DOKTORANDSKÉ DNY 2012

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

16. a 23. listopadu 2012

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2012
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 312, Vydání 1.

ISBN 978-80-01-05138-2

Seznam příspěvků

Předmluva

Workshop Doktorandské dny je tradičním setkáním postgraduálních studentů oboru Matematické inženýrství, který je v rámci doktorského studijního programu Aplikace přírodních věd akreditován na FJFI. Obor je společně zajišťován katedrami matematiky, fyziky a softwarového inženýrství v ekonomii ve spolupráci s několika ústavy Akademie věd ČR. Letošní, již sedmý ročník workshopu se koná ve dnech 16. a 23. listopadu 2012, opět s laskavou podporou vedení KM FJFI.

Na konferenci doktorandi prezentují výsledky své práce za uplynulý rok. Jejich příspěvky jsou publikovány v tomto sborníku buď v plném znění, popřípadě zkrácené ve formě abstraktu, pokud byl obsah přednášky již otištěn v odborném časopise nebo ve sborníku jiné konference.

Vzhledem k širokému rozsahu témat, kterým se doktorandi Matematického inženýrství věnují, je program workshopu rozdělen do několika paralelních sekcí pokrývajících různé oblasti matematického modelování, teoretické informatiky a matematické fyziky. Příspěvky některých doktorandů jsou čistě teoretické, jiné se věnují aplikacím technicko-průmyslovým, socioekonomickým, či biomedicínským.

Za finanční zajištění průběhu konference vděčíme grantu SVK 25/12/F4.

Editoři

Alzheimer Disease Detection Based on FDR Analysis of Spect Images

Kamil Barbierik

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

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

Abstract. In this paper I present a computer-aided diagnosis technique for automatic detection of Alzheimer's disease (AD). It is based on FDR analysis of SPECT images and the results are used for classification task where the Learnig Vector Quantization (LVQ) is employed. The FDR is a tool for multi-hypothesis testing which was used in this work to define the regions of interest (ROIs) by identifying the voxels, where statistically significant difference in intensity was detected. Using this method we obtained compact areas as our ROIs. In further classification task we deal only with intensities in these ROIs. Thus, the computation burden is lowered and what is more, the maps formed by the ROIs which were projected on brain images were accepted as medically explainable by a group of anatomists. The average intensity of the biggest region, average intensity of the whole ROI and other statistics are subjected as futures to the classification task.

Keywords: Alzheimer's disease, SPECT, Multiple hypothesis testing, False discovery rate, Classification, Learning Vector Quantization (LVQ), Neural networks

Abstrakt. Táto práca popisuje metódu na automatické, počítačom riadené rozpoznávanie Alzheimrovej choroby zo SPECT obrázkov mozgu pacientov. Je založená na metóde FDR, ktorej výsledky sú ďalej použité pri klasifikácii obrázkov mozgov. FDR je nástroj pre testovanie multi-hypotéz, ktorý som využil pre vytvorenie oblastí záujmu, ktoré tvoria voxle, v ktorých boli detekované štatisticky významné rozdiely v intenzite. Použitím tejto metódy som získal kompaktné oblasti, ktoré majú význam aj z medicínskeho hľadiska v súvislosti s Alzheimrovou chorobou, čo potvrdila aj skupina neurológov, ktorým som výsledky prezentoval. Do algoritmu klasifikácie potom už vstupujú len voxle z vypočítanej oblasti záujmu, čím sa znižuje výpočetná náročnosť. Hodnoty ako je priemerná intenzita najväčšej oblasti v mape, ktorú tvoria pixle oblasti záujmu, priemerná intenzita celej oblasti záujmu a ďalšie štatistiky vstupujú klasifikačného algoritmu LVQ ako príznaky, podľa ktorých sú mozgy pacientov zaraďované do jednej z dvoch tried (Alzheimrova choroba alebo zdravý).

Klíčová slova: Alzheimrova choroba, SPECT, testovanie štatistických hypotéz, False discovery rate, Klasifikácia, Learning Vector Quantization (LVQ), Neurónové siete

1 Introduction

For last three decades a great progress was made in medical engineering and computing technologies. Also medical imaging technologies have witnessed a tremendous growth that has made a major impact in diagnostic radiology. These advances allow improving

health-care substantially by revealing critical diseases such as cancer, brain tumors etc. in early stages when the treatment is more effective.

One of non-invasive diagnostic tools that provide clinical information regarding biochemical and physiologic processes in patients particularly in patients' brain is called Single-photon emission computed tomography (SPECT). It is a nuclear imaging method based on the distribution of radiopharmaceutical agent in organ of interest. A radiotracer is injected in the patient's vein. As the tracer decays, it emits a photon, which is detected and recorded by the SPECT gamma camera. The computer then reconstructs these detections to produce a 3D tomographic image of blood flow throughout the investigated organ.

In this paper, I propose a method for finding a statistically significant difference between patients' brains suffering from Alzheimer's disease and normal controls i.e. healthy patients' brains. Mathematical statistics offers hypothesis testing to solve this kind of problem. I am able to test hypotheses about equality in each voxel of compared groups of brains and control the type I error for each test. Since there are thousands of voxels and thus thousands of tests, one would like to control the overall error rate, what is much more complicated problem. Nowadays several mechanisms were proposed how to control the compound error in multiple-hypothesis tests [1], [2] and this method was recently employed in more or less similar problems [3], [4], [5].

Statistical features of resulting regions of interest are then used in classification of brains in two groups (Healthy or AD). For classification the LVQ method was used with quite promising results.

2 Multiple hypothesis testing

The single-hypothesis testing is well known procedure. One is testing a null hypothesis H_0 against an alternative H_1 based on a statistic X. Let τ be some given rejection region. When $X \in \tau$, the H_0 is rejected otherwise when $X \notin \tau$ the H_0 is accepted. When H_0 is really true and $X \in \tau$ then the Type I error occurred. On the other hand, Type II error (β) occurs when an alternative H_1 is really true and the test accepts the null hypothesis H_0 because $X \notin \tau$. By choosing a probability α of occurrence of the Type I error, we determine the rejection region τ. Each rejection region has α or less probability of Type I error. Among them, the one with lowest error II is chosen. This is quite successful approach and we can often find a region with very good power $(1 - \beta)$ while maintaining the desired α level.

When one needs to test multiple hypotheses at once and control the overall error, the situation becomes much more complicated. With the increasing number of tests performed on data set the probability of rejecting the null hypothesis when it is true is rising. It means the Type I error is getting larger. This fact arises from the following logic: We reject the null hypothesis if we witness a rare event. But the larger the number of tests, the easier it is to find a rare event and make wrong decision about rejecting the null hypothesis when it is true. This effect is called the inflation of alpha level. To overcome this problem we should correct the original alpha level when performing multiple tests. Lowering the alpha level may be a good idea. It will create fewer errors but if the new alpha is too stringent, it may also make it harder to detect real effects.

Suppose we have a problem of testing m null hypotheses H_0 of which m_0 are true and R hypotheses were rejected. The following table (Table 1) describes the situation. It shows, for example, also a number (V) of null hypotheses that were rejected even when the null hypotheses were true (number of Type I errors).

Hypothesis H_0	Accepted	Rejected	Total
True	U	V	m_0
False	T	S	m_1
	W	R	m

Table 1: Values describing the situation when m hypotheses are tested

Assume that the number of hypotheses m is known in advance. R is an observable random variable and so is W because $W = m - R$. Random variables U, T, V, S are, on the contrary, unobservable. In the following text the lower case of their equivalents will be used for their realized values.

2.1 Family wise error rate (FWER)

The first measure to be suggested to control the overall error rate would be the family wise error rate (FWER). We will call the family of tests the series of tests performed on a set of data. This rate is a probability of making at least one Type I error in the whole family of tests: $P(V > 0)$.

Suppose that we have set the significance level α_T (alpha per test) at some value for each test in the family. The probability of Type I error for one test is then α_T. Events of making the Type I error and not making this error are complementary. Therefore the probability of not making a Type I error is $1 - \alpha_T$. Suppose another m independent events. The probability of not making a Type I error is then $(1 - \alpha_T)^m$. We need a probability of complement to this event; probability of making one or more Type I errors:

$$\alpha_F = 1 - (1 - \alpha_T)^m \tag{1}$$

So we have a value α_F, which is the probability of making at least one Type I error for the whole family of tests. By solving the equation for α_T assuming independence of tests we obtain

$$\alpha_T = 1 - (1 - \alpha_F)^{1/m} \tag{2}$$

This is called the Šidák equation [1]. It shows how to adjust the α_T if we want the α_F be fixed on some value e.g. $\alpha_F = 0.05$. Such control guarantees the following:

$$P(V > 0) \leq \alpha_F. \tag{3}$$

Example: We have 100 tests in a family. We want the α_F the probability of making one or more Type I errors to be $\alpha_F = 0.05$. The question is, how to adjust the α_T value for each test to obtain the required probability over the whole family. Formula (2) gives the answer:

$$\alpha_T = 1 - (1 - 0.05)^{1/100} = 5.128 \times 10^{-4}. \tag{4}$$

Because the Šidák equation is a bit difficult to compute due to the fractional exponent, a simpler expression was derived to compute the approximation using the first linear term of a Taylor expansion of the Šidák equation:

$$\alpha_T \approx \alpha_F/m. \tag{5}$$

This approximation is known as Bonferroni approximation and is related to Šidák expression as follows:

$$\alpha_T = 1 - (1 - \alpha_F)^{1/m} \leq \alpha_F/m. \tag{6}$$

Values in the inequality are very close to each other but Bonferroni approximation is pessimistic. Probably because of easier computation is the Bonferroni approximation more frequently used in practice then the Šidák expression. More powerful FWER controls are available: [6], [7], [8] but still they suffer from low power to detect a specific hypothesis when the number of tests in the family increases.

2.2 The false discovery rate (FDR)

In many situations, especially, when we are dealing with large number of tests, the FWER is much too strict. Benjamini and Hochberg [2] suggested a new approach of controlling the error in multiple hypotheses testing. They proposed the FDR, the expected proportion of erroneous rejections among all rejections.

Definition of FDR
Let Q be the unobservable random variable defined as follows:

$$Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

where V and R are values from Table 1. Then the FDR is simply:

$$FDR \equiv \mathbb{E}\left(Q\right). \tag{8}$$

So instead of controlling an occurrence of at least one erroneous rejection, what is not always as crucial for drawing conclusions from the family tested, the proportion of errors is tested. Thus, when many tests are rejected we are ready to bear with more errors, but with less when fewer tests are rejected. Two properties of this error rate can be easily shown:

a) If all null hypotheses are true then FDR is equivalent with FWER. In such case $s = 0$ and $v = r$ If $v = 0$ then $r = 0$ so there is no rejected hypothesis thus $Q = 0$. If $v > 0$ then v/r is always 1 so $Q = 1$. This leads to $P(V \geq 1) = \mathbb{E}(Q)$. Therefore control of FDR implies control of FWER.

b) In the other hand when $m > m_0$ the FDR is smaller than or equal to FWER: $P(V \geq 1) \geq \mathbb{E}(Q)$. As a result any procedure that controls FWER controls also the FDR but if the procedure controls only the FDR then it can be less stringent and thus the gain in power may be expected.

The procedure of controlling the FDR supposing that the tests in the family are independent is as follows:

1. Sort the p-values of each test in ascending order: $0 \leq p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ being $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ their respective null hypothesis.

2. Search for the maximum $q^* \in \{1, 2, \ldots m\}$ such that

$$p_{(q^*)} \leq \frac{q^* \cdot \alpha_F}{m} \tag{9}$$

3. Reject all hypotheses $H_{(i)}$ for $i \in \{1, 2, \ldots, q^*\}$

In case when we are dealing with dependent tests we need to adjust the formula in the second step of controlling algorithm [11]:

$$p_{(q^*)} \leq \frac{q^* \cdot \alpha_F}{m \cdot C(m)} \tag{10}$$

where

- $C(m) = 1$ if test are positively correlated
- $C(m) = \sum_{i=1}^{m} \frac{1}{i}$ in case the tests are negatively correlated

If we have no a priori knowledge regarding the correlation type, we can assume positive test correlation as kind of optimistic testing approach, meanwhile the supposition of negative tests correlation leads to pessimistic and thus more stringent testing approach.

The described procedure of controlling FDR guarantees the following:

$$FDR \equiv \mathbb{E}\left(\frac{V}{R}\right) \leq \alpha_F \tag{11}$$

3 3D SPECT images analysis

3.1 Subjects

Subject of my analysis are 3D SPECT images of brains of two groups of patients: the group of 38 patients suffering from Alzheimer's disease (AD) and a group of 55 healthy people. I will call the latter group as normal controls (NC). These images were provided, analyzed and classified by professionals in the field, so we may assume correct classification of the patients, whether they are suffering from AD or they are healthy and therefore belong to NC group. Thus the SPECT images of patients' brains are also correctly labeled.

3.2 Methods

My intention is to compare the AD scans against the NC ones and discover some significant difference. For this task I employ the multiple hypotheses testing where the overall error is controlled by the FDR controlling procedure described in section 2.2. To be able to find some differences in brains by comparing them against each other, I need to be sure that the images are registered properly. For this purpose the SPM5 software were employed (Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, London UK - http://www.fil.ion.ucl.ac.uk/spm). The Statistical Parametric Mapping is described for example in [9].

Different amount of radiopharmaceutical agent injected to patient, different absorption properties of body or other factors may influence the global intensity of the resulting image. To reduce the effect of this "false differences" in intensities during comparison, it is necessary to apply some intensity normalization. Among several possibilities we have chosen to divide each voxel by the sum over all voxels of processed image.

When images are correctly registered and transformed in intensities, we create average images separately for AD and NC group of images by summing up values of voxels through every image and dividing them by number of images in the group. These average images are of the same size as the images they were created from: $79 \times 95 \times 69$ voxels i.e. approximately half a million voxels. Thus by testing each voxel from AD average image against NC average image weather the values in particular voxels (mean through the images in the particular group) are different on some significance level, we obtain half million statistical paired t-tests. The null hypothesis is the same for each test:

$$H_0^{(i,j,k)} : \mu_{AD}^{(i,j,k)} = \mu_{NC}^{(i,j,k)}. \tag{12}$$

After computing the p-values of each test we apply the multi-hypothesis testing approach described in section 2.2. The overall significance level was set to $\alpha_F = 0.001$.

3.3 Result

Black and white regions denote voxels where significant difference has been detected between AD and NC pictures. White regions are regions where AD pictures have higher intensities in average and vice versa black regions denote regions with lower average intensities in AD pictures.

4 Experiments on classification using the regions of interest (ROIs) from previous section

The ROIs from previous section are regions where significant difference was discovered between average AD and NC brain. This reduction of voxels to set that we are interested in, reduces the computation burden and makes us focus on areas where most significant changes in blood flow occur by affection of AD. To automatically classify weather examined patient suffers from AD based on these areas I have chosen a method using neural network: the Learning Vector Quantization (LVQ).

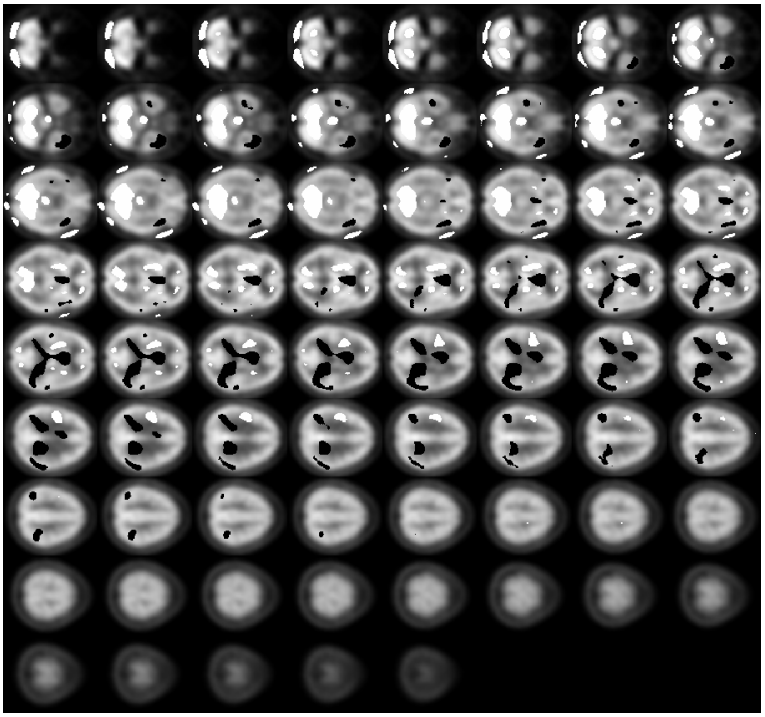


Figure 1: Regions of significant differences between AD and NC brains (black and white). White: higher average intensities in AD images; black: lower average intensities in AD images (in comparison to NC images)

4.1 Classification using LVQ

LVQ method was invented by Teuvo Kohonen [13]. It is related to k Nearest Neighbors algorithm well known from pattern recognition. It is a special case of ANN which applies winner-takes-all learning based approach. The LVQ network architecture is shown on Figure 2.

The input layer contains as many elements as is the number of feature space dimension. The competitive layer contains S^1 elements. Weights in competitive layer need to be initialized during creation of the network in some way. In our case we used Matlab default initialization that set the weights to the "middle" of training data clusters. The linear layer is composed of S^2 output neurons while $S^2 < S^1$. S^2 corresponds to the number of final classes that we want the input data to be classified in. The weights in linear layer are initialized by zeroes and ones according the user input parameter what is a vector of typical class percentages. The purpose of the linear layer is to combine subclasses from the competitive layers and bring results to the output layer i.e. final classes.

The output of competitive layer is a column vector a^1 with one non-zero element in i^*th row. We say that neuron i^* won the competition in competitive layer because the input and weights corresponding to neuron i^* were nearest to each other. Linear layer now classifies this winning neuron to one of the final class. When we are in the learning process, we need to evaluate the result and adjust weights in competitive layer

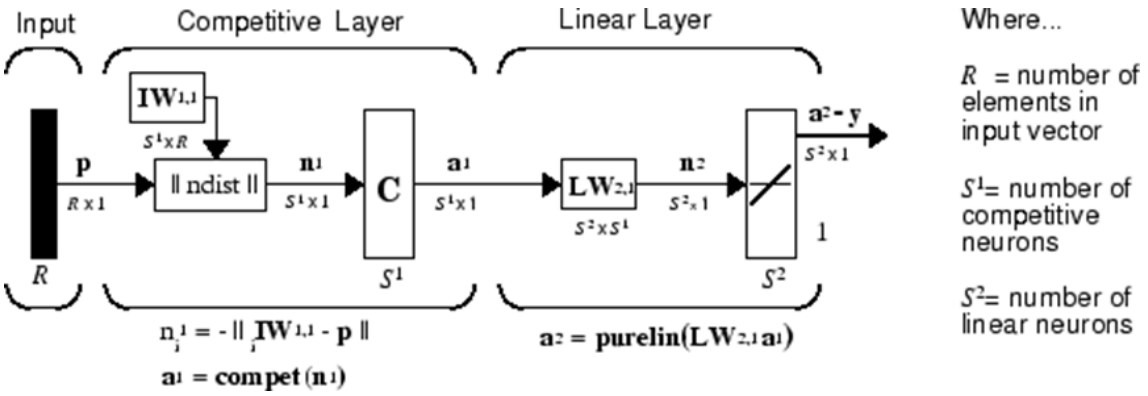


Figure 2: LVQ network architecture

corresponding to winning neuron. If the input was classified correctly we update the i^*th row of matrix $IW^{(1,1)}$ in a way to move this row or hidden neuron closer to the particular input:

$$ {}_{i^*}IW^{1,1}(q) = {}_{i^*}IW^{1,1}(q-1) + \alpha \cdot \big(p(q) - {}_{i^*}IW^{1,1}(q-1)\big) \tag{13} $$

On the other hand if input p is classified incorrectly we move the hidden neuron away from input:

$$ {}_{i^*}IW^{1,1}(q) = {}_{i^*}IW^{1,1}(q-1) - \alpha \cdot \big(p(q) - {}_{i^*}IW^{1,1}(q-1)\big) \tag{14} $$

4.2 Results

With respect to the results of FDR analysis I have chosen the following futures for classification process (see Figure 3 for some feature space cuts):

A. Average intensity of the biggest region

B. Average intensity of the whole ROI

C. Difference of averages of white and black regions

D. 1., 2., 3. coordinate of weighted center of mass of ROI

To classify the brain pictures I set up an LVQ neural network with architecture described by Figure 4 and paramaters (see Table 2):

Epochs	150
Learning step α	0.01
Typical class percentages	(0.60 0.40)

Table 2: LVQ network parameters setting

The network consists of input layer with 6 neurons meaning that we have feature space of dimension 6. The competitive layer is composed by 8 hidden neurons combined

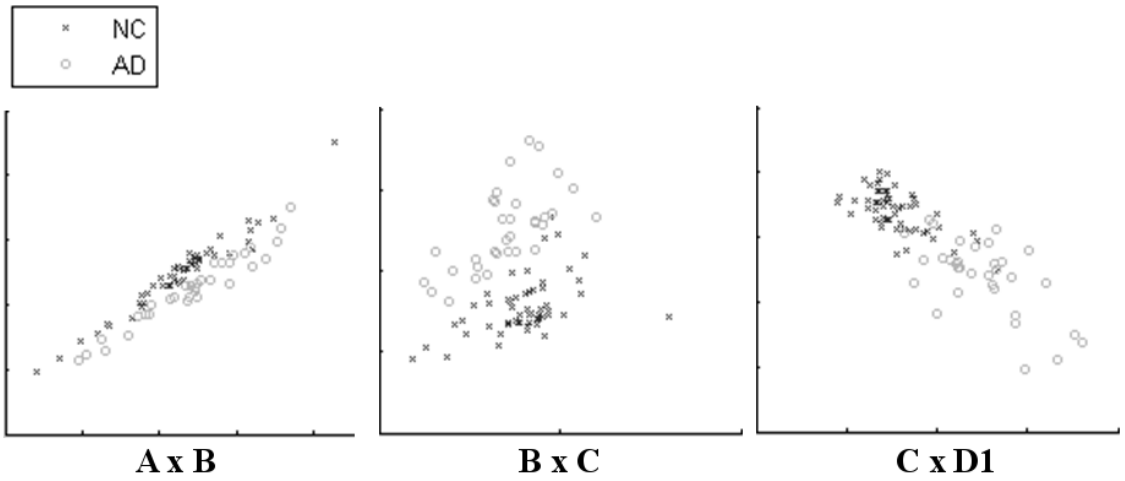


Figure 3: Cuts of feature space

Figure 4: Matlab neural network architecture

in linear layer into two output neurons that represent 2 final classes i.e. class of AD and class of NC. Data, used for training and testing the neural network, are described in the Table 4.

I provided 6 independent runs of the experiment, where in each run a new training set is randomly chosen. The ROI is generated from images in training set. Results of classification by this network using described data are summarized in the following table:

5 Conclusion

I applied the FDR method to discover differences between SPECT pictures of healthy people and people suffering from Alzheimer's disease. As SPECT images provide data about regional blood flow, I am looking for changes in brain perfusion caused by Alzheimer's disease.

As we can see on the resulting image, the areas where significant changes have been detected are compact what is a meaningful observation acknowledged also by neuroanatomists from Faculty Hospital King's Vineyards. Also other information that is provided by the resulting image was accepted as medically explainable after presenting them to the mentioned group of anatomists.

After applying this method, we are left with much less voxels to work with, what

SPECT Alzheimer (AD)	55×
SPECT Normal control (NC)	91×
Training set	60%

Table 3: Available data

Run	Wrong NC classification	Wrong AD classification
1.	**3%** (1/36)	**9%** (2/22)
2.	**8%** (3/36)	**27%** (6/22)
3.	**14%** (5/36)	**14%** (3/22)
4.	**6%** (2/36)	**9%** (2/22)
5.	**11%** (4/36)	**0%** (0/22)
6.	**19%** (7/36)	**23%** (5/22)

Table 4: Available data

reduces the computation burden, while the efficiency regarding the further classification task is not lowered. For classification task the LVQ algorithm was chosen and the results are quite promising.

Further research will be carried out to fine tune this automatic Alzheimer disease detection in order to achieve better results. Using other intensity normalizations or different classification methods may lead to actual results improvement.

Acknowledgement: The support of grant OHK4-165/11 CTU in Prague is gratefully acknowledged as well as valuable notes and reviews from medical point of view by group of neurologists from the Faculty Hospital King's Vineyards namely Aleš Bartoš, Renata Píchová and Helena Trojanová.

References

[1] Z. Šidák. *Rectangular confidence region for themeans of multivariate normal distributions.* In Journal of the American Statistical Association 62, No. 313, 1967, pp. 626-633.

[2] Y. Benjamini, Y. Hochberg. *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.* In J. R. Statist. Soc. B 57, 1995, No. 1, pp. 566-568.

[3] M. Mojzeš, J. Kukal, V. Q. Tran, J. Jablonský. *Performance comparison of heuristic algorithms via multi-criteria decision analysis.* In Proceedings of 17th International Conference on Soft-Computing MENDEL 2011, Brno: FME BUT, 2011, pp. 224-251.

[4] P. N. Jayakumar, G. Venkatasubramanian, N. Gangadhar, N. Janakiramaiah, M. S. Keshavan. *Optimized voxel-based morphometry of gray matter volume in first-episode, antipsychotic-naive schizophrenia.* In Progress in Neuro-Psychopharmacology and Biological Psychiatry 29, 2005, pp. 587-591.

[5] K. Egger, J. Mueller, M. Schocke, CH. Brenneis, M. Rinnerthaler, K. Seppi, T. Trieb, G. K. Wenning, M. Hallet, W. Poewe. *Voxel Based Morphometry Reveals Specific Gray Matter Changes in Primary Dystonia.* In Movement Disorders, vol. 22, no. 11, 2007, pp. 1538-1542.

[6] A. C. Tamhane, Y. Hochberg, Ch. W. Dunnet. *Multiple Test Procedures for Dose Finding.* In Biometrics 52, 1996, pp. 21-37.

[7] J. Shaffer. *Multiple hypothesis testing.* In Annual Review of Psychology, 1995; vol. 46, 561–584.

[8] J. C. Hsu. *Multiple Comparisons: Theory and Methods.* In New York 1996, Chapman & Hall.

[9] K. J. Friston, J. Ashburner, S. J. Kiebel, T. E. Nichols, W. D. Penny. *Statistical Parametric Mapping: The Analysis of Functional Brain Images.* In Academic Press, 2007.

[10] J. D. Storey. *A direct approach to false discovery rates.* In J. R. Statist. Soc. B 64, 2002; Part 3, pp.479-498.

[11] Y. Benjamini, D. Yekutieli. *The control of the false discovery rate in multiple testing under dependency.* In The Annals of Statistics, vol. 29, no. 4, 2001, pp. 1165-1188.

[12] A. P. Dhawan, H. K. Huang, D. Kim. *Principles and advanced methods in medical imaging and image analysis.* In World Scientific Publishing Co. Pte. Ltd., 2008.

[13] T. Kohonen. *Learning Vector Quantization.* In Neural Networks 1, Supplement 1, p. 303, 1988.

Conformal Sets in Neural Network Regression*

Radim Demut

3rd 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 predictive regions in regression models, especially neural networks. We use the concept of conformal prediction (CP) to construct regions which satisfy given confidence level. Conformal prediction outputs regions, which are automatically valid, but their width and therefore usefulness depends on the used nonconformity measure. A nonconformity measure should tell us how different a given example is with respect to other examples. We define nonconformity measures based on some reliability estimates such as variance of a bagged model or local modeling of prediction error. We also present results of testing CP based on different nonconformity measures showing their usefulness and comparing them to traditional confidence intervals.

Keywords: confidence intervals, conformal prediction, regression, neural networks

Abstrakt. Tento článek se zabývá konfidenčními množinami v regresních modelech, speciálně v regresi využívající neuronové sítě. Pro konstrukci konfidenčních oblastí používáme metod konformní predikce. Konformní predikce dává oblasti, které jsou vždy validní, ale jejich velikost, a tedy i užitečnost, závisí na použité míře nekonformity. Míra nekonformity by měla měřit, jak se jednotlivý příklad liší od celé skupiny příkladů. V tomto článku zavedeme několik měr nekonformity založených na odhadech spolehlivosti, jako jsou rozptyl bagged modelu nebo lokální model chyby odhadu. Prezentujeme také výsledky testování konformních oblastí na základě různých měr nekonformity, ukážeme užitečnost těchto oblastí a porovnáme je s tradičními konfidenčními intervaly.

Klíčová slova: konfidenční intervaly, konformní predikce, regrese, neuronové sítě

References

[1] Bosnic, Z., Kononenko, I., "Comparison of approaches for estimating reliability of individual regression predictions", *Data & Knowledge Engineering*, pp. 504–516, 2008.

[2] Gammerman, A., Shafer, G., Vovk, V., "Algorithmic learning in a random world", *Springer Science+Business Media*, 2005.

[3] Uusipaikka E., "Confidence Intervals in Generalized Regression Models", *Chapman & Hall*, 2009.

[4] Papadopoulos, H., Vovk, V., Gammerman, A., "Regression conformal prediction with nearest neighbours", *Journal of Artificial Intelligence Research*, vol. 40, pp. 815–840, 2011.

*The paper was presented at the ITAT 2012 conference and is published in the conference proceedings.

[5] Valero, S., Argente E., et al., "DoE framework for catalyst development based on soft computing techniques", *Computers and Chemical Engineering*, vol. 33, No. 1, pp. 225–238, 2009.

On the Generalizations of the Unit Sum Number Problem*

Daniel Dombek†

3rd 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 representations of algebraic integers of a number field as linear combinations of units with coefficients coming from a fixed finite set, and as sums of elements having small norms in absolute value. These theorems can be viewed as results concerning a generalization of the so-called unit sum number problem, as well. Beside these, extending previous related results we give an upper bound for the length of arithmetic progressions of t-term sums of algebraic integers having small norms in absolute value.

Full version of this contribution, *Representing algebraic integers as linear combinations of units*, will appear in Periodica Mathematica Hungarica [6].

Keywords: number field, linear combinations of units, arithmetic progressions

Abstrakt. Tento příspěvek se zabývá reprezentacemi algebraických celých čísel v číselném tělese jako lineární kombinace jednotek s koeficienty z dané konečné množiny, a dále i jako sumy prvků tělesa s tělesovou normou v absolutní hodnotě omezenou. Uvedené výsledky lze považovat za zobecnění tzv. unit sum number problému. Na závěr je, v návaznosti na dosavadní známé výsledky, odvozena horní mez pro délku aritmetických posloupností t-členných sum algebraických celých čísel s omezenou tělesovou normou.

Nezkrácená verze tohoto příspěvku, *Representing algebraic integers as linear combinations of units*, vyjde v časopise Periodica Mathematica Hungarica [6].

Klíčová slova: číselné těleso, lineární kombinace jednotek, aritmetické posloupnosti

1 Introduction

Let K be an algebraic number field with ring of integers O_K. The problem of representing elements of O_K as sums of units has a long history and a very broad literature. For the sake of brevity, we refer to the excellent survey paper of Barroero, Frei and Tichy [2] and the references there. Now we mention only those results which are most important from our viewpoint.

*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.

†joint work with L. Hajdu and A. Pethő, University of Debrecen

After several partial results due to Ashrafi and Vámos [1] and others, Jarden and Narkiewicz [11] proved that for any number field K and positive integer t, one can find an algebraic integer $\alpha \in K$ which cannot be represented as a sum of at most t units of K.

Observe that if K admits an integral basis consisting of units then clearly every integer of K can be represented as a sum of units. For results in this direction we refer to a paper of Pethő and Ziegler [18]. Showing that (up to certain precisely described exceptions) every number field admits a basis consisting of units with small conjugates, we prove that allowing a small, completely explicit set of (rational) coefficients every integer of K can be expressed as a linear combination of units. We would like to emphasize the interesting property that the set of coefficients allowed depends only on the degree and the regulator of K and that the latter dependence is made explicit.

Further, it is also well-known (see e.g. [2] again) that there are infinitely many number fields whose rings of integers are not generated additively by their units. In other words, in these fields one can find algebraic integers α which cannot be represented as a sum of (finitely many) units at all.

In this paper we extend this investigations to the case where one would like to represent the algebraic integers of K not as a sum of units, but as a sum of algebraic integers of small norm, i.e. using algebraic integers with $|N(\beta)| \leq m$ for some positive integer m. Obviously, taking $m = 1$ we just get back the original question. First we prove that the above mentioned result of Jarden and Narkiewicz extends to this case: for any algebraic number field K and positive integers m and t one can find an algebraic integer $\alpha \in K$ which cannot be obtained as a sum of at most t integers of K of norm $\leq m$ in absolute value. Then we show that in contrast with the original case, one can give a bound m_0 depending only on the discriminant and degree of K, such that if $m \geq m_0$ then already every integer of K can be represented as a sum of integers of K with norm at most m in absolute value. Note that as it is well-known, any number field K contains only finitely many pairwise non-associated algebraic integers of given norm. Hence sums of elements of small norm can be considered as linear combinations of units with coefficients coming from a fixed finite set.

Finally, we also provide a result concerning arithmetic progressions of t-term sums of algebraic integers of small norm in a number field K. This result generalizes previous theorems of Newman (concerning arithmetic progressions of units; see [15] and [16]) and of Bérczes, Hajdu and Pethő (concerning arithmetic progressions of elements of fixed norm; cf. [3]).

2 Main results

From this point on, let K be an algebraic number field of degree k, with discriminant $D(K)$ and regulator $R(K)$. Write O_K for the ring of integers of K, $N(\beta)$ for the field norm of any $\beta \in K$ and U_K for the group of units in O_K.

The unit sum number problem can be considered as a question about linear combinations of units with rational integers. We know that the resulting set is sometimes a proper subset with infinite complementer of O_K. However if we allow that the coefficients have small denominators, then the situation becomes completely different.

At this point let us recall that the field K is called a CM-field, if it is a totally imaginary quadratic extension of a totally real number field.

Theorem 2.1. *Suppose that either K is not a CM-field, or K is a CM-field containing a root of unity different from ± 1. Then there exists a positive integer $\ell = e^{c_1(k)R(K)}$ where $c_1(k)$ is a constant depending only on the degree of K, such that any $\alpha \in O_K$ can be obtained as a linear combination of units of K with coefficients $\{1, 1/2, 1/3, \ldots, 1/\ell\}$.*

Remark 2.1. *The condition that K is not a CM-field or K contains a non-real root of unity is necessary. Indeed, otherwise all units of K are contained in some proper subfield of K, and the statement trivially fails.*

Denote by σ_i $(i = 1, \ldots, k)$ the embeddings of K into $\mathbb{C}$ and for $\alpha \in K$ put $\overline{|\alpha|} = \max_{1 \le i \le k}(|\sigma_i(\alpha)|)$. The following statement is vital for the proof of Theorem 2.1. Moreover, we think that it is interesting also on its own.

Proposition 2.1. *Suppose that either K is not a CM-field, or K is a CM-field containing a root of unity different from ± 1. Then there exists a constant $c_2 = c_2(k)$ depending only on the degree of K, such that K has a basis consisting of units ε_i with $\overline{|\varepsilon_i|} \le e^{c_2(k)R(K)}$, $(i = 1, \ldots, k)$.*

Now we consider the case, where the summands belong to a set of integers of small norm in K. As a motivation, we mention that Newman proved that the length of arithmetic progressions consisting of units of K is at most k (see [15] and [16]). This result has been generalized by Bérczes, Hajdu and Pethő in [3] to arithmetic progressions in the set

$$\mathcal{N}_m := \{\beta \in O_K \ : \ N(\beta) = m\},$$

where $m > 0$. Now we present a result concerning a further generalization of this problem. For $m > 0$ put

$$\mathcal{N}_m^* := \{\beta \in O_K \ : \ |N(\beta)| \le m\},$$

and write

$$t \times \mathcal{N}_m^* := \{\beta_1 + \cdots + \beta_t \ : \ \beta_i \in \mathcal{N}_m^* \ (i = 1, \ldots, t)\}$$

where t is a positive integer.

First theorem gives a bound for the lengths of arithmetic progressions in the sets $t \times \mathcal{N}_m^*$.

Theorem 2.2. *The length of any non-constant arithmetic progression in $t \times \mathcal{N}_m^*$ is at most $c_3(m, t, k, D(K))$, where $c_3(m, t, k, D(K))$ is an explicitly computable constant depending only on m, t, and on the degree k and discriminant $D(K)$ of K.*

Now we present results concerning the above generalization of the unit sum number problem. Slightly modifying the notation of Goldsmith, Pabst and Scott [7] we define the unit sum number $u(O_K)$ as the minimal integer t such that every element of O_K is a sum of at most t units from U_K, if such an integer exists. If it does not, we put $u(O_K) = \omega$ if every element of O_K is a sum of units, and $u(O_K) = \infty$ otherwise. We use the convention $t < \omega < \infty$ for all integers t.

As we have mentioned already, Jarden and Narkiewicz [11] proved that $u(O_K) \geq \omega$ for any number field K. Our next result yields an extension of this nice theorem. To formulate it, we define the m-norm sum number $u_m(O_K)$ as an analogue to $u(O_K)$ with the exception that instead of sums of units we consider sums of elements from $\mathcal{N}_m^*$. Clearly, $u(O_K) = u_1(O_K)$ holds.

Theorem 2.3. *For every number field K and $m > 0$ we have $u_m(O_K) \geq \omega$, i.e. for every $m, t \in \mathbb{N}$ there exists an $\alpha \in O_K$ which cannot be obtained as the sum of at most t terms from $\mathcal{N}_m^*$.*

As it is well-known (see e.g. [2] and the references given there), for infinitely many number fields K we have $u(O_K) = \infty$. In contrast to this result, our next theorem shows that $u_m(O_K) = \omega$ is always valid if m is "large enough" with respect to the discriminant and the degree of K. More precisely, we have the following theorem.

Theorem 2.4. *For every number field K there exists a positive integer $m_0 = m_0(D(K), k)$ depending only on the discriminant and the degree of K, such that for any $m \geq m_0$ we have $u_m(O_K) = \omega$, i.e. any $\alpha \in O_K$ can be obtained as the sum of elements from $\mathcal{N}_m^*$.*

Observe that sums of elements of $\mathcal{N}_m^*$ can be also viewed as linear combinations of units with coefficients coming from a fixed finite set.

References

[1] N. Ashrafi, P. Vámos, *On the unit sum number of some rings*, Q. J. Math. **56** (2005), 1–12.

[2] F. Barroero, C. Frei, R. F. Tichy, *Additive unit representations in rings over global fields - a survey*, Publ. Math. Debrecen **79** (2011), 291–307.

[3] A. Bérczes, L. Hajdu, A. Pethő, *Arithmetic progressions in the solution sets of norm form equations*, Rocky Mountain Math. J. **40** (2010), 383–396.

[4] Y. Bugeaud, K. Győry, *Bounds for the solutions of unit equations*, Acta Arith. **74** (1996), 67–80.

[5] A. Costa, E. Friedman, *Ratios of regulators in totally real extensions of number fields*, J. Number Theory **37** (1991), 288–297.

[6] D. Dombek, L. Hajdu, A. Pethő, *Representing algebraic integers as linear combinations of units*, to appear in Period. Math. Hungar. (2012), 9pp.

[7] B. Goldsmith, S. Pabst, A. Scott, *Unit sum numbers of rings and modules*, Q. J. Math. **49** (1998), 331–344.

[8] L. Hajdu, *Arithmetic progressions in linear combinations of S-units*, Period. Math. Hungar. **54** (2007), 175–181.

[9] L. Hajdu, F. Luca, *On the length of arithmetic progressions in linear combinations of S-units*, Archiv der Math. **94** (2010), 357–363.

[10] H. Hasse, *Number theory*. Translated from the third (1969) German edition. Edited and with a preface by Horst Günter Zimmer. Classics in Mathematics. Springer-Verlag, Berlin (2002).

[11] M. Jarden, W. Narkiewicz, *On sums of units*, Monatsh. Math. **150** (2007), 327–332.

[12] E. Landau, *Abschätzungen von Charaktersummen, Einheiten und Klassenzahlen*, Nachr. Akad. Wiss. Göttingen (1918), 79–97.

[13] K. Mahler, *Inequalities for ideal bases in algebraic number fields*, J. Austral. Math. Soc. **4** (1964), 425–448.

[14] M. R. Murty, J. Van Order, *Counting integral ideals in a number field*, Expo. Math. **25** (2007), 53–66.

[15] M. Newman, *Units in arithmetic progression in an algebraic number field*, Proc. Amer. Math. Soc. **43** (1974), 266–268.

[16] M. Newman, *Consecutive units*, Proc. Amer. Math. Soc. **108** (1990), 303–306.

[17] W. Narkiewicz, *Elementary and analytic theory of algebraic numbers*. Polska Akademia Nauk., Instytut Matematyczny, Monografie matematyczne **57** (1974).

[18] A. Pethő, V. Ziegler, *On biquadratic fields that admit unit power integral basis*, Acta Math. Hungar. **133** (2011), 221–241.

[19] V. G. Sprindžuk, *"Almost every" algebraic number-field has a large class-number*, Acta Arith. **25** (1973/74), 411–413.

Freeconf: A General-purpose Multi-platform Configuration Utility

David Fabian

1st year of PGS, email: fabiadav@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. Configuration tasks can be divided into two groups. Firstly, one has to select right software components from a set and merge them together so that the result meets the customer's requirements. Secondly, the application must be deployed and certain parameters must be tweaked according to the user's needs. Many times, even a very large-scaled professional software does not provide a user-friendly way to set these parameters. This article introduces Freeconf, a new general-purpose multi-platform configuration utility which has been designed to help the user with the deployment and the maintenance of a broad range of existing applications.

Keywords: software, configuration, multi-platform, configuration file, maintenance

Abstrakt. Problémy konfigurace lze dělit na dvě skupiny. Za prvé je nutné vytvořit aplikaci ze samostatných softwarových komponent tak, aby výledek splňoval požadavky zákazníka. Za druhé je výslednou aplikaci nutné nainstalovat a upravit pak některé její parametry podle přání uživatele. Mnohdy však ani rozsáhlé a profesionální aplikace neposkytují přívětivé prostředí pro nastavení těchto parametrů. V tomto článku je popsán nový obecný multiplatformní konfigurační nástroj Freeconf, který byl navržen tak, aby usnadnil instalaci a následnou údržbu celé řady hotových aplikací.

Klíčová slova: software, konfigurace, multiplatformní, konfigurační soubor, údržba

1 Introduction

Nowadays, software configuration is ubiquitous. Large software companies deal with problems of creating a requirement-conforming application from a set of pre-made components which is a typical configuration problem. When the application has been developed and tested, it must be deployed to the customer. This is called installation and during this process the application is provided with the installation specific configuration data. This data can later be modified in reaction to the user's preference or the computer's environment change. In this article, the interest will be focused on configuration problems in software installation and maintenance.

Almost every software application provides a method for its adjustment. Often, the configuration itself is stored in a text file (or multiple files) and sometimes there is a graphical layer to assist the user in filling in the desired changes. This is, of course, the best solution for the beginner, since she does not have to understand the syntax of the configuration file and even know where the file is stored on the hard drive.

A lot of software (especially GNU/Linux software), however, does not provide a graphical user interface (GUI) and the user is forced to use the configuration text file directly. For such a software, since it is in many cases impossible to add GUI to the source code of the application, it would be desirable to have an intermediate layer on top of the existing configuration workflow. There exist such tools, e.g., MenuConfig [6] for setting up the Linux kernel, YaST [3] for setting up the entire openSUSE Linux distribution, or KConfigXT [5] used in KDE environment for modeling of configuration windows. In this article, a new multi-platform general purpose configuration utility Freeconf is presented.

The rest of the article is divided as follows. In Section 2, key concepts of Freeconf are presented. In Section 3, the Freeconf library is described. Section 4 introduces Freeconf client applications and their purpose. In Section 5, the Freeconf configuration package and its structure is described. Section 6 brings a short introduction to the Freeconf package designer and its functions. Lastly, in Section 7 there is a short conclusion.

2 Freeconf

Freeconf is a project started at the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague. Its primary purpose is the simplification and the unification of the configuration process of a great variety of applications (both existing and newly developed) together with the ability to semi-automatically generate a clear GUI layer on multiple platforms.

2.1 Terminology and Requirements

As stated in Section 1, Freeconf is an intermediate layer above the existing infrastructure. The application does not communicate with Freeconf itself but reads the configuration file it understands (in Freeconf's terminology, the file is called the *native configuration file*), while Freeconf, on the other hand, generates the native configuration file from the data it obtains from the user. Figure 1 illustrates the data flow.

The key concept is the transformation from Freeconf's internal configuration file format to the native configuration file format. The transformation will be explained in Section 5.2.

The only thing Freeconf requires is that the native configuration file must come in a text form and that it contains a list (or possibly a tree) of configuration keys and their values, i.e., key-value pairs. The values can have on of the following supported types: boolean, number (integer or float with a restricted precision), string, string-list, and fuzzy (for multi-value choices). For trees of configuration options, the non-leave nodes are called *configuration sections*.

It can happen that some of the sections in the native configuration file will occur multiple times. For example, when configuring the Apache web server, one can create more virtual servers. Such sections, where the key structure is fixed but values differ, is called the *multiple configuration section*.

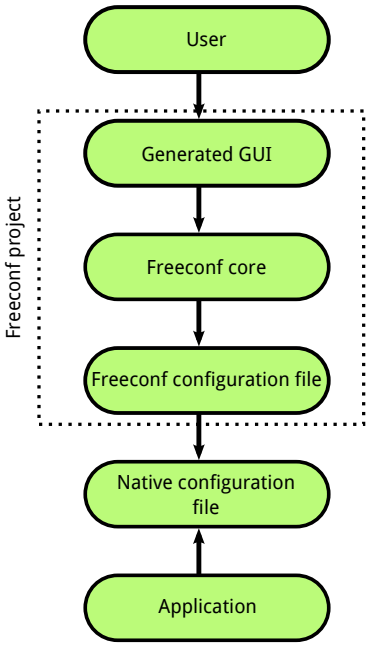


Figure 1: Configuration data flow from the user to the application.

2.2 Parts of the Project

The project has been divided into several parts to better achieve its goals. One of them is to provide the native look&feel on different software platforms. That is why the code has been split into the *Freeconf library* and *graphical clients*. The library is meant to exist in a single implementation for every supported platform, and it should contain the entire logic of the project. The graphical clients are supposed to be light-weight applications, the sole purpose of which should be to present a configuration dialog to the user. A client, when invoked by the user on a certain platform, connects to the Freeconf library and asks it for the configuration data. The library must load an appropriate model of the native configuration file which is called the *configuration package*, process it, and send the resulting data to the client. The client then builds a configuration dialog based on a structure provided by the library. When the user is satisfied with the changes she made, the client requests the library to store them in the package and to issue a transformation to the native configuration file.

Since creating a configuration package is a tedious task, a *package designer* has been developed. It is still a fairly simple program able only to generate a functional skeleton of a package. The designer also uses the library for package manipulation. Figure 2 shows all of the Freeconf components.

3 Freeconf Library

The Freeconf library is written in Python 2.7 (to speed up the development; later, it will be rewritten in C++) and provides the following capabilities:

- It can create, load, and save a configuration package.

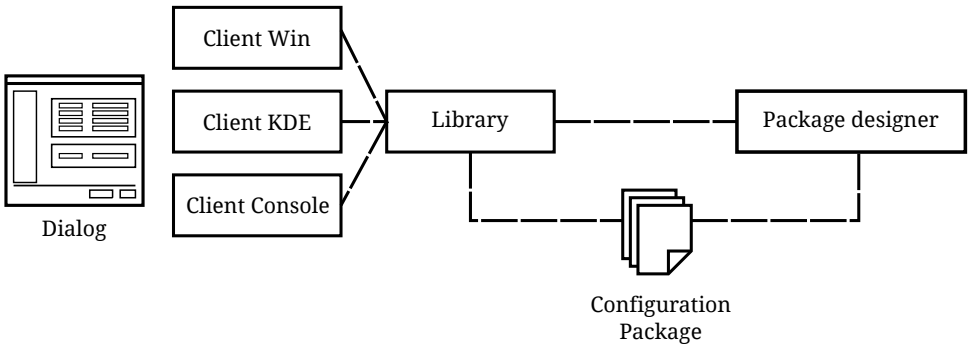


Figure 2: Components of the Freeconf project.

- It can preform a transformation from the Freeconf format to the native configuration file format.
- It processes the loaded package and constructs three in-memory data structures.
- It provides two interfaces: client–library and designer–library.

3.1 Data structures

The library organizes the loaded data in three tree structures.

- *Template tree* stores the key type and its properties. It also holds information about plug-ins and multiple containers.
- *Configuration tree* stores default and current values of the keys. Dependencies and inconsistency checks are evaluated on top of this tree.
- *GUI tree* stores data needed to construct a configuration dialog in the client. It contains various label texts, window dimension, and also takes care of hiding unnecessary keys.

When a package is loaded, the template tree is constructed at first. Then, default values, stored values from the previous configuration (if any), and keys dependencies are read. The configuration tree is constructed according to the template tree. In the past, the trees were identical in terms of structure, but now they can differ because of multiple configuration sections. As the last, the GUI tree is constructed and is linked to the configuration tree in a similar way the template tree is linked to the configuration tree — the corresponding nodes can reach each others through references.

3.2 Interfaces

At the moment, there are two interfaces present in the source code as stated above. There is not a single interface mainly because the client is not allowed to alter the structure of the package and thus certain functions are not populated by the client–library interface. On the other hand, the designer–library interface exposes the entire inner structure of

the package both for reading and writing. It does, however, control the legitimacy of the operations.

The interface is constructed as follows. When the client or the designer connects to the library, they call a predefined method which returns the top of the interface tree. From there, the client or the designer can request other nodes by calling the `children()` method. In the end, there is an interface node between each of the client/designer nodes and the library tree nodes. Figure 3 shows the connection. The interface nodes are proxy objects that dispatch information from the client/designer to the library and back.

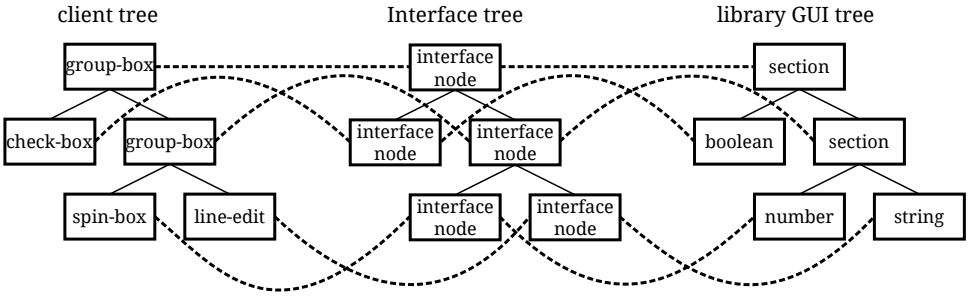


Figure 3: Example of the client–library interface connection.

4 Graphical Clients

There exists only one reference implementation of a client at the moment. It is the Qt/KDE graphical client which displays the configuration window in KDE style. It has been written in PyQt [4]. In Figure 4, one can see the current design of the constructed dialog. It resembles very closely the actual look&feel of the KDE configuration software. On the left side of the dialog, there are configuration tabs. The content of the selected tab is displayed on the right side. One can see a configuration section and some configuration keys there. Down the left, there is a *Show/Hide advanced* button which enables the user to hide or reveal additional expert configuration options that are hidden by default to simplify the configuration dialog (more about the simplification work can be found in [2]). The last three buttons are self-explanatory. *Ok* and *Apply* save the changes to the native configuration file and the first button also closes the dialog. The last button cancels the configuration process and leaves the current native configuration file intact.

When the user makes a change to the configuration, e.g., by checking an unchecked check-box or by filling in a text key, the client will propagate this change through the client–library interface to the corresponding node of the tree structure (in this example, to the configuration tree). The library will test the new value for validity (it is never done in the client except some of the GUI components can perform a simple filtering by themselves), checks the dependencies, and stores the new value. If a dependency changes any other key or its property, a message is sent to the correct client node for it to reload the data and adapt the visual aspect of the matching GUI element. The client–library interface does not allow anything else than loading a package, altering key values, and requesting to save and transform the current state of the configuration.

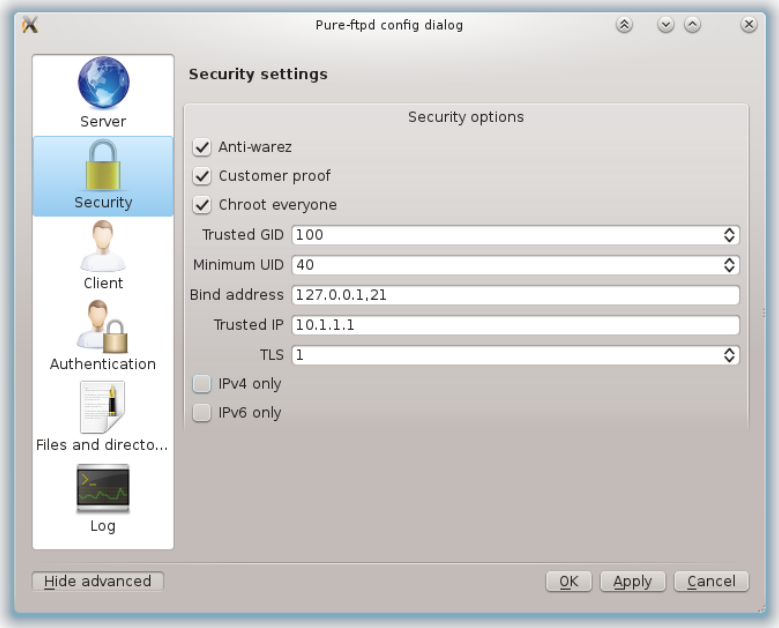


Figure 4: Example of a Freeconf generated configuration dialog.

5 Configuration Package

5.1 Content of Configuration Package

A configuration package is a collection of up to ten XML [8] (actually there can be more due to the optional presence of plug-ins) files which are organized into a directory tree. Many of the files are not mandatory, in fact, only four are necessary for Freeconf to be functional. These files can form up a package:

- *Header file* is an entry-point to the package. It has a fixed name `header.xml` and is placed in the root directory of the package. It contains the list of other components of the package as well as some of the package-level settings parameters (e.g., where to store native configuration files, or whether to store all of the keys or just those whose value has been changed from the default).

- *Template file* describes keys and their properties (like type, number boundaries, regular expressions for string keys etc. For a full set of key properties that Freeconf supports, see [1, 2]), configuration sections, and the entire structure of the configuration.

- *List file* contains definitions of string lists. These files can be stored in the directory `lists` or outside of the package to be shared between packages. For example, a list of encoding tables would be a good candidate for a shared list since it appears in multiple packages.

- *Default values file* holds default values for keys defined in the template. It is possible for some key to not have any default value set. It that case, the value will be undefined and the user might be requested to fill in the value during the first configuration.

- *Help file* contains key labels and tool-tips. The package usually contains more of these files, one for every translation. The files are stored in a directory named the same as the two-letter language code. These directories are placed into the `L10n` directory.

- *List help file* contains tool-tips translations for string-list values. The file is placed similarly to the help file, but the directory structure is itself placed in the `lists` directory.

- *GUI template file* describes the GUI window. It contains hints which the client can use to provide a better look&feel. If the file is not present in the package, the window is built using the information from the template file only.

- *GUI label file* contains translations of tab captions, the window title, and other strings used in the dialog. It is placed in the same directory as the help file.

- *Output file* holds the last saved state of the configuration. The syntax of the file is slightly different from the default file since it can hold more information needed for the transformation process.

- *Transformation file* is a XSL [7] file which is used during the transformation.

From all the files, only the header file, the template file, the transformation file, and the output file are mandatory. Other files are purely optional, though recommended, since the additional information increases the usability of the resulting GUI substantially. The entire package structure is depicted in Figure 5.

The package can be stored at three different places on the file system — two of them are system directories (on GNU/Linux it is `/usr/local/share/freeconf/packages` and `/usr/share/freeconf/packages`) and the last is the user's home directory (on GNU/Linux `~/.freeconf/packages`).

5.2 Transformation

The transformation process is controlled by the transformation file. The Freeconf library assembles the output file in the Freeconf format, loads the transformation file and submits both to the XSL processor (which is a standard `libxslt` library). The expressive power of XSL style-sheets enables to support virtually any text file format.

Should the native output be divided to multiple files, the header file enables to prescribe more XSL style-sheets and native output paths. Each key is then marked by the output group in which it belongs in the template file, so XSL knows where to place it.

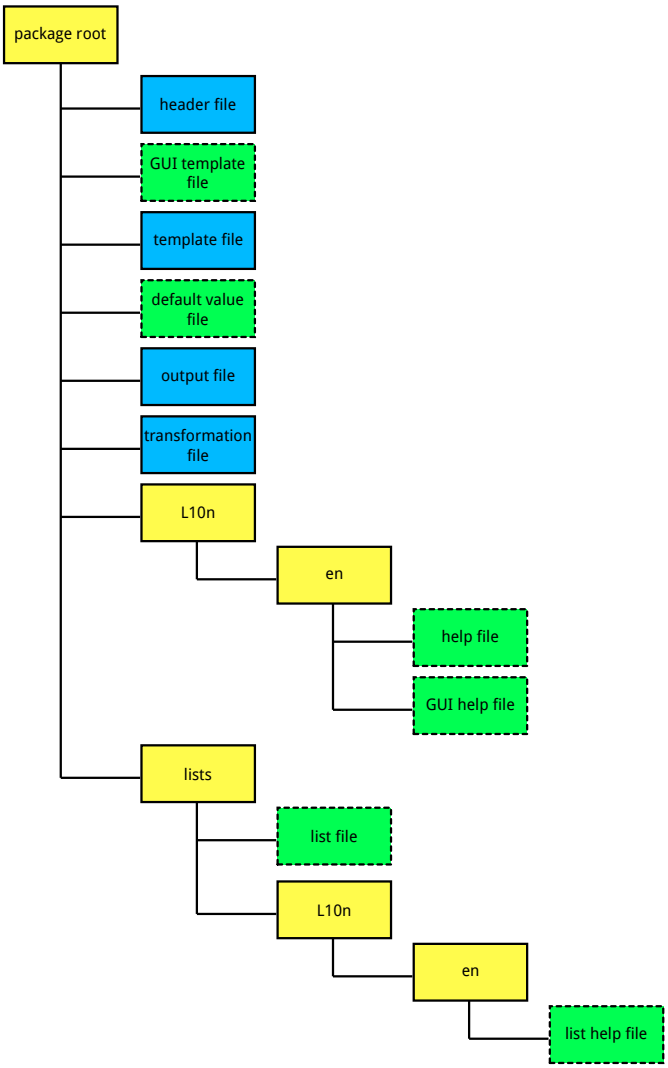


Figure 5: Package file hierarchy. The blue color represents mandatory files, the green boxes with dashed borders represent optional files, and finally the yellow boxes represent directories.

6 Package Designer

Since creating a configuration package manually is a tedious and error prone task, a simple application for package designing has been created. The application can, at the moment, assist only during the creation of the core of the package (the header, template, help, and default values file). Other files must still be written by hand. The current look of the program can be seen in Figure 6.

On the left side, there is the template tree with all of the sections and keys. On the right side, there are widgets for configuring properties of the selected key.

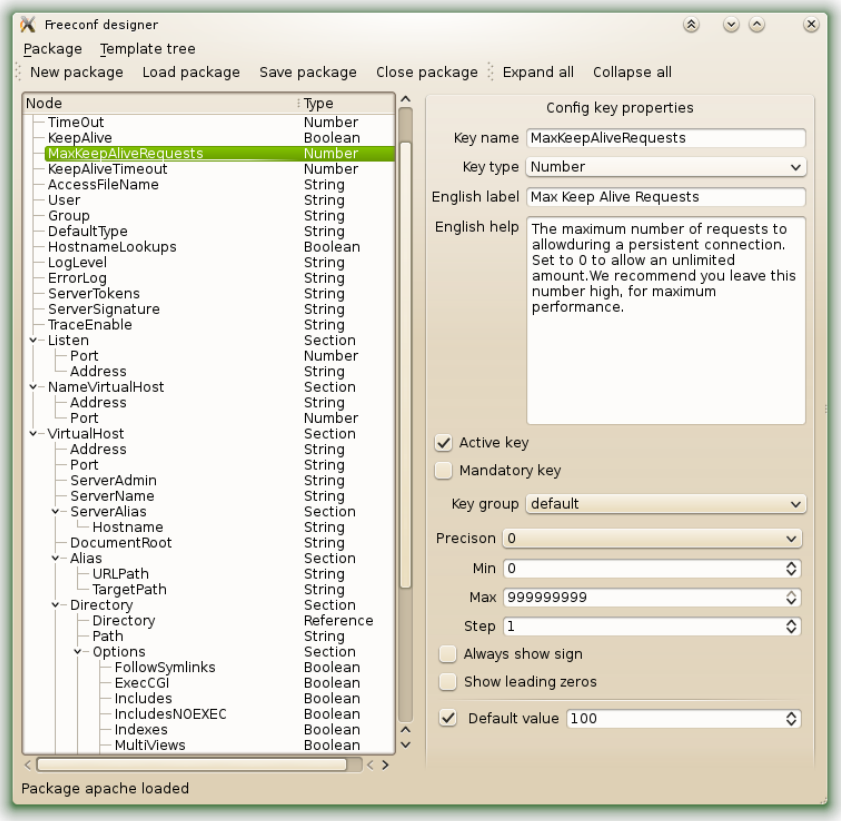


Figure 6: The current look of the package designer.

7 Conclusion

In this article, a general-purpose multi-platform configuration utility Freeconf has been presented. The tool consists of four parts, namely the Freeconf library, the Freeconf clients, the Freeconf packages, and the Freeconf designer. These parts have been studied in grater detail in their respective sections.

In the future, more clients are to be programmed and the Freeconf designer is planned to be able to generate the entire configuration package.

References

[1] D. Fabian. *System for Simplified Generating of Configurations*. Master thesis, Faculty of Nuclear Sciences and Physical Engineering, Prague, (2011). in Czech.

[2] D. Fabian, R. Mařík, and T. Oberhuber. *Toward a Formalism of Configuration Properties Propagation*. In 'Proceedings of the Workshop on Configuration at ECAI 2012 (ConfWS'12)', ECAI2012, 15–20, (2012).

[3] Novell, Inc. *YaST Documentation*, (207). `http://doc.opensuse.org/projects/YaST/SLES11/onefile/yast-onefile.html`.

[4] Riverbank Computing Limited. *PyQt Official Web Page*, (2010). http://www.riverbankcomputing.co.uk/software/pyqt/intro.

[5] K. TechBase. *Using KConfig XT*. http://techbase.kde.org/Development/Tutorials/Using_KConfig_XT, (2012).

[6] S. Vermeulen. *Linux Sea*. http://swift.siphos.be/linux_sea, (2012).

[7] W3C. *XSL Language Documentation*, (2006). http://www.w3.org/TR/xsl/.

[8] W3C. *XML Language Documentation*, (2008). http://www.w3.org/TR/xml/.

Progressive Approaches to Localization and Identification of AE Sources*

Zuzana Farová†

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

advisor: Zdeněk Převorovský[1], Václav Kůs[2],
[1]Institute of Thermomechanics, AS CR,
[2]Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Reliable identification and classification of already localized AE sources is one of the most important and also most difficult problems in AE monitoring. In this paper we suggest new concept of more precise AE source localization and identification in complex structures. The method is based on a Time Reversal (TR) AE signal processing. The theory of TR acoustic is based on the fact that acoustic wave equation in non-dissipative heterogeneous medium is invariant with respect to TR operation. AE signals, recorded by transducers relatively far from the source can be generally considered as a multiple convolution of the source function with the Green's (wave transfer) function and transfer function of sensors along with signal processing devices. Let us consider some point source function $s(t)$ at the position r_0 and a receiver at position r_i. The signal $s_G(t)$ detected at r_i at the time $t \in [0,T]$ arises from the two above mentioned convolutions.

$$s_G = s(t) * G(t, r_0, r_i) * P_i(t), \quad t \in [0,T]\,.$$

The measured signal is then time reversed and rebroadcast from the position r_i to r_0. At the position r_0 we receive resulting TR signal, which is expressed as multiple convolution

$$s_{TR} = s(T-t) * G(T-t, r_0, r_i) * P_i(T-t) * G(t, r_i, r_0) * P_i(t) \quad t \in [0,T]\,.$$

Relation between the signal s_{TR} and the source function $s(t)$ is better understood in frequency domain. With the Fourier transform we convert the signal s_{TR} into $\mathcal{F}(s_{TR})$. Assuming the Green's function in the following form,

$$G(t, r_i, r_0) = \frac{1}{4\pi c^2} \frac{t - \|r_i - r_0\|/c}{\|r_i - r_0\|}\,,$$

we obtain, after some computations, Fourier transform $\mathcal{F}(s_{TR})$ in the form

$$\mathcal{F}(s_{TR}(t)) = \mathcal{F}(s(T-t)) \frac{1}{16\|r_i - r_0\|^2} e^{2i\omega T} \to IFT \to \frac{1}{16\|r_i - r_0\|^2} s(t) = a s(t)\,,$$

*This work is published in CD-Proceedings of 30th European Conference on Acoustic Emission Testing and 7th International Conference on Acoustic Emission, ISBN13:978-84-615-9941-7 and it has been presented as a keynote lecture at the conference EWGAE 2012 in Granada by Ing. Zdeněk Převorovský, CSc.

†This work has been done with co-authors Zdeněk Převorovský, Václav Kůs, Milan Chlada, and Josef Krofta

where IFT denotes inverse Fourier transform and a is a constant of proportionality. So it can be seen that the resulting signal is proportional to the original emitted by the source. This fact is very important for AE source location and further analysis. In the standard AE measurements we mostly analyze signals, which are not directly corresponding to originally emitted waves, but are influenced by traveling through the structure and by characteristics of AE sensors and recording devices and that influence makes classification of AE sources more difficult.

The new (offline) procedures, called "TR AE signal Deconvolution (TRAED)" enable effective solution of both inverse problems dealing with precise source location and partial reconstruction of the source function. The robustness of experimental TR originates in the wave character of the problem and its space-time reciprocity. Recorded AE signals are time-reversed and rebroadcast back to the original source location, where they are detected e.g. by scanning laser vibrometer. Summation of TR signals from more AE transducers substantially enhances signal to noise ratio. Experimental results obtained by TR procedure applied to artificial AE sources on a massive steel plate are discussed in the paper. Realized "deconvolution" shows a high effectiveness of the suggested TR procedure, which does not require any knowledge on elastic wave modes and their propagation velocities and on Green's function of a structure with complex geometry. Any huge computations or numerical simulations are also not necessary. TRAED allows easier and more reliable determination of the sources location (up to 1 mm) and their more reliable identification and statistical classification. Although the theoretical description of all TR effects is relatively complicated and not yet completely formulated, the exploitation of the TR principles can bring to the AE method new possibilities of AE source characterizing and understanding.

Keywords: acoustic emission, source location and identification, time reversal acoustics, signal deconvolution, inverse problem solution

Abstrakt. Spolehlivá klasifikace a identifikace lokalizovaného zdroje akustické emise je jedním z nejdůležitějších, ale také jedním z nejobtížnejších problémů v oblasti akustické emise. V článku navrhujeme novou přesnější metodu lokalizace a identifikace zdrojů AE ve složitých strukturách. Metoda je založená na časově reverzní akustice (Time Reversal Acoustic TRA). Teorie TRA se opírá o fakt, že vlnová rovnice v nedisipativním heterogenním médiu je invariantní vzhledem k časové reverzaci. Signály AE zaznamenané na snímačích ve velké vzdálenosti od zdroje mohou být obecně považovány za výsledek konvoluce zdrojové funkce s Greenovou funkcí a s přenosovou funkcí snímače.
Uvažujme libovolnou zdrojovou funkci $s(t)$ v místě r_0 a snímač v místě r_i. Signál $s_G(t)$ zaznamenaný v místě r_i v čase $t \in [0,T]$ bude výsledkem dvou výše zmíněných konvolucí.

$$s_G = s(t) * G(t, r_0, r_i) * P_i(t), \quad t \in [0,T]\,.$$

Naměřený signál je poté časově obrácen a vyslán zpět z místa r_i do r_0. V místě r_0 pak měříme výsledný časově reverzní signál, který lze vyjádřit opět jako násobnou konvoluci

$$s_{TR} = s(T-t) * G(T-t, r_0, r_i) * P_i(T-t) * G(t, r_i, r_0) * P_i(t) \quad t \in [0,T]\,.$$

Vztah mezi signálem s_{TR} a zdrojovou funkcí $s(t)$ je vhodnější zkoumat ve frekvenční oblasti. Pomocí Fourierovy transformace převedeme signál s_{TR} na spektrum $\mathcal{F}(s_{TR})$. Za předpokladu, že Greenova fuknce má standardní tvar

$$G(t, r_i, r_0) = \frac{1}{4\pi c^2} \frac{t - \|r_i - r_0\|/c}{\|r_i - r_0\|}\,,$$

dostaneme po několika úpravách Fourierovu transformaci $\mathcal{F}(s_{TR})$ v následujícím tvaru

$$\mathcal{F}(s_{TR}(t)) = \mathcal{F}(s(T-t)) \frac{1}{16\|r_i - r_0\|^2} e^{2i\omega T} \rightarrow IFT \rightarrow \frac{1}{16\|r_i - r_0\|^2} s(t) = a s(t)\,,$$

kde IFT označuje inverzní Fourierovu transformaci a a je konstanta proporcionality. Můžeme tedy vidět, že výsledný signál je proporcionální originálnímu signálu vyslaného zdrojem. Tato skutečnost je velmi důležitá pro lokalizaci zdroje AE a další analýzu. Při standardních AE měřeních většinou měříme signály, které často nesouvisí přímo s původně vyslanou vlnou, ale jsou ovlivněny průchodem vlny skrz materiál a charakteristikami AE snímačů a kvůli těmto vlivům je klasifikace zdrojů AE značně obtížná.

Offline metody nazvané jako "TR AE signal Deconvolution (TRAED)" umožnují efektivně najít řešení inverzního problému spolu s přesnou lokalizací a také umožnují částečnou rekonstrukci původní zdrojové funkce. V článku popisujeme též experimentální výsledky získané pomocí TR metody aplikované na umělé zdroje AE na železné desce. Provedená "dekonvoluce" ukazuje vysokou efektivnost navržené TR metody, pro kterou nejsou zapotřebí žádné znalosti módů elastických vln, jejich rychlostí ani znalost Greenovy funkce pro daný vzorek. Rovněž není potřeba žádných numerických simulací a výpočtů. TRAED umožnuje snadnější a přesnější lokalizaci zdroje (až do 1mm) a rovněž přesnější idetifikaci a statistickou klasifikace zdrojů. Ačkoli teorie TR je relativně komplikovaná a stále ještě není kompletně formulována, rozvoj TR principů přináší do metody AE nové možnosti jak charakterizovat a porozumnět zdrojům AE.

Klíčová slova: akustická emise, lokalizace a idetifikace zdroje, časově reverzní akustika, dekonvoluce signálů, řešení inverzního problému

References

[1] B. E. Anderson, M. Griffa, C. Larmat, T. J. Ulrich, P. A. Johnson. *Time Reversal.* Volume 4 (1) of *Acoustics Today* (2008), 4–15.

[2] B. E. Anderson, M. Griffa, C. Larmat, T. J. Ulrich, P. A. Johnson. *Time reversal reconstruction of finite sized sources in elastic media.* Volume 130 (4) of *JASA Express Letters* (2011), 219–225.

[3] C. Bardos, M. Fink. *Mathematical foundations of the time reversal mirror.* Vol 29 (2) of *Asymptot. Anal.* 2002, 157–182.

[4] M. Blahacek, Z. Prevorovsky, J. Krofta, M. Chlada. *Neural network localization of noisy AE events in dispersive media.* Volume 18 (1) of *J. of Acoustic Emission* (2000), 279–285.

[5] M. Blahacek, Z. Prevorovsky. *Advanced AE source location in complex aircraft structures.* Volume 27(1) of *Journal of Acoustic Emission* (2009), 172–177.

[6] M. Blahacek, M. Chlada, Z. Prevorovsky, *Acoustic emission source location based on signal features.* 27th European Conference on AE Testing, EWGAE 2006, Cardiff, UK, volume 13-14 of *Advanced Materials Research* (2006), 77–82.

[7] A. Carpinteri, G. Lacidogna (eds). *Acoustic Emission and Critical Phenomena: From Structural Mechanics to Geophysics.* CRC Press, Taylor&Francis Group (2008).

[8] Z. Farova, Z. Prevorovsky, V. Kus, S. Dos Santos, *Experimental Signal Deconvolution in Acoustic Emission Identification Setup.* Proc. of the 6th Internat. Workshop NDT in Progress, Prague (2011), 33–40.

[9] M. Fink, C. Prada, F. Wu, D. Cassereau. *Self focusing in inhomogeneous media with time reversal acoustic mirrors.* IEEE Ultras. Symp. Proc. 1 (1989), 681–686.

[10] M. Fink. *Time reversal of ultrasonic fields. Part I: Basic principles.* Volume 39 (5) of *IEEE Trans. Ultr. Ferr. Freq. Contr.* (1992), 555-566.

[11] M. Fink. *Time-reversed acoustics*, volume 63 of *Rep. Prog. Phys.* (2000), 1933–1995.

[12] Ch. U Grosse , M. Ohtsu (eds). *Acoustic Emission Testing. Basic for Research - Application in Civil Engineering.* Spriger-verlag, (2008).

[13] M. Chlada, Z. Prevorovsky. *Expert AE Signal Arrival Detection.* Volume 6 (3/4) of *Int, Journal of Material & Product Technology* (2011), 191–205.

[14] M. Chlada, Z. Prevorovsky, M. Blahacek. *Neural network AE source location apart from structure size and material.* 29th EWGAE 2010, Vienna, Volume 28 of *J. of Acoustic Emission* (2010), 99–108.

[15] M. Chlada, Z. Prevorovsky. *AE source recognition. by neural networks with optimized signal parameters*, Volume 27(1) of *Journal of Acoustic Emission* (2009), 250–255.

[16] M. V. Klibanov, A. Timonov. *On the mathematical treatment of time reversal.* Institute of Physics Publishing. Volume 19 of *Inverse Problems* (2003), 1299–1318.

[17] V. Kus, M. Zavesky, Z. Prevorovsky. *Acoustic Emission Defects Localized by Means of Geodetic Iterative Procedure - Algorithms, Tests, AE Experiment.* 30th EWGAE / 7th ICAE, Granada (2012).

[18] G. Muravin, B. Muravin, D. Beilin. *Application of Quantitative Acoustic Emission Method for Non-Destructive Inspection of Metal and Reinforced Concrete Structures - New Opportunities and Prospects.* Volume 13 of *Scientific Israel*, Spec. Issue, (2011), (2–3).

[19] D. Ozevin, Z. Heidary. *Acoustic Emisssion Source Orientation Based on Time Scale.* Volume 29 of *J. Acoustic Emission* (2011), 123–132.

[20] H. W. Park, H. Sohn, K. H. Law, C. R. Farrar. *Time reversal active sensing for health monitoring of a composite plate.* Volume 302 of *J. Sound Vib.* (2007), 50–66.

[21] J.-M. Parot. *Localizing impulse sources in an open space by time reversal with very few transducers.* Volume 69 (4) of *Applied Acoustics* (2008), 311–324 .

[22] Z. Prevorovsky. *Notes on wave and waveguide concepts in AE*, 25th EWGAE 2002, Prague, Proc. Vol II (2002), 83 – 90.

[23] Z. Prevorovsky, M. Chlada, J. Vodicka. *Inverse Problem Solution in Acoustic Emission Source Analysis - Classical and Artificial Neural Network Approach.* In P.P. Delsanto, ed.: 'The universality of Nonclassical Nonlinearity with Applications to Non-destructive Evaluation and Ultrasonics', SPRINGER - Kluwer Academic Publishers, New York , Heidelberg (2007), 515–530.

[24] Z. Prevorovsky, M. Chlada, M. Blahacek, T. Pour. *Ultrasonic Signal Transfer in Thin Extended Aircraft Parts - Experiments and Modeling*. Proc. of the 3rd Internat. Workshop "NDT in Progress", Prague (2005), 332–339.

[25] Z. Prevorovsky, J. Krofta, Z. Farova, M. Chlada. *Structural Health Monitoring in aerospace and civil engineering supported with two ultrasonic NDT methods - AE and NEWS*. Proc. of the 6th Internat. Workshop NDT in Progress, Prague (2011), 237–245.

[26] S. Vejvodova, Z. Prevorovsky, S. Dos Santos. *Nonlinear Time Reversal Tomography of Structural Defects*. 14th Internat. Conf. on Nonlinear Elasticity in Materials, XIV ICNEM, Lisbon, Volume 3 of *ASA POMA*, Issue 1 (2009), 045003–10.

Borders Scanning Algorithm for Solving Total Least Trimmed Squares Estimation*

Jiří Franc

3rd 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, CU in Prague

Abstract. The Total Least Squares is one of the most widely use method for data analysis, where both dependent and independent variables are observed with random errors. If data set contains outliers, the robustified version of TLS such as mixed Least Trimmed Squares - Total Least Trimmed Squares method is used. The disadvantage of this method is absence of exact algorithm that can find solution of the estimation in real time for data sets with large number of observations. In this paper we introduced the BSA algorithm for LTS-TLTS and compare it with BAB algorithm. It is the first introduction to this method for such a problem and only first results from simulation study are shown.

Keywords: robust regression analysis, error in variables model, robustified total least squares, total least trimmed squares, BSA - borders scanning algorithm, branch-and-bound algorithm

Abstrakt. Analýza dat pomocí nejmenších totálních čtverců je jednou z nejpoužívanějších metod pro případy kdy závislé i nezávislé proměnné obsahují chyby měření. Pokud jsou navíc data zatížena odlehlými a vlivnými pozorováními, která mohou odhad úplně zničit, je žádoucí použít robustní přístup. Metoda smíšených nejmenších usekaných čtverců - totálních nejmenších usekaných čtverců je jednou z možností jak se s tímto problémem vypořádat. Nevýhoda této metody spočívá v neexistenci algoritmu, který by dokázal nalézt řešení v reálném čase i pro početné datové soubory. V tomto článku ukážeme použití exaktního algoritmu BSA a jeho srovnání s algoritmem BAB. Jedná se o první seznámení s danou metodou pro tento problém a výsledky algoritmů jsou ukázány na simulovaných a na několika reálných datech.

Klíčová slova: robustí regresní analýza, metoda robustifikovaných totálních čtverců, metoda totálních nejmenších usekaných čtverců, algoritmus BSA, branch-and-bound algoritmus

1 Introduction

The ordinary Total Least Squares (TLS) method is one of several linear parameter estimation techniques that is designed to solve an 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), $\beta^0 \in \mathbb{R}^{p\times 1}$ unknow parameter vector and we have more

*This work has been supported by the Czech CTU grant SGS12/197/OHK4/3T/14 and the MSMT grant INGO II INFRA LG12020

equations than unknows, i.e. $n > p$. In this paper we will assume that $\mathbf{X}$ has full rank.

The idea of the TLS method, to solve mentioned optimization problem, consists in modification of all data points in such a way, that some norm of the modification is minimized subject to the constraint that the modified vectors satisfy a linear relation. The TLS method has a long history in the statistical literature, where the method is known as "errors-in-variables" model or "orthogonal regression", but the progress in computation and application of this methods came in last decades due to work of Golub and Van Loan [4] and Van Huffel and Vandewalle [10] among others.

The TLS method looks for the minimal corrections on the given data $\mathbf{D} = [\mathbf{Y}, \mathbf{X}]$ and the approximation is obtained as a solution of the following optimization problem

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

$\hat{\beta}^{(TLS,n)}$ is called a TLS solution to the problem (1) and $[\varepsilon, \mathbf{\Theta}]$ is called the corresponding TLS correction (residuals, errors). The suitable norm used in previous definitions of the TLS problem is Frobenius norm. The basic algorithm used to solve the problem is based on the singular value decomposition (see [4]) and its generalization together with the discussion when the classical solution exists is described in [10]. In our case, when $\mathbf{X}$ has full rank, $n > p$ and errors are rowwise *iid*, it can be shown that the probability of absence of solution tending to 0 with increasing number of observations. For real data sets is the situation, when the solution does not exists very unlikely.

Let us mentioned, that the TLS problem is equivalent to computing the hyperplane that minimizes the sum of the squared orthogonal distances from the data points to the fitting hyperplane. Then the definition of TLS problem can be formulated as

$$\hat{\beta}^{(TLS,n)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{1 + \|\beta\|^2} \sum_{i=1}^{n} |Y_i - X_i\beta|^2 = \arg\min_{\beta \in \mathbb{R}^p} \frac{\|Y_i - X_i\beta\|}{\sqrt{1 + \|\beta\|^2}}.$$

If the linear modeling problem $\mathbf{Y} \approx \mathbf{X}\beta$ contains the 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 TLS approach is called Mixed Least Squares - Total Least Squares problem (LS-TLS). Let us denote

$$\begin{array}{rll} \text{partition} & \mathbf{X} = \left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right] & \mathbf{X}^{(1)} \in \mathbb{R}^{n \times p_1},\ \mathbf{X}^{(2)} \in \mathbb{R}^{n \times p_2} \\ & \beta^T = \left[{\beta^{(1)}}^T, {\beta^{(2)}}^T\right] & \beta^{(1)} \in \mathbb{R}^{p_1},\ \beta^{(2)} \in \mathbb{R}^{p_2} \end{array}$$

and assume that the columns of $\mathbf{X}^{(1)}$ are error free and $p_1 + p_2 = p$. Then the mixed LS-TLS estimator is defined by

$$\begin{array}{rl} \hat{\beta}^{(LS-TLS,n)} = & \min\limits_{\beta \in \mathbb{R}^p, [\varepsilon, \Theta] \in \mathbb{R}^{n \times (p_2+1)}} \|[\varepsilon, \Theta]\|_F \\ & \text{subject to} \quad \mathbf{Y} + \varepsilon = \mathbf{X}^{(1)}\beta^{(1)} + (\mathbf{X}^{(2)} + \Theta)\beta^{(2)}. \end{array}$$

By varying p_1 from zero to p, the mixed LS-TLS problem can handle also with any ordinary LS or ordinary TLS problem. Since the LS-TLS estimator is very sensitive and can give misleading results when outliers in the dataset occur, other more robust estimator is introduced.

