbatelná.

Protože vzniklé elektrony musí být odvedeny pryč, musí reakce probíhat na povrchu elektrody. Je jasné, že čím větší povrch bude elektroda mít, tím rychleji bude reakce probíhat. Pro zvýšení jejich povrchu se elektrody vyrábí velmi porézní. Mikrostruktura dnes užívaných elektrod zvyšuje jejich účinný povrch o 2 až 3 řády.

Palivový článek je elektrochemické zařízení, které spotřebovává palivo a oxidant a převádí je na vodu a elektrickou energii. Základní princip vodíkového palivového článku může být popsán takto.

Na anodu je přiveden plynný vodík. Ten zde ionizuje.

$$2H_2 \rightarrow 4H^+ + 4e^- \tag{1}$$

Tato reakce je exotermická, ale nazačne samovolně, vždy je potřeba určité množství aktivační energie k jejímu rozběhnutí. Vzniklé protony prochází elektrolytem ke katodě, kde reagují s kyslíkem, který je na katodu obvykle vháněn jako součást vzduchu, a elektrony z elektrody za vzniku vody.

$$O_2 + 4e^- + 4H^+ \rightarrow 2H_2O \tag{2}$$

Pro správné fungování vodíkového palivového článku je nutné, aby tyto reakce probíhaly spojitě. Elektrony vzniklé na anodě procházejí vnějším elektrickým obvodem ke katodě a protony procházejí ke katodě přes elektrolyt. Elektrolyt proto musí být vyroben z takového materiálu, který se vyznačuje vysokou protonovou vodivostí a zároveň neumožňuje

průchod elektronů. Pokud by elektrolytem procházely i elektrony, nedostali bychom požadovaný elektrický proud.

Naše skupina se zabývá vodíkovými palivovými články, kde je elektrolytem pevný polymer na bázi polyfluorethylenu, známý pod obchodní značkou Nafion. Základem tohoto materiálu je tetrafluorethylen, známý také pod názvem Teflon. Tetrafluorethylenu projde polymerizací a poté jsou k němu připojeny boční řetízky obsahující kyselinu siřičnou HSO_3. Mezi H^+ a SO_3^- je iontová vazba. Ionty H^+ a SO_3^- z různých bočních řetízků jsou k sobě velmi silně přitahovány a tvoří shluky. Kyselina siřičná je silně hydrofilní, zatímco polytetrafluorethylen je naopak silně hydrofobní. Tyto hydrofilní shluky řetízků uvnitř hydrofobního materiálu k sobě vážou molekuly vody a tvoří tak jakýsi kanálek, který umožňuje hladký průchod protonům.

Je jasné, že čím více vody tyto shluky bočních řetízků obsahují, tím lepší protonovou vodivostí se elektrolyt vyznačuje.

Kromě vysoké protonové vodivosti je Nafion velmi odolný jak vůči chemickým, tak mechanickým vlivům. Jeho mechanická pevnost umožňuje vyrobit velmi tenkou vrstvičku elektrolytu až do hodnoty $50\mu m$, což podstatně snižuje jeho elektrický odpor.

3 Napětí generované palivovým článkem

Velmi jednoduchou úvahou určíme teoretickou hodnotu napětí generovaného palivovým článkem. Z rovnic (1), (2) je zřejmé, že na 1 mol vodíku připadá $2N_A$ elektronů jdoucích vnějším obvodem, kde N_A je Avogadrovo číslo. Vzniklý náboj je tedy

$$Q = -2N_A e = -2F,$$

kde $F = 96485\mathrm{C.mol}^{-1}$ je Faradayova konstanta.

Práce vykonaná vnějším elektrickým polem je

$$W = QU.$$

Z klasické termodynamiky je známo, že maximální možná práce vykonaná systémem je rovna změně jeho Gibbsovy volné energie

$$W = \Delta g = -2FU.$$

Změna Gibbsovy energie Δg je rovna rozdílu Gibbsových energií produktů a reaktantů a v našem případě je výhodné vztáhnout její množství na 1 mol látky, což právě zdůrazňujeme malým písmenem.

Teoretická hodnota napětí (označíme ji indexem 0) generovaného vodíkovým palivovým článkem je tedy

$$U^0 = -\frac{\Delta g}{2F}. \tag{3}$$

Změna Gibbsovy energie závisí na hodnotách stavových veličin - tlaku a teploty. Při teplotě $t = 25°C$ a standardním tlaku $p = 100kPa$ vzniká voda v kapalné fázi a teoretická hodnota napětí je $U^0 = 1,23V$.

Z klasické termodynamiky je známo, že pro naši chemickou reakci

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O$$

se Gibbsova energie mění s tlakem podle vztahu

$$\Delta g = \Delta g^0 - RT \ln \Big(\frac{a_{H_2} a_{0_2}^{\frac{1}{2}}}{a_{H_2O}} \Big), \tag{4}$$

kde a jsou aktivity jednotlivých složek a Δg^0 značí změnu Gibbsovy energie za standardního tlaku.

Potom můžeme teoretické napětí palivového článku psát ve tvaru

$$U = U^0 + RT \ln \Big(\frac{a_{H_2} a_{0_2}^{\frac{1}{2}}}{a_{H_2O}} \Big), \tag{5}$$

kde indexem 0 u hodnoty napětí opět označujeme, že se jedná o hodnotu za standarního tlaku. Rovnice (5) se nazývá Nernstova rovnice.

Předpokládáme-li, že palivový článek pracuje při vyšších teplotách, voda vzniká ve formě páry a je rozumné předpokládat, že chování reaktantů i produktů bude s dobrou přesností odpovídat chování ideálního plynu. Pro ideální plyn platí, že jeho aktivita je rovna $a = \frac{p}{p_0}$, kde p_0 je hodnota standardního tlaku. Proto platí

$$a_{H_2} = \frac{p_{H_2}}{p_0}, \quad a_{O_2} = \frac{p_{O_2}}{p_0}, \quad a_{H_2O} = \frac{p_{H_2O}}{p_0}$$

a dosazujeme-li tlaky v barech, můžeme psát Nernstovu rovnici ve tvaru

$$U = U^0 + RT \ln \Big(\frac{p_{H_2} p_{0_2}^{\frac{1}{2}}}{p_{H_2O}} \Big). \tag{6}$$

Změna Gibbsovy energie Δg s tlakem je velmi důležitá. Jak už bylo řečeno, na katodu je vháněn kyslík jako součást vzduchu. Kyslík je během chemické reakce spotřebováván, tzn. klesá jeho parciální tlak. Naopak parciální tlak vody během reakce roste, protože je reakcí produkována.

K určení závislosti napětí generovaného palivovým článkem na teplotě, využijeme termodynamickou rovnost

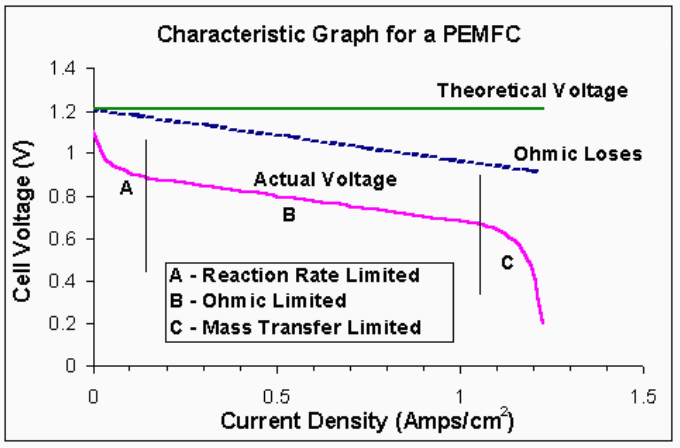
$$\Big(\frac{\partial G}{\partial T} \Big)_p = -S.$$

Rovnici můžeme přepsat do tvaru

$$\Big(\frac{\partial \Delta g}{\partial T} \Big)_p = -\Delta s,$$

odkud plyne

$$U = U^0 + \frac{\Delta s}{2F}(T - T^0), \tag{7}$$



Obrázek 1: Napětí generované palivovým článkem. [5]

kde U^0 je napětí generované článkem za standardní teploty a Δs značí změnu entropie 1 molu látky. S velkou přesností se dá předpokládat, že se entropie s teplotou nemění a její změna závisí především na změně látkového množství plynu během reakce.

$$\Delta s \sim \Delta n_{\text{plyn}} = \Sigma n_{\text{plynprod}} - n_{\text{plynreak}},$$

což v našem případě dává hodnotu $\Delta s \sim 1-1,5 = -0,5$. Ve vodíkovém palivovém článku tedy napětí se vzrůstající teplotou klesá.

4 Napěťové ztráty

Určili jsme, že teoretická hodnota napětí U^0 se dá spočítat jednoduchou formulí (3). Na obrázku můžeme vidět typický průběh napětí generovaného nízkoteplotním (ke kterým patří i vodíkový) palivovým článkem. Z obrázku je zřejmé, že dokonce i napětí naprázdno je podstatně menší než námi vypočtená hodnota. Navíc čím větší odebíráme proud, tím menší napětí jsme schopni získat. Na následujících řádcích popíšu hlavní příčiny napěťových ztrát [3].

Aktivační ztráty jsou způsobeny pomalým průběhem reakce na površích elektrod. Z experimentů vyplynulo, že pro všechny typy palivových článků můžeme aktivační ztráty popsat jednoduchou empirickou rovnicí

$$\Delta U_{akt} = A \ln\left(\frac{i}{i_0}\right), \quad i > i_0.$$

Konstanta A je tím vyšší, čím pomalejší je průběh reakce. Naopak i_0 s rychlejším průběhem reakce vzrůstá, přičemž narůst způsobený i_0 převýší pokles způsobený konstantou A.

Konstanta i_0 má jednoduchý fyzikální význam. I když se nám zdá, že reakce

$$O_2 + 4e^- + 4H^+ \rightarrow 2H_2O$$

neprobíhá, skutečnost je taková, že tato reakce probíhá, ale oběma směry stejně rychle. Existuje tedy neustálý tok elektronů, proudová hustota i_0 udává právě hodnotu tohoto toku a nazývá se proto výměnný proud.

Chceme-li tedy snížit aktivační ztráty, musíme zvýšit výměnný proud. Jedním z možných způsobů je zvýšení teploty, za které chemické reakce v palivovém článku probíhají. Skutečně je pozorováno, že u vysokoteplotních palivových článků je napěťový skok způsobený aktivačními ztrátami mnohem menší než u článků nízkoteplotních. Další způsob, jak zvýšit hodnotu i_0, je užít katalyzátor a zvětšit povrch elektrod.

Vnitřní proud a průchod paliva elektrolytem také způsobují pokles napětí. Jak bylo řečeno na začátku, od elektrolytu požadujeme, aby byl dobrým protonovým vodičem, zároveň ale neumožnil průchod elektronům. V praxi ale vždy nepatrné množství elektronů elektrolytem projde. Tento proud elektronů nazýváme proudem vnitřním. Ještě významnější než ztráty způsobené vnitřním proudem je nevyužité palivo procházející elektrolytem. Stává se, že malá část vodíku na anodě nestihne zreagovat, ale difunduje ke katodě, kde zreaguje s kyslíkem přímo, bez zisku elektrického proudu. Můžeme si představit vnitřní proud jako proud, který se přidá k odebíranému proudu a způsobuje tak aktivační ztráty. Spojíme-li aktivační ztráty a vnitřní proud do jedné rovnice, získáme pro změnu napětí výraz

$$\Delta U_{akt} = A \ln\left(\frac{i + i_n}{i_0}\right).$$

Ohmické ztráty jsou způsobeny z menší části elektrickým odporem elektrod, z větší pak odporem elektrolytu. Jsou popsány Ohmovým zákonem

$$U = ir,$$

kde i je proudová hustota a r plošný elektrický odpor vyjádřený v jednotkách Ωm^2.

Ohmické ztráty je možné snížit užitím elektrod s co nejlepší vodivostí a také, což je u palivových článků běžné, mít mezi elektrodami co nejtenší vrstvičku elektrolytu. Problém je, že tato vrstvička nesmí být tenká moc, protože jinak by mohlo dojít mezi elektrodami ke zkratu.

Koncentrační ztráty Na katodu je vháněn vzduch a je z něj spotřebováván kyslík. Tím klesá jeho parciální tlak a podle Nernstovy rovnice dochází k poklesu napětí. Velikost těchto ztrát závisí na velikosti odebíraného proudu a také na tom, jak rychle vzduch v článku cirkuluje, tj. jak rychle může být spotřebovaný kyslík nahrazen novým.

Stejná situace nastává na anodě, kdy dochází ke snížení parciálního tlaku vodíku.

Bohužel neexistuje fyzikální teorie, která by byla univerzálně použitelná pro všechny typy palivových článků. Proto se užívá empirického vzorce

$$\Delta U_{kon} = m \exp(ni),$$

kde i je proudová hustota a konstanty m, n jdou zvolit tak, aby odpovídaly experimentálně naměřeným hodnotám.

S přihlédnutím k předchozím rovnicím můžeme tedy napětí generované palivovým článkem napsat ve tvaru

$$U = U^0 - ir - A \ln\left(\frac{i + i_n}{i_0}\right) + m \exp(ni). \tag{8}$$

5 Účinnost palivového článku

Je několik způsobů, jak definovat účinnost palivového článku. My budeme mluvit o dvou z nich - o účinnosti termodynamické a účinnosti napěťové. Protože palivové články využívají materiály, které se obvykle pro získání energie spalují, je vhodné srovnat produkovanou elektrickou energii s teplem, které bychom získali spálením vodíku, tj. se změnou entalpie ΔH.

Termodynamická účinnost je tedy definována jako

$$\eta_t = \frac{W}{\Delta H}. \tag{9}$$

Z klasické termodynamiky je známo, že produkovaná elektrická energie W může být maximálně rovna změně Gibbsovy energie ΔG. Proto je maximální možná termodynamická účinnost rovna

$$\eta_{t,max} = \frac{\Delta G}{\Delta H} = 0,83$$

pro reakci probíhající za standardního tlaku a teploty.

Napěťová účinnost $\eta = \frac{W}{\Delta G}$ srovnává produkovanou elektrickou energii s maximální možnou.

Gibbsova energie je spojena s entalpií termodynamickým vztahem

$$G = H - TS. \tag{10}$$

Definujeme-li časovou změnu entalpie jako $\dot{H} = \frac{\Delta W}{\Delta t}$ a výkon jako $\dot{W} = \frac{\Delta W}{\Delta t}$, můžeme psát termodynamickou účinnost ve tvaru

$$\eta_t = \frac{\dot{W}}{\dot{H}} = \frac{\dot{W}}{\dot{G} + T\dot{S}}, \tag{11}$$

kde výkon elektrického pole je ve tvaru

$$\dot{W} = -\vec{j}_{H^+}.F\nabla\phi \tag{12}$$

a protože se jedná o disipativní proces, je produkce entropie kladná.

Z termodynamiky je známo, že produkce entropie má obecně tvar součinu zobecněných toků a jim příslušejících zobecněných sil.

$$\dot{S} = -\vec{j}_{H_2O}.\nabla\Big(\frac{\mu_{H_2O}}{T}\Big) - \vec{j}_{H^+}\nabla\Big(\frac{F\phi}{T}\Big). \tag{13}$$

Díky tomu, že je membrána tak tenká, můžeme předpokládat, že její teplota je konstantní.

Teď využijeme standardního postupu lineární nerovnovážné termodynamiky a napíšeme toky vody a protonů jako

$$\vec{j}_{H_2O} = -L_{11}\nabla\mu_{H_2O} - L_{12}\nabla(F\phi), \tag{14}$$

$$\vec{j}_{H^+} = -L_{21}\nabla\mu_{H_2O} - L_{22}\nabla(F\phi). \tag{15}$$

Fenomenologické koeficienty L_{ij} závisí na stavových proměnných systému, ale nezávisí na tocích, ani zobecněných silách. O jejich fyzikálním významu promluvíme v poslední části.

Postupem z [4] vypočítáme termodynamickou účinnost. Její obrácenou hodnotu můžeme napsat ve tvaru

$$\frac{1}{\eta_t} = \frac{\dot{G}}{\dot{W}} + \frac{T\dot{S}}{\dot{W}} = \frac{1}{\eta} + \frac{1}{\varepsilon}, \tag{16}$$

kde ε označuje relativní disipaci definovanou jako

$$\varepsilon = \frac{\dot{W}}{T\dot{S}} = \frac{\vec{j}_{H^+} F \nabla\phi}{\vec{j}_{H^+} F \nabla\phi + \vec{j}_{H_2O} \nabla\mu_{H_2O}}. \tag{17}$$

$$\eta_t = \frac{W}{\Delta G}\frac{\Delta G}{\Delta H} = \eta\frac{\Delta G}{\Delta H} \tag{18}$$

$$\eta = \frac{\Delta H - \Delta G}{\Delta G}\varepsilon \quad \eta^0 = 0,205\varepsilon \tag{19}$$

Hodnota účinnosti η^0 je opět vyjádřena za standardního tlaku a teploty. Je tedy jasné, že pokud chceme zvýšit účinnost palivového článku, musíme zvýšit hodnotu relativní disipace ε. Dosadíme-li do rovnice (17) hodnoty toků ve tvaru (14), (15) a zavedeme-li nové proměnné

$$y = \sqrt{\frac{L_{11}}{L_{22}}}\left|\frac{\nabla\mu_{H_2O}}{F\nabla\phi}\right|, \quad q = \frac{L_{21}}{\sqrt{L_{11}L_{22}}}, \tag{20}$$

relativní disipace získá tvar

$$\varepsilon = \frac{1 + qy}{1 + 2qy + y^2} \tag{21}$$

Podmínka maxima funkce $\frac{d\varepsilon}{dy} = 0$ vede na kvadratickou rovnici $qy^2 + 2y + q = 0$. Reálným podmínkám odpovídá kořen

$$y = \frac{-1 + \sqrt{1 - q^2}}{q}. \tag{22}$$

6 Transport vody membránou

Z předchozího popisu vodíkového palivového článku je jasné, že aby byla polymerní elektrolytická membrána dobrým protonovým vodičem, musí být dostatečně hydratovaná. Na druhou stranu nesmí být hydratovaná moc, došlo by totiž k zaplavení elektrod.

Ještě jednou popíšeme nejvýznamnější transportní procesy v membráně. Protony se tvoří na anodě a prochází elektrolytickou membránou ke katodě. Protonová vodivost je úměrná obsahu vody v elektrolytu. Voda se tvoří na katodě a difunduje směrem k anodě.

Protože membrána je velmi tenká, při troše snahy může být dosaženo vhodné hydratace celé membrány. Existuje ale několik komplikací. Jedním z nich je tzv. elektroosmotické strhávání. Protony pohybující se od anody ke katodě s sebou strhnou molekuly vody (jeden proton dokáže strhnout až 5 molekul vody). To znamená, že obzvláště při velkých hustotách odebíraného proudu, může být část elektrolytu blíž anodě zcela vysušená, i když katoda sama o sobě je hydratovaná dostatečně.

Ještě větším problémem než elektroosmotické strhávání je vysoušení vzduchem, které při vysokých teplotách nastává. Proto je běžné zvlhčovat vzduch před tím, než vstoupí do palivového článku.

Nyní se podíváme na fyzikální význam fenomenologických koeficientů L_{ij}, které se objevily v rovnicích (14), (15). Je jasné, že v membráně probíhá difúze vody a pohyb protonů. Difúzi můžeme popsat Fickovým zákonem a pohyb protonů zase zákonem Ohmovým.

$$\vec{j}_{H_2O} = -D_{H_2O}\nabla c_{H_2O} - L_1\nabla\phi \tag{23}$$

$$\vec{j}_{H^+} = -\frac{\sigma}{F}\nabla\phi - L_2\nabla c_{H_2O} \tag{24}$$

D_{H_2O} je difúzní koeficient a σ vodivost membrány. Jak už bylo řečeno, pohybující se protony s sebou strhávají molekuly vody a naopak molekuly vody strhávají při svém pohybu protony. Proto předchozí rovnice obsahují i křížové členy. Z předchozího popisu je jasné, že koeficient L_1 můžeme napsat jako $L_1 = n_d\frac{\sigma}{F}$, kde n_d je tzv. koeficient elektroosmotického strhávání a jeho hodnota udává, kolik molekul vody je strženo jedním protonem.

7 Závěr

Cílem mojí práce bylo seznámit se s fungováním vodíkového palivového článku s polymerní elektrolytickou membránou. Pochopila jsem, na jakém principu palivový článek funguje a zajímala jsem se o transportní jevy probíhající v membráně. Tato práce má sloužit jako podklad pro tvorbu numerického modelu, který by byl schopný popsat transportní jevy probíhající v membráně.

Literatura

[1] Y. Demirel. *Nonequilibrium Thermodynamics: Transport and Rate Processes in Physical and Biological Systems.* Elsevier Science, 2002

[2] J. Kvasnica. *Termodynamika.* Státní nakladatelství technické literatury. 1965

[3] J. Larminie. A. Dicks. *Fuel cells systems explained.* JOhn Willey and Sons Ltd., 2003

[4] T. Němec. F. Maršík. O. Mičan *The characteristic thickness of polymer electrolyte membrane and efficiency of fuel cell.* Heat Transfer Engineering, 30(7), 574-581, 2009

[5] http://www.ecocar.mek.dtu.dk/Innovator/Fuel%20cell.aspx

Hadamard Type Infinite Products for Regularized Characteristic Function of Jacobi Operator*

František Štampach

2nd 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 function called $\mathfrak{F}$ defined on a subspace of the space of complex sequences is introduced with its main algebraic properties. With the aid of $\mathfrak{F}$, the characteristic function is defined for a Jacobi matrix from a certain class. Similarly as in the case of finite matrices, zeros of the characteristic function coincide with eigenvalues of the respective Jacobi operator. Further, I restrict myself on a more special subclass of investigated Jacobi matrices and derive a formula for the regularized characteristic function in terms of the infinite product of Hadamard type which forms the major part of the paper.

Keywords: Jacobi matrix, characteristic function, Hadamard factorization

Abstrakt. Na podmnožine prostoru komplexních posloupností definuji funkci $\mathfrak{F}$ a uvedu některé její algebraické vlastnosti. Pomocí $\mathfrak{F}$ zavedu charakteristickou funkci pro Jacobiho matice jistého typu. Podobně jako v případě konečných matic platí, že nuly charakteristické funkce jsou vlastní čísla studovaného Jacobiho operátoru. Hlavní část příspěvku je věnována odvození formule pro regularizovanou charakteristickou funkci ve tvaru nekonečného součinu Hadamardova typu.

Klíčová slova: Jacobiho matice, charakteristická funkce, Hadamardova faktorizace

1 Introduction

Results of this paper are related to the eigenvalue problem of infinite symmetric tridiagonal (Jacobi) matrices. At the start, I introduce a function which is called $\mathfrak{F}$ and is defined on a subspace of the linear space of all complex sequences. This function has many nice and simple algebraic properties. Several of them are presented, first of all, the three-term recurrence relation. Other properties are discussed in [5].

Further, I present the Jacobi operator J of a certain type and, with the aid of $\mathfrak{F}$, define a complex function of one complex variable called the characteristic function of J. Zeros of the characteristic function coincide with the spectrum of J, similarly as in the case of finite matrices. Moreover, a vector-valued function is constructed having the property that its values on spectral points of J are equal to corresponding eigenvectors.

*This work has been supported by the grant 10/210/0HK4/2T/14

Finally, the Green function of the Jacobi operator J is expressed in terms of $\mathfrak{F}$ in a very compact manner.

In the last and main part of this contribution I focus on the derivation of the formula for the regularized characteristic function in the form of Hadamard infinite product where eigenvalues of the studied operator plays an essential role. The approach is based on the theory of regularized determinants which is nicely treated in [4]. General results are demonstrated on a very special Jacobi matrix whose parallels to the diagonal are constant and whose diagonal depends linearly on the index.