2 TLTS - Total Least Trimmed Squares

The robustification of mixed LS-TLS were firstly introduced in [3] and the idea is based on trimming or downweighting high influential points. Let us denote by q_i the sum of the squared orthogonal distance of i-th observation from the hyperplane represented by $\beta^{(2)}$ and the squared vertical distance of ith observation from the hyperplane represented by $\beta^{(1)}$. Then the Mixed Least Trimmed Squares - Total Least Trimmed Squares (LTS-TLTS) estimator minimizes the sum of the h smallest distances q_i

$$\hat{\beta}^{(LTS-TLTS,n)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h} q_{(i)}(\beta),$$

where h is an optional parameter satisfying $\frac{n}{2} \leq h \leq n$ and $q_{(i)}$ is the i-th least mixed distance q_i at β.

LTS-TLTS is so called half-sample estimator and it has 50% breakdown points (for proof see [2]). It has the infinite local sensitivity, which can be improved by adding some continuous weighting function and multiply the distances by a weights from $\langle 0, 1\rangle$ (for definition see [3]). The existence of LTS-TLTS is given by the existence of LS-TLS for subsamples of size h. The exact algorithm based on evaluation of all $\binom{n}{h}$ computations of LS-TLS works in practice only if the number of observations is less than 20. In [3] we proposed the the non-exhaustive exact branch-and-bound (BAB) algorithm that can be used if the number of observations is less than 60. The algorithm is inspired by branch-and-bound algorithm for Least Trimmed Squares (LTS) problem presented by José Agulló [1] and guarantees global optimality. The algorithm passes through the tree with h levels, $(n-h+1)$ roots and $\binom{n}{h}$ terminal nodes and cut given branches. For data sets with more observations and unknowns it is better to use the approximative algorithms that are very fast and give sufficiently good results.

For larger data sets with more observations and unknowns it is necessary to use some approximating algorithm. One of the most popular resampling algorithm for LTS-TLTS is based on the idea of PROGRESS algorithm for LTS proposed by Rousseeuw and Leroy [8] and improved into FAST-LTS algorithm by Rousseeuw and Van Driessen in [9]. The algorithm usually finds a local minimum which is close to the global minimum, but not necessarily equal to that global minimum. In spite of the algorithm gives reasonable estimations and is very fast, Hawkins and Olive [5] proved that elemental concentration algorithms are zero breakdown and that elemental basic resampling estimators are zero breakdown and inconsistent. In this paper we introduced another exact algorithm called Borders Scanning Algorithm (BSA).

3 BSA - Borders Scanning Algorithm

The BSA algorithm was firstly introduced for LTS by Karel Klouda in his master thesis [6] and the detailed description of this algorithm can be also find in [7]. Firstly we describe the algorithm for TLTS. The idea of this algorithm is in scanning of the objective function (cost function) of TLTS, which is continuous, nonconvex, non-differentiable and has multiple local minima, whose number commonly rises with the number of observations and unknowns. The plot of an example of this function is on following Figure 1:

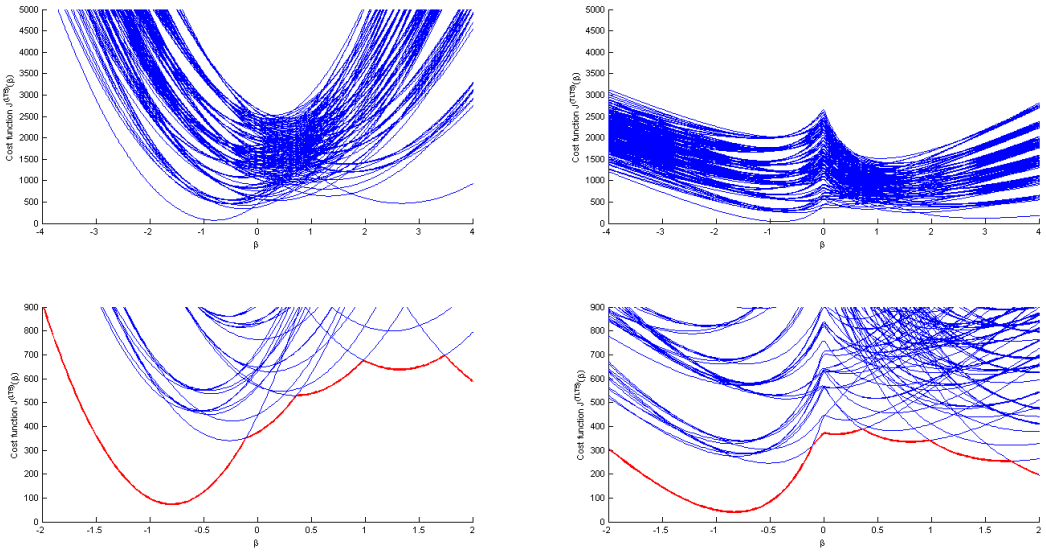


Figure 1: The graph of optional function (red bold line) for LTS and TLTS estimation on data with $n = 10$ observations, $p = 1$ and trimming parameter $h = 6$.

The objective function of LTS is composed from parts of quadratic function

$$J^{(TLTS,n,h)}(\beta) = \sum_{i=1}^{h} r^2_{(i)}(\beta),$$

where $r_j(\beta) = Y_j - X_j^T\beta$ and $r^2_{(1)}(\beta) \leq r^2_{(2)}(\beta) \leq \ldots \leq r^2_{(n)}(\beta)$, while the value of the TLTS objective function, denoted by $J^{(TLTS,n,h)}$, is defined for given parameter β as

$$J^{(TLTS,n,h)}(\beta) = \sum_{i=1}^{h} d^2_{(i)}(\beta),$$

where

$$d_j(\beta) = \frac{\left|Y_j - X_j^T\beta\right|}{\|[-1, \beta^T]\|} \quad \text{and} \quad d^2_{(1)}(\beta) \leq d^2_{(2)}(\beta) \leq \ldots \leq d^2_{(n)}(\beta).$$

The idea of the algorithm is to find all compositions of the objective function, in given part find the local minimum and the global minimum must be in the set of all local

minima. In accordance with [7] let us denote

$$\mathcal{H} = \left\{ \beta \in R \,|\, d^2_{(h)}(\beta) = d^2_{(h+1)}(\beta) \right\}.$$

We are looking for a set containing such a β's that given a hyperplanes which divide into halves the distance between the h-th and $h+1$-th most distant points from a given hyperplane.

Let us denote the set of weighting vectors w's with components from $\{0,1\}$ as

$$Q^{n,h} = \left\{ w \in R \,|\, w^i \in \{0,1\}, i = 1,2,\ldots,n \text{ and } \sum_i w^i = h \right\}$$

and let us define a relation $Z \subset R^p \times Q^{n,h}$ by

$$(\beta, w) \in Z \Leftrightarrow \sum_{i=1}^{h} d^2_{(i)}(\beta) = \sum_{i=n}^{h} w^i d^2_i(\beta)$$

Further we define a set $\mathcal{U} \subset R^p$ as the set where Z is a mapping from R^p to $Q^{n,h}$. Since the set $\mathcal{U}$ is a complement to R^p of the set $\mathcal{H}$, it is obvious, that $\mathcal{H}$ decompose the parameter space R^p into m open subsets U_i, $i = 1,2,\ldots m$. Further it holds that $U_i \cap U_j = \emptyset$ for all i,j where $i \neq j$, $\cap_{i=1}^m U_i = \mathcal{U}$ and $\cap_{i=1}^m \partial U_i = \mathcal{H}$. Last notation to be introduced is the set $W^{(min)}$ which is defined as a set of m vectors from $Q^{n,h}$, i.e. $W^{(min)} = \{w_1, w_2, \ldots, w_m \,|\, w_i = Z(\beta), \text{ where } \beta \in U_i, i = 1,2,\ldots m\}$.

The most important remark is, that the set $\mathcal{H}$ is the same both for the cost function of TLTS and for the cost function of LTS. So we can follow the technique of finding some finite subsets H of candidates of being an element of the set $\mathcal{H}$. Since $H \subset \mathcal{H}$ we can evaluate squared distances $d_i(\beta)$ for all data points and all suspected $\beta \in H$, sort them and if $d^2_{(i)}(\beta) = d^2_{(i+1)}(\beta)$, then $\beta \in \mathcal{H}$ and we find q weights $w^{(1)}, \ldots, w^{(q)} \in Q^{n,h}$. For all weights we evaluate the cost function and the cost function of TLTS estimator is that one with the minimal value.

For $p = 1$ is the situation very easy and we have to check only $4\binom{n}{2}$ weights for suspected βs from the set $H = \{\beta \in R^1 \,|\, \beta(x_i \pm x_j) = (y_i \pm y_j), i \neq j\}$. For $p > 1$, the situation is more complicated and β is a solution of a system of q linear equations. More in [7]. The number of suspected $\beta \in H$ is then greater than $2^p\binom{n}{p+1}$. How is the algorithm fast in comparison with BAB algorithm will be shown in next section.

4 Simulation study and comparison of computation techniques

The test of the algorithm is carried out both on simulated data sets and on some real data benchmarks. Algorithms are written and performed in MATLAB software, mentioned time is measured by the function "cputime", and it express time in seconds that has been used by the MATLAB process.

At first we run several simulations for data sets with intercept and varying number of observation n and number of regression parameters p. The simulation is repeated for each setting and resulting mentioned time is sample mean of all results. We compare time consumption for two different algorithms: BSA and BAB. BAB algorithm use the initial estimation from resampling algorithm (RES) with 1000 starting points and starting level of (BAB) algorithm is chosen to $h/4$. Simulations with $h = 0.8n$ present an example where is 20% occurrence of outliers and simulations with $h = 0.6n$ presents an example with 40% of contamination. Regressors were generated from normal and exponential distributions. Errors are standard normal distributed.

Time consumption in second - median of 10 replications								
n	$p = 3$				$p = 4$			
	$h = 0.8n$		$h = 0.6n$		$h = 0.8n$		$h = 0.6n$	
	BSA	BAB	BSA	BAB	BSA	BAB	BSA	BAB
15	1.887	0.124	2.995	0.187	9.594	0.140		0.168
20	6.474	0.296	9.063	0.452	43.149	0.249	80.324	0.468
25	15.818	1.747	21.964	3.804	138.310	0.811	251.364	2.246
30	35.630	6.489	44.600	22.793	363.482	3.525	602.039	20.092
35	66.862	36.884	82.321	76.378	964.788	24.445	1306.263	114.052
40	112.975	126.627	138.435	251.876	1751.034	122.070	2467.034	622.967

Table 1: Simulation study for different LTS-TLTS estimators, data set with intercept, number of replications is 10 and n, h, p is varying.

As we can see from the previous Table 1 the computation time rapidly increase (nearly exponentially) with increasing number of observations for BAB algorithm. For BSA is more significant the increase in p, while the increase in n is nearly linear for smaller n. The speed of BAB is more dependent on number of observations. As we showed in [3] the BAB is unusable for $n > 60$. Another disadvantage of BAB algorithm is its large variability in time consumption. In simulation is common that for the same settings is one replication four time faster than another. BSA is in this point more stable and the deviation is not more than 20% from the mean value obtained from 10 replications.

It is very surprising that in comparison with simulation results for classical LTS problem presented in [6] the BSA algorithm for TLTS problem needs much more calculations and primarily the time consumption grow up much more faster in relation to number of regression parameters. We were not able to compute the estimation for $p > 6$ on normal PC. The theory of the BSA, number of minimum ordinary TLS calculations, and the estimation of number of corresponding $\beta \in H_p$ has not yet been examined for TLTS problem in detail.

Real data sets are from [8] and let us denote by "Stars" the Hertzsprung-Russell

Diagram of the Star Cluster CYG OB1, which contains 47 stars in the direction of Cygnus, by "Wood" the modified Wood Gravity Data with five independent variables and intercept. It consists of 20 cases and some of them were replaced to contaminate the data by few outliers. And finally by ""Brain" we denote Mammal brain weights data with 28 observations. The time consumption of both algorithms for these three real data sets is shown in the Table 2.

Data	n	p	h	time in seconds	
				BSA	BAB
Stars	47	2	0.8n	4.042	4.973
Wood	20	6	0.6n	235.546	0.187
Brain	28	1	0.8n	2.044	0.515

Table 2: Real data analysis by LTS-TLTS estimator and computational time needed for the evaluation of the estiamte

5 Conclusion

We have modified BSA algorithm for LTS estimator for use in modified LTS-TLTS problem and compare it with another non-exhaustive exact BAB algorithm. It has been the first attempt to use BSA for this type of problem and we have had to cope with lot of problems in programming and in running simulations. Some problems has not been solve and they are tasks of future work.

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

References

[1] J. Agulló. *New algorithms for computing the least trimmed squares regression estimator.* Computational Statistics and Data Analysis **36** (2001), 425—439.

[2] J. Franc. *Introduction to total least trimmed squares estimation.* Doktorandske Dny proceedings 2011, FJFI , Czech Republic (2011).

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

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

[5] 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.

[6] K. Klouda. *Algorithms for computing robust regression estimates, Master thesis.* Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, (2007).

[7] K. Klouda. *Bsa - exact algorithm computing lts estimate.* arXiv:1001.1297 (2010).

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

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

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

Konvergence diskrétních transformací fourierovského typu pro Lieovy algebry ranku 2*

Jan Fuksa

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

školitel: Severin Pošta, Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. Basic used objects are the orbit functions defined on $\mathbb{R}^2$. The orbit functions form an orthogonal basis in the Hilbert space of quadratic integrable functions and determine the orbit function transform on the fundamental region of the affine Weyl group. A discrete version of the orbit function transform is defined on an finite discrete grid in the fundamental region. Applications show that the discrete orbit function transform converges. This contribution puts a target to mathematically support this fact. We show that the discrete orbit function transform converges for $\mathcal{C}^{(6)}$ functions to the expanded function on the grids in the fundamental region with growing density.

Keywords: Weyl group, orbit functions, orbit function transform, discrete orbit function transform, convergence

Abstrakt. Základním používaným objektem jsou funkce na orbitách definované na $\mathbb{R}^2$. Funkce na orbitách tvoří ortogonální bázi v Hilbertově prostoru kvadraticky integrabilních funkcí a určují transformaci fourierovského typu na fundamentální oblasti afinní Weylovy grupy. Na konečné diskrétní mřížce ve fundamentální oblasti se zavádí diskrétní obdoba transformace fourierovského typu. Z aplikací je zřejmé, že diskrétní transformace fourierovského typu konverguje. Tento příspěvek si klade za cíl matematicky podložit tento fakt. Dokážeme, že pro funkce z $\mathcal{C}^{(6)}$ diskrétní transformace fourierovského typu konverguje k rozvíjené funkci na zahušťující se mříži ve fundamentální oblasti.

Klíčová slova: Weylova grupa, funkce na orbitách, transformace fourierovského typu, diskrétní transformace fourierovského typu, konvergence

1 Úvod

Tento příspěvek se zabývá vzájemným vztahem diskrétních a spojitých transformací fourierovského typu. Klíčovým pojmem jsou tzv. funkce na orbitách (*orbit functions*), které se zavádí pomocí Weylovy grupy. Weylovu grupu tvoří z algebraického hlediska množina zrcadlení podle nadploch v euklidovském prostoru $\mathbb{R}^n$. Funkce na orbitách jsou symetrické vůči afinní Weylově grupě. Afinní Weylova grupa vymezuje fundamentální oblast $F \subset \mathbb{R}^n$ tak, že každý bod $\mathbb{R}^n$ lze obdržet jako obraz série zrcadlení nějakého bodu uzávěru $\overline{F}$. Množina funkcí na orbitách tvoří ortogonální bázi Hilbertova prostoru kvadraticky integrabilních funkcí $\mathcal{L}^2(\overline{F})$ na uzávěru $\overline{F}$ fundamentální oblasti F. V prostoru $\mathcal{L}^2(\overline{F})$ lze

*Tato práce byla podpořena grantem SGS12/198/OHK4/3T/14

zavést transformaci fourierovského typu právě pomocí funkcí na orbitách (*orbit function transform*), viz [2]. Na diskrétní mřížce obsažené v $\overline{F}$ lze zavést aproximaci této transformace, která je analogií diskrétní fourierovské transformace. Vztah diskrétních a spojitých transformací fourierovského typu v $\mathbb{R}^2$ zavedených pomocí funkcí na orbitách je blíže popsán v článcích [2, 3, 4]. Z četných aplikací je zřejmé, že pro řadu funkcí konverguje diskrétní transformace ke spojité, dosud však nebyly nikým zveřejněny žádné podmínky konvergence. Cílem tohoto příspěvku je tedy dokázat, že pro jistou třídu funkcí diskrétní transformace ke spojité skutečně konverguje.

Poznamenejme ještě, že když v průběhu textu budeme zmiňovat klasickou fourierovskou transformaci, budeme tím myslet všem dobře známou Fourierovu transformaci, zatímco transformace fourierovského typu bude vždy znamenat *orbit function transform*.

2 Od Lieových algeber k funkcím na orbitách

V této práci se budeme zabývat pouze Lieovými algebrami ranku 2, jmenovitě to jsou algebry $A_1 \times A_1, A_2, C_2$ a G_2. Tyto, jak uvidíme, určují transformaci fourierovského typu.

Poloprosté Lieovy algebry ranku n jsou určeny svým kořenovým systémem $\Delta \subset \mathbb{R}^n$. Kořenový systém obsahuje bázi $\{\alpha_1, \ldots, \alpha_n\} \subset \Delta$. Prvky báze se nazývají prosté kořeny. Vztahy mezi prostými kořeny popisuje Cartanova matice

$$C_{ij} = \frac{\langle \alpha_i | \alpha_j \rangle}{\langle \alpha_j | \alpha_j \rangle} \quad \text{pro } i, j = 1, \ldots, n, \tag{1}$$

kde $\langle \ | \ \rangle$ je skalární součin na $\mathbb{R}^n$. Kořeny vždy nabývají nejvýše dvou různých délek, oblíbená konvence stanovuje pro delší kořeny $\langle \alpha | \alpha \rangle = 2$. Prvky Cartanovy matice jsou při této konvenci celočíselné.

Dále zavádíme bázi fundamentálních vah $\{\omega_1, \ldots, \omega_n\}$ vztahem

$$\frac{\langle \omega_i | \alpha_j \rangle}{\langle \alpha_j | \alpha_j \rangle} = \delta_{ij} \quad \text{pro } i, j = 1, \ldots, n. \tag{2}$$

Přechod mezi těmito dvěma neortogonálními bázemi $\mathbb{R}^n$ je zprostředkován Cartanovou maticí, platí $\alpha_i = C_{ij}\omega_j$.

Zavádíme tzv. kokořeny $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ vztahem $\hat{\alpha}_i = \frac{2\alpha_i}{\langle \alpha_i | \alpha_i \rangle}$ a tzv. kováhy $\hat{\omega}_1, \ldots, \hat{\omega}_n$ vztahem $\hat{\omega}_i = \frac{2\omega_i}{\langle \alpha_i | \alpha_i \rangle}$.

Každý kořen $\alpha \in \Delta$ určuje zrcadlení r_α v $\mathbb{R}^n$ podle k němu kolmé nadplochy vztahem

$$r_\alpha x = x - \frac{\langle x | \alpha \rangle}{\langle \alpha | \alpha \rangle} \quad \text{pro } x \in \mathbb{R}^n. \tag{3}$$

Systém takovýchto zrcadlení se uzavírá do Weylovy grupy W.

Nejdelší kořen kořenového systému označme jako ξ. Zavádíme afinní zobrazení

$$R_\xi = \xi + r_\xi x. \tag{4}$$

Přidáním R_ξ k prvkům Weylovy grupy obdržíme afinní Weylovu grupu W^{aff}. W^{aff} obsahuje abelovskou podgrupu translací ve směru jednotlivých kořenů systému Δ, označme ji

T. Afinní Weylova grupa je polopřímým součinem Weylovy grupy a grupy translací, tj. $W^{\mathrm{aff}} = W \ltimes T$.

Fundamentální oblast F je největší oblast v $\mathbb{R}^n$ taková, že dva libovolné, od sebe různé body uzávěru $\overline{F}$ nepatří do stejné třídy vzhledem k akci afinní Weylovy grupy W^{aff} na $\mathbb{R}^n$. Pro prosté Lieovy algebry je fundamentální oblast tvořena vnitřkem simplexu s vrcholy

$$\left\{0, \frac{1}{q_1}\hat{\omega}_1, \ldots, \frac{1}{q_n}\hat{\omega}_n\right\}, \tag{5}$$

kde $(q_1, \ldots, q_n)$ jsou souřadnice nejdelšího kořenu ξ v bázi $\alpha_1, \ldots, \alpha_n$. Platí $W^{\mathrm{aff}}\overline{F} = \mathbb{R}^n$.

Pro další účely zavádíme tzv. kořenovou mříž Q symbolickým vztahem

$$\mathbb{Z}\alpha_1 + \cdots + \mathbb{Z}\alpha_n, \tag{6}$$

kde $\alpha_1, \ldots, \alpha_n$ jsou prosté kořeny. Pro fundamentální váhy $\omega_1, \ldots, \omega_n$ zavádíme analogicky váhovou mříž P a kladnou váhovou mříž P^+

$$\mathbb{Z}\omega_1 + \cdots + \mathbb{Z}\omega_n \in P, \qquad \mathbb{Z}^{\geq 0}\omega_1 + \cdots + \mathbb{Z}^{\geq 0}\omega_n \in P^+. \tag{7}$$

Prvky P nazýváme váhy, prvky P^+ nazýváme dominantní váhy.

Množinu $W_\lambda \equiv W\lambda$ pro nějaké $\lambda \in P$ nazýváme orbita Weylovy grupy a značíme W_λ. V každé orbitě existuje právě jeden prvek, který náleží P^+, danou orbitu budeme označovat právě dominantním prvkem. Platí, že mohutnost $|W_\lambda|$ je nejvýše rovna mohutnosti $|W|$, přičemž ji ovšem vždy dělí.

Funkce na orbitách pro $\lambda \in P$ jsou definovány jako

$$\Phi_\lambda(x) = |\mathrm{Stab}_W(\lambda)| \sum_{\mu \in W_\lambda} e^{2\pi i \langle \mu | x \rangle}, \tag{8}$$

kde $|\mathrm{Stab}_W(\lambda)|$ je mohutnost stabilizátoru λ vzhledem k akci grupy W na $\mathbb{R}^n$.

Funkce na orbitách tvoří ortogonální množinu. Platí

$$\frac{1}{|\overline{F}|} \int_{\overline{F}} \Phi_\lambda \overline{\Phi_{\lambda'}} \mathrm{d}F = |W_\lambda| |\mathrm{Stab}_W(\lambda)|^2 \delta_{\lambda\lambda'}. \tag{9}$$

Množina funkcí na orbitách $\{\Phi_\lambda |\ \lambda \in P^+\}$ tvoří ortogonální bázi Hilbertova prostoru kvadraticky integrabilních funkcí $\mathcal{L}^2(\overline{F})$ na $\overline{F}$.

Funkci $f \in \mathcal{L}^2(\overline{F})$ lze rozložit do řady funkcí na orbitách

$$f(x) = \sum_{\lambda \in P^+} c_\lambda \Phi_\lambda(x), \tag{10}$$

kde

$$c_\lambda = |W_\lambda|^{-1} |W|^{-1} |\overline{F}|^{-1} \int_{\overline{F}} f(x) \overline{\Phi_\lambda(x)} \mathrm{d}F. \tag{11}$$

Tento rozklad budeme nazývat transformací fourierovského typu.

3 Diskrétní transformace fourierovského typu

Z praktických důvodů se omezíme na Lieovy algebry ranku 2. Nosnou množinou, na které zavádíme diskrétní podobu transformace (11), je mřížka

$$F_M = \left\{ \frac{s_1}{M}\hat{\omega}_1 + \frac{s_2}{M}\hat{\omega}_2 \middle| \; s_0 + q_1 s_1 + q_2 s_2 = M, \; s_0, s_1, s_2 \in \mathbb{Z}^{\geq 0} \right\} \quad \text{pro } M \in \mathbb{N}. \tag{12}$$

F_M je podmnožinou $\overline{F}$. q_1, q_2 jsou souřadnice nejdelšího kořenu ξ v bázi α_1, α_2.

Pro funkce f, g definované svými hodnotami v bodech $s \in F_M$ zavádíme hermitovskou formu vztahem

$$\langle f|g\rangle_M = \sum_{s \in F_M} k_s f(s)\overline{g(s)}. \tag{13}$$

Koeficienty k_s jsou kladná celá čísla závisející na konkrétní algebře, viz [3, 4].

Jistá podmnožina funkcí na orbitách zúžených na mřížku F_M tvoří opět ortogonální množinu vzhledem k hermitovské formě (13). Je jasné, že lineárně nezávislých funkcí na F_M může být nejvýše $|F_M|$, všechny další jsou opakováním předešlých. Takovouto množinou ortogonálních funkcí je

$$S_M = \left\{ \Phi_\lambda \middle| \; \lambda = a\omega_1 + b\omega_2, \; aq_2 + bq_1 \leq M \right\}. \tag{14}$$

Funkce z S_M definují obdobu transformace (11) na mřížce F_M. Funkci f definovanou na F_M lze rozvinout do funkcí na orbitách zúžených na F_M, konkrétně

$$f(s) = \sum_{\lambda \in S_M} d_\lambda^M \Phi_\lambda(s), \quad s \in F_M, \tag{15}$$

kde

$$d_\lambda^M = \frac{\langle f|\Phi_\lambda\rangle_M}{\langle \Phi_\lambda|\Phi_\lambda\rangle_M}. \tag{16}$$

Ve vztahu (15) můžeme diskrétní proměnnou s nahradit spojitou proměnnou x, potom toto spojité rozšíření funkce f označme jako Ψ_M. Ψ_M je funkce hladká na fundamentální oblasti F, dokonce na celém $\mathbb{R}^2$, navíc v bodech F_M nabývá stejných hodnot jako funkce f.

4 Odhad konvergence diskrétní transformace

Předpokládejme, že funkce $f \in \mathcal{L}^2(\overline{F})$. Na $f|_{F_M}$ použijme diskrétní transformaci (16). Spojité rozšíření Ψ_M je dobrou aproximací funkce f, praktické aplikace naznačují, že s rostoucím M se tato aproximace blíží původní funkci f. Vyvstává proto zajímavá otázka, pro jakou třídu funkcí na fundamentální oblasti F lze dokázat, že funkční posloupnost $\{\Psi_M\}_{M=1}^{\infty}$ konverguje k f?

Pokusme se poněkud naivně odhadovat bodový rozdíl

$$|f(x) - \Psi_M(x)| \tag{17}$$

pro nějaké libovolné $x \in F$ a $M \in \mathbb{N}$. Použitím základních vztahů obdržíme odhad

$$|f(x)-\Psi_M(x)| = \left|\sum_{\lambda\in P^+} c_\lambda\Phi_\lambda(x) - \sum_{\lambda\in S_M} d_\lambda^M\Phi_\lambda(x)\right| \le \sum_{\lambda\in S_M}\left|c_\lambda\Phi_\lambda(x)-d_\lambda^M\Phi_\lambda(x)\right| + \\ + \sum_{\substack{\lambda\in P^+\\ \lambda\notin S_M}}|c_\lambda\Phi_\lambda(x)| \le |W|\sum_{\lambda\in S_M}\left|c_\lambda-d_\lambda^M\right| + |W|\sum_{\substack{\lambda\in P^+\\ \lambda\notin S_M}}|c_\lambda|\,. \tag{18}$$

Potřebujeme tedy odhadnout jednotlivé členy $\left|c_\lambda - d_\lambda^M\right|$ pro $\lambda\in S_M$ a sumu $\sum_{\lambda\in P^+\setminus S_M}|c_\lambda|$.

Odhadujme pro $\lambda\in S_M$

$$\left|c_\lambda-d_\lambda^M\right| = \left|c_\lambda - \frac{1}{\langle\Phi_\lambda|\Phi_\lambda\rangle_M}\langle f|\Phi_\lambda\rangle_M\right| = \left|c_\lambda - \frac{1}{\langle\Phi_\lambda|\Phi_\lambda\rangle_M}\sum_{s\in F_M} k_s f(s)\overline{\Phi_\lambda}(s)\right| = \\ = \left|c_\lambda - \frac{1}{\langle\Phi_\lambda|\Phi_\lambda\rangle_M}\sum_{s\in F_M} k_s\sum_{\mu\in P^+} c_\mu\Phi_\mu(s)\overline{\Phi_\lambda}(s)\right| = \\ = \left|c_\lambda - \frac{1}{\langle\Phi_\lambda|\Phi_\lambda\rangle_M}\sum_{\mu\in P^+} c_\mu\sum_{s\in F_M} k_s\Phi_\mu(s)\overline{\Phi_\lambda}(s)\right| = \\ = \left|c_{a,b} - \frac{1}{\langle\Phi_{a,b}|\Phi_{a,b}\rangle_M}\sum_{c,d=0}^{\infty} c_{c,d}\sum_{s\in F_M} k_s\Phi_{c,d}(s)\overline{\Phi_{a,b}}(s)\right| = \\ = \left|c_{a,b} - \frac{1}{\langle\Phi_{a,b}|\Phi_{a,b}\rangle_M}\sum_{c,d=0}^{\infty} c_{c,d}\langle\Phi_{a,b}|\Phi_{a,b}\rangle_M\delta_{a,c(\mathrm{mod}M)}\delta_{b,d(\mathrm{mod}M)}\right| = \\ = \left|c_{a,b} - \sum_{c,d=0}^{\infty} c_{c,d}\delta_{a,c(\mathrm{mod}M)}\delta_{b,d(\mathrm{mod}M)}\right| = \left|\sum_{\substack{m,n=0\\ m+n>0}}^{\infty} c_{a+mM,b+nM}\right|. \tag{19}$$

V průběhu jsme přešli do souřadnic na P^+, konkrétně $\lambda = a\omega_1+b\omega_2$ a $\mu = c\omega_1+d\omega_2$.

Uvažované transformace fourierovského typu lze za jistých předpokladů, které uvedeme dále, převést na klasickou fourierovskou transformaci na $<0,\gamma_1>\times<0,\gamma_2>$, jejíž koeficienty lze snadno odhadnout. Nechť funkce $f\in\mathcal{C}^{(n)}(\mathbb{R}^2)$ je periodická v obou proměnných s periodou γ_1 resp. γ_2, potom její klasický fourierovský koeficient

$$f_{k,l} = \frac{1}{\gamma_1\gamma_2}\int_0^{\gamma_1}\int_0^{\gamma_2} f(x,y)e^{-2\pi i(\frac{kx}{\gamma_1}+\frac{ly}{\gamma_2})}\mathrm{d}x\mathrm{d}y, \tag{20}$$

lze odhadnout jako

$$|f_{k,l}| \le \frac{K_1}{k^r l^{n-r}}, \quad r\in\{0,1,\dots,n\}, \tag{21}$$

kde K_1 je kladná, pro naše úvahy nepodstatná konstanta.

Fundamentální oblast je pro obecnou prostou Lieovu algebru ranku 2 rovna $F = \{x\omega_1^\vee + y\omega_2^\vee | 0<x,y<1,\ q_1x+q_2y<1\}$, kde q_1,q_2 jsou souřadnice nejdelšího kořenu $\xi = q_1\alpha_1+q_2\alpha_2$. Koeficienty c_λ, $\lambda\in P^+$, $\lambda = a\omega_1+b\omega_2$, se podle (11) počítají jako

$$c_{a,b} = |W_\lambda|^{-1}|W|^{-1}|\overline{F}|^{-1}\int_0^{\frac{1}{q_1}}\mathrm{d}x\int_0^{\frac{1-q_1x}{q_2}}\mathrm{d}y\ f(x,y)\overline{\Phi_{a,b}(x,y)}. \tag{22}$$

Nyní za předpokladu, že $f \in \mathcal{C}^{(n)}(\mathbb{R}^2)$ je symetrická vůči W^{aff}, můžeme integrál (22) přepsat do podoby

$$c_{a,b} = N^{-1}|W_\lambda|^{-1}|W|^{-1}|\overline{F}|^{-1}\int_0^{\gamma_1} \mathrm{d}x \int_0^{\gamma_2} \mathrm{d}y\ f(x,y)\overline{\Phi_{a,b}(x,y)}, \tag{23}$$

kde $N|F| = \gamma_1\gamma_2$, tj. N je počet, kolikrát se fundamentální oblast F vejde do obdélníku $< 0, \gamma_1 > \times < 0, \gamma_2 >$. Integrační meze γ_1 a γ_2 se stanoví pro každou konkrétní Lieovu algebru zvlášť tak, aby se jednotlivé exponenty $e^{2\pi i\langle\mu|x\rangle}$, $\mu \in W_\lambda$, $\lambda = a\omega_1 + b\omega_2$, ve funkcích na orbitách $\Phi_{a,b}$ staly periodickými, viz (8). Koeficient $c_{a,b}$ se pak rovná součtu klasických fourierovských koeficientů.

$$\begin{aligned} c_{a,b} &= |\mathrm{Stab}_W(\lambda)||W_\lambda|^{-1}|W|^{-1}| \sum_{\substack{\mu\in W_\lambda \\ \mu=k\omega_1+l\omega_2}} N^{-1}|\overline{F}|^{-1}\int_0^{\gamma_1} \mathrm{d}x \int_0^{\gamma_2} \mathrm{d}y\ f(x,y)e^{-2\pi i\langle\mu|x\rangle} = \\ &= |\mathrm{Stab}_W(\lambda)||W_\lambda|^{-1}|W|^{-1}| \sum_{\substack{\mu\in W_\lambda \\ \mu=k\omega_1+l\omega_2}} f_{k,l}. \end{aligned} \tag{24}$$

Navíc platí, že funkce $f \in \mathcal{C}^{(n)}(\mathbb{R}^2)$, která je podle předpokladu symetrická vůči W^{aff}, se stane na $\mathbb{R}^2$ periodickou v x i y s periodou γ_1 resp. γ_2. Nyní můžeme použít odhad (21) a koeficient $c_{a,b}$ hrubě odhadnout. Za předpokladu, že $f \in \mathcal{C}^{(n)}(\mathbb{R}^2)$ je symetrická vůči W^{aff}, dostáváme

$$|c_{a,b}| \le |\mathrm{Stab}_W(\lambda)||W_\lambda|^{-1}|W|^{-1}| \sum_{\substack{\mu\in W_\lambda \\ \mu=k\omega_1+l\omega_2}} \frac{K_1}{k^r l^{n-r}}, \qquad r \in \{0,1,\ldots,n\}. \tag{25}$$

Protože navíc souřadnice k, l jsou lineárními kombinacemi souřadnic a, b, lze odhad upravit tak, aby byl závislý pouze na a, b, tj.

$$|c_{a,b}| \le \frac{K_2}{a^r b^{n-r}}, \quad r \in \{0,1,\ldots,n\}. \tag{26}$$

Za jednoduchých předpokladů symetrie funkce f vůči W^{aff} a jistého stupně spojité diferencovatelnosti dostáváme velice pěkný odhad koeficientů transformace fourierovského typu. Pro funkci f není dokonce ani nutné předpokládat, aby byla n–krát spojitě diferencovatelná na celém $\mathbb{R}^2$. Vzhledem k symetrii vůči W^{aff} stačí, aby toto f splňovala pouze na jistém otevřeném okolí uzávěru fundamentální oblasti $\overline{F}$.

Pokračujme v odhadu (19). Předpokládejme, že funkce $f \in \mathcal{C}^{(4)}$.

$$\begin{aligned} (19) &\le \sum_{\substack{m,n=0 \\ m+n>0}}^{\infty} \frac{K_2}{(a+mM)^r(b+nM)^{4-r}} = \{r \text{ volíme libovolně podle potřeby}\} \le \\ &\le K_2\left\{\sum_{m,n=1}^{\infty} \frac{1}{(a+mM)^2(b+nM)^2} + \sum_{m=1}^{\infty} \frac{1}{(a+mM)^4} + \sum_{n=1}^{\infty} \frac{1}{(b+nM)^4}\right\} \le \\ &\le K_2\left\{\sum_{m,n=1}^{\infty} \frac{1}{(mM)^2(nM)^2} + \sum_{m=1}^{\infty} \frac{1}{(mM)^4} + \sum_{n=1}^{\infty} \frac{1}{(nM)^4}\right\} = \end{aligned}$$

$$= \frac{K_2}{M^4} \left\{ \sum_{m,n=1}^{\infty} \frac{1}{(m)^2(n)^2} + \sum_{m=1}^{\infty} \frac{1}{(m)^4} + \sum_{n=1}^{\infty} \frac{1}{(n)^4} \right\} = \frac{K_3}{M^4}. \tag{27}$$

Další, co potřebujeme odhadnout, je suma $\sum_{\lambda \in P^+ \setminus S_M} |c_\lambda|$. Pro $x \in \mathbb{R}$ bude $\lfloor x \rfloor$ značit dolní celou část čísla x. Za předpokladu, že $f \in \mathcal{C}^{(6)}$, získáme

$$\begin{aligned}
&\sum_{\substack{\lambda \in P^+ \\ \lambda \notin S_M}} |c_\lambda| = \sum_{j=M+1}^{\infty} \sum_{\substack{k=0 \\ \frac{j-q_2 k}{q_1} \in \mathbb{Z}}}^{\left\lfloor \frac{j}{q_2} \right\rfloor} \left| c_{k, \frac{j-q_2 k}{q_1}} \right| = \\
&= \sum_{\substack{j=M+1 \\ \frac{j}{q_2} \notin \mathbb{Z}}}^{\infty} \sum_{\substack{k=1 \\ \frac{j-q_2 k}{q_1} \in \mathbb{Z}}}^{\left\lfloor \frac{j}{q_2} \right\rfloor} \left| c_{k, \frac{j-q_2 k}{q_1}} \right| + \sum_{\substack{j=M+1 \\ \frac{j}{q_1} \in \mathbb{Z}}}^{\infty} \left| c_{0, \frac{j}{q_1}} \right| + \sum_{\substack{j=M+1 \\ \frac{j}{q_2} \in \mathbb{Z}}}^{\infty} \left| c_{\frac{j}{q_2}, 0} \right| \leq \\
&\leq \sum_{\substack{j=M+1 \\ \frac{j}{q_2} \notin \mathbb{Z}}}^{\infty} \sum_{\substack{k=1 \\ \frac{j-q_2 k}{q_1} \in \mathbb{Z}}}^{\left\lfloor \frac{j}{q_2} \right\rfloor} \frac{q_1^3}{k^3 (j - q_2 k)^3} + \sum_{\substack{j=M+1 \\ \frac{j}{q_1} \in \mathbb{Z}}}^{\infty} \frac{q_1^6}{j^6} + \sum_{\substack{j=M+1 \\ \frac{j}{q_2} \in \mathbb{Z}}}^{\infty} \frac{q_2^6}{j^6} \leq \\
&\leq \sum_{\substack{j=M+1 \\ \frac{j}{q_2} \notin \mathbb{Z}}}^{\infty} \sum_{\substack{k=1 \\ \frac{j-q_2 k}{q_1} \in \mathbb{Z}}}^{\left\lfloor \frac{j}{q_2} \right\rfloor} \frac{q_1^3}{(j-1)^3} + \sum_{\substack{j=M+1 \\ \frac{j}{q_1} \in \mathbb{Z}}}^{\infty} \frac{q_1^2}{(j-1)^2} + \sum_{\substack{j=M+1 \\ \frac{j}{q_2} \in \mathbb{Z}}}^{\infty} \frac{q_2^2}{(j-1)^2} \leq \\
&\leq \sum_{j=M+1}^{\infty} \frac{q_1^3}{(j-1)^2} + \sum_{j=M+1}^{\infty} \frac{q_1^2}{(j-1)^2} + \sum_{j=M+1}^{\infty} \frac{q_2^2}{(j-1)^2} \leq \\
&\leq \left\{ q_1^3 + q_1^2 + q_2^2 \right\} \sum_{j=M}^{\infty} \frac{1}{j(j-1)} = \left\{ q_1^3 + q_1^2 + q_2^2 \right\} \frac{1}{M-1}.
\end{aligned} \tag{28}$$

Celkově tedy bodový rozdíl (18) můžeme odhadnout jako

$$\begin{aligned}
&|f(x) - \Psi_M(x)| \leq |W| \cdot \sum_{\lambda \in S_M} \left| c_\lambda - c_\lambda^M \right| + |W| \cdot \sum_{\substack{\lambda \in P^+ \\ \lambda \notin S_M}} |c_\lambda| \leq \\
&\leq |W| \cdot \sum_{\lambda \in S_M} \frac{K_3}{M^4} + |W| \cdot \left\{ q_1^3 + q_1^2 + q_2^2 \right\} \frac{1}{M-1} \leq \\
&\leq |W| \cdot \frac{K_3}{M^2} + |W| \cdot \left\{ q_1^3 + q_1^2 + q_2^2 \right\} \frac{1}{M-1} \leq \frac{K_4}{M-1}.
\end{aligned} \tag{29}$$

5 Závěr

Výsledek můžeme velice snadno formulovat. Buď $f \in \mathcal{C}^{(6)}(\mathcal{U})$, kde $\mathcal{U}$ je otevřené okolí uzávěru fundamentální oblasti $\overline{F}$, symetrická vůči afinní Weylově grupě W^{aff}, potom funkční posloupnost $\{\Psi_M\}_{M=1}^{\infty}$ konverguje k f stejnoměrně na $\mathcal{U}$.

Funkce z $\mathcal{C}^{(6)}$, které jsme obdrželi jako výsledek, jsou samozřejmě velmi silný předpoklad, ale to je jen důsledek našeho do jisté míry naivního přístupu. Věříme, že v reálu konvergují rozvoje k původní funkci za slabších podmínek. Dále poznamenejme, že uvedený přístup je možný aplikovat i na algebry vyšších ranků. Nicméně předpokládáme, že při aplikaci tohoto přístupu bude stoupat požadavek na spojitou diferencovatelnost rozvíjené funkce. To nám ukázala i zkušenost, kterou jsme získali s algebrou A_1, pro kterou lze tímto způsobem dokázat konvergenci již pro funkce z $\mathcal{C}^{(2)}$.

Literatura

[1] J. Fuksa. *Porovnání dvoudimenzionálních transformací Fourierova typu.* Výzkumný úkol, FJFI ČVUT v Praze (2011). `http://ssmf.fjfi.cvut.cz/studthes/2008/Fuksa_res.pdf`

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

[3] J. Patera, A. Zaratsyan. *Discrete and continuous cosine transform generalized to Lie groups $SU(2) \times SU(2)$ and $O(5)$.* J. Math. Phys. **46** (2005).

[4] J. Patera, A. Zaratsyan. *Discrete and continuous cosine transform generalized to Lie groups $SU(3)$ and $G(2)$.* J. Math. Phys. **46** (2005).

Bidirectional Texture Function Three Dimensional Pseudo Gaussian Markov Random Field Model*

Michal Havlíček†

3rd 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, AS CR

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 angles. Such a function might be represented by thousands of images of given material surface. Original data cannot be used due to its size and some compression is necessary. This paper presents a novel probabilistic model for BTF textures. The method combines synthesized smooth texture and corresponding range map to produce the required BTF texture. Proposed scheme enables very high BTF texture compression ratio and may be used to reconstruct BTF space as well.

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

Abstrakt. Obousměrná funkce textury 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 materiálu. Původní data nelze díky jejich velikosti použít a je třeba je komprimovat. Tento článek představuje nový pravděpodobnostní model pro BTF textury. Tato metoda kombinuje syntetizovanou hladkou texturu a odpovídající hloubkovou mapu výsledkem čehož je požadovaná BTF textura. Navržený postup umožňuje velmi vysokou úroveň komprese BTF textur a může být také využit při rekonstrukci BTF prostoru.

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

1 Introduction

Bidirectional Texture Function (BTF) [3] is recent most advanced representation of real material surface [6]. It is a seven dimensional function describing surface texture appearance variations due to changing illumination and viewing conditions. The arguments of this function are planar coordinates, spectral plane, azimuthal and elevation angles of both illumination and view respectively.

Such a function for given material is typically represented by thousands of images of surface taken for several combinations of the illumination and viewing angles [16]. Direct

*This research was supported by the grant GAČR 102/08/0593.

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

utilization of acquired data is inconvenient because of extreme memory requirements [16]. Even simple scene with only several materials requires about terabyte of texture memory which is still far out of limits for any current and near future graphics hardware.

Several so called intelligent sampling methods, i.e., based on some sort of original small texture sampling, for example [4], were developed to solve this problem, but they still require to store thousands sample images of the original BTF. In addition, they often produce textures with disruptive visual effects except for the Roller algorithm [12]. Another disadvantage is that they are sometimes very computationally demanding [6].

Contrary to the sampling approaches utilization of mathematical model is more flexible and offers significant compression, because only several parameters have to be stored only. Such a model can be used to generate virtually infinite texture without visible discontinuities. On the other hand, mathematical model can only approximate real measurements, which may result in some kind of visual quality compromise.

One possibility is utilization of random field theory [8]. Generally, texture is assumed to be realization of random field. Additional assumptions further vary depending on particular model. BTF theoretically requires seven dimensional model owing to its definition, but it is possible to approximate general BTF model with a set of much simpler less dimensional ones, three [10],[13] and two dimensional [9],[11] in practice. Mathematical model based on random fields provides easy smooth texture generation with huge compression and visual quality ratio for a large set of textures [6].

Multiscale approach (Gaussian Laplacian pyramid (GLP), wavelet pyramid or subband pyramids) provides successful representation of both high and low frequencies present in texture so that the hierarchy of different resolutions of an input image provides a transition between pixel level features and region or global features [9]. Each resolution component is modelled independently.

We propose an algorithm for BTF texture modelling which combines material range map with synthetic smooth texture generated by multiscale three dimensional Pseudo Gaussian Markov Random Field (3D PGMRF) [1]. Overall texture visual appearance during changes of viewing and illumination conditions is simulated using displacement mapping technique [17].

2 BTF 3D PGMRF Model

The overall BTF 3D PGMRF model scheme can be found on Figure 1. First stage is material range map estimation followed by optional data segmentation (k-means clustering with color cumulative histograms of individual BTF images in perceptually uniform CIELAB colour space as the data features) [9]. An analysed BTF subspace texture is decomposed into multiple resolution factors using GLP [9]. Each resolution data are then independently modelled by their dedicated 3D PGMRF resulting with set of parameters. Multispectral fine resolution subspace component can be then obtained from the pyramid collapse procedure, i.e., the interpolation of sub band components which is the inversion process to the creation of the GLP [9]. Resulting smooth texture is then combined with range map via displacement mapping filter of graphics hardware or software.

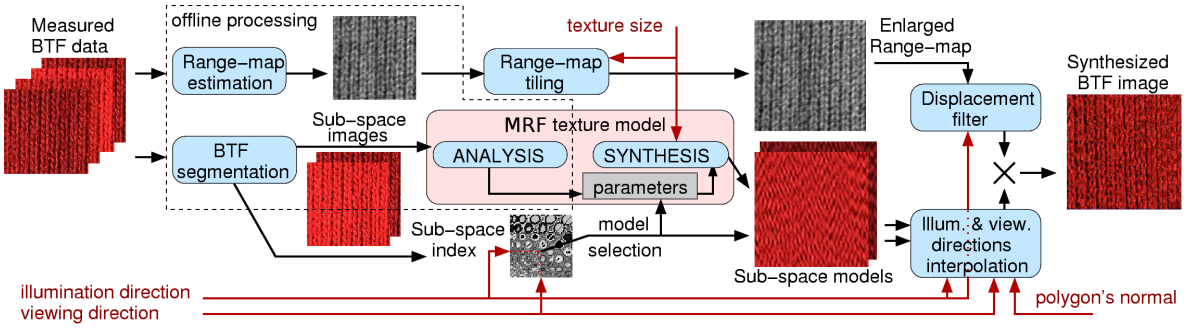


Figure 1: BTF 3D PGMRF model scheme.

2.1 Range Map

The overall roughness of surface significantly influences the BTF texture appearance. This attribute can be specified by range map which comprise information of relative height or depth of individual sites on the surface. Range map can be either measured on real surface or estimated from images of this surface by several existing approaches such as the shape from shading [7], shape from texture [5] or photometric stereo [18]. Since the number of mutually registered BTF measurements for fixed view is sufficient (e.g., 81 in case of the University of Bonn data [16]) it is possible to use over determined photometric stereo to obtain the most accurate outcome. Range map is then stored as a monospectral image where each pixel equals relative height or depth respectively of the corresponding pixel, i.e., point of the surface. If synthesized smooth texture is larger than stored range map then range map is enlarged by the Roller technique [12] chosen for its good properties.

3 3D PGMRF Model

Three dimensional texture random field models are defined as random values representing intensity levels on multiple two dimensional lattices (three in case of widely used colour spaces such as RGB, CIELAB, YUV, YIQ for instance, all of them are widely used in computer graphics, although number of lattices is not limited). The value at each lattice location is considered to be a linear combination of neighbouring ones and some additive noise component. All lattices are considered as double toroidal.

Let a location within an $M \times M$ two dimensional lattice be denoted by (i, j) with $i, j \in J$ where the set J is defined as $J = \{0, 1, \ldots, M-1\}$. The set of all lattice locations is then defined as $\Omega = \{ (i, j) : i, j \in J \}$. Let the value of an image observation at location (i, j) and lattice k be denoted by $y(i, j, k)$ and P equals number of lattices. All random variables forming vector $y(i, j) = (y(i, j, k))\,(i, j) \in \Omega,\ k \in \hat{P}$ are expected to have zero mean. Neighbour sets relating the dependence of points at lattice k on points at lattice l are defined as $N_{kl} = \{ (i, j) : i, j \in \pm J \}$ with the associated neighbour coefficients $\theta(k, l) = \{ \theta(i, j, k, l) : (i, j) \in N_{kl} \}$ where $\pm J = \{-(M-1), \ldots, -1, 0, 1, \ldots, M-1\}$ and $k, l \in \hat{P}$. We also use shortened notation: $\theta = \{ \theta(k, l);\ k, l \in \hat{P} \}$. Our model is defined on symmetric hierarchical contextual

neighbour set (Figure 2), i.e., this holds: $r \in N_{kl} \iff -r \in N_{lk}$. Since all sets N_{kl} are equivalent in our implementation, although generally they do not have to be, we use shortened notation N for simplification purposes.

The 3D PGMRF model relates each zero mean pixel value by a linear combination of neighbouring ones and an additive uncorrelated Gaussian noise component [1]:

$$y(i,j,k) = \sum_{n=1}^{P} \sum_{(l,m)\in N} \theta(l,m,k,n)\ y(i+l, j+m, n) + e(i,j,k) \tag{1}$$

where

$$e(i,j,k) = \sum_{n=1}^{P} \sum_{(l,m)\in \Omega} c(l,m,k,n)\ w(i+l, j+m, n)$$

and $w(i,j,k)$ represents zero mean unit variance i.i.d. variable for $(i,j) \in \Omega,\ k \in \hat{P}$. Rewriting the autoregressive equation (1) to the matrix form, with random fields $y = \{\ y(i,j,k);\ (i,j) \in \Omega,\ k \in \hat{P}\ \}$ and $w = \{\ w(i,j,k);\ (i,j) \in \Omega,\ k \in \hat{P}\ \}$ model equations become $By = w$ where

$$B = \begin{pmatrix} B(\theta(1,1)) & B(\theta(1,2)) & \dots & B(\theta(1,P)) \\ B(\theta(2,1)) & B(\theta(2,2)) & \dots & B(\theta(2,P)) \\ \vdots & \vdots & \ddots & \vdots \\ B(\theta(P,1)) & B(\theta(P,2)) & \dots & B(\theta(P,P)) \end{pmatrix}.$$

Matrix B is in fact $PM^2 \times PM^2$ sized matrix composed of $M^2 \times M^2$ block circulant matrices

$$B(\theta(k,l)) = \begin{pmatrix} B(\theta(k,l))_1 & B(\theta(k,l))_2 & \dots & B(\theta(k,l))_M \\ B(\theta(k,l))_M & B(\theta(k,l))_1 & \dots & B(\theta(k,l))_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ B(\theta(k,l))_2 & B(\theta(k,l))_3 & \dots & B(\theta(k,l))_1 \end{pmatrix} \tag{2}$$

where each element of matrix (2): $B(\theta(k,l)_p)$ is $M \times M$ circulant matrix with elements $b(\theta(k,l))_p(m,n)$ defined as:

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

Let us remark that the selection of an appropriate model support is important to obtain good results in modelling of a given random field. If used hierarchical contextual neighbourhood set is too small then corresponding model cannot capture all details of the random field. Contrariwise inclusion of the unnecessary neighbours increases both time and memory demands with possible model performance degradation as an additional source of noise.

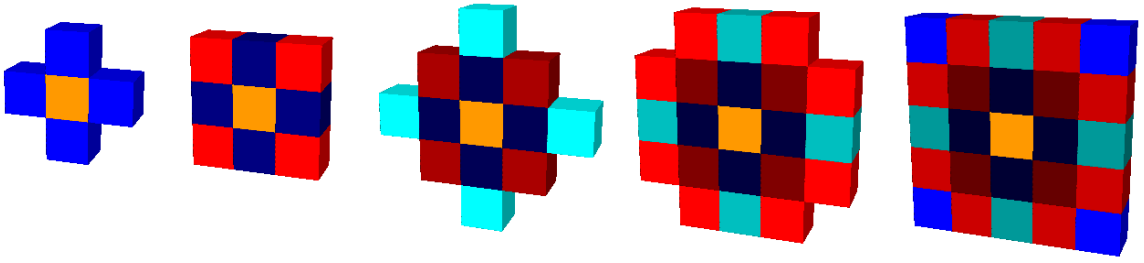


Figure 2: Examples of the used hierarchical contextual neighbourhood sets. The (0,0) position is represented by the central light square while relative neighbour locations are darker surrounding ones. First order neighbourhood to fifth order neighbourhood, from left to right.

3.1 Parameters Estimation

The model is completely specified by parameters $\theta = \{\ \theta(k,l) : k \geq l,\ k \in \hat{P},\ l \in \hat{P}\ \}$ (as $\theta(k,l) = \theta(l,k),\ \forall k,l$ due to symmetry of neighbourhood) and vector ρ where each component $\rho(k),\ k \in \hat{P}$ of ρ specifies variance of noise component of lattice k. These parameters may be estimated by means of the Least Squares (LS) technique [1]. The LS estimates of the neighbour set coefficients $\theta(i,j,k,l),\ (i,j) \in N,\ k,l \in \hat{P}$ of vector θ are independent of the variance vector ρ. It is due to correlation structure of noise component [1]:

$$\varepsilon\{e(i,j,k)e(i+l,j+m,n)\} = \begin{cases} -\theta(l,m,k,n)\sqrt{\rho(k)\rho(n)} & (l,m) \in N\,, \\ \rho(n) & l=0,\ m=0,\ k=n\,, \\ 0 & otherwise\,. \end{cases}$$

If $\rho(k) = \rho(n)\ \forall k,n \in \hat{P}$ then the random field becomes strictly Gaussian Markov with $\hat{\theta}$ depending on $\hat{\rho}$ making impossible non iterative estimation [1].

Estimates may be derived from equating the observed values to their expected ones, i.e., $y(i,j) = Q^T(i,j)\theta,\ (i,j) \in \Omega$ where

$$Q(i,j) = \begin{pmatrix} q(i,j,1,1) & q(i,j,1,2) & \dots & 0 \\ 0 & q(i,j,2,1) & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & q(i,j,P,P) \end{pmatrix}^T$$

$$q(i,j,k,n) = \begin{cases} (y(i+l,j+m,k) + y(i-l,j-m,k),\ (l,m) \in N) & k=n \\ (y(i+l,j+m,n),\ (l,m) \in N) & k<n \\ (y(i-l,j-m,n),\ (l,m) \in N) & k>n \end{cases}$$

The LS solution $\hat{\theta}$ and $\hat{\rho}$ can be found then as [1]

$$\hat{\theta} = \left(\sum_{(i,j)\in\Omega} Q(i,j)Q^T(i,j) \right)^{-1} \left(\sum_{(i,j)\in\Omega} Q(i,j)y(i,j) \right),$$

$$\hat{\rho} = \frac{1}{M^2} \sum_{(i,j)\in\Omega} (y(i,j) - \hat{\theta}^T Q(i,j))^2 \quad .$$

This approximation of real values of parameters allows to avoid expensive numerical optimization method at the cost of accuracy [1].

Additional parameter is mean $\mu = (\mu(k))$, $k \in \hat{P}$. Mean of each spectral plane is estimated as the arithmetic mean and then is subtracted from the plane (prior to estimation of θ and ρ) so that image can be regarded as realization of zero mean random field.

3.2 Image Synthesis

Estimated model parameters $\hat{\theta}$, $\hat{\rho}$ and $\hat{\mu}$ represent original data. So that only their values (several real numbers) need to be stored instead of those data themselves thus this approach offers extreme compression.

A general multidimensional Gaussian Markov random field model has to be synthesized using some of the Markov Chain Monte Carlo (MCMC) method [8]. Due to the double toroidal lattice assumption it is possible to employ efficient non iterative synthesis based on the fast discrete Fourier transformation (DFT) [1].

The model equations (1) may be expressed in terms of the DFT of each lattice as

$$Y(i,j,k) = \sum_{n=1}^{P} \sum_{(l,m)\in N} \theta(l,m,k,n) Y(i,j,n) e^{\sqrt{-1}\omega} + \sqrt{\rho(k)} W(i,j,k) \tag{3}$$

where $Y(i,j,k)$ and $W(i,j,k)$ are the two dimensional DFT coefficients of the image observation $y(i,j,k)$ and noise sequence $w(i,j,k)$, respectively, and $\omega = \frac{2\pi(il+jm)}{M}$ with $(i,j) \in \Omega$ and $k \in \hat{P}$. Model equations (3) can be written in matrix form as $Y(i,j) = \Lambda(i,j)^{-1}\, \Sigma^{\frac{1}{2}}\, W(i,j)$ with the matrices Σ and $\Lambda(i,j)$ defined as [1]:

$$\Sigma = \begin{pmatrix} \rho(1) & 0 & \dots & 0 \\ 0 & \rho(2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \rho(P) \end{pmatrix},$$

$$\Lambda(i,j) = \begin{pmatrix} \lambda(i,j,1,1) & \lambda(i,j,1,2) & \dots & \lambda(i,j,1,P) \\ \lambda(i,j,2,1) & \lambda(i,j,2,2) & \dots & \lambda(i,j,2,P) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(i,j,P,1) & \lambda(i,j,P,2) & \dots & \lambda(i,j,P,P) \end{pmatrix},$$

$$\lambda(i,j,k,n) = \begin{cases} 1 - \sum_{(l,m)\in N} \theta(l,m,k,n)\, e^{\sqrt{-1}\frac{2\pi(il+jm)}{M}} & k = n \\ -\sum_{(l,m)\in N} \theta(l,m,k,n)\, e^{\sqrt{-1}\frac{2\pi(il+jm)}{M}} & k \neq n \end{cases} .$$

The synthesis process begins with generation of two dimensional arrays of white noise w with help of pseudo random number generator for each spectral plane independently. It is followed by two dimensional discrete fast Fourier transform so that arrays W are obtained. Transformation $\Lambda(i,j)^{-1}\ \Sigma^{\frac{1}{2}}\ W(i,j)$ is then computed for each discrete frequency index $(i,j) \in \Omega$. Following step which is inverse two dimensional fast discrete Fourier transform results with image y with zero mean spectral planes so desired mean $\mu(k)$ need to be added to corresponding plane k, $\forall k \in \hat{P}$.

4 Results

We have tested BTF 3D PGMRF model on BTF colour textures from the University of Bonn BTF measurements [16] which represents the most accurate ones available to date [6]. Every material in the database is represented by 6561 images, 800 × 800 RGB pixels each, corresponding to 81 × 81 different view and illumination angles respectively.

The open source project Blender[1] with plugin for BTF texture support [14] was used to render the results. Very simple scene consisting one source of light one three dimensional object represented by polygons and one camera (its coordinates defines view angles) was rendered several times with varying illumination angles while view angles stayed fixed. Synthetic smooth texture combined with range map in displacement mapping filter of Blender was mapped on the object. Several examples may be reviewed on Figures 3 and 4 where visual quality of synthesised BTF may be compared with measured BTF.

The model was also tested on colour textures picked from Amsterdam Library of Textures (ALOT)[2] [2] which consists more coloured, but less dense sampled materials.

5 Conclusion

The main benefit of the presented method is realistic representation of texture colourfulness, which is naturally apparent in case of very distinctively coloured textures. Any simpler two dimensional random field model is not almost able to achieve such results due to colour information loss caused by necessary spectral decorrelation of input data [9]. The multiscale approach is more robust and sometimes allows better results than the single scale one it is when model cannot represent low frequencies properly. This model offers efficient and seamless enlargement of BTF texture to arbitrary size and very high BTF texture compression ratio which cannot be achieved by any other sampling based BTF texture synthesis method while still comparable with other random field BTF models [6]. This can be useful for, e.g., transmission, storing and modelling realistic visual surface texture data with possible application in robust visual classification, human perception study, segmentation, virtual prototyping, image restoration, aging modelling, face recognition and many others [6]. On the other hand the model has still moderate computation complexity. Described approach does not need any kind of time consuming numerical optimisation, e.g., Markov chain Monte Carlo method which is usually employed for such tasks [8]. In addition analysis complexity is not important too much since it is performed

[1] http://www.blender.org

[2] http://staff.science.uva.nl/~ aloi/public_alot/

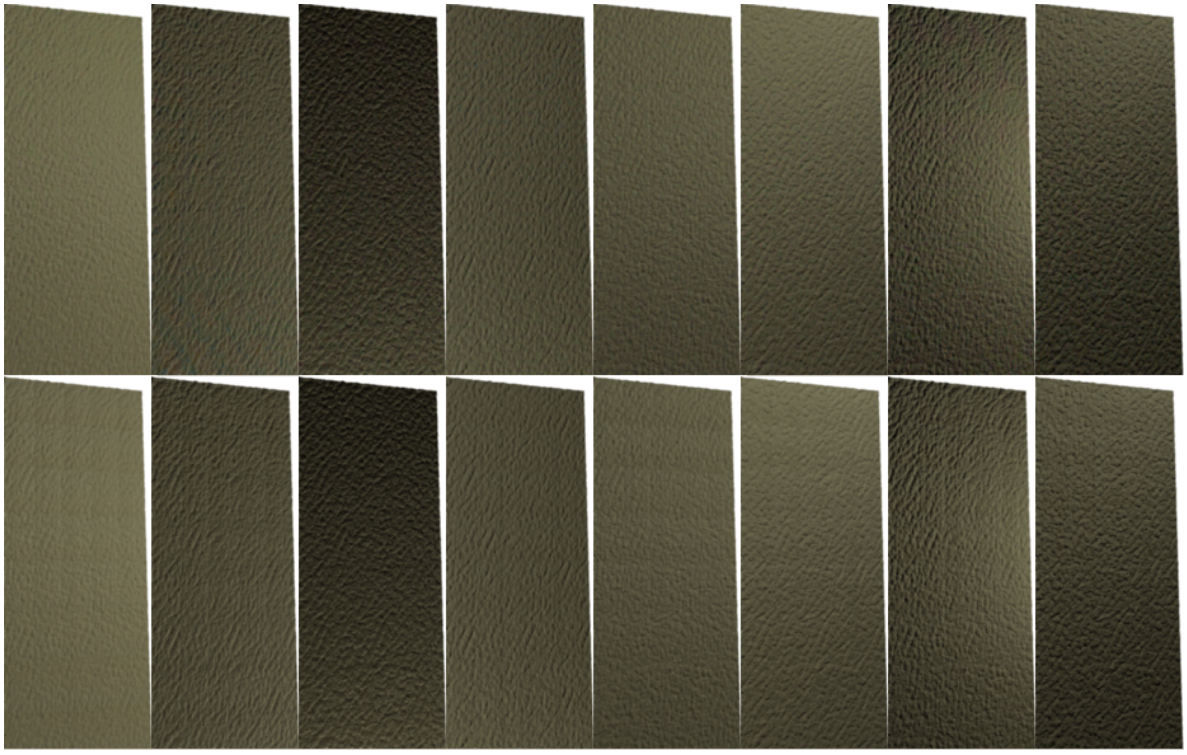


Figure 3: A curved plane with mapped BTF. Bottom row: the original measured BTF (artificial leather). Top row: the synthesised BTF. Each column represents one unique illumination condition. Camera stayed fixed for all shots.

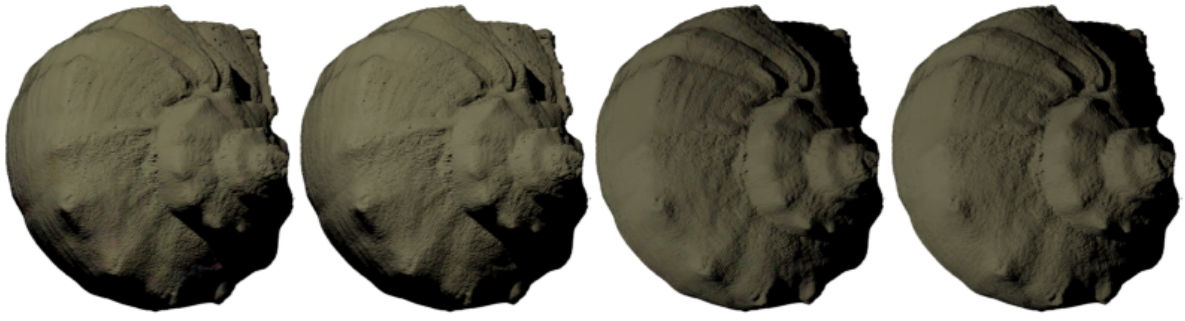


Figure 4: BTF mapped on complex geometry. The original measured BTF of artificial leather (2nd and 4th object from left) and corresponding (same illumination and view angles) synthesised BTF (1st and 3rd object from left).

once per material and offline. Both analysis and synthesis steps may be performed in parallel. Utilizing displacement mapping is both efficient (due to direct hardware support) and improve overall visual quality of the result. In addition, this model may be used to reconstruct BTF space, i.e., synthesize missing parts, previously unmeasured, of the BTF space. Oh the other hand the method is based on the mathematical model in contrast to intelligent sampling type methods and as such it can only approximate realism of the original measurement. The approximation strongly depends on several factors such as size and nature of training data and size of neighbourhood set.

6 Future Work

This BTF model might be further tested and compared with other random field based models. Overall texture visual quality comparison is complex and not yet completely solved problem. We would like to focus on texture overall colour quality comparison because direct pixel to pixel comparison (or based on texture geometry) seems to be inconvenient due to stochastic character of synthesised textures. One possibility might be Generalized Colour Moments [15].

Very interesting task would be extension of current implementation by means of parallel programming, for example with use of OpenMP[3] interface or other multithreading techniques (TBB[4], UPC[5]).

An extensive utilization of graphics processing unit seems to be applicable as well, but requires more sophisticated adaptation of current implementation where all computation is performed in the central processing unit. It would be possible to utilize framework OpenCL[6] or standard OpenGL[7]. Such improvements would notably increase overall performance which would be beneficial especially in case of virtual reality system requiring as fast as possible or even real time render and thus fast texture synthesis as well.

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] G. J. Burghouts, J. M. Geusebroek. *Material-specific Adaption of Color Invariant Features*. Pattern Recognition Letters **30** (2009), 306–313.

[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.

[3] http://openmp.org
[4] http://threadingbuildingblocks.org
[5] http://upc.gwu.edu
[6] www.khronos.org/opencl
[7] www.opengl.org

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

[5] P. Favaro, S. Soatto. *3-D shape estimation and image restoration: exploiting defocus and motion blur*. Springer-Verlag New York Inc. (2007).

[6] J. Filip, J, M. Haindl. *Bidirectional texture function modeling: A state of the art survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(11), (2009) 1921–1940.

[7] R. T. Frankot, R. Chellappa. *A method for enforcing integrability in shape from shading algorithms*. IEEE Trans. on Pattern Analysis and Machine Intelligence **10**(7), (1988) 439–451.

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

[9] 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.

[10] 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.

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

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

[13] M. Haindl, M. Havlíček. *Bidirectional Texture Function Simultaneous Autoregressive Model*. Computational Intelligence for Multimedia Understanding, Lecture Notes in Computer Science **7252**, Springer Berlin / Heidelberg (2012), 149–159.

[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] F. Mindru, T. Tuytelaars, L. Van Gool, T. Moons. *Moment invariants for recognition under changing viewpoint and illumination*. Computer Visual Image Understanding **94**(1–3), Elsevier Science Inc., (2004), 3–27.

[16] 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.

[17] 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.

[18] R. Woodham. *Photometric method for determining surface orientation from multiple images*. Optical engineering **19**(1), (1980) 139–144.

Radiation Tolerance Measurements of Medipix2 Detector

Martin Hejtmánek

2nd 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 radiation tolerance measurements for testing Medipix2 pixel detector, especially its sensor. Radiation tolerance is important property of pixel detector. In many applications, particularly in medicine diagnostics good radiation resistance of detectors can lead to great financial savings. In this article, the methodology of testing of radiation tolerance is developed and verified. The measurements presented here serve as a preparation for upcoming long-term testing of Medipix2 detector in October 2012 at UJP Praha[1] company.

Keywords: pixel detector, Medipix2, radiation tolerance, readout electronics

Abstrakt. Tento příspěvek se zabývá měřením radiační odolnosti pixelového detektoru Medipix2, zejména jeho senzoru. Radiační odolnost je velmi důležitá pro mnoho aplikací, například v lékařské diagnostice, kde může použití dobře radiačně odolného detektoru vést k značným finančním úsporám. V tomto článku je navržena a diskutována správnost metodologie pro taková měření. Zde prezentovaná měření a výsledky slouží jako podklad pro další, tentokrát intenzivnější a delší měření detektoru Medipix2, plánované na říjen roku 2012 v prostorách firmy UJP Praha.

Klíčová slova: pixelový detektor, Medipix2, radiační odolnost, čtecí elektronika

1 Introduction

This article deals with radiation tolerance measurements. The radiation damage of silicon detectors is caused by local defects of crystal structure. By gaining energy, the atoms in crystal lattice can deviate from its position, and, furthermore, these defects can expand to other parts of lattice due to oscillations. In case of silicon, the energy needed for atom to cause defect is 25 eV.

The radiation damage of silicon can influence the operation of the detector. The conductivity of silicon crystals may change and the effectivity of charge acquisition may drop. Therefore the study of radiation tolerance is very important for applications in which the detectors are exposed to radiation for a long time. Typical example of such an application is medical diagnostics – detectors with greater radiation tolerance remain functional longer and thus spare financial resources.

The purpose of measurements, described in this article, was to prove that silicon pixel detector concept (such as Medipix2 detector) is suitable for certain applications, particu-

[1] Ústav jaderných paliv, Praha

larly medical diagnostics. In the case of insufficient results, the aim of such measurements is to propose improvements in technology of constructing silicon pixel detector devices.

2 Measurement setup

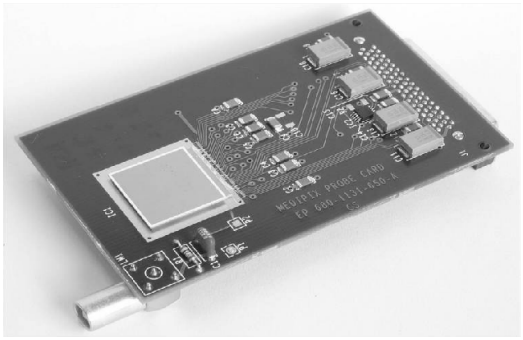


Figure 1: Timepix detector.

Figure 2: Muros2 readout interface.

In this section we will present the setup and devices used during the radiation tolerance measurements in the company UJP Praha. The aim of these measurements was to optimize the parameters and methodics for further radiation tolerance testing in October 2012. Unfortunately, these planed final measurements have not been performed in the time of writing this article. However, the results will be presented on 'Doktorandské dny' conference in November 2012.

For testing purposes Timepix detector from Medipix2 family was used together with Muros readout interface. Further information about Medipix2 and Muros can be found in [1, 4, 6]. The reason for choosing the Muros was its great stability and reliability. The devices can be seen in figures 1 and 2. In typical medical imaging application the detector can be placed outside the radiation area, therefore the Muros was shadowed by robust lead blocks of 5 cm in thickness. The radiation damage was observed only on sensor and close electronics.

As a radiation source was used ^{60}Co from IK Farmer company. The detector was placed 80 cm below the source. The dose rate in the air in this distance was $3.25\,\mathrm{Gy}\cdot\mathrm{min}^{-1}$. The reference area was $10\times 10\,\mathrm{cm}^2$. Between the detector and the source, an aluminum plate was placed in order to filter out the incoming electrons. The setup of the measurement can be seen in figure 4.

In order to be able to see different phase of radiation damage, the detector was covered by several lead plates of 1 mm in thickness. These plates were placed in a way to form small 'stair' structure as can be seen in figure 3. Each stair was formed out of two plates. Another reason for this was the inner construction of Timepix detector which affects the method of reading out the data from detector. Its pixel matrix electronics consists out of 256 columns connected to fast shift register on the bottom side (see picture 5). The data are read out from columns such that they are continuously dropping through the columns to the fast shift register. Thus the destruction of the bottom part of columns could affect the data from the upper part during the read out. By using the stair structure, we are

expecting the upper part to be more degraded by the radiation. Defects should appear more likely and sooner in the upper side of detector.

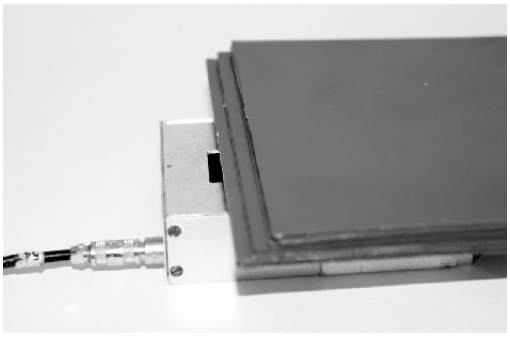


Figure 3: The 'stairs' structure on the detector.

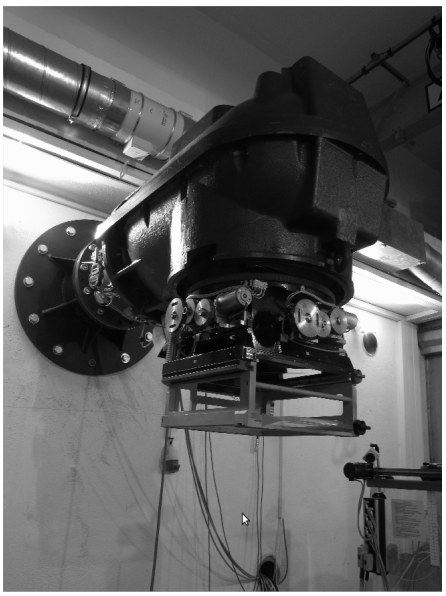


Figure 4: The ^{60}Co source.

3 Methodics of the measurement

First, the bias voltage needed to be optimized. For too high bias voltage setting, the dynamic range of the detector is depleted while for too low bias voltage setting the detector detects almost no signal. Finally, the bias voltage of 30 V was set.

The measurement was performed in three stages. Before each stage, detector was properly calibrated using the threshold equalization procedure (for further details see [7]) with the source turned off. Then the source was turned on and the detector was irradiated for 40 minutes. During the irradiation, the frames were continuously read out with acquisition time of 0.3 s. The total radiation dose which the detector was exposed to is thus

$$3.25\,\mathrm{Gy}\cdot\mathrm{min}^{-1}\times 40\,\mathrm{min}\times 3 = 390\,\mathrm{Gy}\,.$$

Two different effects were investigated:

1. The change of measured pixel matrix values with respect to gained radiation dose: Local defects of the sensor can be easily detected by comparison of the values of the matrices. It is expected that the degradation of a pixel will affect each pixel lying above in column as discussed in section 2.

2. The change of calibration matrix (threshold correction for each pixel) obtained from threshold equalization: These changes can be related to the change of pixel sensitivity. Statistically, the calibration values should be normally distributed. However, irradiation can cause shift of calibration values.

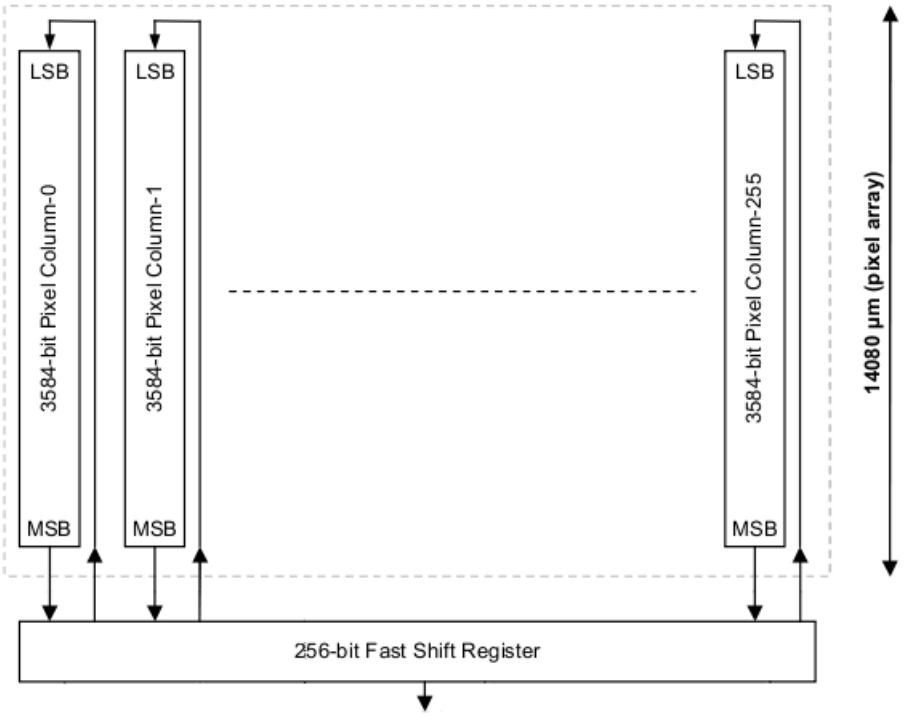


Figure 5: Schematics of pixel electronics below sensor on the Medipix2 detector.

4 Results

In this section, the obtained results are presented. For the two effects discussed above, the end- and start-state were compared by subtracting the corresponding matrices. In order to eliminate noise (in the first case), the states were computed as an average from 10 consecutive frames. First 10 frames were taken from the beginning of each stage and last 10 frames from the end, respectively. Furthermore, the distributions were computed for each subtraction-matrix.

The results can be seen in figures 6-9. As can be seen, during the stages the data differences grow (figures 6, 7). Furthermore, in the subtraction-matrix from third stage of measurement, one can clearly see that in the top part of chip the differences are much more bigger. This is caused by the stair-like shadowing of the detector. The distribution in figure 7 shows that the distribution of differences shifts to the left towards negative values. That means, the sensor sensitivity is lower after gaining radiation dose. At the end of measurement, the pixel counts were lower in comparison with the start of the measurement.

However, the equalization values seem to be nearly constant, as can be seen in figures 8 and 9. That means, the changes probably do not affect the pixel electronics below the sensor.

In figure 9 on the right side typical data matrix obtained during measurements. One can see clearly the stair-like structure on the detector.

5 Proposal for final measurements in October

As already mentioned, these measurements were performed in order to verify the methodics of radiation tolerance measurement. In October, we would like to repeat the measurement with much more gained radiation dose (the detector will be irradiated for several days instead of hours). Furthermore, another interesting variable to measure will be the temperature of the detector and its affect to the obtained results. Since the equalization of the detector does not change with respect to gained radiation dose, the measurement could be performed more continuously and the equalization procedure will not be performed as often. This will ensure that the next measurements will be more automated. However, since with frequency 3 frames per second there will be a lot of data to store on the computer's hard drive, a script written in bash language will be used to remove frames not needed for analysis and to keep just several images in sequence once per specified time period. The first version of the script can be seen in following listing:

```
 #!/bin/bash

IN_DIR=/home/medipix/Desktop/data-pokus/all
OUT_DIR=/home/medipix/Desktop/data-pokus/selected

START=/tmp/start_time
END=/tmp/end_time

touch -d '-10 seconds' $START
touch  $END

FILE_TO_SAVE=$(ls -t1 $(find $IN_DIR ! -newer $END -newer $START -type f)
               | head -n 1)
FILE_TO_SAVE_NAME=$(date -r $FILE_TO_SAVE '+%s')


if [ "$FILE_TO_SAVE" == "" ]; then
   echo $(date)"
        >> /var/log/log-ujp.txt
   echo "No file detected and thus not backupped!"
        >> /var/log/log-ujp.txt
   echo "Check whether Medipix is working correctly."
        >> /var/log/log-ujp.txt
   exit 1;
else
   mv $FILE_TO_SAVE $OUT_DIR/$FILE_TO_SAVE_NAME
   find $IN_DIR ! -newer $END -type f -exec /bin/rm -f '{}' +
fi

rm -f $START $END
```

```
# and process file
# print a frame using gnuplot
gnuplot << EOF
set xrange[0:511]
set yrange[0:511]
set view map

set pm3d map
set size square 1,1
set palette defined ( 0 '#000090',\
                      1 '#000FFF',\
                      2 '#0090FF',\
                      3 '#0FFFEE',\
                      4 '#90FF70',\
                      5 '#FFEE00',\
                      6 '#FF7000',\
                      7 '#EE0000',\
                      8 '#7F0000' )

set xtics 0.0,128.0,511.0 font "Helvetica, 12" scale 0.4 textcolor ls 7
set ytics 0.0,128.0,511.0 font "Helvetica, 12" scale 0.4 textcolor ls 7

set colorbox
set cbrange [0:8192]
set cbtics 0.0,2048.0,8192.0 scale 0.3

set title "Medipix frame '$OUT_DIR/$FILE_TO_SAVE_NAME'" font "Helvetica, 14"
#plot "$OUT_DIR/$FILE_TO_SAVE_NAME" using 1:2:3 with image notitle
plot "$OUT_DIR/$FILE_TO_SAVE_NAME" matrix with image notitle

pause 7

EOF
```

The script will first copy the needed file(s) from frames directory to result directory, then it will remove all other files, and after then it will print the kept frame by using `gnuplot` program. The script will be executed regularly via the standard Linux program `cron`.

6 Conclusion

The methodics of measurement of radiation tolerance was verified. As can be seen from results, the changes in sensitivity of sensor is an effect which is worth of examination. However, it is expected that the pixel electronics is not affect by radiation, at least not as fast as sensor. Therefore there is no need to perform equalization procedure so often.

This fact can be used for better measurement automation.

In October, final measurement will be performed. In addition to currently examinated variables, the temperature of the sensor and the effect of post-irradiation annealing will be investigated. The measurement will also be performed during much longer period (several days) in order to gain sufficient amount of radiation dose. The final results will be presented on Doktorandské dny conference in November.

References

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

[2] C. Grupen. *Particle Detectors.* Cambridge University Press, (1996)

[3] K. Kleinknecht. *Detectors for particle radiation.* Cambridge University Press, 2nd edition (1998)

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

[5] 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`

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

[7] *Pixelman Manual.* `http://aladdin.utef.cvut.cz/ofat/others/Pixelman/Pixelman_manual.html`

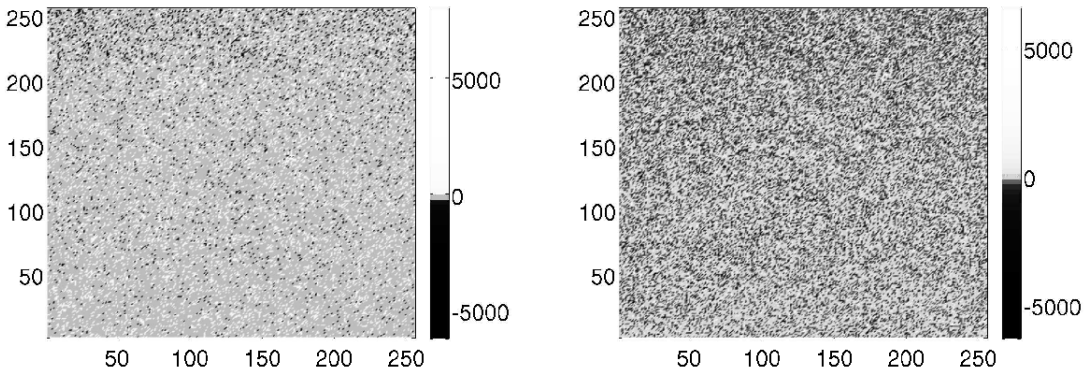


Figure 6: The subtraction-matrix from first stage of measurement (left), and from second stage of measurement (right).

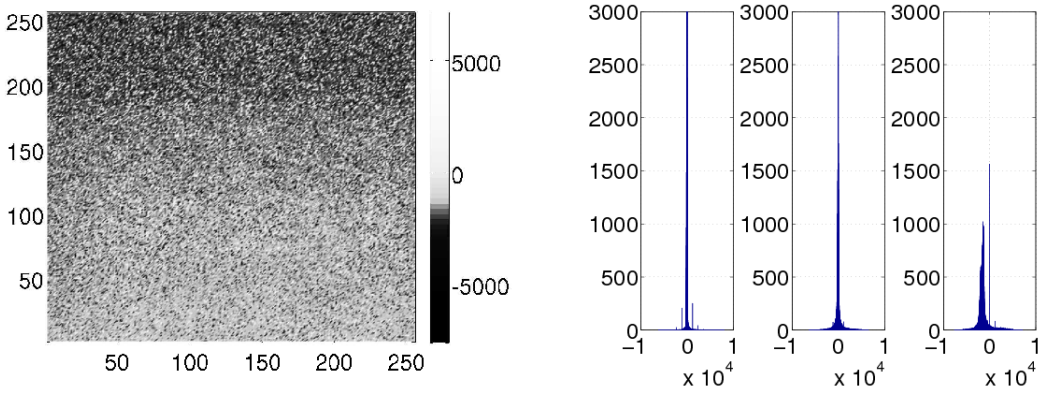


Figure 7: The subtraction-matrix from third stage of measurement (left), and histograms of three subtraction-matrices (right).

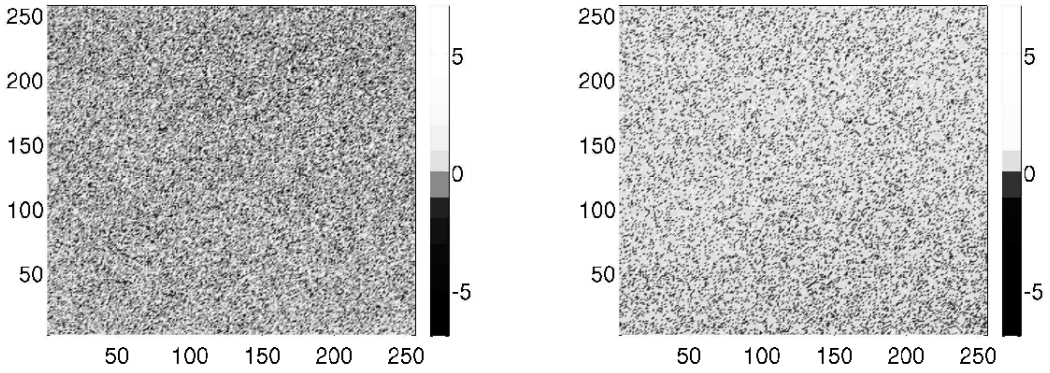


Figure 8: The subtraction-matrix of equalizations before and after first measurement (left), and before and after second measurement (right).

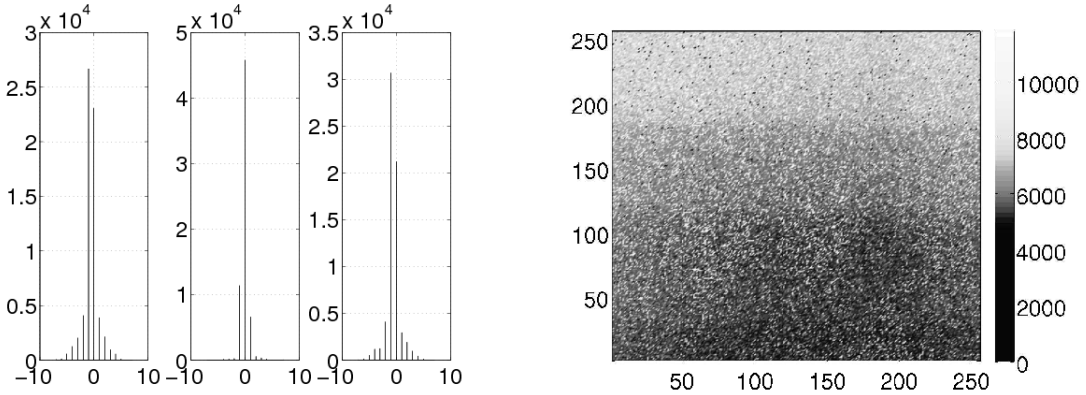


Figure 9: Histograms of two equalization subtraction-matrices (left), and typical frame with significant stair-like structure.

From The Generalization of TASEP in Two Dimensions to the Egress Simulation Model*

Pavel Hrabák

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

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

Abstract. This contribution summarizes the work about two dimensional cellular automata used in pedestrian crowed modeling. It is a compilation of the articles [2] and [3] and concerns with several ideas about phase transitions presented in [1]. The main goal is to introduce the transition from simple one-dimensional TASEP to the model of pedestrian dynamics comparing the basic features of phase transitions induced by density profiles changes.

Keywords: cellular automata, pedestrian dynamics, density profiles

Abstrakt. Tento příspěvek shrnuje výsledky v modelování pohybu chodců pomocí dvojrozměrných celulárních automatů. Jedná se o kompilát článků [2] a [3] a zabývá se studií fázových přechodů prezentovaných v [1]. Hlavním cílem je představit přechod od jednoduchého jednorozměrného modelu TASEP k modelu pohybu chodců porovnáním základních vlastností fázových přechodů spojených s hustotními profily.

Klíčová slova: celulární automaty, modelování chodců, hustotní profily

1 One Dimensional TASEP

The totally asymmetric simple exclusion process (TASEP) is defined on a discrete finite lattice of N cells. The particles move along the lattice in one direction by hopping to the neighboring cell. An particle in the bulk jumps to the cell on the right with probability p if the target cell is empty. New particle enters the lattice by hopping to the first site with probability α, when the site is empty; and a particle at the end of the lattice leaves the system with probability β. Here we implicitly assume that $p, \alpha, \beta \in \langle 0, 1\rangle$ are parameters of the system.

The occurrence of those jumps differs according to the updating procedure. The "playground" of the model can be represented by a weighted oriented graph schematically demonstrated in Figure 1. Vertices of the graph are cells of the lattice denoted by their positions $\{1, 2, \ldots, N\}$ together with the left reservoir 0 and the right reservoir $N + 1$. The edges are given as a set of ordered pairs $(i, i + 1)$, $i = 0, 1, \ldots, N$. The weights

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

$h(i, i+1)$ are given as

$$h(i,i+1) = \begin{cases} p & i = 1, 2, \ldots, N-1\,, \\ \alpha & i = 0\,, \\ \beta & i = N\,. \end{cases} \tag{1}$$

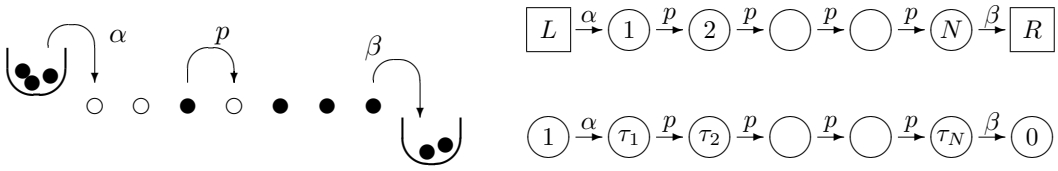


Figure 1: Illustration to the definition of TASEP. Taken from [3]

The current state of the system in time t is expressed by means of state variables $\tau_i(t)$, $i = 1, 2, \ldots, N$, where

$$\tau_i(t) = \begin{cases} 1 & \text{if the site } i \text{ is occupied}\,, \\ 0 & \text{if the site } i \text{ is empty}\,. \end{cases} \tag{2}$$

The left reservoir can be considered as an always occupied cell ($\tau_0 \equiv 1$) and the right reservoir as an always empty cell ($\tau_{N+1} \equiv 0$).

If we consider the system with time continuous dynamics, the hops of a particle are driven by the Poisson process, i.e. a particle in the cell i waits an exponential time with parameter 1, than it hops to the cell $i+1$ with probability $h(i, i+1)$ if the target cell is empty. The time discrete realization of this dynamics is the so called *random sequential* update. In every step one edge $(i, i+1)$ is chosen at random and, if it is possible, a particle hops from i to $i+1$ with probability $h(i, i+1)$.

For simulation purposes, several parallel time-discrete updating schemes are used. Here we summarize the most frequent ones described in [4]:

(a) *fully parallel update*: The exclusion rule is applied on all transitions $(i, i+1)$ simultaneously.

(b) *forward sequential update*: The exclusion rule is applied on transitions in order $(0,1), (1,2), \ldots, (N, N+1)$.

(c) *backward sequential update*: The exclusion rule is applied on transitions in order $(N, N+1), (N-1, N), \ldots, (0,1)$.

All of those updates have a particle oriented variant, i.e. the exclusion rule is applied only on occupied sites.

Considering the average occupancy of the cell i, we obtain the so called density profile $(\varrho_1, \varrho_2, \ldots, \varrho_N)$, where $\varrho_i = \langle \tau_i \rangle$. The density profile of TASEP has been extensively studied in [4] and many other works. By evaluating the density profiles, we can distinguish 3 different phases of the system: The low density phase, high density phase and the maximal current phase. The low and high density phase can be divided into two subphases according to the density profile near the boundary.

The set of parameters (α, β) that fulfils the condition $\alpha = \beta < \frac{1}{2}$ represents the transition line between the low and high density phase. The finite system of low number of cell shows the meltable coexistence of those phases nearby the transition line. For illustration see Figure 2.

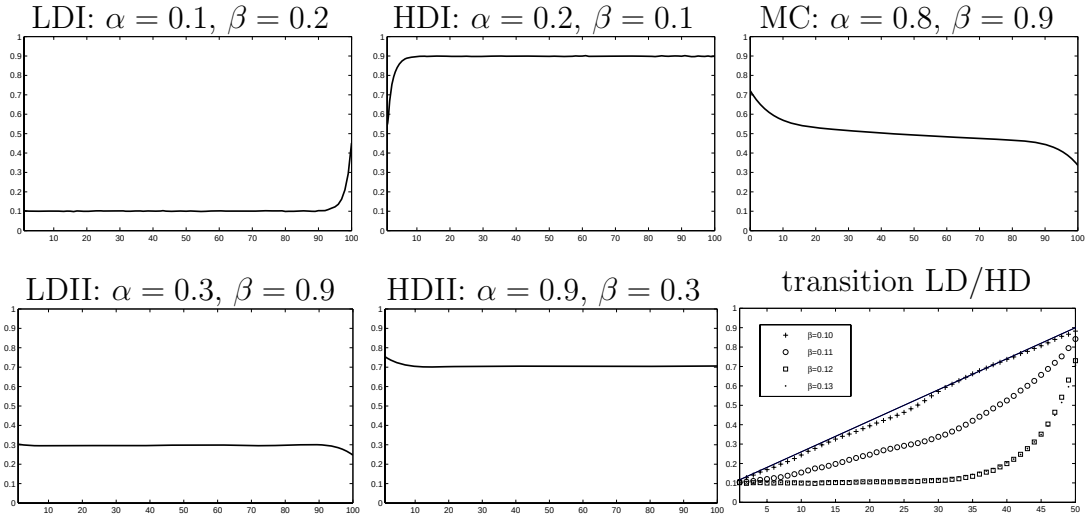


Figure 2: Density profiles in 1D for 5 representatives and nearby the transition line. LDI and LDII stands for low density phase; HDI and HDII stands for the high density phase; the MC stands for maximal current.

2 Generalization in Two Dimensions

Analogically to the one-dimensional case, we will first describe the "playground". We consider a rectangle lattice of $N \times M$ cells. Particles can move along this lattice from the left to the right and from the bottom to the top. New particles enter the system from the left (L) or the lower (D) reservoir with probability α or ε respectively, and can leave the system via the right (R) or the upper (U) reservoir with probability β or δ respectively. Every cell is denoted by its row and column index (i, j). Current state of the cell is given by the state variable $\tau_{i,j}$

$$\tau_{i,j}(t) = \begin{cases} 1 & \text{if the cell is occupied}\,, \\ 0 & \text{if the cell is empty}\,. \end{cases} \tag{3}$$

We will now define the lattice by means of the weighted oriented graph $G = (V, E, h)$. The set of vertices V is given as $V = \mathcal{M} \cup \mathcal{A} \cup \mathcal{B} \cup \mathcal{E} \cup \mathcal{D}$, where

$$\mathcal{A} = \{(i, 0) : i = 1, 2, \ldots, N\}\,, \qquad \mathcal{B} = \{(i, M+1) : i = 1, 2, \ldots, N\}\,, \tag{4}$$

$$\mathcal{E} = \{(0, j) : j = 1, 2, \ldots, M\}\,, \qquad \mathcal{D} = \{(N+1, j) : j = 1, 2, \ldots, M\}\,. \tag{5}$$

The edges are $E = E_{\mathcal{M}} \cup E_{\mathcal{A}} \cup E_{\mathcal{B}} \cup E_{\mathcal{E}} \cup E_{\mathcal{D}}$, where

$$E_{\mathcal{M}} = \big\{\big((i,j),(i+1,j)\big) : i = 1,2,\ldots,N-1, j = 1,2,\ldots,M\big\} \cup \\ \cup \big\{\big((i,j),(i,j+1)\big) : i = 1,2,\ldots,N, j = 1,2,\ldots,M-1\big\} \quad (6)$$

$$E_{\mathcal{A}} = \big\{\big((i,0),(i,1)\big) : i = 1,2,\ldots,N\big\}\ ,$$
$$E_{\mathcal{B}} = \big\{\big((i,M),(i,M+1)\big) : i = 1,2,\ldots,N\big\}\ ,$$
$$E_{\mathcal{E}} = \big\{\big((0,j),(1,j)\big) : j = 1,2,\ldots,M\big\}\ ,$$
$$E_{\mathcal{D}} = \big\{\big((N,j),(N+1,j)\big) : j = 1,2,\ldots,M\big\}\ .$$

Weighting function $h : E \to \langle 0,1\rangle$ is defined as follows

$$h(e) = \begin{cases} 1 & e \in E_{\mathcal{M}}\,, \\ \alpha & e \in E_{\mathcal{A}}\,, \\ \beta & e \in E_{\mathcal{B}}\,, \\ \varepsilon & e \in E_{\mathcal{E}}\,, \\ \delta & e \in E_{\mathcal{D}}\,. \end{cases} \quad (7)$$

Schematically is the graph depicted in Figure 3.

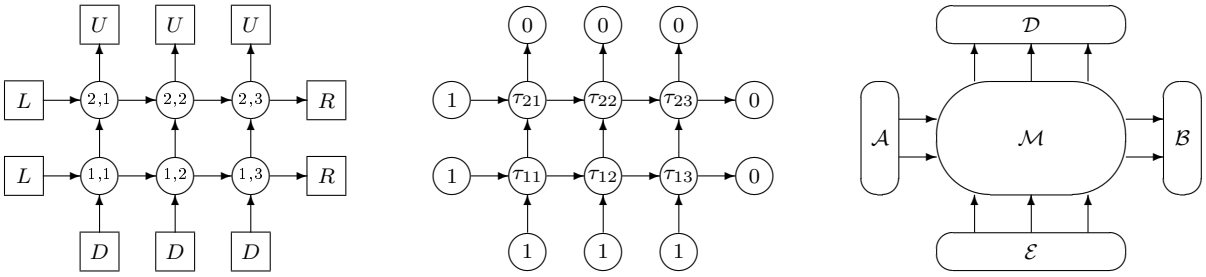


Figure 3: The "playground" for TASEP in 2D

We consider the left (resp. lower) reservoir as a set $\mathcal{A}$ (resp. $\mathcal{E}$) of always occupied cells, and the right (resp. upper) reservoir as a set $\mathcal{B}$ (resp. $\mathcal{D}$) of always empty cells. That means $\tau_{0,j}(t) = \tau_{i,0}(t) \equiv 1$ and $\tau_{N+1,j}(t) = \tau_{i,M+1}(t) \equiv 0$.

Particles move along this "playground" according following rules. An particle in vertex v chooses randomly one of the *unoccupied* neighbors u and then hops from v to u with probability $h(v,u)$. That means, if only one neighboring cell is empty, the particle tries to hop there. Only if both neighbors are empty is the particle forced to chose one of them as a target cell. The time continuous dynamics or random-sequential update can be defined analogically to the one–dimensional case.

The analogical phase transition and phase differentiation to the one dimensional case studied by means of the computer simulation for symmetrical system of $N \times N$ cells with $\alpha = \epsilon$ and $\beta = \delta$ is illustrated in Figure 4.

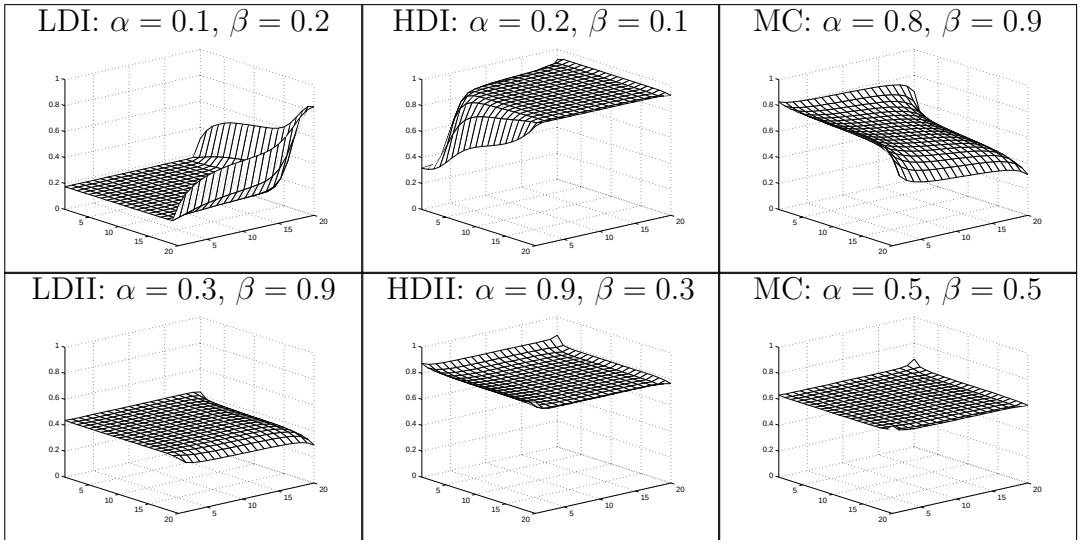


Figure 4: Density profiles in 2D for 6 representatives.

3 Phase Transition in the Floor-Field Model

Similar generalization of the TASEP has been studied in [1] by the group of K. Nishinary. In this work simple Floor Field model is considered. The paticles move along rectangular lattice by choosing one of the neighboring cell (i, j), $(i+1, j)$, $(i-1, j)$, $(i, j+1)$, $(i, j-1)$. The target cell is chosen according the transition probability p_{ij} that is determined as

$$p_{ij} = N\xi_{ij} \exp\{-k_S S_{ij} + k_D D_{ij}\}\,, \tag{8}$$

where S_{ij} is the shortest way in steps to the exit and D_{ij} is the dynamical field corresponding to the motion of other particles. k_S and k_D are sensitivity parameters to the fields. ξ is the indicator of cell availability for particles.

The article [1] studies the propagation of particles through the rectangular room with one injecting place, where particles jump in with probability α, and one exit, where particles leave the system with probability β. The simulations study shows similar behaviour as the 1D TASEP model. Again several phases can be distinguished according to the crowd occupancy of the room. For illustration see Figure 5.

4 Cellular Model of Room Evacuation Based on Occupancy and Movement Prediction

Another approach of two dimensional cellular automata modeling is the evacuation model of single room that has been introduced in [2]. The operational space of the simulation is divided in square-shaped cells with the edge length corresponding to 0.5 m. Each cell $\vec{x} = (x_{\text{column}}, x_{\text{line}})$ may be either empty or occupied by one agent, which is indicated by the *occupation number* $n(\vec{x})$, where $n(\vec{x}) = 0$ if the cell is empty and $n(\vec{x}) = 1$ otherwise. Here we note, the exit cell $\vec{e}$ is presented as always empty, keeping the rule that only one

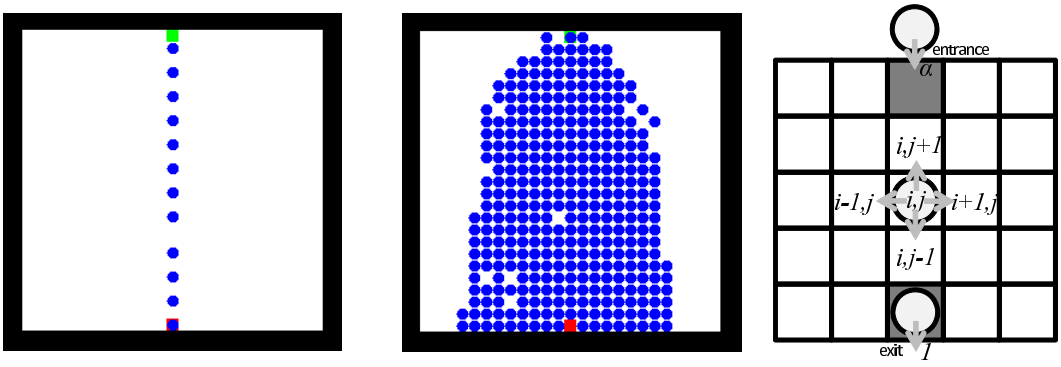


Figure 5: Two different realization of the Floor Field model simulation for different set of parameters α and β. Again, Low and High density phase can be distinguished. Taken from [1].

agent can enter the cell at the time. Each cell carries the *potential* $U(\vec{x})$ indicating the attractiveness of the cell for the agent (see [2] for details), which can be defined as

$$U(\vec{x}) = -F \cdot \varrho(\vec{e}, \vec{x}) \ , \tag{9}$$

where F is the constant determining the potential strength and ϱ is a "distance" of the cell $\vec{x}$ to the exit cell, being often chosen es Euclidian metric, i.e.

$$\varrho(\vec{e}, \vec{x}) = \left(|e_{\text{column}} - x_{\text{column}}|^2 + |e_{\text{line}} - x_{\text{line}}|^2\right)^{\frac{1}{2}} \ . \tag{10}$$

For illustration purposes the coordinates of the exit in presented Figures are set to $\vec{e} = (0, 0)$. To the static properties of the cell belongs the *cell type number* $t(\vec{x})$, which determines, whether the agent can enter the cell ($t(\vec{x}) = 1$), e.g. floor cell, exit, or not ($t(\vec{x}) = 0$), e.g. wall, barrier.

Besides the occupation number, the dynamical status of the cell is determined by the *prediction number* $r(\vec{x}) \in \{0, 1, \dots\}$, which denotes the number of pedestrians being predicted to enter the cell $\vec{x}$. As we will see in (12), the maximum number of entering agents is 8. The principle of prediction will be explained below.

The essence of the CA dynamics lies in the rules, according to which the agent chooses next target cell. In this project, the agent decides stochastically, i.e. the probability $p_{\vec{d}}(\vec{x})$ of choosing the cell $\vec{x} + \vec{d}$ from the "target" surrounding $S_T(\vec{x})$ depends on the current state of the "reaction" surrounding $S_R(\vec{x})$:

$$p_{\vec{d}}(\vec{x}) = \Pr\left\{\vec{x} + \vec{d} | S_R(\vec{x})\right\} \ . \tag{11}$$

In this article, the surrounding according to Moore's definition with range 1 is chosen for both, the target and the reaction surrounding, i.e. $S_T(\vec{x}) = S_R(\vec{x}) = \vec{x} + S_M$, where

$$S_M = \{(-1, 1); (0, 1); (1, 1); (-1, 0); (1, 0); (-1, -1); (0, -1); (1, -1)\} \tag{12}$$

Let us now denote $\vec{d}_r(i)$ the currently predicted direction of the agent i. The movement prediction from the view of the agent i then is

$$r'_i(\vec{d}) = r(\vec{x} + \vec{d}) - \delta_{\vec{d}, \vec{d}_r(i)} \ , \tag{13}$$

where $\delta_{i,j}$ is the Kronecker's symbol. For all $\vec{d} \in S_M$ the indicator $\widetilde{r}_i(\vec{d}) = \delta_{0,r'_i(\vec{d})}$ indicates, whether the cell $\vec{x} + \vec{d}$ is predicted to be entered by another agent than i. Using the notation presented above, the probability that the agent i sitting in the cell $\vec{x}$ chooses the direction $\vec{d}$ is given as

$$p_{\vec{d}}(\vec{x}) = \mathcal{N} \cdot t(\vec{x} + \vec{d}) \cdot \exp\{\alpha \cdot U(\vec{x} + \vec{d})\} \times \\ \times [1 - \beta \cdot n(\vec{x} + \vec{d})] \cdot [1 - \gamma \cdot \widetilde{r}_i(\vec{d})] \ , \tag{14}$$

where $\mathcal{N}$ is the normalization constant ensuring that $\sum_{\vec{d} \in U_M} p_{\vec{d}(x)} = 1$, and coefficients α, β, γ, are coefficients of sensitivity to the potential, occupation number, and prediction number. These parameters are to be determined later and their influence is demonstrated in Figure 6.

Subfigure A visualized wider surrounding of an agent in the cell $\vec{x}$. Integer numbers represent agents and dashed arrows their predicted movement. The probability distribution $p_{\vec{d}}(\vec{x})$ given by (14) is determined by potential, occupation and conflict prediction. The subfigure B visualizes these parameters. The darker color the higher potential (closer to exit), hatched area means penalization in stated category. The final cell attractivity strongly depends on coefficients of sensitivity to stated parameters. While potential represent static conditions, occupation and prediction of conflict reflect agent strategy. Final probabilities for different settings of sensitivity parameters β, γ are shown in subfigure C. For each of them, 2000 decisions were divided into the cells according to (14). The values of potential strength is $F = 3$, and the potential sensitivity $\alpha = 1$. The potential sensitivity plays an important role in the heterogenous system (α_i differs from agent to agent), which is not the demonstrated case.

The theoretical study of the model behaviour has been supported by an experiment performed by 86 volunteers in the study hall T-214. Non-panic egress situation has been considered and by means of this experiment, the model was calibrated. Illustration to the experiment can bee found in Figure 7. Pictures A come from frontal camera, 9 (resp. 6) seconds after initialization, when first person approaches the exit and 15 (resp. 8) seconds after initialization, when compact cluster is developed. Subfigures B project previous pictures to lattice representation and subfigures C represent corresponding realization of the simulation. One time unit of the simulation corresponds to 0,7 s. The time interval between creating the cluster and completing the evacuation was used to create the time-span of the model, because this article focuses on the shape of the cluster in front of the exit. Mean actualization frequency was set to 1 time unit.

As the system is closed, there is no phase transition observed. But the generalization in the way presented in [1] can be simply implemented. This is a motivation for further experimental study of the occurrence of phase transition in the system of pedestrians. As we believe, the behaviour of the individual is strongly influenced by the local density which can be expressed by simple rules presented in this article.

5 Conclusion

This contribution shows the transition from 1D TASEP model to two dimensional problem and discuses basic features of phase transitions in related systems of pedestrian dynamics.

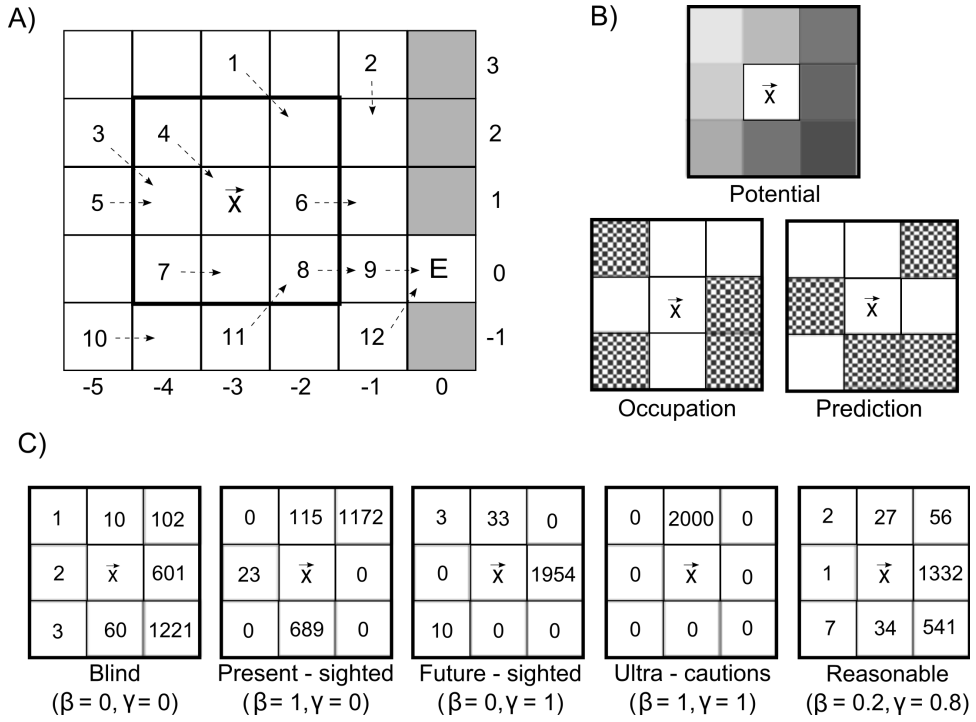


Figure 6: Example illustrating principle of decision of one pedestrian. Taken from [2]

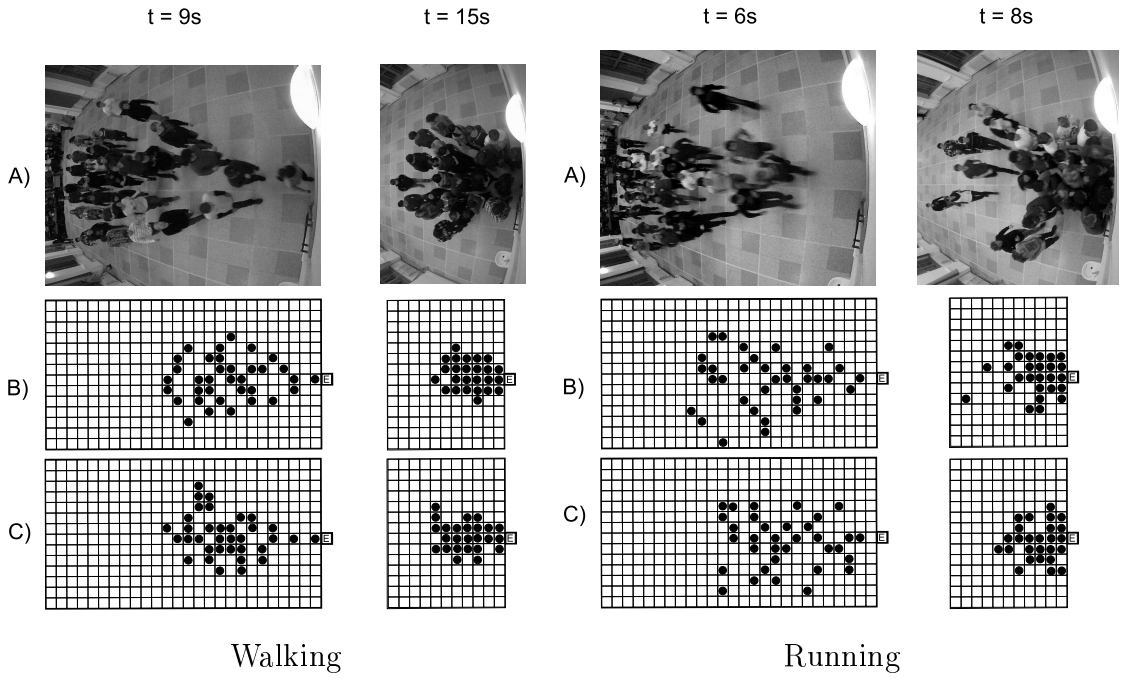


Figure 7: Visualization of progress of one round, pedestrians walked (left) and run (right). Taken from [2].

New ideas of egress simulation is presented and is supported by an experimental study.

Inspired by the work of Nishinari group, we propose an egress experiment with open injection boundary that could enable the study of phase transition in system of pedestrians in non-panic conditions. Such experiment can serve for calibration of introduced model parameters.

Furthermore, The induction of phase transition from low to high density by the microscopical changes of individual behaviour under dense or free conditions can help to understand the crowd behaviour on microscopical bases and could lead to reliable crowd movement prediction by means of real time simulations using cellular automata.

References

[1] T. Ezaki, D. Yanagisawa. *Metastability in Pedestrian Evacuation.* In 'Lecture Notes in Computer Science (Springer Verlag 2012)', G. C. Sirakoulis, S. Bandini, (eds.), volume 7495, p. 776 – 784.

[2] P. Hrabak, M. Bukacek, M. Krbalek. *Cellular Model of Room Evacuation Based on Occupancy and Movement Prediction..* In 'Lecture Notes in Computer Science (Springer Verlag 2012)', G. C. Sirakoulis, S. Bandini, (eds.), volume 7495, p. 709 – 718.

[3] P. Hrabak. *The totally asymmetric simple exclusion process in two–dimensional finite lattice, comparison of density profiles.* Proceedings of SPMS 2010 (2010), p. 91 – 100.

[4] N. Rajewski, L. Santen, A. Schadschneider, M. Schreckenberg. *The asymmetric exclusion process: comparison of update procedures.* Journal of statistical physics (1998), volume 92, p. 151-194.

Kolmogorov–Cramér Type Estimators*

Jitka Hrabáková

3rd year of PGS, email: jitka.hrabakova@fit.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 presented at Joint Meeting of y-BIS–International Young Business and Industrial Statisticians and jSPE–Young Portuguese Statisticians [1] and SPMS 2012. This contributions study properties of minimum distance estimates of unknown parameter θ_0. Two modification of well known Cramér–von Mises distance (namely generalized Cramér–von Mises Distance and Kolmogorov–Cramér distance) was defined in previous work (see [2], [3]). Statistical properties of the corresponding estimators were investigated. In the current work wide family of modification of Kolmogorov–Cramér distance is introduced. All newly defined distances are created so that the corresponding minimum distance estimators remains consistent in L_1–norm. The extensive simulation study concerning robustness was produced.

Keywords: minimum distance estimates, consistency, robustness

Abstrakt. Tento článek shrnuje výsledky prezentované na konferencích y-BIS and jSPE [1] a SPMS 2012. Tyto příspěvky se zabývají vlastnostmi odhadů s minimální vzdáleností nenámeho parametru θ_0. Dvě modifikace dobře známé Cramér–von Mises vzdálenosti (jmenovitě zobecněná Cramér–von Mises a Kolmogorov–Cramér vzdálenost) byli definovány v předchozích pracích (viz [2], [3]). Byly zkoumány statistické vlastnosti příslušných odhadů. V současné práci byla zavedena široká rodina různých zobecnění Kolmogorov–Cramér vzdáleností. Všechny nové vzdálenosti jsou definovány tak aby jim příslušné odhady zůstaly konzistentní v L_1–normě. Byla provedena rozsáhlá simulační studie robustnosti těchto odhadů.

Klíčová slova: odhady s minimální vzdáleností, konzistence, robustnost

1 Summary

We investigate minimum distance estimates based on different modification of Cramér–von Mises distance. In previous work (see [2], [3]) we defined two modification first is generalized Cramér–von Mises distance (1)

$$d_{GCM}(F,G) = \int (F(x) - G(x))^{p/q} \mathrm{d}F(x), \text{ where } p \text{ is even, and } q \text{ is odd}. \qquad (1)$$

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

There are two possibilities how to define minimum distance estimate based on GCM distance, because it is not symmetric. We can search for minimum of (2) or (3)

$$d_{GCM}(F_n, F_\theta) = \int (F_n(x) - F_\theta(x))^{p/q} \mathrm{d}F_n(x) = \frac{1}{n}\sum_{i=1}^{n} (F_\theta(x_i) - F_n(x_i))^{p/q} \tag{2}$$

$$d_{GCM}(F_\theta, F_n) = \int (F_\theta(x) - F_n(x))^{p/q} \mathrm{d}F_\theta(x) \tag{3}$$

$$= \frac{q}{p+q} \sum_{i=1}^{n} \Big[\Big(F_\theta(x_i) - \frac{i-1}{n} \Big)^{\frac{p+q}{q}} - \Big(F_\theta(x_i) - \frac{i}{n} \Big)^{\frac{p+q}{q}} \Big] . \tag{4}$$

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

$$d_i(F_n, F) = (F_n(x_i) - F(x_i))^{p/q} \qquad \text{for } i = 1, \dots, n\,, \tag{5}$$

$$d_{2n+1-i}(F_n, F) = (F_{n-}(x_i) - F_-(x_i))^{p/q} \qquad \text{for } i = 1, \dots, n\,, \tag{6}$$

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

$$d_{KC}(F_n, F) = \frac{1}{m}\sum_{i=1}^{m} d_{(i)}(F_n, F)\,, \tag{7}$$

where $d_{(i)}(F_n, F)$ denotes decreasingly ordered sequence of $(d_i(F_n, F))_1^{2n}$ and m is an integer less or equal to $2n$. Moreover the parameter m can depend on the sample size n. If the parameter m is fixed than the KC estimate is consistent of the order $n^{-1/2}$ in L_1–norm. In case that the parameter m is $O(n^\beta)$ and $\beta \leq \frac{p}{2q}$ then the KC estimate is consistent of the order $n^{\frac{1}{2} - \frac{\beta q}{p}}$. In [2] and [3] are compared properties (robustness and consistency) of this two newly defined estimators with Kolmogorov and original Cramér–von Mises estimator. Current work introduce wide class of modifications of Kolmogorov–Cramér (KC) distance by implementing data based weight functions, random selecting of differences to be summed up and using various coefficient modification. According to the definition of KC distance (7), following class of distances is defined:

$$KC^j(F_n, F) = \frac{1}{km} d_{(1)} + \frac{1}{m^{j+1}} \sum_{i=2}^{m} i^j d_{(i)}, \quad j \in \{0, 1, \dots\},\, k \in \mathbb{R}^+, \tag{8}$$

and parameter m can be either fixed or dependent on sample size n but in both cases less then $2n$. The class is defined so that the corresponding minimum distance estimates remains consistent in L_1–norm. The order depends on choice of parameter m similar as for KC distance for m arbitrary fixed the estimate is consistent of the order $n^{-1/2}$ in L_1–norm. If the parameter m is $O(n^\beta)$ and $\beta \leq \frac{p}{2q}$ then the KC estimate is consistent of the order $n^{\frac{1}{2} - \frac{\beta q}{p}}$. This shape of distance suppresses the influence of Kolmogorov estimate $(d_{(1)})$ by choosing constant k big enough, and contemporary increases the influence of smaller differences for choice $j > 0$. In general, the smaller influence of Kolmogorov

estimate is the more robust estimate we gain. But there are more parameters influencing robustness. For fixed m the influence of power p/q is the same as for KC estimate and the choice $j = 2$ seems to by optimal. Situation significantly differs if m depends on sample size n. We have explored two situations $m = 2n^{\beta}$, $\beta \leq \frac{p}{2q}$ and $m = f \cdot n$, $0 < f < 1$. In both cases the influence of power p/q is weakened by impact of parameters m and j.

Further, random variant of class (8) is defined by taking index $i = [1+2nu_i]$ where $u_i \sim U(0,1)$ is uniformly distributed random variable on $(0,1)$. Results for these random KC_j have similar properties as the non-random form. The biggest impact has the parameter j, but it strongly depends on choice of parameter m. In all investigated situations the theoretical order of consistency is the same, dependent on choice of parameter m. And as simulation shows the real order of consistency coincides with the theoretical one, however, the value of L_1–norm strongly depends on choice of constant k. The bigger the constant is the bigger is the L_1–norm. From this follows that in applications constant k could be chosen very large only for sample size big enough.

References

[1] J. Hanousková and V. Kůs. *Simulation study for robustness and consistency of minimum distance density estimates under physical data framework* Join Meeting of y-BIS–International Business and Industrial Statisticians and jSPE–Young Portugees Statisticians (2012), 168–170

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

[3] J. Hanousková and V. Kůs. *Generalized Cramér–von Mises distance estimators*. Proceedings of SPMS (2011), 73–81.

Requirements Engineering and Project Management

Radek Hřebík

1st year of PGS, email: Radek.Hrebik@seznam.cz
Department of Software Engineering in Economics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. This contribution created for the Proceeding of the workshop Doktorandské dny deals with one development phase of software development – requirements engineering. In introduction is attention paid to requirements analysis and international standards regulating the requirements currently. Special attention is also paid to the current situation of the discipline and its potential benefits in the future. For the evaluation of requirements must be possible to measure them – metrics are used for this propose. Possibilities of utilization of knowledge in Requirements Engineering are inflected in many other areas. In this contribution it is aimed at the improvement of project management methods.

Keywords: requirement, metrics, project management

Abstrakt. Tento příspěvek vytvořený do sborníku workshopu Doktorandské dny se zabývá jedním z vývojových stádií vzniku softwaru a to tzv. Requierements Engineering. V příspěvku je nejprve věnována pozornost úvodu do analýzy požadavků a mezinárodním standardům, které v současné době požadavky nějakým způsobem upravují. Pozornot je věnována rovněž situaci týkající této disciplíny a jejím možným přínosům v budoucnosti. Pro hodnocení požadavků musí existovat možnost jejich měření – pro tento účel se používají metriky. Možnosti využití poznatků z oblasti requirements engineering jsou skloňovány v mnoha dalších oblastech. V tomto příspěvku je věnována pozornost především využití metrik spojených s požadavky ke zlepšení vybraných metod projektového řízení.

Klíčová slova: požadavek, metriky, projektové řízení

1 Introduction

The term requirements engineering consists from two separate terms – requirements and engineering. The first of them informs what is required, what should be fulfilled. The second term is engineering which implies the applying of scientific knowledge to ensure that requirements are exactly as they should be. With the term requirements engineering one often encounters in software developing process. So the obvious definition of requirements engineering is discipline based on understanding software requirements. Another possible definition comes from Laplante [23] and defines requirements engineering as process of eliciting, analyzing, documenting, validating and managing requirements. Requirements engineering represents the early phase of the software life-cycle, but should not be fully

omitted in later phases. Requirements affect the whole process and can be changed. The theme of life-cycle in connection with requirements is detail discussed in [9].

Naturally it has to be confirmed with Laplante [23] that requirements have to be documented. Only documented requirements can lately serve as an evidence. Requirements document has to be processed and readable. To test that requirements are implemented correctly, they should be clear, precise and unambiguous.

There are many kinds of requirements and more kinds of division are possible. One of commonly used division presented for example in [24] is on functional and non-functional. Functional requirements talk about provided services and how they should work. Non-functional requirements primary inform about constraints of the system. Non-functional requirements are discussed in [3] where the treatment with them for future directions is mentioned. It is also possible to talk about architectural, structural and behavioural requirements. Other division is into user and system requirements. As Sommerville says in [24] user requirements are understandable by system users without detailed technical knowledge and system requirements explain how the user requirements should be provided by the system.

The whole system is based on requirements. If they are not well understood the system will not meet expectations and the final version of program will be probably not delivered on time and the costs will be much higher than originally expected. Result is then dissatisfied client. Is there any worse advertisement than dissatisfied clients? Clients may not use facilities of new system simply because they wanted something else. The system may not be badly implemented, but if there is no meeting with client requirements, the system becomes unusable. If such system continues in use, the costs of fixing errors may be very high as discussed for example in [27]. All these problems can be prevented by using knowledge of requirements engineering.

1.1 Requirements

The indisputable need to express requirements leaded to formation of international standards. These standards help to formulate and understand requirements. The standards are needed for reliably and correctly equipment and for making connection between various kinds of equipment. There are two main international organizations focusing on this issue, it is the Institute of Electrical and Electronics Engineers (IEEE) and International Standards Organization (ISO).

The first institute affects requirements engineering directly with two main standards. The first one is standard IEEE 1233-1996 Guide for Developing System Requirements Specifications [12]. It provides very good guidance in working with requirements. The second one is IEEE 830-1998 Recommended Practice for Software Requirements Specifications [13]. The definition of requirement can be found in IEEE 1220-2005 Standard for Application and Management of the Systems Engineering Process [14]. Similar definition from previous version of this standard (IEEE 1220-1998) is in detail discussed in [9].

Talking about ISO standards needs to mention the ISO 9000 process improvement models [2] and the ISO 9001 standard representing standard for quality management systems [4]. In case of model it is also used standard ISO/IEC 15504. Software process assessment and improvement is also mentioned in international standards. It is called

Capability Maturity Model (CMM). While CMM is rather American standard, European version of analogous standard is called BOOTSTRAP [2]. Of course, as mentioned above, the requirements have to be documented. Naturally there are specifications and recommendations what should the documentation contain. It is also included in standard IEEE 830-1998 where the division is recommended into five parts. These are introduction, general description, specific requirements, appendices and index [24].

Of course nothing is fully ideal, but the existence of standards can in most cases only be only helpful and avoid problems. Summary of standards connected with software engineering processes presents also Laplante in [23]. The summary tends to be the cornerstone of the whole requirements research. To do research on requirements engineering without a clue about these standards is something almost impossible.

2 Potential of Requirements Engineering

Natural interest in requirements engineering is of course not only in science. In the following, attention is paid mainly to the software production. Attention is given to the progress in this field in recent years. It seems it really works in practice and the requirements affect not only the area of software, but also many other.

It is very important to mention the results of The Standish Group. The study presented in report from The Standish Group and published in 1995 showed that 31.3 % of software projects is cancelled before they get completed and 52.7 % of projects cost 189 % of their original estimates. Factor number one why the project is impaired are incomplete requirements. [25] In year 2004 a PhD thesis devoted to defects in software and showed that 44 % to 80 % of all defects were inserted in the requirements phase [5].

It is frequented question how much does it cost to fix errors. In answering this question it is very good to mention the study of National Aeronautics and Space Administration (NASA) from 2004. This study puts together data from nine studies that have been performed to determine the software error costs factors and makes the cost data normalized to determine the software error cost factors for each study, along with the overall mean and median values for each life-cycle phase. The result is that if we take the median values then fixing errors costs in design phase 7.3 times, in coding phase 25.6 times and in test phase 177 times more than in phase of requirements. [27] From year 2004 comes also a study devoted to benefits of requirements engineering process improvement at the Australian Center for Unisys Software [4].

One of the last report from The Standish Group comes with optimistic conclusion and indicates some improvement. There is a marked increase in project success rates from 2008 to 2010. These numbers represent an up-tick in the success rates from the previous study, as well as a decrease in the number of failures [26].

Requirements engineering is primary used for working with software requirements. But due to the work with requirements, software engineering knowledge can be evaluated in many areas of human activity. Although the research has progressed significantly in recent years, the situation about requirements is still not fully satisfactory and there is still what to improve. The main field of research will be discussed in next section dealing with metrics. When something should be evaluated, there has to be a kind of evaluation. It is talked about requirements metrics. The research on the requirements metrics and

their use in project management is also supported by what says Kerzner in [18] that only in the last several years has been developed models for measuring the metrics to determine the value on a project. So there is still what to improve and research opportunities are still very wide.

3 Metrics

However it could be perceived as something automatic it is always better to pay special attention to metrics. There have to be special metrics that can inform how good or bad the requirement is. The requirements affect the quality of the whole project. Without any measureable criterions it is not possible to decide about requirements. Metrics can inform what the requirements mean for the project, they can be evaluated and compared. Measurement is the key to improvement, in this case, it is the way to improve the software process. Metrics are also commonly used in software developing process and it is talked about process, product and resources metrics. Requirement metrics play key role in identifying potential project risks. In the requirements phase the metrics show how the new application should be tested. Multiple metrics are needed for comprehensive evaluation. When a potential problem is identified at early development phase the costs are much lower than in the later development phases.

One of the papers talking in detail about requirements engineering and metrics was presented in March 2011 at National conference in India [2]. This paper describes commonly used requirements metrics in detail. There are distinguished the volatility, completeness, traceability and size metrics. It is not possible to say that one of the kind is better than other. The metric selection depends on actual situation and the measurement purpose.

Size metrics are very important and general metrics used not only as requirements metrics, but also as software metrics. They are intuitive and representing one of the simplest metrics. The principle is for example count of lines of code. As code are taken lines with executable commands, executable statements and data declarations. Traceability means the ability to trace requirements in a specification to their origin from higher level to lower level requirements in a set of documented links. Requirements completeness metrics are used to assess whether a requirement is at wrong level of hierarchy or too complex. Volatility of requirements informs us about the degree of requirements changes over a time period [2]. Volatility is checked to know whether the changes are consistent with current development activities. The high degree of volatility indicates changes such as addition, deletion and modifications. Volatility is commonly high in the initial phase of development and as the project flows, the volatility should be reduced so that further development should not be affected.

To illustrate the use of some requirements metrics is nothing better than show some examples. The first metric to show is a size metric that controls unambiguous. The aim is to obtain the percentage of unique requirements. Unique means that they have been identified in a unique manner by reviewers. Metric is expressed as a ratio of number of requirements with the same interpretation to number of total requirements. The interpretation of metric values is obvious, values close to zero indicate ambiguous requirements and for values close to one it is evident that requirements are unambiguous. As second

example can serve metric for measuring of precision that informs about providing a minimum time for acknowledgement. Expression of this metric is a ratio of number of true positive to sum of true and false positives. To one of easily interpretable metric belongs understandable expressed also as fraction. The numerator is number of requirements understood by all reviewers and denominator is count of requirements. Interpretation is that values close to one indicate that all requirements where understood and zero value indicates that no of the given requirements was understood by reviewers. [2]

The number of existing metrics cannot be fixed, because there exist a lot of metrics and some of them can be also named different. The main is to select the right metric for particular measurement. There does not exist any universal metric for all kinds of projects. Metric collection manually is very lengthy and errors caused by human factor are bigger than in case of using a special tool for collecting metrics. Using management tools represents cheaper, faster and more reliable solution.

3.1 Automated Measurement

In some publications, for example in [2], is mentioned The Automated Requirements Measurement tool developed at the National Aeronautics and Space Administration (NASA). This tool is working with natural language. Using such tool helps to write requirements right but not to write right requirements. To this tool from NASA pursues also Laplante in his book [23]. But when searching actual information about this tool there is a problem because do not exist any relevant information about this tool at the NASA official pages. Maybe this project was stopped because austerities. But despite the current situation it is good to know, that the NASA is also resolving requirements metrics. The NASA research is also mentioned in [27].

The theme of metrics is very popular and with a high probability it will be soon not otherwise. Measurement gained unassailable position in software developing process and no wonder that requirements engineering is no exception. It gives a lot of papers talking about this theme. Some overview is presented by Monperrus in [20]. In this publication are pieces of research defining requirements metrics. There is also described how the used measurement tool was prepared. The paper was prepared on the base of many researches and makes very good summary.

Measuring is provided also by tools determined to manage requirements. Some of tools used for requirements engineering are mentioned in the following. They are often determined for a widely range of use and they can serve in a whole developing process. The first selected tool is Jama Contour representing web application which helps users to manage the requirements. On the pages of selected product [16] can be found a report giving information not only about software functions but also general about requirements engineering. The mentioned report ([17]) comes from year 2011 and informs also about the statistics connected with requirements and their right using. The IBM Rational Dynamic Object Oriented Requirements System (DOORS) represents requirements management tool for systems and advanced information technology applications. The tool represents a leading requirements management software product and promises quality improvement by better communication and collaboration within team [10]. Last software to mention comes also from IBM and it is Rational RequisitePro. This requirements management

tool also helps project teams writing and managing good requirements [11].

Range of offered tools is currently very wide and to make summary needs a special research on this topic. As standard is available textual description, user model diagrams and object models. The specific kind of input can differ but currently the tools are very similar in this regard. The main different is in user interface and application design. The facilities not only for elicitation but also for requirements analysis and validation are also offered. It could be said that in most cases the tools can manage requirements quite well. The requirements can also be represented in natural language. Currently, in my opinion, the national language support is miserable. The development of requirements tool, especially the support for treatment with national language requirements, seems to be a big challenge for future development.

4 Project Management and Requirements

The requirements engineering process is used in first phase of product (software) life-cycle. The whole software development process is also question of requirements. The development is firstly by someone required. Somebody needs the software and starts the project of developing required software. The development project is project to be managed. So, is it not a shame to use requirements engineering knowledge just in software developing process? Logically, there is a possibility to use the requirements engineering knowledge in project management. In the following text, attention is paid to the proposals in this area of current research.

About the need of measuring requirements and the automated measurement program helping software project managers to assess progress, mitigate risks, and improve team productivity is talked for example in [15]. There was concluded that requirements engineering is a critical phase for the successful achievement of project objectives. As long as requirements engineering activities are not seen as a tangible outcome in a project, they are likely to be neglected in favor of project activities with tangible (software) products. In [15] is also talked about giving greater emphasis to framing requirements engineering activities and the results from these activities as a desirable and valuable outcome of large research projects.

The need of requirements in case of project management is also discussed for example in [1]. The problem of the right requirements and project management can be also found in [21]. In [8] it is argued that the technical management of the project can take advantages of requirements engineering activities to organize and integrate project activities. The project management is also about time scheduling and there can be seen the main partition recovery. There exist a lot of methods for project management. Because the need of requirements in project management is discussed very often it is appropriate to mention some of the project management methods and discuss the use of requirements metrics.

Very often method used is in project management Critical Path Method (CPM). CPM works with constant times of tasks, the method is based only on deterministic task duration. The principle of method is a critical path. Any delay on this path will cause delay of the whole project. At first sight the requirements metrics do not seem to be beneficial because of inflexible tasks duration. It deals with approaches to task duration

and about a set of performance metrics in [22].

Probably at the same time as CPM was presented other method called Program Evaluation and Review Technique (PERT). This method works with three times for each task. So there is not only deterministic approach as in CPM. The most likely time is supplemented with optimistic and pessimistic time. From these three times is evaluated expected time, which is more realistic. With the values of expected time is worked in using CPM. The method still does not work with flexible time reserves and it is not entirely clear how to use requirements metrics. But as expected and shown in [19] the requirements metrics are also to be utilized.

The main change which came in project management after already mentioned methods is Theory of Constraints [7]. The theory is based on a claim that every achieving is influenced by constraining processes. The project management method affected with Theory of Constraints is so called Critical Chain Method (CCM) which works with something like flexible task duration. There are used some buffers and the tasks reserves with flexible time duration. Critical Chain Method is a schedule network analysis technique that modifies the project schedule to account for limited resources. It mixes deterministic and probabilistic approaches to schedule network analysis. The critical chain concept was coined by Goldratt, the author of Theory of Constraints. [6] The exploitation of requirements metrics is of course possible in previous methods. But in case of CPM and PERT there is no space in algorithms of methods to implement the decision phase based on requirements metric. However, it offers the CCM thanks to flexible tasks.

4.1 Critical Chain Method

Critical Chain Method (CCM) is a schedule network analysis technique that takes account of task dependencies, limited resource availability and so named buffers. Buffers are used to catch possible time to the end of the task. First step in this method is identifying set of tasks that makes the longest path to the end of the project. These tasks are called critical chains. The tasks that form critical chain create in most cases longer path than using CPM schedule. The reason is simple because critical chain tasks include resources. Resources that are used in critical chain are critical resources. Set of tasks that are not included in critical chain but converge to critical chain are feeders. The main principle of the method is based on so-called buffer management.

There are two main kinds of buffers in the project it is project buffer and feeding buffer. Project buffer means the time reserve from the critical chain to end of project. The feeding buffer is the time reserves of other task which leads to the end of critical chain before the ending time of chain. Naturally fundamental question is how to determine size of buffer. This and more about critical chain you can find in book from Goldratt called Critical Chain in which this method was introduced. [6]

The critical chain sets stretch targets for every task duration and thanks to it is the effect of major improvement in task delivery times [28]. So CCM with its buffer management seems as one of the methods that could be improved by using requirements engineering knowledge. This method uses buffers that are not constant length and have not to be used in the final project schedule. If after each step of method, when the task ends, comes as new input some requirements metric, the metric result can serve as

decision what to do with the time stored in the buffer. If the metric results are values that fulfill some criterions, for examples some compliance ratio, it is possible to continue and buffer is not used. But when the metric after some task shows that results are poor, it is possible to use the buffer to save time and costs in future. Fixing errors in early phase means the lower costs. Thanks to buffers, especially feeding buffers, it does not necessarily have to mean the longest schedule and despite this fact it can be very useful because it will prevent the larger losses in the future without any project delay. The use of specific metrics is for further discussion same as the concrete way of implementation to CCM. The proposal for research in project management improvement is based on numerous studies mentioned above.

5 Conclusion

This paper deals with requirements engineering, requirements standards and requirements metrics. The existence of international standards is not surprising because of living in almost standardized world. The main aim of this contribution is to discuss the possible research on the improvement of project management methods by the knowledge of requirements engineering. It was shown the possibility of using requirements metrics to improve scheduling with critical chain method. The main reason for choosing this method is the buffer management. The time reserves are flexible in this case. The time that is given as reserve is not tightly specified but varies. The variability in connection with requirements metrics is the possible way how to improve project schedule and quality of development. Metrics give an advice whether use or not use the time reserves that are available at the moment. This is the way not only to use requirements metrics in time scheduling but this can cause also some savings, because the earlier work with requirements – the cheaper work. The paper focuses on the potential improvement of critical chain method in its buffer management. It was also shown that the strength of requirements metrics can be used in all project management methods. The way of applying requirements metrics depends on future research and at this point seems the critical chain method as the most likely choice.

References

[1] G. Abudi. *Project Managing Business Process Improvement Initiatives.* In: Project Managing Business Process Improvement Initiatives, (2011). [online]. [cited 2012-08-15]. `http://www.bptrends.com/publicationfiles/10-04-2011-ART-Project%20Managing%20Business%20Process%20Improvement%20Initiatives-Abudi-FINAL.pdf`.

[2] M. Bokhari, S. Siddiqui. *Metrics for Requirements Engineering and Automated Requirements Tools.* Computing for Nation Development (2011).

[3] L. Chung, J. do Prado Leite. *On Non-Functional Requirements in Software Engineering.* In: Conceptual modeling: foundations and applications (2009), 363–379.

[4] D. Damian, D. Zowghi, L. Vaidyanathasamy, Y. Pal. *An Industrial Case Study of Immediate Benefits of Requirements Engineering Process Improvement at the Australian Center for Unisys Software*. Empirical Software Engineering 9, (2004), 45–75.

[5] A. Eberlein. *Requirements Acquisition and Specification for Telecommunication Services (PhD Thesis)*. University of Wales, Swansea (1997).

[6] E. Goldratt. *Critical Chain*. The North River Press, Great Barrington (1997).

[7] E. Goldratt. *Theory of Constraints*. North River Press, Great Barrington (1999).

[8] S. Gürses, M. Seguran and N. Zannone. *Requirements engineering within a large-scale security-oriented research project: lessons learned*. Requirements Engineering (2011). [online]. [cited 2012-08-15]. `http://www.springerlink.com/index/10.1007/s00766-011-0139-7`.

[9] E. Hull, K. Jackson, J. Dick. *Requirements Engineering*. Springer, London (2010).

[10] IBM. *Integrate requirements and change management with IBM Rational software*. IBM (2010). [Online]. [Cited: 2012-08-24]. `http://public.dhe.ibm.com/common/ssi/ecm/en/rad14034usen/RAD14034USEN.PDF`.

[11] IBM. *Rational RequisitePro*. IBM (2011). [Online]. [Cited: 2012-08-24]. `http://www-01.ibm.com/software/awdtools/reqpro/`.

[12] IEEE Standards Association: 1233-1996 - IEEE Guide for Developing System Requirements Specifications, `http://standards.ieee.org/findstds/standard/1233-1996.html`.

[13] IEEE Standards Association: 830-1998 - IEEE Recommended Practice for Software Requirements Specifications, `http://standards.ieee.org/findstds/standard/830-1998.html`.

[14] IEEE Standards Association: 1220-2005 - IEEE Standard for Application and Management of the Systems Engineering Process, `http://standards.ieee.org/findstds/standard/1220-2005.html`.

[15] D. Ishigaki. *Effective management through measurement*. IBM (1994). [online]. [cit. 2012-09-29]. `http://www.ibm.com/developerworks/rational/library/4786.html`

[16] Jama Software. *The agile way to communicate requirements and manage complex projects*. Jama Software (2012). [Online]. [Cited: 2012-08-31]. `http://www.jamasoftware.com/contour/`.

[17] Jama Software. *State of Requirements Management 2011*. (2011). [Online]. [Cited: 2012-08-31]. `http://www.jamasoftware.com/media/documents/State_of_Requirements_Management_2011.pdf`.

[18] H. Kerzner. *Project Management Metrics, KPIs, and Dashboards: A Guide to Measuring and Monitoring Project Performance*. Wiley, (2011).

[19] S. Malathi, S.Sridhar. *Analysis of Size Metrics and Effort Performance Criterion in Doftware Cost Estimation.* In: Indian Journal of Computer Science and Engineering, 3(1), (2012), 24-31.

[20] M. Monperrus, et al. *Automated Measurement of Models of Requirements.* (2011).

[21] K. Muppavarapu. *Innovative Quality Measurement System: Ideas for a Project Manager.* (2011). [online]. [cited 2012-09-29]. `http://www.pmi.org/~/media/PDF/Knowledge-Shelf/Muppavarapu_2011.ashx`.

[22] P. Pocatilu, M. Vetrici. *M-applications Development using High Performance Project Management Techniques.* In: Proceedings of the 10th WSEAS Int. Conference on Mathematics and Computers in Business and Economics. Stevens Point, (2009), 123-128.

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

[24] I. Sommerville. *Software Engineering 8th edn.* Addison-Wesley (2006).

[25] The Standish Group. *Chaos.* The Standish Group (c1995).

[26] The Standish Group. *New Standish Group report shows more projects are successful and less projects failing.* The Standish Group (2011).

[27] J. Stecklein, et al. *Error Cost Escalation Through the Project Life Cycle.* In: NASA Technical Reports Server. [online]. [cited 2012-08-21]. `http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100036670_2010039922.pdf`.

[28] P. Weaver. *Why Critical Path Scheduling is Wildly Optimistic.* (2011). [online]. [cited 2012-09-29]. `http://www.mosaicprojects.com.au/PDF_Papers/P117_Why_Critical_Path_Scheduling_is_Wildly_Optimistic.pdf`.

Entopy Estimates of 3D Brain Scans*

Václav Hubata-Vacek

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

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

Abstract. This article deals with generalized definition of entropy to evaluate Hartley, Shannon, and Collision entropies. These methods were tested and used for recognition of Alzheimer's disease, using relationship between entropy and fractal dimension to obtain fractal dimensions of 3D brain scans. Because estimated entropies from limited data set are always biased, it is applied Miller and Harris estimations of Shannon entropy, which are well known bias approaches on Taylor series. Moreover, these estimates were improved by Bayesian estimation of individual probabilities.

Keywords: entropy, fractal dimension, Alzheimer's disease, boxcounting, Rényi entopy

Abstrakt. Článek se zabývá zobecňenou definicí entropie k vyhodnocení Hartleyovy, Shannonovy a Collision entropie. Vzhledem k vztahu mezi entropií a fraktalní dimenzí byly tyto metody použity pro výpocet fraktalní dimenze 3D snímku mozku a následně jsou testovány a použity k rozpoznání Alzheimerovy choroby. Jelikož odhad entropie z omezeného počtu reálných dat je podhodnocen jsou aplikovány Millerovy a Harrisovy odhady Shannonovy entropie, což jsou odhady zalozené na Taylorově řadě. Navíc jsou tyto odhady vylepšeny o Bayesovský odhad jednotlivých pravděpodobností.

Klíčová slova: entropie, fraktální dimenze, Alzheimerova choroba, boxcounting, Rényho entropie

1 Introduction

Before explaning the realtionship between entropy and dimension, we have to introduce the term of dimension. Let $d \in \mathbb{N}$ be dimension of Euclidean space where d-dimensional unit hypercube is placed. Let $m \in \mathbb{N}$ be resolution and $a = 1/m$ be edge length of covering hypercubes of the same dimension d. The number of covering elements is given by

$$N = \mathrm{N}(a) = a^{-D}. \tag{1}$$

The knowledge of N for fixed a enables direct calculation of hypercube dimension according to

$$\ln \mathrm{N}(a) = -D \ln a \tag{2}$$

$$D = \frac{\ln \mathrm{N}(a)}{\ln \frac{1}{a}}. \tag{3}$$

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

Very popular *boxcounting method* [1] is based on the generalization of (3) to the form

$$\ln \mathrm{N}(a) = A_0 - D_0 \ln a \tag{4}$$

and its aplication to boundary of any set $F \subset \mathbb{R}^d$. As will be shown in the next chapter, the quantity $\ln N(a)$ is an estimate of *Hartley entropy*.

2 Rényi Entropy

Using natural logarithm instead of binary one, we can follow in the definition of *Rényi entropy*. Let $k \in \mathbb{N}$ be number of events, $p_j > 0$ be their probabilities for $j = 1, ..., k$ satisfying $\sum_{j=1}^{k} p_j = 1$, and $q \in \mathbb{R}$. We can define *Rényi entropy* [2] as

$$H_q = \frac{\ln \sum_{j=1}^{k} p_j^q}{1-q}, \tag{5}$$

which is a generalization of *Shannon entropy*. With respect of q, we obtain specific entropies

- *Hartley entropy* [3] for $q = 0$ as

$$H_0 = \ln \sum_{p_j > 0} 1 = \ln \sum_{j>0}^{k} 1 = \ln k = \ln \mathrm{N}(a) \tag{6}$$

- *Shannon entropy* [4] for $q \to 1$ as

$$H_1 = \lim_{q \to 1} H_q = -\sum_{j=1} p_j \ln p_j \tag{7}$$

- *Collision entropy* [2] for $q = 2$ as

$$H_2 = -\ln \sum_{p_j > 0} p_j^2 \tag{8}$$

Resulting theoretical entropies can be used for definition of *Rényi dimension* [2] as

$$D_q = \lim_{a \to 0+} \frac{H_q}{\ln \frac{1}{a}}, \tag{9}$$

which corresponds to relationship

$$H_q \approx A_q - D_q \ln a \tag{10}$$

for small covering size $a > 0$.

3 Entropy Estimates

There are several approaches to entropy estimation from experimental data sets. Supposing the experiment number $n \in \mathbb{N}$ is finite, we can count the events and obtain $n_j \in \mathbb{N}_0$ as event frequencies for $j = 1, ..., k$. The first approach to entropy estimation is *naive estimation*. We directly estimate k and p_j as

$$k_{\mathrm{N}} = \sum_{n_j>0} 1 \leq k \tag{11}$$

$$p_{j,\mathrm{N}} = \frac{n_j}{n}. \tag{12}$$

These biased estimates produce also biased entropy estimates

$$H_{0,\mathrm{N}} = \ln k_{\mathrm{N}} \tag{13}$$

$$H_{1,\mathrm{N}} = -\sum_{n_j>0} p_{j,\mathrm{N}} \ln p_{j,\mathrm{N}} \tag{14}$$

$$H_{2,\mathrm{N}} = -\ln \sum_{n_j>0} p_{j,\mathrm{N}}^2. \tag{15}$$

The second approach is based on *Bayesian estimation* of probabilities p_j as

$$p_{j,\mathrm{B}} = \frac{n_j + 1}{n + k_{\mathrm{N}}}. \tag{16}$$

This technique is called here *semi-Bayesian estimation*. We obtain another, but also biased, entropy estimates

$$H_{1,\mathrm{S}} = -\sum_{n_j>0} p_{j,\mathrm{B}} \ln p_{j,\mathrm{B}} \tag{17}$$

$$H_{2,\mathrm{S}} = -\ln \sum_{n_j>0} p_{j,\mathrm{B}}^2. \tag{18}$$

The estimate $H_{2,\mathrm{S}}$ can be improved as

$$H_{2,\mathrm{S2}} = -\ln \sum_{n_j>0} u_j, \tag{19}$$

where $u_j = \frac{(n_j+2)(n_j+1)}{(n+k_{\mathrm{N}}+1)(n+k_{\mathrm{N}})}$ is bayesian estimate of p_j^2. *Direct Bayesian estimate* of H_1 was also calculated as

$$H_{1,\mathrm{B}} = -\sum_{i=1}^{k_{\mathrm{N}}} \frac{n_i + 1}{n + k_{\mathrm{N}}} \left(\psi(n_i + 2) - \psi(n + k_{\mathrm{N}} + 1)\right), \tag{20}$$

where ψ is digamma function.

4 Bias Reduction

Miller [5] modified naive estimate $H_{1,\mathrm{N}}$ using first order Taylor expansion, which produces

$$H_{1,\mathrm{M}} = H_{1,\mathrm{N}} + \frac{k_{\mathrm{N}} - 1}{2n}. \tag{21}$$

Lately, Harris [5] improved the formula to

$$H_{1,\mathrm{H}} = H_{1,\mathrm{N}} + \frac{k_{\mathrm{N}} - 1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{p_j>0}\frac{1}{p_j}\right) \tag{22}$$

From the theoretical point of view, it is prohibited to estimate p_j by its estimates. But we try to investigate biased estimates of H_1 in the forms

$$H_{1,\mathrm{HN}} = H_{1,\mathrm{N}} + \frac{k_{\mathrm{N}} - 1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{n_j>0}\frac{1}{p_{j,\mathrm{N}}}\right) \tag{23}$$

$$H_{1,\mathrm{HS}} = H_{1,\mathrm{N}} + \frac{k_{\mathrm{N}} - 1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{n_j>0}\frac{1}{p_{j,\mathrm{B}}}\right) \tag{24}$$

$$H_{1,\mathrm{HB}} = H_{1,\mathrm{N}} + \frac{k_{\mathrm{N}} - 1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{n_j>0} r_j\right), \tag{25}$$

where $r_j = \frac{n+k_{\mathrm{N}}-1}{n_j}$ is Bayesian estimate of $\frac{1}{p_j}$.

5 Methodology of Estimation

Naive, semi-Bayesian, Bayesian and corrected entropy estimates were subject of testing on 2D and 3D structures with known Hausdorff dimension. The list of involved estimates is included in Tab. 1. Sierpinski carpet with $D_q = 1.8928$ for any $q \geq 0$ of size 81×81 is a typical 2D fractal set model. Using estimates from Tab. 1 and linear regresion model (10), we estimated Rényi dimensions $\hat{D}_q$ and then evaluated its z_{score} as relative measure of bias

$$z_{\mathrm{score}} = \frac{\hat{D}_q - D_q}{S_{D_q}}. \tag{26}$$

The results are included in Tab. 2. The best estimation with $|z_{\mathrm{score}}| \leq 1.960$ are $H_{1,M}$ followed by Harris estimations $H_{1,HN}, H_{1,HS}, H_{1,HB}$. A structure of $D_q = 2.3219$ and size 128×128×128 was then used for testing 3D and the results are also included in Tab. 2. The best estimators are $H_{1,HS}, H_{1,HN}, H_{1,HB}, H_{1,M}, H_{2,S}$

6 Alzheimer's Disease Diagnosis from Fractal Dimension Estimates

These entropy estimators were used for diagnosis of Alzheimer's disease. We tried to separate two different groups of samples of human brains. In the fist group there were brain scans of patients with Alzheimes disease (AD) and in the second group brain scans of patients with amyotrophic lateral sclerosis (ALS). We were testing on 21 samples (11 for AD and 10 for ALS), represented by $128\times128\times128$ matrices of thresholded images ($\theta = 40\%$). We used two-sample t-test for null hypotheses and alternative hypothesis were

$$\mathrm{H}_0 : \mathrm{E}\hat{D}_q(\mathrm{AD}) = \mathrm{E}\hat{D}_q(\mathrm{ALS}) \tag{27}$$

$$\mathrm{H}_\mathrm{A} : \mathrm{E}\hat{D}_q(\mathrm{AD}) \neq \mathrm{E}\hat{D}_q(\mathrm{ALS}). \tag{28}$$

The results are included in Tab.3. The most significant differences between AD and ALS were observed for $H_{0,\mathrm{N}}, H_{1,\mathrm{S}}, H_{1,\mathrm{B}}$. In the figure 1 are results for $H_{0,\mathrm{N}}$. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. And figure 2 represents fractal dimension estimation for 3D brain scan from ALS set via $H_{0,\mathrm{N}}$, where a is edge length of covering hypercubes.

7 Conclusion

In this paper we tested estimates for Hartley, Shannon and Collision entropy. These estimates were improved by Bayesian estimation and tested on fractals with known fractal dimension. Finally, these estimates were used on two groups of samples of brain scans, in order to obtain the best separator. The best separators, with regard to experiment, are $H_{0,\mathrm{N}}, H_{1,\mathrm{S}}, H_{1,\mathrm{B}}$ and they have a 2% level of significance. But the rest of the estimates have also results under a 5% level of significance. The worst results were obtain for $H_{2,\mathrm{N}}$ namely 4.98%. Given the results, entropy can be used for diagnosis of Alzheimer's disease in the future, considering these methods can be still improved especially by the estimation of k_N or image filtering.

References

[1] Theiler, J., *Estimating fractal dimension.* Journal of the Optical Society of America, Vol. 7, No. 6 1990, 1055-1073.

[2] Renyi, A., *On measures of entropy and information.* Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1961, page 547

[3] Hartley, R.V.L., *Transmission of information.* Bell System Technical Journal, Vol. 7, 1928, 535

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

[5] Harris, B., *The statistical estimation of entropy in the non-parametric case.* MRC Technical Summary Report, 1975

[6] Gomez, C., Mediavilla, A., Hornero, R., Abasolo, D., Fernandez, A., *Use of the Higuchi's fractal dimension for the analysis of MEG recordings from Alzheimer's disease patients.* Medical Engineering & Physics, Volume 31, Issue 3, April 2009, Pages 306-313.

[7] Jouny, C.C., Bergey, G.K., *Characterization of early partial seizure onset: Frequency, complexity and entropy.* Clinical Neurophysiology, Volume 123, Issue 4, April 2012, Pages 658-669.

[8] Lopes,R., Betrouni, N., *Fractal and multifractal analysis: A review.* Medical Image Analysis, Volume 13, Issue 4, August 2009, Pages 634-649.

[9] Polychronaki, G.E., Ktonas, P. Y., Gatzonis, S., Siatouni, A., Asvestas, P. A., H Tsekou, H., Sakas, D. and Nikita, K.S., *Comparison of fractal dimension estimation algorithms for epileptic seizure onset detection.* Journal of Neural Engineering 7 (2010).

Table 1: Entropy estimates

Method	H_0	H_1	H_2
Naive	$H_{0,\mathrm{N}}$	$H_{1,\mathrm{N}}$	$H_{2,\mathrm{N}}$
semibayesian (p_j)	*	$H_{1,\mathrm{S}}$	$H_{2,\mathrm{S}}$
semibayesian (p_j^2)	*	*	$H_{2,\mathrm{S2}}$
bayesian	*	$H_{1,\mathrm{B}}$	*
Miller	*	$H_{1,\mathrm{M}}$	*
Harris	*	$H_{1,\mathrm{HN}}$	*
Harris semibayesian (p_j)	*	$H_{1,\mathrm{HS}}$	*
Harris bayesian $(1/p_j)$	*	$H_{1,\mathrm{HB}}$	*

Table 2: Dimension estimates via various entropy estimates

estimate	Sierpinski carpet $D_q = 1.8928$			Five Box Fractal $D_q = 2.3219$		
	$\hat{D}_q$	SD_q	z_{score}	$\hat{D}_q$	SD_q	z_{score}
$H_{0,\mathrm{N}}$	1.8158	0.0064	-12.0577	2.0897	0.0284	-8.1757
$H_{1,\mathrm{N}}$	1.8472	0.0059	-7.7116	2.1853	0.0320	-4.2690
$H_{2,\mathrm{N}}$	1.8578	0.0076	-4.6212	2.1949	0.0298	-4.2568
$H_{1,\mathrm{S}}$	1.8515	0.0058	-7.0853	2.2367	0.0315	-2.7012
$H_{2,\mathrm{S}}$	1.8657	0.0072	-3.7494	2.2927	0.0298	-0.9798
$H_{2,\mathrm{S2}}$	1.7898	0.0077	-13.4269	2.1189	0.0268	-7.5904
$H_{1,\mathrm{B}}$	1.8170	0.0060	-12.6863	2.1654	0.0297	-5.2638
$H_{1,\mathrm{M}}$	1.8930	0.0059	0.0306	2.3315	0.0349	0.2730
$H_{1,\mathrm{HN}}$	1.8921	0.0059	-0.1203	2.3208	0.0347	-0.0332
$H_{1,\mathrm{HS}}$	1.8921	0.0059	-0.1164	2.3226	0.0347	0.0196
$H_{1,\mathrm{HB}}$	1.8920	0.0059	-0.1328	2.3182	0.0346	-0.1084

Table 3: Diagnostic power

Estimate	$\mathrm{E}\hat{D}_q(\mathrm{AD})$	$\mathrm{E}\hat{D}_q(\mathrm{ALS})$	p_{value}
$H_{0,\mathrm{N}}$	1.9748	2.0337	0.0139
$H_{1,\mathrm{N}}$	2.0663	2.1117	0.0200
$H_{2,\mathrm{N}}$	2.0707	2.1056	0.0498
$H_{1,\mathrm{S}}$	2.0979	2.1493	0.0150
$H_{2,\mathrm{S}}$	2.1474	2.1926	0.0241
$H_{2,\mathrm{S2}}$	1.9289	1.9687	0.0266
$H_{1,\mathrm{B}}$	2.0020	2.0527	0.0145
$H_{1,\mathrm{M}}$	2.2621	2.3139	0.0368
$H_{1,\mathrm{HN}}$	2.2443	2.2954	0.0333
$H_{1,\mathrm{HS}}$	2.2467	2.2980	0.0334
$H_{1,\mathrm{HB}}$	2.2380	2.2891	0.0315

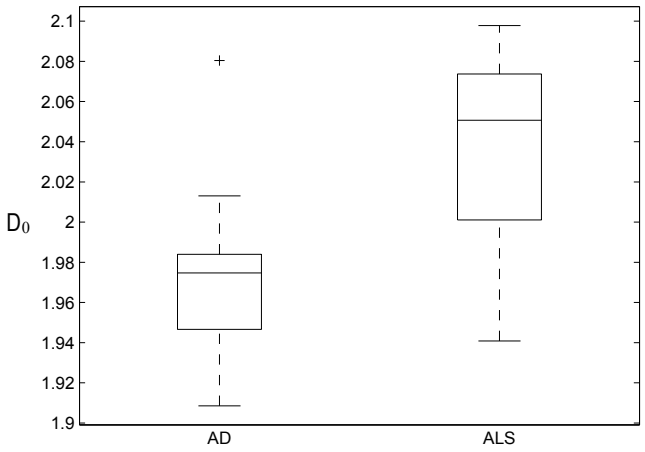


Figure 1: Rényi dimension D_0 for AD and ALS scans

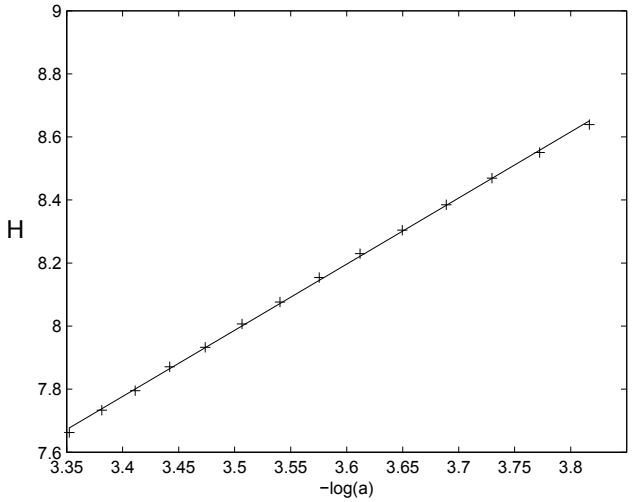


Figure 2: Fractal dimension estimation via entropy

Model-assisted Evolutionary Optimization with Fixed Evaluation Batch Size*

Viktor Charypar

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

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

Abstract. Some black-box optimization problems involve long-running simulations or expensive experiments as the goal function. To enable use of evolutionary algorithms, surrogate models are used to reduce the number of function evaluations. In adaptive model building strategies, some individuals are selected for true function evaluation in order to improve the model. When the experiment or simulation requires a fixed size batch of solutions to evaluate, traditional selection strategies either cannot be used or couple the batch size with the EA generation size. We propose a queue based method for for model-assisted optimization using active learning of a kriging model, where individuals are selected based on the model predictor error estimate. The method was tested on standard benchmark problems and the effects of batch size was studied. Results indicate that the proposed method significantly reduces the number of true fitness evaluation compared to a traditional EA.