2 Function $\mathfrak{F}$ and its Properties

I have introduced a function called $\mathfrak{F}$ and list its main properties in doctoral days 2010 proceedings, [6]. Let me briefly recall the definition and the main properties since these facts are essential for next sections.

Definition 1. Define $\mathfrak{F} : D \to \mathbb{C}$

$$\mathfrak{F}(x) = 1 + \sum_{m=1}^{\infty}(-1)^m \sum_{k_1=1}^{\infty} \sum_{k_2=k_1+2}^{\infty} \dots \sum_{k_m=k_{m-1}+2}^{\infty} x_{k_1}x_{k_1+1}x_{k_2}x_{k_2+1}\dots x_{k_m}x_{k_m+1} \tag{1}$$

where

$$D = \left\{ \{x_k\}_{k=1}^{\infty} \subset \mathbb{C}; \sum_{k=1}^{\infty} |x_k x_{k+1}| < \infty \right\}.$$

For a finite number of complex variables let me identify $\mathfrak{F}(x_1, x_2, \dots, x_n)$ with $\mathfrak{F}(x)$ where $x = (x_1, x_2, \dots, x_n, 0, 0, 0, \dots)$.

Remark 2. Note that the domain D is not a linear space. One has, however, $\ell^2(\mathbb{N}) \subset D$. Further, for $x \in D$ one has estimation

$$|\mathfrak{F}(x)| \leq \exp\left(\sum_{k=1}^{\infty} |x_k x_{k+1}|\right). \tag{2}$$

This inequality follows from the fact that the absolute value of the mth summand in the RHS of (1) is majorized by the expression

$$\sum_{\substack{k\in\mathbb{N}^m \\ k_1<k_2<\dots<k_m}} |x_{k_1}x_{k_1+1}x_{k_2}x_{k_2+1}\dots x_{k_m}x_{k_m+1}| \leq \frac{1}{m!}\left(\sum_{j=1}^{\infty}|x_j x_{j+1}|\right)^m.$$

First, $\mathfrak{F}$ satisfies a very important three-term recurrence relation

$$\mathfrak{F}(x) = \mathfrak{F}(x_1, \dots, x_k)\,\mathfrak{F}(T^k x) - \mathfrak{F}(x_1, \dots, x_{k-1})x_k x_{k+1}\mathfrak{F}(T^{k+1}x), \quad k = 1, 2, \dots \tag{3}$$

where $x \in D$ and T denotes the truncation operator from the left defined on the space of all sequences and which has, for $k = 1$, very simple form

$$\mathfrak{F}(x) = \mathfrak{F}(Tx) - x_1 x_2 \mathfrak{F}(T^2 x). \tag{4}$$

Respective proofs of the above statements and other details can be found in [5]. Second, function $\mathfrak{F}$ restricted on $\ell^2(\mathbb{N})$ is a continuous functional and, finally, for $x \in D$, it holds

$$\lim_{n\to\infty} \mathfrak{F}(T^n x) = 1 \tag{5}$$

and

$$\lim_{n\to\infty} \mathfrak{F}(x_1, x_2, \ldots, x_n) = \mathfrak{F}(x). \tag{6}$$

Although proofs of these properties has not been published yet and exists only in personal notes of the author, they are not presented here.

3 Characteristic Function and Jacobi Operators

In the whole section, suppose sequences $\lambda := \{\lambda_n\}_{n=1}^{\infty} \subset \mathbb{C}$, $w := \{w_n\}_{n=1}^{\infty} \subset \mathbb{C} \setminus \{0\}$, satisfying

$$\sum_{n=1}^{\infty} \frac{|w_n|^2}{|(\lambda_{n+1} - z)(\lambda_n - z)|} < \infty, \tag{7}$$

for some $z \in \mathbb{C}$, are given. Further, let the set of all accumulation points of λ, denoted $\mathrm{der}(\lambda)$, be finite. It can be shown, if the condition (7) holds for one $z \in \mathbb{C} \setminus \overline{\lambda}$ ($\overline{\lambda}$ stands for the closure of λ in $\mathbb{C}$) then it remains true for all $z \in \mathbb{C} \setminus \overline{\lambda}$. Consequently, the function F_J, given by relation

$$F_J(z) := \mathfrak{F}\left(\left\{\frac{\gamma_n^2}{\lambda_n - z}\right\}_{n=1}^{\infty}\right),$$

where $\gamma_1 = 1$ and $\gamma_n\gamma_{n+1} = w_n$, for $n \in \mathbb{N}$, is well defined on $\mathbb{C} \setminus \overline{\lambda}$. For the origin of the gamma sequence see [5]. Let me call F_J the characteristic function for Jacobi operator J introduced below. The reason for this terminology follows from the fact that the characteristic function for a Jacobi matrix of a finite dimension can be expressed in terms of $\mathfrak{F}$ with truncated argument. More precisely, it holds

$$\det(J_n - zI_n) = \left(\prod_{k=1}^{n}(\lambda_k - z)\right) \mathfrak{F}\left(\frac{\gamma_1^2}{\lambda_1 - z}, \frac{\gamma_2^2}{\lambda_2 - z}, \ldots, \frac{\gamma_n^2}{\lambda_n - z}\right), \tag{8}$$

where

$$J_n = \begin{pmatrix} \lambda_1 & w_1 & & & \\ w_1 & \lambda_2 & w_2 & & \\ & \ddots & \ddots & \ddots & \\ & & w_{n-2} & \lambda_{n-1} & w_{n-1} \\ & & & w_{n-1} & \lambda_n \end{pmatrix},$$

which was proved in [6]. Properties of F_J as a complex function of one complex variable are listed in the following proposition without proof.

Proposition 3. *Function F_J is an analytic on $\mathbb{C}\setminus\overline{\lambda}$ and it has poles in points $z \in \lambda\setminus\mathrm{der}(\lambda)$ of finite order less or equal to the number*

$$r_z := \sum_{n=1}^{\infty} \delta_{(z,\lambda_n)}.$$

Function F_J is closely related to the spectrum of Jacobi operator $J := WU^* + UW + \Lambda$ acting on $\ell^2(\mathbb{N})$, where W, Λ are diagonal operators, $We_n = w_n e_n$, $\Lambda e_n = \lambda_n e_n$, U is unilateral shift, $Ue_n = e_{n+1}$, U^* its adjoint ($U^*e_1 = 0$, $U^*e_{n+1} = e_n$), $n = 1, 2, \ldots$, and $\{e_n : n \in \mathbb{N}\}$ is the canonical basis of $\ell^2(\mathbb{N})$. The domain of J is the respective intersection,

$$\mathrm{Dom}(J) = \mathrm{Dom}(UW) \cap \mathrm{Dom}(WU^*) \cap \mathrm{Dom}(\Lambda).$$

For other ways how a Jacobi operator can be constructed from a semi-infinite symmetric matrix see [1].

Example 4. The Bessel function of the first kind can be expressed in terms of $\mathfrak{F}$. More precisely, for $w, \nu, \alpha \in \mathbb{C}$, $\alpha \neq 0$, $\nu/\alpha \notin -\mathbb{N}$, one has

$$J_{\frac{\nu}{\alpha}}\left(\frac{2w}{\alpha}\right) = \frac{1}{\Gamma(\frac{\nu}{\alpha}+1)}\left(\frac{w}{\alpha}\right)^{\frac{\nu}{\alpha}} \mathfrak{F}\left(\left\{\frac{w}{\nu+\alpha k}\right\}_{k=1}^{\infty}\right). \tag{9}$$

This equality can be verified by only slightly modified computation which is work out in [5].

The connection between F_J and $\mathrm{spec}(J)$ and much more is described in the following theorem.

Theorem 5. *Let me denote*

$$\mathfrak{Z}(J) := \left\{ z \in \mathbb{C} \setminus \mathrm{der}(\lambda) : \lim_{\tilde{z}\to z}(z-\tilde{z})^{r_z} F_J(\tilde{z}) = 0, r_z = \sum_{n=1}^{\infty} \delta_{(z,\lambda_n)} \right\},$$

then equalities

$$\mathrm{spec}(J) \setminus \mathrm{der}(\lambda) = \mathrm{spec}_p(J) \setminus \mathrm{der}(\lambda) = \mathfrak{Z}(J) \setminus \mathrm{der}(\lambda) \tag{10}$$

hold. Further, vector-valued function $\xi(z) := \{\xi_n(z)\}_{n=1}^{\infty}$ defined by the relation

$$\xi_n(z) := \lim_{\tilde{z}\to z}(\tilde{z}-z)^{r_z} \prod_{l=1}^{n}\left(\frac{w_{l-1}}{\tilde{z}-\lambda_l}\right) \mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k-\tilde{z}}\right\}_{k=n+1}^{\infty}\right), \tag{11}$$

for $z \notin \mathrm{der}(\lambda)$, has the property that its values on spectral points $z \in \mathrm{spec}_p(J) \setminus \mathrm{der}(\lambda)$ are equal to corresponding eigenvectors, i.e., $J\xi(z) = z\xi(z)$ and $0 \neq \xi(z) \in \ell^2(\mathbb{N})$. Finally, for $z \notin (\mathrm{spec}(J) \cup \mathrm{der}(\lambda))$, and $i, j \in \mathbb{N}$, the matrix element of the Green function $G_{i,j}(z) := (e_i, (J-z)^{-1}e_j)$ is given by the formula

$$G_{i,j}(z) = -\frac{1}{w_{\max(i,j)}} \prod_{l=\min(i,j)}^{\max(i,j)}\left(\frac{w_l}{z-\lambda_l}\right) \frac{\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z}\right\}_{l=1}^{\min(i,j)-1}\right) \mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z}\right\}_{l=\max(i,j)+1}^{\infty}\right)}{\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z}\right\}_{l=1}^{\infty}\right)}. \tag{12}$$

This theorem sums up statements of several propositions which proofs will be published in a future paper.

Remark 6. First, note the set $\mathfrak{Z}(J)$ includes the set of all zeros of F_J. Second, functions $\xi_n(z)$ satisfy an interesting identity,

$$\sum_{k=1}^{\infty}(\xi_k(z))^2 = \xi_0'(z)\xi_1(z) - \xi_0(z)\xi_1'(z),$$

for $z \in \mathbb{C} \setminus \mathrm{der}(\lambda)$. In particular, if λ, w are real and $z \in \mathrm{spec}_p(J)$ then one gets a simple formula for the ℓ^2-norm of the corresponding eigenvector,

$$\|\xi(z)\|^2 = \xi_0'(z)\xi_1(z).$$

Example 7. By Theorem 5 and Example 4, one arrives at the following expression,

$$\mathrm{spec}(J) = \left\{ z \in \mathbb{C};\ J_{-\frac{z}{\alpha}}\left(\frac{2w}{\alpha}\right) = 0 \right\}$$

where J is the Jacobi operator given by the choice $\lambda_n = \alpha n$, $\alpha \neq 0$ and $w_n = w \neq 0$, $n = 1, 2, \ldots$. The formula for the k-th entry of the eigenvector correnspondig to eigenvalue $z \in \mathrm{spec}_p(J)$ then reads

$$v_k(z) = (-1)^k J_{k-\frac{z}{\alpha}}\left(\frac{2w}{\alpha}\right).$$

4 Hadamard Type Infinite Product and Characteristic Function

In this section, I derive a formula for regularized characteristic function in terms of the Hadamard infinite product.

Let $\{w_n\}_{n=1}^{\infty}$, $\{\lambda_n\}_{n=1}^{\infty}$ be real sequences and $\lim_{n\to\infty}\lambda_n = \infty$. In addition, without loss of generality one can assume $\{\lambda_n\}_{n=1}^{\infty}$ to be positive. Otherwise, one adds a positive constant to every element of the sequence to fulfill the positivity. This step means that a multiple of identity is added to the Jacobi operator which only shifts the spectrum. Next, let

$$\sum_{n=1}^{\infty}\frac{w_n^2}{\lambda_n\lambda_{n+1}} < \infty \quad \text{and} \quad \sum_{n=1}^{\infty}\frac{1}{\lambda_n^2} < \infty. \tag{13}$$

To remove poles of finite order of the function F_J, let me define function φ_Λ by relation

$$\varphi_\Lambda(z) := \prod_{n=1}^{\infty}\left(1 + \frac{z}{\lambda_n}\right)e^{-\frac{z}{\lambda_n}}.$$

Since $\sum_n 1/\lambda_n^2 < \infty$ the function φ_Λ is well defined and it is an entire function. Further, φ_Λ has zeros in points $z = -\lambda_n$ with multiplicity equal to the number of repeating of the value λ_n in the sequence $\{\lambda_n\}$ and φ_Λ has no other zeros (see, for instance, [3], Chapter 15). Let me call the function H_J, defined by the relation

$$H_J(z) := \varphi_\Lambda(z)F_J(-z),$$

the regularized characteristic function for operator J.

Theorem 8. *Function $H_J(z)$ is entire and the equality*

$$\{-z \in \mathbb{C} : H_J(z) = 0\} = \operatorname{spec}(J)$$

holds.

Proof. Assumption (13) implies $\operatorname{der}(\lambda) = \emptyset$ .The statement then follows from Proposition 3 and Theorem 5. □

Example 9. By using Example 4 and the well-known formula for the gamma function,

$$\Gamma(z) = \frac{e^{-\gamma z}}{z} \prod_{n=1}^{\infty} \left(1 + \frac{z}{n}\right)^{-1} e^{\frac{z}{n}}, \tag{14}$$

where γ is the Euler–Mascheroni constant, one gets

$$H_J(z) = e^{-\gamma z} w^{-z} J_z(2w)$$

with the choice $\lambda_n = n$, and $w_n = w$.

In Chapters 3,5, and 9 of [4] it is shown how to define determinant of $I + A$ where A is a Schatten class operator, especially, a Hilbert-Schmidt operator. Let me write, for simplicity, W instead of UW and similarly W^* instead of WU^*. Thus, the Jacobi operator J is then equal to $W + W^* + \Lambda$. Since, for $z \in \mathbb{C}$, operator

$$\Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}$$

is represented by matrix

$$\begin{pmatrix} \frac{z}{\lambda_1} & \frac{w_1}{\sqrt{\lambda_1\lambda_2}} & & & \\ \frac{w_1}{\sqrt{\lambda_1\lambda_2}} & \frac{z}{\lambda_2} & \frac{w_2}{\sqrt{\lambda_2\lambda_3}} & & \\ & \frac{w_2}{\sqrt{\lambda_2\lambda_3}} & \frac{z}{\lambda_3} & \frac{w_3}{\sqrt{\lambda_3\lambda_4}} & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

it is, due to (13), the Hilbert-Schmidt operator for all $z \in \mathbb{C}$. Hence, the number

$$\det\left((I + \Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}) \exp\{-\Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}\}\right),$$

which is usually denoted as

$$\det{}_2(I + \Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}),$$

is well defined for all $z \in \mathbb{C}$ (see [4], Chapter 9, for details).

Proposition 10. *It holds*

$$H_J(z) = \det\left((I + \Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}) \exp\{-\Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}\}\right).$$

Proof. First, we verify the formula for truncated finite dimensional operators, $J_N = P_N J P_N$, $\Lambda_N = P_N \Lambda P_N$, where P_N is OG projection on the space spanned by the first N vectors $e_1, \ldots, e_N$ of the canonical basis of $\ell^2(\mathbb{N})$. Thus,

$$\det\left((I + P_N\Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}P_N)\exp\{-P_N\Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}P_N\}\right)$$
$$= \det(\Lambda_N^{-1})\det(J_N + z)\exp\left(-z\,\mathrm{Tr}(\Lambda_N^{-1})\right) = \prod_{n=1}^{N}\left(1 + \frac{z}{\lambda_n}\right)e^{-\frac{z}{\lambda_n}}\mathfrak{F}\left(\left\{\frac{\gamma_n^2}{\lambda_n + z}\right\}_{n=1}^{N}\right),$$

where properties of the determinant, $\det(AB) = \det(A)\det(B)$, $\det(\exp(A)) = \exp(\mathrm{Tr}(A))$, and formula (8) were used. Next, it suffices to sent N to infinity in the above equation. By (6), it is clear the RHS tends to $H_J(z)$ with $N \to \infty$. What remains to be shown is that if A, A_N are Hilbert-Schmidt operators and $\|A_N - A\|_2 \to 0$ ($\|.\|_2$ stands for the Hilbert-Schmidt norm) then it holds

$$\lim_{N\to\infty}\det\left((I + A_N)\exp(-A_N)\right) = \det\left((I + A)\exp(-A)\right).$$

This follows from inequality

$$\left|\det\left((I + A_N)e^{-A_N}\right) - \det\left((I + A)e^{-A}\right)\right| \le \|A_N - A\|_2\exp\{C(\|A\|_2 + \|A_N\|_2 + 1)^2\},$$

see [4], Theorem 9.2. C is a constant and $\|A_N\|_2 \le \|A\|_2$ in this case (otherwise, it is bounded anyway). □

Remark 11. Bearing in mind identities holding for finite matrices, $\det(AB) = \det(A)\det(B)$ and $\det(\exp(A)) = \exp(\mathrm{Tr}(A))$, one would like to write

$$\det\left((I + \Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2})\exp\{-\Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2}\}\right)$$
$$= \det\left((I + \Lambda^{-1/2}(W + W^* + z)\Lambda^{-1/2})\exp\{-z\Lambda^{-1}\}\right),$$

however, the operator in the argument of the determinant on the RHS is not of the form: identity + trace class operator, hence the RHS has not a good sense from the point of view of the Simon's book [4].

To find a formula for H_J in the form of an infinite product one can try to apply the Hadamard factorization theorem (see [2], Theorem 3.4). To use the theorem one has to know the order of the entire function H_J. The order λ of H_J can be computed by using formula

$$\lambda = \limsup_{R\to\infty}\frac{\ln\ln\|H_J\|_{S_R,\infty}}{\ln R}$$

where $\|H_J\|_{S_R,\infty} = \sup\{|H_J(z)| : |z| = R\}$ ([2], Proposition 2.15).

However, I have not been successful in computing the order by this straightforward way. One could guess the order is strictly less then 2. If this would be true the desired formula could have this form

$$H_J(z) = e^{a+bz}\prod_{n=1}^{\infty}\left(1 + \frac{z}{\lambda_n(J)}\right)e^{-\frac{z}{\lambda_n(J)}} \tag{15}$$

where $a, b \in \mathbb{C}$ and $\{\lambda_n(J) : n \in \mathbb{N}\} = \mathrm{spec}_p(J)$. I will show this identity is true.

First, let me assume J to be invertible. This assumption is not too restrictive. Since, according to Theorem 8, $\mathrm{spec}(J)$ coincides with zeros of the entire function H_J, $\mathrm{spec}(J)$ is composed of isolated points (which are simple eigenvalues). Hence, an appropriate constant multiple of identity can be add to J to let zero be in the resolvent set of J.

Second, note that J is invertible if and only if $(I+A)$ is invertible, where

$$A := \Lambda^{-1/2}(W+W^*)\Lambda^{-1/2},$$

which follows from equality $F_A(0) = F_J(0)$ and Theorem 8. Equivalently, one has

$$0 \in \mathrm{spec}_p(J) \Longleftrightarrow -1 \in \mathrm{spec}_p(A).$$

Since, by (12), J^{-1} has symmetric and real matrix representation, it is a hermitian operator. Moreover, $(\Lambda^{-1}(I+A)^{-1})^2$ is a trace class operator (it is multiplication of two Hilbert-Schmidt operators), hence one has

$$\mathrm{Tr}(J^{-2}) = \mathrm{Tr}\left(\Lambda^{-\frac{1}{2}}(I+A)^{-1}\Lambda^{-\frac{1}{2}}\right)^2 = \mathrm{Tr}\left(\Lambda^{-1}(I+A)^{-1}\right)^2 < \infty,$$

thus J^{-1} is Hilbert-Schmidt.

Next, since, for C, D Hilbert-Schmidt operators, it holds

$$\det{}_2(I+C+D+CD) = \det{}_2(I+C)\det{}_2(I+D)\exp(-\mathrm{Tr}(CD)),$$

see [4], Chapter 9, one can write

$$\begin{aligned} H_J(z) &= \det{}_2(I+A+z\Lambda^{-1}) \\ &= \det{}_2(I+A)\det{}_2\left(I+z(I+A)^{-1}\Lambda^{-1}\right)\exp\left(-z\,\mathrm{Tr}(A(I+A)^{-1}\Lambda^{-1})\right). \end{aligned} \tag{16}$$

Lemma 12. *For J invertible, the identity*

$$\det{}_2\left(I+z(I+A)^{-1}\Lambda^{-1}\right) = \det{}_2\left(I+zJ^{-1}\right) \tag{17}$$

holds.

Proof. The proof is based on Plemejl-Smithies formula for $\det_2$ ([4], Theorem 9.3), which, for C a Hilbert-Schmidt operator, reads

$$\det{}_2(I+zC) = \sum_{m=0}^{\infty} \frac{\alpha_m(C)}{m!} z^m,$$

where

$$\alpha_m(C) = \left| \begin{pmatrix} 0 & m-1 & 0 & \dots & 0 & 0 \\ \mathrm{Tr}\,C^2 & 0 & m-2 & \dots & 0 & 0 \\ \mathrm{Tr}\,C^3 & \mathrm{Tr}\,C^2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{Tr}\,C^{m-1} & \mathrm{Tr}\,C^{m-2} & \mathrm{Tr}\,C^{m-3} & \dots & 0 & 1 \\ \mathrm{Tr}\,C^m & \mathrm{Tr}\,C^{m-1} & \mathrm{Tr}\,C^{m-2} & \dots & \mathrm{Tr}\,C^2 & 0 \end{pmatrix} \right|,$$

for $m \geq 1$, $\alpha_0(C) = 1$. Now, since

$$\operatorname{Tr}\left((I+A)^{-1}\Lambda^{-1}\right)^m = \operatorname{Tr}\left(\Lambda^{-\frac{1}{2}}(I+A)^{-1}\Lambda^{-\frac{1}{2}}\right)^m = \operatorname{Tr}(J^{-m}),$$

one has

$$\alpha_m((I+A)^{-1}\Lambda^{-1}) = \alpha_m(J^{-1}),$$

for $m = 0, 1, \ldots$, and the statement is proved by the Plemejl-Smithies formula. □

Finally, according to Theorem 9.2 in [4], a product formula of the form

$$\det{}_2(I+zC) = \prod_{n=1}^{\infty}(1+z\mu_n(C))e^{-z\mu_n(C)}, \tag{18}$$

where C is Hilbert-Schmidt operator with $\operatorname{spec}(C) = \{\mu_n(C) : n = 1, 2, \ldots\}$ (counting up to multiplicity), holds. By putting together (16), (17), and (18), one arrives at the following theorem.

Theorem 13. *Let J is invertible then the product formula*

$$H_J(z) = e^{a+bz}\prod_{n=1}^{\infty}\left(1+\frac{z}{\lambda_n(J)}\right)e^{-\frac{z}{\lambda_n(J)}}, \tag{19}$$

where $a = \ln(\det{}_2(I+A))$, $b = -\operatorname{Tr}(A(I+A)^{-1}\Lambda^{-1})$, and $\operatorname{spec}_p(J) = \{\lambda_n(J) : n = 1, 2, \ldots\}$, holds.