Keywords: optimization, evolutionary algorithm, surrogate model, active learning, Kriging

Abstrakt. Některé optimalizační problémy používají jako cílovou funkci dlouho běžící simulace nebo nákladné experimenty. Aby v takových případech bylo možné vuyžít evolučních algoritmů, používají se ke snížení počtu vyhodnocení skutečné cílové funkce náhradní modely. V adaptivních strategiích učení modelů jsou vybráni někteří jednotlivci, kteří jsou vyhodnoceni skutečnou funkcí, aby zlepšili model. V případě že experiment nebo simulace vyžaduje pevnou dávku řešení k vyhodnocení, tradiční techniky jejich výběru buďto nelze použít, nebo vytvoří závislost mezi velikostí dávky a velikostí populace EA. V této práci navrhujeme metodu optimalizace s náhradním modelem využívající frontu a aktivní učení Kriging modelu, ve které jednotlivá řešení vybíráme k vyhodnocení na základě odhadu chyby predikce modelu. Metoda byla testována na standardních testovacích problémech a byl zkoumán vliv velikosti dávky. Výsledky ukazují, že navržená metoda výrazně sníží počet vychodnocení skutečné funkce v porovnání s tradičním EA.

Klíčová slova: optimalizace, evoluční algoritmy, náhradní modely, aktivní učení, Kriging

1 Introduction

Evolutionary optimization algorithms are a popular class of optimization techniques suitable for various optimization problems. One of their main advantages is the ability to find optima of black-box functions – functions that are not explicitly defined and only

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS12/196/OHK3/3T/14 as well as the Czech Science Foundation grant 201/08/0802.

their input/output behavior is known from previous evaluations of a finite number of points in the input space. This is typical for applications in engineering, chemistry or biology, where the evaluation is performed in a form of computer simulation or physical experiment.

The main disadvantage for such applications is the very high number of evaluations of the objective function (called fitness function in the evolutionary optimization context) needed for an evolutionary algorithm (EA) to reach the optimum, which makes them impractical for many applications.

The typical solution to this problem is performing only a part of all evaluations using the true fitness function and using a response-surface model as its replacement for the rest. This approach is called surrogate modeling. When using a surrogate model, only a small portion of all the points that need to be evaluated is evaluated using the true objective function (simulation or experiment) and for the rest, the model prediction is assigned as the fitness value. The model is built using the information from the true fitness evaluations.

Since the fitness function is assumed to be highly non-linear the modeling methods used are non-linear as well. Some of the commonly used methods include artificial neural networks, radial basis functions, regression trees, support vector machines or Gaussian processes [2].

Furthermore, some experiments require a fixed number of samples to be processed at one time. This presents its own set of challenges for adaptive sampling and is the main concern of this paper. We present an evolutionary optimization method assisted by a variant of a Gaussian-process-based interpolating model called kriging. In order to best use the evaluation budget, our approach uses active learning methods in selecting individuals to evaluate using the true fitness function. The key feature of the approach is support for batch evaluation with arbitrary batch size independent of the generation size of the EA.

2 Model-assisted evolutionary optimization

Since the surrogate model used as a replacement for the fitness function in the EA is built using the results of the true fitness function evaluations, there are two competing objectives. First, we need to get the most information about the underlying relations in the data, in order to build a precise model of the fitness function. If the model does not capture the features of the fitness function correctly, the optimization can get stuck in a fake optimum or generally fail to converge to a global one. Second, we have a limited budget for the true fitness function evaluations. Using many points from the input space to build a perfect model can require more true fitness evaluations than not employing a model at all.

In the general use of surrogate modeling, such as design space exploration, the process of selecting points from the input space to evaluate and build the model upon is called sampling [2]. Instead of a traditional upfront sampling schemes based on the theory of design of experiments (DoE), adaptive sampling strategies are used, where a model is improved during the course of the optimization based on previous fitness function evaluations [2]. In an model-assisted evolutionary optimization algorithm, the adaptive

sampling decisions change from selecting which points to evaluate to whether to evaluate a given point selected by the EA with the true fitness function or not. There are two general approaches to this choice: the generation-based approach and the individual-based approach.

2.1 Generation-based approach

In the generation-based approach the decision whether to evaluate an individual point with the true fitness function is made for the whole generation of the evolutionary algorithm. The optimization takes the following steps.

1. An initial N_i generations of the EA is performed, yielding sets $\mathcal{G}_1, \ldots, \mathcal{G}_{N_i}$ of individuals $(\mathbf{x}, f_t(\mathbf{x}))$, f_t being the true fitness function.
2. The model M is trained on the individuals $(\mathbf{x}, f_t(\mathbf{x})) \in \bigcup_{i=1}^{N_i} \mathcal{G}_i$.
3. The fitness function f_t is replaced by a model prediction f_M.
4. T generations are performed evaluating f_M as the fitness function.
5. One generation is performed using f_t yielding a set $\mathcal{G}_j$ of individuals. (initially $j = N_i + 1$)
6. The model is retrained on the individuals $(\mathbf{x}, f_t(\mathbf{x})) \in \bigcup_{i=1}^{j} \mathcal{G}_i$
7. Steps 4–6 are repeated until the optimum is reached.

The amount of true fitness evaluations in this approach is dependent on the population size of the EA and the frequency of control generations T, which can be fixed or adaptively changed during the course of the optimization [4]. For problems requiring batched evaluation this approach has the advantage of evaluating the whole generation, the size of which can be set to the size of the evaluation batch. The main disadvantage of the generation-based strategy is that not all individuals in the control generation are necessarily beneficial to the model quality and the expensive true fitness evaluations are wasted.

2.2 Individual-based approach

As opposed to the generation-based approach, in the individual-based strategy, the decision whether to evaluate a given point using the true fitness function or the surrogate model is made for each individual separately. There are several possible approaches to individual-based sampling, the most used of which is pre-selection. In each generation of the EA, number of points, which is a multiple of the population size, is generated and evaluated using the model prediction. The best of these individuals form the next generation of the algorithm. The optimization is performed as follows.

1. An initial set of points $\mathcal{S}$ is chosen and evaluated using the true fitness function f_t.
2. Model M is trained using the pairs $(\mathbf{x}, f_t(\mathbf{x})) \in \mathcal{S}$

3. A generation of the EA is run with the fitness function replaced by the model prediction f_M and a population $\mathcal{O}_i$ of size qp is generated and evaluated with f_M, where p is the desired population size for the EA and q is the pre-screening ratio. Initially, $i = 1$.
4. A subset $\mathcal{P} \subset \mathcal{O}$ is selected according to a selection criterion.
5. Individuals from $\mathcal{P}$ are evaluated using the true fitness function f_t.
6. The model M is retrained using $\mathcal{S} \cup \mathcal{P}$, the set $\mathcal{S}$ is replaced with $\mathcal{S} \cup \mathcal{P}$, and the EA resumes from step 3.

The key piece of this approach is the selection criterion (or criteria) used to determine which individuals from set $\mathcal{O}$ should be used in the following generation of the algorithm. There are a number of possibilities, let us discuss the most common.

An obvious choice is selecting the best individuals based on the fitness value. This results in the region of the optimum being sampled thoroughly, which helps finding the true optimum. On the other hand, the regions far from the current optimum are neglected and a possible better optimum can be missed. To sample the areas of the fitness landscape that were not explored yet, space-filling criteria are used, either alone or in combination with the best fitness selection or other criteria.

All the previous criteria have the fact that they are concerned with the optimization itself in common. A different approach is to use the information about the model, most importantly its accuracy, to decide which points of the input space to evaluate with the true fitness function in order to most improve it. This approach is sometimes called active learning.

2.3 Active learning

Active learning is an approach that tries to maximize the amount of insight about the modeled function gained from its evaluation while minimizing the number of evaluations necessary. The methods are used in the general field of surrogate modeling as efficient adaptive sampling strategies. The terms adaptive sampling and active learning are often used interchangeably. We will use the term active learning for the methods based on the characteristics of the surrogate model itself, such as accuracy.

The active learning methods are most often based on the local model prediction error, such as cross-validation error. Although some methods are independent of the model, for example the LOLA-Voronoi method [1], most of them depend on the model used. The kriging model used in our proposed method offers an estimate of the local model accuracy by giving an error estimate of its prediction.

3 kriging meta-models

The kriging method is an interpolation method originating in geostatistics [6], based on modeling the function as a realization of a stochastic process [8].

In the ordinary kriging, which we use, the function is modeled as a realization of a stochastic process

$$Y(\mathbf{x}) = \mu_0 + Z(\mathbf{x}) \tag{1}$$

where $Z(\mathbf{x})$ is a stochastic process with mean 0 and covariance function $\sigma^2\psi$ given by

$$cov\{Y(\mathbf{x}+\mathbf{h}), Y(\mathbf{x})\} = \sigma^2\psi(\mathbf{h}), \tag{2}$$

where σ^2 is the process variance for all $\mathbf{x}$. The correlation function $\psi(\mathbf{h})$ is then assumed to have the form

$$\psi(\mathbf{h}) = \exp\left[-\sum_{l=1}^{d}\theta_l|\mathbf{h}_l|^{p_l}\right], \tag{3}$$

where $\theta_l, l = 1, \ldots, d$, where d is the number of dimensions, are the correlation parameters. The correlation function depends on the difference of the two points and has the intuitive property of being equal to 1 if $\mathbf{h} = \mathbf{0}$ and tending to 0 when $\mathbf{h} \to \infty$. The θ_l parameters determine how fast the correlation tends to zero in each coordinate direction and the p_l determines the smoothness of the function.

The ordinary kriging predictor based on n sample points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with values $\mathbf{y} = (y_1, \ldots, y_n)'$ is then given by

$$\hat{y}(\mathbf{x}) = \hat{\mu_0} + \psi(\mathbf{x})'\mathbf{\Psi}^{-1}(\mathbf{y} - \hat{\mu_0}\mathbf{1}), \tag{4}$$

where $\psi(\mathbf{x})' = (\psi(\mathbf{x}-\mathbf{x}_1), \ldots, \psi(\mathbf{x}-\mathbf{x}_n))$, $\mathbf{\Psi}$ is an $n \times n$ matrix with elements $\psi(\mathbf{x}_i - \mathbf{x}_j)$, and

$$\hat{\mu_0} = \frac{\mathbf{1}'\mathbf{\Psi}^{-1}\mathbf{y}}{\mathbf{1}'\mathbf{\Psi}^{-1}\mathbf{1}}. \tag{5}$$

An important feature of the kriging model is that apart from the prediction value it can estimate the prediction error as well. The kriging predictor error in point $\mathbf{x}$ is given by

$$s^2(\mathbf{x}) = \hat{\sigma}^2\left[1 - \psi'\mathbf{\Psi}^{-1}\psi + \frac{(1 - \psi'\mathbf{\Psi}^{-1}\psi)^2}{\mathbf{1}'\mathbf{\Psi}^{-1}\mathbf{1}}\right] \tag{6}$$

where the kriging variance is estimated as

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \hat{\mu_0}\mathbf{1})\mathbf{\Psi}^{-1}(\mathbf{y} - \hat{\mu_0}\mathbf{1})}{n}. \tag{7}$$

The parameters θ_l and p_l can be estimated by maximizing the likelihood function of the observed data.

For the derivation of the equations 4 - 7 as well as the MLE estimation of the parameters the reader may consult a standard stochastic process based derivation by Sacks et al. in [8] or a different approach given by Jones in [5].

4 Method description

In this section we will describe the proposed method for kriging-model-assisted evolutionary optimization with batch fitness evaluation. Our main goal was to decouple the true fitness function sampling from the EA iterations based on an assumption that requiring a specific number of true fitness evaluations in every generations of the EA forces unnecessary sampling.

The method we propose achieves the desired decoupling by introducing an evaluation queue. The evolutionary algorithm uses the model prediction at all times and when a point, in which the model's confidence in its prediction is low, is encountered, it is added to the evaluation queue. Once there are enough points in the queue, all the points in it are evaluated and the model is re-trained using the results. The optimization takes the following course.

1. Initial set $\mathcal{S}$ of b samples is selected using a chosen initial design strategy and evaluated using the true fitness function f_t
2. An initial kriging model M is trained using pairs $(\mathbf{x}, f_t(\mathbf{x})) \in \mathcal{S}$.
3. The evolutionary algorithm is started, with the model prediction f_M as the fitness function.
4. For every prediction $f_M(\mathbf{x}) = \hat{y}_M(\mathbf{x})$, an estimated improvement measure $c(s^2_M(\mathbf{x}))$ is computed from the error estimate $s^2_M(\mathbf{x})$. If $c(s^2_M(\mathbf{x})) > t$, an improvement threshold, the point is added to the evaluation queue $\mathcal{Q}$.
5. If the queue size $|\mathcal{Q}| \geq b$, the batch size, all points $\mathbf{x} \in \mathcal{Q}$ are evaluated, the set $\mathcal{S}$ is replaced by $\mathcal{S} \cup \{(\mathbf{x}, f_t(\mathbf{x})\}$ and the EA is resumed.
6. Steps 4 and 5 are repeated until the goal is reached, or a stall condition is fulfilled.

The b and t parameters, as well as the function $c(s^2)$, are chosen before running the optimization.

To estimate the improvement, which evaluation of a given point will bring, we use a simple measure of estimated improvement – standard deviation (STD) – based on the kriging predictor error estimate, computed directly as its square root

$$STD(x) = \sqrt{\hat{s^2_M}(\mathbf{x})}. \tag{8}$$

The measure captures only the model's estimate of the error of its own prediction (based on the distance from the known samples). As such, it does not take into account the value of the prediction itself and can be considered a measure of the model accuracy.

An important weakness of the measure is that it is based on the model prediction. If the modeled function is deceptive, the model can be very inaccurate while estimating a low variance. A good initial sampling of the fitness function is therefore very important.

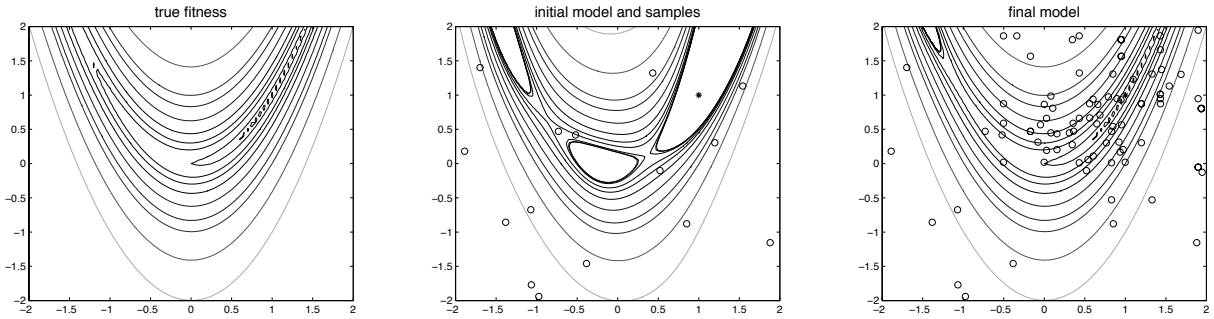


Figure 1: The original fitness function, the initial model and the final model

function	evals (1q)	evals (med)	evals (3q)	goal	reached
De Jong	60	60	120	0.01	1
Rosenbrock	60	125	310	0.1	1
Rastrigin	260	370	580	0.1	0.85

Table 1: GA performance on benchmark functions without a model - number of evaluations to reach the goal and a proportion of 20 runs in which the goal was reached

5 Results and discussion

The proposed method was tested using simulations on three standard benchmark functions. We studied the model evolution during the course of the optimization and investigated the optimal choice of batch size for problems where such a choice is possible.

For testing, we used the genetic algorithm implementation from the global optimization toolbox for the Matlab environment and the implementation of an ordinary kriging model from the SUMO Toolbox [3]. The parameters of the supporting methods, e.g. the genetic algorithm itself, were kept on their default values provided by the implementation.

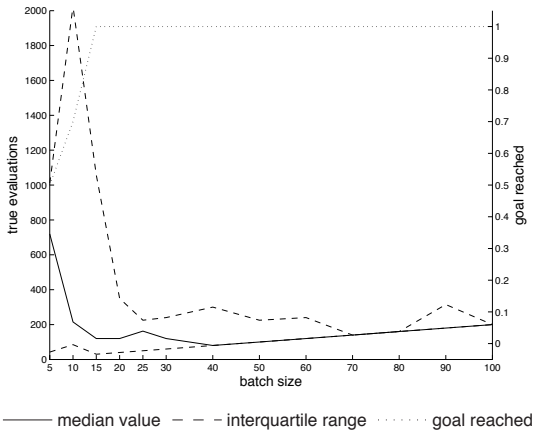


Figure 2: Rosenbrock's function evaluations and proportion of runs reaching the goal with standard GA

Because the EA itself is not deterministic, each test was performed 20 times and the results we present are statistical measures of this sample. As a performance measure we use the number of true fitness evaluations used to reach a set goal in all tests. We also track the proportion of the 20 runs that reached the goal before various limits (time, stall, etc.) took effect.

5.1 Benchmark functions

Since the evolutionary algorithms and optimization heuristics in general are often used on black-box optimization, where the properties of the objective function are unknown, it is not straightforward to asses their quality on real world problems. It has therefore become a standard practice to test optimization algorithms and their modifications on specially designed testing problems.

These benchmark functions are explicitly defined and their properties and optima are known. They are often designed to exploit typical weaknesses of optimization algorithms in finding the global optimum. We used three functions found in literature [7]: the De Jong's function, the Rosenbrock's function and the Rastrigin's function. We performed our tests in two dimensions.

5.2 Model evolution

As the basic illustration of how the model evolves during the course of the EA, let us consider an example test run using the Rosenbrock's function. For this experiment we set the batch size of 15, estimated improvement threshold of 0.001 and the target fitness value of 0.001 as well. The target was reached at the point $(0.9909, 0.9824)$ using 90 true fitness evaluations. A genetic algorithm without a surrogate model needed approximately 3000 evaluations to reach the goal in several test runs.

The model evolution is shown in figure 1. The true fitness function is shown on the left, the initial model is in the middle and the final model on the right. The points where the true fitness function was sampled are denoted with circles an the optimum is marked with a star.

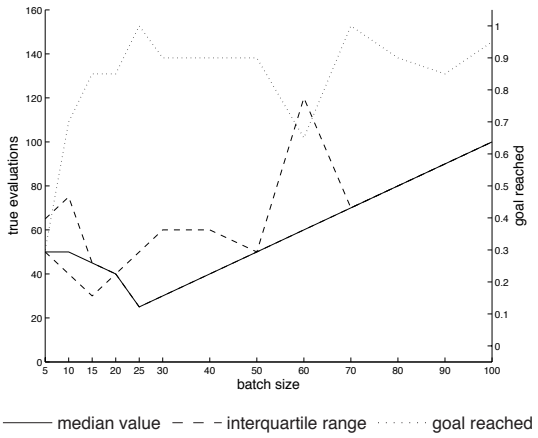


Figure 3: Rosenbrock's function evaluations and proportion of runs reaching the goal with surrogate model

5.3 Batch size

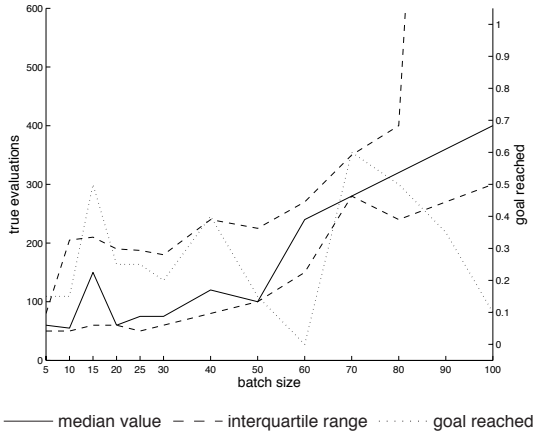


Figure 4: Rastrigin's function evaluations and proportion of runs reaching the goal using SM and normal initial batch size

In order to study the batch size effect on the optimization, a number of experiments were performed with different batch sizes. The only option to achieve a given batch size is to set the population size in a standard GA, in our method however, the settings are independent so a population size of 30, which proved efficient, was used in all of the tests.

For comparison, we also performed tests with the standard genetic algorithm without a model. Results of these simulations are shown in table 1.

The results on the De Jong's functions show that apart from small batch sizes (up to 10), the optimization is successful in all runs. Our method helps stabilize the EA for small batch sizes and for batch sizes above 15 the algorithm finds the optimum using a single batch. For a standard GA this strong dependence arises for batch sizes above 40 and the algorithm reaches the goal in the second generation, evaluating twice as many points.

For the Rosenbrock's function we get the intuitive result that setting the batch size too low leads to more evaluations or a failure to reach the goal, while large batch sizes do not improve the results and waste true fitness evaluations. The comparison is shown in figures 2 and 3 (note the different scales). Overall the method reduces the number of true

evaluations from hundreds to tens for the Rosenbrock's function, while slightly reducing the success rate of the computation.

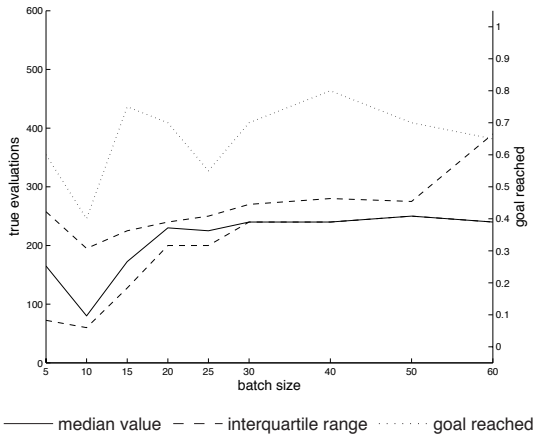


Figure 5: Rastrigin's function evaluations and proportion of runs reaching the goal using SM and double initial batch size

The Rastrigin's function proved difficult to optimize even without a surrogate model. The number of true fitness evaluations was reduced approximately three times in the area of the highest success rate with batch size of 70 (figure 4). We attribute the method's difficulty optimizing the Rastrigin's function to the fact that the kriging model is local and thus it requires a large number of samples to capture the function's complicated behavior in the whole input space. When the initial sampling is misleading, which is more likely for the Rastrigin's function, both the model prediction and estimated improvement are wrong.

In order to prevent bad initial sampling a subset of tests was conducted using an integer multiple of batch size. Figure 5 shows results for Rastrigin function with double initial batch size. Larger initial batch size stabilizes the method. Success rate increased from around 30% to 60% even for smaller batch sizes, which is close to what a simple GA achieved, while maintaining the number of true evaluations low. The fact that a larger initial batch will be evaluated even in cases where a small batch would suffice can be considered a disadvantage of this approach.

The results suggest that the best batch size is highly problem-dependent. The experimental results support the intuition that batches too small are bad for the initial sampling of the model and batches too large slow down the model improvement by evaluating points that it would not be necessary to evaluate with smaller batches. The proposed method is also very sensitive to good initial sample selection, which is the most usual reason for it to fail to find the optimum. Combining a larger initial batch with a smaller batch during the optimization helps alleviate the problem.

6 Conclusions

In this paper we presented a method for model-assisted evolutionary optimization with a fixed batch size requirement. To decouple the sampling from the EA iterations and support an individual-based approach while keeping a fixed evaluation batch size, the method uses an evaluation queue. The candidates for true fitness evaluations are selected by an active learning method using a measure of estimated improvement of the model quality based on the model prediction error estimate.

The results suggest that small batch sizes perform better when the objective function is simple, while causing bad initial sampling, which can be successfully solved using a larger initial batch. The future development of this work should include experiments with a different initial sample distribution than random as well as comparison of the

method with other ways of employing a surrogate model in the optimization and other model-assisted optimization methods.

The method brings promising results, reducing the number of true fitness evaluations to a large degree for some of the benchmark functions, however its success is highly dependent on the optimized function and its initial sampling.

References

[1] K. Crombecq, L. De Tommasi, D. Gorissen, and T. Dhaene. *A novel sequential design strategy for global surrogate modeling.* In 'Winter Simulation Conference', WSC '09, 731–742. Winter Simulation Conference, (2009).

[2] D. Gorissen. *Grid-enabled Adaptive Surrogate Modeling for Computer Aided Engineering.* PhD thesis, Ghent University, University of Antwerp, (2009).

[3] D. Gorissen, I. Couckuyt, P. Demeester, T. Dhaene, and K. Crombecq. *A surrogate modeling and adaptive sampling toolbox for computer based design.* The Journal of Machine Learning Research **11** (2010), 2051–2055.

[4] Y. Jin, M. Olhofer, and B. Sendhoff. *Managing approximate models in evolutionary aerodynamic design optimization.* In 'Evolutionary Computation, 2001. Proceedings of the 2001 Congress on', volume 1, 592–599. Ieee, (2001).

[5] D. Jones. *A taxonomy of global optimization methods based on response surfaces.* Journal of Global Optimization **21** (2001), 345–383.

[6] G. Matheron. *Principles of geostatistics.* Economic geology **58** (1963), 1246–1266.

[7] M. Molga and C. Smutnicki. *Test functions for optimization needs.* Test functions for optimization needs (2005).

[8] J. Sacks, W. Welch, T. Mitchell, and H. Wynn. *Design and analysis of computer experiments.* Statistical science **4** (1989), 409–423.

Database Optimization at COMPASS Experiment*

Vladimír Jarý

4th year of PGS, email: Vladimir.Jary@cern.ch
Department of Software Engineering in Economics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. The COMPASS experiment at CERN laboratory employs the database service to manage information about data taking process and about condition of detectors, triggers, and beamline. During the year 2009, the database service experienced performance and stability issues caused by increases of the data rates. This paper summarizes various optimization techniques that have been proposed and implemented in order to guarantee requested high availability and high reliability of the service. At first, a new database architecture of the experiment based on the proxy software, replication, regular backups, and continuous monitoring is presented. Then, various possible optimizations of the structure of tables and queries are analyzed. Finally, several features of the new version of the database software that could be used to increase scalability and reliability of the system are discussed.

Keywords: database, storage, COMPASS, high availability, high reliability

Abstrakt. Experiment COMPASS v laboratoři CERN využívá databázovou službu pro správu informací o sběru dat a o stavu detektorů, terče nebo svazku při měření. Během roku 2009 způsobilo zvýšení datového toku problémy s výkonem a se stabilitou databázové služby. V tomto článku jsou shrnuty optimalizace databázové služby, které byly navrženy a implementovány s cílem zajistit požadovanou vysokou spolehlivost a dostupnost služby. Nejprve je představena nově implementovaná databázová architektura založená na proxy serveru, replikaci, pravidelném zálohování a nepřetržitém dohledu. Poté jsou analyzovány rozličné techniky optimalizace struktury tabulek a dotazů. V závěru článku jsou diskutovány vlastnosti nově implementované do databázového software a je navrženo jejich možné využití pro další navýšení spolehlivosti a škálovatelnosti služby.

Klíčová slova: databáze, úložiště, COMPASS, vysoká dostupnost, vysoká spolehlivost

1 Introduction

COMPASS is a fixed target high energy physics experiment situated at the Super Proton Synchrotron particle accelerator at laboratory CERN in Geneva, Switzerland. The scientific program of the experiment was approved in 1997 by the CERN scientific council; it includes studies of the gluon and quark structure and the spectroscopy of hadrons using high intensity muon and hadron beams, [1]. After several years of preparations and commissioning, the data taking started in 2002. Currently, the experiment is already in its

*This work has been supported by the MŠMT grants LA08015 and SGS 11/16

second phase known as the COMPASS-II that is designed to study Primakoff scattering or Drell-Yan effect, [2].

COMPASS experiment uses the MySQL database server software to manage information about data taking process and about configuration of the various equipment. At first, the original database service of the experiment is presented. The service experienced performance and stability issues caused by increases in trigger rates during the data taking in 2009. Therefore, we have designed and implemented more robust architecture. Then, we present results of optimizations of the database queries and database structure. Finally, we propose a further improvements based on features included into the recent version of the server software that should increase scalability and reliability of the service.

2 Optimization of the database architecture

MySQL server used by the COMPASS database service manages approximately 20 logical databases, however the most important and the most frequently used data are stored in the *beamdb*, the *runlb*, and the *DATE_log* databases. The *beamdb* database contains information about state of triggers, detectors, and beamline. The *runlb* database stores the data of the electronic logbook of the experiment. Finally, the *DATE_log* database holds software messages produced by various components of the *DATE* data acquisition system, [3].

In the original architecture, the database service was powered by two physical servers called *pccodb01* and *pccodb02*. These servers were synchronized using the master-master replication, i.e. the *pccodb01* server acted as a replication master of the slave server *pccodb02* and at the same time, the *pccodb02* server also acted as a replication master of the slave server *pccodb01*. Clients connected to the service through the virtual address *pccodb00* that normally pointed directly to the server *pccodb01*. A watchdog process continuously monitored a health of the servers and in the case it detected a failure of one server, it rewrote the virtual address to point to the remaining server.

After increase of the trigger rate in 2009, the database service experienced performance issues. We have investigated the architecture and concluded that the issues had been caused by combination of obsolete database software and outdated hardware. In [8], we have shown that the amount of random access memory (RAM) is essential for optimal performance of the database servers; the original servers were equipped only by 3 GB of RAM. Therefore, we have proposed to migrate the service to more powerful servers equipped with multicore processors and 16 GB of RAM and also to update database software to the most recent stable version. Furthermore, the 64-bit operating system would be used on the new servers. The new architecture consists of two new servers *pccodb11* and *pccodb12* that are synchronized using the master-master replication. Third machine *pccodb10* is mainly used as a proxy, monitoring, and web server. In the new architecture, the virtual address *pccodb00* points to the proxy software on the *pccodb10* server. By using the same virtual address, there was no need to reconfigure clients during the migration to the new architecture.

The MySQL Proxy software deployed on the *pccodb10* server can be used to log, modify, or filter both queries sent to server and result sets returned by the server. Furthermore, the proxy software is able to change a backend server within active client

connection. We have used this feature together with a monitoring system to implement a failover solution. As a monitoring system, we have decided to use the Nagios software that is designed to monitor state of services and resources on remote hosts. Besides the reporting via the web interface, the Nagios is also capable of sending e-mail or SMS notification and of executing predefined action in the case it detects an accident. We have configured Nagios to monitor the state of MySQL process, the state of replication, available RAM and disk space, temperature of CPU cores, or state of the system scheduler *cron* on the servers *pccodb11* and *pccodb12*. In the case Nagios detects failure of one of the servers, it changes the address of the backend server of the MySQL Proxy. We use the above mentioned scheduler to regularly create snapshots of the database that should be used as a backup. Furthermore, during the replication, the statements that modify data or structure are recorded into the binary log on the master server; this log can be regarded as an incremental backup.

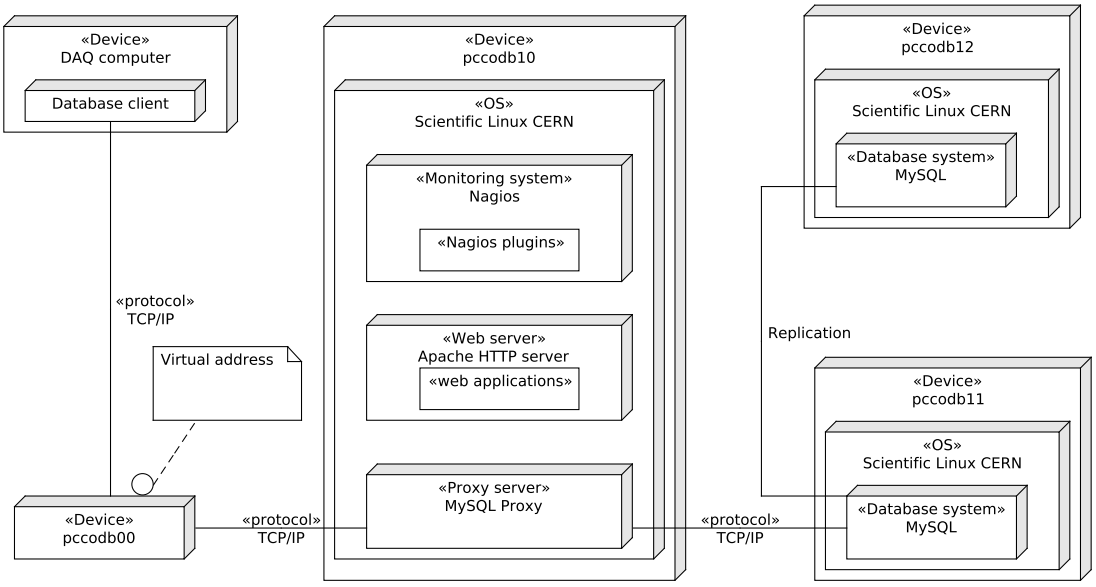


Figure 1: Deployment diagram of the newly implemented database architecture

We have presented the proposal to the COMPASS collaboration, [7] and after approval by a technical coordinator of the experiment, we have successfully implemented it just before the start of data taking in the year 2010, [6]. The issues solved during the migration to the new architecture are summarized in [9].

3 Optimization of table structure and queries

During configuration of the MySQL servers on the new machines, we have enabled logging of the *slow* queries. Evaluation of the slow query requires time longer than a predefined value. Knowledge of the slow queries is important as it can be used to identify improperly designed table indexes which can cause performance issues. MySQL software contains the *EXPLAIN* tool that analyzes the query evaluation plan of given query. The tool

displays which (if any) index is used during query evaluation, number of rows that need to be searched, whether a result set needs to be sorted, or whether a temporary file is required for this sorting. The output of the *EXPLAIN* tool should be used to design proper table indexes or modify structure of the queries to reduce query evaluation time.

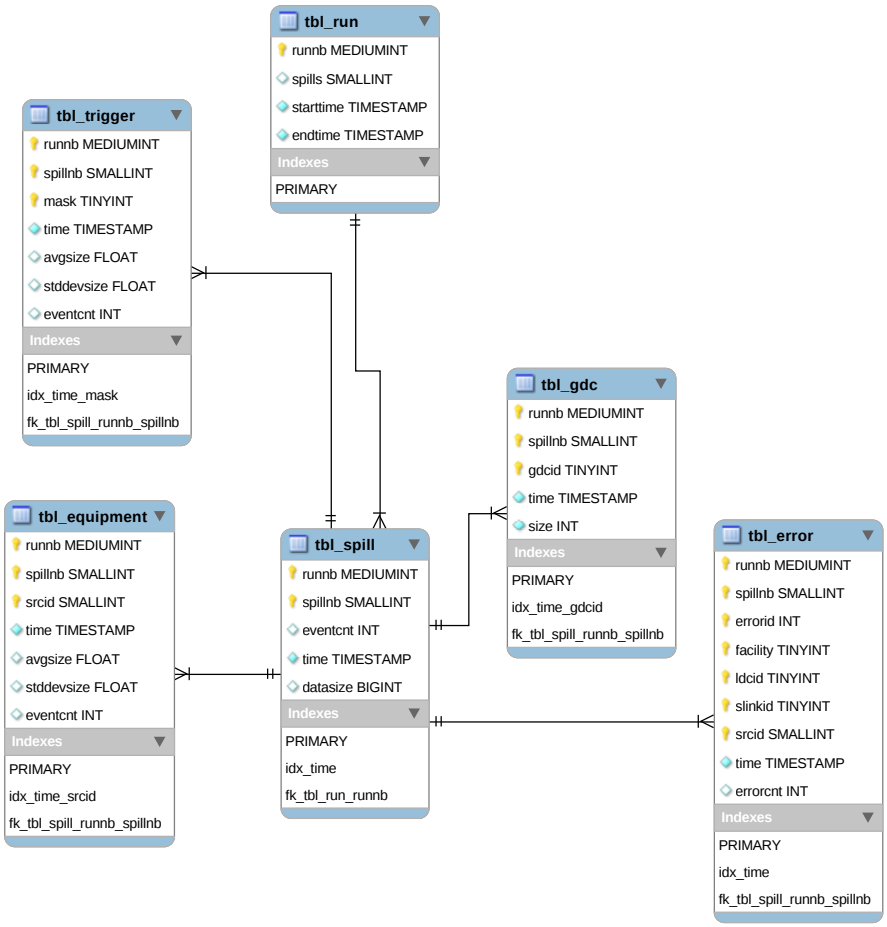


Figure 2: Schema of the *daqmon* database

We have been asked to develop a database part of the new application called *daqmon* designed for monitoring of the performance of the various parts of the data acquisition system. The database tables would be filled by an online filter process, the data would be visualized by a custom graphical interface based on the ROOT framework, [4]. We have used the MySQL Workbench tool to design a structure of the *daqmon* database. The database consists of six tables: *tbl_run* and *tbl_spill* with information about periods of data taking called runs and spills, *tbl_gdc* with information about global data collectors (i.e. computers that gather data), *tbl_trigger* with information about occurrences of triggers, *tbl_equipment* with information about subdetectors, and *tbl_error* with information about errors.

We have used the *EXPLAIN* tool to propose proper indexes of the tables in the *daqmon* database. The structure of the *tbl_trigger* table is shown in the following listing:

```
CREATE  TABLE IF NOT EXISTS `daqmon`.`tbl_trigger`(
```

```
‘runnb‘ MEDIUMINT NOT NULL COMMENT ’Run number’ ,
‘spillnb‘ SMALLINT NOT NULL COMMENT ’Spill number’ ,
‘mask‘ TINYINT NOT NULL COMMENT ’Trigger mask’ ,
‘time‘ TIMESTAMP NOT NULL DEFAULT CURRENT_TIMESTAMP COMMENT ’Timestamp’,
‘avgsize‘ FLOAT NULL COMMENT ’Avg. event size for the  mask in spill’,
‘stddevsize‘ FLOAT NULL COMMENT ’Standard deviation
                    of the event size for the mask in spill’ ,
‘eventcnt‘ INT NULL
       COMMENT ’Number of times the mask appeared in spill’,
PRIMARY KEY(‘mask‘, ‘runnb‘, ‘spillnb‘),
INDEX idx_time_mask(‘time‘, ‘mask‘)) ENGINE = MyISAM;
```

Suppose that the monitoring application built on the *daqmon* database should display the trigger mask, the timestamp, and the average size of all the records with given run number (e.g. 85626) ordered by the time. The corresponding rows can be retrieved from the table using the following query in the SQL language:

```
SELECT mask, time, avgsize FROM tbl_trigger WHERE runnb=85626
                           ORDER BY time;
```

The output of the EXPLAIN command is summarized in Table 1.

id	*select_type*	*table*	*type*	*key*	*rows*	*Extra*
1	SIMPLE	tbl_trigger	ALL	NULL	1127528	Using where; Using filesort

Table 1: The result of the EXPLAIN command on an non–optimized table

The result of the EXPLAIN statement revealed several problems: the type *All* means that all the 1127528 rows in the table must be searched, none key/index can be used. Moreover, an additional pass is required to sort the result. The type *Simple* of the query means that nor union nor subqueries are used during evaluation of the query. The primary key of the table *(mask, runnb, spillnb)* cannot be used because the *runnb* is not its prefix. If the columns in the primary key are reorganized into the following order *(runnb, spillnb, mask)*, the primary key could be used to retrieve the desired rows. This assumption can be confirmed by the EXPLAIN command (see Table 2). This time, only 2383 records are searched, though the file sorting is still performed.

id	*select_type*	*table*	*type*	*key*	*rows*	*Extra*
1	SIMPLE	tbl_trigger	ref	idx_runnb_spillnb_mask	2383	Using where; Using filesort

Table 2: The result of the EXPLAIN command on the table with the optimized index

Under certain circumstances, it is possible to satisfy the ORDER BY clause using the index, thus eliminating the need of the file sorting. According to the documentation [12], this is valid for the queries with the following structure:

```
SELECT * FROM table WHERE keypart1=constant ORDER BY keypart1;
```

Unfortunately, the examined query does not have this structure because the key that retrieves the rows (*Primary Key*) is different from the key that is used in the ORDER BY clause (the *idx_time_mask*).

The *runnb* in the WHERE clause can be replaced by the time interval between the start and the end of the run. The information about the start and the end of the run is stored in the table *tbl_run* described in the following listing:

```
CREATE TABLE IF NOT EXISTS `daqmon`.`tbl_run` (
 `runnb` MEDIUMINT UNSIGNED NOT NULL COMMENT 'Run number',
 `spills` SMALLINT DEFAULT NULL COMMENT 'Number of spills in run',
 `starttime` TIMESTAMP NOT NULL DEFAULT CURRENT_TIMESTAMP
                 COMMENT 'Time when run started',
 `endtime` TIMESTAMP NULL DEFAULT NULL COMMENT 'Time when run ended',
 PRIMARY KEY (`runnb`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8;
```

The start time of the run X is returned using the following query:

```
SELECT starttime FROM tbl_run WHERE runnb=X;
```

In a similar fashion, one can also obtain the time when the given run ended. The *runnb* is the PRIMARY KEY of the table, therefore at most one record with the given run number can exist in the table. This means that only one record needs to be searched to return the start/end time of the given run. By substituting the run number with the corresponding time interval in the original query, we get the following query:

```
SELECT mask, time, avgsize
FROM tbl_trigger
WHERE time>=(SELECT starttime from tbl_run WHERE runnb=85626)
  AND time<=(SELECT   endtime from tbl_run WHERE runnb=85626)
ORDER BY TIME;
```

The results of the EXPLAIN (see table 3) statement confirm that both subqueries are indeed searching only 1 row in the *tbl_run* table as was expected. Additionally, the file sort is not needed anymore and the query is executed faster. However, the speed improvement is not very significant in this particular case. The result of the query contains approximately 2000 rows and the file sort can be done in the memory buffer so it is reasonably fast. In case the size of memory buffer is exceeded, a temporary table must be created and sorted on the disk and the file sorting is slow. The maximal size of the memory buffer is controlled by the variable *sort_buffer_size*. To sum it up, the file sorting should be avoided, if possible.

We have been also asked to analyze the most frequently used queries over the *ecal_mon* table. The queries are regularly issued by the *Detector Control System* every 15 minutes. The *ecal_mon* table in the *beamdb* database contains information about state of blocks that form the electromagnetic calorimeter. With over one billion rows, it is the

largest table in the database, therefore proper indexing of the table is essential for the smooth operation of the database service. Using the EXPLAIN tool, we have verified that the table indexes are correctly used during evaluation of the queries. Additionally, the EXPLAIN tool contained the *Select tables optimized away* value in the *Extra* column for several queries. This means that the query contains some aggregate function such as MIN or MAX that can be resolved using the table index, therefore no rows are browsed and only one row is returned.

id	*select_type*	*table*	*type*	*key*	*rows*	*Extra*
1	SIMPLE	tbl_trigger	range	idx_time_mask	1932	Using where;
2	SUBQUERY	tbl_run	const	PRIMARY	1	
3	SUBQUERY	tbl_run	const	PRIMARY	1	

Table 3: The result of the EXPLAIN command on the modified query

We have also analyzed different storage engines available in the MySQL. We have compared performance of the InnoDB and MyISAM engines in the most frequently used operations: row inserting, table indexing, and query evaluation. InnoDB is a transactional engine (i.e. engine that supports fully ACID compliant transactions), therefore transaction handling reduces speed of queries that modify data. On the other hand, the engine offers better support for indexing, therefore retrieving rows is faster. MyISAM engine does not support transactions, therefore it is faster on queries that modify data. As the COMPASS database is characterized by frequent updates, we have decided to use the MyISAM engine [8]. However, we also propose to convert tables with historical data to the ARCHIVE engine to save disk space.

4 Proposal of the update of the database architecture

The newly implemented database architecture is in operation since year 2010. During data taking in years 2010 and 2011 no serious problem occurred. However, the server *pccodb11* crashed due to a hardware failure in May 2012 during the shutdown of the experiment. Monitoring system Nagios detected the incident and changed the address of the backend server of the MySQL proxy to the *pccodb12* server. Unfortunately, as a result of the crash, the binary log on the *pccodb11* server became corrupted and therefore, it was not possible to restart the replication process. Thus, it was required to shutdown the database service and manually resynchronize both servers.

Although no data were lost, the availability of the service was affected during the recovery. Therefore, we recommend to add more database servers into the architecture in order to increase the redundancy of the system. Additionally, we propose to change the current master-master replication topology to the master-multiple slaves topology and enable the load balancing mode of the MySQL Proxy software. The proxy supports the read-only load balancing; in this mode all queries that modify data or structure (i.e. *INSERT*, *UPDATE*, *DELETE*, *ALTER*, *DROP*, *TRUNCATE* statements) are sent to the master server by the proxy software. Queries that only retrieve data (i.e. *SELECT* statement) are distributed between the replication slaves by the proxy. Load balancing

contributes to the scalability of the architecture – if a higher performance of the service is required, more replication slaves are added into the system.

The new database servers are powered by the MySQL software in version 5.1. We have investigated new features implemented into the more recent version 5.6 of the MySQL. This version improves the replication technology by implementing global transaction identifiers (GTID) and introducing new *mysqlfailover* and *mysqlrpadmin* tools, [11]. GTIDs are used to simplify tracking of replication progress between master and slave servers. The *mysqlfailover* utility monitors the replication topology and in case it detects failure of the master server, it automatically promotes the most updated slave to the master role; the tool uses the GTIDs to ensure that no transaction is lost during failover. The *mysqlrpadmin* utility provides slave discovery in the replication environment and replication monitoring, it enables disconnection of the master server for the maintenance purposes.

The support for the partitioned tables has been added into MySQL 5.1. Partitioning enables distribution of table data into several partitions. MySQL software supports horizontal partitioning, i.e. table is distributed into partitions by rows. Division of rows into partitions is based on the value of the *partitioning function* which is based on selected type of partitioning. Depending on the type, the partitioning function takes as a parameter a column value, a set of column values, or a function of one or more column values. MySQL supports *range*, *list*, *hash*, and *key* partitioning, each type is described in the MySQL manual, [12].

Partitioned tables are used in the optimization technique known as a *partition pruning*. The technique is based on a fact that the query evaluation engine only browses the partition(s) that can contain the desired data instead of performing a scan of the full table. We have conducted a simple test that would verify the benefit of the partition pruning. We have defined a simple test table *employees1* with information about employees:

```
CREATE TABLE employees1 (
  id INT NOT NULL,
  salary int(11) NOT NULL);
```

The table *employees2* has the same structure, however it is distributed into 4 partitions by range on the *salary* column.

```
CREATE TABLE employees2 (
  id INT NOT NULL,
  salary int(11) NOT NULL
) PARTITION BY RANGE (salary)(
  PARTITION p0 VALUES LESS THAN (25000),
  PARTITION p1 VALUES LESS THAN (50000),
  PARTITION p2 VALUES LESS THAN (75000),
  PARTITION p3 VALUES LESS THAN MAXVALUE);
```

In the range partitioning, the database administrator needs to define division of possible values of given column (*salary* in this case) into several continuous, non-overlapping intervals. The partitioning function takes a value of the given column as its parameters and according to the value it places the row into the corresponding partition.

We have filled both tables with 10 000 000 random records using the script in the Perl language. Then we have measured time required to calculate number of employees with salary from the interval [26 000, 49 000]. The test have been performed in the *qemu* virtual system and the laptop powered by Intel Core2 Duo T9600 processor (two cores running at 2.8 GHz) supported by 4 GB of RAM. The results of the test are summarized in Table 4. On the physical hardware, the execution of query is almost four times faster on the partitioned table. This is caused by the fact that only the partition *p1* that contains approximately 1/4 of rows is searched whereas all 10 million rows must be browsed in the non-partitioned table.

Configuration	*employees1*	*employees2*
Core2 Duo CPU T9600 @ 2.80GHz	0.92 s	0.26 s
QEmu	32.64 s	13.48 s

Table 4: Partition pruning in MySQL 5.1.42

At the COMPASS experiment, the range partitioning could be used for the *messages* table in the *DATE_log* database. The table contains debug and information messages produced by the DATE software package. Very often, the data acquisition experts need to know the behaviour of the system in a certain time period. Thus it would be possible to define range partitioning based on the values from the *timestamp* column. In fact, this behaviour is emulated by a special *cron* job; the job creates a new table for messages every day and puts older message into archive tables.

As the support for partitioning has been introduced in MySQL 5.1, we have decided not to use it yet. However, in newer versions of MySQL server (5.5, 5.6) improvements of the partitioning have been implemented.

5 Summary

After the increase of the trigger rates in 2009, the original database system of the COMPASS experiment experienced performance and stability problems. We have investigated the system and concluded that the problems had been caused by old hardware and software. We have proposed and implemented new database architecture based on more recent software and more powerful hardware. The new architecture uses replication, continuous monitoring, regular backups, and proxy software to guarantee high availability and high reliability of the service. Then, we have tested different storage engines supported by the MySQL software and decided that the MyISAM engine is the most suitable candidate for the needs of the experiment. We have also used the *Explain* tool to design proper table indexes.

However, the redundancy of the database service is low as only two machines are used as a database servers. Therefore, we recommend to add more servers and enable the load balancing mode of the proxy software. We also propose to update the database software to MySQL 5.6 that improves replication technology. During the update, several tables should also be partitioned to make use of the partition pruning optimization. The proposal needs to be discussed within the COMPASS collaboration and in case it is

approved, it may be implemented during the planned shutdown of the CERN accelerators in 2013.

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] C. Adolph, . . . , V. Jarý et al. (the COMPASS collaboration): *COMPASS-II proposal.* CERN-SPSC-2010-014; SPSC-P-340 (May 2010)

[3] T. Anticic et al. (the ALICE collaboration): *ALICE DAQ and ECS User's Guide.* CERN, ALICE internal note, ALICE-INT-2005-015, 2005.

[4] R. Brun, F. Rademakers: *ROOT - An Object Oriented Data Analysis Framework.* In: Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) pp. 81–86.

[5] L. Fleková, V. Jarý, T. Liška: *Mass Data Processing Optimization on High Energy Physics Experiments*, In: 4th International Conference on Advanced Computer Theory and Engineering, Dubai, 2010, ISBN 978-07-918-5993-3.

[6] L. Fleková, V. Jarý, T. Liška, M. Virius: *Proposal and results of COMPASS database upgrade*, In: Stochastic and Physical Monitoring Systems, Děčín, 2010, ISBN 978-80-01-04641-8. pp. 45–50.

[7] L. Fleková, V. Jarý, T. Liška: *Proposal on the COMPASS database upgrade*, In: COMPASS Frontend Electronics meeting, March 2010, Geneva.

[8] L. Fleková, V. Jarý, T. Liška, M. Virius: *Využití databází v rámci fyzikálního experimentu COMPASS*, In: 36th Software Development, Ostrava: VŠB – Technická univerzita Ostrava, 2010, ISBN 978-80-248-2225-9. pp. 68–75.

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

[10] V. Jarý: *Highly available and reliable database for the COMPASS experiment*, In: Advanced Studies Institute, Symmetries and Spin, Prague 2012.

[11] M. Keep: *MySQL 5.6 Replication - Enabling the Next Generation of Web & Cloud Services* [online]. August 2012. Available at: `http://dev.mysql.com/tech-resources/articles/mysql-5.6-replication.html`

[12] *MySQL 5.1 Reference Manual* [online]. August 2012. Available at: `http://dev.mysql.com/doc/refman/5.1/en/`

Diferenciální rovnice s danými symetriemi*

Dalibor Karásek

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

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

Abstract. This paper describes method of finding systems of differential equations with prescribed symmetries. It briefly defines infinitesimal symmetries and then present algorithm of searching such systems. This method is subsequently used to obtain differential equation, whose algebra of symmetries contain specific nilpotent Lie algebra, from certain infinite series of algebras.

Keywords: Lie algebras, series of algebras, solvable extensions, infinitesimal symmetries

Abstrakt. Tento příspěvek popisuje jak nalézt diferenciální rovnice, po kterých požadujeme, aby měly předepsané symetrie. Ve stručnosti zadefinuje, co jsou infinitezimální symetrie a potom prezentuje metodu hledání těchto rovnic. Tato metoda je vzápětí aplikována na nalezení systémů rovnic, jejichž symetrie se uzavírají do jedné z algeber, jež jsou členy jisté nekonečné série nilpotentních algeber.

Klíčová slova: Lieovy algebry, série algeber, řešitelná rozšíření, infinitezimální symetrie

1 Úvod

Symetrie diferenciálních rovnic stály u zrodu teorie Lieových algeber [4] a není tedy divu, že se stále zkoumají způsoby, jak tyto dva obory propojit. Je to dáno i tím, že symetrie jsou jedním z nejdůležitějších nástrojů moderní fyziky. Využívají se jak při konstrukci modelů mikrosvěta, tak při modelování různých dějů probíhajících v našem každodenním životě.

Teorie Lieových algeber je velmi pestrá oblast sahající od diferenciální geometrie až po kvantovou mechaniku. Lieovy algebry již byly částečně zklasifikovány Cartanem a Levim [2, 3], kteří zklasifikovali poloprosté Lieovy algebry a dokázali, že každá Lieova algebra jde rozložit na součet poloprosté a řešitelné. Řešitelné algebry ovšem nejsou doposud klasifikovány. Jejich úplná klasifikace je známá jen pro nízké dimenze.

Rozdílný přístup použil Pavel Winternitz, Libor Šnobl a další autoři v sérii prací (např. [5–7]), ve kterých rozšiřovali jistou posloupnost nilpotentních algeber na algebry řešitelné a zkoumali jejich vlastnosti. Tímto postupem získali pro libovolně velké $n \in \mathbb{N}$ několik algeber dimenze n, na nichž lze testovat různé hypotézy, které by měly platit obecně. Některé z těchto algeber se také používají ve fyzice (například zobecněná Heisenbergova algebra).

*Tato práce byla podpořena grantem SGS10/295/OHK4/3T/14

Jak už bylo řečeno, Lieovy algebry byly objeveny při studiu symetrií diferenciálních rovnic. Je znám algoritmický postup, jak pro daný systém (parciálních) diferenciálních nalézt takzvané infinitezimální symetrie, což jsou jistá vektorová pole, které se uzavírají pomocí komutátoru do Lieovy algebry. Tento postup má svá úskalí, ale je docela dobře prozkoumán a implementován v symbolických výpočetních programech.

Tento příspěvek si klade za cíl opačný proces, t.j. pro dané infinitezimální symetrie nalézt co největší, ideálně celou, třídu diferenciálních rovnic s těmito symetriemi. Tento proces pak lze použít třeba pro tvorbu fyzikálních modelů, po kterých vyžadujeme některé symetrie.

2 Infinitezimální symetrie, silné invarianty, realizace

V této sekci stručně zavedeme definice infinitezimálních symetrií, prolongací a invariantů vektorových polí. Abychom nedělali porozumění definicí zbytečně složité, omezíme se na případ soustav obyčejných diferenciálních rovnic se dvěma závislými proměnnými. Na složitější případy můžeme použité vzorce snadno zobecnit.

Definice 2.1. Mějme soustavu diferenciálních rovnic N-tého řádu

$$F_a(x, y, z, y', z', \dots, y^{(N)}, z^{(N)}) = 0, \tag{1}$$

kde $a = 1, \dots, K$.

$\mathbb{R}^{2N+3}$ se souřadnicemi $x, y_0, z_0, \dots, y_N, z_N$ nazveme **N-tý jetový prostor** J^N. Tuto definici je výhodné intuitivně rozšířit až na J^∞. Díky tomu, že máme diferenciální rovnice konečného řádu, nenastanou žádné komplikace s nekonečným počtem souřadnic.

Na J^∞ máme význačné vektorové pole

$$D_x := \partial_x + \sum_{i=0}^{\infty} y_{i+1}\partial_{y_i} + \sum_{j=0}^{\infty} z_{j+1}\partial_{z_j}, \tag{2}$$

které nazveme **operátor totální derivace**

Poznámka 2.2. Na systém diferenciálních rovnic (1) se lze nyní dívat jako na systém algebraických rovnic na J^N.

Poznámka 2.3. D_x jde interpretovat jako zobrazení $\mathcal{F}(J^k) \to \mathcal{F}(J^{k+1})$, t.j. zobrazení, které vezme funkci na k. jetovém prostoru a vyrobí z ní funkci na $k+1$. jetovém prostoru.

Poznámka 2.4. D_x se říká operátor totální derivace, protože jeho aplikací na funkci $F \in \mathcal{F}(J^n)$ a vyhodnocením výsledku na zobecněném grafu funkce $(y(x), z(x))$, t.j. množině $\{(x, y(x), z(x), y'(x), z'(x), \dots, y^{(n)}(x), z^{(n)}(x))\}$ dostaneme tentýž výsledek, jako kdybychom zderivovali $F \circ y$

$$\frac{\mathrm{d}}{\mathrm{d}x} F(x, y(x), z(x), y'(x), z'(x), \dots) = (D_x F)(x, y(x), z(x), y'(x), z'(x), \dots). \tag{3}$$

Stejně jako lze každý graf funkce $\{(x, y(x), z(x))\}$ pomocí zobecněného grafu rozšířit na celé J^N, můžeme na jetový prostor rozšířit pomocí takzvané prolongace libovolné vektorové pole z $\mathbb{R}^3$.

Definice 2.5. Buď $X = \xi(x,y,z)\partial_x + \phi_0^y(x,y,z)\partial_y + \phi_0^z(x,y,z)\partial_z$ vektorové pole na $\mathbb{R}^3$. Definujme rekurentně funkce

$$\begin{aligned} \phi_j^y &:= D_x\phi_{j-1}^y - y_j(D_x\xi), \\ \phi_j^z &:= D_x\phi_{j-1}^z - z_j(D_x\xi). \end{aligned} \tag{4}$$

N-tá **prolongace** vektorového pole X je vektorové pole

$$\mathrm{pr}^N X := \xi\partial_x + \sum_{i=0}^{N} \phi_i^y \partial_{y_i} + \sum_{j=0}^{N} \phi_j^z \partial_{z_j}. \tag{5}$$

Nyní můžeme konečně přistoupit k definici infinitezimálních symetrií.

Definice 2.6. Mějme zadanou soustavu rovnic (1). **Infinitezimální symetrie** této soustavy je vektorové pole na $\mathbb{R}^3$ splňující podmínku

$$(\mathrm{pr}^N X)F_a\big|_{F=0} = 0, \forall a \in \hat{K} \tag{6}$$

tedy aby prolongovaná vektorová pole anihilovaly funkce F_a na množině řešení.

Takováto vektorová pole se automaticky uzavírají do Lieovy algebry.

Rovnice (6) je v definici použita v situaci, kde známe F_a a hledáme X. Naším cílem bude použít ji opačně, tedy dívat se na ní jako na rovnici určující F_a pro zadaná vektorová pole X. Ukazuje se, že většinou stačí uvažovat lépe řešitelný tvar rovnice (6).

Definice 2.7. Mějme zadaná vektorová pole $X_j \in \mathfrak{X}(\mathbb{R}^3)$. Funkce $I : J^\infty \to \mathbb{R}$ je **silný invariant**, pokud splňuje rovnici

$$(\mathrm{pr}\, X_j)I = 0, \forall j. \tag{7}$$

Nás zajímají především netriviální silné invarianty. Pokud existují, není třeba hledat takzvané slabé invarianty, podrobnosti lze najít v [1]. Navíc z charakteru rovnic (7) plyne, že lze najít funkcionální bázi silných invariantů, to jest množinu funkcionálně nezávislých silných invariantů $I_1, \ldots, I_k$, ze kterých lze pomocí nějaké funkce nakombinovat libovolný další invariant J řádu N (obsahující nejvýše N-té derivace). Tedy $J = G(I_1, \ldots, I_k)$. Všechny rovnice, které mají pole X_j jako své symetrie jdou pak zapsat právě ve tvaru

$$G_a(I_1, \ldots, I_k) = 0. \tag{8}$$

Poslední věc, jež je třeba definovat je realizace Lieovy algebry pomocí vektorových polí. Proč nás to vlastně zajímá? Protože když hledáme rovnice pro fyzikální model, jenž má mít nějaké symetrie, obvykle nemáme zadaná vektorová pole, ale známe jen abstraktní Lieovu algebru, t.j. komutační relace. A pro hledání silných invariantů je potřeba mít vektorová pole.

Definice 2.8. Realizací Lieovy algebry L pomocí vektorových polí na M nazveme věrnou reprezentaci L do algebry vektorových polí na varietě M.

Na realizacích je definována relace ekvivalence pomocí bodových transformací variety M.

Naše zadání je tedy následující: Pro zadanou abstraktní Lieovu algebru L najděte všechny diferenciální rovnice N-tého řádu se dvěma závislými proměnnými, jež mají symetrie, které se uzavírají do algebry izomorfní s L.

Jako první krok při řešení provedeme klasifikaci realizací zadané Lieovy algebry vektorovými poli na $\mathbb{R}^3$. Poté najdeme silné invarianty pro reprezentanta z každé třídy realizací, a to pomocí rovnice (7). Hledané rovnice jsou pak funkcionální kombinace těchto invariantů a to pro každého reprezentanta zvlášť.

3 Demonstrace metody

Metodu hledání rovnic s předepsanými symetriemi, kterou jsme popsali na konci předchozí sekce budeme demonstrovat na sérii konečněrozměrných nilpotentních Lieových algeber, která byla podrobněji zkoumána v [5]. Pro dané $n \geq 4$ máme Lieovu algebru $\mathfrak{n}_{n+4,3}$ s dimenzí $n+4$, jejíž nenulové komutační relace jsou

$$\begin{aligned} [e_k, d_1] &= e_{k-1}, \text{ pro } k \in \hat{n}, \\ [f, d_f] &= e_0, \\ [d_f, d_1] &= f, \end{aligned} \tag{9}$$

kde $(e_0, \ldots, e_n, d_1, f, d_f)$ je báze $\mathfrak{n}_{n+4,3}$. Bazické vektory jsou, v zájmu jednoduchosti vzorců a kvůli rozdílnému charakteru vektorů e_i od ostatních, značeny trochu jinak než v [5].

První krok naší metody je nalézt a klasifikovat všechny realizace $\mathfrak{n}_{n+4,3}$ na $\mathbb{R}^3$. Díky tomu, že je tato algebra nilpotentní, existuje v ní posloupnost do sebe vnořených ideálů s kodimenzemi jedna. Jednodimenzionální algebru lze vždy realizovat pomocí pole ∂_y. Pak postupujeme induktivně. Nechť máme realizovaný ideál J_i. Přidáme k němu obecné vektorové pole a zjistíme, v jakém tvaru musí být, aby mělo správné komutační relace a tím vznikla realizace ideálu J_{i+1} s dimenzí o jedna větší. Pokud takové pole existuje, je naším dalším úkolem zjistit, jak vypadají bodové transformace, jež nechají invariantní tvar vektorových polí realizujících J_i. Tyto transformace využijeme ke klasifikaci realizací J_{i+1}. V zásadě se snažíme co nejvíce zjednodušit tvar přidávaného vektorového pole.

Neekvivalentní realizace pro $\mathfrak{n}_{n+4,3}$ jsou dvě.

Realizace A

$$\begin{aligned} E_k &= \rho_a(e_k) = \frac{x^k}{k!}\partial_y, \\ D_1 &= \rho_a(d_1) = -\partial_x, \\ F^a &= \rho_a(f) = \partial_z, \\ D_f^a &= \rho_a(d_f) = z\partial_y + x\partial_z. \end{aligned} \tag{10}$$

Realizace B

$$\begin{aligned} E_k &= \frac{x^k}{k!}\partial_y, \\ D_1 &= -\partial_x, \\ F^b &= z\partial_y, \\ D_f^b &= xz\partial_y - \partial_z. \end{aligned} \tag{11}$$

Dalším krokem je napočítání prolongací. To se nemusí vždy povést, obzvlášť, když ho máme provést pro obecnou dimenzi algebry a pro prolongaci libovolného stupně. Ale v pro náš případ to naštěstí jde.

Prolongace realizace A

$$\begin{aligned}
\mathrm{pr}^N E_k &= \sum_{j=0}^{N} \frac{x^{k-j}}{(k-j)!}\partial_{y_j},\\
\mathrm{pr}^N D_1 &= -\partial_x,\\
\mathrm{pr}^N F^a &= \partial_{z_0},\\
\mathrm{pr}^N D_f^a &= x\partial_{z_0} + \partial_{z_1} + \sum_{j=0}^{N} z_j\partial_{y_j}.
\end{aligned} \tag{12}$$

Prolongace realizace B

$$\begin{aligned}
\mathrm{pr}^N E_k &= \sum_{j=0}^{N} \frac{x^{k-j}}{(k-j)!}\partial_{y_j},\\
\mathrm{pr}^N D_1 &= -\partial_x,\\
\mathrm{pr}^N F^b &= \sum_{j=0}^{N} z_j\partial_{y_j},\\
\mathrm{pr}^N D_f^b &= -\partial_{z_0} + xz\partial_{y_0} + \sum_{j=1}^{N}(xz_j - jz_{j-1})\partial_{y_j}.
\end{aligned} \tag{13}$$

Teď když máme napočítané prolongace již nám nic nebrání v hledání silných invariantů řešením rovnice (7). Používá se k tomu metody charakteristik nebo metody pohyblivých reperů[1]. Je výhodné postupovat v podobném pořadí ve kterém jsme konstruovali použitou realizaci.

V následujícím seznamu je uvedena funkcionální báze silných invariantů. Pokud nás zajímají konkrétní tvary obecných diferenciálních rovnic, stačí vzít invarianty do N-tého řádu, vytvořit z nich libovolnou funkci, položit ji rovnu nule a nakonec nahradit y_i a z_i i-tými derivacemi.

Výsledky pro realizaci A

$$\begin{aligned}
&z_2,\ z_3,\ z_4,\ \ldots\\
&y_{n+1} - z_{n+1}z_1,\ y_{n+2} - z_{n+2}z_1,\ y_{n+3} - z_{n+3}z_1,\ \ldots
\end{aligned} \tag{14}$$

Výsledky pro realizaci B

$$\begin{aligned}
&z_1,\ z_2,\ z_3,\ \ldots\\
&y_{n+1+j} - z_{N+1+j}\frac{y_{n+1}}{z_{n+1}} - (n+1)\frac{z_n z}{z_{n+1}} + (n+1+j)z_{(n+j)}z,
\end{aligned} \tag{15}$$

kde $j \in \mathbb{N}$.

[1] Neplést s metodou pohyblivých rapperů.

Pro ilustraci jen uveďme, že jeden ze systému rovnic, mající za symetrie realizaci A (pro $n = 4$) je třeba

$$\begin{aligned} y^{(5)} - z^{(5)} z' &= z^{(4)} \sin(z'' \cdot z'''), \\ z^{(5)} &= \cos(z'' \cdot z''' + z^{(4)}). \end{aligned} \tag{16}$$

4 Závěr

V práci je po nezbytném úvodu, kde se definují základní věci týkající se infinitezimálních symetrií, ve stručnosti popsána metoda hledání diferenciálních rovnic s předepsanou algebrou symetrií. Uvedená metoda je vzápětí úspěšně demonstrována na sérii nilpotentních algeber.

Další potenciálně zajímavé směry, kterými se můžeme z tohoto místa vydat a prozkoumat je, jsou například systémy diferenčních rovnic, které potom lze použít na tvorbu takzvaných diferenčních schémat. To jsou diferenční rovnice, jež jistým způsobem, který respektuje symetrie, aproximuje jisté diferenciální rovnice. Diferenčních schémat pak můžeme využít pro numerické výpočty.

Literatura

[1] J. F. Cariñena, M. A. del Olmo, and P. Winternitz. *On the relation between weak and strong invariance of differential equations.* Lett. Math. Phys. **29** (1993), 151–163.

[2] E. J. Cartan. *Sur la structure des groupes de transformations finis et continus.* PhD thesis, École Normale Supérieure, Paris, (1894).

[3] E. E. Levi. *Sulla struttura dei gruppi finiti a continui.* Atti Accad. Sci. Torino **40** (1905), 551–565.

[4] M. S. Lie and F. Engel. *Theorie der Transformationsgruppen I.* Teubner, Leibzig, (1888).

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

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

[7] L. Šnobl and P. Winternitz. *All solvable extensions of a class of nilpotent Lie algebras of dimension n and degree of nilpotency n − 1.* J. Phys. A **42** (2009), 105201, 16 pp.

Použití metody Verlet pro simulaci dopravy

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 v Praze

Abstract. In this article a commonly used integration method named Verlet is introduced. This numerical method frequently used to integrate Newton's equations of motion, and so to calculate trajectories of particles in molecular dynamics simulations, offers great stability, is time invariant, energy conserving and preserve the symplectic structure of the phase space. The next step is to explore the possibility of using the modifying called the adaptive Verlet method. This approach is based on a time reparametrization, which led to improvement in the behavior of the numerical method. This method could be the proper scheme for numerical integration of one-dimensional short-range thermodynamic partical gas called Dyson gas, used for traffic modeling.

Keywords: verlet integration, Dyson gas, traffic modeling

Abstrakt. V tomhle příspěvku je představena běžně používána metoda pro integraci nazývaná Verlet. Tuhle numerickou metodu lze využít na integraci Newtonových pohybových rovnic a výpočet trajektorií částic při simulaci molekulární dynamiky. Nabízí dostatečnou stabilitu, časovou invariantnost, zachování energie a symplektické struktury fázového prostoru. Dalším krokem je prozkoumání modifikace metody nazývané adaptivní Verlet. Tenhle přístup je založen na reparametrizaci, která vede ke zlepšení chování numerické metody. Tahle metoda může být správnou možností pro numerickou integraci jednodimenzionálného krátkodosahového termodynamického částicového plynu nazývaného Dysonův plyn využívaný pro modelování dopravy.

Klíčová slova: integrace Verlet, Dysonův plyn a modelování dopravy

1 Úvod

Tato práce se zaobírá Verletovým algoritmem, který představuje často využívanou metodu na numerickou integraci pohybových rovnic. Pojmenování získala po francouzském fyzikovi Loupovi Verletovi, který tuhle metodu v roce 1967 zpopularizoval pomocí svého slavného díla "Computer Experiments on Classical Fluids". Avšak již předtím byla používána norským matematikem Carlem Stromerem k výpočtu trajektorií částic pohybujících se v magnetickém poli, a proto je také známá jako Stromerova metoda. Výhodou Verletovy integrace je větší stabilita v porovnání s jednodušší Eulerovou metodou. Také disponuje důležitými vlastnostmi jako časová reverzibilita a zachování plochy.

2 Základní Verlet

Základní Verletův algoritmus se používá na integraci Newtonových pohybových rovnic pro uzavřený systém N částic

$$M\ddot{\mathbf{x}}(t) = F(\mathbf{x}(t)) = -\nabla V(\mathbf{x}(t))$$

kde $\mathbf{x}(t)$ je soubor polohových vektorů, V je skalární funkce pro potenciál, F je negativní gradient potenciálu a M je hmotnostní matice. Po zjednodušení dostáváme rovnici

$$\ddot{\mathbf{x}}(t) = A(\mathbf{x}(t)),$$

kde A reprezentuje vektorovou funkci zrychlení závislou na pozicích částic.
Ve většině případů je dána počáteční konfigurace $\vec{x}_0 = \vec{x}(0)$ a počáteční rychlosti $\vec{v}_0 = \dot{\vec{x}}(0)$. Pak je vybrán vhodný časový krok $\Delta t > 0$ a pozice částic $\vec{x}_n$ jsou počítány v okamžicích $t_n = n\Delta t$. Sekvence $\vec{x}_n$ by pak měla nýt dostatečně blízko přesnému řešení $\vec{x}(t_n)$.
Verletova metoda používá centrální diference pro aproximaci druhé derivace, zatímco u Eulerovy metody dopředí diference aproximují první derivace.

$$A(\vec{x}_n) = \frac{\Delta^2 \vec{x}_n}{\Delta t^2} = \frac{\frac{\vec{x}_{n+1}-\vec{x}_n}{\Delta t} - \frac{\vec{x}_n - \vec{x}_{n-1}}{\Delta t}}{\Delta t} = \frac{\vec{x}_{n+1} - 2\vec{x}_n + \vec{x}_{n-1}}{\Delta t^2}.$$

Tedy, pro Verletův algoritmus je potřeba znát dva předchozí vektory pozic částic pro spočítání nové konfigurace.

$$\vec{x}_{n+1} = 2\vec{x}_n - \vec{x}_{n-1} + \vec{a}_n \Delta t^2, \tag{1}$$

kde $\vec{a}_n$ je zkrácený zápis pro $A(\vec{x}_n)$. Základem je jednoduše Taylorův rozvoj do třetího řádu

$$\vec{x}_{n+1} = 2\vec{x}_n + \vec{v}_n \Delta t + \frac{1}{2}\vec{a}_n \Delta t^2 + \frac{1}{6}\vec{b}_n \Delta t^3 + O(\Delta t^4)$$

$$\vec{x}_{n+1} = 2\vec{x}_n - \vec{v}_n \Delta t + \frac{1}{2}\vec{a}_n \Delta t^2 - \frac{1}{6}\vec{b}_n \Delta t^3 + O(\Delta t^4).$$

Je zřejmé, že předchozí rovnice vznikla sečtením zmíněných Taylorových rozvojů a lokální chybový člen představuje $O(\Delta t^4)$.

2.1 Rychlostní Verlet

Základní Verletův algortimus nezahrňuje výpočet rychlostí. Avšak často je vyhodnocení rychlosti potřeba, čemu vděčí za oblibu rychlostní Verletova metoda. Tenhle postup je velice blízce příbuzný klasickému Verletově algortimu.
Standardní implementace obsahuje čtyři kroky

1. $\vec{v}(t + \frac{1}{2}\Delta t) = \vec{v}(t) + \frac{1}{2}\vec{a}(t)\Delta t$
2. $\vec{x}(t + \Delta t) = \vec{x}(t) + \vec{v}(t + \frac{1}{2}\Delta t)\Delta t$

3. calculate $\vec{a}(t+\Delta t)$ as a function of $\vec{x}(t+\Delta t)$ from the interaction potential
4. $\vec{v}(t+\Delta t) = \vec{v}(t+\frac{1}{2}\Delta t) + \frac{1}{2}\vec{a}(t+\Delta t)\Delta t$.

Postup lze skrátit na

1. $\vec{x}(t+\Delta t) = \vec{x}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}(t)\Delta t^2$
2. calculate $\vec{a}(t+\Delta t)$ as a function of $\vec{x}(t+\Delta t)$ from the interaction potential
3. $\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{1}{2}(\vec{a}(t) + \vec{a}(t+\Delta t))\Delta t$.

I když je výpočet rychlostí součástí algoritmu, zrychlení $\vec{a}(t+\Delta t)$ je pořád závislé pouze na pozicích částic $\vec{x}(t+\Delta t)$.

3 Adaptivní Verlet

U popsané Verleté metody může docházet k problémům v případě nelineárních systémů, především v okolí singularit vektorového pole.
To bylo motivací pro pokusy vyvinout algoritmus, který by dovoloval změnu časového kroku v závislosti na aktuálně potřebné přesnosti. V blízkosti fixních bodů a singularit integrace vyžaduje velice krátký časový krok. Velikost šasového kroku může být korigována funkcí závisející na lokálním odhadu chyby. Takové schéma by navíc mělo zajistit zachování geometrických struktur systému.
Zmíněné pokusy přinesli mimo jiné algoritmus nazvaný Adaptivní Verlet [1], metodu založenou na dynamickém přeškálování vektorového pole. Pro reparametrizaci času se používá vhodná hladká funkce $R(u)$, přičemž $0 < m < R < M$, kde m a M udávají minimální a maximální poměr u změny časového kroku.
Autonómní diferenciální rovnice v $\mathbf{R}^N$

$$\frac{d}{dt}u = f(u)$$

je tedy změněna pomocí reparametrizující funkce $R(u)$ na

$$\frac{d}{ds}u = \frac{f(u)}{R(u)}, \frac{dt}{ds} = \frac{1}{R(u)}.$$

Uvedená reparametrizace j irelevantní z pohledu orbitů fázového prostoru, protože tok zachovává všechny integrální invarianty. Je možné uvažovat několik přístupů k volbě funkce R. Jednou z intuitivních variant je výběr heuristiky $R = \|f\|$, kde $\|.\|$ představuje Euclidovou normu. Jinou možností je R založeno na lokálním odhadu chyby nebo rozmístnění částic v mnohočásticovém systému.
Pro předvedení Adaptivní Verletovy metodyjsou obzvlášť vhodné systémy s odděleným Hamiltoniánem $H(p,q) = \frac{1}{2}p^t M^{-1} p + V(q)$, kde pohybové rovnice mají tvar:

$$\dot{q} = M^{-1}p, \dot{p} = -\nabla V.$$

Po reparametrizaci se uvažuje funkce pro změnu časového parametru ρ jako nová proměnná s vlastní diferenciální rovnicí. To vede k systému

$$\frac{dq}{ds} = \frac{1}{\rho} M^{-1} p,$$
$$\frac{dp}{ds} = -\frac{1}{\rho} \nabla V(q),$$
$$\rho = R(q, p).$$

Diskretizace popsaného systému vede k postupu vyžadujícímu pouze jeden výpočet $-\nabla V(q_n)$ v každém kroku.

$$q_{n+1} = q_n + \frac{\Delta s}{\rho_{n+\frac{1}{2}}} M^{-1} p_{n+\frac{1}{2}},$$
$$p_{n+\frac{1}{2}} = p_{n-\frac{1}{2}} - \frac{\Delta s}{2} \nabla V(q_n) \left(\frac{1}{\rho_{n-\frac{1}{2}}} + \frac{1}{\rho_{n+\frac{1}{2}}} \right),$$
$$\rho_{n+\frac{1}{2}} + \rho_{n-\frac{1}{2}} = R(q_n, p_{n+\frac{1}{2}}) + R(q_n, p_{n-\frac{1}{2}}),$$
$$t_{n+1} = t_n + \frac{\Delta s}{\rho_{n+\frac{1}{2}}}.$$

4 Dysnonův plyn

Dále se bude pozornost věnovat možnostem využití zmíněných metod pro numerickou integraci Dysnova plynu, který slouží mimo jiné pro modelovaní dopravy. Jedná se o jednodimenzionální termodynamický částicový plyn. Mikroskopická struktura vykazuje stejné statistické rozdělení jako u reálné cestné dopravy, proto se používá pro její modelování.

Systém se skládá z N identických vozidel reprezentujících jednotlivé vozidla umístněné na kruhu s obvodem $L = N$. Potenciální energie obecně závisí na odpudivém potenciálu V, přičemž v Dysonově plyně interakcím částic odpovídá Coulombův potenciál

$$V = -\sum_{i=j+1,j+2,\ldots,j+h} \ln(|x_i - x_j|),$$

kde x_i znamená souřadnici označující polohu $i$té částice a h představuje počet sousedících částic. U klasického Dysonova modelu byl použitý dalekodosahový potenciál, což značí, že všechny částice navzájem reagovali. Avšak nedávné výzkumy zavádí trend preferovat při modelování dopravy krátkodosahový potenciál, který lépe odpovídá reálným interakcím pozorovaným v dopravních vzorkách. zvolena bylo konkrétně hodnota $h = 1$, čemu odpovídá Hamiltonián tvaru

$$H = \frac{1}{2} \sum_{i=1}^{N} (v_i - \bar{v})^2 + C \sum_{i=1}^{N} V(r_i),$$

kde v_i označuje rychlost $i$té částice a $\bar{v}$ průměrnou rychlost ve vzorku. Funkce odpudivého potenciálu je zjednodušena na $V(r_i) = -\ln(r_i)$, přičemž r_i reprezentuje vzdálenost mezi $i$tou a předchozí částicí. Celý soubor je navíc umístněn v teplotné lázni s termodynamickou teplotou T. Bylo dokázáno [2], že popsaný model má souvislost jak s dopravním systémem tak i steorií náhodných matic.
Zajímat nás bude stav po dosažení termální rovnováhy, konkrétně rozdělení vzdáleností sousedících částic (tzv. spacing distribution),které má tvar

$$P_\beta(r) = \frac{(\beta+1)^{\beta+1}}{\Gamma(\beta+1)} r^\beta \exp\left[-(\beta+1)r\right],$$

kde β je inverzní termodynamická teplota získaná pomocí vztahu $\beta = \frac{1}{kT}$, kde k reprezentuje Boltzmannův faktor.

4.1 Numerická integrace

Numerické schéma pro integraci Dysonova plynu musí být časově invariantní, zachovávat energii a symplektickou strukturu fázového prostoru. To splňují zmíněné verze Verletova algoritmu. Jako nejvhodnější se jeví Adaptivní Verlet s funkcí pro reparametrizaci $R(x)$. Pak uvažujeme rovnice

$$\frac{d}{ds}x = \frac{v}{R(x)},$$
$$\frac{d}{ds}v = -\frac{\nabla V(x)}{R(x)},$$
$$\frac{d}{ds}t = \frac{1}{R(x)},$$

kde x je vektor pozic, v vektor rychlostí, V funkce odpudivého potenciálu a s nová časová proměnná.
Funkce reparametrizace je dána předpisem $R(x) = \max_i |a_i(x)|$, kde $a(x)$ reprezentuje zrychlení závislé na pozicích. Difernciální rovnice jsou převedeny na numerické schéma

$$x_{n+1} = x_n + \frac{\Delta s}{\rho_{n+\frac{1}{2}}} v_{n+\frac{1}{2}},$$

$$v_{n+\frac{1}{2}} = v_{n-\frac{1}{2}} - \frac{\Delta s}{2}\nabla V(v_n)\left(\frac{1}{\rho_{n-\frac{1}{2}}} + \frac{1}{\rho_{n+\frac{1}{2}}}\right),$$
$$\rho_{n+\frac{1}{2}} = 2R(x_n) - \rho_{n-\frac{1}{2}},$$
$$t_{n+1} = t_n + \frac{\Delta s}{\rho_{n+\frac{1}{2}}},$$

kde ρ_n představuje zkrácený zápis pro $R(x_n)$.
K dosažení rovnováhy je potřeba přidat termalizační krok představující přeškálování rychlostí

$$v = \frac{v}{\sqrt{2\beta E_{kin}}},$$

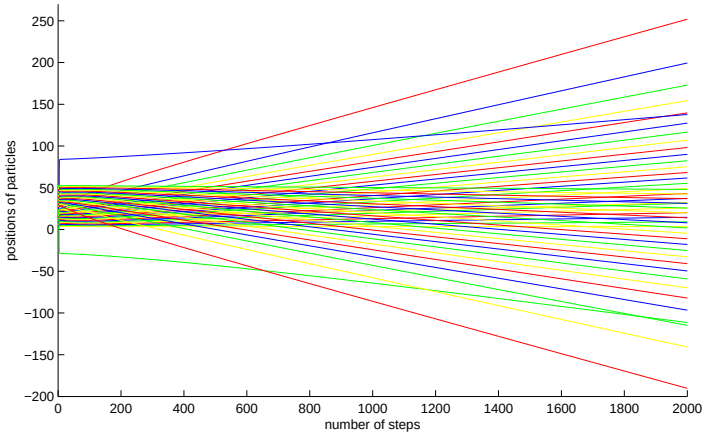
kde E_{kin} označuje kinetickou energii vzorku a β je již zmíněná inverzní termodynamická teplota.

5 Problematika počátečních hodnot

Implementace popsaného postupu přináší další problém v podobě volby vhodných počátečných hodnot.
K demonstraci téhle problematiky nám nejlíp poslouží implementace rychostního Verletova algoritmu. Je potřeba definovat počátečné rozmístnění pomocí vektoru pozic x_0, dále vektor počátečných rychlostí v_0 a časový krok Δt.

5.1 Časový krok Δt

Časový krok má v porovnání s ostatními počátečnými hodnotami veličin vetší váhu. Krátký časový krok znamená víc vyčíslení hodnot veličin a větší výpočetní náročnost. na druhou stranu příliš dlouhý časový krok může způsobit předbíhání, ke kterému v uvažovaném systému nemá docházet. Změna pořadí částic navíc způsobí nestandardní situaci a algoritmus neobsahuje postupy pro její zvládnutí. Předbíhaní tedy způsobí, že obě částice, které změnily pořadí, dramaticky zvýší svou rychlost a předbíhající částice se bude pohybovat ve směru pohybu ostatných částic zatímco předběhnutá částice se začne pohybovat opačným směrem a obě budou protínat trajektorie dalších vozidel. Stejný efekt nastane v případě implementace Adaptivní Verletovy metody, kdy je zvolený příliš dlouhý počátečný časový krok, který způsobí předbíhání.



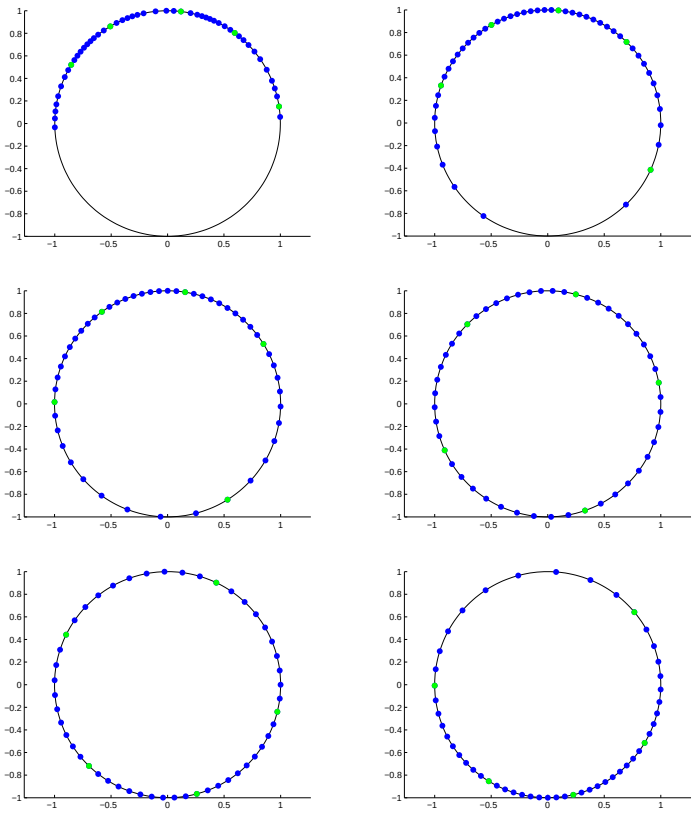
Trajektorie částic v případě, že dojde k předjíždění

5.2 Počátečné rychlosti v_0

Díky termalizačnímu kroku, který přeškáluje rychlosti, odeznívá vliv počátečných rychlostí obzvláště rychle, vetšinou hned po prvním přeškálování. proto k problémům způsobeným volbou počátečních rychlostí může prakticky dojít pouze v průbehu prvního kroku, než jsou přeškálovány termalizačním krokem. U počátečných rychlostí není ani tak důležitá jejich absolutní hodnota, jak jejich rozdělení. Velký rozdíl v počátečných

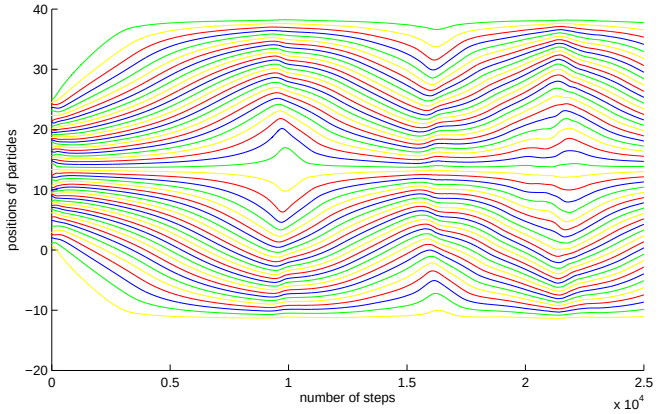
rychlostech dvou sousedících částic může vést k předbíhání a pak dalšímu nepříznivému vývoji popsanému v predchozím odstavci.

5.3 Počáteční konfigurace x_0



Vývoj při nevyvážené počáteční konfiguraci s konstantními časovými rozestupy.

Jelikož systém spěje k termální rovnováze je zřejmé že algoritmus se spíš vyrovná s počáteční konfigurací bližší rovnovážnému stavu. zajímavé je tedy podívat se, jestli si poradí s vysoce nevyváženou konfigurací.
Vhodným kandidátem, kde leze lehce vypozorovat onu nevyváženost je situace, kde jsou všechny částice umístněné pouze na jedné polovině uvažovaného kruhu. Jak se můžeme přesvědčit provedením odpovídající simulace, i v tomhle případě systém konverguje k rovnovážnému stavu. Částice se budou vzájemně odpuzovat co způsobí jejich přesun na druhou polovinu kruhu dokud se jejich koncentrace výrazně nezvýší, pak dojde k opačnému pohybu a další fluktuaci částic z jedné poloviny kruhu na druhou a zpátky, avšak s klesající tendencí až nakonec fluktuace úplně vymizí a systém přejde do rovnovážného stavu.



Trajektorie částic při nevyvážené počáteční konfiguraci.

6 Závěr

Jak klasická Verletova metoda, tak Adaptivní Verlet představují účinné nástroje pro numerickou simulaci pohybu částic Dysonova plynu. Tahle simulace má analogii s pohybem reálných vozidel v dopravním vzorku a proto může sloužit k modelování dopravy. Jediným úskalým je problematika volby vhodných počátečných parametrů.

Literatura

[1] W. Huang and B. Leimkuhler. The Adaptive Verlet method. *SIAM J. SCI. COMPUT*, **18**, 239–256, 1997.

[2] M. Krbálek, P. Šeba, P. Wagner. Headways in traffic flow: Remarks from a physical perspective. *PHYSICAL REVIEW E*, **64**, 066119, 2001.

[3] M. Krbálek. Equilibrium distributions in a thermodynamical traffic gas. *J. Phys. A: Math. Theor.*, **40**, 5813, 2007.

Numerical Programming on GPU

Vladimír Klement

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

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

Abstract. This article deals with the use of modern graphics cards for numerical computations, specifically with the means of acceleration of iterative solvers for sparse matrices. The first part is devoted to graphics cards in general, their advantages and reasons why they should be suitable hardware architecture for numerical problems. Following are described the main aspects of graphics cards, that distinguish them from other parallel architectures and which should be taken into account for effective parallelization on GPUs. In the rest of the article two specific numerical problems, their implementation on the GPU, and achieved speed-ups are presented. The first one is the problem of image segmentation on structured mesh using level set method, where the resulting linear system is solved by the SOR method. The second problem is concerned with simulation of 2D incompressible air flow governed by the Navier-Stokes equations and solved by multigrid method. The results show that in both problems GPU implementation achieved significant speed-up compared to CPU version.

Keywords: GPU, Multigrid, Red-Black Gauss Seidel, Red-Black SOR

Abstrakt. Tento článek se zabývá využitím moderních grafických karet k numerickým výpočtům, konkrétně možností zrychlení iteračních řešičů pro řídké matice. První část se věnuje grafickým kartám obecně jejich přednostem a důvodům, proč by měly být vhodný hardware pro numerické vypočty. Následně jsou popsány hlavní specifika grafických karet, jež je odlišují od ostatních paralelních architektur a která je potřeba vzít v potaz pro efektní paralelizaci na GPU. Ve zbytku článku jsou pak rozebrány dva konkrétní numerické problémy, jejich implementace na GPU a dosažené zrychlení. První problém je segmentace obrazu na strukturované síti pomocí level set metody, kde výsledný lineární systém je řešen SOR metodou. Druhý problém se potom zabývá simulací 2D nestlačitelného proudění tekutin pomocí Navierových-Stokesových rovnic a po diskretizaci je řešen metodou multigridu. Naměřené výsledky ukazují, že pro oba řešené problémy se implementací na GPU podařilo dosáhnout značného urychlení.

Klíčová slova: GPU, Multigrid, Red-Black Gauss Seidel, Red-Black SOR

1 Introduction

Graphics cards are special piece of hardware designed to improve visual quality of computer games. Their architecture differs significantly from that of a normal processors and due to their highly parallel nature, they are expected to outperform CPUs by an increasing margin in parallelizable calculations [10] [7]. At first their capabilities were limited to few fixed types of graphical computations, but nowadays graphics cards contain fully

programmable graphical computation units (*GPUs*) and can be used for a large range of problems.

Main advanteges of graphics cards compared to standard CPUs are:

- More processing units
- Faster arithmetic computations
- Higher memory bandwidth
- Faster growth of computational power

Our goal, is to find whether graphics card can be used to speed-up computation of numerical problems leading to large system of linear equations with sparse matrix. Solution of such problems is typically limited by computational power and memory bandwidth. Since graphics cards greatly outperform processors in both these parameters, they should be very suitable for this tasks.

2 GPU programming

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

There are several technologies, that lets programmer create application for GPU but most important are [6]:

- OpenGL - It is cross-platform graphical API so basic knowledge about computer graphics is needed and general problems have to be inconveniently masked as a graphical ones. This was the first way how graphics card can be used to solve general problems, but nowadays this isn't commonly used any more.
- CUDA - Is a technology from NVidia company designed specifically for general purpose computing on graphics card, main disadvantage is that it only works for NVidia graphics cards. Pluses are that it is quickly developed and there exist a lot of example and documentation for it.
- OpenCL - Newest technology for general computation on graphics card, same as OpenGL, it is an industry standard and so it can be used for almost all new devices ranging from graphics cards to cell phones.

In our programs we use CUDA rather than OpenCL. However core parts of both these technologies are very similar, in essence, the main difference is naming of the functions.

Compared to other types of parallel programming (i.e. OpenMp, MPI), programming for graphics cards have some specifics given by the type of calculations graphics cards were designed for. It is important to know them and keep them in mind when creating program for GPU in order to fully utilize it's potential. In rest of this section the most important ones will be point out.

2.1 Limited communication

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

2.2 Branching

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

2.3 Coalescing

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

- Best solution, if it is possible, is to reorder data so that access to them will be coalesced. One classic example is to use structure of arrays instead of array of structures (i.e. group data by type, not by the thread they belong to).
- Threads in the same block can pre-fetch data to shared memory (shared within block), even random accesses to this memory are very cheap. This is especially usefull when needed data form a continuous region, but are accessed randomly.
- If data are needed to be ordered differently in different kernels they can be duplicated (unless memory is a strong concern) this can be especially useful in case of constant data (for example data describing mesh on which problem is solved).

2.4 Transports between GPU and CPU memory

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

3 Image segmentation

First problem, that we will present is considered with image segmentation, which is one of the main parts of image recognition. It deals with the problem of dividing image to the number of non-overlapping regions corresponding to the objects in input image.

One of the possibilities, how to compute segmentation of an image is the so called *Level Set Method* (first introduced in [9]). This method represents solution of the problem as zero level set of some implicit functions which's development in time is governed by the level set equation. Main advantages of this method is no need for parametrization of segmented area and ability to change topology, main drawback is great computational complexity, since the equation has to be solved for every pixel of the image.

3.1 Linear system

Level set problem has the form of

$$\begin{aligned} u_t = \sqrt{\epsilon^2 + |\nabla u|^2}\nabla \cdot \left(g^0 \frac{\nabla u}{\sqrt{\epsilon^2 + |\nabla u|^2}} \right) \qquad na \quad (0,T\rangle \times \Omega, \\ u(0,x) = u^0(x) = |s - x| - R \quad na \quad \Omega, \\ u(t,x) = u^0 \qquad pro \quad (0,T\rangle \times \partial\Omega, \end{aligned} \tag{1}$$

Where u is the level set function, ϵ is small regularization constant, g^0 is edge function of input image and Ω is the area the input image is defined on.

By the time discretization of this equation, we obtain

$$\frac{h^2}{|\nabla u_{i,j}|^{n-1}} \frac{u^n_{i,j} - u^{n-1}_{i,j}}{\tau} = \sum_{(i',j') \in C_{i,j}} \left(\frac{g^0_T}{|\nabla u^{i',j'}_{i,j}|^{n-1}} \right) (u^n_{i',j'} - u^n_{i,j}). \tag{2}$$

which can be altered to the form of

$$\left(\frac{h^2}{|\nabla u_{i,j}|^{n-1}} + \tau \sum_{(i',j') \in C_{i,j}} \left(\frac{g^0_T}{|\nabla u^{i',j'}_{i,j}|^{n-1}} \right) \right) u^n_{i,j} - \tau \sum_{(i',j') \in C_{i,j}} \left(\frac{g^0_T}{|\nabla u^{i',j'}_{i,j}|^{n-1}} \right) u^n_{i',j'} = \frac{h^2}{|\nabla u_{i,j}|^{n-1}} u^{n-1}_{i,j},$$

which is a linear system with solution u^n. This system has same number of rows as there are values $u^n_{i,j}$ and each row has only 5 non-zero elements (one on diagonal and one for each neighbouring pixel). We will denote matrix of this system A diagonal, elements (belonging to pixel with coordinates (i,j)) $A^{i,j}_{i,j}$, and non-diagonal elements $A^{i',j'}_{i,j}$. Right hand side will be denoted b and its elements $b_{i,j}$. All these values can be computed as:

$$A^{i',j'}_{i,j} = -\tau \frac{g^0_T}{|\nabla u^{i',j'}_{i,j}|^{n-1}} \quad pro \quad (i',j') \in C_{i,j},$$

$$A^{i,j}_{i,j} = \frac{h^2}{|\nabla u_p|^{n-1}} + \sum_{(i',j') \in C_{i,j}} \left(\frac{g^0_T}{|\nabla u^{i',j'}_{i,j}|^{n-1}} \right) = \frac{h^2}{|\nabla u_p|^{n-1}} - \frac{1}{\tau} \sum_{(i',j') \in C_{i,j}} A^{i',j'}_{i,j},$$

$$b_{i,j} = \frac{h^2}{|\nabla u_{i,j}|^{n-1}} u_{i,j}^{n-1},$$

Matrix A is sparse, symmetric, and diagonally dominant.

This scheme is semi-implicit because solution from the previous time-step is used to create values of A. It is also possible to recompute A from current solution during each iteration of matrix solver and so proceed to implicit scheme. Such a version needs more computations but less memory bandwidth.

3.2 SOR method

For solving this linear system standard SOR method was used. Each iteration of this method starts with approximate solution $\vec{x^k}$ and finds new better one $\vec{x^{k+1}}$ via the formula

$$x_i^{k+1} = (1-\omega)x_i^k + \frac{w}{A_{ii}}\left(b_i - \sum_{j>i} A_{ij}x_j^k - \sum_{j<i} A_{ij}x_j^{k+1}\right),$$

where ω is a chosen constant (relaxation factor), $i = 1, 2, ..., n$ and n is the size of $\vec{x}$.

This iterations are repeated until the solution error is sufficiently small. Squared error after l-th iteration is given by

$$R^l = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} A_{ij}x_j^l - b_i\right)^2.$$

3.3 Parallelization

Issue with SOR method is that it is inherently sequential and so it can't be used on GPU. Therefore we had to switch to the so called Red-Black SOR[3] method. This method consist in dividing SOR iteration to two steps (red and black). During each step only elements which's actualizations are independent (and so can be done in parallel) are processed. Because in our problem only neighbouring elements affect each other during actualization, the final splitting must fulfil the condition that no two neighbouring elements can be in the same group. On structured square grid this can be easily achieved by dividing elements based on the sum of their indices to odd and even.

Implementation of this algorithm on GPU wasn't particularly complicated, but there were some issues with memory coalescing and border communication that needed to be taken into account. Our final implementation works like this (example is given only for the red step, the black one would be similar):

1. Because our system represents 2D domain we will use 2D coordinates.
2. Launch SOR kernel with blocks of 16x16 threads and grid with enough blocks to have one thread per red element (or little more if number of elements isn't multiple of 256).
3. Each block will load 34x18 (32x16 active area with margin of one element) values of x^k to shared memory, because this region is compact whole read can be coalesced.

4. Following instructions are given for each thread
5. Compute coordinates of red element from 32x16 area belonging to this thread.
6. Fetch(semi-implicit) or compute(implicit) values of A for this element. In semi-implicit case this data can be ordered, so that this operation will be coalesced.
7. Compute new value of x^{k+1} for this element.
8. Save the new value, this won't be coalesced because value for only red elements don't form continuous block.

3.4 Results

The results were obtained on computer equipped with AMD 2.4GHz processor, 4GB RAM, and GeForce GTX 480 graphics card. First we will compare the speed of CPU vs GPU for both methods. Where CPU version uses standard SOR method, and GPU version uses Red-Black SOR method.

Semi-implicit	64x64 px	128x128px	256x256px	512x512px
GPU time	4 s	4 s	11 s	33 s
CPU time	18 s	87 s	349 s	1438 s
Speed-up	4,5	21,75	31,72	43,5

Table 1: Comparison of semi-implicit version on CPU and GPU for different image sizes.

Implicit	64x64 px	128x128px	256x256px	512x512px
GPU time	6 s	9 s	22s	67 s
CPU time	239 s	1097 s	4454 s	17832 s
Speed-up	39,8	121,8	202,2	266

Table 2: Comparison of implicit version on CPU and GPU for different image sizes.

And then we will compare all the version together.

From tables 1 a 2 it is apparent, that GPU has better speed-up for larger problems. From table 3 it can be also seen that implicit method is much slower than semi-implicit method, even though the difference on GPU isn't that large. This shows that GPU is extremely well suited for tasks which's bottleneck is computational power.

4 Airflow simulation

The second problem, which's implementation on GPU will be presented, is the simulation of air flow over urban canopy governed by the system of viscous incompressible Navier-Stokes equations (taken from [1]).

	64x64 px	128x128px	256x256px	512x512px
CPU Semi-implicit	1	1	1	1
GPU Semi-implicit	4,5	21	32	43
CPU Implicit	0,075	0,079	0,078	0,080
GPU Implicit	3	9,66	15,86 7	21,46

Table 3: Relative speed-up of all methods.

4.1 Problem

The problem is given by the system:

$$\frac{\partial \mathbf{u}(t,\mathbf{x})}{\partial t} + \mathbf{u}(t,\mathbf{x}) \cdot \nabla \mathbf{u}(t,\mathbf{x}) - \nu \triangle \mathbf{u}(t,\mathbf{x}) + \nabla p(t,\mathbf{x}) = 0, \tag{3a}$$

$$\nabla \cdot \mathbf{u}(t,\mathbf{x}) = 0. \tag{3b}$$

where $\mathbf{x} = (x, y)$, $\mathbf{u}$ stands for the flow velocity and p for pressure.

The semi-implicit Oseen scheme is used for the time discretization of (3).

$$\frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{\tau} + \mathbf{u}^{n-1} \cdot \nabla \mathbf{u}(t,\mathbf{x}) - \nu \triangle \mathbf{u}(t,\mathbf{x}) + \nabla p(t,\mathbf{x}) = 0, \tag{4}$$

where $\tau > 0$ is the time step, $\mathbf{u}^n(\mathbf{x}) = \mathbf{u}(n\tau, \mathbf{x})$ and $\mathbf{p}^n(\mathbf{x}) = \mathbf{p}(n\tau, \mathbf{x})$.

In space, the problem is discretized by the non-conforming Crouzeix-Raviart finite elements and the convective term is stabilized through upwinding. At each time level linear system of the form

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{5}$$

is solved.

4.2 Geometric multigrid

Final system is solved by *geometric multigrid* method. Term multigrid covers group of methods for solving systems of partial differential equations with use of hierarchical structure of grids with different numbers of elements.

They are typically used for numerical solution of partial differential equations of elliptic type in two or more dimensions. They are compatible with all common discretization techniques and can be used for unbalanced and non-linear systems of equations, such as our system of Navier-Stokes equations.

Conventional iterative methods quickly eliminate the oscillatory components of the error (high frequencies), while the smooth components have low reduction rate of $1-O(h^2)$, this renders them inefficient for large systems. To overcome this issue multigrid methods solve problem simultaneously on smaller grids, where former smooth components become more oscillatory and thus can be easily eliminated by conventional solvers (referred to as

smoothers in the context of multigrid algorithms). This makes multigrid methods much faster than standard solvers.

But to formulate problem on coarser grids and use solution from them to improve solution on the finer grids the so called *transition operators* will be needed. The exact appearance of these operators depends on the chosen discretization.

4.3 Parallelization

First part that was implemented on GPU was smoother, because most of the computational time was spent in it. We used block Gauss-Seidel type smoother in our original program which, same as SOR method, is inherently sequential. So again we have used the red-black version. However since this problem is solved on unstructured triangular grid, which couldn't be generally coloured by two colors, this imposed a restriction on type of grids we can use. This is quite an issue mainly for meshes that are unevenly refined, and we are currently trying to solve it, but this is beyond the scope of this article. Other issue was related to the updates of velocities which are stored for edges not for triangles. Since the smoother is parallelized over triangles, each velocity is actualized by two threads (related to to two triangles, sharing the given edge). In order to prevent possible clashes, both red and black part of the smoother had to be divided to two steps, fist parallelized over triangles to compute and save all actualization values and second parallelized over edges to sum these values and actualize respectful velocities.

Next step was to convert transition operators, because they became slower than smoothing part, and moreover constant transfers of data from and to VRAM created quite an overhead. Implementation of these operators was straightforward, only issue being again actualization of edge values.

Unfortunately, after implementing whole multigrid part on the GPU creating the system between each steps became slowest and most limiting part of the computation. So in order to achieve best possible results we had to also implement system creation on GPU. Although it wasn't very complicated, since there were no parallel issues, it makes further modification of the program more problematic because now very large portion of the code has to be implemented on both CPU and GPU.

4.4 Results

The computations were done on system equipped by two CPUs AMD Opteron 6172 each having 12 cores running on 2.1 GHz. We have tested our GPU implementation on two cards. One was Nvidia Geforce GTX480 with 1.5 GB RAM and the other was Nvidia Tesla C2070 with 6GB RAM and ECC turned off. All simulations were computed in double precision.

Comparison of GPU and single core computation is in Table 4. The card GTX480 performs better but it is equipped with rather small amount of global memory. It did not allow us to run the largest simulation on this card. Tesla C2070 is a bit slower. However, on larger meshes the speed-up 25 can be achieved as well.

Table 5 presents speed-up of the GPU solver versus parallel multicore one. The multicore algorithm is very similar to the GPU one. The red-black colouring was applied

DOF	CPU	GTX 480		Tesla C2070	
	Time	Time	Speed-up	Time	Speed-up
310,848	59.7	3.4	**17.5**	4.4	**13.5**
1,244,288	393	14.9	**26.4**	19.26	**20.4**
4,978,944	2390	out of memory		96.2	**24.8**

Table 4: Performance comparison of the solver running on single core and on the GPU on three different meshes.

Cores	310,848 DOFs		1,244,288 DOFs		4,978,944 DOFs	
	Time	GPU Speed-up	Time	GPU Speed-up	Time	GPU Speed-up
1	83.3	18.9	561	29.1	3230	33.5
2	45.7	10.4	308	15.9	1770	18.4
4	25.5	5.8	175	9.1	1100	11.4
8	16.4	**3.7**	111	**5.8**	677	**7**
16	12.1	2.7	82.3	4.3	518	5.4
24	11.3	2.5	75.4	3.9	484	5

Table 5: Comparison of the computation time on GPU (Tesla C2070) and multicore Opteron. The times for 8 cores are in bold font because for more cores the efficiency is smaller than 0.5.

as well and the data were reorganised in memory with respect to the color of each triangle so that different cores do not compete for the same piece of memory. The code was parallelised by the OpenMP pragmas. Our code scales well only up to 8 cores. The reason is that the algorithm does not exhibit high arithmetic intensity and mainly the memory bandwidth is the limiting factor. The speed-up 7 was attained on 8 cores and if we omit the fact, that the parallel multicore algorithm has low efficiency on 24 cores, the speed-up here is 5. We also would like to comment different times measured with one core computation in Table 5 and Table 4. The reason is that in Table 5 the red-black colouring is used. This shows us impact of the colouring on the solver effectivity.

5 Summary

This article presented key principles of GPU programming and demonstrated them on two numerical problems. Both these problems were successfully implemented on GPU with significant speed-up. Therefore, conclusion of this article is that modern graphics cards are very suitable hardware for solving numerical problems and it can be very expedient to use them, even though they differ from standard parallel architectures and this differences has to be taken in account, when designing optimal algorithms.

References

[1] P. Bauer, *Dissertation thesis: Mathematical modelling of pollution transport in urban canopy*, ČVUT-FJFI, 2011

[2] V. Klement, *Diplomová práce: Implementace řešičů řídkých matic na GPU*, ČVUT-FJFI, 2011

[3] K. Mikula, A. Sarti, *Parallel co-volume subjective surface method for 3D medical image segmentation, in: Parametric and Geometric Deformable Models: An application in Biomaterials and Medical Imagery, Volume-II*, Springer Publishers, 2007

[4] K. Mikula, *Numerical Solution, analysis and application of geometrical nonlinear diffusion equations*, STU Bratislava, 2006

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

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

[7] J. D. Owens et al., *A survey of General-Purpose Computation on Graphics Hardware*, Computer Graphics Forum 26(1):80-113, 2007

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

[9] J. A. Sethian, *Level Set Methods: Evolving Interfaces in Computional Geometry*, Fluids Mechanics, Computer Vision, and Materials Science, Cambridge University Press, 1996

[10] J. Vacata, *Diplomová práce: Obecné výpočty na grafických procesorech*, ČVUT-FJFI, 2008

Application of a Degenerate Diffusion Method in 3D Medical Image Processing*

Radek Máca

3rd 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 [10]. This paper presents a 3D (2D+time) segmentation of the real cardiac MRI data using an algorithm based on a numerical solution of the partial differential equation of the level set type. The algorithm is derived from the level set equation using a semi-implicit complementary volume numerical scheme approximation. To apply the algorithm the correct set up of algorithm parameters is provided. In particular, the application is focused on the segmentation of the heart ventricles from the cine MRI data.

Keywords: cardiac MRI, co-volume method, image segmentation, level set method, PDE

Abstrakt. Tento příspěvek je rozšířeným abstraktem článku [10]. Tématem tohoto článku je segmentace 3D (2D+t) obrazových dat pomocí parciální diferenciální rovnice vrstevnicového typu. Algoritmus je odvozen z vrstevnicové rovnice při použití semi-implicitníhí časové diskretizace. K prostorové diskretizaci je použito schéma duálních objemů. Práce se dále zabývá vhodným nastavením výpočetních parametrů k dosažení co nejlepších výsledků při segmentaci levé a pravé srdeční komory na snímcích získaných pomocí magnetické rezonance.

Klíčová slova: metoda duálních objemů, segmentace obrazu, vrstevnicová rovnice

1 Introduction

In this paper we focus on the segmentation of the heart ventricles from cardiac MRI (CMR) data. CMR is a highly specialized imaging technique to examine the heart. In comparison with MR imaging of other organs, the CMR has to take into account the motion of the heart, breathing motion and the blood flow in the heart cavities. The images are usually acquired over several cardiac cycles triggered by the patient's ECG (Electrocardiography). There are several Cardiac MRI sequences used in clinical practice. The image data, we are focusing on, are obtained by the Cine MRI referring to an examination of the heart kinematics. The heart is covered by 2D planes with the spatial resolution about $2 \times 2 \times 10$mm. Therefore, the ventricles can be entirely covered by 10–15 slices. The temporal resolution ranges between 20ms and 60ms, i.e. the cardiac

*This work was supported by the project "Advanced Supercomputing Methods for Implementation of Mathematical Models" of the Student Grant Agency of the Czech Technical University in Prague No. SGS11/161/OHK4/3T/14 and the HPC-EUROPA2 project (project number: 228398) with the support of the European Commission – Capacities Area – Research Infrastructures.

cycle is usually covered by 15–50 time frames. For a detailed information about MRI and the heart ventricles segmentation from a medical point of view see [3], [5].

In terms of the CMR data dimension, there are three possibilities to segment this data. First, the data can be segmented separately each of other (see [1], [2], [8], [14], [16]). Second, we can join 2D images for a given time phase to get a 3D image of the ventricle. As we mentioned above, it means the resolution in the third dimension ranges between 10–15. On the other hand, we can join 2D images for a given slice to get a 3D (2D+t) image of the resolution in time between 15–50 (see [6], [12]). Last, we could join all 2D images together to get 4D (3D+t) image ([9], [13]).

A natural way to proceed the CMR data is to segment the data separately. One of the drawbacks of the 2D approach is a time discontinuity of the segmentation results. This problem could be solved using a 3D (2D+t) segmentation. Specifically, the 3D image is built of the time sequences of 2D images. This approach ensures the time continuity in the segmented data.

2 Mathematical model

The detection of image object edges belongs to main tasks in image segmentation. Edges in the input image $I^0 : \Omega \to \{0, 1, 2 \ldots, I_{\max}\}$, represented by the matrix $n_x \times n_y \times n_z$, where the third direction corresponds to the time of the processed data

$$\Omega = (0, n_x/n) \times (0, n_y/n) \times (0, n_z/n)\,, \qquad n := \max\{n_x, n_y, n_z\}\,,$$

can be recognized by the magnitude of its spatial gradient. We will use the following format of the CMR data size: $n_x \times n_y \times n_z \times n_s$, where n_s denotes number of slices. The level set equation operating in Ω can be modified as follows

$$\partial_t u = |\nabla u|_\varepsilon \nabla \cdot \left(g\left(\left|I^0 * \nabla G_\sigma\right|\right) \frac{\nabla u}{|\nabla u|_\varepsilon} \right) - g\left(\left|I^0 * \nabla G_\sigma\right|\right) |\nabla u|_\varepsilon F\,, \tag{1}$$

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 ([17] 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 $\lambda \geq 0$. $G_\sigma \in \mathcal{C}^\infty(\mathbb{R}^3)$ is a smoothing kernel, e.g. the Gauss function with zero mean and variance σ^2

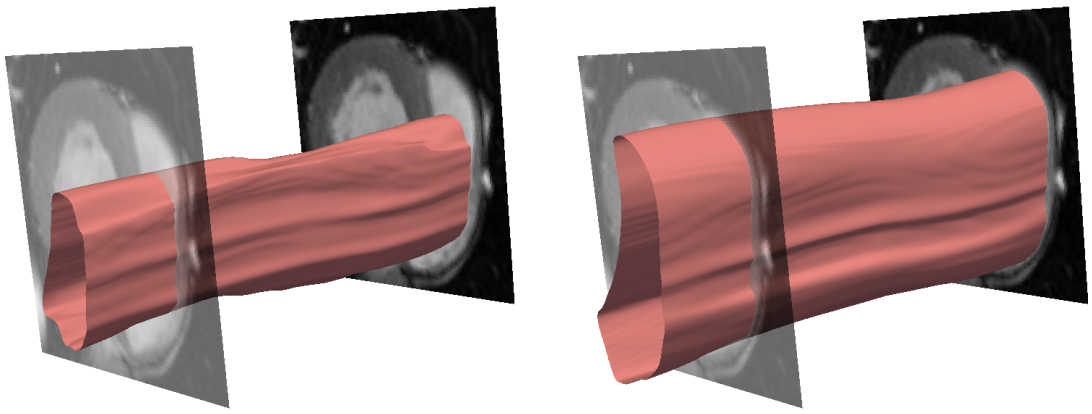
$$G_\sigma(\vec{x}) = \frac{1}{(2\pi)^{3/2}\,\sigma_x^3\sigma_y^3\sigma_z^3} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2} - \frac{z^2}{2\sigma_z^2} \right), \tag{2}$$

which is used to pre-smoothing (denoising) of image gradients by convolution

$$(I^0 * \nabla G_\sigma)(\vec{y}) = \int_{\mathbb{R}^3} \bar{I}^0(\vec{y} - \vec{x}) \nabla G_\sigma(\vec{x})\, \mathrm{d}\vec{x}\,, \tag{3}$$

where $\bar{I}^0$ is the extension of I^0 to $\mathbb{R}^3$ by, e.g., mirroring, periodic prolongation or zero padding. Let us note that equation (1) can be rewritten into the advection-diffusion form

$$\partial_t u = \underbrace{g^0 |\nabla u|_\varepsilon \nabla \cdot \left(\frac{\nabla u}{|\nabla u|_\varepsilon} \right)}_{(D)} + \underbrace{\nabla g^0 \cdot \nabla u}_{(A)} - \underbrace{g^0 |\nabla u|_\varepsilon F}_{(F)}\,. \tag{4}$$



(a) (b)

Figure 1: Right ventricle segmentation: (a) the segmentation surface after 20 time iterations, (b) final shape of the segmentation surface after 200 time iterations.

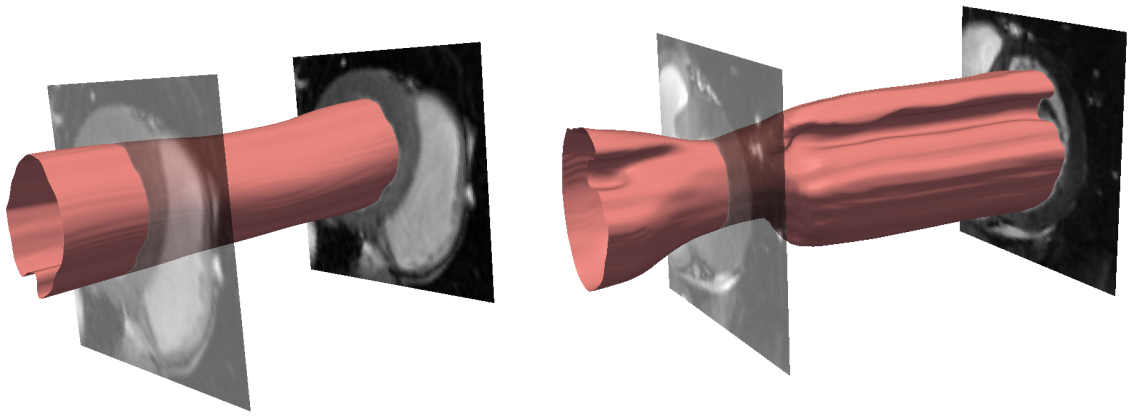
For convenience, the abbreviation $g^0 = g(|I^0 * \nabla G_\sigma|)$ is used. (D) in (4) denotes the diffusion term, (A) the advection term and (F) the external force term. The term g^0 is called the edge detector. The value of the edge detector is approximately equal to zero close to image edges (high gradients of input image). Consequently, the evolution of the segmentation function slows down in the neighbourhood of image edges. On the contrary, in parts of the image with constant intensity the edge detector equals one. The advection term attracts the segmentation function to the image edges.

3 Results

Given the extent of this contribution, the data for a single patient are chosen as an example of the segmentation results. The size of CMR data for this patient equals $128 \times 128 \times 26 \times 12$.

The result of the right ventricle segmentation is depicted in Fig. 1. In Fig. 1a we can see the shape of the segmentation surface after 20 time iteration. Stopping criterion terminated the segmentation process after 200 time iteration; Fig. 1b presents the result of segmentation process.

In the same way as we presented the results of the right ventricle segmentation, the results of the left ventricle segmentation is shown. The final shape of the segmentation surface after 250 time iteration shows Fig. 2a. The patient, we chose to present the result, has low contractility of myocardium. To see the difference between the final shapes of the segmentation surfaces for hearts with low and high contractility we apply our algorithm on a healthy volunteer. The result is depicted in Fig. 2b. We can clearly see that a shape of the segmentation surface is similar to the "cylinder" for a heart with low contractility, whereas the shape for the heart with high contractility could be compared to the "hourglass".



(a) (b)

Figure 2: Left ventricle segmentation: (a) result of segmentation for the patient (250 time iterations), (b) final shape of the segmentation surface for the healthy volunteer after 300 time iterations (data size: $128 \times 128 \times 80 \times 1$).

References

[1] Beneš, M., Chalupecký, V., Mikula, K.: Geometrical image segmentation by the Allen-Cahn equation, *Applied Numerical Mathematics 51*, 187–205 (2004)

[2] Beneš, M., Kimura, M., Pauš, P. and Ševčovič, D., Tsujikawa, T., Yazaki, Sh.: Application of a Curvature Adjusted Method in Image Segmentation, *Bulletin of the Institute of Mathematics*, Academia Sinica (New Series) 2008, 509–523 (2008)

[3] Bogaert, J., Dymarkowski, S., Taylor, A. M.: Clinical cardiac MRI, *Springer-Verlag*, Berlin Heidelberg, (2005)

[4] Cao, F.: Geometric Curve Evolution and Image Processing, *Lecture Notes in Mathematics*, No 1805, Springer Verlag, Février (2003)

[5] Chabiniok, R.: Personalized Biomechanical Heart Modeling for Clinical Applications, *Université Pierre et Marie Curie - Paris 6*, PhD thesis, 2011

[6] 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

[7] Evans, L. C., and Spruck, J.: Motion of level sets by mean curvature I, *J. Diff. Geom.*, Vol. 33, 381–635 (1991)

[8] Loucký J., Oberhuber T.: Graph cuts in segmentation of a left ventricle from MRI data, *Proceedings of Czech-Japanese Seminar in Applied Mathematics 2010, COE Lecture Note, 2012*, vol. 36, 46–54

[9] Lynch, M., Ghita, O., Whelan, P. F.: Segmentation of the Left Ventricle of the Heart in 3D+t MRI Data Using an Optimized Non-Rigid Temporal Model, *IEEE Transactions on Medical Imaging 27(2)*, 195–203 (2008)

[10] Máca R., Beneš M.: Application of a degenerate diffusion method in 3D medical image processing, *Algoritmy 2012 Proceedings of Contributed Papers and Posters*, Slovak University of Technology, Faculty of Civil Engineering, Bratislava, 427–437 (2012)

[11] Mikula, K.: Numerical solution, analysis and application of geometrical nonlinear diffusion equations, *Edition of Scientific Publications*, No. 34, Publishing House of the Slovak University of Technology, Bratislava (2006)

[12] 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, 583–626 (2005)

[13] Montagnat, J., Delingette, H.: 4D deformable models with temporal constraints: application to 4D cardiac image segmentation, *in Medical Image Analysis (MedIA)*, Vol. 9 (1), 87–100 (2005)

[14] Oberhuber T.: Complementary finite volume scheme for the anisotropic surface diffusion flow *Proceedings of Algoritmy 2009, Handlovičová A., Frolkovič P., Mikula K. and Ševčovič D. (ed.)*, 153–164 (2009)

[15] Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces, *Springer Verlag*, (2003)

[16] Paragios, N.: Variational Methods and Partial Differential Equations in Cardiac Image Analysis, *Invited Publication : IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, (2004)

[17] Perona, P., Malk, J.: Scale space and edge detection using anisotropic diffusion, *Proc. IEEE Computer Society Workshop on Computer Vision*, (1987)

[18] Sapiro, G.: Geometric Partial Differential Equations and Image Processing, *Cambridge University Press*, (2001)

[19] Sethian, J. A.: Level Set Methods, *Cambridge University Press*, (1996)

[20] Zhao, H. K., Osher, S., Chan, T., Merriman, B.: A variational level set approach to multiphase motion, *J. Comput. Phys. 127*, 179–195 (1996)

Distributed Data Processing in High-energy Physics*

Dzmitry Makatun†

1st year of PGS, email: makatun@rcf.rhic.bnl.gov
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Šumbera[1], Jérôme Lauret[2],
[1]Nuclear Physics Institute, AS CR;
[2]STAR, Brookhaven National Laboratory

Abstract. In this paper a brief introduction into requirements for a computational network in the High Energy Physics experiment STAR is done. The main part of the text describes a development of cache management for such a system. Different caching policies are discussed. The data access pattern was studied. Results of a computer simulation of cache performance for several policies are presented. In the end of the paper, the future direction of project is described.

Keywords: data transfer, cache, optimization, grid

Abstrakt. V tomto článku je krátký úvod do požadavků na výpočetní sítě v experimentu STAR. Hlavní část textu popisuje přípravu řízení mezipaměti pro takový systém. Dále představíme několik různých algoritmů ovládání mezipaměti a analýzu přístupu k datům. Výsledky počítačové simulace několik algoritmů pro praci s mezipamětí jsou prezentovány. Na konci tohoto článku je nastiněn budoucí směr projektu.

Klíčová slova: přenos dat, mezipaměť, plánování, grid

1 Introduction

The increased volume of data (up to an order of magnitude) provided by a new data acquisition system (DAQ1000), combined with the needs of an increasingly complex, resource hungry analysis and simulation planned for the future, lead to a projected storage need of 6482 TB and 117605 of CPU's for STAR experiment at RHIC by 2015. To meet these needs within the funding guidance of the BNL mid-term plan, careful optimization of the use of resources is required [1].

Research and implementation have been seldom in the field of efficient data distribution over the Wide Area Network (WAN). For data-intensive applications (as one used in High Energy Physics) effective use of available storage, computational and network resources is crucial for end-to-end application performance. Furthermore, in the advent of new distributed computing paradigms such as Cloud computing, the harvesting of widely fluctuating and volatile resources, accessible through Cloud providers, has been

*This work has been supported by the grant SGS12/198/OHK4/3T/14
†Nuclear Physics Institute ASCR

hindered by the lack of integration of such resources in a global planning strategy. In other words, a planner that takes into account the available CPU resources, storage and the interconnection between elements before the data processing, is a key to the best use of widely distributed resources.

During the previous period of work in the collaboration of the experiment *STAR at Brookhaven National Laboratory (USA)* with *Nuclear Physics Institute of Academy of Science of Czech Republic*, the software for optimization of data transfer in a distributed system was implemented [2], and the principles for fair-share scheduling of requests were developed [3]. These approaches can be applied to the data processing to fill the knowledge and availability gap in the area of global planning by extending the existing research. The final result will be the realization of a global data management system able to function independently and reason based on the available resources as well as adapt to fluctuations (such as network downtime or addition of cloud resources in a global resource pool).

The Reasoner for Intelligent File Transfer (RIFT) is software which allows optimization of data transfer in a computational network with the use of available transfer mechanisms (FDT [4], HPSS [5], Xrootd [6], etc.). It was developed in collaboration between NPI and BNL by PhD student Michal Zerola [2]. It consists of several components running at a central server and locally at each server participating in the network.

The system works in the following way:

1. Users submits requests for file transfers.
2. The central component (scheduler) processes requests and generate a transfer plan.
3. The components installed at each server perform transfers according to the plan.

The optimization of resource usage is due to the plan generated by the scheduler. This is the main component of the system. It is based on constrain programming. It takes a bunch of requests, data about network configuration and speed from a central database and generates an optimal plan.

2 Opened questions

Although RIFT has shown good performance in testing evaluation [2], there are several improvements to be done in order to include RIFT into production. These improvements are: caching, fair-share algorithms and coupling with CPU's.

After a file has been transferred by RIFT, its copy remains in cache at each server on the transfer path. These copies can be used after as a source for the next transfers. But since the cache space on a particular server is limited, it has to be periodically cleaned. The dedicated algorithms which selects files to delete or to remain in cache can improve the efficiency of caching and the performance of the whole system.

Fair-share policy prevents users subscribing many requests from blocking other users. But fair-share policy in some cases is in contradiction with the requirement for the optimizing utilization of the resource. Some balance between fair-share and utilization should be found.

In real world, users of computing facilities do not usually transfer files to the server where the processing takes place. Instead, they submit a job on some dataset to the

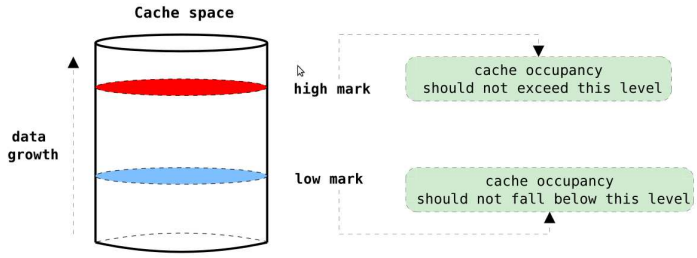


Figure 1: A schema of the watermarking concept.

job-scheduling system, the system allocates CPU for the job and then transfers data to it. This means RIFT should be either coupled with the existing job-scheduling software, or the logic for reasoning about the CPU allocation should be added.

In order to be applicable in the real world, the software should fulfil three more requirements: it must be scalable, automatic and flexible. The scalability means that new nodes (servers) can be added without rebuilding the system. The automation means that it can provide it's functionality without human intervention after the software is installed and configured . Lastly, the flexibility means that components of software running on a particular server can be configured for that particular server. The RIFT was developed on the basis of these principles and all further development also needs to fulfil them.

3 Cache study

The main recent improvement done to the RIFT is the addition of caching. The basic principles of caching, cache performance simulation and its implementation are discussed in this section.

3.1 Introduction to caching

The cache cleaning algorithms can be applied to keep the cache of data-transfer tools within defined limit or for cleaning local data storages. In the first case, the size of cache is small (several percent of the entire dataset) and the clean up has to take place regularly. In the second case, the task can be, for example, to delete a part of local data replica. In this case, the amount of data can scale up to the size of the entire dataset. In both cases, the problem is to select and delete files which are the least probable to be used. An investigation to find the most appropriate algorithm is required.

The water-marking is an approach dedicated for setting up the threshold when the cache clean-up starts and stops. It considers the current disk space occupied by cache. A high-mark and low-mark for cache size are externally set up. When the cache size exceeds the high-mark, the cache clean-up starts, and files are deleted until the cache size gets below the low-mark. The figure 1 illustrates the water marking concept.

The following cache algorithms were studied:

Least-Used (LU): evicts the set of files which were requested less times since they entered cache.

Least-Recently-Used (LRU): evicts the set of files which were not used for the longest period of time.

Most-Recently-Used (MRU): evicts the set of files which were used most recently. This algorithm can bring benefit for certain access patterns. For example, if files are requested sequentially, the last accessed file is the least likely to be requested next.

Least-Frequently-Used (LFU): evicts the set of files which were used least often.

Most Size (MS): evicts the set of files which have the largest size. It is preferable to keep smaller files in the cache in the case when retrieving a file from a storage produces an overhead which does not depend on the file size. A tape storage is an example.

Least Size (LS): evicts the set of files which have the smallest size. This policy can be efficient if larger files are being requested more often.

The selection of cache policy depends on user access pattern and disk space available. The efficiency of caching can be estimated with cache hits and cache hits per megabyte of data (cache data hits). The definition of parameters used for cache performance evaluation is given below.

3.2 Cache performance simulation

Let us consider a number of users requests N_{req} for files (possibly repeated) within a certain time window. Let each request have a count j, time of submission t_j and request for a file f_j of size S_j. Then the total size of requested files is

$$S_{req} = \sum_{j=1}^{N_{req}} S_j \tag{1}$$

Since many requests can ask for the same file, a set of N_{set} unique files can be selected. Let i be a count of each file in this set, S_i be a size of this file and R_i be the number of times the file was requested. Then the following equality takes place

$$N_{req} = \sum_{i=1}^{N_{set}} R_i \tag{2}$$

The total size of a dataset is

$$S_{set} = \sum_{i=1}^{N_{set}} S_i \tag{3}$$

Let us assume that the system has a single cache of size D_{cache}. Then for each request j a binary variable b_j can be assigned in a way that $b_j = 1$ (cache hit) if the file f_j appeared

Table 1: Average parameters of the access pattern in the experiment STAR in a period from 07.06 to 05.09.2012.

Number of requests	33×10^6
Amount of data transferred	49×10^{15} bytes
Maximal number of requests for one file	192
Minimal number of requests for one file	1
Average number of requests for one file	19

in cache at the time of request t_j and $b_j = 0$ (cache miss) if not. Then the total number of cache hits is

$$N_{cache} = \sum_{j=1}^{N_{req}} b_j \tag{4}$$

and the total size of files transferred from cache is

$$S_{cache} = \sum_{j=1}^{N_{req}} b_j \times S_j \tag{5}$$

Let us assume that at the initial moment the cache is empty. This means that files have to be requested for the first time before they can appear in the cache. Therefore, we can conclude, that the first request for the file should not be taken into account when evaluating cache performance. Then cache hits ratio (H) can be defined in a following way:

$$H = \frac{\sum_{j=1}^{N_{req}} b_j}{\sum_{i=1}^{N_{set}} (R_i - 1)} = \frac{N_{cache}}{N_{req} - N_{set}} \tag{6}$$

where equality 2 was applied. The cache hits ratio (H) defined this way can get values in [0,1]. We can also define cache data hits ratio (Hd) in a similar way:

$$Hd = \frac{\sum_{j=1}^{N_{req}} b_j \times S_j}{\sum_{i=1}^{N_{set}} (R_i - 1) \times S_i} = \frac{S_{cache}}{S_{req} - S_{set}} \tag{7}$$

Cache data hits ratio is a parameter to measure cache efficiency from the point of view of data flow, it can be explained as the ratio of data flow from cache to overall data flow. Since files in the set have different size, the cache hits ratio (H) and the cache data hits ratio (Hd) are not equal.

3.3 Access pattern

A real access pattern obtained from the experiment STAR was used for the simulation of cache performance. This pattern was extracted from the access log of the entire dataset. The access log of a period of 3 months was studied. Relevant parameters of the user access pattern are listed in table 1. The average number of requests per file is 19, which is promising for implementation of cache algorithm. We can also make a conclusion that

Table 2: Average parameters of the set of accessed files in the experiment STAR in a period from 07.06 to 05.09.2012.

Number of files	1.8×10^6
Total size	1.45×10^{15} bytes
Minimal file size	296 bytes
Maximal file size	5.3×10^9 bytes
Average file size	815×10^6 bytes

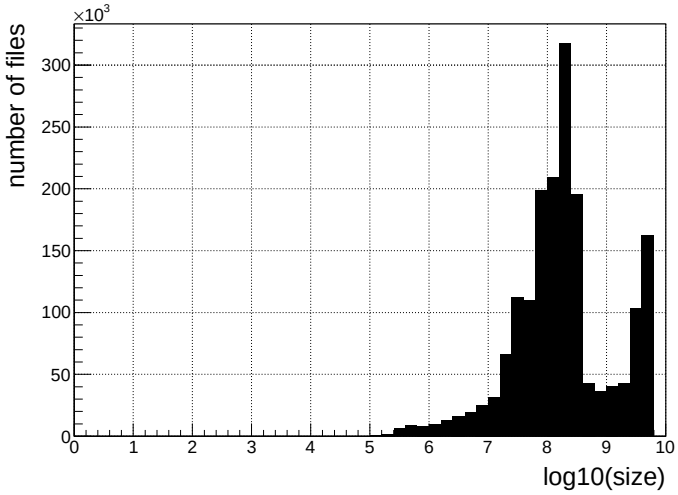


Figure 2: The set of accessed files in the experiment STAR in a period from 07.06 to 05.09.2012. Distribution of files by the logarithm of size.

requests are not distributed uniformly among files. This means there are files requested much more often than average, and that the algorithm which is able to keep those files in cache can deliver high cache performance.

From the list of all accessed files, a unique set of files can be subtracted. In a simulation this set plays the role of storage. Characteristics of the dataset obtained from the access pattern are presented in table 2. The distribution of files by size is presented on histogram in figure 2.

The access pattern can be represented as a contour plot (figure 3) where the axes are the number of requests for a particular file and the size of that file. Colour represents the number of files with the same coordinates on a plot. From the hot spots on this contour plot we can conclude that the most of the files in the set are small files that have been accessed only several times. Other hot spots of the contour plot are presented in table 3.

The further analysis of access pattern has shown that a subset of files that is 6.5% of the storage size can be selected in a way that 20% of requests are for the files from this subset. In the same time, this requests make 18% of the data traffic. In other words, with a cache size of 6.5% of the storage size a cache hit of 20% and a cache data hit of

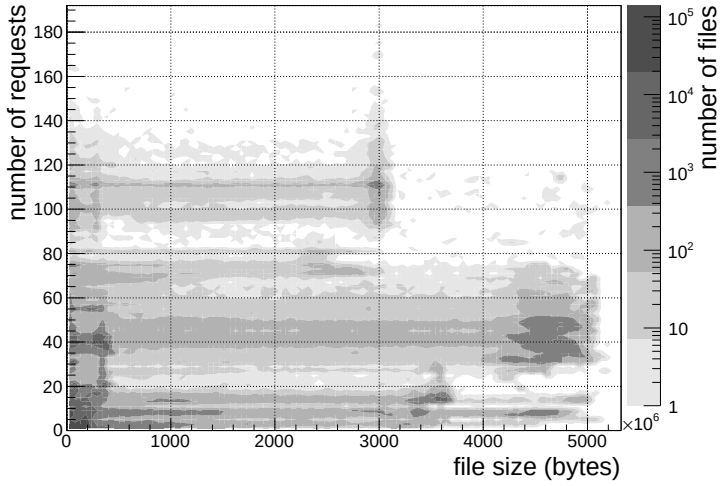


Figure 3: The access pattern in the experiment STAR in a period from 07.06 to 05.09.2012 as a contour plot. The colour represents the number of files with same coordinates on a plot. Several hot spots can be observed.

Table 3: Hot spots of the data access pattern. Several groups of files can be selected from the entire dataset to represent the most typical cases. The values are averaged.

Size	Requests	Number of files	Total size	Total requests	Total data flow
$(\times 10^9 bytes)$	-	(% of the data set)		(% of requests)	
0.1	10	58	8	16	1
0.1	40	15	6	32	4
3	110	0.8	3	5	10
3.6	15	1.5	7	1	3
4.8	40	6	37	15	46

18% can be achieved.

All the above mentioned makes it possible to conclude that implementation of cache into data-transfer system for high-energy physics can be efficient.

3.4 Results of simulation

A computer simulation of basic cache algorithms for the access pattern of high-energy physics data-processing was made. A cache hit ratio and a cache data hits ratio were obtained for different configurations of cache size and low mark (water marking) for different algorithms. The results of the simulation are presented in figures 4 and 5. There is a set of plots for different low marks (as ratio to the cache size) in this figures. The cache size is given as ratio to the size of storage (unique set of files).

There are two trivial cases which can help to verify results: if the cache size is close

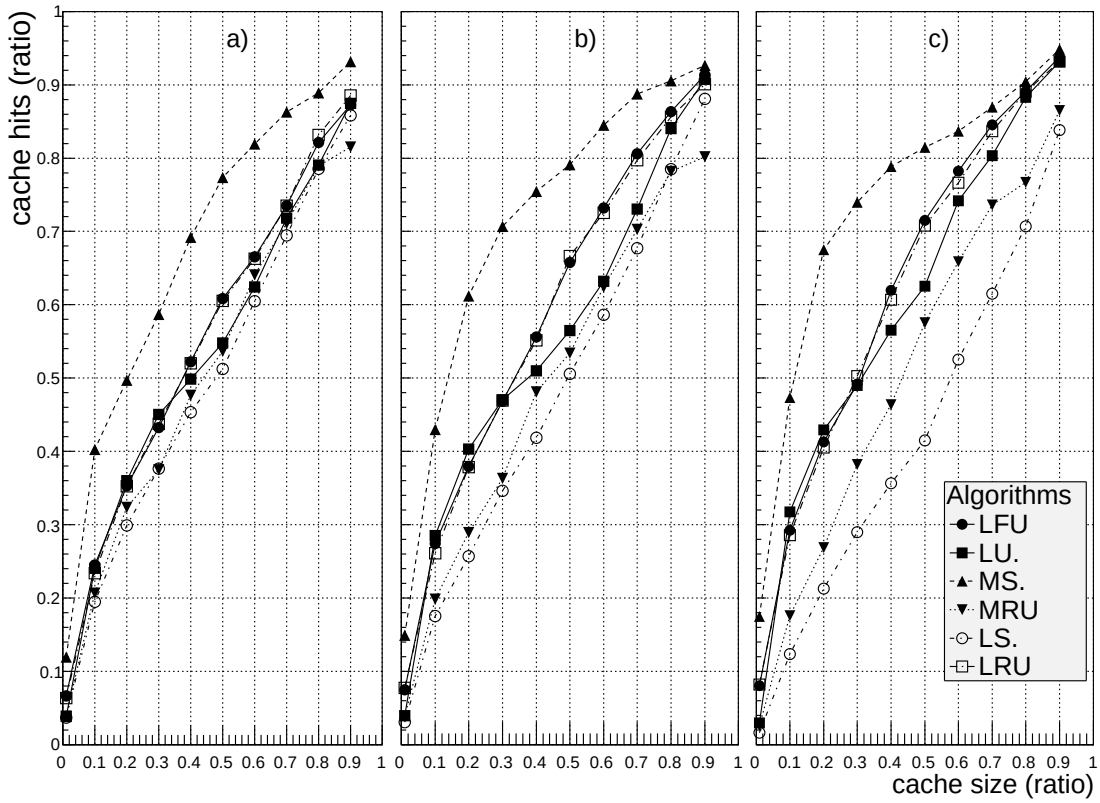


Figure 4: Results of simulation for 6 different algorithms. Cache hits as a function of cache size. Plots are given for different values of low mark: a) 25%, b) 50%, c) 75%.

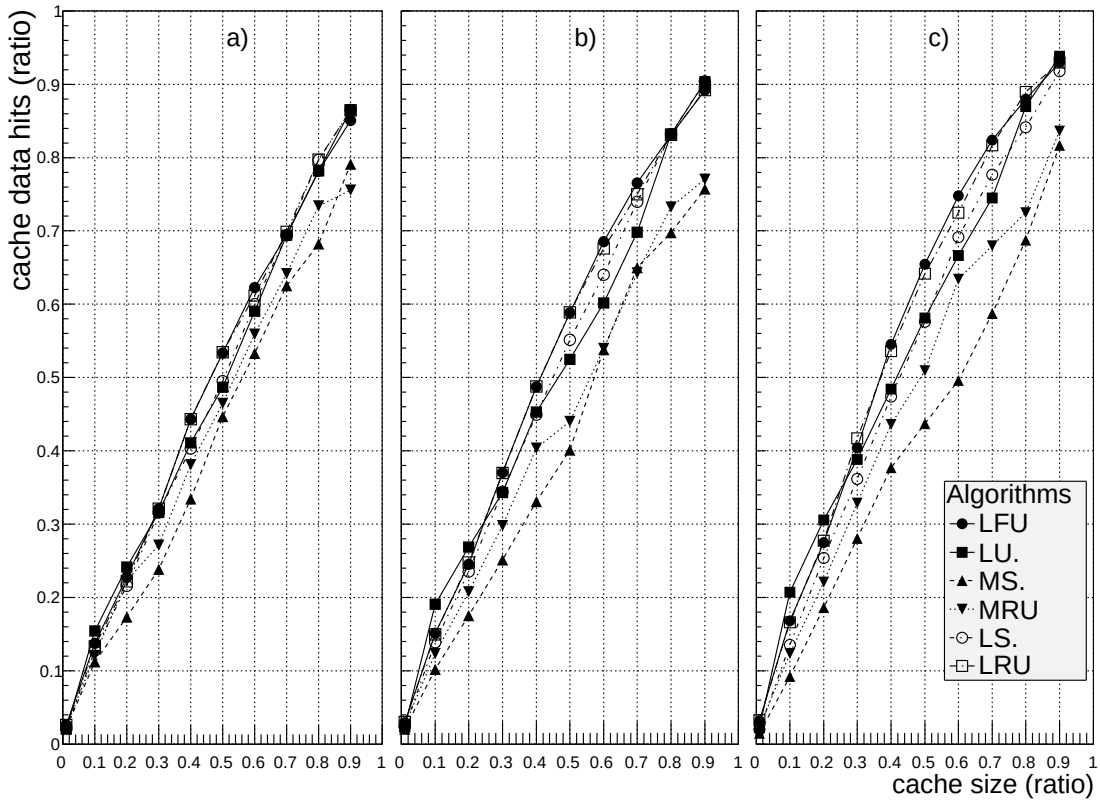


Figure 5: Results of simulation for 6 different algorithms. Cache data hits as a function of cache size. Plots are given for different values of low mark: a) 25%, b) 50%, c) 75%.

to 0 the cache hit is also close to 0, and if the cache size is 100% the cache hit is also close to 100% for all algorithms. It can be also observed on plots that when the low mark is larger, the difference between algorithms is more notable. That is because between clean-ups, the cache is being populated by files according to the access pattern, and it does not depend on the cache algorithm. With the larger low mark, clean-up takes place more often, which means the cache content is more controlled by the algorithm.

From the results of the simulation we can conclude:

- LFU and LRU algorithms have close efficiency both in terms of cache hits and cache data hits.
- LU has better efficiency than the two algorithms named above for small cache (up to 30%) but worse for larger.
- MS algorithm has the highest cache hit but the lowest cache data hits. This is appropriate for replicating data from storages with high latencies not dependent on file size (e.g. HPSS).
- MRU is less efficient than all named algorithms in all cases. Therefore we can conclude that this algorithm is not suitable for the studied access pattern.

3.5 Cache implementation

A cache algorithm was implemented to RIFT in a manner that the policy can be changed according to available cache space, access pattern and other parameters. This algorithm calculates a value of utility function for each file and then evicts files with the smallest value. By changing the utility function, the different policies can be applied.

A tool for cache management was implemented as a part of a component called watcher, which runs locally at each node in order to have actual information on cache status and be able to remove files. The tool also uses connection to the central database to use it's information, and to provide information to other components of the system. It sends requests to the central database in order to perform the following tasks:

- updates records in the database when new files appear in cache.
- deletes records from the database when files are deleted from cache.
- verifies that the content of the cache corresponds to records in the database, and if not, notifies what files or records are missing.
- before the deletion of selected file, verifies that this file is not being currently transferred. This prevents from deletion of required files.
- receives the data for the utility function calculation.

4 Conclusion and further plans

A data access pattern at the high-energy physics experiment STAR was analysed. The analysis showed a potential for implementation of caching for data transferring.

Based on the log files of access for the entire dataset, the computer simulation of cache performance was done. The cache hit for several algorithms was measured as a function of cache size and low mark setting. The comparison of existing algorithms and their suitability for the study case was made grounded on results of simulation.

Obtained results can also be applied to managing local data storages containing replication of the main dataset.

The cache management was implemented into RIFT. It allows to use the disk space which is available at servers in computational network, in order to decrease waiting time for data requests and to reduce network load. The cache management includes watermarking concept. The cache algorithm was implemented in a way that the cache policy can be selected depending on available cache space, access pattern and other parameters.

The final goal of the project is to develop a complete end-to-end global optimization system for data-processing, that automatically submits requested jobs to CPUs and delivers data to that CPUs.

In order to achieve this, the following steps should be taken: resolve the optimization problem with constrained programming method for allocating CPU's, integrate RIFT with job submitting environment, implement a fair-share algorithm and to continue development on principles of scalability, flexibility and automation.

Acknowledgements

The support of the grant SGS12/198/OHK4/3T/14 is gratefully acknowledged. The author would also like to thank his supervisors Michal Šumbera from NPI ASCR and Jérôme Lauret from STAR BNL in USA and graduated PhD student of CTU in Prague Michal Zerola for provided help and all the members of Ultra-relativistic Heavy Ion Group at Bulovka for collaboration.

References

[1] Jérôme Lauret, Tim Hallman *The Solenoidal Tracker At RHIC (STAR) Computing Resource Plan* Jan. 14, 2009

[2] Michal Zerola . *Distributed Data Management in Experiments at RHIC and LHC* PhD thesis, CVUT 2012

[3] Pavel Jakl *Efficient access to distributed data: A "many" storage element paradigm* PhD thesis, CVUT 2010

[4] Fast Data Transfer *Project web-site: http://monalisa.cern.ch/FDT/*

[5] High Performance Storage System *Project web-site: http://www.hpss-collaboration.org/*

[6] Xrootd *Project web-site: http://xrootd.slac.stanford.edu/*

[7] Jagdish Prasad Achara , Abhishek Rathore , Vijay Kumar Gupta and Arti Kashyap. *An improvement in LVCT cache replacement policy for data grid.* LNMIIT (Jaipur, India, 2010)

[8] Song Jiang, Xiaodong Zhang *Efficient Distributed Disk Caching in Data Grid Management.* Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER'03) 0-7695-2066-9/03, 2003 IEEE

Quality of Fractographic Sub-Models via Cross-Validation*

Matej Mojzeš

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

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

Abstract. Fatigue crack growth rate may be explained using linear regression to model the relationship between fatigue crack growth rate and fracture surface textural features. It may be useful to add to the model non-linear transformations of the basic linear features. However, the resulting extended model will probably be significantly more complex. Therefore an optimization heuristic, which is proposed in this paper, could be utilized to evaluate quality of different subsets of these explanatory variables using statistical tests or information criteria. As a conclusion of cross-validation analysis on our experimental results we are providing a list of evaluation methods that could be generally used.

Keywords: Sub-model, fractographic analysis, linear regression, heuristics, statistical testing, information criterion, cross-validation

Abstrakt. Rýchlosť šírenia únavovej trhliny môže byť vysvetľovaná lineárnou regresiou modelujúcou vzťah medzi rýchlosťou rastu trhliny a texturálnymi vlastnosťami povrchu trhliny. Mohlo by byť prínosné pridať do modelu nelineárne transformácie pôvodných lineárnych vlastností, avšak výsledný rozšírený model bude pravdepodobne podstatne zložitejší. Preto môže byť použitá v publikácii navrhovaná optimalizačná heuristika na hodnotenie kvality rôznych submodelov vysvetľujúcich premenných využívajúc štatistické a informačné kritériá. Ako záver krížovej validácie na experimentálnych dátach ponúkame zoznam hodnotiacich metód, ktoré by mohli byť všeobecne použiteľné.

Kľúčové slová: Submodel, fraktografia, lineárna regresia, heuristika, štatistické testovanie, informačné kritérium, krížová validácia

1 Introduction

One of the tasks of quantitative fractography consists in modelling of the relation between fatigue *crack growth rate* (velocity, CGR) and *textural features* of images of fatigue fracture surfaces [10]. For this purpose, either a multilinear regression model or a neural network may be used. Of these two possibilities the latter allows us to analyze the structure of the model obtained and to describe and better imagine the textural subset which is mutually related with the CGR.

The parameters of respective regression model may be estimated using the least squares method. However, in real-world applications the basic linear model is not flexible

*This paper has been supported by the grant OHK4-165/11 CTU in Prague

enough to fit the data. This can be solved by adding terms defined by non-linear functions of basic features, e.g. logarithm, second root, etc. However, adding such features is soon limited by the given number of images.

According to [10] one possible way around this limitation is a two-phase stepwise regression with the first stage being a bottom–up stepwise regression beginning with constant model and terminating at a given over-fitting level p_0. In each iteration a new explanatory variable is included - the one which maximally decreases the sum of squares of residui. The second stage is top–down stepwise regression beginning with the final sub–model from the first stage and terminating at given final over-fitting level p_F. In this procedure, an explanatory variable is selected for the elimination via Wald test on a selected critical level.

While keeping in mind the relevant motivation to this problem, we suggest that instead of the stepwise regression, an alternative statistical approach could be based on the method of sub-model multiple testing. There is a vast set of possible criteria that evaluate the quality of a given sub-model and are to be minimized. Selection and assessment of some of them, which are interesting in the fractographic context, but may be applied generally in multi-parametric recognition, etc., is elaborated further in this paper.

2 Linear model

Let denote v_j the crack growth rate assigned to the j-th image of the fracture surface, and f_{uj} the set of image features. The simplest form of a multilinear model is

$$\log_{10} v_j \approx c_0 + \sum_u c_u f_{uj} \ . \tag{1}$$

Parameters c_u can be estimated by the least squares method. Since the linear model is not flexible enough to fit the data we may add different non-linear functions of basic features and therefore modify the model to the following form:

$$\log_{10} v_j \approx c_0 + \sum_q c_q h_q \ . \tag{2}$$

where h's are selected from an extended set of features containing the features f_u and a selection of basic non-linear functions of them, e.g.

$$\{h_i\} \subset \{f_u, \log_{10} f_u, f_u^{-1}, f_u^{1/2}, f_u^2\} \ . \tag{3}$$

The next task will consist of defining a specific methodology how to select and assess a distinct combination of explanatory variables from the extended feature set, or a sub-model.

3 Sub-model selection

The sub-model should be regarded as a nested subset of the full model including all the explanatory variables from the entire set of extended features. The are two extreme cases - first is the full model and the second one corresponds to constant model.

Let $n \in \mathbb{N}$ be the length of the vector v (the number of observations), $m \in \mathbb{N}$ the cardinality of the extended feature set and $k \in \{0, 1, \ldots, m\}$ the number of explanatory variables from the extended feature set used in the sub-model. Moreover, let $\boldsymbol{c} = (c_0, c_1, \ldots, c_k)$ be the vector representing coefficients of sub-model calculated solving eq. (2), $\boldsymbol{c}_{\text{red}} = (c_1, c_2, \ldots, c_k)$ its significant part and $\boldsymbol{c_0} = (c_0)$ the coefficient of the constant term.

Then, we may denote SSQ the sum of squares for the optimum $\boldsymbol{c}$ of given sub-model and SSQ_0 the sum of squares for $\boldsymbol{c_0}$. Last, but not least, we will make use of the error of sub-model defined as follows:

$$s_{\text{e}}^2 = \frac{SSQ}{n-k-1} \ . \tag{4}$$

At this point, we should choose some of the many possibilities for testing a sub-model quality. We have selected a few of them, that can be divided in two sets, based on the concepts they are based on. The first one comprises traditional statistical tests and the criterion that will reflect the quality of a sub-model will be logarithm of the p_{value}. On the other hand, the second set contains different statistical information criteria regarding model selection. In the latter case, we are simply minimizing the value of the respective information criterion.

3.1 Sub-model testing

In this case we will be testing significance of the vector $\boldsymbol{c}_{\text{red}}$ representing the given sub-model. Corresponding hypotheses may be defined as:

- $H_0 : \boldsymbol{c}_{\text{red}} = \boldsymbol{0}$
- $H_1 : \boldsymbol{c}_{\text{red}} \neq \boldsymbol{0}$

and we will be using the McFadden R-square test and Wald test and their testing criteria.

McFadden R-square test

Should we use the McFadden R^2 test to analyse sub-model and constant model according to variance analysis [4], we should define a stochastic variable F as follows:

$$F = \frac{SSQ_0 - SSQ}{SSQ} \cdot \frac{n-k-1}{k} \ . \tag{5}$$

Variable F has distribution $\text{F}_{k,n-k-1}$ and the corresponding p_{value} is then calculated as $p_{\text{value}} = 1 - \text{F}_{k,n-k-1}(F)$.

Wald test

Alternatively, if we decide to incorporate the Wald test to compare distinct sub-models [7], following variable Z is to be considered:

$$Z = \frac{1}{k\, s_{\text{e}}^2} \cdot \boldsymbol{c}^{\text{T}}\, \boldsymbol{W}^{-1} \boldsymbol{c} \ . \tag{6}$$

Matrix $\boldsymbol{W}$ represents the matrix resulting from $(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X})^{-1}$ without both the first row and first column. Then, the variable Z has distribution $\mathrm{F}_{k,n-k-1}$ and the $p_{\mathrm{value}} = 1 - \mathrm{F}_{k,n-k-1}(Z)$.

Finally, for both of the tests, the resulting value of sub-model quality criterion to be minimized can be defined as

$$CRIT = \log_{10} p_{\mathrm{value}} \ . \tag{7}$$

Since the values of p_{value} may get very close to one, it is necessary to handle potential numerical problems and express p_{value} in terms of incomplete gamma distribution.

3.2 Information criteria

A different approach to comparing the sub-model quality is based on statistical information criteria. The criteria we have selected are sorted from the least stringent to the most one.

Wilks Information Criterion
Ralston [2] according to Wilks [8] recommends to search for a sub-model with minimal error s_{e}^2. Corresponding logarithmic form, which will enable us to compare the criterion with the following ones, can be defined as:

$$WIC = n \ln s_{\mathrm{e}}^2 \ . \tag{8}$$

It is obvious that k, the number of explanatory variables included in sub-model, is already indirectly penalizing the information quality in this basic criterion .

Akaike Information Criterion

Furthermore, an additional penalty for adding explanatory variables is included in the Akaike criterion which measure of the relative goodness of the sub-model [3] may be denoted as:

$$AIC = 2k + WIC \ . \tag{9}$$

Bayesian Information Criterion

Under the assumption of $n \geq 8$ the Bayesian criterion [1] generates stronger penalty for extra explanatory variables, thus preventing over-fitting even more. Following the previous terminology, the criterion may be defined as:

$$BIC = k \ln n + WIC \tag{10}$$

As opposed to the logarithm of p_{value}, the final criterion $CRIT$ to be minimized will be directly equal to the value of respective information criterion.

4 Data description

For image textural features, energies of 2D discrete wavelet transform were taken [10]. Decomposition using the Type 3 Daubechies wavelet at 8 levels was computed by Matlab function `wavedec2`. Energy is the mean square of wavelet coefficients for a given level and direction.

The basic sequence of features, $x_1, x_2, \ldots, x_{24}$, may be regarded as a set of $H_1, V_1, D_1, \ldots, H_8, V_8, D_8$ where H_j, V_j, D_j are wavelet decomposition energies at j-th level in horizontal, vertical and diagonal directions. The vector y represents decimal logarithm of crack growth rate $y = \log_{10} v$.

To minimize potential numerical errors when working with the data, input data standardization was implemented as follows:

$$x_k = \frac{h_k - \mathrm{E}\, h}{\sqrt{\mathrm{D}\, h}}, \tag{11}$$

using $\mathrm{E}\, h$ and $\mathrm{D}\, h$ as mean value and dispersion of the explanatory data.

Last, but not least – apart from the significance of the data we can make use also of physical distribution of the data in given data set. Due to the fact that data are representing fatigue crack growth rate of three different materials the data set is divided into three separate groups. This will be especially useful when dealing with the cross-validation.

5 Selection heuristic

Searching for the best available sub-model is a binary optimization task that can be defined as minimization of the objective function f: $\mathbf{D} \to \mathbb{R}$ where

$$\mathbf{D} = \{\boldsymbol{x} \in \{0,1\}^m \mid \mathbf{0} \leq \boldsymbol{x} \leq \mathbf{1}\} \tag{12}$$

is binary domain. Here, the binary vector $\boldsymbol{x}$ is directly representing utilization of the extended feature set, i.e. its components that are equal to "one" are included in the corresponding sub-model. Therefore $\mathbf{0}$ means the constant model and $\mathbf{1}$ the full model. Furthermore, 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

$$\mathbf{G} = \{\boldsymbol{x} \in \mathbf{D} \mid \mathrm{f}(\boldsymbol{x}) \leq f^*\} \tag{13}$$

where

$$f^* \geq \min\{\, \mathrm{f}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbf{D} \,\}\ . \tag{14}$$

For that purpose, we may utilize some of the well-known heuristic algorithms. We have chosen physically motivated Fast Simulated Annealing (FSA) [5] with reputable efficiency in the case of integer optimization tasks. FSA performs mutation on the ring neighbourhood

$$\mathrm{N}(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbf{D} \mid \|\boldsymbol{y} - \boldsymbol{x}\|_1 = 1\}\ . \tag{15}$$

Beginning with $k = 0$, $T_k > 0$ and initial solution vector generated by uniform distribution $\boldsymbol{x_0} \sim \mathrm{U}(\mathbf{D})$ we perform FSA mutation as uniformly generated random binary vector $\boldsymbol{y}_k \sim \mathrm{U}(\mathrm{N}(\boldsymbol{x}_k))$. Using $\eta_k \sim \mathrm{U}([-1, +1])$ we set

$$\boldsymbol{x}_{k+1} = \begin{cases} \boldsymbol{y}_k & \mathrm{f}(\boldsymbol{y}_k) < \mathrm{f}(\boldsymbol{x}_k) + T_k \tan(\frac{\pi\eta}{2}) \\ \boldsymbol{x}_k & \mathrm{f}(\boldsymbol{y}_k) \geq \mathrm{f}(\boldsymbol{x}_k) + T_k \tan(\frac{\pi\eta}{2}) \end{cases} \tag{16}$$

until a solution from the goal set is found or the pre-defined number of objective function evaluations is exhausted. The cooling strategy is represented by non-increasing sequence of positive temperatures T_k.

We were slightly inspired the by increased efficiency of hybrid heuristics in the case of combination of differential evolution and steepest descent [6] and since the previously defined set of optimization problems has many local minima, we have enhanced the FSA algorithm by a hybrid part - steepest descent, which may increase the probability of reaching the global optimum.

In our approach to hybrid heuristic optimization, instead of $\mathrm{f}(\boldsymbol{x})$ optimization we were optimizing $\mathrm{g}(\boldsymbol{x}) = \mathrm{f}(\boldsymbol{h})$ where $\boldsymbol{x} = \boldsymbol{x}_0$, $\boldsymbol{h} = \boldsymbol{x}_H$ are the first and last members of any series $\{\boldsymbol{x}_k\}_{k=0}^{H}$ satisfying $\boldsymbol{x}_j \in \mathrm{N}(\boldsymbol{x}_{j-1}), \mathrm{f}(\boldsymbol{x}_j) < \mathrm{f}(\boldsymbol{x}_{j-1})$ for $j = 1, \ldots, H$. I.e. $\boldsymbol{h}$ is the best solution, in terms of steepest descent heuristic. Before any problem solution vector is evaluated, its nearest local neighbourhood is iteratively searched for a better solution, until no further advance in terms of objective function value can be made (or until a pre-defined maximum number of local evaluations is exceeded).

This way we were able to set a higher temperature T_0 and to use more benevolent cooling strategy. In other words, the algorithm was able to prevent getting stuck in a local minimum and still not loose the ability to fine-tune a given solution. Thus the FSA performance, on this specific task, was improved.

6 Cross-validation

As aforementioned, we have the data divided into three groups according to the material being analysed. This allows us to perform a rather strong cross-validation to assess how the results of a specific criterion will generalize to an independent data set.

We will perform the optimization on two out of three groups (training group) and validate the analysis on the remaining third group (verification group). To improve overall consistency, multiple rounds of cross-validation will performed using different permutations of the data sets and the verification results will be averaged over the rounds.

As the goodness of fit measure we propose to use R as the correlation coefficient between the original data and the data proposed by respective sub-model. However, when optimizing, we will be still using the original objective function based on the minimization of $CRIT$ value.

7 Experimental results

Analysed data consisted of $n = 162$ observations and a total of 120 features in the expanded feature set. That means the standard, linear, features and four non-linear

transformations, as stated in (3).

To be able to compare results gained with the hybrid heuristic, mentioned in the previous section, against a stepwise approach, we have implemented a simple stepwise approach. Starting from the constant model the algorithm was trying to improve the sub-model quality by adding or removing one feature at a time until no further improvement was possible. Despite having multiple methods, stepwise approach was always outperformed by hybrid heuristic in terms of quality of the best found sub-model ($CRIT$) [9]. Furthermore, the traditional greedy stepwise approach generates sub-models with less explanatory variables, k_{opt}, because of the stop condition that is effective too soon [9].

Aggregated results may be found in Tab. 1. In here, column R represents correlation coefficient between the original data and the data proposed by respective sub-model. Also, basic performance measures, such as mean number of evaluations (MNE), standard deviation of the number of evaluations (SNE) and reliability (REL - number of runs during which the algorithm found a solution from the goal set before exceeding 1 500 000 evaluations, compared to the total number of runs), of the implemented heuristic that led to the stated results may be found in Tab. 2.

Table 1: Optimal sub-model quality and features using hybrid heuristic

				Level								**Direction**			**Term**				
Method	*CRIT*	*R*	k_{opt}	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**H**	**V**	**D**	f_u	$f^{1/2}$	f^2	f^{-1}	$\log_{10} f_u$
R^2 test	-106.43	0.9850	23	0	2	1	0	4	8	4	4	8	5	10	4	4	7	4	4
Wald test	-93.26	0.9765	11	0	2	1	0	2	2	3	1	3	6	2	3	0	3	3	2
WIC	-865.20	0.9971	89	14	15	12	9	7	12	14	6	28	29	32	16	16	18	19	20
AIC	-679.29	0.9908	41	3	5	4	0	13	9	1	6	14	8	19	9	7	7	11	7
BIC	-585.11	0.9771	12	0	2	1	1	2	2	3	1	3	6	3	0	2	6	2	2

Table 2: Hybrid heuristic performance measures

Method	*MNE*	*SNE*	*REL*
R^2 test	717 056.38	120 582.41	0.81
Wald test	385 850.75	49 034.88	0.42
WIC	1 254 669.71	99 941.24	0.77
AIC	638 683.60	113 380.11	0.56
BIC	596 040.33	102 196.46	0.70

As far as the cross-validation is concerned, the full data set was divided into three groups of data, each having 59 (I. group), 53 (II. group) and 50 (III. group) observations. For each permutation of training and verification groups the hybrid heuristic did optimize the sub-model to make the model fit the training data as well as possible according to respective method. Same settings and conditions were used as in the case of full data set without cross-validation. Detailed results are organized in Tab. 3. The most important results are in the column of correlation coefficient R_{verify} which measures the quality of fit on the verification data set.