Corollary 14. *For each $\alpha > 0$ there is $R_\alpha > 0$ such that, for $|z| > R_\alpha$,*

$$|H_J(z)| < \exp\left(\alpha|z|^2\right).$$

Proof. By Theorem 13, H_J is an entire function of genus one. The statement then follows from Theorem 2.6 in [2]. □

Example 15. With the aid of Theorem 13 and Example 9 one can rediscover the infinite product formula for the Bessel function of the first kind considered as a function of its order. The formula reads

$$\frac{w^{-z}J_z(2w)}{J_0(2w)} = e^{\gamma z}\prod_{n=1}^{\infty}\left(1+\frac{z}{\lambda_n(J)}\right)e^{-\frac{z}{\lambda_n(J)}}$$

where $\lambda_n(J)$, $n = 1, 2, \ldots$ are eigenvalues of Jacobi operator J with linear diagonal and constant parallels, i.e. $\lambda_n = n, w_n = w$, $z, w \in \mathbb{C}$, $J_0(2w) \neq 0$. To verify this identity one only has to show the constant b from Theorem 13 is zero in this special case. By using the series expansion for $J_z(2w)$ in w and by setting $w = 0$ in (19) (note (19) holds even if $w = 0$), one arrives at the equality

$$\frac{e^{-\gamma z}}{\Gamma(z+1)} = e^{bz}\prod_{n=1}^{\infty}\left(1+\frac{z}{n}\right)e^{-\frac{z}{n}},$$

for $\lambda_n(J) \to n$ as $w \to 0$. Finally, it suffices to use identity (14).

Remark 16. For $\{\lambda_n\}_{n=1}^{\infty}$ positive, satisfying

$$\sum_{n=1}^{\infty} \frac{1}{\lambda_n \lambda_{n+1}} < \infty,$$

the formula

$$F_z(w) := \mathfrak{F}\left(\left\{\frac{w}{\lambda_n - z}\right\}_{n=1}^{\infty}\right) = \prod_{k=1}^{\infty}\left(1 - \frac{w^2}{\zeta_k^2(z)}\right), \tag{20}$$

where $\pm\zeta_1(z), \pm\zeta_2(z), \ldots$ are all zeros of F_z, is also true. The proof is ready to be published in a future work, however, is omitted here. In fact, this result was known before Theorem 13 and served as a motivation for the work presented here.

References

[1] B. Beckerman: *Complex Jacobi matrices*, J. Comput. Appl. Math., 127, (2001), 17-65.

[2] J. B. Conway: *Functions of One Complex Variable*, Second Edition, Springer (New York, 1978).

[3] W. Rudin: *Analýza v reálném a komplexním oboru*, Academia (Praha, 2003).

[4] B. Simon: *Trace Ideals and Their Applications*, Second Edition, Mathematical surveys and monographs, vol. 120, (2005).

[5] F. Štampach, P. Šťovíček: *On the eigenvalue problem for a particular class of finite Jacobi matrices*, Lin. Alg. App., 434, (2011), 1336-1353

[6] F. Štampach: *The Characteristic Function for a Particular Class of Infinite Jacobi Matrices*, Doktorandské dny 2010, sborník ČVUT, (2010).

Interaction-Sensitive Fuzzy Measure in Dynamic Classifier Aggregation: an Experimental Comparison

David Štefka

7th year of PGS, email: david.stefka@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Martin Holeňa, Institute of Computer Science, ASCR, v.v.i.

Abstract. As the literature shows, the performance of a pattern recognition system can be improved by introducing the concept of classifier aggregation [2]. Instead of using a single classifier ϕ, we create a team of classifiers $\phi_1, ..., \phi_r$, let the classifiers predict independently, and aggregate the results using an aggregation operator. One of the popular aggregation operators is the *fuzzy integral* [4, 1], which aggregates the outputs of the individual classifiers in the team with respect to a *fuzzy measure*. Fuzzy measure is a set function representing the classification confidence of the prediction for a given set of classifiers.

Fuzzy measure is a generalization of the additive probabilistic measure, where the additivity is replaced by a weaker condition, monotonicity ($A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$) – this gives us a tool which can model interactions between different elements of the fuzzy measure space. However, due to the lack of additivity, the fuzzy measure needs to be defined on all subsets of the fuzzy measure space, resulting in 2^r defining values for finite cases, where r is the size of the universe (the number of classifiers). There are several approaches to overcome this weakness: *additive measures*, corresponding to the probabilistic measure, *symmetric fuzzy measures*, for which the value of the measure depends only on the number of elements in the argument, and $\perp$-*decomposable* fuzzy measures, including *Sugeno* λ-*measure* , for which the fuzzy measure values are computed from the fuzzy measure values for the singletons (called *fuzzy densities*) using a fixed t-conorm $\perp$. However, it can be shown that none of the aforementioned approaches can reasonably model interactions between the different elements of the universe, i.e., to model similarities of the individual classifiers.

In the literature of classifier aggregation, fuzzy integral is usually used with Sugeno λ-measure. There is usually no explicit reason for the choice of this measure other than its simplicity. Sugeno λ-measure is a special case of a $\perp$-decomposable fuzzy measure, and as such, it cannot model similarities between the individual classifiers, and thus the contribution of using fuzzy integral in the agregation is unclear.

In classifier aggregation, we usually try to create a team of classifiers that are not similar. This property is called *diversity* [3]. There are many methods for building a diverse team of classifiers; however, the team always contains classifiers that are similar. If we use the fuzzy integral with a symmetric or $\perp$-decomposable fuzzy measure, we are not able to incorporate the diversity into the measure (and thus to the aggregation process), because the fuzzy measure of a union of two sets is a function only of the fuzzy measures of the two sets, regardless of the similarity of the elements in the sets.

To overcome this weakness, we have introduced an *Interaction-Sensitive Fuzzy Measure* (ISFM) [6, 5], which is defined using the fuzzy measure values for the singletons (fuzzy densities),

and the similarities of the elements in the universe. If the fuzzy measure space corresponds to the team of classifiers, the fuzzy measure incorporates both the classification confidence (fuzzy densities), and the diversity of the team of classifiers (mutual similarities of the classifiers). Using ISFM in fuzzy integral as an aggregation operator in classifier aggregation, the aggregation process involves all the important properties: the predictions of the classifiers, the classification confidences, and the diversity of the team.

The theoretical results with preliminary experiments with ISFM were published in [6]. In [5], the experiments were extended to cover the Choquet and the Sugeno integral, and also to cover other classification models, namely Random Forests, ensembles of k-Nearest Neighbor classifiers created by bagging and ensembles of Quadratic Discriminant Classifiers created by the Multiple feature subset method. The methods were evaluated on 23 benchmark datasets, and the results show that ISFM outperforms the Sugeno λ-measure for both Choquet and Sugeno fuzzy integrals.

Keywords: classier aggregation, fuzzy integral, fuzzy measure

References

[1] Michel Grabisch and Hung T. Nguyen. *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[2] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[3] Ludmila I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles. *Machine Learning*, 51:181–207, 2003.

[4] Vicenç Torra and Yasuo Narukawa. *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer, 2007.

[5] David Štefka. Interaction-sensitive fuzzy measure in dynamic classifier aggregation: an experimental comparison. In D. Kuželová and F. Hakl, editors, *Proceedings of the XVI. PhD. Conference, October 2011*. Institute of Computer Science, ASCR.

[6] David Štefka and Martin Holeňa. Dynamic classifier aggregation using fuzzy integral with interaction-sensitive fuzzy measure. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications, ISDA 2010, November 29 - December 1, 2010, Cairo, Egypt*, pages 225–230. IEEE, 2010.

Factor Analysis of Scintigraphic Image Sequences with Integrated Probabilistic Mask of Factor Images

Ondřej Tichý*

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

advisor: Václav Šmídl, Department of Adaptive Systems,
Institute of Information Theory and Automation, AS CR

Abstract. Factor analysis is a well established mathematical method for factor separation in the analysis of scintigraphical sequences. The results are typically an input to the next step, e.g. factor analysis for computing significant diagnostic coefficients. However, this computing highly depends on proper identification of factors and their biological meaning, which is not ensure only by factor analysis. The main issue is separation overlaping factors from themselves and from tissue background covering the whole sequence. Factor analysis highly depends on prior information which allows us to set biologically reasonable conditions to a mathematical model. In this paper, we propose a mathematical model which estimates the probability mask of each image factor and sets it as a prior information for the next step of iterative algorithm based on Variational Bayes method. The new proposed model provides more realistic estimates of factors than the standard factor analysis.

Keywords: Nuclear Medicine, Scintigraphy, Factor Analysis, Factor Separation

Abstrakt. Jednou ze známých matematických metod pro analýzu scintigrafických obrazových sekvencí je faktorová analýza. Cílem diagnostiky je určit důležité diagnostické koeficienty, k tomu je ovšem potřeba detekovat jednotlivé, biologicky smysluplné, faktory, což nelze zajistit samotnou faktorovou analýzou. Základním problémem při analýze sekvence je překryv jednotlivých orgánů a jejich částí a odseparování krevního a tkáňového pozadí, které se vyskytují v celé sekvenci, přičemž faktorová analýza umožňuje zabudovat biologické předpoklady vedoucí ke smysluplnému řešení problému. V tomto příspěvku je představen nový matematický model, který odhaduje pravděpodobnost příslušnosti jednotlivých pixelů k faktorovým obrázkům a tuto informaci využívá k nastavení apriorna pro další krok výpočtu založeném na metodě Variační Bayes. Tento model dává realističtější odhady faktorů než standardní faktorová analýza.

Klíčová slova: Nukleární Medicína, Scintigrafie, Faktorová Analýza, Separace Faktorů

1 Introduction

Scintigraphy is a well known and very important method in nuclear medicine. Diagnosis using scintigraphy includes following steps. At first, a tagged radiopharmaceutical is applied into a human body lying under the scintillation camera. At second, in every

*Institute of Information Theory and Automation, Department of Adaptive Systems, AS CR

10 seconds an image of distribution of radiopharmaceutical is saved; consequently, the functional image sequence with the scanned region of interest is obtained. Further analysis of measurement is necessary for propper diagnosis. In this paper, we are focused on renal scintigraphy.

A kidney is composed of parenchyma and pelvis. In biological constraint, in about the first 120 - 180 seconds fills only parenchyma of kidney [2]; then the radiopharmaceutical passes from parenchyma to pelvis and next to the urinary bladder. This is very important information for biologically meaningful solution and verification of a mathematical model, see Section 4.1. Another assumption, the shape of convolution kernel of factor curve, will be studied in Section 4.2. For further analysis, factor identification is necessary. This is typically done by expert manually or by factor analysis automatically [1]. Finally, the resulting factors can be analyzed to set the proper diagnosis. This analysis can be done by expert or by semi-automatic algorithm based on more or less sophisticated mathematical background: Patlak-Rutland plot [4], or post-processing by deconvolution [6]. The result highly depends on the first step, correct separation, identification, and detection of factors.

Factor analysis is a statistical method based on data decomposition to the factors. Its usage is mostly scintigraphy [1], ultrasound [7], or PET [5]. However, the solution of factor analysis is ambiguous and allows infinitely many solutions. Some restrictions have been made for biologically meaningful solution, e.g. positivity of factors [10]; nevertheless, the uniqueness of solution or even biologically meaningful solution is not guaranteed only by positivity. Uniqueness can be guaranteed when each factor has at least one pixel where the others have no activity [11], but this assumption does not hold in scintigraphy because of residue activity in the whole sequence. Additional constraints are necessary to restrict the space of possible solutions.

The analytical solution of the presented model is intractable; therefore, an additional approximations have been made. The Variational Bayes approximation methodology [8] was successfully used in fields related to factor decomposition, e.g. principal component analysis, factor analysis, or models with convolution. In addition, Variational Bayes approximation offers reasonable ratio between options of modeling and computation difficulties.

2 Variational Factor Analysis (FA)

We briefly review Variational Factor Analysis. The sequence obtained by scintillation camera contains n images taken at time $t = 1 \dots n$, typically after 10 seconds. Every image is a compound of p pixels; consequently, the images are saved in p-dimensional vectors and data matrix $D \in \mathbf{R}^{p \times n}$ is generated. Let us assume that each observed image is a linear combination of r factor images, aggregated in matrix $A \in \mathbf{R}^{p \times r}$. Typically, $r < n \ll p$ is expected. Every factor image has its time-activity curve, $x_j = [x_{1,j}, \dots, x_{n,j}]'$; therefore, time-activity matrix $X \in \mathbf{R}^{n \times r}$ is created. The only that we have is data-storage matrix D, and we would like to estimate factor image matrix A and factor curve matrix X.

The model of the factor analysis can be written in matrix form as

$$D = AX' + E, \tag{1}$$

where $E \in \mathbf{R}^{p\times n}$ is noise matrix with i.i.d. elements with variance ω^{-1}. Matrix D aggregates measurements of radioactive particles with Poisson distributions which can be approximated by Gauss normal distribution; therefore, covariance matrix of noise matrix E can be found using correspondence analysis [3] as

$$f(D|A, X, \omega) = \mathrm{tN}_D(AX', \omega^{-1}\Omega_p \otimes \Omega_n), \tag{2}$$

$$\Omega_p = \mathrm{diag}(D\mathbf{1}_{n,1}), \Omega_n = \mathrm{diag}(\mathbf{1}_{1,p}D), \tag{3}$$

where tN(.) denotes truncated normal distribution, diag(.) denotes square diagonal matrix with diagonal vector as an argument, $\mathbf{1}_{k,l}$ denotes matrix of ones of subscripted dimensions, and $\otimes$ denotes Kronecker matrix product.

A prior model of parameters follows as:

$$f(\omega) = G_\omega(\vartheta_0, \rho_0), \tag{4}$$

$$f(X|\Upsilon) = \mathrm{tN}_X(0_{n,r}, \Omega_n \otimes \Upsilon^{-1}), \tag{5}$$

$$\Upsilon = \mathrm{diag}(v), v = [v_1, \ldots, v_r]', \tag{6}$$

$$f(v) = \prod_{j=1}^{r} G_{v_j}(\alpha_{j,0}, \beta_{j,0}), \tag{7}$$

$$f(A) = \mathrm{tN}_A(0_{p,r}, \Omega_p \otimes I_r), \tag{8}$$

where $\vartheta_0, \rho_0 \in \mathbf{R}$ are scalar prior parameters, v is vector of hyper-parameters with prior parameters $\alpha_0, \beta_0 \in \mathbf{R}$, G(.) is gamma distribution, and I_r is identity matrix of dimensions $r \times r$.

The difference between principal component analysis (PCA) and factor analysis is truncation in equations (5) and (8); in addition, for non-truncated distributions in (5) and (8), variational solution converges to the PCA solution [8].

With respect to Variational Bayes method [8], a logarithm of joint distribution $f(D, A, X, \Upsilon, \omega|r)$ is computed and the resulting approximate posterior marginals are recognized in form:

$$\tilde{f}(\omega|D, r) = G_\omega(\vartheta, \rho), \qquad \tilde{f}(X|D, r) = \mathrm{tN}_X(\mu_X, \Sigma_X \otimes \Upsilon), \tag{9}$$

$$\tilde{f}(v|D, r) = \prod_{j=1}^{r} G_{v_i}(\alpha_i, \beta_i), \qquad \tilde{f}(A|D, r) = \mathrm{tN}_A(\mu_A, \Omega_p^{-1} \otimes \Sigma_A), \tag{10}$$

and the associated shaping parameters are

$$\mu_A = \hat{\omega}\Omega_p D\Omega_n \widehat{X}\Sigma_A, \qquad \Sigma_A = \left(\hat{\omega}\widehat{X'\Omega_n X} + I_r\right)^{-1},$$

$$\mu_X = \hat{\omega}\Omega_n D'\Omega_p \widehat{A}\Sigma_X, \qquad \Sigma_X = \left(\hat{\omega}\widehat{A'\Omega_p A} + \widehat{\Upsilon}\right)^{-1},$$

$$\alpha = \alpha_0 + \frac{n}{2}\mathbf{1}_{r,1}, \qquad \beta = \beta_0 + \frac{1}{2}\operatorname{diag}\left(\widehat{X'\Omega_n X}\right),$$
$$\vartheta = \vartheta_0 + \frac{np}{2}, \qquad \rho = \rho_0 + \frac{1}{2}\operatorname{tr}\left(DD' - \widehat{A}\widehat{X}'D' - D\widehat{X}\widehat{A}'\right) + \frac{1}{2}\operatorname{tr}\left(\widehat{A'A}\widehat{X'X}\right).$$

The necessary moments of previous distributions are $\widehat{\Upsilon} = \operatorname{diag}(\alpha \circ \beta^{-1})$, where $\circ$ denotes Hadamard product, $\widehat{\omega} = \frac{\vartheta}{\rho}$ and moments of truncated normal distribution are computed with respect to Appendix A.

3 Factor Analysis with a Prior Mask on Factor Images (FAM)

In the previous section, we revised classical factor analysis without any additional assumptions. Our long-way intention is to automatically analyse a scintigraphical sequence, not only set out factor images and factor curves. This section models a prior probabilistic mask on factor images, i.e. matrix A. This is motivated by unsatisfactory separation of tissue background from other organs, parenchyma and pelvis at most, in the previous methods.

3.1 Modeling of Factor Images

Our new model should better separate tissue background and the proper organ; consequently, the relation factor curve will be better too. In addition, a probability mask of location of a factor will be obtained.

Consider prior probability mask of A of the same size as A, $\mathbf{i} \in \mathbf{R}^{p\times r}$, where

$$\mathbf{i}_{i,j} = \begin{cases} 1 & i\text{th pixel belongs to the } j\text{th factor} \\ 0 & i\text{th pixel not belongs to the } j\text{th factor} \end{cases}, \text{ with prior}$$

$$f(\mathbf{i}_{i,j}) = \operatorname{Exp}(\lambda_{ij,0}). \tag{11}$$

In places with pixels which not belong to the related factor, the noise with normal distribution with zero mean value is expected. For the jth factor, these pixels have distribution $N(0, \xi_{0,j}^{-1})$. Here, $\xi_{0,j}$ is covariance of these zero-mean-pixels of the jth factor hyperparametrized by ϕ and ψ as gamma distribution; for $\xi_0 = [\xi_{0,1}, \ldots, \xi_{0,r}]', \Xi_0 = \operatorname{diag}(\xi_0)$ is

$$f(\xi_0) = \prod_{j=1}^{r} G_{\xi_{0,j}}(\phi_{j,0}, \psi_{j,0}). \tag{12}$$

In case of non-zero-value-pixels of the jth factor, uniform distribution is expected in the form $U(0, A_j^{\max})$ for $A_j^{\max} = \max_i A_{i,j}$. In general, the second parameter of uniform distribution can be replaced by Pareto distribution or Gamma distribution, but the maximum of the jth column of matrix A is almost the same.

Generally, matrix A is modeled as independent elements as

$$f(A) = \prod_{i=1}^{p} \prod_{j=1}^{r} f(a_{i,j}), \tag{13}$$

and each element is modeled as

$$f(a_{i,j}) = U(0, A_j^{\max})^{\mathbf{i}_{i,j}} \, \mathrm{tN}_{a_{i,j}}(0, \xi_{0,j}^{-1})^{(1-\mathbf{i}_{i,j})}, \tag{14}$$

where exponentation of $\mathbf{i}_{i,j}$ or $(1 - \mathbf{i}_{i,j})$ provides an affilation to the informative or non-informative part of the factor image.

3.2 Variational Solution

The joint likelihood for the new model, $f(D, A, X, \Upsilon, \Xi_0, \mathbf{i}, \omega | r)$, is obtained by replacing (8) in model (2) - (8) with prior information (11), (12), and (14). Using Variational Bayes method, the following posterior densities are identified:

$$\tilde{f}(X|D,r) = N(\mu_X, I_n \otimes \Sigma_X), \qquad \tilde{f}(v|D,r) = \prod_{j=1}^{r} G_{v_j}(\alpha_j, \beta_j),$$
$$\tilde{f}(\omega|D,r) = G_{\omega}(\vartheta, \rho), \qquad \tilde{f}(a_i|D,r) = N_{a_i}(\mu_{a_i}, \Sigma_{a_i}),$$
$$\tilde{f}(\xi_0|D,r) = G_{\xi_0}(\phi, \psi), \qquad \tilde{f}(\mathbf{i}_{i,j}|D,r) = \mathrm{Exp}_{\mathbf{i}_{i,j}}(\lambda_{i,j}),$$

with shaping parameters

$$\Sigma_X = \left(\widehat{\omega}\widehat{A'A} + \widehat{\Upsilon}\right)^{-1}, \qquad \mu_X = \widehat{\omega} D' \widehat{A} \Sigma_X,$$
$$\alpha = \alpha_0 + \frac{n}{2}\mathbf{1}_{r,1}, \qquad \beta = \beta_0 + \frac{1}{2}\,\mathrm{diag}(\widehat{X'X}),$$
$$\vartheta = \vartheta_0 + \frac{pn}{2}, \qquad \rho = \rho_0 + \frac{1}{2}\,\mathrm{tr}\left(DD' - \widehat{A}\widehat{X}'D' - D\widehat{X}\widehat{A}'\right) + \frac{1}{2}\,\mathrm{tr}\left(\widehat{A'A}\widehat{X'X}\right),$$
$$\Sigma_{a_i} = \left(\widehat{\omega}\sum_{k=1}^{n}(\widehat{x_k' x_k}) + \widehat{\Xi_0}(I_r - \widehat{\iota_i})\right)^{-1}, \qquad \mu_{a_i} = \left(\Sigma_{a_i}\left(\widehat{\omega}\sum_{k=1}^{n}(\widehat{x}_k d_{i,k})'\right)\right)',$$
$$\phi_j = \left(\phi_{j,0} + \frac{1}{2}\sum_{i=1}^{p}(1 - \widehat{\mathbf{i}_{i,j}})\right), \qquad \psi_j = \left(\psi_{j,0} + \frac{1}{2}\sum_{i=1}^{p}(1 - \widehat{\mathbf{i}_{i,j}})\widehat{a_{i,j}^2}\right),$$
$$\lambda_{i,j} = \lambda_{ij,0} - \ln A_j^{\max} - \frac{1}{2}\ln\widehat{\xi_0} + \frac{1}{2}\widehat{a_{i,j}\xi_0 a_{i,j}},$$

where $\iota_i = \mathrm{diag}(\mathbf{i}_{i,:})$.

The required moments are $\widehat{\Upsilon} = \mathrm{diag}(\alpha \circ \beta^{-1})$, $\widehat{\Xi_0} = \mathrm{diag}(\phi \circ \psi^{-1})$, $\widehat{\omega} = \frac{\vartheta}{\rho}$, $\widehat{\mathbf{i}_{i,j}} = \frac{1}{\lambda_{i,j}}$, and moments of truncated normal distribution are computed with respect to Appendix A.