These results are aggregated using mean of respective methods and furthermore expanded by comparing the data composed from distinct verification data sets to the original one in Tab. 4. Also, the outcomes of composed verification data are depicted in Fig. 1.

Table 3: Cross-validation detailed results

Method	**Training & verification grp.**	$CRIT$	R_{train}	R_{verify}	k_{opt}
R^2 test	I.+II. & III.	-77.33	0.9880	0.8857	17
R^2 test	I.+III. & II.	-72.26	0.9868	0.9356	18
R^2 test	II.+III. & I.	-71.20	0.9832	0.6472	7
Wald test	I.+II. & III.	-67.73	0.9849	0.8788	13
Wald test	I.+III. & II.	-63.00	0.9829	0.9193	13
Wald test	II.+III. & I.	-67.46	0.9804	0.5729	4
WIC	I.+II. & III.	-706.90	0.9991	0.2966	67
WIC	I.+III. & II.	-680.31	0.9992	0.6961	73
WIC	II.+III. & I.	-746.96	0.9997	0.2073	78
AIC	I.+II. & III.	-557.82	0.9989	-0.2424	66
AIC	I.+III. & II.	-437.49	0.9873	0.9386	19
AIC	II.+III. & I.	-472.67	0.9917	-0.1066	23
BIC	I.+II. & III.	-420.31	0.9869	0.8633	16
BIC	I.+III. & II.	-387.14	0.9829	0.9193	13
BIC	II.+III. & I.	-431.81	0.9804	0.5729	4

Table 4: Cross-validation summary

Method	**Mean** R	**Composed** R
R^2 test	0.8228	0.7745
Wald test	0.7903	0.7375
WIC	0.4000	0.0773
AIC	0.1965	-0.0297
BIC	0.7852	0.7336

8 Conclusion

The benefits of heuristic approach to sub-model testing in fractographics described above are considerable. Nearly an unlimited set of explanatory variables may be offered without any respect to the original number of observations in a given case. Very good models were obtained also in previously unsolvable cases with a very small number of observations.

Of course, the final result is mostly dependent on the sub-model selection approach. As it is apparent from the results of cross-validation and also based on our experience we are recommending BIC, Wald test and potentially also McFadden R-square test and WIC. Nevertheless, there are significant differences between these four and more specifically we are suggesting:

- BIC as an universal criterion,
- Wald test as a well balanced criterion, similar to the BIC, but only as far as linear regression models are concerned,

- McFadden R-square test as a legitimate criterion with respect to the variance analysis approach,
- WIC as a criterion that leads to considerable adherence to the data, however, as opposed to aforementioned criteria, lacks the ability to generalize.

References

[1] G. Schwarz. *Ann. Statist.*. 5, 461, (1978).

[2] A. Ralston, P. Rabinowitz. *A First Course in Numerical Analysis*. Courier Dover Publications, (2001).

[3] H. Akaike. *A new look at the statistical model identification*. IEEE Trans. Automat. Contr., AC-19:716–23, (1974).

[4] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, (2002).

[5] V. Kvasnička, J. Pospíchal, P. Tiňo. *Evolutionary Algorithms* (in Slovak). STU Bratislava, (2000).

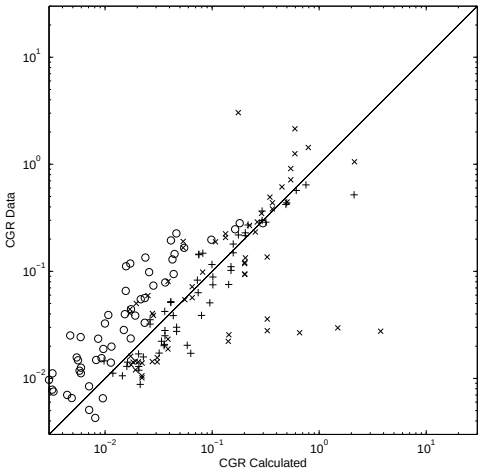
[6] J. Tvrdík, I. Křivý. *Hybrid Adaptive Differential Evolution in Partitional Clustering*. Proceedings of Mendel 2011 Conference, Brno University of Technology, Brno, (2011), pp. 1–8.

[7] J. Anděl. *Mathematical Statistics* (in Czech). SNTL/Alfa, Praha, (1978).

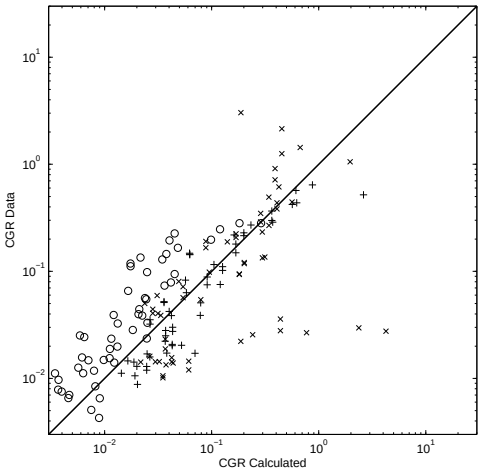
[8] S. S. Wilks. *Mathematical Statistics*, rev. ed.. John Wiley and & Sons, Inc., New York, (1962).

[9] M. Mojzeš, J. Kukal, H. Lauschmann. *Sub-model Testing in Fractographic Analysis*. Proceedings of Mendel 2012 Conference, Brno University of Technology, Brno, 2012, pp. 350–355.

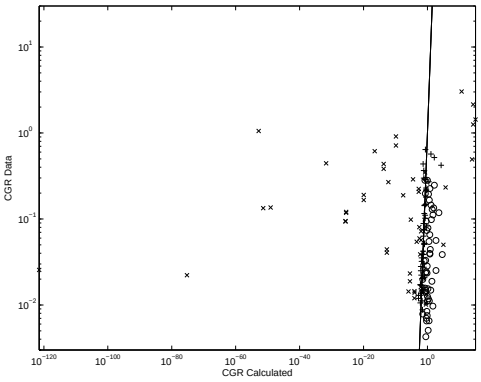
[10] H. Lauschmann, N. Goldsmith. *Textural Fractography of Fatigue Fractures*. Fatigue Crack Growth: Mechanics, Behavior and Prediction. Alphonse F. Lignelli, ed., Nova Science Publishers, Inc., (2009), pp. 125-166.



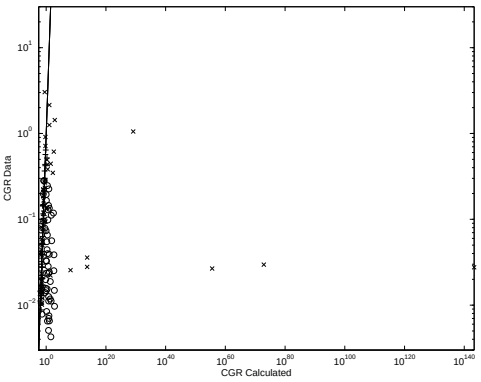
R^2 test



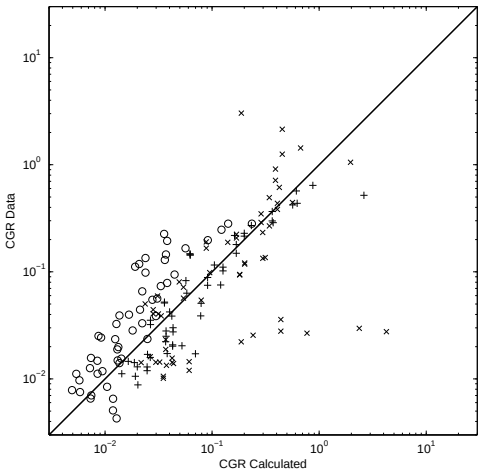
Wald test



WIC



AIC



BIC

Figure 1: Composed verification data (markers used: cross for I. group, plus sign for II. group, circle for III. group)

Rima Glottidis Segmentation by Thresholding Using Graph Cuts

Adam Novozámský

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

advisor: Stanislav Saic, Institute of Information Theory and Automation, AS CR

Abstract. In 1996 was invited videokymography as high-speed medical imaging method to visualize the human vocal cords vibrations in voice disorders. This method provides a good visualization of vocal fold vibration, frequency and amplitude of oscillation, the duration of each phase of the cycle-opening and closing of the glottis, or propagation of mucosal waves. Manual data extraction is time-consuming and depends on the correct identification of the features of physician. We proposed a new segmentation method base on thresholding for detection rima glottidis. A proper search for glottis is very important for further analysis of the features.

Keywords: medical imaging, vocal chords, videokymography, segmentation, thresholding

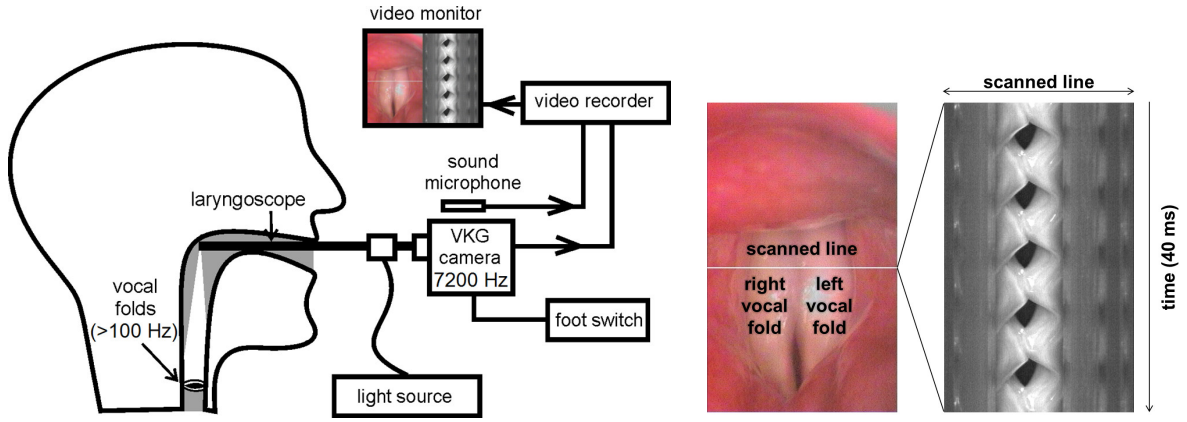
Abstrakt. V roce 1996 byla navržena videokymografie jako vysokorychlostní lékarská zobrazovací technika k vizualizaci poruch vibrací lidských hlasivek. Tato metoda dobre zobrazuje kmitání hlasivek, frekvenci a amplitudu kmitu, trvání jednotlivých fází cyklu-otevírání a zavírání glotis, nebo šírení sliznicních vln. Rucní extrakce dat je casove nárocná a závisí na správné identifikaci príznaku doktorem. My zde predstavujeme novou segmentacní metodu k detekci hlasivkové šterbiny založenou na prahování. Správné nalezení této šterbiny je velmi duležité pro další analýzu príznaku.

Klíčová slova: medicínské zobrazování, hlasivky, Videokymografie, segmentace, prahování

1 Introduction

The quality of voice is critically determined by vibration of the vocal folds. Revealing small changes in the vibration can help early detection of various diseases, including cancer of the larynx. Therefore the objective evaluation and quantification is an important issue.
In general, the frequency of vocal folds vibrations varies within the range of 100 to 500 Hz in males and 130 to 1,000 Hz in females (extreme position is reached only when singing). This speed can not be captured with cameras used current television standard, which rate is 25-60 frames per second (fps). To capture so fast phenomena are primarily used two techniques:

- *The Stroboscopy*: Vibrating vocal cords are illuminated by a flashing light source. It looks like they were motionless when synchronizing light flashes with the vibrations of vocal cords. If we illuminate the vocal chords in a different phase of the vibration we



(a) Videokymographic examination (b) Vocal folds & Videokymogram

Figure 1: The videokymographic examination

reach apparent slowdown oscillations. This allows us to observe the oscillations of the vocal cords also with the help of slow cameras working with the television standard. Although the stroboscopy is good for recording fast processes, it has a serious limitation. It works only with periodic vibration, therefore every frequency disturbance of the vibration also disturbs the resulting stroboscopic output and irregular vibrations of the vocal folds cannot be studied at all.

- *The Ultra High-Speed Photography*: This method is time-consuming and these devices is very expensive. The rate is from 1,000 to 100,000 fps. It means, if we have this camera with 50,000 fps and the examination will takas 5 second, we have 250,000 frames. Normal video playback speed is 32 frames per second, this means that the five-second examination of the patient produces approximately 2 hour video. It is not feasible for most laboratories.

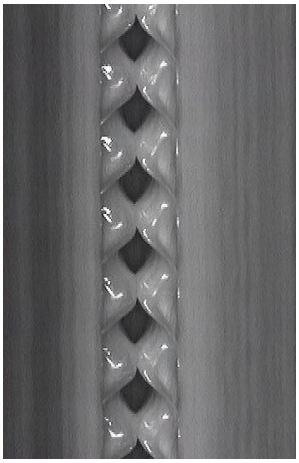
In 1996 Dr. Švec et al. [5] suggested a new method of recording the vocal cords, which he called videokymography (**VKG**). Here is used standard camera operates in two modes. In standard mode, the camera works as well as standard TV cameras, recording 50 fps with a resolution of 768x576 pixels. In the second videokymographic mode, the camera captures only one line (top) and the frequency reaches 7812.5 lines per second with a resolution of 768 pixels.

2 Dataset

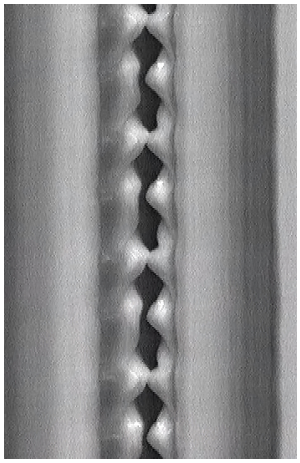
Dr. Švec gave fifty pictures taken from different patients. This data set consists mainly of records with some vocal defect. Thanks this was achieved the great variability of records and were covered major damage to the vocal cords.

3 Videokymogram and its characteristics

The scheme examination with videokymographic camera is shown in Figure 1a. The thin white horizontal line in Figure 1b(left) indicates the position of the recording line. The



(a) Vocal fold vibration recorded by the videokymographic mode of the system. We can detect the opening and closing movement of the glottis, the frequency and other features



(b) Irregular vibration of vocal folds. We recognize the left-right asymmetry in he vibration as well as the uncomplete closing

Figure 2: Examples of Videokymography Examination

Videokymographic image (Videokymogram) in Figure 1b(right) is two-dimensional image composed of this line captured in time sequence.

The Figure 2 shows regular and irregular vocal fold vibrations of another patients scanned in the middle of the glottis. After analyzing VKG images, Dr. Švec[6] created a collection of features for characteristic of patients vocal folds. Here is their list with a brief description, taken from [6]:

Absence of Vibration of Vocal Fold: We distinguish completely absent vibration of the vocal fold or only partly, shown in Figure 3a.

Interference of Surroundings With Vocal Folds: We can divided this into two category: 1) co-vibrations of the ventricular folds or other laryngeal tissues with the vocal folds and 2) co-vibration of fluids with the vocal folds.

Cycle-to-Cycle Variability: This characteristic refers to dissimilarity of consecutive vibration cycles in duration, amplitude, and overall shape, shown in Figure 3b.

Duration of Glottal Closure: The duration of closure divided by the duration of the glottal cycle.

Left-Right Asymmetry: It is caused by any difference in the oscillation, but the most serious behavioral expression of asymmetry are frequency differences, in which the left and right vocal folds vibrate with different, shown in Figure 3c. frequencies,

Shape of Lateral Peaks: Sharpness of the lateral peaks is a sign of vertical phase differences, ie. a delayed movement of the upper margin behind the lower margin of the vocal fold, shown in Figure 3d.

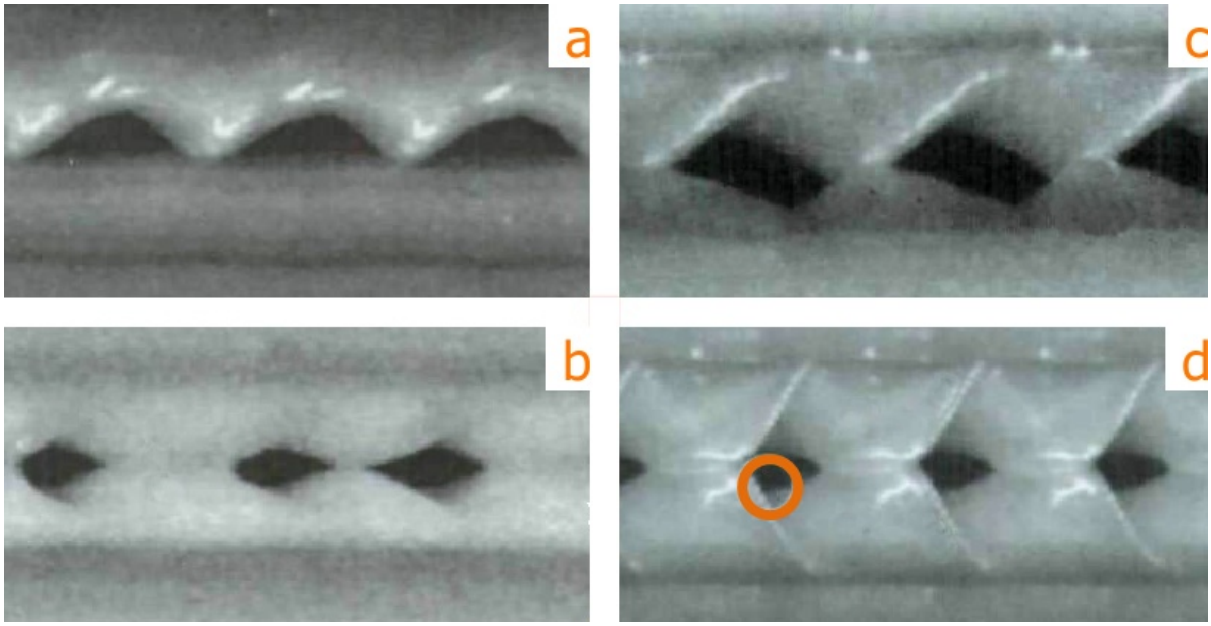


Figure 3: (**a**)Absence of vocal fold vibration. (**b**) Large left-right synchronous cycle-to-cycle variability. (**c**) Phase differences and axis shift. (**d**) Sharp lateral peak.

Laterally Traveling Mucosal Waves: These can be defined as the lateral movements on the vocal folds that occur during the medial movement of the glottal edge.

Opening Versus Closing Duration: This characteristic compares the time during which the vocal fold edge moves in the lateral direction (*opening*) to the time during which it moves in the medial direction (*closing*).

Shape of Medial Peaks: Similar to the lateral peaks, the shape of the medial peaks was found to occur in two types: *rounded* or *sharp*.

Cycle Aberrations: This is feature that disturb the simple shape of the vibratory cycle of the vocal fold while not necessarily disturbing the periodicity of the vibration.

4 Analysis of the Basic Features

To extract all properties of the vocal folds mentioned in the previous section, we first have to find correctly the rima glottidis. Thanks to various voice disorders this task very difficult, although at the first glance it seems easy. During work, we tried a number of methods that did not lead to the goal:

Classical Thresholding: The rima glottidis is on all frames very well recognized with human eye, due to its contrast to the vocal chords, which is lightened. Despite this it is impossible to find a global threshold for all tested images, which could be applied to segmentation. This is caused by different brightness and high-noise images. The Otsu's

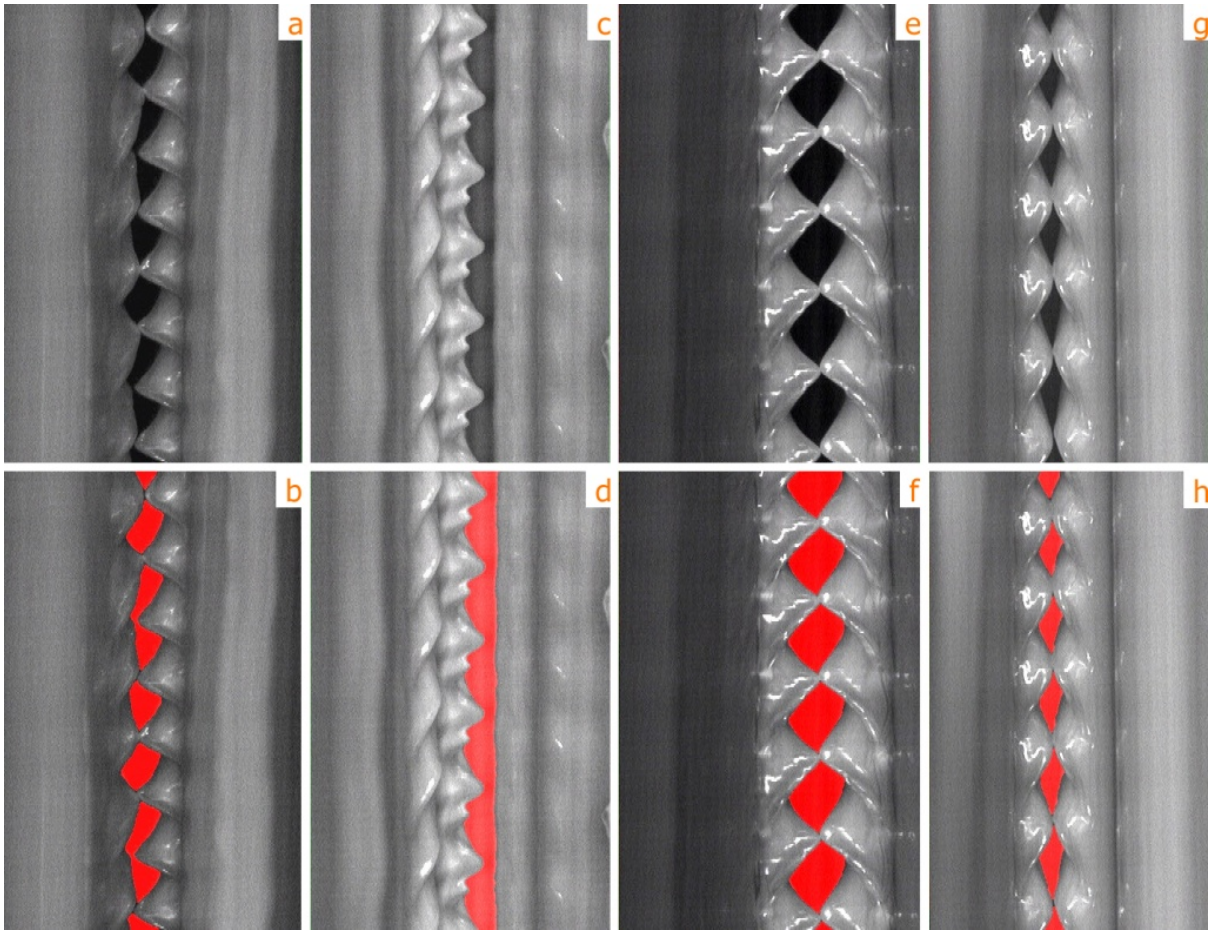


Figure 4: Rima Glottidis Segmentation via Graph Cuts. Analyzed images (**a**, **c**, **e**, **g**) and their segmentations (**b**, **d**, **f**, **h**).

method gave also poor results. This algorithm assumes that the processed image contains two classes of pixels or bi-modal histogram and this condition is not fulfilled.

Level Set: The level set method was developed in the 1980s and is widely used for segmentation. We implemented it according to this Approach [2]. Unfortunately, on our data does not work quite well .

Graph Cuts: Very good method implemented according to the approach Boykov, Kolmogorov and Zabih [1]. On our data works well, but time-consuming computation is large (minutes on a single picture). The resulting segmentation using graph cuts we can see in the Figure 4.

During testing, we found the method of adaptive threshold searching based on minimizing the graph cut. Its description is in the following section.

4.1 Thresholding Using Graph Cuts

In 2008 Wenbing Tao [4] introduced a novel thresholding algorithm. The proposed method uses a normalized graph cut measure as thresholding principle to distinguish an object

from background.
Consider a weighted undirected graph $G = (V, E)$, where V is the set of vertices, E is the set of edges. Each edge has its weight $w(u, v)$ describing the similarity between two nodes u and v. The graph cut means the division this graph into two disjoint complementary sets A and $B = V - A$. We can quantify this distribution as a total weight of the edges connecting the two parts

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v). \tag{1}$$

The goal is to find the optimal bipartitioning of a G. In 2000 Shi and Malik [3] proposed a new measure of disassociation between two sets. They named normalized cuts (Ncut)

$$Ncut(A, B) = \frac{cut(A, B)}{asso(A, V)} + \frac{cut(A, B)}{asso(B, V)} \tag{2}$$

, where $asso(A, V) = \sum_{u \in A, t \in V} w(u, t)$ is the total connection from nodes in A to all nodes in the graph.

4.1.1 Algorithm Construction [4]

• Let $V = \{(i, j) : i = 0, 1, ..., n_h - 1; j = 0, 1, ..., n_w - 1\}, L = \{0, 1, ..., 255\}$, where n_h and n_w are the height and width of the image.

$$f(x, y) \in L \quad \forall (x, y) \in \quad V_k = \{(x, y) : f(x, y) = k, (x, y) \in V\} \quad k \in L \tag{3}$$

$$\bigcup_{k=0}^{255} V_k = V \quad V_j \cap V_k = \Phi \quad k \neq j \quad j, k \in L \tag{4}$$

• Construct undirected weighted graph $G = (V, E)$, where nodes are pixels and weight is defined as follows

$$w(u, v) = \begin{cases} e^{-[\frac{\|F(u) - F(v)\|_2^2}{d_I} + \frac{\|X(u) - X(v)\|_2^2}{d_X}]}, & \text{if } \|X(u) - X(v)\|_2 < r \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where d_I and d_X are positive scaling factors defining the relationship of $w(u, v)$ to the intensity difference or spatial location two nodes, $r \in \mathbf{R}^+$ determines the number of neighboring nodes, and $\|.\|$ denotes the vector norm. These parameters are set to $d_I = 625$, $d_X = 4$, and $r = 2$.

• For all $t \in L$ we have a unique bisection $V = A, B$ of the graph $G = (V, E)$, where A and B is defined as follows

$$A = \bigcup_{k=0}^{t} V_k, \qquad B = \bigcup_{k=t+1}^{255} V_k, \qquad k \in L. \tag{6}$$

Then the graph cut by definition (1) becomes

$$cut(A, B) = \sum_{i=0}^{t} \sum_{j=t+1}^{255} cut(V_i, V_j), \tag{7}$$

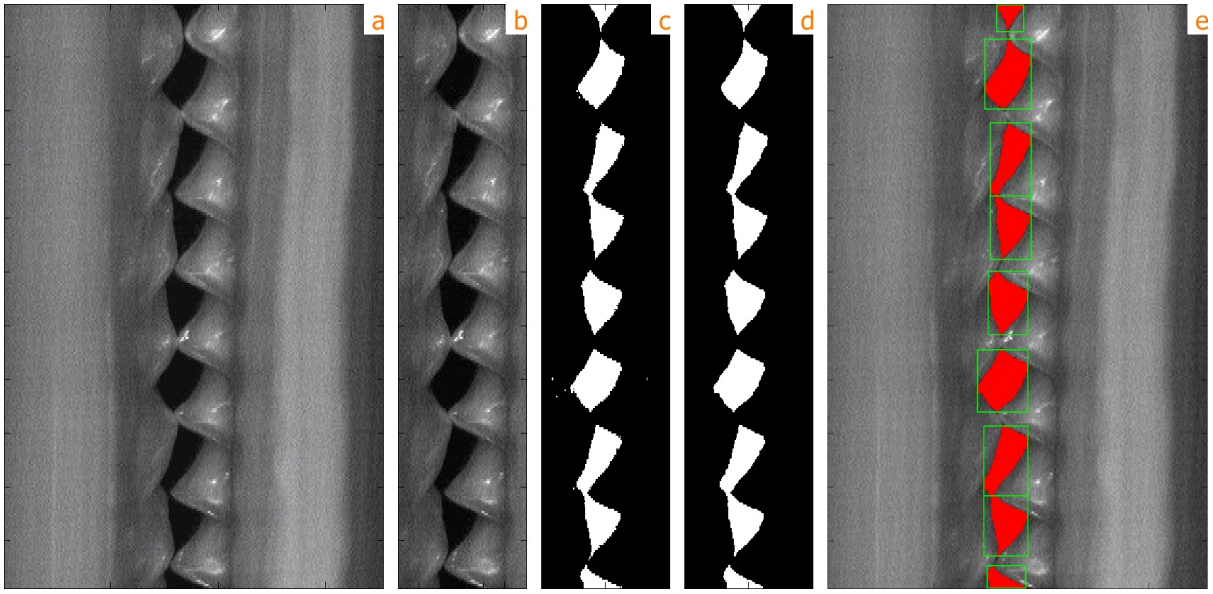


Figure 5: Rima Glottidis Segmentation (**a**) Analyzed image. (**b**) Center of the vocal chords. (**c**) Threshold found using 4.1.1. (**d**) Morphological Operations.(**d**) Complete segmentation with opening and closing of rima glottidis.

where $cut(V_i, V_j) = \sum_{u \in V_i, v \in V_j} w(u, v)$ is the sum of the weights of the total connection between all nodes with gray level i and all nodes with gray level j. Similarly, we can write the following relations

$$asso(A, A) = \sum_{i=0}^{t} \sum_{j=i}^{t} cut(V_i, V_j) \quad \text{and} \quad asso(B, B) = \sum_{i=t+1}^{255} \sum_{j=i}^{255} cut(V_i, V_j) \tag{8}$$

and also $asso(A, V) = asso(A, A) + cut(A, B)$ and $asso(B, V) = asso(B, B) + cut(A, B)$. Then finally

$$Ncut(A, B) = \frac{cut(A, B)}{asso(A, A) + cut(A, B)} + \frac{cut(A, B)}{asso(B, B) + cut(A, B)}, \tag{9}$$

which we minimize with respect to t.

For a more detailed description of the algorithm we refer to [4].

5 Rima Glottidis Segmentation

In this section we describe our algorithm to find the rima glottidis and its segmentation step by step.

a ***Analyzed Image***

b ***Find Center by Deviation*** First we have to find the approximate center of rima glottidis. This can by achieved by counting the standard deviations in column of the

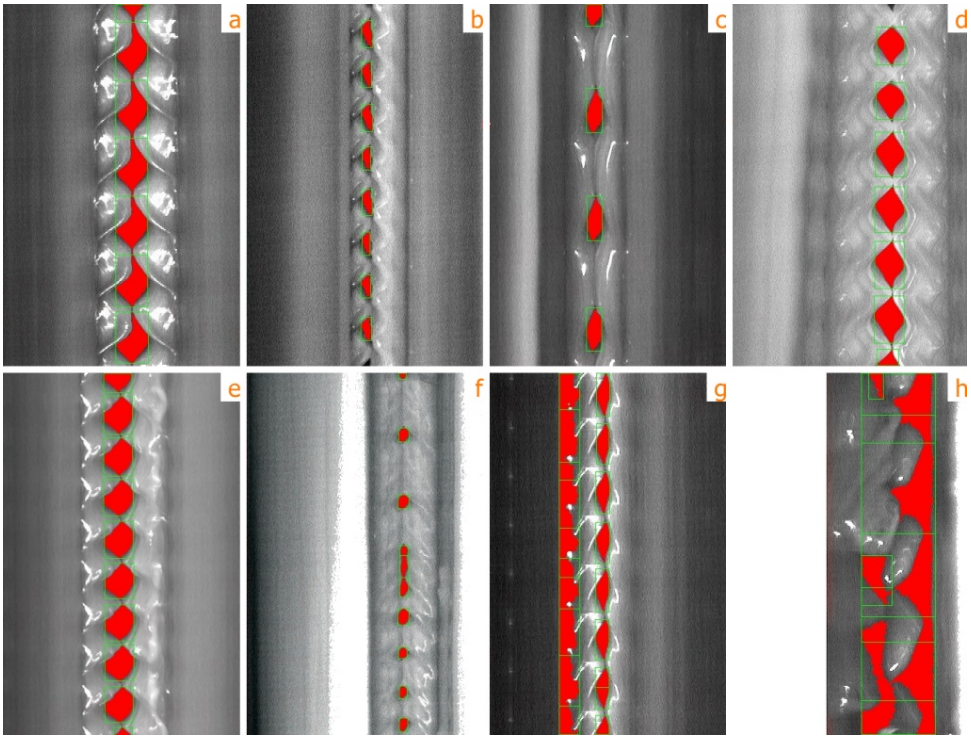


Figure 6: Rima Glottidis Segmentation with our method - well (**a**, **b**, **c**, **d**, **e**, and **f**) and poorly analyzed (**g**, **h**).

image and finding the greatest. This calculation marks the column, where it appears both two extremal values of light and dark, which is exactly our opening and closing of the vocal cords. Vocal cords is approximately in the middle of all the images and takes at most one quarter, so we crop both edges of the image.

c ***Thresholding*** Now we find the threshold by the algorithm described in 4.1.1 and we do the thresholding with this value.

d ***Morphological Operations*** These operations are a necessary step to remove unwanted artifacts (holes and false response) after thresholding.We use morphological operations namely *opening* and *closing*.

e ***Opening and Closing of the Rima Glottidis*** There are two situation that may occur. First and easier way the cycles vocal cords are separated, so there is completely closing movement of vocal cords. The second is a little bit complicated, there is no closure of the vocal cords. We solved this issue by analyzing the function of the row sum, where we are looking for local minima, which means the end of the cycle without vocal cord closure.

In the Figure 5 are shown individual steps of the above described algorithm. The whole segmentation takes average 0.1796 second with resolution of 350 x 550 pixels.

6 Conclusion

Rima Glottidis Segmentation via thresholding was studied in this paper. The experimental results in the Figure 6 show good ability to detect a variety of vocal cords. For fifty test images ill patients were correctly detected 84 percent. 8 images were falsely detected mainly due to the poor quality of the images, or abnormal vocal chords.

7 Acknowledgment

The author would like to thank Dr. Švec for providing images and helping us in medical term.

References

[1] V. Kolmogorov and R. Zabin. *What energy functions can be minimized via graph cuts?* Pattern Analysis and Machine Intelligence, IEEE Transactions on **26** (feb. 2004), 147 –159.

[2] C. Li, C. Xu, C. Gui, and M. Fox. *Level set evolution without re-initialization: a new variational formulation.* In 'Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on', volume 1, 430 – 436 vol. 1, (june 2005).

[3] J. Shi and J. Malik. *Normalized cuts and image segmentation.* In 'Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on', 731 –737, (jun 1997).

[4] W. Tao, H. Jin, Y. Zhang, L. Liu, and D. Wang. *Image thresholding using graph cuts.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on **38** (sept. 2008), 1181 –1195.

[5] J. G. Švec and H. K. Schutte. *Videokymography: High-speed line scanning of vocal fold vibration.* Journal of Voice **10** (1996), 201 – 205.

[6] J. G. Švec, F. Šram, and H. K. Schutte. *Videokymography in voice disorders: What to look for?.* Annals of Otology, Rhinology & Laryngology **116** (2007), 172 – 180.

Limiting Normal Operator*

Miroslav Pištěk

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

advisor: Jiří Outrata, Institute of Information Theory and Automation, AS CR

Abstract. A new approach to quasi-convex analysis is proposed here with a clear relation to the general variational analysis. The basic idea is to deal with the sublevel mapping instead of the epigraph of a quasi-convex function. Such approach leads to a notion of the normal operator, which, however, lacks continuity for a general quasi-convex function. Here, we resolved this problem by introducing limiting variant of the normal operator, which is outer-semicontinuous. Moreover, the basic properties of the limiting normal operator are examined together with its relation to the limiting subdifferential.

Keywords: variational analysis, quasi-convex function, limiting normal operator

Abstrakt. V této práci prezentujeme nový přístup ke kvazi-konvexní analýze, jež úzce souvisí s obecnou variační analýzou. Základní myšlenkou je vyšetřovaní tzv. sublevel zobrazení namísto zkoumání epigrafu kvazi-konvexní funkce. Tento přístup vede k pojmu normalového operátoru, který však pro obecnou kvazi-konvexní funkci není spojitý. To je vyřešeno přechodem k limitní verzi normálového operátoru, která je zvně polospojitá (outer-semicontinuous). V článku jsou představeny i další zakladní vlastnosti limitního normalového operátoru spolu s jeho vztahem k limitnímu subdiferenciálu.

Klíčová slova: variační analýza, kvazi-konvexní funkce, limitní normálový operátor

1 Introduction

This work aims at obtaining a tool of generalized differentiation adopted for the class of quasi-convex functions. Since the general notion of limiting subdifferential [1, 4] does not benefit from convexity of sublevel sets of quasi-convex functions, there is a need to find an alternative way. The sublevel approach was pioneered in [2] where the notion of normal operator and strict normal operator was introduced. The first operator is outer-semicontinuous and the other quasi-monotone, however, none of them has both these important properties together. On that account, an unifying approach of adjusted normal operator was developed in [3], which is outer-semicontinuous and quasi-monotone at the same time. Moreover, it is capable of full characterization of the optimality conditions for quasi-convex optimization. Albeit these qualities, it is hard to find proper calculus rules for adjusted normal operator. The reason is its non-local nature, which is inherited from strict normal operator used in the definition of the adjusted normal operator. Therefore,

*The presented research is a result of the joint PhD programme co-supervised by Professor Didier Aussel, Lab. PROMES, University of Perpignan, France

we decided to follow a different way here. The lack of outer-semicontinuity of the normal operator may be untangled using the far-reaching idea of Mordukhovich [4]. This way we obtain *limiting normal operator* which is outer-semicontinuous and quasi-monotone at the same time. Further, the relation of limiting normal operator and limiting subdifferential is established at the end of this article.

2 Elements of Variational Analysis

First, we introduce basic elements of modern variational analysis which we then adopt to quasi-convex setting. The full motivation of the following notions is out of scope of this article, an interested reader is referred to excellent monograph [1]. We deal with finite-dimensional case only, extensions to infinite dimensions may be developed by following [4].

A generalized differentiation is based on the notion of a cone, which may contain directional derivatives or normal vectors, for instance.

Definition 2.1 (Cone). *A set $C \subset \mathbb{R}^m$ is called a cone if $0 \in C$ and for all $\lambda \geq 0$ we have $\lambda C \in C$.*

This is the first time we used the so-called Minkowski notation for basic set operations. For a general sets $A, B \subset \mathbb{R}^m$ we denote

$$A + B \equiv \{a + b : a \in A, b \in B\} \tag{1}$$

and considering any $\lambda \in \mathbb{R}$ also

$$\lambda A \equiv \{\lambda a : a \in A\}. \tag{2}$$

The smallest cone containing set C is its positive hull $\mathrm{pos}\{C\}$.

Definition 2.2 (Positive hull). *For a set $C \subset \mathbb{R}^m$, a positive hull $pos\{C\}$ is defined as*

$$pos\{C\} \equiv \{0\} \cup \bigcup_{\lambda > 0} \lambda C. \tag{3}$$

For a convex hull of set $A \subset \mathbb{R}^m$, $\mathrm{conv}\{A\}$ is used. Next, there exists an important dual representation of closed convex cones, which is based on the notion of a polar cone.

Definition 2.3 (Polar cone). *For $C \subset \mathbb{R}^m$ we define a (negative) polar cone C^o as*

$$C^o \equiv \left\{ y \in \mathbb{R}^m : \underset{x \in C}{\forall} \langle y, x \rangle \leq 0 \right\}. \tag{4}$$

For any set $C \subset \mathbb{R}^m$, its polar set C^o is a convex closed cone. Especially, for a closed convex cone $K \subset \mathbb{R}^m$ we have $K^{oo} = (K^o)^o = K$.

For the subject of variational analysis, the cone-valued mappings are fundamental. Thus, we have to develop several notions of set-valued analysis. First to say, we denote $M[\mathbb{R}^m \rightrightarrows \mathbb{R}^n]$ a multivalued mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$, i.e. $M(x) \subset \mathbb{R}^n$ for $x \in R^m$. Then, the following concept of outer-semicontinuity of such mappings is of high importance.

Definition 2.4 (Outer limit of multivalued mapping). *For a multivalued mapping $M[\mathbb{R}^m \rightrightarrows \mathbb{R}^n]$ we define outer limit as*

$$\operatorname*{limsup}_{x\to\bar{x}} M(x) = \left\{ y \in \mathbb{R}^n : \exists_{x_m\to x} \exists_{y_m\in M(x_m)} y_m \to y \right\}. \tag{5}$$

Definition 2.5 (Outer-semicontinuous multivalued mapping). *We say that a multivalued mapping $M[\mathbb{R}^m \rightrightarrows \mathbb{R}^n]$ is outer-semicontinuous at $\bar{x} \in \mathbb{R}^m$ if*

$$\operatorname*{limsup}_{x\to\bar{x}} M(x) \subset M(\bar{x}), \tag{6}$$

or, equivalently, $\operatorname{limsup}_{x\to\bar{x}} M(x) = M(\bar{x})$.

The following lemma will be helpful in the next section.

Lemma 2.6 (Outer limit of linear images). *For a linear mapping $L[\mathbb{R}^m \to \mathbb{R}^n]$ and cone-valued outer-semicontinuous multi-mapping $M[\mathbb{R}^k \rightrightarrows \mathbb{R}^m]$ it holds*

$$L(M(\bar{x})) \subset \operatorname*{limsup}_{x\to\bar{x}} L(M(x)). \tag{7}$$

This inclusion is an equality if $L^{-1}(0) \cap M(\bar{x}) = \{0\}$.

Proof. This statement follows directly from [1, Theorem 4.26] applied to arbitrary sequence $x_n \to \bar{x}$ with condition of equality adopted to the case of linear mapping L and cone-valued outer-semicontinuous mapping M. □

Now, we may continue with local analysis of sets. For a closed set $K \subset \mathbb{R}^m$ and $x \in K$ we define tangent cone $T_K(x)$ as follows

Definition 2.7 (Tangent cone). *For a closed set $K \subset \mathbb{R}^m$ and $x \in K$ we define tangent cone $T_K(x)$ at point x as*

$$T_K(x) \equiv \operatorname*{limsup}_{\lambda\searrow 0} \frac{K - x}{\lambda}. \tag{8}$$

Tangent cone $T_K(x)$ contains such directional vectors $v \in T_K(x)$ that a point x remains within the set K when moving in the direction of v, at least in the sense of outer limit. A dual concept to tangent cone is regular normal cone.

Definition 2.8 (Regular normal cone). *For a closed set $K \subset \mathbb{R}^m$ and $x \in K$ we define regular normal cone $\widehat{N}_K(x)$ at point x as*

$$\widehat{N}_K(x) \equiv T_K(x)^o. \tag{9}$$

We see that for a general set K the regular normal cone $\widehat{N}_K(x)$ is convex by definition. However, $\widehat{N}_K(x)$ is not outer-semicontinuous, which complicates its calculation in applications. On that account, the limiting normal cone $N_K(x)$ was introduced, see [4]

Definition 2.9 (Limiting normal cone). *For a closed set $K \subset \mathbb{R}^m$ and $\bar{x} \in K$ we define limiting normal cone $N_K(\bar{x})$ at point $\bar{x}$ as*

$$N_K(\bar{x}) \equiv \underset{\substack{x \to \bar{x} \\ K}}{limsup}\, \widehat{N}_K(x). \tag{10}$$

The limiting normal cone is outer-semicontinuous by definition, nonetheless, on general it is no more convex.

Further, various notions of subdifferentials follows inheriting properties of the closely related normal cones.

Definition 2.10 (Regular subdifferential). *For any lower-semicontinuous function $f[\mathbb{R}^m \to \mathbb{R}]$ we may define regular subdifferential $\hat{\partial} f(x)$ using regular normal cone to the epigraph of f at the point in question*

$$\hat{\partial} f(x) \equiv (\mathbb{R}^m \times \{-1\}) \bigcap \widehat{N}_{epif}(x, f(x)). \tag{11}$$

Therefore, regular subdifferential $\hat{\partial} f$ is convex-valued owning to convexity of $\widehat{N}_f$.

Definition 2.11 (Limiting subdifferential). *For any lower-semicontinuous function $f[\mathbb{R}^m \to \mathbb{R}]$ we may define limiting subdifferential $\partial f(x)$ via limiting normal cone to the epigraph of f at the point in question*

$$\partial f(x) \equiv (\mathbb{R}^m \times \{-1\}) \bigcap N_{epif}(x, f(x)). \tag{12}$$

Similarly to normal cones, limiting subdifferential $\partial f(x)$ allows more practicable calculus rules at the price of not being convex-valued in opposite to $\hat{\partial} f(x)$. For analysis of non-Lipschitz function, another notion of subdifferential is necessary.

Definition 2.12 (Singular subdifferential). *For any lower-semicontinuous function $f[\mathbb{R}^m \to \mathbb{R}]$ we may define singular subdifferential $\partial^\infty f(x)$ using limiting normal cone to the epigraph of f at the point in question*

$$\partial^\infty f(x) \equiv (\mathbb{R}^m \times \{0\}) \bigcap N_{epif}(x, f(x)). \tag{13}$$

Even though subdifferentials play a primary role in applications, here we preferably deal with normal cones since they may be more easily applied to sublevel sets of quasiconvex functions in the next section. Therefore, the following lemma is useful.

Lemma 2.13 (Subdifferentials as projection of normal cone). *For any $x \in dom(f)$ for f lower-semicontinuous we have*

$$pos\{\partial f(x)\} \cup \partial^\infty f(x) = Proj_{domf}[N_{epif}(x, f(x))], \tag{14}$$

where $Proj_{domf}$ is a canonical projection on domain of function f.

Proof. This follows directly from (12) and (13). □

3 Limiting Normal Operator

In this section we adapt the general notions of variational analysis to the class of quasi-convex functions. We decided to borrow the notation from [1] to stress the newly established relation of quasi-convex analysis and modern variational analysis. For some terms, it was unavoidable to change the notation usual in quasi-convex analysis, we will comment on such cases. We analyse a quasi-convex function in terms of its sublevel set.

Definition 3.1 (Sublevel set)**.** *For function $f(x)$ we define sublevel set $S_f(x)$ for any $x \in dom(f)$ as*

$$S_f(x) \equiv \{y \in dom(f) : f(y) \leq f(x)\}. \tag{15}$$

We note that function $f(x)$ is quasi-convex if and only if $S_f(x)$ is convex for all $x \in \mathrm{dom}(f)$. Moreover, since we are interested in lower-semicontinuous functions, we see that sublevel set $S_f(x)$ is closed for all $x \in \mathrm{dom}(f)$. On that account, we define even the strict sublevel set as a closed set to ease further notation.

Definition 3.2 (Strict sublevel set)**.** *The (closed) strict sublevel set $\bar{S}_f^<(x)$ is defined as follows*

$$\bar{S}_f^<(x) \equiv \overline{\{y \in dom(f); f(y) < f(x)\}}. \tag{16}$$

Whatever sublevel set we use, we may define tangent operator for a quasi-convex function f at point x as follows.

Definition 3.3 (Tangent operators)**.** *Tangent operator $T_f[X \rightrightarrows X]$ and strict tangent operator $T_f^<[X \rightrightarrows X]$ to a quasi-convex lower-semicontinuous function f at point $x \in dom(f)$ are defined as*

$$\begin{aligned} T_f(x) &\equiv pos\{S_f(x) - x\}, \\ T_f^<(x) &\equiv pos\big\{S_f^<(x) - x\big\}. \end{aligned} \tag{17}$$

In this article, tangent operator T_f substitute tangent cone to epigraph T_{epif} used for analysing general lower-semicontinuous functions in variational analysis. Following this analogy, we define regular normal operator and strict normal operator.

Definition 3.4 (Normal operators)**.** *For a quasi-convex lower-semicontinuous function f we define regular normal operator $\widehat{N}_f[\mathbb{R}^m \rightrightarrows \mathbb{R}^m]$ and strict normal operator $N_f^<[\mathbb{R}^m \rightrightarrows \mathbb{R}^m]$ at point $x \in dom(f)$ as*

$$\begin{aligned} \widehat{N}_f(x) &\equiv T_f(x)^o, \\ N_f^<(x) &\equiv T_f^<(x)^o. \end{aligned} \tag{18}$$

We note that regular normal operator $\widehat{N}_f$ was originally called 'normal operator' and denoted N_f, see [2]. We decided to reserve this name and notation for a normal operator introduced further to establish and emphasize relation to modern variational analysis [1]. Next, we show basic properties of $\widehat{N}_f$ and $N_f^<$.

Definition 3.5 (Quasi-monotone operator). *We say that a set-valued operator $N[\mathbb{R}^m \rightrightarrows \mathbb{R}^m]$ is quasi-monotone if implication*

$$\langle x^\star, y - x\rangle > 0 \Rightarrow \langle y^\star, y - x\rangle \geq 0 \tag{19}$$

holds for all $x, y \in X$, $x^\star \in N(x)$, $y^\star \in N(y)$.

Lemma 3.6 (Quasi-monotonicity of $\widehat{N}$). *Regular normal operator $\widehat{N}_f$ is quasi-monotone for all quasi-convex lower-semicontinuous functions f.*

Proof. See [3]. □

Lemma 3.7 (Outer-semicontinuity of $N^<$). *Strict normal operator $N_f^<$ is outer-semicontinuous for all quasi-convex lower-semicontinuous functions f.*

Proof. According to [2, Proposition 2.1], $\operatorname{gph} N_f^<$ is closed, which is equivalent to outer-semicontinuity, see [1, Theorem 5.7]. □

There are, however, well-known examples of lower-semicontinuous quasi-convex functions where $\widehat{N}$ is not outer-semicontinuous and $N^<$ is not quasi-monotone, see [2, Example 2.2] and [3, Example 2.1], respectively. The first normal operator satisfying both these properties is adjusted normal operator N^a defined in [3]. However, it lacks calculus rules because of its non-local nature. This was the ultimate motivation for introducing the new notion of limiting normal operator in a way similar to the limiting normal cone, see Definition 2.9.

Definition 3.8 (Limiting normal operator). *For a quasi-convex function $f(x)$ we define the limiting normal operator $N_f[\mathbb{R}^m \rightrightarrows \mathbb{R}^m]$ at point $\bar{x} \in dom(f)$ as*

$$N_f(\bar{x}) \equiv \operatorname*{limsup}_{x \to \bar{x}} \widehat{N}_f(x). \tag{20}$$

This variant of normal operator possesses both important properties of quasi-monotonicity and outer-semicontinuity. Indeed, N_f is outer-semicontinuous by definition, and at the same time it attains quasi-monotonicity of $\widehat{N}_f$.

Theorem 3.9 (Quasi-monotonicity of N_f). *Limiting normal operator N_f is quasi-monotone for any quasi-convex lower-semicontinuous function f.*

Proof. Take any $x, y \in X$ and $x^\star \in N_f(x)$, $y^\star \in N_f(y)$. There exist sequences $x_m \to x$, respective $y_n \to y$, such that $x^\star \in \widehat{N}_f(x_m)$ for all m, respective $y^\star \in \widehat{N}_f(y_n)$ for all n. Next, we assume $\langle x^\star, y - x\rangle > 0$ and we need to show that $\langle y^\star, y - x\rangle \geq 0$. For m and n large enough, we have $\langle x^\star, y_n - x_m\rangle > 0$ with $x^\star \in \widehat{N}_f(x_m)$ and $y^\star \in \widehat{N}_f(y_n)$, and thus we may apply quasi-monotonicity of $\widehat{N}_f$. This way we obtain $\langle y^\star, y_n - x_m\rangle \geq 0$ and so the proof is finished if we consider limit $n, m \to \infty$. □

Next, we establish relation of the newly introduced $N_f(x)$ to limiting and singular subdifferentials $\partial f(x)$ and $\partial^\infty f(x)$, respectively. To this end, several auxiliary statements are necessary. First, we introduce a concept of the attentive convergence.

Definition 3.10 (Attentive convergence)**.** *For function $f[\mathbb{R}^m \to \mathbb{R}]$ we say that x_n converges to x f-attentively, $x_n \underset{f}{\to} x$, if $x_n \to x$ and $lim_{n\to\infty} f(x_n) = f(x)$.*

For a continuous function f, the topology of f-attentive convergence coincides with the topology generated by norm, i.e. $x_n \to x$ is equivalent to $x_n \underset{f}{\to} x$. On general, however, norm topology is finer.

We will see that the concept of f-attentive convergence is helpful in subsequent analysis of limiting notions.

Theorem 3.11 (Attentiveness of N_f)**.** *For any lower-semicontinuous quasi-convex function f, limiting normal operator N_f may be defined using f-attentive convergence only, i.e. the following equation holds*

$$N_f(\bar{x}) = \underset{x \underset{f}{\to} \bar{x}}{limsup}\, \widehat{N}_f(x). \tag{21}$$

Proof. First, we fix point $\bar{x}$ and denote

$$A = \underset{x \underset{f}{\to} \bar{x}}{\operatorname{limsup}}\, \widehat{N}_f(x). \tag{22}$$

By the definition of $N_f(\bar{x})$, it holds $A \subset N_f(\bar{x})$. Thus we have to show that $y \in N_f(\bar{x})$ implies $y \in A$ to fully prove our statement. For such y there exist $x_n \to \bar{x}$ and $y_n \to y$ satisfying $y_n \in \widehat{N}_f(x_n)$. We may assume that $f(x_n) \not\to f(\bar{x})$, for otherwise $y \in A$ directly by the definition. Now, we denote $\sigma = \operatorname{limsup}_n f(x_n)$. Then, according to the lower-semicontinuity of f, one has $\sigma > f(\bar{x})$. Indeed, either $\sigma \geq \operatorname{liminf}_n f(x_n) > f(\bar{x})$ or $\operatorname{liminf}_n f(x_n) = f(\bar{x})$ and then, since the sequence $f(x_n)$ doesn't converges to $f(\bar{x})$, $\operatorname{limsup}_n f(x_n) > \operatorname{liminf}_n f(x_n) = f\bar{x})$. Thus, we may take subsequence of x_n such that $f(x_n) \to \sigma$ and $f(x_n) > f(\bar{x})$ for all n. It implies $S_f(\bar{x}) \subset S_f(x_n)$ and so

$$y_n \in \widehat{N}_f(x_n) = (S_f(x_n) - x_n)^o \subset (S_f(\bar{x}) - x_n)^o\,. \tag{23}$$

Further, we take any $z \in S_f(\bar{x})$ and rewrite the previous inclusion as

$$\langle y_n, z - x_n \rangle \leq 0. \tag{24}$$

Now, letting $n \to \infty$ we obtain $\langle y, z - \bar{x} \rangle \leq 0$ for any $z \in S_f(\bar{x})$ and so

$$y \in (S_f(\bar{x}) - \bar{x})^o = \widehat{N}_f(\bar{x}). \tag{25}$$

Thus, the proof is finished since $\widehat{N}_f(\bar{x}) \subset A$ by the definition. □

Since we need to to clarify the relation of N_f and ∂f, the previous lemma is of no use until we obtain a similar result for limiting normal cone.

Lemma 3.12 (Attentiveness of N_{epif})**.** *For a lower-semicontinuous function f it holds*

$$N_{epif}(\bar{x}, f(\bar{x})) = \underset{x \underset{f}{\to} \bar{x}}{limsup}\, \widehat{N}_{epif}(x, f(x)). \tag{26}$$

Proof. We denote the right-hand side of (26) as A

$$A = \limsup_{\substack{x \to \bar{x} \\ f}} \widehat{N}_{epif}(x, f(x)). \tag{27}$$

From the definition of $N_{epif}(\bar{x}, f(\bar{x}))$ it follows that $A \subset N_{epif}(\bar{x}, f(\bar{x}))$. Thus, we need to show that $y \in N_{epif}(\bar{x}, f(\bar{x}))$ implies $y \in A$ to prove the statement. For such y there exists $(x_n, z_n) \underset{epif}{\to} (\bar{x}, f(\bar{x}))$ and $y_n \in \widehat{N}_{epif}(x_n, z_n)$ satisfying $y_n \to y$. Observing $\limsup_{n\to\infty} f(x_n) \leq f(\bar{x})$ implied by $f(x_n) \leq z_n$, we have also $x_n \underset{f}{\to} \bar{x}$ using lower-semicontinuity of f. We finish the proof establishing $y_n \in \widehat{N}_{epif}(x_n, f(x_n))$. We observe $epif - (x_n, f(x_n)) \subset epif - (x_n, z_n)$ and thus $T_{epif}(x_n, f(x_n)) \subset T_{epif}(x_n, z_n)$. In other words, $\widehat{N}_{epif}(x_n, f(x_n)) \supset \widehat{N}_{epif}(x_n, z_n)$ and so $y \in A$. □

Now, we may state and prove the final theorem of this article.

Theorem 3.13 (Relation of $N_f(x)$ and $\partial f(x)$)**.** *For a lower-semicontinuous function f we have*

$$pos\{\partial f(\bar{x})\} \cup \partial^\infty f(\bar{x}) \subset N_f(\bar{x}), \tag{28}$$

where equality holds provided $0 \notin \partial f(\bar{x})$.

Proof. Inclusion (28) may be verified directly. We observe that

$$(S_f(x) - x) \times \mathbb{R}^+ \subset \mathrm{epi}(f) - (x, f(x)). \tag{29}$$

Thus also $T_f(x) \times \mathbb{R}^+ \subset T_{epif}(x, f(x))$ and so $\widehat{N}_f(x) \times \mathbb{R}^- \supset \widehat{N}_{epif}(x, f(x))$. Applying outer limit with respect to f-attentive convergence $x \underset{f}{\to} \bar{x}$ on both sides we have

$$N_f(\bar{x}) \times \mathbb{R}^- \supset N_{epif}(\bar{x}, f(\bar{x})) \tag{30}$$

using Theorem 3.11 and Lemma 3.12. Projection on $\mathrm{dom}(f)$ together with Lemma 2.13 completes the proof of (28).

The opposite inclusion is more difficult. For any lower-semicontinuous f we have

$$\widehat{N}_f(x) = \widehat{N}_{S_f(x)}(x) \subset N_{S_f(x)}(x), \tag{31}$$

and so we may adopt [1, Proposition 10.3] to our notation obtaining

$$\widehat{N}_f(x) \subset \mathrm{pos}\{\partial f(x)\} \cup \partial^\infty f(x) \tag{32}$$

valid whenever $0 \notin \partial f(x)$ and thus also for all x near $\bar{x}$. We rewrite the right hand side of (32) according to Lemma 2.13

$$\widehat{N}_f(x) \subset \mathrm{Proj}_{domf}[N_{epif}(x, f(x))], \tag{33}$$

where Proj_{domf} is a canonical projection from $\mathbb{R}^m \times \mathbb{R}$ to $\mathbb{R}^m$. Further, we have also

$$N_f(\bar{x}) = \limsup_{\substack{x \to \bar{x} \\ f}} \widehat{N}_f(x) \subset \limsup_{\substack{x \to \bar{x} \\ f}} \mathrm{Proj}_{domf}\,[N_{epif}(x, f(x))]. \tag{34}$$

Since $\mathrm{Proj}_{dom f}$ is linear and $N_{epi f}(x, f(x))$ outer-semicontinuous and cone-valued, we may apply Lemma 2.6 holding with equality as

$$\mathrm{Proj}^{-1}_{dom f}(0) \cap N_{epi f}(\bar{x}, f(\bar{x})) = \{0\} \tag{35}$$

owning to the assumption $0 \notin \partial f(\bar{x})$. Then

$$N_f(\bar{x}) \subset \mathrm{Proj}_{dom f}[N_{epi f}(\bar{x}, f(\bar{x}))], \tag{36}$$

which proves our statement regarding Lemma 2.13 again. □

Theorem 3.13 is important as it indicates to which extent N_f may bring some novelty when compared with ∂f and $\partial^\infty f$. We see that outside stationary points, $0 \notin \partial f(x)$, both approaches are equivalent in a sense. The next example shed a more light upon this relation.

Example 3.14 (Necessity of $0 \notin \partial f(x)$)**.** *Consider $f(x) = x^3$ at $x = 0$. Then, $N_f(0) = [0, \infty)$ whereas $\partial f(0) = \partial^\infty f(0) = \{0\}$. Therefore, the equality in Theorem 3.13 does not hold. Strict inclusion $\partial f(0) \subsetneq N_f(0)$, however, does not mean that N_f is less informative than ∂f. In this example it indicates non-stationarity of $f(x)$ at 0 in opposite to $\partial f(0)$.*

Finally, we note that conditions for equality in Theorem 3.13 are substantially weaker than in the case of regular normal operator $\widehat{N}_f$, see again [1, Proposition 10.3].

4 Conclusion

We introduced a new notion of the limiting normal operator and shown its outer-semicontinuity and quasi-monotonicity. Then, also a clear relation to limiting subdifferential was established. However, the calculus rules, which truly verify the real usability of the limiting normal operator, are to be developed in the future. Nonetheless, such program should be feasible owning to the local nature of the limiting normal operator.

References

[1] R.T. Rockafellar and R.J.B. Wets. *Variational analysis.* Grundlehren der mathematischen Wissenschaften. Springer, 1998.

[2] J. Borde and J. P. Crouzeix. Continuity properties of the normal cone to the level sets of a quasiconvex function. *Journal of Optimization Theory and Applications*, 66:415–429, 1990. 10.1007/BF00940929.

[3] D. Aussel and N. Hadjisavvas. Adjusted sublevel sets, normal operator, and quasiconvex programming. *SIAM J. on Optimization*, 16:358–367, June 2005.

[4] Boris S. Mordukhovich. *Variational analysis and generalized differentiation. I: Basic theory. II: Applications.* Grundlehren der Mathematischen Wissenschaften 330/331. Berlin: Springer. xxii, 579 p., xxii, 610 p., 2006.

[5] A. D. Ioffe and Jiří V. Outrata. On metric and calmness qualification conditions in subdifferential calculus. *Set-Valued Analysis*, 16:199–227, 2008.

Homogeneous Droplet Nucleation Modeled Using the Gradient Theory Combined with the PC-SAFT Equation of State*

Barbora Planková

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

advisor: Jan Hrubý, Department of Thermodynamics, Institute of Thermomechanics AS CR, v.v.i.

Abstract. In this work, we used the density gradient theory (DGT) combined with the cubic equation of state (EoS) by Peng and Robinson (PR) and the perturbed chain (PC) modification of the SAFT EoS developed by Gross and Sadowski [1]. The PR EoS is based on very simplified physical foundations, it has significant limitations in the accuracy of the predicted thermodynamic properties. On the other hand, the PC-SAFT EoS combines different intermolecular forces, e.g., hydrogen bonding, covalent bonding, Coulombic forces which makes it more accurate in predicting of the physical variables. We continued in our previous works [2, 3] by solving the boundary value problem which arose by mathematical solution of the DGT formulation and including the boundary conditions. Achieving the numerical solution was rather tricky; this study describes some of the crucial developments that helped us to overcome the partial problems. The most troublesome were computations for low temperatures where we achieved great improvements compared to [2]. We applied the GT for the n-alkanes: n-heptane, n-octane, n-nonane, and n-decane because of the availability of the experimental data. Comparing them with our numerical results, we observed great differences between the theories; the best results gave the combination of the GT and the PC-SAFT. However, a certain temperature-dependent deviation was observed that is not satisfactorily explained by the present theories.

This work will be presented at **Experimental fluid mechanics 2012** in Liberec (20.11.2012 - 23.11.2012) and whole text subsequently published in **The European Physical Journal**.

Keywords: Density gradient theory, nucleation, PC-SAFT, Cahn-Hilliard theory

Abstrakt. V této práci jsme zkombinovali gradientní teorii s Pengovou-Robinsonovou (PR) stavovou rovnicí a stavovou rovnicí PC-SAFT vytvořenou Grossem a Sadowskou [1]. Rovnice PR je založena na jednoduchých fyzikálních zákonitostech, takže přesnost, s jakou je schopna předpovědět termodynamické vlastnosti je tedy omezená. Rovnice PC-SAFT na druhou stranu kombinuje různé mezimolekulární síly jako vodíkové můstky, kovalentní vazby, či Coulombovy síly, které ji dělají daleko přesnější. Navázali jsme na naše předchozí práce [2, 3] tím, ze jsme řešili okrajovou úlohu, která vznikla matematickým vyřešením problému formulovaného gradinentí teorií a při zahrnutí okrajových podmínek. Dosáhnout numerického řešení bylo poněkud komplikované; tato studie popisuje některé zásadní invence, které nám pomohly překonat dílčí problémy. Nejobtížnější byly výpočty pro nízké teploty, kde jsme dosáhli velkých zlepšení oproti [2]. Aplikovali jsme gradientní teorii pro n-alkany: n-heptan, n-oktan, n-nonan a n-dekan. Tyto

*The project has been supported by grants GA ASCR No. IAA200760905, GACR Nos. 101/09/1633 and GPP101/11/P046 and MSMT LA09011.

látky byly zvoleny proto, že jejich experimentální data jsou k dispozici. Když jsme je porovnali s numerickými výsledky, objevili jsme velké rozdíly mezi oběma teoriemi. Nejlepšího výsledku dosáhla kombinace gradientní teorie a rovnice PC-SAFT. V porovnání dat se ovšem vystkytla odchylka závislá na teplotě; tato odchylka není současnými teoriemi vysvětlena.

Tato práce bude prezentována na konferenci **Experimental fluid mechanics 2012** v Liberci (20.11.2012 - 23.11.2012) a celý text následně publikován v žurnálu **The European Physical Journal**.

Klíčová slova: Gradientní teorie, nukleace, PC-SAFT, Cahn-Hilliardova teorie

References

[1] J. Gross and G. Sadowski. *Perturbed-chain saft: An equation of state based on a perturbation theory for chained molecules.* Ind . Eng. Chem. Res. **40** (2001), 1244–1260.

[2] J. Hrubý, D. G. Labetski, and M. E. H. van Dongen. *Gradient theory computation of the radius-dependent surfae tension and nucleation rate for n-nonane.* J. Chem. Phys. **127** (2007), 164720.

[3] V. Vinš, J. Hrubý, and B. Planková. *Droplet and bubble nucleation modeled by density gradient theory - cubic equation of state versus saft model.* EPJ Web of Conferences **25** (2012).

Numerická simulace dvoufázového stlačitelného proudění směsi v porézním prostředí*

Ondřej Polívka

3. 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 v Praze

Abstract. The paper deals with the numerical modeling of compressible two-phase flow of a mixture composed of several components in a porous medium. The mathematical model is formulated by means of extended 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 the total flux discretization and the finite volume method for the discretization of the transport equations. A new approach to flux approximation is proposed, allowing us not to determine the corresponding phases between elements. This approach provides exact local mass balance. The time discretization is carried out by the backward Euler method. The resulting large system of nonlinear algebraic equations is solved by the Newton-Raphson iterative method. Methane injection into a homogeneous 2D reservoir filled with propane in two phases is simulated in a horizontal and vertical cut.

Keywords: mixed-hybrid finite element method, finite volume method, Newton-Raphson method, two-phase compressible multicomponent flow, miscible displacement

Abstrakt. Článek pojednává o numerickém modelování stlačitelného dvoufázového proudění směsi o několika složkách v porézním prostředí. Matematický model je formulován pomocí rozšířeného 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 celkového toku a metody konečných objemů pro diskretizaci transportních rovnic. K aproximaci toků navrhujeme vlastní upwind přístup, který odbourává určování korespondujících fází mezi elementy. Tento přístup poskytuje přesnou lokální bilanci hmoty. Časová diskretizace je provedena zpětnou Eulerovou metodou. Výsledná soustava nelineárních algebraických rovnic je řešena Newtonovou-Raphsonovou iterační metodou. Na závěr je simulováno vtláčení metanu do homogenního 2D rezervoáru naplněného propanem ve dvou fázích v horizontálním a vertikálním řezu.

Klíčová slova: smíšená hybridní metoda konečných prvků, metoda konečných objemů, Newtonova-Raphsonova metoda, dvoufázové stlačitelné vícekomponentní proudění, mísitelné proudění

*Tato práce byla podpořena grantem „Development of Computational Models for Simulation of CO2 Sequestration" P105/11/1507 Grantové agentury České republiky a projektem „Computational methods in thermodynamics of multicomponent mixtures" Kontakt LH12064 Ministerstva školství, mládeže a tělovýchovy České republiky.

1 Úvod

Spolehlivá simulace dvoufázového transportu vícekomponentní směsi v podzemním porézním prostředí je důležitá při řešení řady problémů, jako je např. těžba ropy nebo sekvestrace CO_2. Klíčové pro tento druh proudění je správné rozhodnutí o počtu fází a jejich složení na každém výpočetním elementu. Dále je to pak správné provázání fázových toků mezi jednotlivými elementy tak, aby bylo splněno zachování hmoty mezi elementy. Tradiční přístupy [3] se snaží na základě jistých vlastností zkoumané směsi složitě propojovat jednotlivé fáze mezi elementy. Tento postup však často selhává (např. v nadkritické oblasti p-V diagramu nelze rozlišovat mezi fázemi).

V této práci se zabýváme numerickým modelováním stlačitelného dvoufázového 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) s použitím upwind metody pro diskretizaci toků na hranách elementů triangulace. Výsledné numerické schéma pak zaručuje lokální bilanci hmoty a korektní ošetření fázových toků mezi elementy. Odpadá tak nutnost složitého určování korespondujících fází mezi jednotlivými elementy. Tlak a rozdělení směsi mezi fáze je určeno prostředky rovnovážné termodynamiky.

2 Matematická formulace

Nechť $\Omega \subset \mathbb{R}^2$ je omezená oblast s porozitou ϕ [-] a (t_0, τ) je časový interval [s]. Uvažujme dvoufázové stlačitelné proudění tekutiny o n_c složkách v oblasti při konstantní teplotě T [K]. Při zanedbání difúze je transport jednotlivých složek v oblasti Ω a čase (t_0, τ) (dle [10]) popsán následujícími rovnicemi

$$\frac{\partial(\phi c_i)}{\partial t} + \nabla \cdot \mathbf{q}_i = F_i\,, \quad i = 1, \ldots, n_c\,, \tag{1}$$

$$\mathbf{q}_i = \sum_\alpha c_{\alpha,i}\,\mathbf{v}_\alpha\,, \tag{2}$$

$$\mathbf{v}_\alpha = -\lambda_\alpha \mathbf{K}(\nabla p - \varrho_\alpha \mathbf{g})\,, \tag{3}$$

kde neznámé veličiny $c_{\alpha,i}$, jsou molární koncentrace fáze α komponent směsi [$\mathrm{mol\,m^{-3}}$]. V rovnici (1) je ϕ porozita [-] a F_i zdrojový člen [$\mathrm{mol\,m^{-3}s^{-1}}$]. V rozšířeném Darcyho zákoně (3) je $\lambda_\alpha = \lambda_\alpha(S_\alpha)$ mobilita fáze α závislá na saturaci S_α, $\mathbf{K} \in [L^\infty(\Omega)]^{2\times 2}$ vlastní permeabilita [$\mathrm{m^2}$] (obecně symetrický stejnoměrně eliptický tenzor [9]), ∇p gradient tlaku p [Pa], $\mathbf{g}$ vektor gravitačního zrychlení [$\mathrm{m\,s^{-2}}$] a $\varrho_\alpha = \sum_{i=1}^{n_c} c_{\alpha,i} M_i$ hustota tekutiny ve fázi α [$\mathrm{kg\,m^{-3}}$] (M_i je molární hmotnost komponenty i [$\mathrm{kg\,mol^{-1}}$]) . Pomocí Darcyho zákona (3) můžeme spočítat fázový tok $\mathbf{q}_\alpha$ a celkový tok $\mathbf{q}$ jako

$$\mathbf{q} = \sum_\alpha \mathbf{q}_\alpha = \sum_\alpha c_\alpha \mathbf{v}_\alpha\,. \tag{4}$$

Rozdělení komponent mezi fáze je dáno následujícími termodynamickými vztahy

$$\sum_\alpha c_{\alpha,i}\, S_\alpha = c_i\,, \quad \sum_\alpha S_\alpha = 1\,, \tag{5a}$$

$$p\,(T, c_{\alpha,i}\,, \ldots, c_{\alpha,n_c}) = p\,(T, c_{\beta,i}\,, \ldots, c_{\beta,n_c})\;, \quad \forall \alpha \neq \beta\,, \tag{5b}$$

$$\mu_i\,(T, c_{\alpha,i}\,, \ldots, c_{\alpha,n_c}) = \mu_i\,(T, c_{\beta,i}\,, \ldots, c_{\beta,n_c})\;, \quad \forall \alpha \neq \beta\,, \quad \forall i \in \widehat{n_c}\,. \tag{5c}$$

Rovnice (5a) vyjadřují bilanci hmoty a objemu, (5b) mechanickou rovnováhu, kde je tlak dán Pengovou-Robinsonovou stavovou rovnicí (PR EOS) [13]. Rovnice (5c) představuje chemickou rovnováhu, přičemž μ_i je chemický potenciál i-té komponenty. Přesný tvar (5c) a způsob řešení (5) je popsán v [11]. Postup jak určit počet fází lze nalézt v [12].

Počáteční a okrajové podmínky jsou následující

$$c_i(\mathbf{x}, 0) = c_i^0(\mathbf{x})\,, \quad \mathbf{x} \in \Omega\,,\ i = 1, \ldots, n_c\,, \tag{6a}$$

$$p(\mathbf{x}, t) = p^D(\mathbf{x}, t)\,, \quad \mathbf{x} \in \Gamma_p\,,\ t \in (t_0, \tau)\;, \tag{6b}$$

$$\mathbf{q}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}) = 0\,, \quad \mathbf{x} \in \Gamma_q\,,\ t \in (t_0, \tau)\;, \tag{6c}$$

kde $\mathbf{n}$ je jednotkový vektor vnější normály k hranici $\partial\Omega$. Rovnice (6b), (6c) určují Dirichletovy a Neumannovy okrajové podmínky na částech hranice Γ_p, resp. Γ_q, přičemž platí $\Gamma_p \cup \Gamma_q = \partial\Omega$ a $\Gamma_p \cap \Gamma_q = \emptyset$.

3 Numerické řešení

Systém rovnic (1)–(6) je řešen numericky kombinací MHFEM aplikovanou na celkový tok (4) a FVM aplikovanou na transportní rovnice (1). Časová diskretizace je provedena zpětnou 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 celkového toku

Celkový tok $\mathbf{q}$ lze aproximovat v Raviartově-Thomasově prostoru nejnižšího řádu (RT^0_K) nad elementem $K \in \mathcal{T}_\Omega$ jako

$$\mathbf{q} = \sum_{E \in \partial K} q_{K,E} \mathbf{w}_{K,E}\,, \tag{7}$$

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 [1, 2, 10]).

Po dosazení z rovnice (3) do (4) můžeme vyjádřit gradient tlaku jako

$$\nabla p = -\frac{\mathbf{K}^{-1}\mathbf{q}}{\sum_\alpha c_\alpha \lambda_\alpha} + \varrho \mathbf{g}\,, \quad \varrho = \frac{\sum_\alpha c_\alpha \lambda_\alpha \varrho_\alpha}{\sum_\alpha c_\alpha \lambda_\alpha}\,, \tag{8}$$

kde ϱ je celková hustota. Vynásobením vztahu (8) pro ∇p bazickou funkcí $\mathbf{w}_{K,E}$, integrací přes K, využitím vlastností prostoru RT^0_K, vztahu (7), Greenovy věty a věty o střední hodnotě odvodíme diskrétní tvar celkového toku

$$q_{K,E} = \sum_{\alpha\in\Pi(K)} c_{\alpha,K}\lambda_{\alpha,K}\left(\alpha^K_E p_K - \sum_{E'\in\partial K}\beta^K_{E,E'}p_{K,E'} + \gamma^K_E \varrho_K\right), \quad E\in\partial K. \tag{9}$$

V rovnici (9) značí $\Pi(K)$ všechny fáze na elementu K; $\alpha^K_E, \beta^K_{E,E'}$ a γ^K_E jsou 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'; $c_{\alpha,K}$, $\lambda_{\alpha,K}$, ϱ_K značí střední hodnotu koncentrace fáze α, mobility fáze α a celkové hustoty na trojúhelníku K.

Ve smíšená formulaci požadujeme spojitost normálové složky toku a tlaku na hraně E mezi sousedícími elementy $K, K' \in \mathcal{T}_\Omega$, což lze zapsat jako

$$q_{K,E} + q_{K',E} = 0\,, \tag{10}$$
$$p_{K,E} = p_{K',E} =: p_E\,. \tag{11}$$

Okrajové podmínky (6b), (6c) vyjádřené v diskrétním tvaru jsou

$$p_{K,E} = p^D(E)\,, \quad \forall E\subset\Gamma_p\,, \tag{12a}$$
$$q_{K,E} = 0\,, \quad \forall E\subset\Gamma_q\,, \tag{12b}$$

kde $p^D(E)$ je předepsaná hodnota tlaku p na hraně E.

Tok můžeme eliminovat dosazením $q_{K,E}$ ze vztahu (9) do rovnic (10) a (12b). Pro další odvození označme časově závislé veličiny v čase t_{n+1} horním indexem $n+1$. Pak rovnice (9)–(12) přejdou na následující soustavu n_e lineárních algebraických rovnic

$$\mathcal{F}_E \equiv \begin{cases} \displaystyle\sum_{K:E\in\partial K}\left(\sum_{\alpha\in\Pi(K)} c^{n+1}_{\alpha,K}\lambda^{n+1}_{\alpha,K}\right)\left(\alpha^K_E p^{n+1}_K - \sum_{E'\in\partial K}\beta^K_{E,E'}p^{n+1}_{K,E'} + \gamma^K_E\varrho^{n+1}_K\right) = 0 & \forall E\not\subset\partial\Omega\,, \\ \displaystyle\sum_{\alpha\in\Pi(K)} c^{n+1}_{\alpha,K}\lambda^{n+1}_{\alpha,K}\left(\alpha^K_E p^{n+1}_K - \sum_{E'\in\partial K}\beta^K_{E,E'}p^{n+1}_{K,E'} + \gamma^K_E\varrho^{n+1}_K\right) = 0 & \forall E\subset\Gamma_q\,, \\ p^{n+1}_{K,E} - p^D(E) = 0 & \forall E\subset\Gamma_p\,. \end{cases} \tag{13}$$

Zde symbol $\sum_{K:E\in\partial K}$ značí sčítání přes elementy obsahující hranu E. Podobný postup vedoucí ke smíšené hybridní formulaci lze nalézt v [9].

3.2 Aproximace transportních rovnic

Transportní rovnice (1) s počátečními a okrajovými podmínkami (6) jsou diskretizovány pomocí FVM [8]. Integrací (1) přes libovolný element $K\in\mathcal{T}_\Omega$ a použitím Greenovy věty dostaneme

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_K \phi(\mathbf{x})c_i(\mathbf{x},t) + \int_{\partial K}\mathbf{q}_i(\mathbf{x},t)\cdot\mathbf{n}_{\partial K}(\mathbf{x}) = \int_K F_i(\mathbf{x})\,, \quad i=1,\ldots,n_c\,. \tag{14}$$

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 (14) na

$$\frac{\mathrm{d}(\phi_K c_{i,K})}{\mathrm{d}t}|K| + \sum_{E\in\partial K}\int\limits_E \mathbf{q}_i \cdot \mathbf{n}_{K,E} = F_{i,K}|K|\,, \tag{15}$$

kde $\mathbf{q}_i$ lze dosazením z (8) do (3) a vzniklého výrazu pak do (2) vyjádřit jako

$$\mathbf{q}_i = \left(\sum_\beta c_\beta\,\lambda_\beta\right)^{-1} \sum_\alpha c_{\alpha,i}\,\lambda_\alpha \left(\mathbf{q} - \sum_\beta c_\beta\,\lambda_\beta\,(\varrho_\beta - \varrho_\alpha)\,K\mathbf{g}\right). \tag{16}$$

Integrál v (15) můžeme pomocí (16) aproximovat (upwind) jako

$$\int\limits_E \mathbf{q}_i \cdot \mathbf{n}_{K,E} \approx \sum_{\alpha\in\Pi(K,E)^+} q_{\alpha,i,K,E} - \sum_{\beta\in\Pi(K',E)^+} q_{\beta,i,K',E}\,, \quad \forall E \notin \partial\Omega\,, \tag{17}$$

kde $K \cap K' = E\,,\ \Pi(K,E)^+ = \{\alpha \in \Pi(K) : q_{\alpha,i,K,E} > 0\}$ a

$$q_{\alpha,i,K,E} = \frac{c_{\alpha,i,K}\,\lambda_{\alpha,K}}{\sum\limits_{\alpha'\in\Pi(K)} c_{\alpha',K}\,\lambda_{\alpha',K}} \left(q_{K,E} - \sum_\beta c_{\beta,K}\,\lambda_{\beta,K}\,(\varrho_{\beta,K} - \varrho_{\alpha,K})\,\gamma_E^K\right). \tag{18}$$

Vzhledem k okrajovým podmínkám (6b), (6c) (a neuvažováním úlohy s vtokovou částí hranice) lze vztah (17) rozšířit i na hrany z hranice, vynecháme-li v něm druhý člen.

Časová derivace $c_{i,K}$ v (15) je aproximována časovou diferencí s časovým krokem Δt_n. Při použití zpětné Eulerovy metody [8], máme pro každé n, všechny elementy $K \in \mathcal{T}_\Omega$ a komponenty $i = 1, \ldots, n_c$

$$\mathcal{F}_{K,i} \equiv \phi_K|K|\frac{c_{i,K}^{n+1} - c_{i,K}^n}{\Delta t_n} + \sum_{E\in\partial K}\left(\sum_{\alpha\in\Pi(K,E)^+} q_{\alpha,i,K,E} - \sum_{\beta\in\Pi(K',E)^+} q_{\beta,i,K',E}\right)^{n+1} - F_{i,K}|K| = 0\,, \tag{19}$$

kde $q_{\alpha,i,K,E}$ je dáno (18). Poznamenejme, že schéma je plně implicitní.