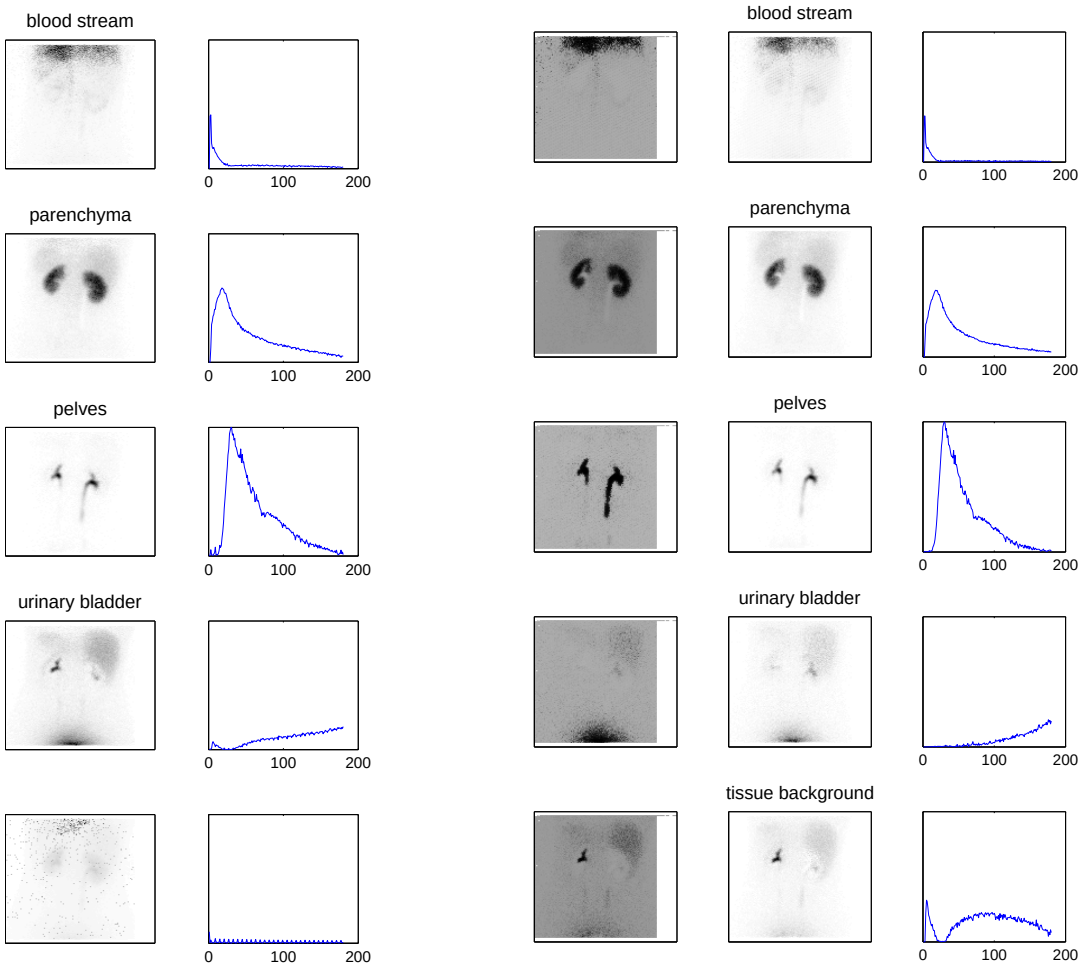


Figure 1: The results from the FA algorithm (left) and from the FAM algorithm (right)

4 Feasibility Study with Clinical Data

The previous algorithms were tested on a scintigraphic study. Factor images and factor curves were estimated in the case of FA and FAM algorithms; next, the resulting estimates and computed convolution kernels of parenchyma are studied.

4.1 Estimation of Factor Images and Curves

The first task is an estimation and separation of factors. Figure 1 shows the results from the FA algorithm, section 2, and from the FAM algorithm, section 3. From the left, FA estimates factor images, i.e. $\widehat{A}$, and factor curves, i.e. $\widehat{X}$; FAM estimates probability mask of factor images, i.e. $\widehat{\mathbf{i}}$, factor images, i.e. $\widehat{A}$, and factor curves, i.e. $\widehat{X}$. Both algorithms automatically estimated as the strongest factors blood background, renal parenchyma,

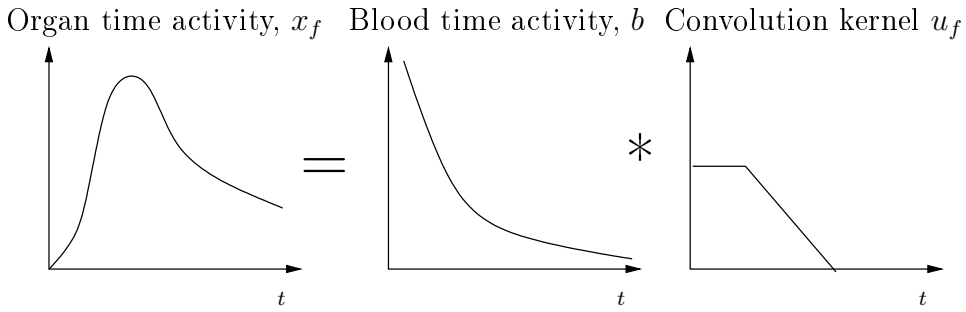


Figure 2: Theoretical decomposition of a factor curve

renal pelves, and urinary bladder; in addition, FAM estimated tissue background as the last significant factor.

The main differences between FA and FAM are in the beginning of the factor curves, especially of renal pelves and urinary bladder. In biological restriction, pelves curve should be at a zero level for the first $2-3$ minutes, i.e. $12-18$ frames. This restriction is well satisfied by FAM in contrast with FA with significant activity at the beginning of the curve. The same can be seen by urinary bladder; here, non-zero beginning is caused by improper separation of tissue background and bladder by FA algorithm. This zero-level-plateaus are very important from the biological view. In addition, this fact implies that the factor images of pelves and urinary bladder are undoubtedly better separated from tissue background by FAM then by FA.

4.2 Estimation of Convolution Kernel of Parenchyma

In the biological point of view, each time activity curve of factor is a convolution between blood and its specific convolution kernel [6, 2, 9]. Moreover, this convolution kernel is positive and its shape is shown in Figure 2. There should be a constant positive plateau and then linear or exponential decline to zero. From the length the plateau can be identified an important diagnostic coefficient - the transit time.

Figure 3 shows convolution kernels of parenchyma computed using Fourier transform. The result of FA is in the top, the result of FAM is the bottom row. In FA case, the peak at the beginning of the convolution kernel implies that separation of parenchyma and tissue background are not perfect [2]. From this point of view, FAM gives more appropriate results.

5 Discussion

The results presented in Section 4 suggest that factor analysis with integrated probability mask on factor images has a potential to improve the whole estimative procedure. However, more improvement is necessary for automatic estimation of diagnostic coefficients, which can be compared with experts. For example, the information from probabilistic mask $\widehat{\mathbf{i}}$ can be adopted for automatic selection of position of the single organs and consecutive computations. Study and usage of this fact is suggestion for future work.

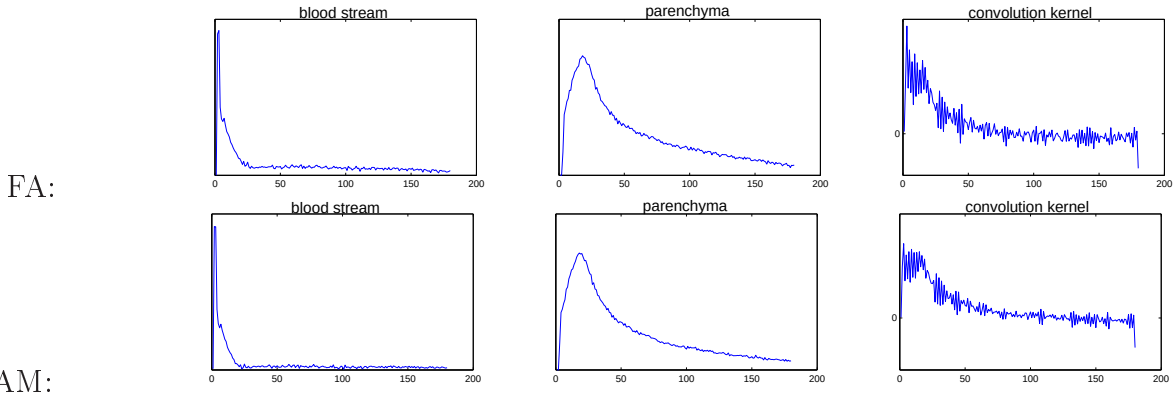


Figure 3: Convolution kernel of parenchyma (right) obtained from blood curve (left) and parenchyma curve (center).

Modeling of non-zero pixels as uniform distribution (14) is motivated by observation and seems to be better than modeling as normal distribution. However, more appropriate distribution or method should be used for modeling of histogram of the matrix A. It could lead to better separation of the factors.

With respect to study of non-zero priors of factor images, factor curves can be studied in the same way. A prior zero mean value of X is chosen due to computable reasons; nevertheless, more appropriate mean value can be computed [9].

6 Conclusion

A new model of factor images in functional analysis of scintigraphic dynamic sequences is proposed. The main addition is the dividing of pixels of factor images into informative and non-informative parts. The resulting algorithm is obtained using Variational Bayes method based on modeling parameters as independent components. Feasibility of solution is shown on clinical data from renal scintigraphy and compared with classical factor analysis where is demonstrated an improvements over previous methods. An automatic estimation of important diagnostic parameters will follow so as an extensive clinical study.

Appendix

A Moments of truncated Normal Distribution

Scalar truncated normal distribution

$$tN_x(x|\mu, r) = \alpha\sqrt{2}\exp\left(-\frac{(x-\mu)^2}{2r}\right), \; x > 0, \tag{15}$$

has moments

$$\widehat{x} = \mu + r\alpha\sqrt{2}\exp\left(-\frac{\mu^2}{2r}\right), \quad \widehat{x^2} = r + \mu\widehat{x},$$

where $\alpha^{-1} = \sqrt{\pi r}(1 - \mathrm{erf}(-\frac{\mu}{\sqrt{2r}})$ and erf is the error function.

References

[1] I. Buvat, H. Benali, and R. Di Paola. *Statistical distribution of factors and factor images in factor analysis of medical image sequences.* Physics in Medicine and Biology **43** (1998), 1695–1711.

[2] E. Durand, M. Blaufox, K. Britton, O. Carlsen, P. Cosgriff, E. Fine, J. Fleming, C. Nimmon, A. Piepsz, A. Prigent, et al. *International Scientific Committee of Radionuclides in Nephrourology (ISCORN) consensus on renal transit time measurements.* In 'Seminars in nuclear medicine', volume 38, 82–102. Elsevier, (2008).

[3] J. Fine and A. Pouse. *Asymptotic study of the multivariate functional model. application to the metric of choise in principal component analysis.* Statistics **23** (1992), 63–83.

[4] J. Fleming and P. Kemp. *A comparison of deconvolution and the Patlak-Rutland plot in renography analysis.* Journal of Nuclear Medicine **40** (1999), 1503.

[5] R. Klein, R. Beanlands, A. Adler, and R. deKemp. *Model-based factor analysis of dynamic sequences of cardiac positron emission tomography.* In 'Nuclear Science Symposium Conference Record, 2008. NSS'08. IEEE', 5198–5202. IEEE, (2009).

[6] A. Kuruc, W. Caldicott, and S. Treves. *An improved deconvolution technique for the calculation of renal retention functions.* Computers and Biomedical Research **15** (1982), 46–56.

[7] G. Lueck, T. Kim, P. Burns, and A. Martel. *Hepatic perfusion imaging using factor analysis of contrast enhanced ultrasound.* Medical Imaging, IEEE Transactions on **27** (2008), 1449–1457.

[8] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing.* Springer, (2005).

[9] O. Tichý. Integral models for dynamic renal scintigraphy, (2010). Thesis, FNSPE CTU.

[10] M. Šámal, M. Kárný, H. Sůrová, E. Maříková, and Z. Dienstbier. *Rotation to simple structure in factor analysis of dynamic radionuclide studies.* Physics in Medicine and Biology **32** (1987), 371–382.

[11] M. Šámal, M. Kárný, H. Sůrová, P. Pěnička, E. Maříková, and Z. Dienstbier. *On existence of unambiguous solution in factor analysis of dynamic studies.* Physics in Medicine and Biology **34** (1989), 223–228.

Some New Applications of Logistic Regression*

Tran Van Quang

3rd year of PGS, email: tran@vse.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. Logistic regression has been widely used in many areas. In this contribution, it is extended to two other applications. First, it is used as a tool for reducing the size of a hidden layer of an artificial neural network (ANN), which can be generalized to a universal instrument to find the optimal structure of an ANN. Second, the logistic regression is used to solve a multi-class partition task. In this case, when the number of input variables is fixed and the patterns are non-separable, the task can be solved via likelihood maximization with two fundamental extensions. First, Bayesian approach is used to formulate an alternative, a regularized optimization task with the corresponding objective function smooth and convex, which makes the problem easy to be solved. Then, the structure of multi-classifier is pruned to obtain the best model. The implementation of pruning process leads to binary optimization task, which is solved via fast simulated annealing heuristics. In both cases, the likelihood ratio test is used to the evaluation criterion to reach the objective.

Keywords: Logistic regression, LR test, ANN pruning, Regularization, Binary optimization, Multi-class classifier, FSA.

Abstrakt. Logistická regrese je široce využita v mnoha oblastech. V tomto příspěvku využití této regrese je rozšířena o další dvě možnosti. Nejdříve je využita jako nástroj úpravy velikosti skryté vrstvy umělé neuronové sítě, který dále lze zobecnit na univerzální nástroj k nalezení optimální struktury umělé neuronové sítě. Pak logistická regrese je využitá jako nástroj k řešení úlohy vícenásobné klasifikace. V tomto případě, když je počet vstupních proměnných je fixní a vzory jsou neseparabilní, tato úloha je řešitelná přes maximalizaci věrohodnosti s dvěma důležitými rozšířeními. Zaprvé, bayesovský přístup je použit k formulování alternativní účelové funkce, která je hladká a konvexní (tzv. regulovaná optimalizační úloha), což značně usnadňuje řešení problému. Následně struktura vícenásobného klasifikátoru bude vytříbena, aby byl zjištěn nejlepší model. Implementace vytříbení vede k binární optimalizační úloze, která se řeší pomocí heuristické metody rychlého simulovaného žíhání. V obou případech je používán test poměru věrohodností jako evaluační kritérium k dosažení vytyčeného cíle.

Klíčová slova: Logistická regrese, LR test, ANN ořezání, Regularizace, Binární optimalizace, Vícenásobný klasifikátor a Rychlé simulované žíhání

1 Introduction

The traditional probit regression proposed by Bliss [1] , which is based on the Gaussian normal distribution assumption, is a well-known two-class classifier. Later Berkson [2]

*This work has been supported by the grant OHK4-165/11 CTU in Prague

introduced the logistic regression based logistic distribution. From then on, logistic regression is widely used [4], [5], [9] and it has also been generalized to multi-class classifier [4], [9] and several tests have been invented to verify the relevance of a variable or a group of them. The most often used is the likelihood ratio (LR) test. In this work, first, the logistic regression and the LR test are used to examine if an input or a group of inputs are significant for the hierarchical decision process inside an ANN and by this way, it can help to eliminate redundant inputs or generate a hidden layer of a multilayer perceptron (MLP). Secondly, in this work, a multi-class classifier is developed. Unlike the one in Kukal and Vyšata [6], which is a soft multi-classifier with constrained gain together with maximum sensitivity and specificity and designs its learning as multi-criteria optimization task, this one can deal with the difficulties resulting from the fact that the number of input variables is higher than two, but fixed and the patterns are non-separable. The problem is solved via likelihood maximization with two important extensions. First, the Bayesian approach is used to set an alternative, a regularized optimization problem, which is easy to be solved because the objective function smooth and convex. The second extension is to prune the structure of multi-classifier to obtain the best model by solving a binary optimization problem with the help of the heuristic method called fast simulated annealing. In this very case, the likelihood ratio test is also used to achieve the best model.

2 Logistic regression with binary outcome

In this section, first, I will go over the general binary response index model, then I will show several versions of the binary index model with the emphasis on the logistic regression. After that, I will show how parameters of the logistic regression are estimated. Finally, in this part, I will show the essence of the LR test.

2.1 The general binary response index model

Let's suppose a model with m real inputs $\boldsymbol{x}$ and a single binary output y in the form: $y = \mathrm{h}(\boldsymbol{x}\beta + e)$, where $\mathrm{h}(z) = 1$ if $z \ \ 0$ and $\mathrm{h}(z) = 0$ if $z \leq 0$ is the Heaviside's unit step function and $\boldsymbol{x}$, $\beta \ \in R^{m+1}$, $x_0 = 1$, e is a continuous random variable with positive and symmetric probability density function $\mathrm{g}(z)$ around zero. Its cumulative distribution function is:

$$\mathrm{G}(z) = \int_{-\infty}^{z} \mathrm{g}(u)\mathrm{d}u. \tag{1}$$

By definition, the output variable y is of stochastic nature and it can be defined through probability as follows:

$$\mathrm{p}(\boldsymbol{x}) = \mathrm{p}(y = 1 \mid \boldsymbol{x}) = \mathrm{p}(\boldsymbol{x}\beta + e > 0) = \mathrm{p}(\boldsymbol{x}\beta > -e) = 1 - \mathrm{G}(-\boldsymbol{x}\beta) = \mathrm{G}(\boldsymbol{x}\beta). \tag{2}$$

This is the well-known formula for binary model with logistic probability density function.

2.2 Special cases of binary model

The binary model was first introduced by Bliss [1] in 1934 as the probit model with the following cumulative distribution function and the corresponding probability density function are:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u, \tag{3}$$

$$\mathrm{g}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right). \tag{4}$$

In 1944, Berkson [2] presented the logit model with logistic cumulative distribution function and the corresponding density function of logistic distribution defined as follows:

$$\mathrm{G}(z) = \frac{1}{1+\exp(z)}, \tag{5}$$

$$\mathrm{g}(z) = \frac{\exp(-z)}{(1+\exp(-z))^2}. \tag{6}$$

The logistic model is the most frequently used model in many applications. Since the density function of logistic distribution has fatter tails than the density function of normal distribution, it is often used when the normal distribution assumption is not appropriate. Other distributions also can be used for the binary response index model. One of them is the so called Cauchy distribution and the corresponding cumulative distribution function defined as follows:

$$\mathrm{g}(z) = \frac{1}{\pi(1+z^2)}, \tag{7}$$

$$\mathrm{G}(z) = \frac{1}{2} + \frac{1}{\pi}\arctan. \tag{8}$$

Like logistic distribution, Cauchy distribution also has heavy tails and provides some interesting features for modelling binary responses. The existence of various probability density functions suitable for binary response modelling provides a unique opportunity to choose the most appropriate one for this purpose.

2.3 The estimation of parameters of a binary response model

The parameters of a binary response model are estimated by the maximum likelihood method. Let's N be the number of observations, and $(\boldsymbol{x}_k, y_k)$ be the individual observation for $k = 1, ..., N$. The density of y_k for individual $\boldsymbol{x}_k$ is:

$$\mathrm{f}(y_k|\boldsymbol{x}_k\beta) = [\mathrm{G}(\boldsymbol{x}_k\beta)]^{y_k} [1 - \mathrm{G}(\boldsymbol{x}_k\beta)]^{1-y_k}. \tag{9}$$

The logarithmic likelihood function over all observations is defined as:

$$\mathrm{L}(\beta) = \sum_{k=1}^{N} (y_k \mathrm{log}\mathrm{G}(\boldsymbol{x}_k\beta) + (1 - y_k)\mathrm{log}[1 - \mathrm{G}(\boldsymbol{x}_k\beta)]). \tag{10}$$

The point estimate of the vector of parameters is the solution of the maximization problem of the objective function L on a closed convex domain $\mathbf{D}$ as $\boldsymbol{b} = \hat{\beta} = \text{argmax}_{\beta \in \text{D}} \text{L}(\beta)$. It is necessary to remind that $\mathbf{b}$ exists but is not unique. Replacing $\mathbf{D}$ by an open set $\mathbf{R}^n$ will be problematic when the observations are separable, which is the ideal case for classifier tuning but not for parameter estimation. The analysis of asymptotic variance begins with matrix:

$$\mathbf{U} = \sum_{k=1}^{N} \frac{\text{g}(\boldsymbol{x}_k \boldsymbol{b}) \boldsymbol{x}_k \boldsymbol{x}_k^T}{\text{G}(\boldsymbol{x}_k \boldsymbol{b})(1 - \text{G}(\boldsymbol{x}_k \boldsymbol{b}))}. \tag{11}$$

When matrix $\boldsymbol{U}$ is regular, it is positive definite and the asymptotic variance of estimate $\boldsymbol{b}$ is:

$$\text{Avar}(\boldsymbol{b}) = \boldsymbol{V} = \boldsymbol{U}^{-1}. \tag{12}$$

The asymptotic standard error of estimate $\boldsymbol{b}$ is:

$$\text{Astd}(\boldsymbol{b}) = \boldsymbol{s} = \text{diag}\,(\boldsymbol{V})^{1/2}. \tag{13}$$

The corresponding approximate 95% confidence interval is:

$$\beta \in [\boldsymbol{b} - \mathbf{1.96}\boldsymbol{s}, \boldsymbol{b} + \mathbf{1.96}\boldsymbol{s}]. \tag{14}$$

The confidence interval is important for the final report of the significance of the estimates ant it is sensitive to the singularity of matrix $\boldsymbol{U}$.

2.4 Hypothesis testing

There are several methods to test for the significance of some input or a group of inputs. One of them is the likelihood ratio test which compares a given model to its sub-models. Let's $\boldsymbol{r} = (1, r_1, ..., r_m) \in \{0, 1\}^{m+1}$ be the selection vector which describes whether the corresponding components of vector $\boldsymbol{x}$ will be present in the model. Vector $\boldsymbol{r}$ divides vector $\boldsymbol{x}$ into 2 vectors: vector $\boldsymbol{u}$ of active inputs with $\boldsymbol{u} \in \boldsymbol{R}^{K+1}$ and vector $\boldsymbol{v}$ of eliminated inputs with $\boldsymbol{v} \in \boldsymbol{R}^{Q}$. It is obvious that $K + Q = m$. Similarly, vector of parameters β can be decomposed into $\mu \in \boldsymbol{R}^{K+1}$ and $\eta \in \boldsymbol{R}^{Q}$. If $Q = 0$, then $K = m$ and we get the full model:

$$\text{p}(\boldsymbol{x}) = \text{G}(\boldsymbol{x}\beta) = \text{G}(\boldsymbol{u}\mu + \boldsymbol{v}\eta) \tag{15}$$

with the optimal likelihood value L_{full}. And if $Q > 0$ and $K < m$, then the model is reduced to a sub-model in the form:

$$\text{p}(\boldsymbol{u}) = \text{G}(\boldsymbol{u}\mu) \tag{16}$$

with the optimal likelihood value L_{sub}. The likelihood ratio test examines the validity of the null hypothesis $H_0 : \eta = \mathbf{0}$ against the alternative hypothesis $H_A : \eta \neq \mathbf{0}$. The test statistic is calculated as:

$$LR = 2(L_{\text{full}} - L_{\text{sub}}) \tag{17}$$

and has the chi-squared distribution with Q degrees of freedom. The corresponding p-value can be calculated as

$$p_{\text{value}} = 1 - \text{F}_Q(LR), \tag{18}$$

where F_Q is the cumulative distribution function of chi-squared distribution. The LR test evaluates how good the sub-model is compared to the full model. There are two important cases. The first one is when $K = 0$ and $Q = m$ and we get a model only with a constant. In this case, the maximum likelihood procedure is trivial:

$$\mathrm{p}(\boldsymbol{u}) = \mathrm{G}(\mu_0) = p_\mathrm{c} = \frac{1}{N}\sum_{k=1}^{N} y_k. \tag{19}$$

The likelihood value of the model with a constant and the p-value of the model with a constant compared to the full model are as follows:

$$L_\mathrm{const} = N(p_\mathrm{c}\log p_\mathrm{c}) + (1-p_\mathrm{c})(1-\log p_\mathrm{c}) \tag{20}$$

$$p_0 = 1 - F_m(2(L_\mathrm{full} - L_\mathrm{const})). \tag{21}$$

The lower the value of p_0 is, the higher the significance of a model is and various models can be sortd by p_0. The second important case is when $K = m - 1$ and $Q = 1$. In this case, we can compare the full model with a model with k-th input taken away and the optimal likelihood value is denoted as L_k. We can test the significance of k-th input by using LR test again. The null hypothesis H_0 is : $\beta_k = 0$ against the alternative H_A : $\beta_k \neq 0$. The p-value is:

$$p_1 = 1 - F_1(2(L_\mathrm{full} - L_k)). \tag{22}$$

We can use stepwise strategy to eliminate irrelevant inputs to get the model, in which each parameter is significant in the sense that $p_k < \alpha$ as well as $p_0 < \alpha$. If there are more models satisfied this criterion, the best on is the model with minimum p_0.

3 Application of binary model to ANN

In this part, the binary logistic model and its selection with the help of LR test are used first to prune a given ANN and then to build an optimal hidden layer in an ANN.

3.1 Pruning an ANN by using logistic model

Let's have a hierarchical ANN classifier with a single hidden layer and one binary output. The hidden layer is supposed to be fixed, i. e. the weights from input layer are constant without any opportunity to learn. The hidden layer can be designed as a result of systematic or sophisticated preprocessing. Under this setting, the logistic regression can be applied as a sophisticated preprocessor. Using the LR test, it prunes those redundant connections between the hidden layer and the output layer. An ANN with binary output can be expressed as follows:

$$y = \mathrm{h}\left(\sum_{k=0}^{N} \beta_k \varphi_k(\boldsymbol{x}) + e\right). \tag{23}$$

It is similar to the logistic model $\mathrm{p}(\boldsymbol{x}) = \mathrm{G}\left(\Phi(\boldsymbol{x})\beta\right)$, where instead of $\boldsymbol{x}$ we have $\Phi(\boldsymbol{x})$ and $\varphi_0(\boldsymbol{x})$. The domain of optimization must be closed and convex and should be defined

as $\mathbf{D} = \left\{\beta \in \boldsymbol{R}^{N+1} | \parallel \beta \parallel \leq \rho\right\}$. Statistical meaning of ANN pruning is to find the best sub-model in the defined domain and log-likelihood optimization is applied in the inner loop. There are 2^N possibilities how to design selection vector $\boldsymbol{r}$ and any integer heuristics (FSA, GO) can be used to find the best pruning as a global minimum of objective function:

$$\mathrm{q}(\boldsymbol{r}) = p_0. \tag{24}$$

A stronger form of the previous problem is the one when we require the significance of parameters as a set constraints:

$$\mathrm{p}_k(\boldsymbol{r}) \leq \alpha, \text{for} \quad k = 0, 1, ..., N. \tag{25}$$

The result of this integer optimization problem is an ANN classifier in the form (23) with only significant weights, therefore the network has only $N_{\text{opt}} \leq N$ neurons in the hidden layer.

3.2 Hidden layer of an ANN via logistic regression

Logistic regression can be also used as a tool for building a hidden layer of ANN. In order to do so, first, we need to define the notions a local minimum and a degenerated point in the problem of constrained binary optimization. In this problem, the searching domain is set $\mathbf{S} = \{0, 1\}^{N+1}$. The feasible domain is set $\mathbf{T} = \{\boldsymbol{r} \in \mathbf{S} | r_0 = 1, \forall k = 0, ..., N : p_k < \alpha\}$. The feasible neighborhood of a point $\boldsymbol{r} \in \mathbf{T}$ is set $\mathbf{N}(\boldsymbol{r}) = \{\boldsymbol{q} \in \mathbf{T} | \parallel \boldsymbol{q} - \boldsymbol{r} \parallel = 1\}$. Then, a local minimum of function q on the feasible domain is any point $\boldsymbol{r}_{\text{loc}} \in \mathbf{T}$ satisfying

$$\mathrm{q}(\boldsymbol{r}_{\text{loc}}) < \min\left\{\mathrm{q}(\boldsymbol{r}) | \boldsymbol{r} \in \mathbf{T}(\boldsymbol{r}_{\text{loc}})\right\}. \tag{26}$$

Similarly, a degenerated point of function q on the feasible domain is any point $\boldsymbol{r}_{\text{deg}} \in \mathbf{T}$ satisfying

$$\mathrm{q}(\boldsymbol{r}_{\text{deg}}) = \min\left\{\mathrm{q}(\boldsymbol{r}) | \boldsymbol{r} \in \mathbf{T}(\boldsymbol{r}_{\text{deg}})\right\}. \tag{27}$$

The global minimum of (24) satisfying (25) is either a local minimum or a degenerated point. This method is based on finding local minima and degenerated points. They are easily found as a result of random walk or steepest descent algorithm from a random initial point. The feasibility is solved by a penalization of objective function (24) with respect to constraint (25).By repeating the local searching we get a set of various solutions. The final three layer ANN is formed by them through the formula:

$$\mathrm{p}(\boldsymbol{x}) = \mathrm{p}(y = 1 \mid \boldsymbol{x}) = \mathrm{G}_0\left(w_0 + \sum_{k=1}^{H} w_k \mathrm{G}_k(\boldsymbol{x}\beta_k)\right), \tag{28}$$

where H is the number of local minima or degenerated points found by each of them is characterized by selection vector $\boldsymbol{r}_k$ and complete vector β_k and the unknown vector of weights $\boldsymbol{w}$ is estimated with the help of logistic regression. The ANN built upon the logistic regression was used to predict the probability of increase of the number of sunspots. The annual dataset from 1700 to 1987 was used. As input, the relative differences was introduced to the ANN. For the pruning study, the number of input was 30, and a global solution was found which consists of 1 significant positive and 5 negative weights. The optimal topology is 29 - 6 - 1. In the case of the hierarchical approach, 12 various local minima were found and the final topology is 29 - 12 - 1.

4 Multi-class logistic regression

The binary response model in the previous part is in fact a two-class classifier which can be extended to classify more classes. Let $n, N \in \mathbf{N}$ be the number of properties and the number of classes. Then, the classifier f has n inputs and N outputs. It realizes a partition among N classes and can be described as a function f: $\mathbf{R}^N \rightarrow \boldsymbol{Q}_N$ where $\boldsymbol{Q}_N = \{\, \boldsymbol{y} \in \mathbf{P}_N | \parallel \boldsymbol{y} \parallel_1 = 1 \,\} \subseteq \mathbf{P}_N = [0.1]^N$. The original logit model can be generalized for N classes as follows:

$$y_i = \frac{\exp(s_i)}{\sum_{k=1}^{N} \exp(s_k)}, \text{where} \quad s_i = \sum_{j=1}^{n} v_{i,j} x_j \quad \text{for } i = 1, ..., N, \tag{29}$$

with $\boldsymbol{x} \in \mathbf{R}^N$, $\boldsymbol{V} \in \mathbf{R}^{N(n+1)}$ and $x_0 = 1$. Since there are N classes, when we identify $N - 1$ classes, the last one will be identified as well, therefore we can rearrange formula [29] to the following form:

$$y_i = \frac{\exp(s_i - s_1)}{\sum_{k=1}^{N} \exp(s_k - s_1)} = \frac{\exp(h_i)}{\sum_{k=1}^{N} \exp(h_k)} \tag{30}$$

$$\text{where} \quad h_i = \sum_{j=1}^{n} (v_{i,j} - v_{1,j}) x_j = \sum_{j=1}^{n} w_{i,j} x_j \quad \text{for } i = 1, ..., N, \tag{31}$$

where $\boldsymbol{W} \in \mathbf{R}^{N(n+1)}$ and $w_{1,j} = 0$ for $j = 0, ..., n$. Model (30) with unknown matrix $\boldsymbol{W}$ has only $(N \check{} 1)(n + 1)$ free parameters. The sensitivity of this model to input variables depends only on reduced matrix $\boldsymbol{W}_{\text{red}}$, which is defined via relationship

$$\boldsymbol{W} = \begin{bmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{b} & \boldsymbol{W}_{\text{red}} \end{bmatrix}, \tag{32}$$

where $\boldsymbol{W}_{\text{red}} \in \mathbf{R}^{n(N-1)}$ and $\boldsymbol{b} \in \mathbf{R}^{N-1}$ is a bias vector for N classes.

4.1 Maximum likelihood estimate and its regularization

As usual, the estimation of model parameters is frequently performed via maximization of its likelihood function. Let m be the number of patterns, $(\boldsymbol{x}_k, c_k)$ be a pattern and $c_k = 1, ..., N$ be the class index, then the MLE estimate of $\boldsymbol{W}$ is:

$$\Phi(\boldsymbol{W}) = -\ln \mathrm{L}(\boldsymbol{W}) = \min, \tag{33}$$

where L is the likelihood function for a given pattern set. Since $\boldsymbol{y}_k = \mathrm{f}(\boldsymbol{x}_k)$ is a vector of class membership probability and denote its i^{th} component as $(\boldsymbol{y}_k)_i$, then

$$\Phi(\boldsymbol{W}) = \sum_{k=1}^{m} \ln(\boldsymbol{y}_k)_{c_k}, \tag{34}$$

which is smooth and convex, thus a unimodal function. But in many cases, the minimum of (34) does not exist and the norm of W approaches infinity during search process. In such cases task regularization is needed. The regularization is based on statistical theory, we can convert the task to M-estimate finding and testing. Bayesian approach is used to regularize. Based on an a priori knowledge of distribution of $\boldsymbol{W}_{\mathrm{red}}$ while without this knowledge for the bias vector $\boldsymbol{b}$, the conditional probability defined in (11) is:

$$\Psi(\boldsymbol{W}) = \frac{1}{2\sigma^2} \| \boldsymbol{W}_{\mathrm{red}} \|_{\mathrm{F}}^2 - \sum_{k=1}^{m} \ln(\boldsymbol{y}_k)_{c_k} = \min . \tag{35}$$

Here, $\| \ldots \|_F$ is Frobenius norm and resulting function in (35) is smooth and convex, again. But in this case, the optimum of (35) exists in all cases. After some rearrangement we get:

$$\Psi(\boldsymbol{W}) = \sum_{k=1}^{m} \left(- \ln(\boldsymbol{y}_k)_{c_k} + \frac{1}{2m\sigma^2} \| \boldsymbol{W}_{\mathrm{red}} \|_{\mathrm{F}}^2 \right) = \sum_{k=1}^{m} \Psi_k(\boldsymbol{W}) = \min, \tag{36}$$

where $\boldsymbol{W}_{\mathrm{red}}$ is defined in (32). We can use the theory of M-estimates for model and sub-model testing. The minimization of (35) is easy to perform in the Matlab environment using functions *fminunc* or *fminsearch*.

4.2 Model and sub-model testing

The likelihood ratio test can be used to compare the full model with its sub-models, as it was in the previous section. If Ψ and Ψ_0 are the optimal values from (35) for a model and a model only with a constant respectively, then the significance of the model different with the one with a constant is:

$$p_0 = 1 - \mathrm{F}_Q(2(\Psi_0 - \Psi)), \quad \text{where } \Psi_0 = -m \sum_{k=1}^{N} y_k^0 \ln y_k^0, \quad y_k^0 = \frac{m_k}{m} \tag{37}$$

for $k = 1, ..., N$. The lower value of p_0 indicates the higher significance of given model and various models can be ordered according to p_0 to obtain the best one.

4.3 Model pruning as a binary optimization task

The minimization of (36) is easy task of convex programming but it is only a subject of inner loop. Statistical meaning of model pruning is in finding of the best model or its control matrix $\boldsymbol{B}$, respectively. There are $2^{(N\check{}1)n}$ possibilities how to design the control matrix $\boldsymbol{B}$. We find the best pruning as a global minimum of this objective function:

$$\mathrm{q}(\boldsymbol{W}) = \log_{10} p_0, \tag{38}$$

The probability of state changing from $\boldsymbol{W}$ to $\boldsymbol{W}_{\mathrm{new}}$ is:

$$p_{\mathrm{new}} = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{\mathrm{q}(\boldsymbol{W}) - \mathrm{q}(\boldsymbol{W}_{\mathrm{new}})}{T_k} \tag{39}$$

where T_k is the cooling strategy set by this formula:

$$T_k = T_0 \left(1 + \frac{k}{n_0}\right)^{-1} \tag{40}$$

where T_0, n_0 and k are initial temperature, index scale and index of state change respectively.

4.4 Iris flower classification task

The dataset from [3] consists of 50 samples from each of three species of Iris flowers (Iris Setosa, Iris Virginica and Iris Versicolor). Four features were measured on each pattern. They are: sepal length, sepal width, petal length and petal width. Based on the combination of the four features, Fisher developed a linear discriminant model to determine which species they are. In our setting, the multi-classification task has N = 3 classes, n = 4 inputs, m = 150 patterns and 10 free parameters including biases. Thus $0 < Q < 8$ and there are only $2^8 = 256$ states for binary optimization by FSA. We used $T_0 = 0.01$, $n_0 = 10$, $p_{\text{mut}} = 0.2$ to reach global optimum after less than 200 function evaluations. The apriori value of parameter σ changes the optimum solution of multi-classification task: Q, $\log_{10} p_0$, n_{err}, $\| \boldsymbol{W} \|_{\text{red}}$ as the number of free parameters, the quality of model, the number of miss-classified patterns and the sensitivity to input signals. When σ is small, the model is over-regularized and imprecise. When $\sigma > 2$, the number of classification errors is suppressed to 3 from 150 and the multi-classifier use only four active weights from three inputs (excluding sepal length), but the sensitivity to inputs is higher than 20. In the medium range of regularization, the multi-classifier used only three active weights from two inputs (only petal length and width) with slightly increased number of classification errors but with decreased sensitivity to inputs. The classification problem then was proceeded with a quadratic preprocessing by adding squared terms and cross terms of the 4 original inputs. By doing so, we obtained a new data set with 3 classes, 14 inputs and 150 patterns. The number of states has increased to $Q = 2^2 8$. We kept the parameters of FSA and the value of σ unchanged and monitored the same set of output variables: Q, $\log_{10} p_0$, n_{err}, $\| \boldsymbol{W} \|_{\text{red}}$ as in the case without preprocessing. The most interesting feature in this case is that the number of misclassification was suppressed to 0. At the same time, as the misclassification disappeared, the sensitivity to inputs of the model increased significantly (see table 1).

5 Conclusion

Logistic regression has had a wide scope of applications. In this paper, the use is extended into two directions. First, it is used to prune a given ANN and then to build an optimal hidden layer in an ANN. The cutting off the redundant connections between the hidden layer and the output layer is based on the LR test. In the same fashion it is also used to build a optimal hidden layer in an ANN. Then logistic regression was used to construct a regularized multi-classifier, whose parameter estimation was converted to convex optimization task for free minimization. Even in this case, the likelihood ratio test is used to select the best sub-model with the help of Fast Simulated Annealing. This

Table 1: The results of Iris classification using logistic approach

	without preprocessing				with preprocessing			
σ	Q	$\log_{10} p_0$	n_{err}	$\|\boldsymbol{W}_{\text{red}}\|_{\text{F}}$	Q	$\log_{10} p_0$	n_{err}	$\|\boldsymbol{W}_{\text{red}}\|_{\text{F}}$
0.1	3	-7.32	24	0.82	2	-43.54	7	0.65
0.2	3	-33.64	5	3.31	2	-57.43	6	1.20
0.5	3	-56.25	6	8.26	2	-64.10	5	1.95
1	3	-62.70	6	12.55	2	-66.57	4	5.57
2	4	-65.16	3	20.46	2	-67.54	4	6.41
5	4	-66.22	3	28.05	2	-67.84	4	6.82
10	4	-66.41	3	31.32	2	-67.89	4	6.76
20	4	-66.47	3	33.81	4	-68.31	1	6.84
50	4	-66.48	3	37.00	4	-68.96	0	3199.6
100	4	-66.48	3	39.16	4	-69.21	0	4986.8
1000	4	-66.49	3	70.47	4	-69.34	0	10678.0
∞	4	-66.49	3	70.55	4	$-\infty$	0	∞

regularized multi-classifier then was employed to perform iris flower classification task as well as to examine the effect of model structure pruning. The novel method and program library in the Matlab environment can act as a universal tool for multi-classification.

References

[1] C. I. Bliss. *The method of probits.* Science **79**, No.2037, (1934), 38—39.

[2] J. Berkson. *Application of the logistic function to bio-assay.* J. Am. Stat. Assoc. **39** (1944), 357—365.

[3] R. A. Fisher. *The Use of Multiple Measurements in Taxonomic Problems.* Annals of Eugenics **7** (1936), 179—188.

[4] J. M. Hilbe. *Logistic Regression Models.* Chapman & Hall - CRC Press (2009).

[5] D. W. Hosmer, L. Stanley, *Applied Logistic Regression.* New York; Chichester, Wiley, (2000).

[6] J. Kukal , O. Vyšata. *Learning of Soft Classifier via Differential Evolution.* Proc. 14th Int. Conf. on Soft Computing, Mendel 2008, 181–185, VUT Brno, Brno, (2008).

[7] Q. V. Tran, J. Kukal, J. Kalčevová, J. Boštík. *Logistic Regression as Bridge Between Statistics and Artificiall Inteligence.* In Proc. 16th Int. Conf. on Soft Computing, Mendel 2010, 491–494, VUT Brno, Brno (2010).

[8] Q. V. Tran, J. Kukal, M. Mojzes. *Multi-class logistic model and its pruning.* In Proc. 17th Int. Conf. on Soft Computing, Mendel 2011, VUT Brno, Brno (2011).

[9] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press (2002).

Transversality Condition in Sufficient Stochastic Maximum Principle

Petr Veverka

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

advisor: Bohdan Maslowski,
Department of Probability and Mathematical Statistics, MFF UK

Abstract. In this article, the sufficient Pontryagin's maximum principle for infinite horizon discounted stochastic control problem is given. The sufficiency is ensured by an additional assumption of concavity of the Hamiltonian function. In the paper, it is assumed that the control domain U is a convex set and the control enters also the diffusion part of the state equation. Due to our setting, the Hamiltonian function has to be modified using an additional term coming from Lyapunov function for the FBSDE system. The result of this paper extends the one in [17] where the knowledge of the terminal condition of the associated BSDE is assumed. In this paper, to overcome this unrealistic assumption, we establish a so called transversality condition. In the end, we apply the result to an example from finance with known solution to conclude that our approach gives the same result.

Keywords: Stochastic maximum principle, discounted control problem, BSDE, transversality condition

Abstrakt. V tomto článku se zabýváme stochastickým principem maxima ve smyslu postačující podmínky pro optimalitu řízení, která je zajištěna dodatečný předpokladem konkavity Hamiltonianu. Předpokládáme, že řízení je z konvexního stavového prostoru a že též vstupuje do difuzního členu stavové rovnice. Díky tvaru uvažovaného funkcionálu modifikujeme Hamiltonian dodatečným členem, který přichází z Ljapunovské funkce pro řízenou FBSDE. Tento článek rozšiřuje výsledek uvedený v [17], kde autor předpokládá znalost koncové podmínky příslušné BSDE. Tento v praxi těžko zaručitelný předpoklad je v tomto článku nahrazen příslušnou podmínkou transverzality. Na závěr ukážeme užití stochastického principu maxima na úlohu z financí, u které je řešení známo a pro kterou dává předkládaná metoda stejný výsledek.

Klíčová slova: Stochastický princip maxima, diskontovaná úloha řízení, BSDE, podmínka transverzality

1 Introduction

In this paper, the discounted stochastic control problem is considered. This kind of problem is very popular and plentifully used in many domains, especially in stochastic finance since it leads to maximizing the average discounted agent's utility. The approach to the solution here is the maximum principle which, in deterministic setting, was formulated in 1950s by the group of L.S.Pontryagin. For diffusions, the maximum principle has been studied by many researchers. The earliest versions of a maximum principle for such

process were given by Kushner [7] and Bismut [8]. Further progress on the subject was subsequently made by Bensoussan [9], Peng [10], and Cadenillas and Haussmann [12]. Originally, the main technical tool used when considering maximum principle was the calculus of variations which was not easy to apply to real examples and was difficult to simulate. This was the reason why the approach via maximum principle was rather theoretical and discomfited by the dynamic programming approach. The turning point which led to its intensive study was the paper [4] by Pardoux and Peng who formulated the general problem of Backward Stochastic Differential Equation (BSDE in short) and proved the existence and uniqueness theorems. BSDE's provide an elegant and easy-to-handle tool to describe the adjoint (shadow price) processes to the control problem and to formulate the maximum principle using the Hamiltonian function. For diffusions with jumps, a necessary maximum principle on the finite time horizon was formulated by Tang and Li [13] whereas sufficient optimality conditions on finite time horizon were specified by Øksendal, Sulem and Framstad [1].

The paper is organized as follows: in the second section, some known results on Forward-Backward Stochastic Differential Equations with infinite tim e horizon are provided. The formulation of the discounted problem is in the third section. Fourth section contains the main result of the paper - the formulation and proof of the sufficient infinite time maximum principle for the discounted problem. In the last section, one example on agent's optimal consumption with known solution due to [16] is used to be compared with our approach.

2 Preliminaries

We are given a basic probability space $\big(\Omega, \mathscr{F}, \mathbf{P}\big)$, $\mathbb{R}^d$-valued standard Wiener process $W = \big(W_t\big)_{t\geq 0}$. Let $\big(\mathscr{F}_t^W\big)_{t\geq 0}$ be the canonical filtration of W, i.e. $\mathscr{F}_t^W = \sigma\big(W_s; s \leq t\big)$, and $\big(\mathscr{F}_t\big)_{t\geq 0}$ be its $\mathbf{P}-$null sets augmentation. We denote $\mathscr{F}_\infty = \bigvee_{t\geq 0} \mathscr{F}_t \subset \mathscr{F}$. Further, to simplify the notation, we write just 'a.s.' instead of '$\mathbf{P}$ -a.s.'. We denote $|\cdot|$ and $||\cdot||$ the Euclidean norms in $\mathbb{R}^n$ and $\mathbb{R}^{n\times d}$ respectively.

3 Formulation of the problem

3.1 Controlled state equation

The controlled state process $(X_t)_{t\geq 0}$ is a strong solution to the following controlled SDE on $\mathbb{R}_+$

$$\begin{aligned} dX_t &= b(X_t, u_t, \omega)dt + \sigma(X_t, u_t, \omega)dW_t, \quad \forall t \geq 0 \text{ a.s.} \\ X_0 &= x, \end{aligned} \tag{1}$$

where U is a compact convex subset of $\mathbb{R}^k$, the random functions $b : \mathbb{R}^n \times U \times \Omega \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \times U \times \Omega \to \mathbb{R}^{n\times d}$ are continuous in variables (x, u). Further we assume that

b and σ satisfy some assumptions ensuring the existence of strong solution to (1). For example, these can be

$$\langle x_1 - x_2, b(x_1, u, \omega) - b(x_2, u, \omega)\rangle \leq \mu|x_1 - x_2|^2, \tag{2}$$
$$|\sigma(x_1, u, \omega) - \sigma(x_2, u, \omega)| \leq c|x_1 - x_2|, \tag{3}$$
$$|b(x, u, \omega)| + |\sigma(x, u, \omega)| \leq K(1 + |x| + |u|), \tag{4}$$

for every $x, x_1, x_2 \in \mathbb{R}^n,\ u \in U$ and some constants $\mu \in \mathbb{R},\ c, K > 0$. The above conditions hold $a.s.$ The condition (2) means some kind of monotonicity of b in x, (3) is the standard (global) Lipschitz condition, (4) controls growth of b, σ in (x, u) as most as linearly. In the latter, we omit the notation of the dependence on ω and we denote as u both the element of the set U and the admissible control process $u = (u_t)_{t\geq 0}$ as defined bellow.

We denote as $\mathcal{U}_{ad}$ the set of all admissible controls which satisfy

$$\mathcal{U}_{ad} = \left\{u = \big(u_t\big)_{t\geq 0} : u \in \mathbf{L}^2_{\mathscr{F}}\big(\mathbb{R}_+; U\big)\right\}, \tag{5}$$

where $\mathbf{L}^2_{\mathscr{F}}\big(\mathbb{R}_+; U\big)$ denotes the Hilbert space of $\big(\mathscr{F}_t\big)$-adapted, $U-$valued processes u with $\mathbf{E}\int_0^{+\infty} |u_t|^2 dt < +\infty$. Any process $u \in \mathcal{U}_{ad}$ is called an admissible control. The functional considered is of the form

$$J(u) = \mathbf{E}\int_0^{+\infty} e^{-\beta t} f(X_t, u_t) dt, \tag{6}$$

where $f : \mathbb{R}^n \times U \to \mathbb{R}$ is the penalization (or appreciation) function over $\mathbb{R}_+$ such that $J(u)$ converges for all every admissible control. $\beta > 0$ is the discount factor.
Further, we define the cost function v by

$$v = \sup_{u\in\mathcal{U}_{ad}} J(u). \tag{7}$$

The goal is to find such a strategy $u^* \in \mathcal{U}_{ad}$ so that the supremum in (7) is attained in u^*, i.e. $v = J(u^*)$.