Počáteční podmínku (6a) v diskrétním tvaru můžeme psát jako

$$c_{i,K}^0 = c_i^0(K)\,, \quad \forall K \in \mathcal{T}_\Omega\,,\ i = 1, \ldots, n_c\,. \tag{20a}$$

3.3 Linearizace schémat z MHFEM a FVM

V rovnicích (13) a (19) jsme označili $\mathcal{F}_E$ a $\mathcal{F}_{K,i}$, (pro hranu $E \in \{1, \ldots, n_e\}$, element $K \in \{1, \ldots, n_k\}$ a komponentu $i \in \{1, \ldots, n_c\}$) výrazy, které budou tvořit složky vektoru $\boldsymbol{\mathcal{F}}$. Použitím NRM pak řešíme nelineární soustavu algebraických rovnic o $n_k \times n_c + n_e$ rovnic

$$\boldsymbol{\mathcal{F}} = [\mathcal{F}_{1,1}, \ldots, \mathcal{F}_{1,n_c}, \ldots, \mathcal{F}_{n_k,1}, \ldots, \mathcal{F}_{n_k,n_c}; \mathcal{F}_1, \ldots, \mathcal{F}_{n_e}]^T = \mathbf{0} \tag{21}$$

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}, $E \in \{1, \ldots, n_e\}$. Jacobiho matice $\mathbf{J}$ linearizované soustavy je řídká, avšak nesymetrická. Je rozdělena na 4 bloky, jejichž prvky lze napočítat analyticky podle následujících vztahů

$$(\mathbf{J}_{K,K'})_{i,j} = \frac{\partial \mathcal{F}_{K,i}}{\partial c_{j,K'}^{n+1}}, \quad (\mathbf{J}_{K,E})_i = \frac{\partial \mathcal{F}_{K,i}}{\partial p_E^{n+1}}, \quad (\mathbf{J}_{E,K})_j = \frac{\partial \mathcal{F}_E}{\partial c_{j,K}^{n+1}}, \quad J_{E,E'} = \frac{\partial \mathcal{F}_E}{\partial p_{E'}^{n+1}}, \tag{22}$$

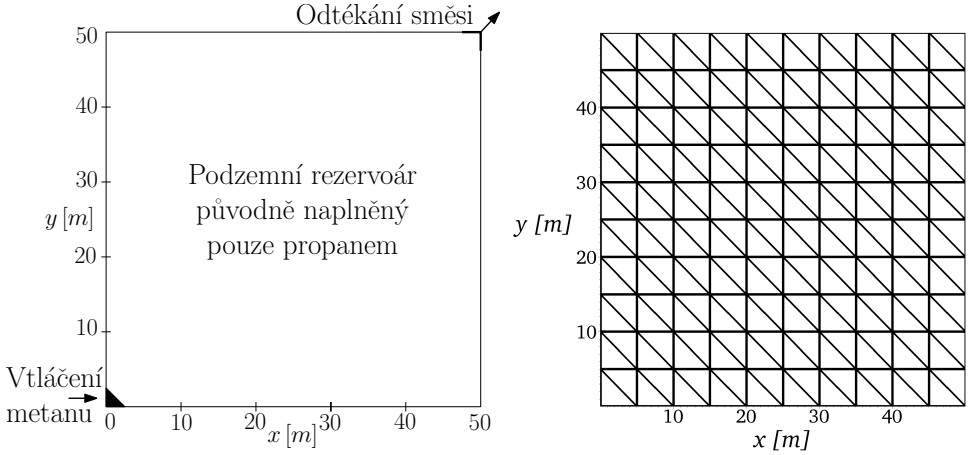
kde $J_{E,E'}$ je prvek matice $\mathbf{J}_{E,E'}$, $i, j = 1, \ldots, n_c$; $K, K' = 1, \ldots, n_k$; $E, E' = 1, \ldots, n_e$. Vektor neznámých obsahuje korekce molárních koncentrací $\delta c_{i,K}$ a tlaků na hranách δp_E, které jsou spočtené v každé iteraci NRM a přičteny k hodnotám p_E^{n+1}, $c_{i,K}^{n+1}$ z předchozí iterace. Iterační procedura končí při splnění podmínky

$$\|\boldsymbol{\mathcal{F}}\| < \varepsilon \tag{23}$$

pro zvolené $\varepsilon > 0$ (viz [14]).

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 $p = 6.9 \cdot 10^6\,\mathrm{Pa}$ a teplotě $T = 311\,\mathrm{K}$. V levém dolním rohu rezervoáru je vtláčen metan a v pravém horním rohu směs metanu a propanu odtéká (obr. 1). Hodnota vtláčení $F_{1,K}$ je $42.5\,\mathrm{m}^2/\mathrm{den}$ při tlaku 1 atm a teplotě 293 K. Fyzikálně-chemické vlastnosti směsi jsou shrnuty v tab. 1. Při tomto nastavení se směs během proudění může nacházet ve dvoufázovém stavu. Hranice oblasti je nepropustná kromě odtokového rohu, kde je udržován tlak $p = 6.9 \cdot 10^6\,\mathrm{Pa}$. Struktura výpočetní sítě o $2 \times 10 \times 10$ elementech je zobrazena na obr. 1. Parametr ε z konvergenčního kritéria NRM (23) byl zvolen 10^{-6}. K řešení soustavy lineárních rovnic byla použita knihovna UMFPACK [4, 5, 6, 7].



Obrázek 1: Schéma simulovaného rezervoáru a struktura výpočetní sítě.

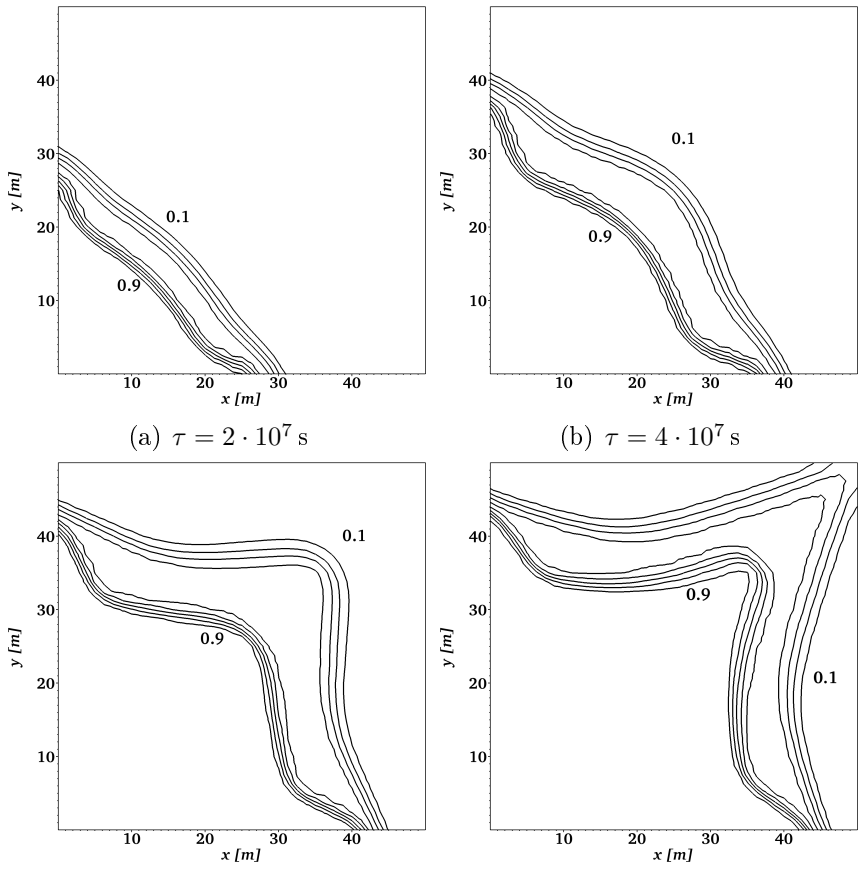
i (složka směsi)	p_{ci} [Pa]	T_{ci} [K]	V_{ci} [$\mathrm{m^3 mol^{-1}}$]
1 (CH_4)	$4.58373 \cdot 10^6$	$1.89743 \cdot 10^2$	$9.897054 \cdot 10^{-5}$
2 (C_3H_8)	$4.248 \cdot 10^6$	$3.6983 \cdot 10^2$	$2.000001 \cdot 10^{-4}$

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 Úloha bez gravitace

V první úloze budeme simulovat vtláčení metanu do horizontálního rezervoáru (tj. s nulovou gravitací) naplněného propanem. Na obr. 2 jsou zobrazeny izočáry molárního zlomku $\frac{c_1}{c_1+c_2}$ v různých časech. Nejblíže k vtláčecímu vrtu je vždy hodnota molárního zlomku 0.9 a s každou další izočárou směrem k odtokovému rohu se hodnota zmenší o 0.1. Výpočet byl proveden na síti $2 \times 40 \times 40$ elementů.



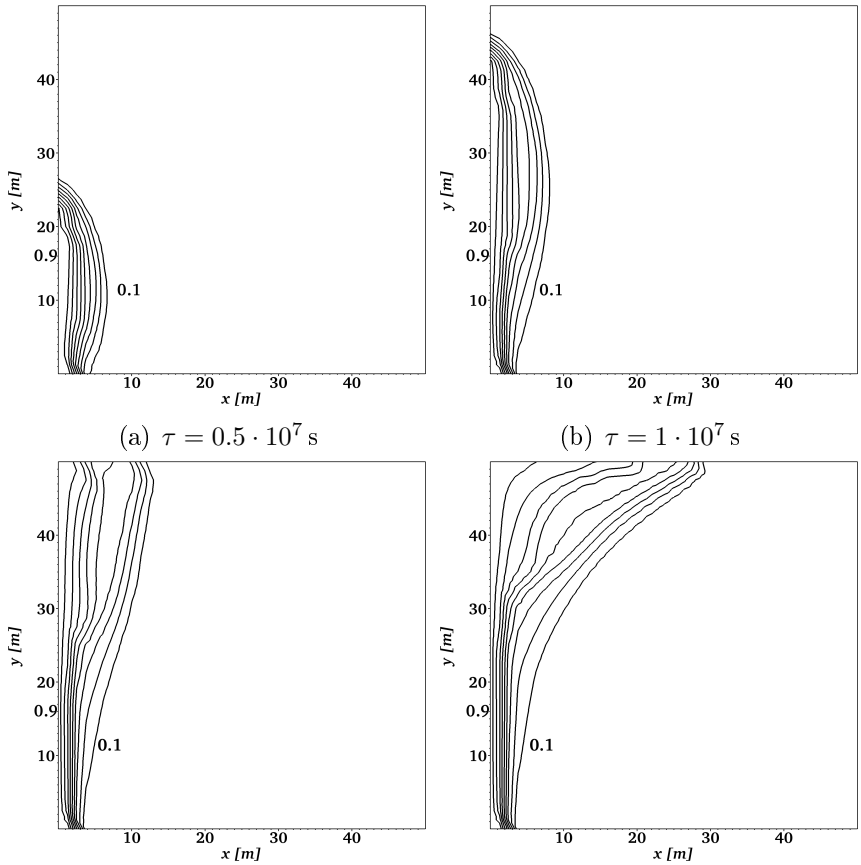
(a) $\tau = 2 \cdot 10^7$ s (b) $\tau = 4 \cdot 10^7$ s

(c) $\tau = 6 \cdot 10^7$ s (d) $\tau = 8 \cdot 10^7$ s

Obrázek 2: Molární zlomky metanu $c_1/(c_1+c_2)$ na síti $2 \times 40 \times 40$. Izočáry jsou rozloženy rovnoměrně mezi dvěma zobrazenými hodnotami.

4.2 Úloha s gravitací

Ve druhé úloze budeme simulovat vtláčení metanu do vertikálního rezervoáru (tj. s gravitací) naplněného propanem. Na obr. 3 jsou zobrazeny izočáry molárního zlomku $\frac{c_1}{c_1+c_2}$ v různých časech. Nejblíže k vtláčecímu vrtu je vždy hodnota molárního zlomku 0.9 a s každou další izočárou směrem k odtokovému rohu se hodnota zmenší o 0.1. Výpočet byl proveden na síti $2 \times 40 \times 40$ elementů.



(a) $\tau = 0.5 \cdot 10^7$ s (b) $\tau = 1 \cdot 10^7$ s

(c) $\tau = 1.5 \cdot 10^7$ s (d) $\tau = 2 \cdot 10^7$ s

Obrázek 3: Molární zlomky metanu $c_1/(c_1 + c_2)$ na síti $2 \times 40 \times 40$. Izočáry jsou rozloženy rovnoměrně mezi dvěma zobrazenými hodnotami.

5 Závěr

V této práci jsme popsali numerické schéma založené na kombinaci MHFEM a FVM pro řešení dvoufázového stlačitelného proudění směsi v porézním prostředí. Oproti tradičním přístupům nemusíme složitě určovat odpovídající si fáze na hraně mezi dvěma elementy, protože to navržená upwind technika ani diskretizace numerických toků nevyžaduje. Přesto náš přístup zaručuje lokální bilanci hmoty, která je důležitá zejména při řešení problémů v heterogenním prostředí. Numerický model jsme použili pro simulaci dvousložkové směsi – metan, propan proudící dvoufázově v horizontálním nebo vertikálním rezervoáru.

Literatura

[1] F. Brezzi, M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, New York Inc. (1991).

[2] 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).

[3] Z. Chen, G. Ma Y. Huan. *Computational Methods for Multiphase Flows in Porous Media.* SIAM, Philadelphia (2006).

[4] 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.

[5] T. A. Davis. *Algorithm 832: UMFPACK, an unsymmetric-pattern multifrontal method.* ACM Transactions on Mathematical Software, vol 30, no. 2 (2004), pp. 196–199.

[6] 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.

[7] 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.

[8] R. J. Leveque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge University Press, Cambridge (2002).

[9] 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.

[10] J. Mikyška, A. Firoozabadi. *Implementation of higher-order methods for robust and efficient compositional simulation.* Journal of Computational Physics, 229 (2010), 2898–2913.

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

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

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

[14] A. Quarteroni, R. Sacco, F. Saleri. *Numerical Mathematics.* Springer-Verlag, New York (2000).

Design of Refactoring Tool for C++ Language*

Michal Rost

2nd year of PGS, email: rost.michal@gmail.com
Department of Software Engineering in Economics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. Refactoring is widely utilized by programmers to improve the existing code. However, this process, if performed manually, consumes much time; this is the reason why automated refactoring tools appeared in many integrated development environments during the last decade. Refactoring of a code of some programming language requires previous syntactic analysis of the code. Therefore, refactoring of a C++ code is a complex issue with regard to hardly recognizable context of C++. This paper summarizes main refactoring methods then focuses primarily on the process of syntax analysis with respect to the C++ language. At the end of the paper the current progress in refactoring tool development is described.

Keywords: C++, refactoring, syntactic analysis

Abstrakt. Refaktorování je proces, který je programátory široce využíván ke zlepšení vlastností již existujícího zdrojového kódu. Pokud je refaktorování prováděno ručně, může trvat poměrně dlouhou dobu. Z tohoto důvodu se v posledních letech stávají automatické refaktorovací nástroje běžnou součástí vývojových prostředí. Refaktorování kódu ve zvoleném jazyce je závislé na předešlé syntaktické analýze tohoto kódu. Proto není vytvoření refaktorovacího nástroje pro jazyk C++ jednoduchou záležitostí, zejména kvůli jeho obtížně rozpoznatelnému kontextu. Tento článek shrnuje hlavní refaktorovací techniky, následně se zaměřuje především na proces syntaktické analýzy jazyka C++. V závěru článek dokumentuje průběh práce na refaktorovacím nástroji.

Klíčová slova: C++, refaktorování, syntaktická analýza

1 Introduction

Refactoring is a process during which internal structure of software is changed, but behaviour (functionality) of refactored software remains unchanged [5]; in other words, during refactoring is a poorly-designed code transformed into a well-designed [5] form which is easily readable, maintainable and which does not contain duplicated parts.

1.1 Code smells and refactoring methods

In order to distinguish between the badly designed code and the good one a term *code smell* has been introduced [5]. There are various kinds of code smells; each kind refers to a specific design issue. Moreover, each type of smell is connected with one or more

*This work has been supported by the grant SGS 11/167

refactoring techniques that are used to fix the related issue. In Table 1 the most common refactoring techniques are listed together with brief description; Table 2 shows a list of frequent code smells together with refactoring techniques that should be used for their removal. More sophisticated taxonomy of code smells was introduced by Mäntylä and Lassenius [10] who divided code smells into five groups with respect to negative contributions of each smell.

Table 1: Common refactoring techniques

	Name	*Description*	*Reverse*
1	Encapsulate field	Creates *setter* and *getter* for a selected attribute of a given class.	
2	Extract class	Extracts selected attributes and methods into a new class.	8
3	Extract interface	In C++ this is equivalent to extracting a superclass with virtual methods.	
4	Extract superclass	Extracts selected attributes and methods into a new parent class.	
5	Extract method	Extracts reusable part of some method into a new one.	9
6	Form template	Creates template for a given method or class.	
7	Hide delegate	Hides given class to user and makes its methods available through "middle man" class.	11
8	Inline class	Moves all its attributes/methods into another class and deletes it.	2
9	Inline method	Puts the method's body into the body of its callers and remove the method.	5
10	Move method	Moves a method from one class to another one.	
11	Remove middle man	Makes methods of given class available to user directly without the "middle man".	7
12	Rename	Changes the name of a selected class, method, or variable in all its occurences.	
13	Replace conditional logic with polymorphism	Replaces a conditional with a call of virtual method of a polymorphic object.	
14	Replace temp with query	Replaces a read-only temporary variable in a method with a query function (getter) call.	

1.2 The state of the art

The detection of smells is upon a programmer who is expected to discover smells in the code and then to fix them. Transformation of code via refactoring techniques may be performed manually by the programmer, or with utilization of automated tools offered by many present-day integrated development environments (IDE). Despite the fact that large portion of present-day IDEs contain advanced refactors, so far, there is no non-proprietary IDE or tool which allows full refactoring of the C++ language [8].

Table 2: Frequent code smells

Name	*Related techniques*
Alternative Classes with Different Interfaces	Extract interface
Divergent Change	Extract class
Duplicated code	Extract class/superclass/method, Form template
Conditional complexity	Replace conditional logic with polymorphism
Large class	Extract class, Move method
Lazy class	Inline class
Long method	Extract method
Long parameter list	Extract class
Message chains	Hide delegate
Middle man	Remove middle man
Uncommunicative name	Rename

2 Decomposition of refactoring tool

To perform the automated refactoring of a selected part of code, the original code has to be transformed into a form of *abstract syntax tree* (AST) [1]; as shown in Figure 1, mentioned transformation, often referred to as *parsing*, consists of two steps: lexical analysis and syntactic analysis. Once the syntax analysis of the code is completed and AST is produced, the refactoring process may start.

2.1 Lexer

During the lexical analysis [1] a stream of characters is read from the input. Consequently individual characters are grouped into meaningful sequences (*lexemes*). Finally, for each lexeme an output *token* [1] is created. Application or module which performs the lexical analysis is referred to as *Lexer*. Each lexer has to be provided with a list of valid tokens so as the input stream can be split into them.

2.2 Syntactic analyser

Within the syntax analysis [1], performed by syntactic analyser, a sequence of tokens is taken as the input and a tree-like representation of code (syntax tree) is created to show the grammatical structure of the token stream. The input token stream is transformed into the syntax tree on the basis of a grammar.

2.2.1 Grammar

A *formal grammar* (grammar) [3, 9] is defined as (1) where T is a finite set of *terminal symbols*, N is a finite set of *nonterminal symbols*, P represents a finite set of *production rules* and S denotes an unique nonterminal *start symbol*.

$$G = (T, N, P, S) \quad N \cap T = \emptyset \quad S \in N \tag{1}$$

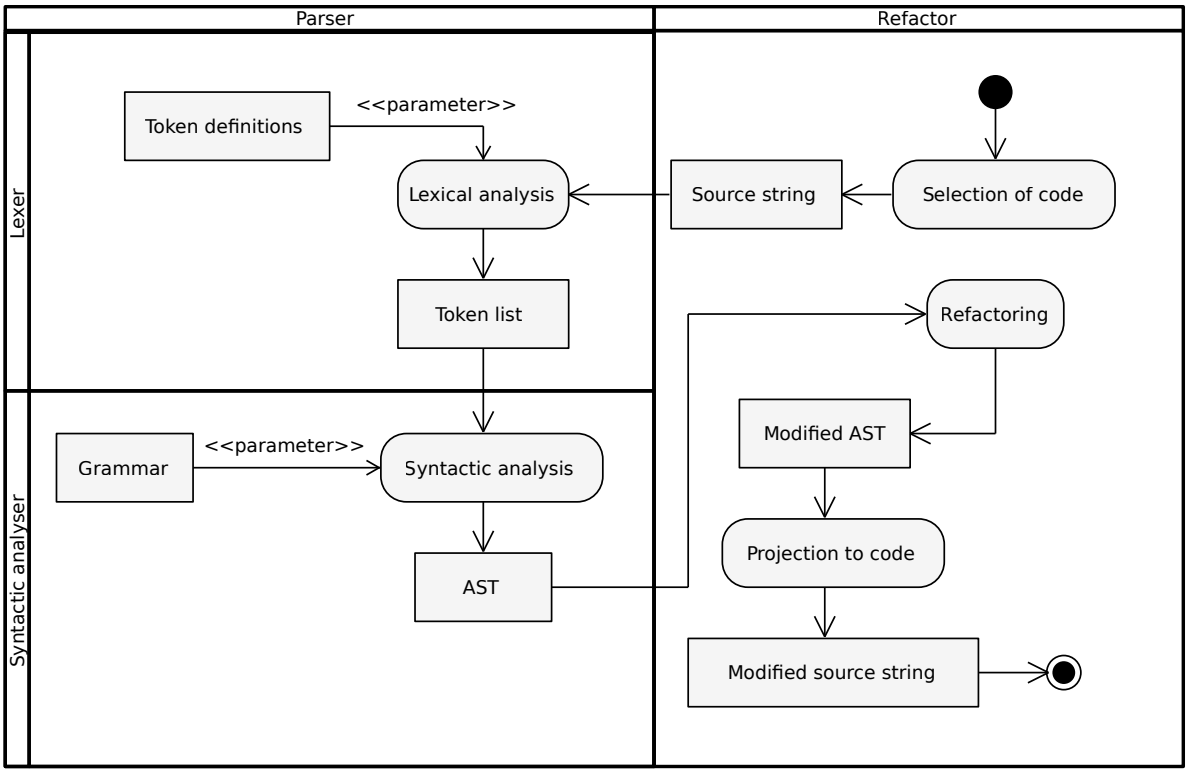


Figure 1: Activity diagram of the refactoring tool

Terminals are elementary symbols (keywords and literals) of the analysed language; they are represented by tokens and often marked by lowercase letters.

Nonterminals , by convention marked by capital letters, are sometimes aptly called *syntactic variables*, suggesting that nonterminal is a variable that substitutes other nonterminals and terminals; this "substitution" is performed via production rules.

Production rules (rules) are described in the form of $\alpha \to \beta$ which means that left-hand side string α is going to be replaced by right-hand side string β. The form of rules differs based on the type of the grammar; four types are recognized in so-called *Chomsky classification*. They are listed together with corresponding form of rule in Table 3 where symbol $*$ is a *Kleene star* (zero or more occurences) and symbol | is an acronym for "or".

Table 3: Chomsky classification of grammars

Type	*Grammar name*	*Rule form*		
0	Unrestricted	any form		
1	Context-sensitive	$\alpha_1 A \alpha_2 \to \alpha_1 \beta \alpha_2$	$A \in N$	$\alpha_1, \alpha_2, \beta \in (N \cup T)^*$
2	Context-free	$A \to \beta$	$A \in N$	$\beta \in (N \cup T)^*$
3	Regular	$A \to aB \vert a$	$A, B \in N$	$a \in T$

Grammar is a language generator; the language L that is generated by the grammar G is defined as (2) where symbol + is a *Kleene plus* (one or more occurences) and $S \Rightarrow_G^+ \omega$ means that after one or more subsequent *derivations* [9] of the starting symbol S a string ω will be produced.

$$L(G) = \{\omega \in T^* : S \Rightarrow_G^+ \omega\} \tag{2}$$

2.3 Refactor

When the syntax tree is constructed, it can be used in the refactor which has to be able to search in this tree. Moreover, since the refactoring process changes the original code, a mechanism that rewrites AST and projects it back to the source code has to be present in the refactor. Last but not least, the tool has to be provided with user interface which will allow to choose a refactoring method and to select a portion of the code on which the method will be applied. An interesting way of searching through the XML representation of AST via XQuery language as well as the rewriting of AST was introduced in [13].

3 Parsing C++ language

3.1 Grammar of C++ language

As mentioned in [4] C++ is the context-sensitive language; the following code will be used as example.

```
void function(int b) {
  (a) (b);
}
```

If literal `a` was previously declared as a type then expression `(a)(b)` would be recognized as a cast expression. On the other hand, if `a` was declared as a function then same expression would be recognized as a function call. Other ambiguities are for example: `A*B;` (dynamic type declaration vs multiplication) or `vector<vector<T>> matrix;` (templated type declaration[1] vs comparison).

In contrast to other context-sensitive languages, in the case of C++, it can sometimes be very difficult to determine the right context [4]. Let us consider the following example in which the context depends on template instantiation.

```
template <unsigned N>
struct A : A<N-2> {
};

template <>
struct A<0> {
```

[1]Since the introduction of C++11 standard, empty space is no longer necessary to separate individual characters within >>.

```
  enum { a };
};

template <>
struct A<1> {
  typedef int a;
};

void function() {
  int x(A<42>::a);
  int y(A<41>::a);
}
```

As can be seen in the above code `x` and `y` have the different meaning with regard to whether the template parameter `N` is even or odd; `x` is a variable initialized with a value of the enumeration literal `a`, while `y` is a nested function declaration.

Because there are a number of tools for lexical and syntactic analysis that would be pointless to re-create, it was decided to use an existing analysis tool.

3.2 Tools for lexical and syntactic analysis of C++

A search for tools capable of lexical and syntactic analysis was performed. The most interesting alternatives were examined in order to choose the most suitable one for subsequent analysis of C++ language.

3.2.1 Considered alternatives

GNU G++ compiler, which source code is available to public, contains highly optimized lexical and syntax analysers of C++ language. Moreover, if G++ were utilized for parsing of C++, it would guarantee that the top level refactoring tool would meet the contemporary C++ standards. On the other hand G++ source code is little structured and badly readable. However, in 2004 a team from the University of Sannio conducted research [2] about extracting syntax tree from GCC compiler which resulted into creation of tool *XOGastan*; unfortunately this tool is no longer being developed.

ANTLR (acronym of Another Tool for Language Recognition) [11] is one of the tools that allow to generate source code of the lexical and syntactic analyser based on the provided grammar and token specification. Furthermore, it allows to choose the language in which will be source codes of analysers generated. Last but not least, ANTLR is used by a large group of users, so that many user-defined input grammars, including C language grammar, are available. The disadvantage of ANTLR is that the generation is one-way process and each change in grammar requires a new generation. Moreover, the generated code is not too easy to read. These two facts make it difficult for the programmer to manually intervene in the parser code.

Eclipse CDT is plugin for *Eclipse*, the widely known Java IDE, representing an environment for the C++ language with built-in code-insight or simple refactor. Part of CDT module is C++ parser with object representation of AST; both the Parser and the AST structure are written in Java and used by CDT internally.

Boost Spirit is a part of *Boost* a large extension library for C++ language. Spirit [6] comprises both lexical and syntactic analyser. The grammar rules can be described in the modified[2] *Extended Backus Normal Form* (EBNF) [12] and embedded to the syntax analyser directly in C++, so they can mix freely with other C++ code; due to this fact the programmer is allowed to write a clear, easily changeable and immediately executable parser without the need to constantly re-generate the badly readable code. Character input is not limited to 8-bit ASCII, but 16-bit and 32-bit characters are supported such as Unicode. Spirit also provides macros that allow to turn on a detailed debugging of the syntax analysis process.

3.2.2 Chosen alternative

The Boost Spirit library was chosen for further implementation of the parser. The main reason was that the parser together with rules can be written directly in C++, so both can be constantly improved over the time without delays with the re-generation of the parser. Moreover, a library that provides automated unit tests may be included to the parser project. Regardless of fact that Spirit allows to describe rules in approximate EBNF form, in which only context-free rules can be formulated [12], it allows to inject semantic actions to the syntactic analyser via C++ function pointers or function objects.

4 Development progress

According to the decomposition in the section 2 the development process was divided into three major phases. First, the lexer and syntax analyser are going to be developed as a single module. Next, the basic refactor module is planned to be created. During the third step, both modules are going to be continuously improved in order to implement the largest possible number of refactoring methods.

4.1 Parser construction

The parser is developed in the C++ language. CMake [7] is utilized for automated generation of platform dependent makefiles.

So far two C++ structures were created `CppLexer` and `CppGrammar`; CppLexer is the instance of `lexer` structure which contains definitions of tokens and information about order in which they should be matched; `CppGrammar` is the instance of `grammar` structure and contains definitions of rules. Rules were divided into six groups according to individual C++ statements: block, selection, iteration, jump, declaration, expression. Currently

[2]The symbols has been changed in order to be expressed by C++ operators.

rules for block statement and several types of expression (assignment, unary, postfix, primary) are created and remaining expression rules are developed. By this time semantic actions are not utilized and the code is parsed without the context disambiguation.

4.2 Parser testing

As the grammar evolves with the growing number of rules during the development process, the parser has to be constantly tested to make sure that newly developed rules are valid and old ones has not been negatively affected by the current changes in the grammar. With regard to these reasons, it was decided that unit tests will be utilized. For unit testing QTest framework from Qt library [14] was chosen and included to the project; a separate module *test* was also created in the project. This module contains expression strings and source files that are iteratively passed to the parser in order to test it.

5 Conclusion

In this paper basic refactoring techniques were summarized together with corresponding *code smells*. Next, the refactoring tool was decomposed into three parts: lexer, syntax analaser and refactor. Moreover, the process of syntax analysis was discussed in more detail with respect to difficulties with analysing of the C++ language presented in the next section. Subsequently, the paper devoted to choosing a suitable tool for syntactic and lexical analysis of C++; individual alternatives were described; and finally, the Boost Spirit library was chosen. In the last section the current state of the work on the refactoring tool was described. Further work will focus primarily on formulation of rules for declaration statements and disambiguation of C++ context via semantic actions.

References

[1] A. V. Aho, M. S. Lam, R. Sethi and J. D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley, (2006), 2nd edition.

[2] G. Antoniol, M. Di Penta, G. Masone and U. Villano. *Compiler Hacking for Source Code Analysis*. In 'Software Quality Journal', Springer, volume 12, issue 4, (2004), 383–406.

[3] N. Chomsky. *Three Models for the Description of Language*. In 'IRE Transactions on Information Theory', volume 2, issue 3, (1956), 113–124.

[4] V. David. *Language Constructs for C++-like languages*. Dissertation thesis, University of Bergen, (2009).

[5] M. Fowler. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley, (2006), 2nd edition.

[6] J. Guzman and D. Nuffer. *The Spirit Library: Inline Parsing in C++*. In 'C/C++ Users Journal', volume 21, issue 9, (2003).

[7] B. Hoffman, K. Martin. *Mastering CMake*. Kitware, (2010).

[8] ISO/IEC 14882:2011. *Information technology – Programming languages – C++*. ISO, (28 September 2012).

[9] T. Jiang, M. Li, B. Ravikumar and K. W. Regan. *Formal Grammars and Languages*. In 'Algorithms and Theory of Computation Handbook', M. J. Atallah (ed.), CRC Press, (1998).

[10] M. V. Mäntylä and C. Lassenius. *Subjective Evaluation of Software Evolvability Using Code Smells: An Empirical Study*. In 'Journal of Empirical Software Engineering', Springer, volume 11, issue 3, (2006), 395–431.

[11] T. Parr. *The Definitive ANTLR Reference*. Pragmatic Bookshelf, (2007).

[12] R. E. Pattis. *EBNF: A Notation to Describe Syntax*. http://www.cs.cmu.edu/ pattis/misc/ebnf.pdf, (27 September 2012).

[13] J. Smolka. *Refactoring tool for Java programs*. Master's thesis, Czech Technical University, (2010).

[14] M. Summerfield. *Advanced Qt Programming*. Addison-Wesley, (2010).

Využití lambda kalkulu v metodě BORM*

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 v Praze

Abstract. The article is about the method of data modeling BORM, which is very popular among developers, analysts and consultants due its complexity and understandability. Attention is given to the lambda-calculus, which this method, unlike other methods of data modeling supports by built in an interpreter of language C.C. These enable to solve many problems of simulations, transformation of models and others, which allow a better understanding of reality.

Keywords: data modelling, BORM method, lambda-calculus

Abstrakt. Příspěvek je věnován metodě datového modelování BORM, která je pro svoji komplexnost a zároveň snadnou srozumitelnost velmi oblíbena mezi vývojáři, analytiky i konzultanty. Pozornost je věnována zejména lambda-kalkulu, který tato metoda na rozdíl od ostatních metod datového modelování podporuje prostřednictím zabudovaného interpretu jazyka C.C. Díky tomu je možno vyřešit řadu problémů simulace, transformace modelů a dalších, což umožňuje lépe pochopit realitu.

Klíčová slova: datové modelování, metoda BORM, lambda-kalkul

1 Úvod

Existuje celá řada metod datového modelování (BPMN, EPC, IDEF, UML...) které umožňují uživatelům a vývojářům co nejlépe porozumět modelovanému světu. Každá z těchto metod má svoje výhody i nevýhody. V tomto příspěvku se budu zabývat metodou BORM, která je pro svoji komplexnost velmi oblíbena mezi analytiky, konzultanty i vývojáři. Pozornost budu věnovat zejména lambda-kalkulu, který tato metoda, resp. CRAFT.CASE, který je nejčastěji používaným softwarovým CASE nástrojem pro tuto metodu, na rozdíl od většiny ostatních metod podporuje.

2 Metoda BORM

Metoda BORM (Business Object Relation Modeling) je vyvíjena od roku 1993. Od počátku je orientována na podporu tvorby objektově orientovaných softwarových systémů založených na čistě objektově orientovaných programovacích jazycích a vývojových prostředích (například prostředí Smalltalku - VisualWorks, VisualWave, VisualAge, ...) a objektových databázích (Gemstone, Artbase, ...). Metoda je podporována i CRAFT.CASE

*Tato práce byla podpořena grantem SGS2012

nástrojem vyvíjeným firmou e-FRACTAL, s.r.o. CRAFT.CASE implementuje i funkční programovací jazyk C.C, který je založený na lambda kalkulu.

BORM pokrývá všechny fáze vývoje softwaru. Základní odlišnosti metody od ostatních podle V. Merunky [3] jsou:

- 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.
- 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.
- Nevyžaduje oddělování od sebe statických a dynamických pohledů na systém do různých typů diagramů s rozdílnou notací, je možno je v jednotlivých diagramech kombinovat.

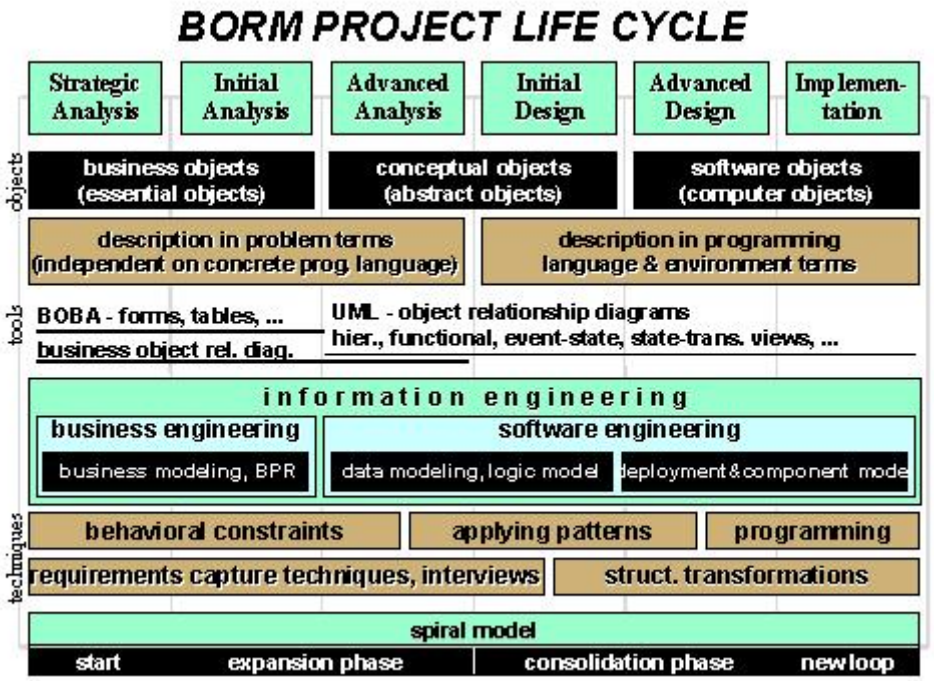
Fáze životního cyklu podle BORM [3, 5]

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.

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. Pracuje rovněž s hierarchií objektů (polymorfismus, is-a vztah, závislost objektů).

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 [4].



Obrázek 1: Životní cyklus projektu podle metody BORM podle V. Merunky [3]

3 Lambda kalkul

Lambda kalkul (označovaný také jako λ-kalkul) je formální systém a výpočetní model používaný v teoretické informatice a matematice pro studium funkcí a rekurze. Jeho autory jsou Alonzo Church a Stephen Cole Kleene. Lambda kalkul je teoretickým základem funkcionálního programování a příslušných programovacích jazyků, obzvláště Lispu. Analyzuje funkce nikoli z hlediska původního matematického smyslu zobrazení z množiny do množiny, ale jako metodu výpočtu [6].

Základem syntaxe λ-kalkulu je *výraz*. Ukážeme si pro srovnání jednoduchý výraz zapsaný v λ-kalkulu, Smalltalku a Javě:

λ-kalkul	Smalltalk	Java
$(\lambda x \mid x + 2)$	[:x | x + 2]	x += 2
$(\lambda x \mid x + 2) \triangleleft :12$	[:x | x + 2] value: 12	x = 12; x += 2

Tabulka 1: Srovnání zápisu výrazu v λ-kalkulu, Smalltalku a javě

Každý λ-výraz je složen ze dvou částí, oddělených nějakým znakem, např. |. První část se nazývá hlavička a obsahuje seznam všech proměnných použitých ve výrazu uvozený znakem λ, druhá část je vlastní tělo, které se zapisuje i chová stejně jako běžný matematický zápis.

3.1 α-konverze

se používá tam, kde by při skládání více λ-výrazů mohlo dojít k záměně stejně pojmenovaných proměnných a "λ-počítač" ji vykonává automaticky:
(λx | x + 2)◁:(λx λy | x + y)
⇒ (λx = (λx λy | x + y) | x + 2)
⇒ ???

V tomto případě je nutno v jednom výrazu proměnnou přejmenovat, např.:
(λx | x + 2) ⇒ (λz | z + 2)
a nyní už lze bez problémů aplikovat druhý výraz:
(λz | z + 2)◁:(λx λy | x + y)
⇒ (λz = (λx λy | x + y) | z + 2)
⇒ ((λx λy | x + y) + 2).

3.2 η-redukce

Mějme výraz (λx | (*výraz*◁: x)), kde *výraz* označuje libovolný jiný λ-výraz. Při aplikaci jakékoli hodnoty do tohoto výrazu postupně dostaneme:
(λx | (*výraz*◁: x))◁:*hodnota*
⇒ (λx = *hodnota* | (*výraz*◁: x))
⇒ *výraz*◁:*hodnota*.
Můžeme tedy prohlásit, že (λx | (*výraz*◁: x)) = *výraz*. Toto zjednodušení se nazývá η-redukce.

3.3 Objektově orientovaný přístup v λ-kalkulu

Základem Objektově orientovaného přístupu (OOP) je objekt. Objekt představuje nějakou část reálného světa, obvykle je možno ji v popisu modelu identifikovat jako podstatné jméno (např. auto, řidič...).

3.3.1 Atributy a metody objektu

Objekty mají *hodnoty* (atributy) např. značka auta, objem, výkon... a dokáží reagovat na požadavky, které jsou jim zasílány pomocí *zpráv*.

Data, která objekt uchovává můžeme např. pro objekt *auto* zapsat:
Δ(*auto* = [*model*: Renault Clio III, *objem*: 1598, *výkon*: 65, *SPZ*: 1AC2889, ...]

Při zaslání zprávy *auto stáří* dostaneme číselnou hodnotu udávající stáří automobilu. Tato hodnota nebude v systému uložena, protože by bylo nutné zajistit pravidelné aktualizace, ale bude vypočítávána pomocí metody. Metoda se skládá ze dvou částí, které označujeme <*hlavička, tělo*>. Metoda pro výpočet stáří automobilu může být například <*stáří,* (*dnešní datum* - σ*datum výroby*) / 365.2422>.
Symbolem σ označujeme objekt, kterém tato metoda patří.

Funkce *Meth* označuje možinu všech metod daného objektu. Fakt, že metoda stáří patří objektu auto můžeme zapsat jako <*stáří,* (*dnešní datum* - σ*datum výroby*) / 365.2422> ϵ *Meth(auto).*

λ-kalkul	Smalltalk	Java
auto$\lhd$*objem*	*auto objem*	*auto.objem()*
auto$\lhd$*objem:1598*	*auto objem: 1598*	*auto.objem(1598)*
auto$\lhd$*objem* $\Rightarrow$ *1598*	*auto objem*	*auto.objem()*

Tabulka 2: Srovnání zápisu zaslání zprávy objektu auto λ-kalkulu, Smalltalku a javě

3.3.2 Protokol objektu

Protokolem objektu rozumíme množinu všech zpráv, které je možné příslušnému objektu poslat. Pro náš objekt auta to je například:
$\Pi(auto) = [model,\ objem,\ výkon,\ SPZ,\ stáří,\ VIN]$.

3.3.3 Dědičnost

Funkce *super* označuje nadtřídu k dané třídě. Např. fakt. že nadtřídou třídy osobní auto je třída motorové vozidlo můžeme zapsat jako $super(OsobniAuto) = MotoroveVozidlo$.
Symbolem Ω označme množinu všech objektů v systému.

Dědění nyní můžeme zapsat jako
$\forall\ a,\ b\ \epsilon\ \Omega\ a = super(b) \rightarrow Meth(a) \subseteq Meth(b)$, resp.
$\forall\ a,\ b\ \epsilon\ \Omega\ a = super(b) \rightarrow \Pi(a) \subseteq \Pi(b)$.

3.3.4 Polymorfismus

Polymorfismus v OOP znamená, že stejná zpráva může vyvolávat různé operace, které se z pohledu toho, kdo zprávu poslal jeví jako stejné, i když samy o sobě stejné nejsou. Například u vozidel máme metodu SPZ, která vrací SPZ vozidla. U přípojných vozidel bez SPZ pak vrací SPZ vozidla, za které je přípojné vozidlo připojeno. Fakt, že přípojné vozidlo bez SPZ patří k nějakému vozidlu může být uložen v objektu přípojné vozidlo s VIN 239KI4959EBK398O jako $[SPZtahač:\ 1AC2889] \subset \Delta$(239KI4959EBK398O).

Zápis polymorfní metody SPZ v λ-kalkulu vypadá
<*SPZ,* σ*SPZtahač*$\lhd$*SPZ*> ϵ *Meth*(239KI4959EBK398O).

3.3.5 Kolekce objektů

V programovacích jazycích jako je Java nebo Smalltalk je více než 100 druhů kolekcí. Základní jsou 3 druhy:

- Set - neuspořádané prvky bez duplicit,
- Bag - neuspořádané prvky, mohou obsahovat duplicitní hodnoty
- List - uspořádané prvky, mohou být duplicitní hodnoty.

Typ kolekce lze pomocí konverzí měnit. Konverze zapisujeme jako

- *Set(A)* - přemění kolekci A na Set
- *Bag(B)* - přemění kolekci A na Bag
- *List(A)* - přemění kolekci A na List
- *List(A, Λ)* - přemění kolekci A na List s prvky setříděnými podle hodnoty dané λ-výrazem Λ.

Setřídění seznamu auta podle objemu zapíšeme pomocí λ-výrazu jako
$List(Auta,\ (\lambda x \mid x \triangleleft objem))$.

4 Λ-kalkul v metodě BORM

Součástí nejrozšířenějšího softwarového nástroje CRAFT.CASE pro metodu BORM je programovací jazyk C.C, který implementuje λ-kalkul. CRAFT.CASE provádí transformace modelů prostřednictvím interpreteru C.C, který umožňuje nejen tvorbu skriptů a implementaci pravidel. C.C interpreter umožňuje navíc vykonávat všechny operace nad modelem včetně simulace, refactoringu, tvorby nových diagramů...) [2].

V C.C můžeme funkce zapisovat jako λ-výrazy ve složených závorkách - $(\lambda x\ \lambda y \mid x^2 + y)$ v C.C zapíšeme jako {:X, :Y | X^2 + Y}. Argumenty můžeme funkci předat v kulatých závorkách: {:X, :Y | X^2 + Y}(5, 10).

Selekci a projekci v C.C můžeme pomocí λ-výrazů zapsat jako
[100, 200, 300, 400, 500] // {:X | X > 200} = [300, 400, 500].
[100, 200, 300, 400, 500] ≫ {:X | X + 1} = [101, 201, 301, 401, 501].

Pomocí λ-výrazů můžeme v C.C zapisovat i těla cyklů:
for [1, 2, 3, 4, 5] do {:X | console:print-nl(X)}.

Obdobně je možné zapsat i rekurzivní funkce, např. funce pro výpočet faktoriálu může vypadat takto:

```
| Faktorial |
Faktorial :=   {:X |   if X = 0
                       then {1}
                       else {X * Faktorial(X - 1)}}.
```

Výběr instančních proměnných, které mají být přesunuty z původní třídy do nové:
JmenaAtributu := dialog:choose-multiple
("Vyberte atributy pro přesun...", OldClass → 'Composition' ≫ {:X | X[name]}).

5 Závěr

Λ-kalkul, který je součástí jazyka C.C v softwarovém nástroji CRAFT.CASE umožňuje uživateli vytvářet vlastní funkce umožňující práci s daty a funkcemi systému. Díky tomu je možno vyřešit např. mnoho problémů simulace business procesů, transformace modelů, ověřování modelů apod., což umožňuje lépe pochopit modelovanou realitu. Tyto možnosti budou předmětem dalšího výzkumu.

Literatura

[1] V. Merunka. *Datové modelování.* Praha: Alfa Publishing (2006), 9 - 23. http://www.google.cz/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCcQFjAA&url=http%3A%2F%2Fhome.czu.cz%2Fwebdav.php%3Fseo%3Dlinhart%2Fkestazeni%2F%26file%3D%2Fmerunka6.pdf&ei=iWlfUOvhOIaxtAbH2oDYCA&usg=AFQjCNH0470nuyJKDw_T_ZUT2cnOYO3cQw&sig2=EvmcbV9VbVooGJRAZsMcdQ

[2] V. Merunka. *Programming language C.C.* Automatizace **51** (2008), 562 - 565.

[3] 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.

[4] A. Rývová *Datové modelování.* konference Doktorandské dny 2011, FJFI ČVUT Praha (2011), 193 - 202.

[5] 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.

[6] *Lambda kalkul.*
Wikipedie, Otevřená encyklopedie, http://cs.wikipedia.org/wiki/Lambda_kalkul.

Model of Bacterial Colony Evolution in the Presence of Another Bacterial Body*

Josef Smolka

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

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

Abstract. This paper presents a problem of bacterial colony evolution in the neighborhood of another evolving bacterial body of the same species. The research shows that the colony shape and pattern are influenced in a way that point out at advanced communication capabilities. A reaction-diffusion model of bacterial colony interactions is introduced. Example of a model implementation in a newly created domain-specific language is given and simulation results are presented.

Keywords: bacterial colony simulation, object-oriented database, Java, Groovy

Abstrakt. Příspěvek představuje problematiku vývoje bakteriální kolonie v blízkosti jiného vyvíjejícího se bakteriálního tělesa stejného druhu. Výzkum ukazuje, že způsob, jakým je ovlivněn tvar a povrchový vzor kolonie, je důsledkem pokročilých komunikačních schopností, které člověk u takto jednoduchých organismů nečekal. Článek představuje reakčně difúzní model vzájemných interakcí jednotlivých kolonií a jeho implementaci v nově vytvořeném doménově specifickém jazyce.

Klíčová slova: simulace bakteriální kolonie, objektová databáze, Java, Groovy

1 Introduction

Bacterial colony of gram-negative, facultative anaerobic, rod-shaped bacteria *Serratia rubidaea* growing on an agar plate shows a variety of complex patterns of color and structure. Patterns are influenced by both colony-autonomous developmental and regulatory processes and by environmental influences [1]. Modeling of relationship between external influences and colony evolution is quite a common task in the field of microbiology, see Zwietering [13, 12], Wijtzes [11] and Houtsma [2].

This paper presents a reaction-diffusion model that describes a distribution of two substances (bacteria and signal – will be explained later). The distribution is influenced by two related processes: chemical reactions, which express the transformation of substances into each other, and diffusion which induces the substances to spread out over the agar plate [4]. The colony is then modeled as heterogeneous body (heterogeneous within the meaning of local concentrations of bacteria).

*Creation of the paper was supported by the grant SGS 11/167.

1.1 Modeled Experiments

Several phenotypes of wild-type strain of *Serratia rubidaea* are recognized in [1]. Experiments modeled in this paper were carried out with R (red) phenotype which forms circular red glossy colonies, see Fig. 1. Despite the fact, that colonies in Fig. 1 are

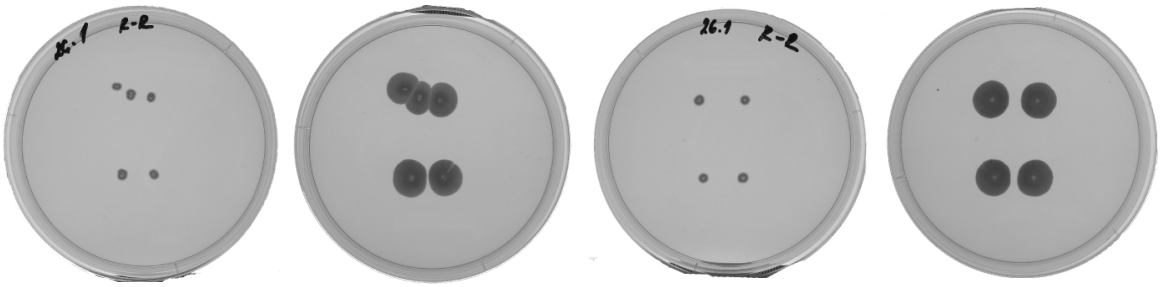


Figure 1: A colony formation example of R phenotype of *Serratia rubidaea* bacterium. Two sets of images – the first and the fifth day is captured in each set.

monoclonal, some kind of behaviour that resembles some sort of communication is observed. How bacteria, in some cases, recognize that neighborhood identical colony is not the same bacterial body? This behaviour could be explained by unknown chemical substance, so-called signal, that is diffusing in the substrate and in the air. An excess of the concentration level has impact on metabolism of bacteria. Consider the following

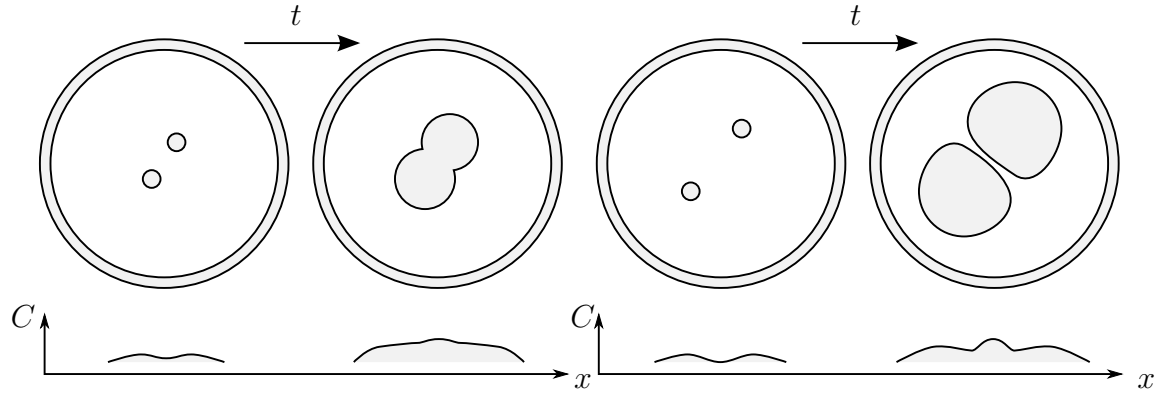


Figure 2: The first test scenario (left) – two foundations of the future colonies are placed on the agar near each other. Observations: the two colonies blend together and both are considerable smaller. The second test scenario (right) – two foundations of the future colonies are placed at a greater distance then in the first case. Observations: clearly visible border is formed between colonies. Colonies have shape of letter D.

test scenario, two foundations of the future colonies are placed on the agar near each other (see Fig. 2 left). After some time, the two original independent colonies blend together. This can be explained by the concentration of the signal substance that did not manage to reach the critical level (due to the distance). On the other hand (see Fig. 2 right), when there is a greater distance between the two colony foundations, clearly visible border between two colonies is formed. This can be explained by the exceeding concentration of the signal that inhibits bacteria division ability.

2 Diffusion-reaction Model

If the colony is addressed as a multicellular body, the model may cover only the processes observable from a macro view and does not have to deal with the processes on a micro scale level as in the already presented individual-based model [8]. Changes in the assumptions are:

- Only the local concentrations of substances (bacteria, signal) are modeled.
- As colony is evolving on rich medium, the influence of nutrients supply can be ommited.

A core of the model is based on a system of two reaction diffusion equations [10], where the Eq. 1 expresses the diffusion and division of bacteria and the Eq. 2 describes diffusion and production of the signal substance.

$$\frac{\partial B}{\partial t} = D_1 \left(\frac{\partial^2 B}{\partial x^2} + \frac{\partial^2 B}{\partial y^2} \right) + R_B(B, G) \tag{1}$$

$$\frac{\partial G}{\partial t} = D_2 \left(\frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2} \right) + R_G(B) \tag{2}$$

This system of partial differential equations can be approximated by finite difference method and simplified to be more suitable for the simulation.

$$\Delta B_{S_{ij}} = c_1 \left(\sum_{s \in O(S_{ij})} \bar{B}_s - 4\bar{B}_{S_{ij}} \right) + R_B(.) \tag{3}$$

$$\Delta G_{S_{ij}} = c_2 \left(\sum_{s \in O(S_{ij})} G_s - 4G_{S_{ij}} \right) + R_G(.) \tag{4}$$

Coefficients c_i are rates of diffusion and are equal to $\frac{D_i \Delta t}{(\Delta x)^2}$. Original equation 1 is transformed into delayed variation, where the $\bar{B}_{S_{ij}}$ is increment of bacteria concentration in the previous generation. To complete the model formulation, the function of bacteria growth $R_B(B, G)$ and function of signal production $R_G(B)$ have to be defined.

2.1 Bacteria Growth

Model of bacteria growth is based on a curve of exponential and logistic growth [13, 9]. The equation 5 characterizes increase of bacteria concentration in the location S_{ij} in dependence on bacteria and signal concentration in neighborhood.

$$R_B(B, G) = r_1 \frac{\Delta t}{T_c} (1 - g) b_1 b_2 B_{S_{ij}} \tag{5}$$

$$g = \frac{G(S_{ij}) + G(O_3(S_{ij})}{G_{\max}(S_{ij}) + G_{\max}(O_3(S_{ij})}$$

$$b_1 = 1 - \frac{B(S_{ij})}{B_{\max}(S_{ij})}$$

$$b_2 = \frac{B(S_{ij}) + B(O(S_{ij})}{B_{\min}(S_{ij}) + B_{\min}(O(S_{ij})}$$

The r_1 is random variable from $N(c_b, \sigma_b^2)$, where c_b is calibration coefficient of bacteria reproduction and T_c is a length of cell-division cycle. Member $(1-g)$ represents ratio of inactive bacteria due to the exceeding signal concentration in the neighborhood. The b_1 is a reduction in growth rate due to the high concentration in S_{ij} and the b_2 is a reduction in growth rate due to the low concentration in $O(S_{ij})$. Coefficient b_2 is significant for the colony shape.

2.2 Signal Production

The signal production model just exhibits dependency between bacteria concentration in location S_{ij} and signal generation. This dependency is described by Eq. 6, where r_2 is random variable from the normal distribution with mean equal to signal production of one bacteria in time Δt.

$$R_G(B) = r_2 g_2 B_{S_{ij}}^{g_1} \tag{6}$$

3 Model Implementation and Simulation

For the purpose of simulation, specialized software written in the Java programming language and with the help of the SWT library was created. A core of the simulator consists of simple in-memory object database system with optional file-based persistence for storing the simulation state. The database contains an implementation of quadtree which serves as spatial grid index for fast grid locations access. In order not to delay the simulation during the evaluation with allocation of new objects, system of object pools (as in Object Pool design pattern) with sufficient initial capacity was proposed [6]. To work with the object database specialized query API based on Groovy is exposed.

3.1 Query Language Introduction

The query language, through the object database is accessed, is based on the Groovy language. For a better understanding of presented statements, it is necessary to embrace class diagram in Fig. 3.

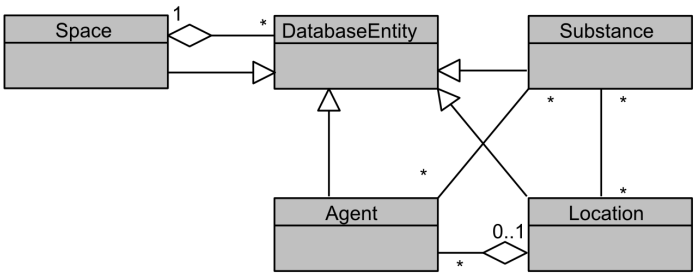


Figure 3: UML class diagram of simulation model core classes.

All entities stored in the database are derived from the abstract class *DatabaseEntity* which implements basic interface *DenotedEntity*, so every entity in the system can be addressed by its name. Every instance of any database entity is associated with instance

of so-called *Space*; it is something like database schema in classic relational database. The grid is composed of set of locations (instances of class *Location*). Every location can contain set of substances (support for diffusion-reaction models) and set of agents (support for individual-based models).

Let us take a look at query example [7] – implementation of the diffusion part from Eq. 3.

```
def S = { L, w -> return d.s.loc(L.x+w?.x?:0, L.y+w?.y?:0)?.subst(p.signal)
  ?: 0.0}

def SD = { L ->
  def SI = SP * (p.moore.sum {w -> S(L, w)} - 4 * S(L, null))
  L.add(p.signal, SI)
}

d.s.loc.each{L -> SD(L)}
```

The query itself is just an application of defined closure (lambda expression in fact) SD to all the locations in the grid. The first line is a closure definition for getting signal concentration in the specified direction. Respective location in a grid is obtained with a help of spatial index, which is accessed by the *loc* method of the current space s in the database d. The location is addressed by x and y coordinates, which are evaluated as a sum of position of the current location L and a specified way w, which can be null. Note the use of *Safe Navigator Operator ?.* to avoid *NullPointerException* when accessing property on a null object and the *Elvis Operator ?:* to shorten the classic ternary operator if one of the results is null. The value of the concentration is returned by the *subst* method for the specified substance instance p.signal. The second closure SD deals with the signal concentration in the actual location L. The variable SI is the signal fluctuation in the location [7].

4 Results

State of the simulation of diffusion-reaction model, stored in the object database, was rendered using simulator graphical output and saved as an image in predefined points of time (1st day, 3rd day, 5th day). Those images were then compared to experimental data in the form of camera pictures. The Fig. 5 compares experimental data and the model output. Results are complemented by Fig. 4 which shows the 7th day of colony evolution in the second scenario to show the predicted behaviour – colonies has shape of letter D.

The presented method of model implementation based on lambda expressions serves as an intermediate language for the future development of end-user declarative language for bacteria simulation description.

5 Discussion

The previously presented individual-based model [8], although it seemed like a natural choice at first, proved to be very difficult to estimate. The reasons might be substantial

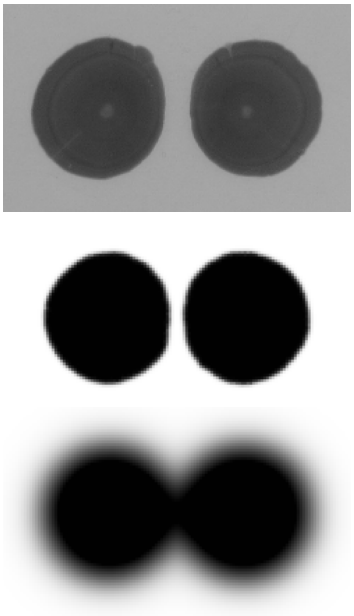


Figure 4: Comparison of an experiment and the reaction-diffusion model. The simulation was performed with the following parameters' setting: $\Delta t = 20min$, $\Delta x = 1mm$, $c_1 = 0.18$, $c_2 = 0.16$, $r_1 \in N(0.35, 0.15)$, $T_c = 40min$, $r_1 \in N(1, 0.3)$, $g_1 = 0.11$ and $g_2 = 0.25$. The seventh day of colonies evolution.

greater number of parameters and also the level of detail that must be modeled. The gap between the level of detail of experimental data and modeled processes then results in difficulty verifiable model. On the other hand, reaction-diffusion model proved to be the appropriate selection. The model output shows characteristics observable during the experiments.

6 Conclusion

The paper introduced the model of bacteria interactions through the signal substance. The presented reaction-diffusion model is based on the statement that a colony can be viewed as multicellular body. Results of reaction-diffusion model simulation were compared to experimental data and presented. The future work consists of more research in individual-based model and development of declarative language for simulation description as well as in extension of the models to cover more complicated situations.

References

[1] Cepl, J., Patkova, I., Blahuskova, A., Cvrckova, F., Markos, A.: Patterning of mutually interacting bacterial bodies: close contacts and airborne signals. In *BMC Microbiology*, 10:13, 2010

[2] Houtsma, P. C., Kusters, B. J. M., De Wit, J.C., Rombouts, F. M., Zwietering, M. H.: Modelling growth rates of Listeria innocua as a function of lactate concentration. In *International Journal of Food Microbiology*, Vol. 24, No. 1–2, 1994, pp. 113–123

[3] Melke, P., Sahlin, P., Levchenko, A., Jansson, H.: A Cell-Based Model for Quorum Sensing in Heterogeneous Bacterial Colonies. In *PLoS Comput Biol*, 6(6): e1000819, 2010

[4] Mimura, M., Sakaguchi, H., Matsushita, M.: Reaction-diffusion modelling of bacterial colony patterns. In *Physica A: Statistical Mechanics and its Applications*. Vol. 282, No. 1–2, 2000, pp. 283-303

[5] Prats, C., Ferrer, J., Lopez, D., Giro, A., Vives-rego, J.: On the evolution of cell size distribution during bacterial growth cycle: Experimental observations and individual-based model simulations. In *Journal of Microbiology*. Vol. 4, No. 5, 2010, pp. 400-407

[6] Smolka, J.: Using Java and Groovy in Simulation of Mutually Interacting Bacterial Bodies. In *Objects 2011*. Zilina: University of Zilina, Faculty of Management Science and Informatics, 2011, pp. 24-31

[7] Smolka, J.: Groovy as a Swiss Knife – from Enterprise to Science. In *38th Software Development 2012*. Ostrava: VSB - Technicka univerzita Ostrava, 2012, pp. 114-120

[8] Smolka, J: Simulace interakce bakterialnich kolonii. In: *Doktorandske dny 2011*. Praha: Ceska technika - nakladatelstvi CVUT, 2011, pp. 203-211

[9] Sugiura, K., Kawasaki, Y., Kinoshita, M., Murakami, A., Yoshida, H., Ishikawa, Y.: A mathematical model for microcosms: formation of the colonies and coupled oscillation in population densities of bacteria. In *Ecological Modelling*. Vol. 168, No. 1–2, 2004, pp. 173–201

[10] Walther, T., Reinsch, H., Grose, A., Ostermann, K., Deutsch, A., Bley, T.: Mathematical modeling of regulatory mechanisms in yeast colony development. In *Journal of Theoretical Biology*. Vol. 229, No. 3, 2004, pp. 327–338

[11] Wijtzes, T., De Wit, J. C., Intveld, J. H. J., Van Riet., K., Zwietering, M. H.: Modelling Bacterial Growth of Lactobacillus curvatus as a Function of Acidity and Temperature. In *Applied and Environmental Microbiology*, Vol. 61, No.7, 1995, pp. 2533–2539

[12] Zwietering, M. H., De Koos, J. T., Hasenack, B. E., De Wit, J. C., Van't Riet, K.: Modeling of Bacterial Growth as a Function of Temperature. In *Applied and Environmental Microbiology*, Vol. 57, No. 4, 1991, pp. 1094–1101

[13] Zwietering, M. H., Jongenburger, I., Rombouts, F. M., Van Riet., K.: Modeling of the Bacterial Growth Curve. In *Applied and Environmental Microbiology*, Vol. 56, No. 6, 1990, pp. 1875–1881

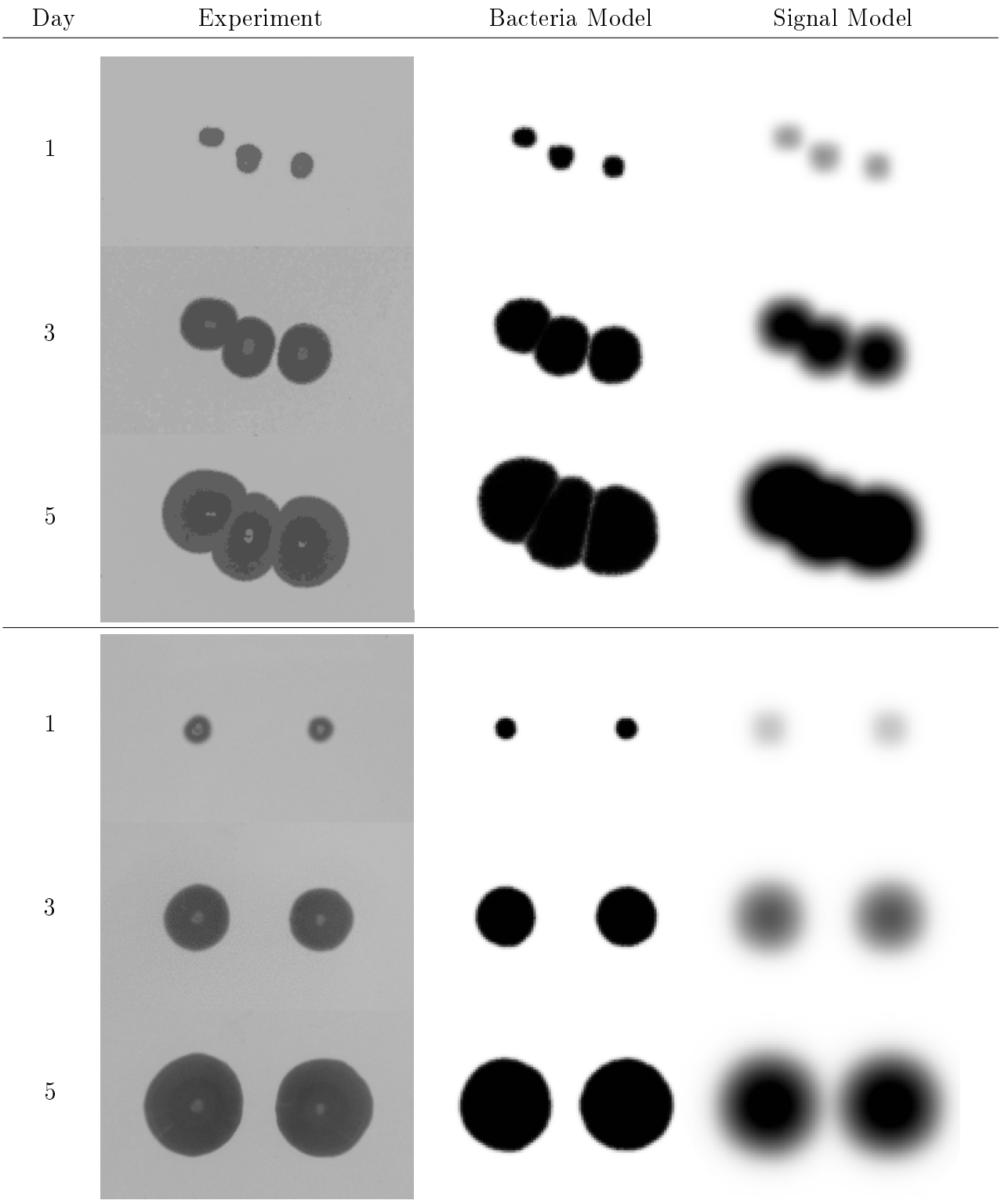


Figure 5: Comparison of an experiment and the reaction-diffusion model. The simulation was performed with the following parameters' setting: $\Delta t = 20min$, $\Delta x = 1mm$, $c_1 = 0.18$, $c_2 = 0.16$, $r_1 \in N(0.35, 0.15)$, $T_c = 40min$, $r_1 \in N(1, 0.3)$, $g_1 = 0.11$ and $g_2 = 0.25$. The top part corresponds to the first test scenario as in Fig. 2. The bottom part corresponds to the second test scenario as in Fig. 2.

Simulations in Hydrogen Fuel Cells*

Lucie Strmisková

3rd year of PGS, email: lucka.strmiskova@seznam.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: František Maršík, Institute of Thermomechanics, AS CR

Abstract. This contribution is a review on modeling techniques used for hydrogen fuel cells' research. At the beginning, the basic principles and the detailed structure of a fuel cell will be described in order to see the processes, that can occur there. Deep understanding of these processes is necessary for improving fuel cell design. Three different simulation methods are presented: methods based on continuum models, molecular dynamics and methods using quantum mechanics for describing processes on an atomic level. Some results obtained by continuum models are presented, but the highest attention will be given to the methods based on quantum mechanics.

Keywords: hydrogen fuel cells, modeling, ab initio models

Abstrakt. Tento příspěvek je přehledový článek o modelovacích technikách využívaných při studiu vodíkových palivových článků. Na začátku budou popsány základní principy fungování palivových článů a jejich detailní struktura, abychom viděli, jaké jevy v palivových článcích nastávají. Pochopení těchto jevů je nutným předpokladem pro práci na vylepšení palivových článků. Budou představeny tři různé simulační metody: metody založené na mechanice kontinua, molekulární dynamika a metody, které využívají kvantovou mechaniku pro popis jevů na atomární úrovni. Budou prezentovány některé výsledky získané pomocí mechaniky kontinua, i když největší pozornost bude věnována metodám založeným na kvantové mechanice.

Klíčová slova: vodíkové palivové články, modelování, ab initio modely

1 Introduction

Hydrogen fuel cells are considered as one of the most promising power sources, that can replace internal combustion engines in automotive industry. But their usage is not limited only to replacement of engines. They are used as power back-up systems and combined heat and power systems for households.

Apart from almost unlimited sources of hydrogen, hydrogen fuel cells have many other advantages over combustion engines. They are not limited by Carnot efficiency and they convert hydrogen and oxygen energy into electricity without combustion, so their resulting efficiency is 50 – 60%, almost double of combustion engine's efficiency.

They are much more silent than combustion engines, because fuel cell has no moving parts inside. Their silent operation make them perfect backup power in hospitals or hotels, that are placed in quiet locations.

*This work has been supported by the grant CZ.1.05/2.1.00/03.0088

There are no pollutant emissions except water, if the hydrogen is pure. We, of course, do not have to forget, that if we want to use hydrogen, we first have to generate it from some chemical component containing it. We can never produce pure hydrogen and these small impurities from the production of hydrogen can react in fuel cell and thus also produce environment pollution. More serious problem is, that exhaustion gases are produced while generating hydrogen and the amount of exhaustion gases can be sometimes higher than the amount of exhaustion gases, when using classical combustion engines. The best option how to produce hydrogen is by electrolysis, when the electricity is produced by wind or solar power plants. Unfortunately, this method is incredibly expensive at the present time.

And the last, but not one reason for using hydrogen fuel cells is, that they are modular, they can provide power over a large range, from a few watts to megawatts.

The widespread commercialization of fuel cells as sources of electrical energy is primarily limited by two factors: high cost and bad performance. In order to fight against these limitations, we need to optimize fuel cell design and introduce cheaper materials (without loss of efficiency and durability). But the design optimization and the introduction of new materials significantly depend also on the development of physical models, that reliably simulate all processes in fuel cells under realistic conditions.

Within last 20 years, high attention was given to the fuel cell modeling. Fuel cell models helped engineers to predict the behavior of fuel cell with given geometric parameters, materials and operating conditions. These models have many advantages over experimental methods. Their cost is not so high and time consuming. Moreover experiments are limited only to currently used designs. Also the environment in fuel cells is very reactive and it is difficult to measure important parameters like temperature, pressure or species concentration in the cell, so we would like to have models, that can predict these parameters.

A review of modeling techniques used in fuel cells and important results obtained in last 20 years will be presented. But first, all parts of fuel cells will be described in a detail, because deep knowledge of fuel cell structure and processes, that occur inside, is necessary for choosing the best modeling procedure.

2 Basic principles of hydrogen fuel cells

The basic operation of hydrogen fuel cell is quite simple. It is a reversed electrolysis of water. Hydrogen gas is driven to the anode, where it ionizes to electrons and protons.

$$2H_2 \rightarrow 4H^+ + 4e^-$$

While protons migrate to the cathode through electrolyte, the produced electrons pass there through external electrical circuit, creating thus required electrical current.

The cathode is fed by oxygen, usually in the form of humidified air. Oxygen and hydrogen react there and produce water and heat.

$$O_2 + 4H^+ + 4e^- \rightarrow 2H_2O$$

The hydrogen fuel cell consists of a current collector with gas channels, gas diffusion layer (GDL) and catalyst layer (CL) on the anode and cathode sides as well as a proton conducting polymer membrane in the middle of the cell, which serves as an electrolyte.

Air and hydrogen enter the cell through gas channels and travel to the gas diffusion layers. GDLs have to uniformly distribute reactants on the surface of the catalyst layers and provide structural support for CLs. GDLs also have to transport water to or from CLs and to provide an electrical connection between catalyst layers and current collectors. The common materials for GDLs are carbon paper or carbon cloth, porous materials with typical thickness of 100–300 μm. GDLs are usually coated with Teflon to reduce flooding, that unable reactant gases to travel to CL.

The catalyst layers are the place, where the electrochemical reactions occur, so it has to be designed in order to transport well protons, electrons, as well as gaseous reactants. The thickness of the catalyst layer is typically between 5 and 20 μm and the most commonly used catalyst is platinum.

Since the activity of the catalyst occurs on the surface, we need to increase it. The typical procedure is to spread platinum onto the surface of larger particles of carbon support.

Oxidation of hydrogen at the anode produce protons, that are transported through ion conducting polymer within the catalyst layer to the membrane, and electrons, that travel through the electrically conductive part of the catalyst layer to the gas diffusion layer, then to the collector plates and finally through the load to the cathode. Diffusion and advection through pores are the main ways of transport of gaseous reactants in the catalyst layers. Water produced at the cathode may be liquid or gaseous. Mechanism similar to capillary flow is cause of the liquid water transport through the pores in CL and GDL. When water reaches gas channels, it is dragged out by gas flow.

Membrane plays a vital role in fuel cells. The membrane has to prevent mixing of reactant gases and provide good transport of protons from the anode to the cathode. To minimize losses, the membrane should be very bad electronic conductor. It also has to have high chemical and thermal stability and low production cost.

Although there was developed a plenty of alternative polymer electrolytes, the most common material used for the membrane is still material known under its commercial name Nafion, which was developed by DuPont company in the late 1960s. Nafion consists of a polytetrafluoroethylene backbone and perfluorinated side chains ending by a sulfonate ionic group. The bonds between the fluorine and the carbon make Nafion very durable and chemical-resistant. The conductivity of Nafion depends almost linearly on water content, so we need to keep membrane fully and uniformly humidified at all times. The thickness of the membrane is also crucial factor for optimum fuel cell performance. Thinner membrane has lower ohmic losses, on the other hand, if the membrane is too thin, hydrogen will cross over it to the cathode and react with the oxygen without creating a required electrical current.

Typically, the thickness of the membrane lies in the range of 50-300 μm.

2.1 Processes in fuel cells

We have familiarized ourselves with basic principles and detailed structure of fuel cells. Now we would like to describe transport processes inside the heart of the cell – the membrane.

As we have said, sufficient water content inside the membrane is necessary for good proton conductivity. There are four main causes of water transport inside the membrane: diffusion, electro-osmotic drag and transport caused by pressure and temperature gradients, but the last two are negligible in comparison with diffusion and drag.

Water is created at the cathode and it diffuses to the anode due to the concentration gradient. Protons travel from anode to the cathode and when a proton meets a water molecule, it bounds it by hydrogen bridges forming thus H_3O^+. The higher ions $H_5O_2^+$ and $H_9O_4^+$ can be also created. These ions continue in the earlier proton direction to the cathode. The average number of water molecules dragged by proton is called electro-osmotic drag coefficient and its value is obtained from the experiments. The problem is, that different experimental techniques gives us significantly different values of this coefficient (between one and five water molecules per proton)[3]. So in this case, using modeling techniques for determining electro-osmotic drag coefficient is more than welcomed.

At higher current densities, the produced protons do not allow the water to reach the anode and although the cathode side of membrane is flooded, the anode side can be completely dry. Insufficient water level inside the membrane leads to the poor proton conductivity and thus to lower fuel cell performance. Dry membrane is also more prone to pinhole formation and the degradation process is more fast. Humidifying of the anode is not so easy solution, because excessive liquid water (on both sides) can block the pores in CL or GDL and limited mass transport leads to significant voltage losses. Therefore good water management is one of the main goals in fuel cells design.

Good proton conductivity is a result of the fact that Nafion is a combination of highly hydrophobic polytetrafluoroethylene and highly hydrophilic sulfonic acid. These acid groups are attracted to each other and they form nanoscale hydrophilic domains inside Nafion. If Nafion is sufficiently hydrated, these domains create something like 'water channels,' that allow protons to travel through the membrane, while hydrophobic domains gives the membrane its morphological stability.

It is expected, that there are two main ways, how protons can move within these channels: Grotthuss mechanism and vehicular mechanism.

The vehicular mechanism is a diffusion of hydrated proton ($H^+(H_2O)_x$) due to gradient of electrochemical potential.

Grotthuss mechanism is also known under the name proton hopping. Proton produced at the anode sticks to the water molecule presented in the catalyst-membrane interface creating thus H_3O^+. When this ion is close to another water molecule, proton hops to it. Original ion turns again into water molecule and water molecule changes to hydronium ion. This way, proton hopping continues until it reaches cathode.

The dominance of one the mechanism against the other depends on water level and the precise modeling of all steps of these processes still has to be done.

3 Fuel cell models

This section provides a very brief description of methods and models, that are used for understanding the detailed structure of materials used in fuel cells and transport phenomena inside them.

Fuel cell modeling is a multi-scale problem. To respect this, three different methods in collaboration are used: ab initio models based on quantum mechanics, classical molecular dynamics and continuum models.

The aim of ab initio models is to find the solution of Schrödinger equation

$$H(\vec{r_i}, \vec{R_j})\psi(\vec{r_i}, \vec{R_j}) = E(\vec{R_j})\psi(\vec{r_i}, \vec{R_j}), \tag{1}$$

where the wave function ψ, which describes the state of a molecular system, depends on $3n$ coordinates $\vec{r_i}$ of n electrons and $3N$ coordinates $\vec{R_j}$ of N nuclei.

Born-Oppenheimer approximation, which separates slow nuclear motion from fast electronic motion, is used and the Hamiltonian $H(\vec{r_i}, \vec{R_j})$ is separated to two effective Hamiltonians corresponding to electronic (H_{el}) and nuclear (H_{nuc}) part.

$$H_{el} = \sum_{i=1}^{n} \big[-\frac{\hbar}{2m}\frac{\partial^2}{\partial \vec{r_i}^2} + \sum_{j=1}^{N} V_{el-ion}(\vec{r_i}, \vec{R_j})\big] + \sum_{i,j=1;i>j}^{n} \frac{e^2}{|\vec{r_i} - \vec{r_j}|}, \tag{2}$$

$$H_{nuc} = \sum_{j=1}^{N} \big[-\frac{\hbar}{2M_j}\frac{\partial^2}{\partial \vec{R_j}^2} + \sum_{i=1}^{n} V_{ion-ion}(\vec{r_i}, \vec{R_j})\big] + E_{tot}(\vec{R_j}), \tag{3}$$

where m is the mass of the electron, M_j mass of the jth ion, V_{el-ion}, $V_{ion-ion}$ is potential, which describes direct electron-ion, ion-ion interaction respectively and E_{tot} is total energy of electrons in the field created by ions.

But solving the partial differential equation (1) with Hamiltonian (2) with $3n$ unknowns is still impossible to do exactly, therefore some other approximation has to be made.

To simplify the equations, as first Hartree-Fock approximation is widely used. Because Pauli principle is valid, wave function under this approximation is written as an antisymmetrized product of n molecular orbitals (MO). The choice of optimal MOs is made by variationally minimizing $E(\vec{R_j})$.

In this approximation, the wave function, which solves (1), is reduced to n functions called molecular orbitals(MO). Each molecular orbital describe the probability distribution of a single electron moving in a average field of the other electrons.

Often these unknown MOs are written as a linear combination of a finite set of well known functions, usually Gaussians.

Solving Hartree-Fock equations is still time-demanding and difficult problem, therefore density function theory (DFT) is used very often nowadays. The basis of DFT are famous Hohenberg-Kohn theorems. These theorems claim, that the total energy E of many electron system in an external potential $V_{ex}(\vec{r})$ is a unique functional of the electron density $n(\vec{r})$ and this functional has a minimum at the ground-state density $n_0(\vec{r})$.

$$E[n(\vec{r})] = \int V_{ex}(\vec{r})n(\vec{r})d\vec{r} + f[n(\vec{r})], \tag{4}$$

$$E[n(\vec{r})] \geq E[n_0(\vec{r})]. \tag{5}$$

The functional f has a form

$$f[n(\vec{r})] = T[n(\vec{r})] + \int \frac{n(\vec{r})n(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r} d\vec{r}' + E_{xc}[n(\vec{r})], \tag{6}$$

where $T[n(\vec{r})]$ is the kinetic energy, the second term (often called Hartree term) corresponds to energy of Coulombic repulsion and $E_{xc}[n(\vec{r})]$ represents exchange correlation energy.

The electron density is only function of 3 variables, so the calculations are dramatically simplified, which is the reason, why DFT is now the preferred method for treating large molecules. But there is still a great challenge in determining the functional (6).

Molecular dynamics describes the motion of the molecular system with Newton's second laws. The potential of the molecular system is not a function of electronic wave functions like in ab initio models, but it is a function of the positions of nuclei $U(\vec{R}_j)$. These functions $U(\vec{R}_j)$ are evaluated by methods of quantum mechanics or empirically. Atoms and molecules are considered as classical particles moving in this potential field.