3.2 Hamiltonian of the system

We define a generalized Hamiltonian function $\mathcal{H}$ associated to control problem (1) - (7) by $\mathcal{H} : \mathbb{R}^n \times U \times \mathbb{R}^n \times \mathbb{R}^{n\times d} \to \mathbb{R}$ and

$$\mathcal{H}(x, u, y, z) = \big\langle b(x, u), y\big\rangle + Tr(\sigma(x, u)'z) + f(x, u) - \beta\big\langle x, y\big\rangle, \tag{8}$$

where $\big\langle \cdot, \cdot \big\rangle$ denotes inner product in $\mathbb{R}^n$, z' is the transpose of z and $Tr(\cdot)$ denotes trace of a matrix in $\mathbb{R}^{d\times d}$. The Hamiltonian is an analogy of the Lagrange function in the theory of constrained optimization since the variables y and z can be viewed as 'generalized Lagrange multiplicators' and the functions b and σ as the constraints for the dynamics of the space process X_t. The additional term $-\beta\big\langle x, y\big\rangle$ comes up from the

Lyapunov function of the FBSDE system, see [11]. We note that the extremal point of $\mathcal{H}$ w.r.t. u does not depend on the last term $-\beta\langle x, y\rangle$ and neither the concavity/convexity w.r.t. (x, u).

We further suppose that $\mathcal{H}$ is differentiable in x (with the gradient denoted as $\nabla_x \mathcal{H}$) and we consider the following BSDE

$$Y_t = \int_t^{+\infty} \nabla_x \mathcal{H}(X_s, u_s, Y_s, Z_s)ds - \int_t^{+\infty} Z_s dW_s, \quad \forall t \geq 0 \text{ a.s..}$$

3.3 Some discussion of FBSDE on infinite time horizon

In the previous subsection, we have seen that when applying the Hamiltonian formalism to stochastic control problems, the class of Forward-Backward Stochastic Differential Equations (FBSDE in short) naturally arises in form of a partially-coupled system of the state (forward) equation for the controlled diffusion and the adjoint backward equation for 'generalized Lagrange multiplicators'. Forward-Backward stochastic differential systems with infinite (or random) time horizon are today still under study. The question which is quite delicate is the behaviour of the solution processes at infinity. There are several papers answering this question under different assumptions both on the coefficients of the FBSDE and on the terminal condition. All those approaches naturally assume that the terminal condition of the Backward equation is given in advance with suitable properties which consequently determine properties of the solution processes. The problem when considering stochastic control problems in infinite time is that one does not know this terminal condition. Therefore, we have to introduce some kind of transversality condition specifying the behavior of the processes at infinity. On the other hand, if we would be able to state the terminal condition (which is an open question), we could use the existing theory as mentioned above. For illustration, we mention in short those known result on FBSDE in view of the transversality condition.

In all the papers, the fully coupled system of equations is considered. Let us have

$$\begin{aligned} dX_t &= b(t, X_t, Y_t, Z_t, \omega)dt + \sigma(t, X_t, Y_t, Z_t, \omega)dW_t, \quad \forall t \geq 0 \text{ a.s.} \\ X_0 &= x \in \mathbb{R}^n, \\ Y_t &= \int_t^{+\infty} h(s, X_s, Y_s, Z_s, \omega)ds - \int_t^{+\infty} Z_s dW_s, \quad \forall t \geq 0, \text{ a.s.} \end{aligned} \tag{9}$$

(10)

In Peng and Shi [11], very strong conditions on b, σ and h are imposed. Namely, it is assumed that they are monotone and globally Lipschitz in all the variables. In that case, the solution process is vanishing in infinity and therefore, the terminal condition is zero a.s. In the paper by Wu [18], a different monotonicity condition is assumed to obtain a solution process with non zero (in general) yet still a.s. constant terminal condition. The most general result is due to Yin [6] who weakens the assumptions to obtain the solution in $\mathbf{L}^2$ spaces with some exponential weight.

We note that assumptions laid on the terminal condition ξ in all those papers are implied by $\xi \in \mathbf{L}^2(\Omega, \mathbb{P})$.

4 Infinite horizon sufficient maximum principle

In this section, the main result is given.

Theorem 1 (Sufficient stochastic maximum principle)**.** *Let $\hat{u} \in \mathcal{U}_{ad}$ and $\hat{X}$ be the associated controlled diffusion process. Let us suppose that there exists a solution $(\hat{Y}, \hat{Z})$ to the associated BSDE (9) such that*

- $\mathcal{H}(\hat{X}_t, \hat{u}_t, \hat{Y}_t, \hat{Z}_t) = \max\limits_{u \in U} \mathcal{H}(\hat{X}_t, u, \hat{Y}_t, \hat{Z}_t), \quad \mathbb{P} \otimes dt - \text{a.e.},$
- $(x, u) \to \mathcal{H}(x, u, \hat{Y}_t, \hat{Z}_t)$ *is a concave function for all t,*
- *the transversality condition*

$$\overline{\lim_{t \to +\infty}} \mathbf{E}\Big[e^{-\beta t}\big\langle \hat{X}_t - X_t, \hat{Y}_t \big\rangle\Big] \leq 0, \tag{11}$$

holds for every $X = X^u, u \in \mathcal{U}_{ad}$.

Then $\hat{u} = u^*$, i.e. $\hat{u}$ is the optimal control strategy to the stochastic control problem (1) - (7).

Proof. Let us take an arbitrary $u \in \mathcal{U}_{ad}$ and examine the difference $J(\hat{u}) - J(u)$. The goal is to show that this quantity is nonnegative. Using the definition of $J(u)$ and $\mathcal{H}$ we have

$$\begin{aligned} J(\hat{u}) - J(u) &= \mathbf{E} \int_0^{+\infty} e^{-\beta t}\big(f(\hat{X}_t, \hat{u}_t) - f(X_t, u_t)\big)dt = \\ &\mathbf{E} \int_0^{+\infty} e^{-\beta t} \Big[\big(\mathcal{H}(\hat{X}_t, \hat{u}_t, \hat{Y}_t, \hat{Z}_t) - \mathcal{H}(X_t, u_t, \hat{Y}_t, \hat{Z}_t)\big) \\ &+\big\langle b(X_t, u_t) - b(\hat{X}_t, \hat{u}_t), \hat{Y}_t \big\rangle + Tr\Big\{\big(\sigma'(X_t, u_t) - \sigma'(\hat{X}_t, \hat{u}_t)\big)\hat{Z}_t\Big\} \\ &+\beta\big\langle \hat{X}_t - X_t, \hat{Y}_t \big\rangle\Big] dt. \end{aligned} \tag{12}$$

The Lebesgue integral over $\mathbb{R}_+$ can be expressed (from Fubini's theorem and existence of the original integral) as the following limit

$$\mathbf{E} \int_0^{+\infty} \mathcal{I}_t dt = \lim_{T \to +\infty} \mathbf{E} \int_0^T \mathcal{I}_t dt, \tag{13}$$

where $\mathcal{I}_t$ is the integrand of (12).

Now we take into account the transversality condition (11) with the sequence $T_n \nearrow +\infty$ realizing the lim sup. Applying the Itô formula we arrive at

$$0 = \lim_{T_n\to+\infty} \mathbf{E}\Big[e^{-\beta T_n}\big\langle \hat{X}_{T_n} - X_{T_n}, \hat{Y}_{T_n}\big\rangle\Big] = \lim_{T_n\to+\infty} \mathbf{E}\int_0^{T_n} \Big[\big\langle \hat{X}_t - X_t, d\big(e^{-\beta t}\hat{Y}_t\big)\big\rangle$$
$$+e^{-\beta t}\big\langle \hat{Y}_t, d\big(\hat{X}_t - X_t\big)\big\rangle + e^{-\beta t} Tr\Big\{\big(\sigma'(\hat{X}_t,\hat{u}_t) - \sigma'(X_t,u_t)\big)\hat{Z}_t\Big\}\Big]\,dt =$$
$$= \lim_{T_n\to+\infty} \mathbf{E}\int_0^{T_n} e^{-\beta t}\Big[\big\langle \hat{X}_t - X_t, -\nabla_x \mathcal{H}(\hat{X}_t,\hat{u}_t,\hat{Y}_t,\hat{Z}_t)\big\rangle - \beta\big\langle \hat{X}_t - X_t, \hat{Y}_t\big\rangle$$
$$-Tr\Big\{\big(\sigma'(X_t,u_t) - \sigma'(\hat{X}_t,\hat{u}_t)\big)\hat{Z}_t\Big\} - \big\langle b(X_t,u_t) - b(\hat{X}_t,\hat{u}_t), \hat{Y}_t\big\rangle\Big]\,dt. \tag{14}$$

Then, putting (14) into (13) we get

$$J(\hat{u}) - J(u) = \lim_{T_n\to+\infty} \mathbf{E}\int_0^{T_n} e^{-\beta t}\Big[\mathcal{H}(\hat{X}_t,\hat{u}_t,\hat{Y}_t,\hat{Z}_t) - \mathcal{H}(X_t,u_t,\hat{Y}_t,\hat{Z}_t)$$
$$-\big\langle \hat{X}_t - X_t, \nabla_x \mathcal{H}(\hat{X}_t,\hat{u}_t,\hat{Y}_t,\hat{Z}_t)\big\rangle\Big]\,dt. \tag{15}$$

>From the concavity of $\mathcal{H}$ in (x,u), we know that

$$\mathcal{H}(\hat{X}_t,\hat{u}_t,\hat{Y}_t,\hat{Z}_t) - \mathcal{H}(X_t,u_t,\hat{Y}_t,\hat{Z}_t) - \big\langle \hat{X}_t - X_t, \nabla_x \mathcal{H}(\hat{X}_t,\hat{u}_t,\hat{Y}_t,\hat{Z}_t)\big\rangle \geq 0. \tag{16}$$

Therefore, we deduce that

$$J(\hat{u}) - J(u) \geq 0, \quad \forall u \in \mathcal{U}_{ad},$$

which proves that $\hat{u}$ is indeed the optimal control. □

5 Example - Optimal consumption rate

The problem of optimal consumption rate is taken from [16] where it is solved using a bit different definition of Hamiltonian. For our setting, we can, in fact, exactly follow the solution process step by step showing the same result.

Let us consider an agent whose wealth evolves according to the following controlled bilinear SDE in $\mathbb{R}$

$$\begin{aligned} dX_t &= X_t\big(\mu_t - u_t\big)dt + X_t\sigma_t dW_t, \quad \forall t \geq 0 \text{ a.s.},\\ X_0 &= x_0 > 0, \end{aligned} \tag{17}$$

where u_t is the consumption rate process, μ_t and σ_t are some deterministic functions. The aim is to maximize over all $u(\cdot) > 0$ the discounted cost functional

$$J(u) = \mathbf{E}\int_0^{+\infty} e^{-\beta t}\ln\Big(u_t X_t\Big)dt, \tag{18}$$

where $\beta > 0$ is a discount factor.

The Hamiltonian of this control problem is

$$\mathcal{H}(x, u, y, z) = x(\mu_t - u)y + x\sigma_t z + \ln(xu) - \beta xy \tag{19}$$

which is a concave function in (x, u). The driver of the backward adjoint equation is obtained as the derivative of $\mathcal{H}$ w.r.t x, i.e.

$$h(x, u, y, z) = (\mu_t - u - \beta)y + \sigma_t z + \frac{1}{x}. \tag{20}$$

To find the maximal point of $\mathcal{H}$ we lay

$$\frac{\partial}{\partial u}\mathcal{H}(x, u, y, z) = -xy + \frac{1}{u} = 0, \tag{21}$$

which leads to

$$\hat{u}_t = \frac{1}{\hat{X}_t \hat{Y}_t}, \tag{22}$$

and the associated BSDE is

$$Y_t = \int_t^{+\infty} \big((\mu_s - u_s - \beta)Y_s + \sigma_s Z_s + \frac{1}{X_s}\big)ds - \int_t^{+\infty} Z_s dW_s, \quad \forall t \geq 0 \text{ a.s.} \tag{23}$$

Further, we will try to find the solution to BSDE (23) for general control $u(\cdot)$. We denote Φ_t the fundamental solution of the following equation

$$\begin{aligned} &\dot{\Phi}_t = -(\mu_t - u_t - \beta)\Phi_t, \quad \forall t \geq 0 \text{ a.s.} \\ &\Phi_0 = 1. \end{aligned} \tag{24}$$

It is easy to show that

$$\Phi_t = e^{-\int_0^t (\mu_s - u_s - \beta)ds} =: e^{-S_t}. \tag{25}$$

Using Φ_t, (23) can be rewritten as

$$Y_t = e^{-S_t}Y_0 - e^{-S_t}\int_0^t e^{S_r}\big(\sigma_r Z_r + \frac{1}{X_r}\big)dr + e^{-S_t}\int_0^t e^{S_r} Z_r dW_r, \quad \forall t \geq 0 \text{ a.s.} \tag{26}$$

As in [16] we lay

$$Y_0 = \frac{1}{x_0\beta}, \quad Z_t = -\frac{\sigma_t}{X_t\beta}, \tag{27}$$

and we show that the solution to (23) is

$$Y_t = \frac{1}{X_t\beta}.$$

Indeed, laying $F_t = e^{-S_t} X_t$ one can easily verify that

$$dF_t = \beta F_t dt + F_t \sigma_t dW_t,$$

and therefore

$$d\left(\frac{1}{F_t \beta}\right) = -\frac{1}{F_t} dt + \frac{\sigma_t^2}{\beta F_t} dt - \frac{\sigma_t}{\beta F_t} dW_t. \tag{28}$$

Integrating (28) from 0 to t we finally observe that

$$\frac{1}{F_t \beta} = \frac{1}{e^{-S_t} X_t \beta} = \frac{1}{x_0 \beta} - \int_0^t e^{S_r} \frac{1}{X_r} dr + \int_0^t e^{S_r} \frac{\sigma_r^2}{\beta X_r} dr - \int_0^t e^{S_r} \frac{\sigma_r}{\beta X_r} dW_r = \frac{Y_t}{e^{-S_t}}, \tag{29}$$

by (26) and (27). Therefore we conclude that

$$Y_t = \frac{1}{X_t \beta} \quad \text{and} \quad \hat{u}_t = \beta. \tag{30}$$

The last thing which remains to be shown is that our solution fulfils the transversality condition (11). We know that

$$\mathbf{E}\left[e^{-\beta t} \hat{Y}_t \left(\hat{X}_t - X_t\right)\right] = \frac{1}{\beta} \mathbf{E}\left[e^{-\beta t}\left(1 - \frac{X_t}{\hat{X}_t}\right)\right] = \frac{1}{\beta} \mathbf{E}\left[e^{-\beta t} - e^{-\int_0^t u_s ds}\right] \leq \frac{1}{\beta} \mathbf{E}\left[e^{-\beta t}\right] \underset{t \to +\infty}{\to} 0, \tag{31}$$

which was to prove.

6 Further work

The next step will be finding an example with nonlinear state dynamics and deriving the necessary maximum principle. Some other generalizations could be introducing jumps into the model, considering relaxed controls or solving the problem in spaces of infinite dimension.

I am grateful to my supervisor, prof. RNDr. Bohdan Maslowski, DrSc. for his helpful comments and his kind guidance.

References

[1] Øksendal B., Sulem A. and Framstad N. C. *A sufficient stochastic maximum principle for optimal control of jump diffusions and applications to finance. J. Optimization Theory and Applications*, **121**, 77-98, 2004. Errata: *J. Optimization Theory and Applications* **124**, 511-512, 2005.

[2] Øksendal B. *Stochastic Differential Equations. An Introduction with Applications, 4th edition. Berlin, Springer-Verlag 1995.*, ISBN 3-540-60243-7 (Universitext)

[3] Pardoux E. *BSDEs weak convergence and homogenizations of semilinear PDEs. Nonlinear Analysis Differential Equations and Control. Clark, F.H., Stern, R.J. (Eds.), Kluwer Academic, Dordrecht*, 503-549, 1999.

[4] Pardoux E., Peng S. G. *Adapted solution of a backward stochastic differential equation. Systems & Control Letters.* **14**, 55-61, 1990.

[5] R. W. R. Darling, Pardoux E. *Backwards SDE with Random Terminal Time and Applications to Semilinear Elliptic PDE. The Annals of Probability.* Vol. 25, No. 3, 1135-1159, 1997.

[6] Yin J. *On solutions of a class of infinite horizon FBSDE's. Statistics and Probability Letters.* **78**, 2412-2419, 2008.

[7] Kushner H. J. *Necessary conditions for continuous parameter stochastic optimization problems. SIAM J. Control.* **10**, 550-565, 1972.

[8] Bismut J.-M. *Conjugate convex functions in optimal stochastic control. J. Math. Anal. Appl.* **44**, 384?04, 1973.

[9] Bensoussan A. *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions. Stochastics.* **9**, 169-222, 1983.

[10] Peng S. G. *A general stochastic maximum principle for optimal control problems. SIAM J. Control Optim.* **28**, 966-979, 1990.

[11] Peng S. G., Shi Y. *Infinite horizon forward–backward stochastic differential equations. Stochastic Process. Appl.* **85**, 75–92, 2000.

[12] Cadenillas A. and Haussmann U. G. *The stochastic maximum principle for a singular control problem. Stochastics Rep.* **49**, 211-237, 1994.

[13] Tang S. J. and Li X. J. *Necessary conditions for optimal control of stochastic systems with random jumps. SIAM J. Control Optim.* **32**, 1447-1475, 1994.

[14] Pham, H. *Continuous-time stochastic control and optimization with financial applications. Stochastic Modelling and Applied Probability* **61**, Springer-Verlag, Berlin, 2009.

[15] Borkar V. S. *Controlled diffusion processes. Probability Surveys.* **2**, 213-244, 2005.

[16] Haadem S., Proske F. and Øksendal B. *A maximum principle for jump diffusion processes with infinite horizon. To appear.*

[17] Veverka P. *Sufficient stochastic maximum principle for discounted control problem. Stochastic & Physical Monitoring Systems (27.6. – 1.7.2011, Křižánky), editor T. Hobza*

[18] Wu Z. *Forward-Backward Stochastic Differential Equations with Stopping Time. ACTA MATHEMATICA SCIENTIA.* **24**, 91-99, 2004.

Towards Routing in Robotics: Using Constraint Programming in Anytime Path Planner

Michal Zerola

4th year of PGS, email: michal.zerola@ujf.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Šumbera, Nuclear Physics Institute, ASCR
Jérôme Lauret, Brookhaven National Laboratory, USA
Roman Barták, Faculty of Mathematics and Physics, Charles University

Abstract. Path planning is one of the critical tasks for autonomous robots. We will study the problem of finding the shortest path for a robot collecting waste spread over the area such that the robot has a limited capacity and hence during the route it must periodically visit depots/collectors to empty the collected waste. This is a variant of often overlooked vehicle routing problem with satellite facilities. We present the approach based on Constraint Programming techniques driven by the concept of finite state automaton. The experimental comparison and enhancements of models are discussed.

Keywords: vehicle routing, autonomous robots, constraint programming, optimization

Abstrakt. Plánovanie cesty je jednou z kritických úloh v atonómnej robotike. Zameriame sa na problém nájdenia najkratšej trajektórie robota zbierajúceho odpadky v prostredí, kde je kapacita robota obmedzená a preto musí robot pravidelne navštevovať tzv. depoty, kde je jeho zásobník vyprázdnený. Jedná sa o variantu často prehliadaného problému - VRP so satelitmi. Predstavíme prístup založený na programovaní s obmedzujúcimi podmienkami inšpirovaný konečnými stavovými automatmi. Prediskutujeme tiež experimenálne porovnania a rozšírenia modelov.

Kľúčové slová: logistika, autonómni roboti, programovanie s obmedzujúcimi podmienkami, optimizácia

1 Introduction

Recent advances in robotics have allowed robots to operate in cluttered and complex spaces. However, to efficiently handle the full complexity of the real-world tasks, new deliberative planning strategies are required. We solve the problem of planning a route for a single robot such that all waste is collected, robot's capacity is never exceeded, and the route is as short as possible. We assume the environment to be known and not changing, in particular, the location of waste and depots is known and the robot knows how to move between these locations. To handle changes in the environment we focus on anytime planning algorithms that can be re-run when the initial task changes. We propose to use Constraint Programming (CP) to solve the problem because of the flexibility of CP. The task we are dealing with is to develop a robot solving a specific routing problem - an often overlooked variant of the standard Vehicle Routing Problem

(VRP). In our setting, the robot has to clean out a collection of waste spread in a building, but under the condition of not exceeding its internal storage capacity at any time. The storage tank can be emptied in one of available collectors. The goal is to come up with the routing plan minimizing the travelled trajectory. This is a similar setting to a Vehicle Routing Problem with Satellite Facilities (VRPSF), where the task is to deliver goods rather than to collect waste. However, this variant of VRP problem, has not been solved extensively. The latest published results can be found in [1], and are mentioned also in [3]. Authors presented an exact procedure for solving the VRPSF that combines heuristics with methods from polyhedral theory in a branch and cut framework. They focused on obtaining an optimal solution, that took almost an hour on 15-customer case instances.

Our primary goal is to develop an algorithm that returns good solutions in a short time (almost anytime algorithm) and that can be easily extended by additional constraints. Neither of existing CP-oriented works solves the above problem, but we can use them as the initial motivation for the design of our constraint model. Most of the routing models are based on the formulation of the problem using network flows (Simonis [7]) so we also proposed a constraint model based on this standard technique. Very often, authors use CP methods embedded into Local Search (LS) in order to achieve better running time while preserving good quality of the solution. We will present also the LS extension and comparison to the pure CP model. Nevertheless, the performance of this model was not satisfactory in our experiments so we proposed a radically new approach to model the problem using a finite state automaton. In our experiments, this model outperforms the traditional model and can solve larger instances of the problem.

2 Problem formulation

The robot's environment consists of the navigation points defined by the locations of waste and collectors. We use a mixed weighted graph (V, E) with both directed and undirected edges to represent this environment. The reason for using undirected edges is minimizing the size of the representation. The set of vertices $V = \{I\} \cup W \cup C \cup \{D\}$ consists of the initial position I, the set W of waste vertices, the set C of collectors and the destination vertex D. From the initial position the robot has to visit some waste so we have directed arcs from I to all vertices in W. The robot can travel between the waste vertices so we assume a complete undirected graph between vertices in W. From any waste vertex the robot can go to a collector so we use a directed edge there and from any collector we can go to any waste which is again modelled using a directed edge. We need directed edges here as we need to count the number of incoming and ongoing edges for the collectors. There are no edges between the collector vertices. As mentioned, we use a dummy destination vertex that is connected to all collector vertices by a directed edge. The weight of each edge describes the distance between the navigation points. The edges going to the dummy destination vertex D have zero weight so the robot can actually finish at any collector. The task is to find a minimal-cost path starting at I, finishing at D and visiting each vertex in W exactly once such that the number of any consecutive vertices from W does not exceed the given capacity of the robot.

3 CP model based on finite state automata

The model that we propose brings a radically new approach not seen so far when modelling VRPs or TSPs. We can base the model on the existing regular constraint (Pesant [4]). This constraint allows a more global view of the problem so the hope is that it can infer more information than the network-based model and hence decreases the search space to be explored. First, it is important to realize that the exact path length is unknown in advance. Each waste vertex is visited exactly once, but the collector vertices can be visited more times and it is not clear in advance how many times. Nevertheless, it is possible to compute the upper bound on the path's length. Let us assume that the path length is measured as the number of visited vertices, the robot starts at the initial position and finishes at some collector vertex (we will use the dummy destination in a slightly different meaning here), and the weight/cost of arcs is non-negative. Let $K = |W|$ be the number of waste vertices and $cap \geq 1$ be the robot's capacity. Then the maximal path length is $2K + 1$. This corresponds to visiting a collector vertex immediately after visiting a waste vertex. Recall that each waste vertex must be visited exactly once and there is no arc between the collector vertices.

Our model is based on four types of constraints. First, there is a restriction on the existence of a connection between two vertices - a *routing constraint*. This constraint describes the routing network. It roughly corresponds to the *Kirchoff's* constraints from the network-based model. Note that the sub-tour elimination constraints are not necessary here. Second, there is a restriction on the robot's capacity stating that there in no continuous subsequence of waste vertices whose length exceeds the given capacity - a *capacity constraint*. Third, each waste must be visited exactly once, while the collectors can be visited more times (even zero times) - an *occurrence constraint*. Finally, each arc is annotated by a weight and there is a constraint that the sum of the weights of used arcs does not exceed some limit - a *cost constraint*. This constraint is used to define the total cost of the solution.

In the constraint model we use three types of variables. Let $N = 2K + 1$ be the maximal path length. Then we have N variables $Node_i$, N variables Cap_i, and N variables $Cost_i (i = 1, \ldots, N)$ so we assume the path of maximal length. Clearly, the real path may be shorter so we introduce a dummy destination vertex that fills the rest of the path till the length N. In other words, when we reach the dummy vertex, it is not possible to leave it. This way, we can always look for the path of length N and the model gives flexibility to explore the shorter paths too.

The semantic of the variables is as follows. The variables $Node_i$ describe the path hence their domain is the set of numerical identifications of the vertices. We use positive integers $1, \ldots, K (K = |W|)$ to identify the waste vertices, $K + 1, \ldots, K + L$ for the collector vertices ($L = |C|$), and 0 for the dummy destination vertex. In summary, the initial domain of each variable $Node_i$ consists of values $0, \ldots, K + L$. Cap_i is the used capacity of the robot after leaving vertex $Node_i (Cap_1 = 0$ as the robot starts empty), the initial domain is $\{0, \ldots, cap\}$. $Cost_i$ is the cost of the arc used to leave the vertex $Node_i (Cost_N = 0)$, the initial domain consists of non-negative numbers. Formally:

$$\forall i = 1, \ldots, N(N = 2K + 1): \begin{array}{rcl} 0 & \leq & \mathrm{N}ode_i \leq K + L \\ 0 & \leq & \mathrm{C}ap_i \leq cap, \mathrm{C}ap_1 = 0 \\ 0 & \leq & \mathrm{C}ost_i, \mathrm{C}ost_N = 0 \end{array} \tag{1}$$

We will start the description of the constraints with the *occurrence constraint* saying that each waste vertex is visited exactly once. This can be modelled using the global cardinality constraint (Régin [5]) over the set $\{\mathrm{N}ode_1, \ldots, \mathrm{N}ode_N\}$. The constraint is set such that the each value from the set $\{1, \ldots, K\}$ is assigned to exactly one variable from $\{\mathrm{N}ode_1, \ldots, \mathrm{N}ode_N\}$ - each waste node is visited exactly once. The values $\{0, K + 1, \ldots, K + L\}$ can be used any number of times. Formally:

$$\begin{array}{rc} & \mathrm{g}cc(\{\mathrm{N}ode_1, \ldots, \mathrm{N}ode_N\}, \\ \{v: & [1,1]\,\forall v = 1, \ldots, K, \\ 0: & [0, \infty], \\ v: & [0, \infty]\,\forall v = K + 1, \ldots, K + L\}) \end{array} \tag{2}$$

where $v : [\min, \max]$ means that value v is assigned to at least min and at most max variables from $\{\mathrm{N}ode_1, \ldots, \mathrm{N}ode_N\}$. The *gcc* constraint allows specifying the number of appearances of the value using another variable rather than using a fixed interval as in 2. Let D be the variable describing the number of appearances of value 0 (identification of the dummy vertex) in the set $\{\mathrm{N}ode_1, \ldots, \mathrm{N}ode_N\}$, then we can use the following constraints instead of 2:

$$\begin{array}{rc} & \mathrm{g}cc(\{\mathrm{N}ode_1, \ldots, \mathrm{N}ode_N\}, \\ \{v: & [1,1]\,\forall v = 1, \ldots, K, \\ 0: & D, \\ v: & [0, \infty]\,\forall v = K + 1, \ldots, K + L\}) \end{array} \tag{3}$$

$$\mathrm{N}ode_{N-D} > 0 \tag{4}$$

The constraint 4 says that $\mathrm{N}ode_{N-D}$ is not a dummy vertex; actually it is the last real vertex in the path. We can also set the upper bound for D by using the information about the minimal path length (*MinPathLength* is a constant computed in advance):

$$D \leq N - \mathrm{M}inPathLength \tag{5}$$

These additional constraints 4 and 5 are not necessary for the problem specification but they improve inference (we use them in experiments).

The *cost constraint* can be easily described as

$$\mathrm{O}bj = \sum_{1,\ldots,N} \mathrm{C}ost_i \tag{6}$$

so we can use the constraints $\mathrm{O}bj < \mathrm{B}ound$ in the branch-and-bound procedure exactly the same way as in the network-based model.

For the cost constraint to work properly we need to set the value of $\mathrm{C}ost_i$ variables. Recall that $\mathrm{C}ost_i$ is the cost/weight of the arc going from vertex $\mathrm{N}ode_i$ to vertex $\mathrm{N}ode_{i+1}$. Hence, we can connect the *Cost* variables with the *Node* variables when specifying the *routing constraint*. In particular, we use the ternary constraints over the variables

$\mathrm{N}ode_i, \mathrm{C}ost_i, \mathrm{N}ode_{i+1}\ i = 1, \ldots, N-1$. This set of constraints corresponds to the idea of slide constraint (Bessiere et al. [2]). We implement the constraint between the variables $\mathrm{N}ode_i, \mathrm{C}ost_i, \mathrm{N}ode_{i+1}$ as a ternary tabular (extensionally defined) constraint; let us call it link, where the triple (p, q, r) satisfies the constraint if there is an arc from the vertex p to the vertex r with the cost q. In other words, this table describes the original routing network with the costs extended by the dummy vertex. Formally:

$$\begin{aligned} link(p,q,r) \equiv \quad & \exists e \in E : e = (p,r), q = weight(e) \\ & \vee (q = r = 0 \wedge (p = 0 \vee p > K) \end{aligned} \tag{7}$$

$$\forall i = 1, \ldots, 2K : link(\mathrm{N}ode_i, \mathrm{C}ost_i, \mathrm{N}ode_{i+1}) \tag{8}$$

It remains to show how the *capacity constraint* is realized. Briefly speaking, we use a similar approach as for the routing constraint. The capacity constraint is realized using a set of ternary constraints over the variables $\mathrm{C}ap_i, \mathrm{N}ode_{i+1}, \mathrm{C}ap_{i+1}\ i = 1, \ldots, N-1$, again exploiting the idea of slide constraint. The constraint is implemented using a tabular constraint, let us call it *capa*, with the following semantics. Triple (p, q, r) satisfies this constraint if and only if:

- q is an identification of a collector vertex $(q > K)$ or a dummy vertex $(q = 0)$ and $r = 0$
- q is an identification of a waste node $(0 < q \leq K)$ and $r = p + 1$.

Recall that the domain of capacity variables is $\{0, \ldots, cap\}$ so we never exceed the capacity of the robot. Formally:

$$\begin{aligned} capa(p,q,r) \equiv \quad & (q = r = 0) \\ & \vee (q > K \wedge r = 0) \\ & \vee (0 < q \leq K \wedge r = p + 1) \end{aligned} \tag{9}$$

$$\forall i = 1, \ldots, 2K : capa(\mathrm{C}ap_i, \mathrm{N}ode_{i+1}, \mathrm{C}ap_{i+1}) \tag{10}$$

Any solution to the above described constraint satisfaction problem defines a valid solution of our single robot path planning problem with the capacity constraint. Vice versa, any solution to the path planning problem is also a feasible solution of the specified constraint satisfaction problem. We omit the formal proof due to limited space.

3.1 Search procedure

It is important to specify the search strategy. Only the variables $\mathrm{N}ode_i$ are the decision variables - they define the search space. It is easy to realize that the inference through the routing constraints 8 decides the values of the $\mathrm{C}ost_i$ variables and the inference through the capacity constraints 10 decides the values of the $\mathrm{C}ap_i$ variables provided that the values of all variables $\mathrm{N}ode_i$ are known.

When searching for the solution we first use a greedy approach to find the initial solution (the initial cost). This greedy algorithm instantiates the variables $\mathrm{N}ode_i$ in the order of increasing i in such a way that the arc with the smallest cost is preferred. We

select the node to which the least expensive arc from the previously decided node leads. Naturally, the capacity constraint is taken into account so only the nodes such that the capacity is not exceeded are assumed.

To find the optimal solution we use a standard branch-and-bound approach with restarts. To instantiate the N*ode* variables we use the *min-dom* heuristic for the variable selection, that is, the variable with the smallest current domain is instantiated first. We select the values in the order defined in the problem (the waste nodes are tried before the collector nodes). After finding a solution with the total cost *Bound*, the constraint O*bj* < B*ound* is posted and search continues until any solution is found. The last found solution is the optimum. Note that using the well known and widely applied min-dom heuristic for the variable selection is meaningful in this model because we have larger domains.

4 Embedding CP models into local search

The current state of the art techniques for solving VRPs are frequently based on hybrid approaches. For example the paper (Rousseau et al. [6]) suggests using CP techniques to explore the neighborhood within Large Neighborhood Search (LNS). We decided to apply a similar approach with our CP models to check, if the solution quality can be improved in comparison with the pure branch-and-bound approaches presented above.

The basic elements in the neighborhood local search are the concept of the neighborhood of a solution and the mechanism for generating neighborhoods. It is eminent that the performance and "success" of the local search algorithm strongly depends on the neighborhood operator and its state space. In our case, the state corresponds to the plan - a valid path for the robot. The local search algorithm is repeatedly choosing another solution in the neighborhood of the current solution with the goal to improve the value of the objective function. This move is realized by a so called *neighborhood operator*.

We have implemented an operator that is successfully used for solving the Travelling Salesman Problems. The operator relaxes the solution by removing an induced path of a given length and then it calls the CP solver to complete the solution. It means that we add to a given constraint model additional constraints that fix some edges (for the model based on network flows) or forbid using some edges (for the model based on finite state automata). These fixed edges correspond to the edges in the original solution that were not removed by the neighborhood operator. The role of the CP solver is to optimally complete this partial solution by adding the missing edges. The new solution is the state to which the local search procedure moves. As the local search repeatedly chooses a move that improves the value of the objective function (we are minimizing the value), it can get "trapped" in the first local minimum it encounters. In order to escape the local minimum, a controlled method of accepting an ascending move is required. In this paper, we examined the simplified simulated annealing. Note finally, that as the initial solution for local search we used the first solution obtained from the pure CP model (see the description of the search procedures above).

5 Experimental results

In this section we will present the experimental evaluation of the presented solving techniques. As there is no standard benchmark set for the studied problem, we generated own problem instances. We used a square-sized robot arena where the positions of the waste and the initial location of the robot were uniformly distributed. The collectors were uniformly distributed along the boundaries of the arena and the weights set up as a point-to-point distance using the Euclidean metric. All the following measurements were performed on `Intel Xeon CPU@2.5GHz` with 4GB of RAM, running a Debian GNU Linux operating system.

5.1 Performance of the network flow model

As stated earlier, the model based on network flows corresponds to the traditional operations research approach, but we modified the model to describe specifics of our robot routing problem. The model was implemented in *Java SE 6* [1] using **Choco** [2], an open-source constraint programming library. The optimization search strategy uses the built-in branch-and-bound method, while all constraints correspond to the mathematic formulations described earlier.

Figure 1 (left) shows the runtime (a logarithmic scale) to obtain the optimal solution as a function of the instance size measured by the number of waste and by the number of collectors. We generated 15 instances for each problem size and the graph shows the average time the solver needs for finding and proving the optimality of the solution. The capacity of robot was 3.

As already mentioned in (Bard et al. [1]), the satellite facilities in VRP (or collectors in robotics case) heavily increase the complexity of the problem. The initial experiment shows that the runtime increases exponentially with the number of waste but the runtime is not significantly affected by the increased number of collectors. In fact it seems that for different quantities of waste there are different numbers of collectors where the best runtime is achieved. This is an interesting observation claiming that for a given number of waste there is some number of collectors that gives the best result. Nevertheless, this observation requires additional experiments to confirm it.

While the graph in Fig. 1 (left) represents the total time the solver needs for finding and proving the optimality of the solution, we are also interested in how fast a "good enough" solution can be found. This characteristic can be seen in Fig. 1 (right), where the graph displays the convergence of the solution during search. We can see that even a simple greedy heuristic performs very well and the difference from the optimal solution was less than 5% within first seconds for the instance $7 + 3$.

5.2 Performance of the network flow model within local search

As mentioned above, the CP model can be used within the Large Neighborhood Search procedure to solve larger instances but obviously without any guarantee of optimality.

[1] Java: `http://www.java.com`http://www.java.com

[2] Choco: `http://choco.sourceforge.net`http://choco.sourceforge.net

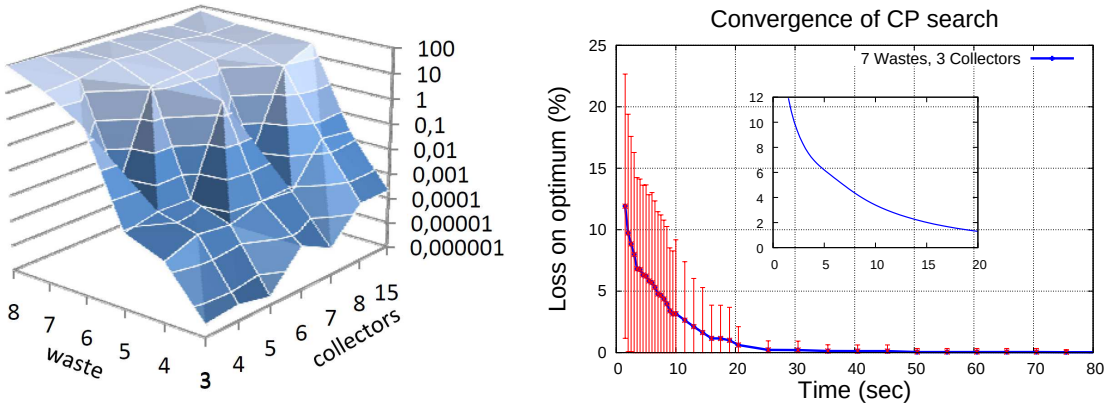


Figure 1: Network flow model. **Left.** Runtime (seconds). **Right.** Quality convergence.

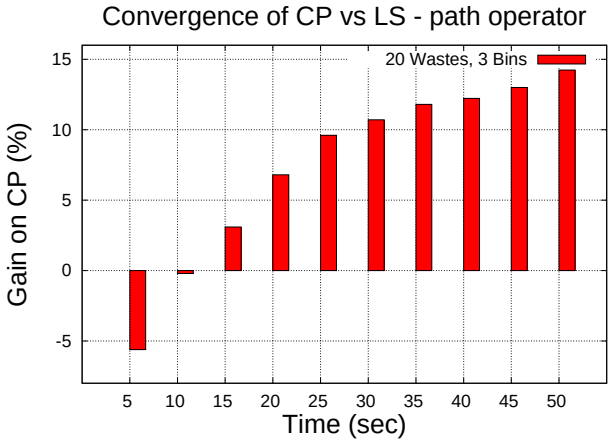


Figure 2: Comparison of the quality convergence of the network flow model in the pure CP approach and the CP model embedded into local search.

We generated 50 independent problem instances with 20 wastes and 3 collectors (referred to as $20 + 3$). The capacity of the robot was set to 7 units. The neighborhood operator was allowed to remove 5 randomly selected consecutive edges during the search and the embedded CP solver was allowed to search for 1 second. The graph in Fig. 2 shows an average one-to-one performance of the pure CP method and the LS method (with the embedded CP model) applied to the produced instances. The graph shows the difference in the quality of a solution found in the corresponding time from the LS viewpoint.

The local search procedure performed better in the long run, when compared to the pure CP method relaying only on its inner heuristic. However, CP beat LS in the first seconds where the convergence drop was steeper. As a consequence, CP seems to be a more appropriate method under very short time constraints, while reasonably good solutions can be found with a combination of LS for larger instances.

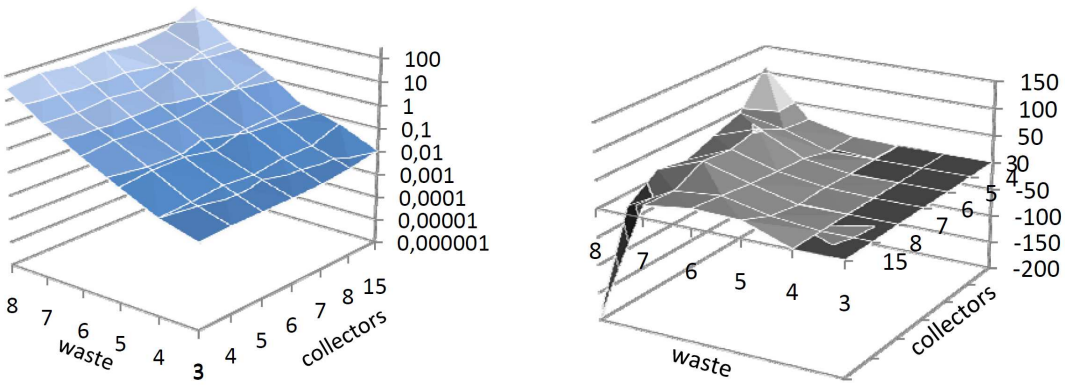


Figure 3: **Left.** Runtime (seconds) for the model based on finite state automata. **Right:** Time difference (seconds) between the CP models. Positive values means that the model based on finite state automata is faster.

5.3 Performance of the finite state automaton model

The network flow model represents a standard approach to solving the Vehicle Routing Problems so we compared our novel constraint model based on the finite state automaton directly to this approach. The second model was implemented in **SICStus Prolog** [3]. Figure 3 (left) shows the runtime (a logarithmic scale) to obtain the optimal solution using the constraint model based on finite state automata using the same problems as for the model based on network flows (Fig. 1 (left)). The result also shows the exponential grow with the increased number of waste and weaker dependence on the number of collectors.

To directly compare both models, we generated a difference graph showing the difference of runtimes for the network model and for the automata model - the values above zero mean faster automata model, while the times below zero mean faster network model. Figure 3 (right) shows these difference times. The automata-based model is visibly better for a smaller number of collectors where the problem is more constrained and the capacity constraints can prune more of the search space. A bit surprisingly, it seems that the network-based model is better when the number of collectors becomes larger. This feature will require a further investigation.

Since in robotic, finding a good plan fast is more important than having the optimal one late, we started to investigate again the quality of the plans found by the CP solver in a limited time. In particular, we embedded the new CP model in the LNS procedure as described above and we tried to compare the pure CP model with this LS approach on much bigger instances with 40 wastes and 3 collectors. To our surprise, the LS method was not able to improve the solution found by the CP model in the 2 minutes runtime. As we need to produce a good solution in seconds, the pure CP model based on finite state automaton seems more appropriate for our purpose.

[3]SICStus Prolog: `http://www.sics.se/sicstus`http://www.sics.se/sicstus

6 Conclusions

We developed the robotic architecture incorporating both purely reactive execution and deliberative planning that works in complex and dynamic environment. The goal of the robot is to pick up all wastes in a given environment and put them to collectors while assuming a limited capacity of the robot. We used a constraint model based on network flows that is traditionally applied to this type of routing problems and we developed a completely new model based on finite automata. We further studied local search techniques that are traditionally used to improve the runtime performance of CP models for vehicle routing problems and we have found that our novel model based on finite automata performs better without local search. The experiments showed some interesting behavior of the model in relation to the number of collectors that we are going to further investigate. In summary, there are three novel contributions. First, we reformulated the traditional network flow model to solve the waste collecting problem with limited capacity of the robot. Second, we proposed a novel constraint model based on finite automata (state transitions) and we experimentally showed that it outperforms the traditional approach, if the number of waste collecting places is small. Finally, we integrated the proposed models with a reactive planner to show that deliberative planning based on CP can be used in real robots and environments.

References

[1] J. F. Bard, L. Huang, M. Dror, and P. Jaillet. *A branch and cut algorithm for the VRP with satellite facilitie s.* IIE Transactions **30** (1998), 821–834.

[2] C. Bessiere, E. Hebrard, B. Hnich, Z. Kiziltan, C.-G. Quimper, and T. Walsh. *Reformulating global constraints: the slide and regular constraints.* In 'Proceedings of the 7th International conference on Abstraction, reformulation, and approximation', SARA'07, 80–92, Berlin, Heidelberg, (2007). Springer-Verlag.

[3] J.-F. Cordeau, G. Laporte, M. W. Savelsbergh, and D. Vigo. *Vehicle routing.* In 'Transportation, Handbooks in Operations Research and Management Science', C. Barnhart and G. Laporte, (eds.), volume 14, Elsevier (2007), 367–428.

[4] G. Pesant. *A Regular Language Membership Constraint for Finite Sequences of Variables.* In 'Principles and Practice of Constraint Programming', 482–495, (2004).

[5] J.-C. Régin. *Generalized arc consistency for global cardinality constraint.* In 'Proceedings of the 13th national conference on Artificial intelligence - Volume 1', AAAI'96, 209–215. AAAI Press, (1996).

[6] L.-M. Rousseau, M. Gendreau, and G. Pesant. *Using Constraint-Based Operators to Solve the Vehicle Routing Problem with Time Windows.* Journal of Heuristics **8** (2002), 43–58.

[7] H. Simonis. *Constraint applications in networks.* In 'Handbook of Constraint Programming', F. Rossi, P. van Beek, and T. Walsh, (eds.), Elsevier (2006), chapter 25, 875–903.

Implementation of the Schur Complement Method for the Stokes Problem*

Vítězslav Žabka

2nd year of PGS, email: zabkavit@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. The aim of this article is to investigate the Schur complement method for parallel numerical solution of the two-dimensional Stokes problem. Parallel MPI implementation of the method is developed and tested on the lid driven cavity flow. The implementation relies on the LU decomposition and the unpreconditioned conjugate gradient method. Possible ways of improving the implementation are discussed.

Keywords: Schur complement method, Stokes problem, domain decomposition, FEM

Abstrakt. Cílem tohoto příspěvku je implementovat metodu Schurova doplňku pro paralelní numerické řešení Stokesova problému ve 2D. S použitím MPI byla implementována metoda Schurova doplňku založená na nepředpodmíněné metodě konjugovaných gradientů a LU rozkladu. Funkčnost implementace metody byla ověřena na problému proudění v kavitě. V článku jsou také rozebrány možnosti vylepšení implementace.

Klíčová slova: Metoda Schurova doplňku, Stokesův problém, doménová dekompozice, MKP

1 Introduction

Solving the Stokes problem by means of the mixed finite element method leads to a symmetric saddle point system of linear equations. Because of their indefiniteness, such systems are difficult to solve. Numerous techniques have been proposed in recent years [2]. So far, we have used a multigrid solver developed by P. Bauer [1]. The main drawback of this solver is its sequential nature which limits the size of problems it is applicable to.

In this article, we investigate the Schur complement method. This non-overlapping domain decomposition method is suitable for parallel implementation on distributed memory systems. We implement the method using MPI. Our implementation is based on the conjugate gradient method and the LU decomposition, and it is tested on the lid driven cavity flow problem.

The article is organized as follows. In Section 2, we introduce the Stokes problem with the Dirichlet boundary condition and its solution by the mixed finite element method. In Section 3, we describe the Schur complement method. In Section 4, we present our parallel implementation of the Schur complement method for the Stokes problem. In Section 5,

*The work has been performed under the Project HPC-EUROPA2 (Project number 228398), with the support of the European Community — under the FP7 "Research Infrastructures" Programme.

we compare the implementation with a corresponding sequential GMRES solver and a multigrid solver.

2 Stokes problem

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain filled with a viscous fluid of constant density. Low-speed incompressible flow of the fluid can be modelled by the following system of Stokes equations with the Dirichlet boundary condition:

$$\begin{aligned} -\nu\Delta\vec{u} + \nabla p &= \vec{f} \quad \text{in } \Omega, \\ -\nabla\cdot\vec{u} &= 0 \quad \text{in } \Omega, \\ \vec{u} &= \vec{g} \quad \text{on } \partial\Omega, \end{aligned} \tag{1}$$

where, x being the spatial variable, $\vec{u} = \vec{u}(x)$ is the vector of flow velocity, $p = p(x)$ is the pressure divided by the fluid mass density, $\vec{f} = \vec{f}(x)$ denotes the density of volume forces per mass unit, ν denotes the kinematic viscosity of the fluid, $\partial\Omega$ denotes the boundary of Ω and $\vec{g} = \vec{g}(x)$ is a given function satisfying $\int_{\partial\Omega} \vec{g}\cdot\vec{n}\,\mathrm{d}s = 0$ with $\vec{n}$ denoting the unit normal vector to the boundary $\partial\Omega$.

We now introduce the mixed weak formulation of the Stokes problem (1). Let us consider the usual Sobolev space $H^1(\Omega)$ and denote

$$\begin{aligned} V &= \left\{\vec{v} \in \left(H^1(\Omega)\right)^2 : \vec{v}|_{\partial\Omega} = 0\right\}, \\ V_{\vec{g}} &= \left\{\vec{v} \in \left(H^1(\Omega)\right)^2 : \vec{v}|_{\partial\Omega} = \vec{g}\right\}, \\ Q &= L_2(\Omega), \end{aligned} \tag{2}$$

where the restriction $\vec{v}|_{\partial\Omega}$ is understood in the sense of traces. Then, the mixed weak formulation of the problem (1) reads: Find $\vec{u} \in V_{\vec{g}}$ and $p \in Q$ such that

$$\begin{aligned} \nu\int_\Omega \nabla\vec{u} : \nabla\vec{v}\,\mathrm{d}\Omega - \int_\Omega p\,\nabla\cdot\vec{v}\,\mathrm{d}\Omega &= \int_\Omega \vec{f}\cdot\vec{v}\,\mathrm{d}\Omega \qquad \forall\vec{v} \in V, \\ -\int_\Omega q\,\nabla\cdot\vec{u}\,\mathrm{d}\Omega &= 0 \qquad \forall q \in Q. \end{aligned} \tag{3}$$

To obtain a discrete analogue of (3), the sets V, $V_{\vec{g}}$ and Q are replaced by their finite dimensional subsets $V^h \subset V$, $V^h_{\vec{g}} \subset V_{\vec{g}}$ and $Q^h \subset Q$. Denoting $\vec{u}_{0h} = \vec{u}_h - \vec{g}_h \in V^h$, where $\vec{g}_h \in V^h_{\vec{g}}$ represents the Dirichlet boundary condition $\vec{g}$ in (1), the discrete Stokes problem reads: Find $\vec{u}_{0h} \in V^h$ and $p_h \in Q^h$ such that

$$\begin{aligned} \nu\int_\Omega \nabla\vec{u}_{0h} : \nabla\vec{v}_h\,\mathrm{d}\Omega - \int_\Omega p_h\,\nabla\cdot\vec{v}_h\,\mathrm{d}\Omega &= \int_\Omega \vec{f}\cdot\vec{v}_h\,\mathrm{d}\Omega - \nu\int_\Omega \nabla\vec{g}_h : \nabla\vec{v}_h\,\mathrm{d}\Omega \qquad \forall\vec{v}_h \in V^h, \\ -\int_\Omega q_h\,\nabla\cdot\vec{u}_{0h}\,\mathrm{d}\Omega &= \int_\Omega q_h\,\nabla\cdot\vec{g}_h\,\mathrm{d}\Omega \qquad \forall q_h \in Q^h. \end{aligned} \tag{4}$$

Let $\left(\vec{\phi}_j\right)_{j=1}^{n_V}$ be a basis of V^h and $\left(\psi_j\right)_{j=1}^{n_Q}$ a basis of Q^h. The functions $\vec{u}_{0h}$ and p_h can then be expressed as linear combinations of the corresponding basis functions:

$$\vec{u}_{0h} = \sum_{j=1}^{n_V} u_{0h,j}\vec{\phi}_j, \qquad p_h = \sum_{j=1}^{n_Q} p_{h,j}\psi_j. \tag{5}$$

We introduce matrices

$$\mathbf{A} = (A_{ij})_{i,j=1}^{n_V}, \qquad A_{ij} = \nu \int_\Omega \nabla\vec{\phi}_i : \nabla\vec{\phi}_j \,\mathrm{d}\Omega, \tag{6}$$

$$\mathbf{B} = (B_{ij})_{i,j=1}^{n_Q,n_V}, \qquad B_{ij} = -\int_\Omega \psi_i \,\nabla\cdot\vec{\phi}_j \,\mathrm{d}\Omega, \tag{7}$$

right-hand side vectors

$$\mathbf{b} = (b_i)_{i=1}^{n_V}, \qquad b_i = \int_\Omega \vec{f}\cdot\vec{\phi}_i \,\mathrm{d}\Omega - \nu\int_\Omega \nabla\vec{g}_h : \nabla\vec{\phi}_i \,\mathrm{d}\Omega, \tag{8}$$

$$\mathbf{c} = (c_i)_{i=1}^{n_Q}, \qquad c_i = \int_\Omega \psi_i \,\nabla\cdot\vec{g}_h \,\mathrm{d}\Omega \tag{9}$$

and vectors of unknowns $\mathbf{u}_0 = (u_{0h,i})_{i=1}^{n_V}$, $\mathbf{p} = (p_{h,i})_{i=1}^{n_Q}$. Now, taking $\vec{v}_h = \vec{\phi}_i$ for $i = 1,\ldots,n_V$ and $q_h = \psi_i$ for $i = 1,\ldots,n_Q$, the discrete formulation (4) leads to the following symmetric indefinite system of linear equations:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_0 \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix}. \tag{10}$$

We discretize the problem (3) by the mixed finite element method using triangular $P_1^{NC} - P_0$ Crouzeix-Raviart elements [4], i.e. a combination of non-conforming piecewise linear elements for the components of velocity and piecewise constant elements for pressure. Velocity unknowns are located in the centers of mesh edges, pressure unknowns at the centers of mesh cells. Such a combination satisfies the Babuška-Brezzi condition [3].

3 Schur complement method

The Schur complement method (see, e.g., [6]) belongs to non-overlapping domain decomposition methods. The mesh obtained by triangulation of the domain Ω is split into N non-overlapping submeshes, i.e. submeshes with no common cells (see Figure 1). However, the submeshes can share certain edges and, consequently, velocity unknowns. The union of all unknowns associated with at least two submeshes is called the interface, and it is denoted by Γ. Other unknowns are called interior unknowns.

Viewing the system (10) as

$$\mathbf{C}\mathbf{x} = \mathbf{d}, \tag{11}$$

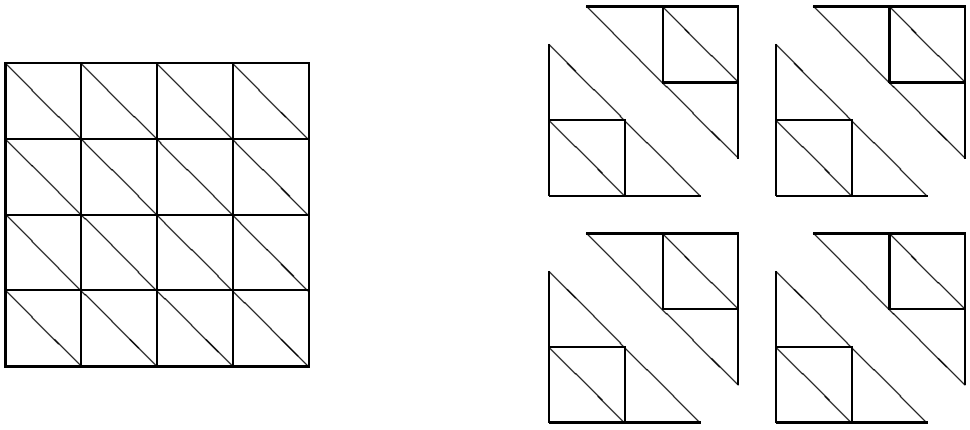


Figure 1: Computational mesh and its decomposition into 8 non-overlapping submeshes.

where

$$\mathbf{C} = \begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & 0 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} \mathbf{u}_0 \\ \mathbf{p} \end{pmatrix}, \qquad \mathbf{d} = \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix}, \tag{12}$$

we can now formally reorder its unknowns into $N+1$ blocks. The i-th block, $i = 1, \ldots, N$, contains the interior unknowns associated with the i-th submesh, whereas the last block corresponds to the interface unknowns. Then, system (11) can be rewritten as follows:

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{1\Gamma} \\ \mathbf{C}_{\Gamma 1} & \mathbf{C}_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_\Gamma \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_\Gamma \end{pmatrix}. \tag{13}$$

Here, $\mathbf{x}_1$ contains only the interior unknowns, $\mathbf{x}_\Gamma$ is composed of the interface unknowns, the matrices $\mathbf{C}_{11}$ and $\mathbf{C}_{\Gamma\Gamma}$ are square matrices, and $\mathbf{C}_{11}$ is invertible and has a block diagonal structure with N blocks. Applying blockwise Gaussian elimination to (13), we get

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{1\Gamma} \\ 0 & \mathbf{C}_{\Gamma\Gamma} - \mathbf{C}_{\Gamma 1}\mathbf{C}_{11}^{-1}\mathbf{C}_{1\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_\Gamma \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_\Gamma - \mathbf{C}_{\Gamma 1}\mathbf{C}_{11}^{-1}\mathbf{d}_1 \end{pmatrix}. \tag{14}$$

The solution of (13) is obtained in two steps. First, the interface unknowns $\mathbf{x}_\Gamma$ are computed as the solution of

$$\mathbf{S}\mathbf{x}_\Gamma = \mathbf{f}, \tag{15}$$

where $\mathbf{S}$ is the Schur complement matrix defined as

$$\mathbf{S} = \mathbf{C}_{\Gamma\Gamma} - \mathbf{C}_{\Gamma 1}\mathbf{C}_{11}^{-1}\mathbf{C}_{1\Gamma} \tag{16}$$

and the condensed right-hand side $\mathbf{f}$ as

$$\mathbf{f} = \mathbf{d}_\Gamma - \mathbf{C}_{\Gamma 1}\mathbf{C}_{11}^{-1}\mathbf{d}_1. \tag{17}$$

The matrix $\mathbf{C}_{11}$ is typically inverted using an LU decomposition, i.e. $\mathbf{C}_{11}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$. Consequently, the Schur complement matrix $\mathbf{S}$ does not have to be formed explicitly; its multiplication by a vector can be accomplished by three sparse matrix multiplications and one forward and back substitution instead. Once the vector of interface unknowns $\mathbf{x}_\Gamma$ is known, the interior unknowns are found using the relation

$$\mathbf{C}_{11}\mathbf{x}_1 = \mathbf{d}_1 - \mathbf{C}_{1\Gamma}\mathbf{x}_\Gamma. \tag{18}$$

There are two main advantages of the Schur complement method. First, Krylov subspace methods usually converge faster for the system (15) than for the original system (11); see [7]. And second, the matrix $\mathbf{C}_{11}$ has a block diagonal structure. Thus, computations involving $\mathbf{C}_{11}$ can be performed in parallel.

4 Implementation

We implemented a parallel algorithm of the Schur complement method using MPI. Our implementation relies on a parallel conjugate gradient solver [6] for the Schur complement system (15) and an LU decomposition of $\mathbf{C}_{11}$. In this section, we describe the implementation including the way of assembling the system (13).

Let us begin with a brief analysis of (13). Since the matrix $\mathbf{C}_{11}$ is block diagonal, we can rewrite (13) in the following form:

$$\begin{pmatrix} \mathbf{C}_{11}^{(1)} & & & \mathbf{C}_{1\Gamma}^{(1)} \\ & \ddots & & \vdots \\ & & \mathbf{C}_{11}^{(N)} & \mathbf{C}_{1\Gamma}^{(N)} \\ \mathbf{C}_{\Gamma 1}^{(1)} & \cdots & \mathbf{C}_{\Gamma 1}^{(N)} & \mathbf{C}_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1^{(1)} \\ \vdots \\ \mathbf{x}_1^{(N)} \\ \mathbf{x}_\Gamma \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1^{(1)} \\ \vdots \\ \mathbf{d}_1^{(N)} \\ \mathbf{d}_\Gamma \end{pmatrix}, \tag{19}$$

where the vectors $\mathbf{x}_1^{(i)}$, $i = 1, \ldots, N$, are composed of the interior unknowns associated with the i-th submesh, the matrices $\mathbf{C}_{11}^{(i)}$ describe interactions between the interior unknowns associated with the i-th submesh, the matrices $\mathbf{C}_{1\Gamma}^{(i)}$ and $\mathbf{C}_{\Gamma 1}^{(i)}$ describe interactions between the interior unknowns associated with the i-th submesh and the interface. Moreover, the matrix $\mathbf{C}_{\Gamma\Gamma}$ can be expressed as a sum of matrices describing interactions between the interface unknowns associated with each submesh:

$$\mathbf{C}_{\Gamma\Gamma} = \sum_{i=1}^{N} \mathbf{C}_{\Gamma\Gamma}^{(i)}. \tag{20}$$

As a result, the system (19) can be easily decomposed.

Considering (20), the Schur complement matrix $\mathbf{S}$ is, in accordance with (16), given by

$$\mathbf{S} = \sum_{i=1}^{N} \left(\mathbf{C}_{\Gamma\Gamma}^{(i)} - \mathbf{C}_{\Gamma 1}^{(i)} \left(\mathbf{C}_{11}^{(i)} \right)^{-1} \mathbf{C}_{1\Gamma}^{(i)} \right). \tag{21}$$

Similarly, the condensed right-hand side $\mathbf{f}$ defined by (17) takes the form

$$\mathbf{f} = \mathbf{d}_\Gamma - \sum_{i=1}^{N} \mathbf{C}_{\Gamma 1}^{(i)} \left(\mathbf{C}_{11}^{(i)} \right)^{-1} \mathbf{d}_1^{(i)}, \tag{22}$$

and the relation (18) splits into

$$\mathbf{C}_{11}^{(i)} \mathbf{x}_1^{(i)} = \mathbf{d}_1^{(i)} - \mathbf{C}_{1\Gamma}^{(i)} \mathbf{x}_\Gamma, \qquad i = 1, \ldots, N. \tag{23}$$

We can now proceed to the description of our parallel implementation. Let us assume $N + 1$ MPI processes labeled $0, \ldots, N$ are available. The 0-th process is referred to as

the root process; the other processes are called non-root processes. In the beginning, the root process creates the computational mesh, decomposes it into N non-overlapping submeshes (see Section 3), and distributes them to the non-root processes, so that the i-th process owns the i-th submesh. The root process also computes the right-hand side vector of (19) and distributes $\mathbf{d}_1^{(i)}$ to the other processes. Next, each non-root process assembles the four matrices $\mathbf{C}_{11}^{(i)}$, $\mathbf{C}_{1\Gamma}^{(i)}$, $\mathbf{C}_{\Gamma 1}^{(i)}$ and $\mathbf{C}_{\Gamma\Gamma}^{(i)}$, and computes an LU decomposition of $\mathbf{C}_{11}^{(i)}$.

At this moment, the conjugate gradient solver is applied to the system (15). We use the zero vector as the initial guess, so there are two operations to be performed in parallel by the non-root processes: evaluation of the right-hand side vector $\mathbf{f}$ and matrix-vector multiplication $\mathbf{Sp}$ with $\mathbf{p}$ being an arbitrary vector of the same size as $\mathbf{x}_\Gamma$ stored in the memory of the root process. The outcome of both operations is a vector stored in the memory of the root process. All other vector operations are performed sequentially by the root process.

To evaluate $\mathbf{f}$ given by (22), each non-root process computes

$$\mathbf{f}^{(i)} = \mathbf{C}_{\Gamma 1}^{(i)} \left(\mathbf{C}_{11}^{(i)}\right)^{-1} \mathbf{d}_1^{(i)} \tag{24}$$

and sends the result $\mathbf{f}^{(i)}$ to the root process. This can be done in parallel. Then, the root process obtains $\mathbf{f}$ by

$$\mathbf{f} = \mathbf{d}_\Gamma - \sum_{i=1}^{N} \mathbf{f}^{(i)}. \tag{25}$$

The $\mathbf{Sp}$ multiplication for the matrix $\mathbf{S}$ decomposed according to (21) starts with sending the vector $\mathbf{p}$ from the root process to the non-root processes. Afterwards, each non-root process computes

$$\mathbf{s}^{(i)} = \mathbf{C}_{\Gamma\Gamma}^{(i)}\mathbf{p} - \mathbf{C}_{\Gamma 1}^{(i)} \left(\mathbf{C}_{11}^{(i)}\right)^{-1} \mathbf{C}_{1\Gamma}^{(i)}\mathbf{p} \tag{26}$$

and sends $\mathbf{s}^{(i)}$ to the root process which then sums the partial results $\mathbf{s}^{(i)}$:

$$\mathbf{Sp} = \sum_{i=1}^{N} \mathbf{s}^{(i)}. \tag{27}$$

Once the Schur system (15) is solved, the resulting vector of interface unknowns $\mathbf{x}_\Gamma$ is distributed by the root process to the non-root processes, and the interior unknowns are found in parallel using (23).

Our implementation of the Schur complement method stores all matrices in the CSR sparse matrix format. It is written in the C++ programming language, and all data transfers between the root process and the non-root processes are accomplished using the blocking MPI functions `MPI_Send()` and `MPI_Recv()` [5].

5 Results

We tested our implementation of the Schur complement method for the Stokes problem on the lid driven cavity flow problem with the following setting: $\Omega = [0,1]^2$, $\nu = 0.01$,

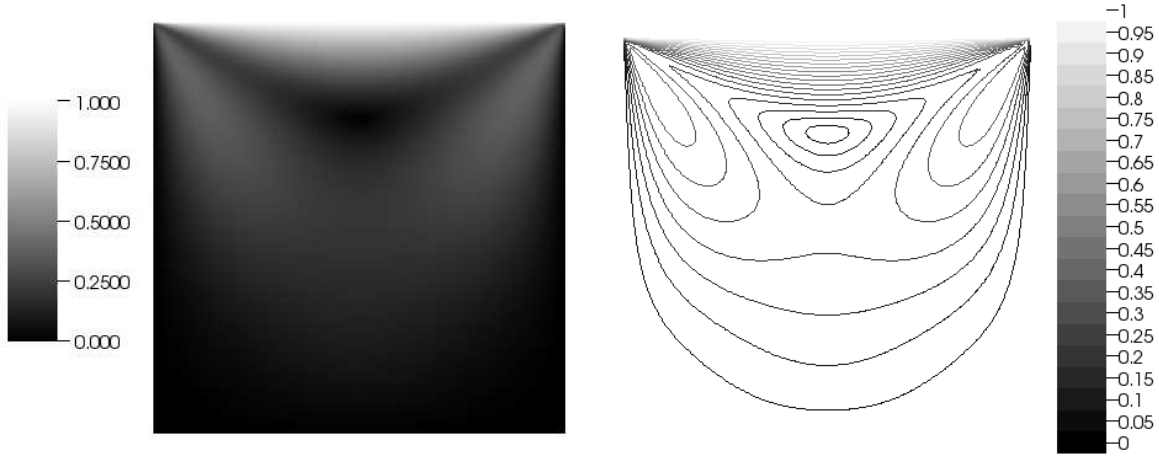


Figure 2: Solution of the lid driven cavity flow. Left: velocity magnitude, right: velocity magnitude contours.

$\vec{f} = 0$, $\vec{g}$ is zero except for its horizontal component on the upper part of the boundary, which is 1. For the FEM discretization, meshes obtained by uniform refinements of the mesh depicted in Figure 1 were used. The results are shown in Figure 2.

We measured running times of the solver and compared it with a sequential implementation of the GMRES method applied to (10). The computations were performed on the IBM SP6 cluster at CINECA. The maximum error tolerance was set to 10^{-8} times the Euclidean norm of the right-hand side of the respective system for all the methods, and the GMRES method was restarted after each 20 iterations. The results are presented in Table 1 including those achieved by a sequential multigrid method [1].

It follows from the results that the Schur complement method needed much less iterations to converge than the GMRES method, which agrees with our expectations. The performance of our implementation of the Schur complement method was strongly affected by the underlying LU decomposition of the matrices $\mathbf{C}_{11}^{(i)}$. If larger submeshes were used, i.e. $N = 2$ or $N = 8$, the LU decomposition consumed a significant amount of time and memory. Otherwise, the number of iterations increased, but the total time decreased (with the exception of the smallest problem). However, the multigrid solver still performed much better.

6 Conclusion

Our implementation of the Schur complement method performs worse than the multigrid method. The results suggest that it suffers from two issues: computational and memory cost of the LU decomposition of the matrices $\mathbf{C}_{11}^{(i)}$ and slow convergence of the conjugate gradient solver for the Schur complement system (15). In addition, in the case of large submeshes, the LU decomposition requires such huge amount of memory that its allocation fails. This restricts the usability of the implementation.

When dealing with large submeshes, avoiding the LU decomposition for inverting the matrices $\mathbf{C}_{11}^{(i)}$ might be beneficial. There are several options available: an iterative solver would be easy to implement on the CPU and the GPU; a multigrid-based solver

Problem DOFs	N	Iterations	LU-decomp. time	Total time
8 064	2	66	45.0	52.1
	8	706	0.8	5.0
	32	4 752	0.013	2.64
	128	15 807	0.000 4	16.5
	GMRES	65 992	–	155
	multigrid	–	–	0.07
32 512	2	80	2 730	2 860
	8	962	50.0	152
	32	6 127	0.76	38.6
	128	22 729	0.01	31.5
	GMRES	273 118	–	2 650.0
	multigrid	–	–	0.19
130 560	2	*LU decomposition out of memory*		
	8	2 110	2 710	6 200
	32	11 206	46.5	1 268
	128	30 237	0.7	222
	GMRES	*did not converge within time limit*		
	multigrid	–	–	0.61

Table 1: Comparison of solution methods. The time values are in seconds.

proved to be effective when solving the global system (10). Another possibility is to employ the Schur complement method recursively, i.e. for both, the global system and the subsystems. Such approach would allow to exploit another level of parallelism using, e.g., OpenMP.

To improve convergence of the conjugate gradient solver for the Schur complement system (15), a suitable preconditioner can be used (e.g. BDDC [8]).

References

[1] P. Bauer. *Mathematical modelling of pollution transport in urban canopy.* Dissertation thesis, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, (2010).

[2] M. Benzi, G. H. Golub, and J. Liesen. *Numerical solution of saddle point problems.* Acta Numerica **14** (2005), 1–137.

[3] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods.* Springer-Verlag, (1991).

[4] M. Crouzeix and P.-A. Raviart. *Conforming and nonconforming finite element methods for solving the stationary Stokes equations I.* Revue française d'automatique, informatique, recherche opérationnelle **7** (1973), 33–75.

[5] Message Passing Interface Forum. *MPI: A message-passing interface standard*, (2009). Version 2.2.

[6] Y. Saad. *Iterative methods for sparse linear systems.* Society for Industrial and Applied Mathematics, Philadelphia, (2003).

[7] A. Toselli and O. B. Widlund. *Domain decomposition methods – algorithms and theory.* Springer-Verlag, (2005).

[8] J. Šístek, B. Sousedík, P. Burda, J. Mandel, and J. Novotný. *Application of the parallel BDDC preconditioner to the Stokes flow.* Computers & Fluids **46** (2011), 429–435.