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{F}_i = -\nabla U \tag{7}$$

The good choice of potential U (often called as a force field) is a crucial point in molecular dynamics and it is determined by the bond types, desired accuracy and of course our computational resources. Also comparison with measurements on thermophysical properties and vibration frequencies is necessary for choosing the most suitable force field.

The forcefields use a combination of internal coordinates (bond distances, bond angles, torsions, etc.) for covering the part of the potential energy connected with interactions between bonded atoms and non-bond terms for describing the van der Waals, electrostatic and other interactions between atoms. Forcefields can contain famous Morse potentials, Lennard-Jones potentials, etc.

Continuum models completely ignore the microscopic structure of the substance and assume that the matter continuously fills the space it occupies. On the length scales much greater than that of inter-atomic distances, these models are generally very accurate. Equations, that are able to describe macroscopic behavior of objects, are derived from fundamental physical laws such as the conservation of mass, the conservation of momentum and the conservation of energy. Some other information about object of study is added through constitutive relations.

Ab initio models can reveal us the detailed structure of used materials and microscopic properties and their accuracy depends on approximations we made and on our computational resources and time, we are willing to wait for the results. But the price for accuracy and detailed information about microscopic structure is really high. We are able to analyze only small clusters of few nanometers size within only few picoseconds, so ab initio models are unable to represent real-world macroscopic phenomena.

Unlike ab initio methods, the computational requirements of molecular dynamics are not so high. We are able to investigate systems up to 100 nm length scale and few nanoseconds time scale with still quite good precision. So although molecular dynamics

can display trajectories of atoms and molecules in microscopic system, it is unable to show us the collective behavior of all atoms in real world time scale ($1s$), because the capacity of currently used computers is insufficient and the prognosis about possibility to model a molecular system of macroscopic size (10^{24} atoms) and time ($10^3 s$) within the visibility range of future is not optimistic at all.

If we will summarize it, ab initio models are suitable for understanding breaking and formation of chemical bonds and precise description of chemical reactions. Molecular dynamics can reveal the details of mass transport inside fuel cells and show the proton transport as a function of temperature, water content and other parameters. Continuum models are very good in describing water management and voltage losses in fuel cells.

3.1 Obtained results in last 20 years

The major issue, that scientists interested in fuel cells are facing, is the catalysis of oxygen reduction in cathodic catalyst layer. The oxygen reduction is about 6 orders slower than hydrogen oxidation. This slow reaction rate limits the overall efficiency of the fuel cell. The transport processes and the reaction path in the specific interface structure between the polymer and catalyst particles order the electrochemical activity of the catalyst layer.

Exactly the lack of understanding of catalyst layer structure is the main cause, that despite 30-years' effort, the development of better catalyst did not resulted in a desired progress. There were presented growing number of ab initio studies about oxygen reduction on metals and alloys during last 10 years, they helped with better understanding these mechanisms, but still ab initio modeling did not succeed in providing the detailed reaction description with all of its steps and also did not reveal the structure of interface between hydrated membrane and carbon supported platinum particles.

The oxygen diffusion is first necessary step before the reaction can proceed. So it is essential to understand, how is oxygen transported through this interface, especially how is its transport influenced by the water and polymer clusters distribution at the interface, by carbon support and finally by the electrical field at the interface. The second step is adsorption/desorption of oxygen to catalyst particles. It is commonly accepted, that these processes determine the rate of oxygen reduction. The last step of the reaction is the forming of water and its transport out of catalyst layer (both to GDL and to membrane). So detailed understanding of the interface structure and consequently all reaction steps is a great challenge for ab initio modeling, because so far the presented ab inito models were able to model the interface only as a sheet of catalyst with a water layer, no polymer was involved, although the role of polymer on the reaction is significant.

Molecular dynamics enables to calculate with more atoms than ab initio models, therefore there were attempts to model the interface between CL and polymer electrolyte with it. Currently existing molecular dynamics interface models involve platinum catalyst with its carbon support, water and polymer clusters and these models can give us reliable picture of this interface. These models showed very well, how the presence of polymer cluster changes the water distribution in CL and how is their presence affecting oxygen adsorption. But there is a problem with an electrical double layer and electrode potential, because the picture given by molecular dynamics is not correct. The effect of polymer side groups on electrical double layer and the shape of electrode potential is not described

properly.

We have said, that the proton conductivity is due to Grotthuss or vehicular mechanism. These mechanisms were found, while trying to understand, why is the proton conductivity in water $5-8$ times higher than the conductivity of other cations. The scientists successfully applied these mechanism also for explaining the proton conductivity of Nafion. Ab initio methods were found very useful for proving these mechanisms in Nafion.

Molecular dynamics can create very realistic model of proton diffusion in Nafion, because in last decade, there was published a series of paper describing new force fields, that proved themselves very suitable for modeling of transport processes in Nafion. The movement of proton was studied by tracking the trajectories of protons in the membrane. These new force fields also revealed the formation of water clusters around Nafion side chains and their changing to water channels with a growing water content. The simulation data were consistent with experimental results in a large range of water content. So molecular dynamics can serve as a guide for optimizing Nafion properties or in better case, it can show us the way for developing new cheaper materials, that can replace Nafion in the future.

There are many commerical numerical programs based on continuum mechanics. And because fuel cells became very popular between scientists, there was naturally demand also for software suitable for modeling fuel cells. One of the companies, that satisfied this demand, was COMSOL AB. They created Battery and fuel cell module suitable for modeling transport processes inside fuel cells [2]. This module is suitable for modeling mass transport, current-density distribution on the electrode surfaces or the influence of the gas channels in current collectors on the current-density distribution and the distribution of reactant gases over catalysts.

We are interested in processes in CL, so it was naturally to use COMSOL for modeling the mass transport through catalyst layer. It is assumed, that mass transport can be described by Maxwell-Stefan diffusion and the electrochemical reactions at the cathode are expected to be Tafel-like. The details about all assumptions can be found in [2].

Figure 3.1 shows a geometry of cathode, where the pink volume corresponds to the small domain, we will model. The holes in the collectors represent the places, where humidified oxygen enters the modeling domain. The reactive layer has a porous structure and it is a mixture feed-gas, carbon support carrying platinum catalyst and electrolyte. The electrolyte layer represents Nafion, no reaction can occur there, it is also not allowed oxygen and electrons to go to the electrolyte. Both layers are $75\mu m$ thick.

It is obvious from Figure 3.1, that oxygen concentration along thickness of reactive layer is almost constant, but it is significantly decreasing while moving apart from the hole. Because of this, the reaction rate is nonuniform in the reactive layer. This nonuniformity has an influence also on the current density distribution. The current density is also highly nonuniform, as can be seen in Figure 3.1.

4 Conclusions

The aim of this paper was to familiarize myself with a current state of the art in the fuel cell modeling. This review should help me to find a particular problem I should

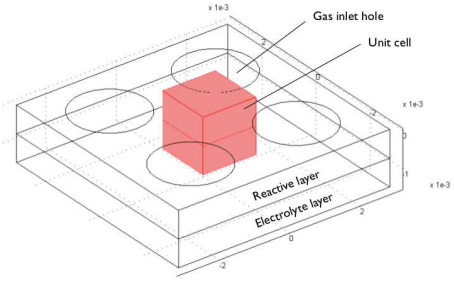


Figure 1: Hydrogen fuel cell cathode. [2]

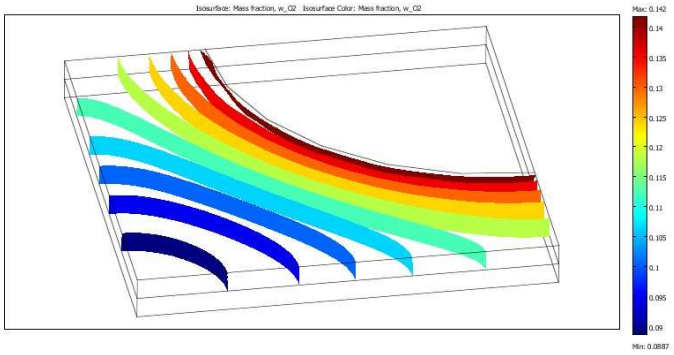


Figure 2: Oxygen concentration.

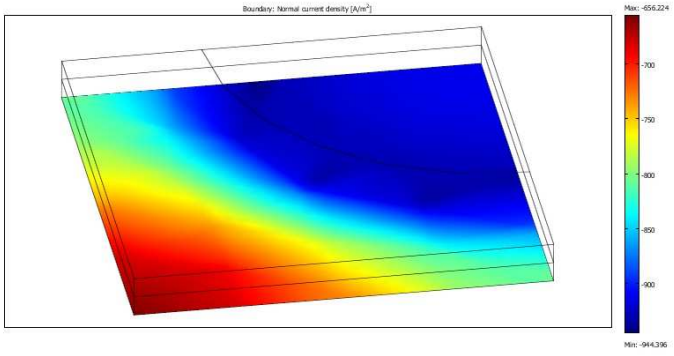


Figure 3: Produced current density.

concentrate on. The problem of chemical reaction in the catalyst layers offers too many interesting problems to solve. I made very simple model of chemical reactions in COMSOL and I would like to make more detailed model in order to be in correspondence with our measurement data. But I am mainly interested in modeling methods based on quantum mechanics, so I would like to use them for modeling the chemical reactions in the catalyst layers and gain thus much more reliable data then from simple modeling based only on continuum mechanics.

References

[1] M.S. Al-Baghdadi. *PEM Fuel Cell Modeling* In 'Fuel Cell Research Trends', L.O.Vasquez (ed.) Nova Science Publishers, Inc. (2007), 273–380

[2] COMSOL AB. *Chemical Engineering Module Model Library* Version 2007, COMSOL 3.4

[3] W. Dai, H. Wang, X.-Z. Yuan, J. Martin, D. Yang, J. Qiao, J. Ma. *A review on water balance in the membrane electrode assembly of proton exchange membrane fuel cells* International Journal of Hydrogen Energy 34 (2009), 9461–9478

[4] L. Kalvoda, P. Sedlák, M. Dráb. *Počítačové simulace kondenzovaných látek* notes from lecture presented at the Department of Solid State Engineering, Faculty of Nuclear Sciences and Physical Engineering

[5] K.-D. Kreuer, S.J. Paddison, E. Spohr, M. Schuster. *Transport in Proton Conductors for Fuel-Cell Applications: Simulations, Elementary Reactions, and Phenomenology* Chemichal Reviews 2004, 104, 4637–4678

[6] S.J. Peighambardoust, S. Rowshanzamir, M. Amjadi. *Review of the proton exchange membranes for fuel cell applications* Internation Journal of Hydrogen Energy 35 (2010), 9349–9384

[7] X. Zhou, J. Zhou, Y. Yin. *Atomistic modeling in Study of Polymer Electrolyte Fuel cells – A Review* In 'Modern aspects of electrochemistry, No.49: Modeling and diagnostics of polymer electrolyte fuel cells', C.-Y.Wang, U.Pasaogullari (eds.) Springer (2010), 307–376

Conserved Quantities in Repeated Interaction Quantum Systems*

Helena Šedivákova

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

advisor: David Krejčiřík, Nuclear Physics Institute, AS CR

Abstract. In the model of repeated interaction quantum systems, a reference system interacts successively with a chain of identical quantum systems and its long-time behavior is studied. Mathematically, the asymptotic state of the reference system corresponds to the states invariant with respect to certain operator describing the dynamics of the composed system. Such states were found in previous works, however, in this paper we give a simple way how to obtain them only from the knowledge of quantities that commute with the total Hamiltonian.

Keywords: Repeated interactions, conserved quantities, invariant states

Abstrakt. Kvantové systémy s opakovanou interakcí modelují situaci, kdy určitý referenční systém interaguje postupně s řetězcem identických kvantových systémů a zkoumá se jeho chování v limitě dlouhého času. Matematicky asymptotický stav referenčního systému odpovídá stavům invariantním vůči jistému operátoru popisujícímu dinamiku složeného systému. Hledáním takových stavů se zabývala již řada prací, avšak v tomto článku navrhujeme jednoduchý způsob, jak invariantní stavy najít pouze na základě znalosti veličin, které komutují s celkovým Hamiltoniánem.

Klíčová slova: Opakované interakce, zachovávající se veličiny, invariantní stavy

1 Introduction

Motivated by the setup of "one-atom maser" experiment [5], repeated interaction quantum systems (RIQS) have been studied mathematically in the last years. In this model, we consider a "reference" or "small" system $\mathcal{S}$ that interacts successively with the elements $\mathcal{E}$ of a chain $\mathcal{C}$ of independent quantum systems, and the state of $\mathcal{S}$ after great number of such interactions is studied.

In [1], so called "repeated interaction asymptotic state" of the small system was found for the general setting and this state was proved to be independent of the initial state of $\mathcal{S}$. It was shown that the asymptotic state corresponds to the eigenvalue 1 eigenstate of certain operator describing the dynamics, *i.e.* the state that is invariant under the action of this operator. The speed of the convergence to the asymptotic state was determined to be exponential when $\mathcal{S}$ is finite-dimensional. To extend these results, in [2] the randomness in interaction time and state of incoming atoms was taken into account, while in [3] the environment was interacting with the small system besides the atoms.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS11/132/OHK4/2T/14.

In papers [1], [2], and [3], particular examples of the small system, the atoms, and their interaction are given and the particular formula for the asymptotic state is found. However, only finite-dimensional small systems are considered, hence the case of maser cavity from [5] being the small system is not included, as the electromagnetic field is usually modeled by the harmonic oscillator, *i.e.* an infinite-dimensional quantum system. Furthermore, for most of the interactions considered in the papers above, the perturbation regime for the small coupling constant has to be used. On the contrary, [4] deals with harmonic oscillator as a small system from the beginning and since the interaction is described by simple Jaynes-Cummings Hamiltonian, the asymptotic state can be found without using the perturbation theory.

Our paper deals with the following result of [4] on the thermalization of the small system. If the atoms in the chain are assumed to be "thermal" (*i.e.* to be in stationary states that are parameterized by an inverse temperature $\beta > 0$) the asymptotic state of $\mathcal{S}$ was proved (under some non-resonance condition on the system) to be a thermal state with respect to a certain temperature β^*. We will show that the last result does not hold for more general systems since if the Hamiltonian of the small system is slightly perturbed, then the asymptotic state is no more the thermal one. On the other hand, we show that the asymptotic state is closely related to the quantities that are conserved in the interaction.

In Theorem 1, we state that the relation mentioned above appears in general; we state that if a quantity M of certain form is conserved in the interaction (*i.e.* if M commutes with the total Hamiltonian H), then there is an invariant state of the dynamics which can be expressed by means of the conserved quantity. The proof is very simple, however, this statement may be very useful for studying the RIQS. The assumptions of the theorem may be even weakened, it is enough for M to commute with $e^{i\tau H}$ for all $\tau \in \mathbb{R}$ (see Theorem 4). Subsequently, we apply the theorems mentioned above to examples.

The paper is organized as follows. In Section 2.1 the general setup of the repeated interaction quantum systems (RIQS) is given and the importance of invariant state is explained. As an example of RIQS, we describe the model for one-atom maser in Section 2.2 and we add several results on the behavior of this system obtained in [4]. As our main result, a theorem on the relation between conserved quantities and invariant states is given and proved in Section 3.1, where it is also explained how this theorem works in the case of the atom-field interaction. In Section 3.2 we summarize the results on the example of perturbed atom-field interaction: in Theorem 3, we give all the diagonal invariant states, which is a result obtained by explicit calculations. Then we state Theorem 4 which also enables us to find all the diagonal invariant states, however, in much simpler way. In Section 3.3, the example of spin-spin interaction is studied. Finally, the results are summed up and commented on in Section 4.

2 Preliminaries

2.1 General setup of RIQS

Let us consider a small system $\mathcal{S}$ interacting with a chain $\mathcal{C}$ of identical elements $\mathcal{E}$. The states of $\mathcal{S}$, resp. $\mathcal{E}$ are represented by vectors from the Hilbert space $\mathcal{H}_\mathcal{S}$, resp. $\mathcal{H}_\mathcal{E}$. We

suppose that the elements of the chain interact with the small system successively, the m-th element interacting with the small system in the time interval $((m-1)\tau, m\tau)$, and we do not consider any direct interaction between different elements of the chain. The dynamics of the system composed of $\mathcal{S}$ and arbitrary element $\mathcal{E}$ is given by Hamiltonian H acting on the Hilbert space $\mathcal{H}_\mathcal{S} \otimes \mathcal{H}_\mathcal{E}$ where H includes the free dynamics of $\mathcal{S}$ and $\mathcal{E}$ and the interaction of these two systems.

Let $\rho_0 \in \mathcal{J}_1(\mathcal{H}_\mathcal{S})$ (a trace one operator on $\mathcal{H}_\mathcal{S}$) be an initial state of the small system (state in time $t = 0$) and let $\rho_\mathcal{E} \in \mathcal{J}_1(\mathcal{H}_\mathcal{E})$ be the state of incoming elements of the chain. We assume that $\rho_\mathcal{E}$ is invariant with respect to the free evolution of $\mathcal{E}$. Using this fact and basic rules of quantum mechanics we find that the state ρ_n of the small system after interaction with n elements of the chain (state in time $n\tau$) is given by $\rho_n = \mathcal{L}^n(\rho_0)$ where

$$\mathcal{L}(\rho) := \mathrm{Tr}_{\mathcal{H}_\mathcal{E}} \left(e^{-iH\tau}(\rho \otimes \rho_\mathcal{E}) e^{iH\tau} \right) \tag{1}$$

(for detailed derivation see [1]). Notice that this means that the system is Markovian, *i.e.* that the state ρ_n depends only on the state ρ_{n-1}.

We say that $\rho_* \in \mathcal{J}_1(\mathcal{H}_\mathcal{S})$ is invariant with respect to $\mathcal{L}$ if

$$\mathcal{L}(\rho_*) = \rho_*. \tag{2}$$

Looking for invariant states is the main topic of this paper, so let us explain why they are important for the dynamics of RIQS.

In the special case when there is unique invariant state of $\mathcal{L}$ and when $\mathcal{H}_\mathcal{S}$ is finite dimensional, it is easy to show that $\mathcal{L}^n(\rho_0)$ converges to ρ_* when $n \to \infty$ (*i.e.* ρ_* is the asymptotic state of the small system), the speed of the convergence is exponential (*i.e.* $\|\mathcal{L}^n(\rho_0) - \rho_*\| \propto e^{-\gamma n}$ for some constant $\gamma > 0$), and this holds independently of the initial state ρ_0.

In [4] $\mathcal{H}_\mathcal{S} = \ell^2(\mathbb{N})$ is infinite dimensional, but still invariant states play important role for the asymptotics. If there exists unique invariant state then the relation

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} (\mathcal{L}^n(\rho_0))(A) = \rho_*(A) \tag{3}$$

holds for any initial state ρ_0 and any observable $A \in \mathcal{B}(\mathcal{H}_\mathcal{S})$ due to the ergodic theorem and we say that the small system converges to ρ_* in the ergodic sense.

2.2 Example: atom-field interaction

As an example of RIQS, the setup of [4] is described in this section and also a few results of this paper are given as we will work with them later.

In the model of "one atom maser", the elements of the chain are two-level atoms, hence their states are described by vectors from Hilbert space $\mathcal{H}_\mathcal{E} = \mathbb{C}^2$ and the free Hamiltonian reads

$$H_\mathcal{E} = \omega_0 b^* b = \begin{pmatrix} 0 & 0 \\ 0 & \omega_0 \end{pmatrix}. \tag{4}$$

Here b, resp. b^* read the annihilation, resp. creation operators and ω_0 is the difference of the two energy levels of the atoms. We will denote the ground and excited atom states by

$$|-\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |+\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{5}$$

in this notation

$$b = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = |-\rangle\langle+|.$$

The role of the small system is played by a single mode of electromagnetic field in a cavity tuned to the excitation energy of the atoms. Hence $\mathcal{S}$ is a harmonic oscillator of frequency $\omega \approx \omega_0$ (more precisely $\Delta := \omega - \omega_0$ is assumed to satisfy $|\Delta| \ll \min(\omega_0, \omega)$), *i.e.* the Hilbert space is $\mathcal{H}_\mathcal{S} = \ell^2(\mathbb{N})$ and

$$H_\mathcal{S} = \omega N = \omega a^* a = \sum_{n=0}^{\infty} \omega |n\rangle\langle n| \tag{6}$$

is the free Hamiltonian written in terms of number operator N, creation and annihilation operators a^* and a, or in the bra-ket formalism in the energy representation, respectively.

So called "rotating wave approximation" is used to describe the coupling of the field and the atoms, hence the Jaynes-Cummings Hamiltonian describes the dynamics of the coupled system:

$$\begin{aligned} H &= H_\mathcal{S} \otimes \mathbb{1}_\mathcal{E} + \mathbb{1}_\mathcal{S} \otimes H_\mathcal{E} + \lambda V, \\ V &= \frac{1}{2}\left(a \otimes b^* + a^* \otimes b\right) \end{aligned} \tag{7}$$

From mathematical point of view, H is convenient since it commutes with the "total number operator"

$$M = a^* a \otimes \mathbb{1}_\mathcal{E} + \mathbb{1}_\mathcal{S} \otimes b^* b \tag{8}$$

which allows the explicit diagonalization of H. Moreover, the relation $[M, H] = 0$ will be important for application of our Theorem 1.

The incoming atoms are in the thermal state

$$\rho_\mathcal{E}^\beta = \frac{e^{-\beta H_\mathcal{E}}}{\mathrm{Tr} e^{-\beta H_\mathcal{E}}} \tag{9}$$

and we define the operator analogous to (1), $\mathcal{L}_\beta(\rho) := \mathrm{Tr}_{\mathcal{H}_\mathcal{E}}\left(e^{-iH\tau}(\rho \otimes \rho_\mathcal{E}^\beta)e^{iH\tau}\right)$ where H reads (7). $\mathcal{L}_\beta$ can be written down explicitly due to the simple form of the Hamiltonian and also correspondent invariant states can be found just by algebraic computations, without any perturbation expansion. In this way it was obtained in [4] that if

$$\frac{\tau}{2}\sqrt{\lambda^2 n + \Delta^2} \neq k\pi \qquad \forall k, n \in \mathbb{N} \tag{10}$$

then there exists unique invariant state

$$\rho_\mathcal{S}^{\beta^*} = \frac{e^{-\beta^* H_\mathcal{S}}}{\mathrm{Tr} e^{-\beta^* H_\mathcal{S}}} \tag{11}$$

where $\beta^* = \beta\frac{\omega_0}{\omega}$. The model with constants satisfying (10) is said to be "non-resonant" and for such setup the field in the cavity converges to the state (11) in the ergodic sense.

It may be said that the small system is drown to a thermal state by the interaction with the thermal atoms. However, we will see in the following that this thermalization will not occur if the Hamiltonian of the small system is slightly perturbed and that the form of the invariant state corresponds rather to the quantity (8) conserved by the dynamics of the composed system then to the free Hamiltonian of the small system which has here by accident the same form as the part of M acting on $\mathcal{H}_\mathcal{S}$, *i.e.* multiple of a^*a.

In case when (10) is not satisfied (simply resonant, resp. fully resonant systems), there exist two, resp. infinite number of invariant states, but (11) is always included.

3 Results

3.1 Invariant states induced by conserved quantities

In this section a general theorem on the connection between the conserved quantities (*i.e.* quantities that commute with the Hamiltonian) and the invariant states is stated, a simple proof is given and the theorem is then applied on the example from Section 2.2.

Theorem 1. *Let M be a self-adjoint operator on $\mathcal{H}_\mathcal{S}\otimes\mathcal{H}_\mathcal{E}$ that satisfies $[M,H]=0$, and that can be written in the form $M = M_\mathcal{S}\otimes\mathbb{1}_\mathcal{E} + \mathbb{1}_\mathcal{S}\otimes M_\mathcal{E}$. Let $\alpha\in\mathbb{R}$ be a constant such that both $\mathrm{Tr}(e^{\alpha M_\mathcal{E}})<\infty$ and $\mathrm{Tr}(e^{\alpha M_\mathcal{S}})<\infty$. If we put*

$$\mathcal{L}_\alpha(\rho) := \mathrm{Tr}_{\mathcal{H}_\mathcal{E}}\left(e^{-iH\tau}\left(\rho\otimes\rho_\mathcal{E}^\alpha\right)e^{iH\tau}\right)$$

where $\rho_\mathcal{E}^\alpha = \frac{e^{\alpha M_\mathcal{E}}}{\mathrm{Tr}e^{\alpha M_\mathcal{E}}}$ then $\rho_*^\alpha := \frac{e^{\alpha M_\mathcal{S}}}{\mathrm{Tr}e^{\alpha M_\mathcal{S}}}$ is an invariant state of $\mathcal{L}_\alpha$.

Proof. Since

$$e^{\alpha M} = \exp\left[\alpha\left(M_\mathcal{S}\otimes\mathbb{1}_\mathcal{E}+\mathbb{1}_\mathcal{S}\otimes M_\mathcal{E}\right)\right] = \left(e^{\alpha M_\mathcal{S}}\otimes\mathbb{1}_\mathcal{E}\right)\left(\mathbb{1}_\mathcal{S}\otimes e^{\alpha M_\mathcal{E}}\right) = e^{\alpha M_\mathcal{S}}\otimes e^{\alpha M_\mathcal{E}},$$

and since M and H commute, we get

$$\begin{aligned}\mathcal{L}_\alpha\left(\frac{e^{\alpha M_\mathcal{S}}}{\mathrm{Tr}e^{\alpha M_\mathcal{S}}}\right) &= \mathrm{Tr}_{\mathcal{H}_\mathcal{E}}\left(e^{-iH\tau}\left(\frac{e^{\alpha M_\mathcal{S}}}{\mathrm{Tr}e^{\alpha M_\mathcal{S}}}\otimes\frac{e^{\alpha M_\mathcal{E}}}{\mathrm{Tr}e^{\alpha M_\mathcal{E}}}\right)e^{iH\tau}\right)\\ &= \frac{1}{\mathrm{Tr}e^{\alpha M_\mathcal{S}}\mathrm{Tr}e^{\alpha M_\mathcal{E}}}\mathrm{Tr}_{\mathcal{H}_\mathcal{E}}\left(e^{-iH\tau}e^{\alpha M}e^{iH\tau}\right)\\ &= \frac{1}{\mathrm{Tr}e^{\alpha M_\mathcal{S}}\mathrm{Tr}e^{\alpha M_\mathcal{E}}}\mathrm{Tr}_{\mathcal{H}_\mathcal{E}}e^{\alpha M} = \frac{e^{\alpha M_\mathcal{S}}}{\mathrm{Tr}e^{\alpha M_\mathcal{S}}}.\end{aligned}$$

□

Let us go back now to the example of atom-field interaction from Section 2.2 where the conserved quantity M in the desired form really occur (see (8)). If we realize that $H_\mathcal{S} = \omega M_\mathcal{S}$ and $H_\mathcal{E} = \omega_0 M_\mathcal{E}$ holds (see (6) and (4)), and if we put $\alpha = -\beta\omega_\mathcal{E}$ then $\rho_\mathcal{E}^\alpha = \frac{e^{-\beta H_\mathcal{E}}}{\mathrm{Tr}e^{-\beta H_\mathcal{E}}}$ which corresponds to (9) and we get by the Theorem 1 an invariant state

$$\rho_*^\alpha = \frac{e^{-\beta\omega_0 M_\mathcal{S}}}{\mathrm{Tr}e^{-\beta\omega_0 M_\mathcal{S}}} = \frac{e^{-\beta\frac{\omega_0}{\omega}H_\mathcal{S}}}{\mathrm{Tr}e^{-\beta\frac{\omega_0}{\omega}H_\mathcal{S}}}.$$

This is exactly the formula for $\rho_{\mathcal{S}}^{\beta^*}$ from (11).

In conclusion we have found an invariant state of $\mathcal{L}_\beta$ in very simple way in comparison with the computations from [4]. Of course, the uniqueness can not be proved in this way. On the other hand, we will see in the next section (where we study a model which includes the setup of Section 2.2 as a special case) that it is possible to find all the diagonal invariant states using (slightly modified version of) Theorem 1.

3.2 Example: Perturbed atom-field interaction

Let us now consider a small perturbation of the dynamics in the following way:

$$H'_{\mathcal{S}} = \sum_{n=0}^{\infty} (n\omega + \delta_n)|n\rangle\langle n| \tag{12}$$
$$V' = \frac{1}{2}\left[\left(\sum_{n=0}^{\infty} \lambda_n\sqrt{n+1}|n+1\rangle\langle n|\right)\otimes b + \left(\sum_{n=1}^{\infty}\lambda_n\sqrt{n}|n-1\rangle\langle n|\right)\otimes b^*\right],$$

$\delta_n, \lambda_n \in \mathbb{R}$. Here $\{|n\rangle\}_{n=0}^{\infty}$ are the eigenstates of $H'_{\mathcal{S}}$, which are assumed to form an orthogonal basis of $\mathcal{H}_{\mathcal{S}}$. The total Hamiltonian is given by $H' = H'_{\mathcal{S}}\otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}}\otimes H_{\mathcal{E}} + V$ and we define

$$\mathcal{L}'_\beta(\rho) := \mathrm{Tr}_{\mathcal{H}_{\mathcal{E}}}\left[e^{-i\tau H'}(\rho\otimes\rho_{\mathcal{E}}^{\beta})e^{i\tau H'}\right], \qquad \rho\in\mathcal{J}_1(\mathcal{H}_{\mathcal{S}}) \tag{13}$$

($\rho_{\mathcal{E}}^{\beta}$ was defined in (9)). We again look for states ρ_* satisfying

$$\mathcal{L}'_\beta(\rho_*) = \rho_*. \tag{14}$$

It can be seen that the operator

$$M' = \sum_{n=0}^{\infty} n|n\rangle\langle n|\otimes\mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}}\otimes b^*b = M'_{\mathcal{S}}\otimes\mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}}\otimes M'_{\mathcal{E}}, \tag{15}$$

analogous to (8), commutes with H', hence the state

$$\frac{e^{-\beta\omega_0 M'_{\mathcal{S}}}}{\mathrm{Tr}e^{-\beta\omega_0 M'_{\mathcal{S}}}} \tag{16}$$

is again an invariant state of the dynamics according to Theorem 1.

Remark 2. *Notice that due to the change in the Hamiltonian of the small system, the relation $H'_{\mathcal{S}} = \omega M'_{\mathcal{S}}$ does not hold and the invariant state can not be interpreted as the thermal state of the small system.*

By now, the question if (16) is a unique invariant state remains open. That is why, we looked for the invariant states explicitly by solving equation (14), and in the theorem below we summarize the results. The proof closely follows the procedure of [4], and we do not give it in this paper. To state the theorem, we have to start with several definitions.

As in non-perturbed case, the number of solutions of (14) (at least among diagonal matrices) depends on condition similar to (10). Hence we define

$$\mathcal{R} := \left\{ n \in \mathbb{N} \middle| \exists k \in \mathbb{N}, \quad \frac{\tau}{2}\sqrt{(\Delta + \tilde{\delta}_n)^2 + n\lambda_n^2} = k\pi \right\} \tag{17}$$

which is analogue of the set of Rabi resonances from [4]. Similarly as in [4] we decompose $\mathbb{N}_0$ according to set $\mathcal{R} = \{n_1, n_2 ...\}$ as

$$I_1 = \{0, ..., n_1 - 1\}, I_2 = \{n_1, ..., n_2 - 1\}, ...$$

If we denote $\mathcal{H}_{\mathcal{S}}^{(k)} = \ell^2(I_k)$, then we can decompose the the Hilbert space of the small system as $\mathcal{H}_{\mathcal{S}} = \bigoplus_{k=1}^{r} \mathcal{H}_{\mathcal{S}}^{(k)}$ where $r-1$ is the number of integers in the set $\mathcal{R}$ (of course, this number may be infinite). We denote by P_k the orthogonal projection on $\mathcal{H}_{\mathcal{S}}^{(k)}$ and we use the notation $N = \sum_{n=0}^{\infty} n|n\rangle\langle n| (= M'_{\mathcal{S}})$.

Theorem 3. *Let $\beta > 0$. Then all the diagonal invariant states of $\mathcal{L}'_\beta$ are*

$$\rho_*^{(k)} = \frac{e^{-\beta\omega_0 N} P_k}{\mathrm{Tr} e^{-\beta\omega_0 N} P_k}, \qquad k \in \{1, 2, ..., r\}. \tag{18}$$

Moreover, let ρ_* satisfy (14)*. Then the diagonal of ρ_* is a linear combination of states* (18)*.*

For simplicity we restricted here to the case $\beta > 0$, however, the case $\beta \leq 0$ may be included easily.

All the invariant states (18) may be found using following modification of Theorem 1.

Theorem 4. *Let M be a self-adjoint operator on $\mathcal{H}_{\mathcal{S}} \otimes \mathcal{H}_{\mathcal{E}}$ that satisfies $\left[M, e^{i\tau H}\right] = 0$ for any $\tau \in \mathbb{R}$, and that can be written in the form $M = M_{\mathcal{S}} \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes M_{\mathcal{E}}$. Let $\alpha \in \mathbb{R}$ be a constant such that both $\mathrm{Tr}(e^{\alpha M_{\mathcal{E}}}) < \infty$ and $\mathrm{Tr}(e^{\alpha M_{\mathcal{S}}}) < \infty$. If we put*

$$\mathcal{L}_\alpha(\rho) := \mathrm{Tr}_{\mathcal{H}_{\mathcal{E}}} \left(e^{-iH\tau} \left(\rho \otimes \rho_{\mathcal{E}}^\alpha \right) e^{iH\tau} \right)$$

where $\rho_{\mathcal{E}}^\alpha = \frac{e^{\alpha M_{\mathcal{E}}}}{\mathrm{Tr} e^{\alpha M_{\mathcal{E}}}}$ then $\rho_*^\alpha := \frac{e^{\alpha M_{\mathcal{S}}}}{\mathrm{Tr} e^{\alpha M_{\mathcal{S}}}}$ is an invariant state of $\mathcal{L}_\alpha$.

The proof is identical as the proof of Theorem 1, however, this modified version enables us to consider the observables $M_k = \sum_{n=n_{k-1}}^{n_k - 1} n|n\rangle\langle n| \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes b^* b$ that commute with $e^{i\tau H}$ but not with H. Using Theorem 4, we come to all the invariant states (18).

3.3 Example: Spin-spin interaction

In this section we apply Theorem 1 to the example mentioned in Section 3.3 of [1]. In [1], explicit calculations were made and the asymptotic state was found for general case, whereas here we obtain some results in special cases only. On the other hand, the computation becomes much simpler.

In this model, the small system as well as the incoming atoms are just two-level systems, hence $\mathcal{H}_{\mathcal{S}} = \mathcal{H}_{\mathcal{E}} = \mathbb{C}^2$ and the Hamiltonians read

$$H_{\mathcal{S}} = \begin{pmatrix} 0 & 0 \\ 0 & E_{\mathcal{S}} \end{pmatrix}, \qquad H_{\mathcal{E}} = \begin{pmatrix} 0 & 0 \\ 0 & E_{\mathcal{E}} \end{pmatrix}$$

respectively (for simplicity, we do not introduce any new notation for the quantities like $H_{\mathcal{S}}$ or M in this section). In [1], the initial state of incoming atoms was assumed to be

$$\rho_{\mathcal{E}}^{\beta} = \frac{e^{-\beta H_{\mathcal{E}}}}{\mathrm{Tr} e^{-\beta H_{\mathcal{E}}}} = \begin{pmatrix} \frac{1}{1+e^{-\beta E_{\mathcal{E}}}} & 0 \\ 0 & \frac{1}{1+e^{+\beta E_{\mathcal{E}}}} \end{pmatrix}$$

for some inverse temperature β, however, by applying Theorem 1 we may get also different initial state (however, we will see that it will not happen). The coupling acting on the total Hilbert space $\mathcal{H} = \mathcal{H}_{\mathcal{S}} \otimes \mathcal{H}_{\mathcal{E}}$ reads in general:

$$V = I \otimes a + I^* \otimes a^*. \tag{19}$$

Here I is a general complex matrix, and a, resp. a^* are annihilation, resp. creation operators of the atoms:

$$I = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad a = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

The total Hamiltonian (for interaction of the small system with one atom) reads $H = H_{\mathcal{S}} \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes H_{\mathcal{E}} + V$ and to apply Theorem 1 we will look for general matrices

$$M = \begin{pmatrix} x_{\mathcal{S}} & z_{\mathcal{S}} \\ z_{\mathcal{S}}^* & y_{\mathcal{S}} \end{pmatrix} \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes \begin{pmatrix} x_{\mathcal{E}} & z_{\mathcal{E}} \\ z_{\mathcal{E}}^* & y_{\mathcal{E}} \end{pmatrix} \qquad x_{\mathcal{S}}, y_{\mathcal{S}}, x_{\mathcal{E}}, y_{\mathcal{E}} \in \mathbb{R};\ z_{\mathcal{S}}, z_{\mathcal{E}} \in \mathbb{C} \tag{20}$$

that satisfy

$$[M, H] = 0. \tag{21}$$

While solving this equation, we assume that $E_{\mathcal{E}}, E_{\mathcal{S}} \neq 0$.

In the following we give all the solutions of equation (21). The solution exists only for particular choices of the coupling V_j (given by (19) with appropriate constants A, B, C, D), hence we divide the analysis into several cases. For each coupling V_j the solution of (21) M_j is given, and the corresponding form of incoming atoms $\rho_{\mathcal{E}}^{(j)}$ and invariant state $\rho_*^{(j)}$ is derived.

Case 1. *General values of* $A, B, C, D \in \mathbb{C}$.

For coupling of this form the only solution is

$$M_1 = \begin{pmatrix} x_{\mathcal{S}} & 0 \\ 0 & x_{\mathcal{S}} \end{pmatrix} \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes \begin{pmatrix} x_{\mathcal{E}} & 0 \\ 0 & x_{\mathcal{E}} \end{pmatrix} \qquad x_{\mathcal{S}}, x_{\mathcal{E}} \in \mathbb{R}$$

which corresponds to the case

$$\rho_{\mathcal{E}}^{(1)} = \rho_*^{(1)} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Hence for any coupling of form (19), it holds that if atoms with "infinite temperature" ($\beta \to 0$) interact with the small system, then the small system comes also to the thermal state corresponding to the infinite temperature.

Case 2. $B = C = 0,\ A, D \in \mathbb{C}$

Here the total Hamiltonian commutes with any matrix

$$M_2 = \begin{pmatrix} x_{\mathcal{S}} & 0 \\ 0 & y_{\mathcal{S}} \end{pmatrix} \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes \begin{pmatrix} x_{\mathcal{E}} & 0 \\ 0 & x_{\mathcal{E}} \end{pmatrix} \qquad x_{\mathcal{S}}, y_{\mathcal{S}}, x_{\mathcal{E}} \in \mathbb{R}.$$

This suggests that the incoming atoms "with infinite temperature" leave the small system invariant in any diagonal state, *i.e.*

$$\rho_{\mathcal{E}}^{(2)} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \qquad \rho_{*}^{(2)} = \begin{pmatrix} t & 0 \\ 0 & 1-t \end{pmatrix}, \qquad t \in (0,1).$$

Let us note that the coupling V_2 has the property $[V_2, H_{\mathcal{S}}] = 0$, hence it is quite natural that the state of $\mathcal{S}$ is preserved. On the other hand, the coupling is not trivial as $[V_2, H_{\mathcal{E}}] \neq 0$.

Case 3. $A = C = D = 0,\ B \neq 0 \in \mathbb{C}$.

This coupling generalizes the toy-model Jaynes-Cummings coupling (where $B \in \mathbb{R}$) and the matrix that commutes with the Hamiltonian is then

$$M_3 = \begin{pmatrix} x_{\mathcal{S}} & 0 \\ 0 & x_{\mathcal{S}} + y_{\mathcal{E}} - x_{\mathcal{E}} \end{pmatrix} \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes \begin{pmatrix} x_{\mathcal{E}} & 0 \\ 0 & y_{\mathcal{E}} \end{pmatrix} \qquad x_{\mathcal{S}}, x_{\mathcal{E}}, y_{\mathcal{E}} \in \mathbb{R}$$

This admits the "thermal" incoming atoms. If we denote $y_{\mathcal{E}} - x_{\mathcal{E}} =: E_{\mathcal{E}}$ and use the parametrization by the inverse temperature β as the parameter $-\alpha$ from Theorem 1, then

$$\rho_{\mathcal{E}}^{(3)} = \frac{1}{1 + e^{-\beta E_{\mathcal{E}}}} \begin{pmatrix} 1 & 0 \\ 0 & e^{-\beta E_{\mathcal{E}}} \end{pmatrix}. \tag{22}$$

Corresponding invariant states are then identical with the incoming atoms, which may be interpreted as thermal state of the small system with inverse temperature $\beta_* = \frac{E_{\mathcal{E}}}{E_{\mathcal{S}}}\beta$:

$$\rho_{*}^{(3)} = \frac{1}{1 + e^{-\beta E_{\mathcal{E}}}} \begin{pmatrix} 1 & 0 \\ 0 & e^{-\beta E_{\mathcal{E}}} \end{pmatrix} = \frac{1}{1 + e^{-\beta_* E_{\mathcal{S}}}} \begin{pmatrix} 1 & 0 \\ 0 & e^{-\beta_* E_{\mathcal{S}}} \end{pmatrix}.$$

Case 4. $A = B = D = 0,\ C \neq 0 \in \mathbb{C}$.

In the last case where a solution of (21) exists we get

$$M_4 = \begin{pmatrix} x_{\mathcal{S}} & 0 \\ 0 & x_{\mathcal{S}} - y_{\mathcal{E}} + x_{\mathcal{E}} \end{pmatrix} \otimes \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{S}} \otimes \begin{pmatrix} x_{\mathcal{E}} & 0 \\ 0 & y_{\mathcal{E}} \end{pmatrix} \qquad x_{\mathcal{S}}, x_{\mathcal{E}}, y_{\mathcal{E}} \in \mathbb{R}.$$

If we again parameterize the state of incoming atoms as in (22), *i.e.* $\rho_{\mathcal{E}}^{(4)} = \rho_{\mathcal{E}}^{(3)}$, then

$$\rho_{*}^{(4)} = \frac{1}{1 + e^{\beta E_{\mathcal{E}}}} \begin{pmatrix} 1 & 0 \\ 0 & e^{\beta E_{\mathcal{E}}} \end{pmatrix}$$

which suggests that in this case it would hold $\beta_* = -\frac{E_{\mathcal{E}}}{E_{\mathcal{S}}}\beta$. Hence the thermalization works in a sense "upside down". For example if all the incoming atoms are excited ($\beta \to -\infty$), then $\beta_* \to +\infty$ and the the small system stays in the ground state.

4 Conclusion

As we explained in Section 2.1, invariant states are important for the long time behavior of the repeated interaction quantum systems. By Theorem 1 (or its improved version, Theorem 4) we suggested a method for obtaining an invariant state of the dynamics when the total Hamiltonian commutes with an observable in a certain form. In case of the perturbed atom-field interaction, it follows from Theorem 3 that we were able to find all the diagonal invariant states using this simple method. Of course, the whole problem of the long time behavior is solved only after proving that we have found all the invariant states, which needs different techniques. On the other hand, our approach may give an insight into the origin of invariant states. This is demonstrated in the example from Section 3.2, where the invariant states are not thermal states of the small system as it was the case in [4], but in both examples the invariant states are generated by the conserved total number operator (8), resp. (15).

In the Section 3.3, we looked for the invariant states for the model of spin-spin interaction only on the basis of the Theorem 1. We were looking for quantities that commute with the total Hamiltonian, *i.e.* we were solving the equation (21). Unfortunately, the solution exists for special choices of the interaction only, hence we have found the invariant states only for these particular examples.

As a task for future, it would be interesting to apply the method from Section 3.3 on more complicated examples. The inverse problem could be also studied, we might try to determine if the invariant states induce a conserved quantity for the dynamics.

References

[1] L. Bruneau, A. Joye, and M. Merkli. *Assymptotics of repeated interaction quantum systems.* Journal of Fuctional Analysis **239** (2006), 310–344.

[2] L. Bruneau, A. Joye, and M. Merkli. *Random repeated interaction quantum systems.* Communications in Mathematical Physics **284** (2008), 553–581.

[3] L. Bruneau, A. Joye, and M. Merkli. *Repeated and continuous interactions in open quantum systems.* Annales Henri Poincaré **10** (2010), 1251–1284.

[4] L. Bruneau and C.-A. Pillet. *Thermal relaxation of a QED cavity.* Journal of Statistical Physics **134** (2009), 1071–1095.

[5] D. Mechede, H. Walther, and G. Müller. *One-atom maser.* Physical Review Letters **54(6)** (1985), 551–554.

Comparison of CPU and CUDA Implementation of Matrix Multiplication*

Vladimír Španihel

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

advisor: František Hakl, Institute of Computer Science, AS CR

Abstract. This paper deals with a comparison of different kinds of matrix-matrix multiplication. Two main approaches are investigate: nonparallel implementation on CPU vs. massively parallel implementation on GPU using NVIDIA CUDA architecture.

On CPU a naive algorithm and a function from scientific library GSL (GNU Scientific Library) are considered against three algorithms on GPU, namely a simple kernel not using shared memory, a kernel using shared memory, and a function from library CUBLAS (CUDA Basic Linear Algebra Subroutines).

It is supposed that the function from CUBLAS will have best performace, and this paper confirms it.

Full version of this contribution has been published in the proceedings of the Doktorandské dny 2012 ÚI AV ČR, 24.–26.9.2012, Jizerka.

Keywords: CPU, CUDA, GPU, Matrix-matrix multiplication

Abstrakt. Tento článek porovnává různé způsoby implementace algoritmu maticového násobení. Porovnáváme dva hlavní přístupy, tj. neparalelní implementaci na CPU oproti masivně paralelním algoritmům na GPU za použití architektury NVIDIA CUDA.

Na straně algoritmů spouštěných na CPU uvažujeme naivní algoritmus a implementaci z knihovny GSL (GNU Scientific Library), zatímco na straně GPU uvažujeme jednoduchý kernel, kernel využívající sdílenou paměť a nakonec implementaci z knihovny CUBLAS (CUDA Basic Linear Algebra Subroutines).

Předpokládá se, že funkce z knihovny CUBLAS bude dosahovat nejlepších výsledků, což tato práce potvrzuje.

Plná verze tohoto příspěvku byla publikována ve sborníku Doktorandské dny 2012 ÚI AV ČR, 24.–26.9.2012, Jizerka.

Klíčová slova: CPU, CUDA, GPU, Maticové násobení

*This work has been supported by the grant No. LG12020 of the Czech Ministry of Education, Youth and Sport.

Orthogonal Polynomials with Discrete Measure of Orthogonality

František Štampach

3rd 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. For certain class of orthogonal polynomials defined via three-recurrence rule, we derive the orthogonality relation formula in terms of a function $\mathfrak{F}$. The definition and basic properties of $\mathfrak{F}$ are summarized in [8, 7, 9]. In the case under investigation, the measure of orthogonality is discrete and is supported by the point spectrum of a Jacobi operator which corresponds to the class of orthogonal polynomials. Further, a new class of orthogonal polynomials related to regular Coulomb wave function is introduced. These polynomials are generalization of the well known Lommel polynomials. Several identities together with a description of the measure of orthogonality for these polynomials are presented.

Keywords: orthogonal polynomials, Jacobi matrix, Lommel polynomials, Coulomb wave function

Abstrakt. Pro jistou třídu rekurentně zadaných orthogonálních polynomů odvodíme tvar míry ortogonality. Formule je popsána pomocí funkce $\mathfrak{F}$, která je definována a studována v [8, 7, 9]. Ve zkoumaném případě je tato míra vždy diskrétní a jejím nosičem je bodové spektrum odpovídajícího Jacobiho operátoru. Dále zavedeme novou třídu ortogonálních polynomů, která souvisí s regulární Coulombovou funkcí. Tato třída je zobecněním dobře známých Lommelových polynomů. Odvodíme pro ni řadu identit, včetně formule pro vztah ortogonality.

Klíčová slova: ortogonální polynomy, Jacobiho matice, Lommelovy polynomy, Coulombova vlnová funkce

1 Introduction

Results of this paper are based on some author's work concerning spectral analysis of certain Jacobi operators that have been published in [8, 7, 9]. As it is well known (see [2, 3]), Jacobi matrices are closely related with the theory of orthogonal polynomials (=OPs), and thus some of our results have consequences also in the theory of OPs.

At the start we recall our main algebraic tool, called function $\mathfrak{F}$, that have already been introduced in the last Doctoral Days proceeding [8] as well as the notion of the characteristic function for a Jacobi operator. Further, an important formula for the Weyl m-function in terms of $\mathfrak{F}$ is stated.

In Section 3, we provide a formula for the characteristic function with concrete choice of a compact Jacobi matrix. This formula involves a function from a decomposition of regular Coulomb wave function that has the same roots with the only possible exception

being 0. Consequently, the spectrum of the Jacobi matrix can be described in terms of nonzero roots of regular Coulomb wave function.

Next, we recall general OPs and the Favard's theorem. We show that any OPs can be expressed by $\mathfrak{F}$ applied on a special truncated sequence. In the second part of Section 4 we state the main theorem that gives a description for the measure of orthogonality in terms of $\mathfrak{F}$ for certain class of OPs.

As an application, we define a new class of OPs in Section 5. These OPs generalize Lommel OPs, the deeply investigated polynomials in the theory of Bessel functions. We reveal several identities for these OPs, mostly involving Coulomb wave functions. Finally, under certain assumption, we provide a description of the respective orthogonality measure which is discrete and supported by the set of reciprocal values of nonzero roots of regular Coulomb wave function.

This text serves as an overview of the author's recent progress in the theory of OPs. Several things are only indicated and longer proofs are omitted. There are still few aspects that wait for completion, however, results presented here will provide a core for a future publication.

2 Preliminaries

In this section we give a review of results which have already been presented in [8] (and with much more details in [9]) and which are essential for nowadays development.

2.1 Main tool

First of all, we extensively use a function, called $\mathfrak{F}$ which have been introduced in [7] for the first time. The definition is as follows,

$$\mathfrak{F}(x) = 1 + \sum_{m=1}^{\infty} (-1)^m \sum_{k_1=1}^{\infty} \sum_{k_2=k_1+2}^{\infty} \dots \sum_{k_m=k_{m-1}+2}^{\infty} x_{k_1} x_{k_1+1} x_{k_2} x_{k_2+1} \dots x_{k_m} x_{k_m+1}. \tag{1}$$

This function is defined on domain

$$D = \left\{ \{x_k\}_{k=1}^{\infty} \subset \mathbb{C}; \sum_{k=1}^{\infty} |x_k x_{k+1}| < \infty \right\}.$$

For a finite number of complex variables we identify $\mathfrak{F}(x_1, x_2, \dots, x_n)$ with $\mathfrak{F}(x)$ where $x = (x_1, x_2, \dots, x_n, 0, 0, 0, \dots)$.

Function $\mathfrak{F}$ possesses many nice algebraic and combinatorial properties. Recall here, for example, the recurrence rule

$$\mathfrak{F}(x) = \mathfrak{F}(x_1, \dots, x_k)\, \mathfrak{F}(T^k x) - \mathfrak{F}(x_1, \dots, x_{k-1}) x_k x_{k+1} \mathfrak{F}(T^{k+1} x), \quad k = 1, 2, \dots \tag{2}$$

which holds for any $x \in D$. T denotes the truncation operator from the left. Other useful identity reads

$$\begin{aligned} &\mathfrak{F}(x_1, x_2, \dots, x_d)\mathfrak{F}(x_2, x_3, \dots, x_{d+s}) - \mathfrak{F}(x_1, x_2, \dots, x_{d+s})\mathfrak{F}(x_2, x_3, \dots, x_d) \\ &= \left(\prod_{j=1}^{d} x_j x_{j+1} \right) \mathfrak{F}(x_{d+2}, x_{d+3}, \dots, x_{d+s}) \end{aligned} \tag{3}$$

where $d, s \in \mathbb{Z}_+$. Formula (3) is a special case of a more general identity, see [9, Subsection 2.3]. By sending $s \to \infty$ in (3), one arrives at the equality

$$\mathfrak{F}(x_1, \ldots, x_d)\mathfrak{F}(Tx) - \mathfrak{F}(x_2, \ldots, x_d)\mathfrak{F}(x) = \left(\prod_{k=1}^{d} x_k x_{k+1}\right) \mathfrak{F}(T^{d+1}x) \tag{4}$$

which is true for any $d \in \mathbb{Z}_+$ and $x \in D$.

2.2 Characteristic function

In [9] we introduce characteristic function F_J for Jacobi matrix J which is given by formula

$$F_J(z) := \mathfrak{F}\left(\left\{\frac{\gamma_n^2}{\lambda_n - z}\right\}_{n=1}^{\infty}\right),$$

provided that there exists $z_0 \in \mathbb{C}$ such that

$$\sum_{n=1}^{\infty} \left| \frac{w_n^2}{(\lambda_n - z_0)(\lambda_{n+1} - z_0)} \right| < \infty. \tag{5}$$

Sequence $\{\gamma_n\}_{n=1}^{\infty}$ is defined recursively by equations $\gamma_1 = 1$ and $\gamma_n \gamma_{n+1} = w_n$. Further $\lambda := \{\lambda_n\}_{n=1}^{\infty} \subset \mathbb{C}$ denotes the diagonal sequence of J, and $w := \{w_n\}_{n=1}^{\infty} \subset \mathbb{C} \setminus \{0\}$ stands for the off-diagonal sequence of J. Hence J has the form

$$J = J(\lambda, w) = \begin{pmatrix} \lambda_1 & w_1 & & & \\ w_1 & \lambda_2 & w_2 & & \\ & w_2 & \lambda_3 & w_3 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

We show in [9] that, under assumption (5), zeros of the characteristic function coincide with the point spectrum of J. More precisely, we prove equalities ([9, Theorem 14])

$$\operatorname{spec}(J) \setminus \operatorname{der}(\lambda) = \operatorname{spec}_p(J) \setminus \operatorname{der}(\lambda) = \mathfrak{Z}(\mathcal{J}) \tag{6}$$

where we denote by

$$\mathfrak{Z}(\mathcal{J}) := \left\{ z \in \mathbb{C} \setminus \operatorname{der}(\lambda);\ \lim_{u \to z} (u - z)^{r(z)} F_J(u) = 0 \right\}, \tag{7}$$

an extended zero set for F_J. Symbol $\operatorname{der}(\lambda)$ stands for the set of all finite accumulation points of the sequence λ and

$$r(z) := \sum_{k=1}^{\infty} \delta_{z,\lambda_k}$$

is the number of members of the sequence λ coinciding with z.

Moreover in [9, Subsection 3.3], we introduce a vector-valued function

$$\xi(z) := (\xi_1(z), \xi_2(z), \xi_3(z), \ldots)$$

where we put

$$\xi_k(z) := \lim_{u\to z}(u-z)^{r(z)}\left(\prod_{l=1}^{k}\frac{w_{l-1}}{u-\lambda_l}\right)\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-u}\right\}_{l=k+1}^{\infty}\right), \quad (w_0 := 1). \tag{8}$$

This function has the property that for $z \in \mathfrak{Z}(J)$ it coincides with the corresponding eigenvector to the eigenvalue z (see [9, Proposition 11]).

Finally, in [9, 8], we express the Green function for J in terms of $\mathfrak{F}$. Especially, for the Weyl m-function $m(z)$ we find

$$m(z) = \frac{\mathfrak{F}\left(\left\{\frac{\gamma_j^2}{\lambda_j-z}\right\}_{j=2}^{\infty}\right)}{(\lambda_1-z)\mathfrak{F}\left(\left\{\frac{\gamma_j^2}{\lambda_j-z}\right\}_{j=1}^{\infty}\right)}, \tag{9}$$

for $z \notin \operatorname{spec}(J)$.

3 Example with regular Coulomb wave function

Recall that regular Coulomb wave function $F_L(\eta,\rho)$ is one of two linearly independent solutions of the second-order differential equation

$$\frac{d^2u}{d\rho^2} + \left[1 - \frac{2\eta}{\rho} - \frac{L(L+1)}{\rho^2}\right]u = 0$$

where $\rho > 0, \eta \in \mathbb{R}$, and $L \in \mathbb{Z}_+$ (see [1, chap. 14]). These ranges for parameters ρ, η, and L are, however, too restrictive and can be generalized. $F_L(\eta,\rho)$ can be decomposed as follows,

$$F_L(\eta,\rho) = C_L(\eta)\rho^{L+1}\phi_L(\eta,\rho),$$

where

$$C_L(\eta) = \sqrt{\frac{2\pi\eta}{e^{2\pi\eta}-1}}\frac{\sqrt{(1+\eta^2)(4+\eta^2)\ldots(L^2+\eta^2)}}{(2L+1)!!L!}$$

and

$$\phi_L(\eta,\rho) = e^{-i\rho}{}_1F_1(L+1-i\eta, 2L+2, 2i\rho),$$

see [1, 14.1.3]. ${}_1F_1$ denotes confluent hypergeometric series.

Let us now consider a concrete Jacobi matrix J with

$$w_n := \frac{\sqrt{(n+1)^2+\eta^2}}{(n+1)\sqrt{(2n+1)(2n+3)}} \quad \text{and} \quad \lambda_n := \frac{\eta}{n(n+1)}. \tag{10}$$

In this case one can compute the characteristic function F_J. For $n \in \mathbb{Z}_+$, $\eta,\rho \in \mathbb{C}$, $\eta\rho \neq -k(k+1)$, $k \geq n+1$, the formula reads

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k+1/\rho}\right\}_{k=n+1}^{\infty}\right) = \frac{\Gamma\left(\frac{3}{2}+n-\frac{1}{2}\sqrt{1-4\eta\rho}\right)\Gamma\left(\frac{3}{2}+n+\frac{1}{2}\sqrt{1-4\eta\rho}\right)}{n!(n+1)!}\phi_n(\eta,\rho). \tag{11}$$

The proof of this equality is based on three-recurrence rule [1, 14.2.3] for $F_L(\eta,\rho)$, and it will be published in a future paper.

By using general results summarized in Subsection 2.2 one finds out the matrix

$$J_L = \begin{pmatrix} -\lambda_{L+1} & w_{L+1} & & & \\ w_{L+1} & -\lambda_{L+2} & w_{L+2} & & \\ & w_{L+2} & -\lambda_{L+3} & w_{L+3} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{12}$$

where $\{\lambda_n\}_{n=L+1}^{\infty}$ and $\{w_n\}_{n=L+1}^{\infty}$ are defined in (10), is a compact operator (since λ and w have zero limit) which nonzero eigenvalues are reciprocals of roots of $\phi_L(\eta,\rho)$. This is the same set as reciprocals of nonzero roots of $F_L(\eta,\rho)$. Hence the spectrum of J_L can be expressed as follows,

$$\operatorname{spec}(J_L) = \{1/\rho : \phi_L(\eta,\rho) = 0\} \cup \{0\} = \{1/\rho : F_L(\eta,\rho) = 0\} \cup \{0\}. \tag{13}$$

Moreover, the formula for the respective eigenvector (multiplied by a constant) to the eigenvalue $1/\rho$ reads

$$v(1/\rho) = \left(\sqrt{2L+3}F_{L+1}(\eta,\rho), \sqrt{2L+5}F_{L+2}(\eta,\rho), \sqrt{2L+7}F_{L+3}(\eta,\rho), \dots\right)^T.$$

These results are, however, known. They have been published by Ikebe in [5].

4 Orthogonal polynomials with discrete measure of orthogonality

4.1 Orthogonal polynomials

Orthogonal polynomials (=OPs) are defined as a family of polynomials $\{P_n\}_{n=1}^{\infty}$ that obey an orthogonality relation

$$\int_{\mathbb{R}} P_n(x)P_m(x)d\mu(x) = \delta_{mn}$$

with respect to a positive measure μ on $\mathbb{R}$. The theory of OPs is deeply developed and there are plenty of books written on the topic. Let us mention at least monographs [2, 3]. Any family of OPs satisfies a three recurrence

$$xP_n(x) = w_{n-1}P_{n-1}(x) + \lambda_n P_n(x) + w_{n+1}P_{n+1}(x) \tag{14}$$

where $\{\lambda_n\}_{n=1}^{\infty}$ is a real sequence and $\{w_n\}_{n=1}^{\infty}$ is a positive sequence (one sets here $P_0 = 0$ and w_0 arbitrary).

However, due to the well known Favard's theorem, the opposite statement is also true. Any family of polynomials that fulfills recurrence (14) forms OPs. OPs are related to $\mathfrak{F}$ through identities

$$P_{n+1}(x) = \prod_{k=1}^{n}\left(\frac{x-\lambda_k}{w_k}\right)\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l - x}\right\}_{l=1}^{n}\right), \quad n = 0, 1\dots, \tag{15}$$

which can be verified by using property (2). Formula (15) determines the solution of (14) with initial conditions $P_0 = 0$ and $P_1 = 1$.

4.2 Orthogonality relation

Having OPs of the first kind $P_n(x)$ defined via recurrence rule, i.e., via identity (15), a crucial question is how does the measure of orthogonality looks like? The following theorem gives the answer for a certain class of OPs.

Theorem 1. *Let (5) holds for some $z_0 \in \mathbb{C}$. Next, let Jacobi operator J be self-adjoint and either J has discrete spectrum or it is an invertible compact operator. Then, for $m, n \in \mathbb{N}$, the orthogonality relation reads*

$$\int_{\mathbb{R}} P_n(x)P_m(x)d\mu(x) = \delta_{mn} \tag{16}$$

where $d\mu$ is purely discrete positive measure supported by the set $\mathfrak{Z}(\mathcal{J})$. The step function $\mu(x)$ has jumps of magnitude

$$\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l - x}\right\}_{l=2}^{\infty}\right)\left[(x-\lambda_1)\frac{d}{dx}\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l - x}\right\}_{l=1}^{\infty}\right)\right]^{-1} \tag{17}$$

at $x \in \mathfrak{Z}(\mathcal{J})$.

Proof. Let $\{e_n : n \in \mathbb{N}\}$, stands for the standard basis in $\ell^2(\mathbb{N})$. Then one easily verifies equality

$$e_n = P_n(J)e_1, \tag{18}$$

holds for any $n \in \mathbb{N}$. The proof proceed by mathematical induction in n.

Further let λ denotes a non-degenerate isolated eigenvalue of J and $E_J(\lambda)$ stands for the corresponding Riezs spectral projection, i.e.,

$$E_J(\lambda) = -\frac{1}{2\pi i}\oint_{|\lambda - z| = \epsilon}(J-z)^{-1}dz$$

where $\epsilon > 0$ such that $\{z \in \mathbb{C} : |\lambda - z| \leq \epsilon\} \cap \operatorname{spec}(J) = \{\lambda\}$. For the Weyl m-function it holds $m(z) = (e_1, (J-z)^{-1}e_1)$. Hence, according to the Residue Theorem, one has

$$(e_1, E_J(\lambda)e_1) = -\frac{1}{2\pi i}\oint_{|\lambda - z| = \epsilon} m(z)dz = -\operatorname{Res}(m, \lambda), \tag{19}$$

since $m(z)$ has a simple pole in λ. Finally, due to identity (9), one can express the residuum as

$$\operatorname{Res}(m,\lambda) = \mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l - \lambda}\right\}_{l=2}^{\infty}\right)\left[(\lambda_1 - \lambda)\frac{d}{dx}\Big|_{x=\lambda}\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l - x}\right\}_{l=1}^{\infty}\right)\right]^{-1}.$$

The rest then follow from the Spectral Theorem applied on the self-adjoint operator J,

$$\delta_{mn} = (e_m, e_n) = (e_1, P_m(J)P_n(J)e_1) = \int_{\mathbb{R}} P_m(\lambda)P_n(\lambda)d(e_1, E_J(\lambda)e_1).$$

□

Remark 2. If J is self-adjoint compact but not invertible, i.e., $0 \in \operatorname{spec}_p(J)$, the step function $\mu(x)$ has one more jump at 0 of magnitude

$$\left(\sum_{n=1}^{\infty}|P_n(0)|^2\right)^{-1}.$$

5 Generalized Lommel polynomials

In this section we introduce a new class of OPs related to the Coulomb wave function that can be viewed as a generalization of Lommel OPs.

5.1 Well known facts on Lommel polynomials

Recall the Lommel "polynomials" arise in the theory of Bessel function (see [10, §9.6-9.73]). They may be given explicitly in the form

$$R_{n,\nu}(x) = \sum_{k=0}^{[n/2]} \binom{n-k}{k} (-1)^k \frac{\Gamma(\nu+n-k)}{\Gamma(\nu+k)} \left(\frac{2}{x}\right)^{n-2k}.$$

One can easily check the identity

$$R_{n,\nu}(x) = \left(\frac{2}{x}\right)^n \frac{\Gamma(\nu+n)}{\Gamma(\nu)} \mathfrak{F}\left(\left\{\frac{x}{2(\nu+k)}\right\}_{k=0}^{n-1}\right) \tag{20}$$

holds for $n = 0, 1, \ldots$. Alternatively, Lommel OPs can be expressed in terms of Bessel functions,

$$R_{n,\nu}(x) = \frac{\pi x}{2} \left(Y_{-1+\nu}(x) J_{n+\nu}(x) - J_{-1+\nu}(x) Y_{n+\nu}(x)\right),$$

or equivalently,

$$R_{n,\nu}(x) = \frac{\pi x}{2\sin(\pi\nu)} \left(J_{1-\nu}(x) J_{n+\nu}(x) + (-1)^n J_{-1+\nu}(x) J_{-n-\nu}(x)\right).$$

Another well known property of Lommel OPs is that they play a role of coefficients in the formula

$$R_{n,\nu}(x) J_\nu(x) - R_{n-1,\nu+1}(x) J_{\nu-1}(x) = J_{\nu+n}(x) \tag{21}$$

where $n \in \mathbb{N}$, $\nu, x \in \mathbb{C}$, see again, for instance, [10, Chp. 9.6].

The explicit orthogonality relation for the Lommel polynomials have been determined in terms of zeros of the Bessel function of order $\nu - 1$, which one can find, for example, in [4]. This relation can be rederived by using Theorem 1, however, we only state the result since we obtain a more general formula below. The orthogonality relation for Lommel OPs reads

$$\sum_{k \in \pm\mathbb{N}} j_{k,\nu}^{-2} R_{n,\nu+1}(j_{k,\nu}) R_{m,\nu+1}(j_{k,\nu}) = \frac{1}{2(n+1+\nu)} \delta_{mn} \tag{22}$$

where $j_{n,\nu}$ denotes the n-th nonzero root of J_ν, $\nu > -1$ and $m, n \in \mathbb{Z}_+$.

5.2 Orthogonal polynomials related to $F_L(\eta, \rho)$

Let us denote $\{P_n^{(L)}(\eta; z)\}_{n=1}^\infty$ OPs given by three-recurrence (14) with coefficients from matrix (12), i.e.,

$$z P_n^{(L)}(\eta; z) = w_{n-1+L} P_{n-1}^{(L)}(\eta; z) - \lambda_{n+L} P_n^{(L)}(\eta; z) + w_{n+L} P_{n+1}^{(L)}(\eta; z)$$

with $P_0^{(L)}(\eta;z) = 0$ and $P_1^{(L)}(\eta;z) = 1$. These polynomials are not included in the Askey-scheme [6]. Further let us denote

$$R_n^{(L)}(\eta;\rho) := P_n^{(L)}(\eta;\rho^{-1})$$

for $\rho \neq 0$, $n \in \mathbb{Z}_+$. According to (15), for $n \in \mathbb{N}$, we have the expression

$$P_n^{(L)}(\eta;z) = \left(\prod_{k=1}^{n-1} \frac{z+\lambda_{k+L}}{w_{k+L}}\right) \mathfrak{F}\left(\left\{\frac{\gamma_{l+L}^2}{\lambda_{l+L}+z}\right\}_{l=1}^{n-1}\right). \tag{23}$$

Alternatively, polynomials can be expressed in terms of Coulomb wave functions,

$$R_n^{(L)}(\eta;\rho) = \frac{\sqrt{(L+1)^2+\eta^2}}{L+1}\left(F_L(\eta,\rho)G_{L+n}(\eta,\rho) - F_{L+n}(\eta,\rho)G_L(\eta,\rho)\right) \tag{24}$$

where $G_L(\eta,\rho)$ is irregular Coulomb wave function (see [1, Chp. 14]). To verify this identity it suffices to check the RHS fulfills the same recurrence rule as $R_n^{(L)}(\eta,\rho)$ (see [1, 14.2.3]) with the same initial conditions. One needs the formula for the Wronskian [1, 14.2.5], which reads

$$F_{L-1}(\eta,\rho)G_L(\eta,\rho) - F_L(\eta,\rho)G_{L-1}(\eta,\rho) = \frac{L}{\sqrt{L^2+\eta^2}}.$$

Further, polynomials $R_n^{(L)}(\eta;\rho)$ can be viewed as a generalization of Lommel polynomials $R_{n,\nu}(x)$ since, by setting $\eta = 0$ and $L = \nu - 1/2$, it holds

$$R_n^{(\nu-1/2)}(0;\rho) = \sqrt{\frac{\nu+n}{\nu+1}} R_{n-1,\nu+1}(\rho) \tag{25}$$

where $n \in \mathbb{N}$ and $\rho \neq 0$.

Next, one obtains a generalization of formula (21) by using identity (4) together with (11) and (23). This formula reads

$$\begin{aligned} &R_{n+1}^{(L-1)}(\eta,\rho)F_L(\eta,\rho) - \sqrt{\frac{2L+3}{2L+1}}\frac{L+1}{L}\frac{\sqrt{\eta^2+L^2}}{\sqrt{\eta^2+(L+1)^2}} R_n^{(L)}(\eta,\rho)F_{L-1}(\eta,\rho) \\ &= \sqrt{\frac{2L+2n+1}{2L+1}} F_{L+n}(\eta,\rho) \end{aligned}$$

where $n \in \mathbb{Z}_+$, $L \in \mathbb{N}$, $\eta \in \mathbb{R}$, $\rho \neq 0$.

Even one more identity is to be presented. By setting $d = L-1$, $s = n$, and

$$x_k = \frac{\gamma_k^2}{\lambda_k + z}$$

into (3) and taking into account (23), one finds the identity

$$P_L^{(0)}(\eta;z)P_{L+n-1}^{(1)}(\eta;z) - P_{L+n}^{(0)}(\eta;z)P_{L-1}^{(1)}(\eta;z) = \frac{w_1}{w_L} P_n^{(L)}(\eta;z) \tag{26}$$

holds for any $n \in \mathbb{Z}_+$.

Finally, we use Theorem 1 to obtain the orthogonality relation for $R_n^{(L)}(\eta, \rho)$ that is the generalization of (22). However, we have to assume J_L to be invertible since it is an assumption of Theorem 1. Till now, we have not been able to determine whether J_L is invertible and if so, for what parameters η and L? This problem still remains open. In the special case of Lommel OPs, i.e., if $\eta = 0$ and $L = \nu - 1/2$, it is quite easy to verify that zero is not an eigenvalue of the corresponding Jacobi matrix by solving respective eigenvalue equations. The zero diagonal simplifies the case significantly.

Thus let us assume J_L to be invertible. Let $\rho_n = \rho_n(\eta, L)$, $n \in \mathbb{N}$ (arbitrarily indexed) stands for roots of $\phi_L(\eta, \rho)$, i.e., nonzero roots of $F_L(\eta, \rho)$. They are infinite (and even simple with no finite accumulation point) by the Hilbert-Schmidt Theorem, for example. According to [1, 14.2.2], regular Coulomb wave function $F_L(\eta, \rho)$ fulfills identity

$$(L+1)\partial_\rho F_L(\eta, \rho) = \left(\frac{(L+1)^2}{\rho} + \eta\right) F_L(\eta, \rho) - \sqrt{(L+1)^2 + \eta^2} F_{L+1}(\eta, \rho). \tag{27}$$

Consequently, one has

$$\partial_\rho \phi_L(\eta, \rho_n) = -\frac{(L+1)^2 + \eta^2}{(2L+3)(L+1)^2} \rho_n \phi_{L+1}(\eta, \rho_n)$$

and the weight function (17) in the orthogonality relation simplifies considerably,

$$\frac{F_{J_{L+1}}(\rho_n^{-1})}{\left(\rho_n^{-1} + \frac{\eta}{(L+1)(L+2)}\right) \frac{\partial}{\partial \rho} F_{J_L}(\rho_n^{-1})} = \frac{(2L+3)(L+1)^2}{(L+1)^2 + \eta^2} \frac{1}{\rho_n^2}. \tag{28}$$

Hence the orthogonality relation now reads

$$\sum_{k=1}^{\infty} \rho_k^{-2} R_n^{(L)}(\eta; \rho_k) R_m^{(L)}(\eta; \rho_k) = \frac{(L+1)^2 + \eta^2}{(2L+3)(L+1)^2} \delta_{mn} \tag{29}$$

where $m, n \in \mathbb{N}$, $\eta \in \mathbb{R}$, and $L \in \mathbb{Z}_+$. By setting $\eta = 0$ and $L = \nu - 1/2$ in (29) and using (25) together with [1, 14.6.6], one easily checks (29) coincides with (22).

References

[1] M. Abramowitz, I. A. Stegun: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, (Dover Publications, New York, 1972).

[2] N. I. Akhiezer: *The Classical Moment Problem and Some Related Questions in Analysis*, (Oliver & Boyd, Edinburgh, 1965).

[3] T. S. Chihara: *An Introduction to Orthogonal Polynomials*, (Gordon and Breach, Science Publishers, Inc., New York, 1978).

[4] D. Dickinson, H. O. Pollak, G. H. Wannier, *On a class of polynomials orthogonal over a denumerable set*, Pacific J. Math. 6, (1956), 239-247.

[5] Y. Ikebe: *The Zeros of Regular Coulomb Wave Functions and of Their Derivatives*, Math. Comp., 29(131), (1975), 878-887.

[6] R. Koekoek, R. F. Swarttouw: *The Askey-scheme of hypergeometric orthogonal polynomials and its q-analogue*, arXiv:math/9602214.

[7] F. Štampach, P. Šťovíček: *On the eigenvalue problem for a particular class of finite Jacobi matrices*, Lin. Alg. App., 434, (2011), 1336-1353

[8] F. Štampach: *Hadamard Type Infinite Products for Regularized Characteristic Function of Jacobi Operator*, Doktorandské dny 2011, sborník ČVUT, (2011).

[9] F. Štampach, P. Šťovíček: *The characteristic function for Jacobi matrices with applications*, preprint, arXiv:1201.1743.

[10] G. N. Watson: *A treatise on the theory of Bessel functions*, Second Edition, (Cambridge University Press, Cambridge, 1944).

Model Considerations for Blind Source Separation of Medical Image Sequences

Ondřej Tichý*

3rd 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. The problem of functional analysis of medical image sequences is studied. The obtained images are assumed to be a superposition of images of underlying biological organs. This is commonly modeled as a Factor Analysis (FA) model. However, this model alone allows for biologically impossible solutions. Therefore, we seek additional biologically motivated assumptions that can be incorporated into the model to yield better solutions. In this paper, we review additional assumptions such as convolution of time activity, regions of interest selection, and noise analysis. All these assumptions can be incorporated into the FA model and their parameters estimated by the Variation Bayes estimation procedure. We compare these assumptions and discuss their influence on the resulting decomposition from diagnostic point of view. The algorithms are tested and demonstrated on real data from renal scintigraphy; however, the methodology can be used in any other imaging modality.

Keywords: Blind Source Separation, Factor Analysis, Convolution, Regions of Interest, Image Sequence

Abstrakt. V příspěvku je studován problém funkční analýzy obrazových sekvencí v medicíně. Získaný obraz je tvořen superpozicí obrázků jednotlivých orgánů ve snímané oblasti, což je typicky modelováno jako model faktorové analýzy, který však v základním tvaru dovoluje biologicky nesmysluplná řešení. Proto je studována možnost zavést do modelu biologicky motivované předpoklady. V tomto příspěvku je uveden přehled dosavadních předpokladů, konkrétně konvolučního modelu časových křivek, automatický výběr oblastí zájmu a analýza šumu. Tyto předpoklady jsou zabudovány do modelu faktorové analýzy, jehož parametry jsou odhadovány pomocí Variační Bayesovy metody. Jednotlivé modely jsou porovnány a je diskutován vliv předpokladů z hlediska diagnostiky. Algoritmy jsou testovány na reálných scintigrafických datech, nicméně mohou být použity i v jiných zobrazovacích modalitách.

Klíčová slova: Slepá Separace, Faktorová Analýza, Konvoluce, Oblasti Zájmu, Obrazová Sekvence

1 Introduction

In many imaging modalities, the original organs are not observed directly but only via observing the activity of radioactive particles and scan of their superposition. In this paper, we are concerned with modalities, where the images are superposed in all observed

*Institute of Information Theory and Automation, Department of Adaptive Systems, AS CR

pictures in the series. The task of source separation is to recover the original images of the biological organs (sources) from the observed images.

One of the first methods of source separation is Factor Analysis (FA). It has been used in functional medical imaging such as scintigraphy, Positron Emission Tomography, or functional Magnetic Resonance Imaging [8]. The factor analysis model is based on a simple assumption that the observed image is a linear combination of the underlying factor image weighted by its time-activity curves. This model is also the basis of other methods, such as the Independent Component Analysis (ICA). The FA and ICA as methods have the same basic model but differ in additional assumptions.

The additional assumptions has potential to change the results significantly. If they are justified for the studied problem, they improve the results of separation. In medical imaging, the additional assumptions are needed to recover biologically meaningful solutions of the separation problem. One of the first additional assumptions was positivity of the images and the time-activity curves [9]. It comes from the physical meaning of measurements of radioactive particles. However, even with this restriction, the model allows for biologically impossible solutions. Therefore, we seek additional assumptions and constraints that restrict the space of possible solutions to those with biological meaning. However, the assumption must be also very general to allow for a great variability that is exhibited by a living body.

All assumptions are translated into parameters of a mathematical model, which needs to be estimated from the data. We are concerned with Bayesian estimation, specifically by an approximate solution provided by the Variational Bayes approximation [11]. It offers a reasonable ratio between possibilities of mathematical modeling and computational difficulties.

2 Mathematical Models

The objective is to analyze a sequence of n images obtained at time $t = 1, \ldots, n$ and stored in vectors $\mathbf{d}_t$ with pixels stacked columnwise. The number of pixels in each image is p, thus $\mathbf{d}_t \in \mathbf{R}^p$. The important assumption is that every observed image is a linear combination of r factor images, stored in vectors $\mathbf{a}_j \in \mathbf{R}^p$, $j = 1, \ldots, r$, using the same order of pixels as in $\mathbf{d}_t$. The dimensions of the problem are typically ordered as $r < n \ll p$. Each factor image has its respective time-activity curve stored in vector $\mathbf{x}_j \in \mathbf{R}^n$, $j = 1, \ldots, r$, $\mathbf{x}_j = [x_{1,j}, \ldots, x_{n,j}]'$, $\mathbf{x}'$ denotes transpose of vector $\mathbf{x}$. With these assumptions, the model of Factor Analysis is:

$$\mathbf{d}_t = \sum_{j=1}^{r} \mathbf{a}_j x_{t,j} + \mathbf{e}_t, \tag{1}$$

where vector $\mathbf{e}_t$ denotes the noise of the t-th observed image. Note that vectors $\mathbf{a}_j$ and $\mathbf{x}_j$, are unknown and must be estimated from measurements $\mathbf{d}_t$ so as the variance of a noise, ω.

For the purpose of medical image analysis we already imposed restrictions on the elements of the probabilistic model of FA (1): (i) all elements of the observed vectors $\mathbf{d}_{t\in 1,\ldots,n}$ are positive, (ii) all elements of the factor images $\mathbf{a}_{j\in 1,\ldots,r}$ and the factor curves $\mathbf{x}_{j\in 1,\ldots,r}$ are also positive, and (iii) the number of relevant factors, r, is unknown. These

assumptions are translated into probabilistic model as follows [11]: the positivity in (i) and (ii) is imposed using truncation of priors of the parameters, i.e. $\mathbf{d}$, $\mathbf{a}$, and $\mathbf{x}$, to the positive numbers; and (iii) the number of factors is estimated using Automatic Relevance Detection (ARD) procedure via hyper-parameters, see [2].

Additional assumptions that are known about the problem are: (i) The time activity curves represent flow of fluids in the human body. The flow is a result of different pressures on the input and output of a biological organ. The output flow is then modeled as convolution of the input flow and convolution kernel of the biological organ. (ii) The biological organ is covers only an area in the full image. When selected manually, these areas are called regions-of-interest. (iii) The noise within the observed image is not isotropic. Good model of the noise properties is required.

These assumptions will be now described as parameters of mathematical models. Discussion of classical methods for their estimation is also provided.

2.1 Regions of Interest

The FA assumption of linear combination (1) are typically not valid over the full size of the images but only in a limited area. This can be modeled by an indicator variable for each pixel of the factor image. Specifically, each pixel of the jth factor, $\mathbf{a}_{i,j}$, has its indicator variable $\mathbf{i}_{i,j}$ which is 1 if the ith pixel belongs to the jth factor and 0 if the ith pixel does not belong to the jth factor. Once again, the indicator variable is unknown and must be estimated from the data.

This task is also standard and the estimation of the indicator variable is known as selection of Regions of Interest (ROI). This is often done manually and it is considered to be a necessary preprocessing step of factor analysis after which it yields much better results [7]. Several automatic and semi-automatic methods were proposed, however, the ROI selection is almost exclusively done by specialists in clinical practice. The incorrect selection of the ROI has significant impact on the following factor analysis. Often, the ROI must be selected iteratively until an acceptable solution is found. This procedure is very time consuming and strongly depends on the experience of specialists and chosen method [4].

2.2 Convolution Model

The assumption that factor curve is a result of convolution of an input function and a kernel is well established [6]. The kernels are organ-specific and are useful in diagnostic parameters estimation [5]. Illustration of the assumption is displayed in Fig. 1.

Mathematically formulated, the time-activity curve of the fth factor, $\mathbf{x}_f$, is modeled as

$$x_{t,f} = \sum_{m=1}^{t} \mathbf{b}_{t-m+1} u_{m,f}, \tag{2}$$

where $\mathbf{b}$ is the input activity, common to all factors, and $\mathbf{u}_f$ is the convolution kernel of the factor. Following [6], we consider the kernel elements $u_{m,t}$ to be decreasing, hence they are modeled by a sum of non-negative increments.

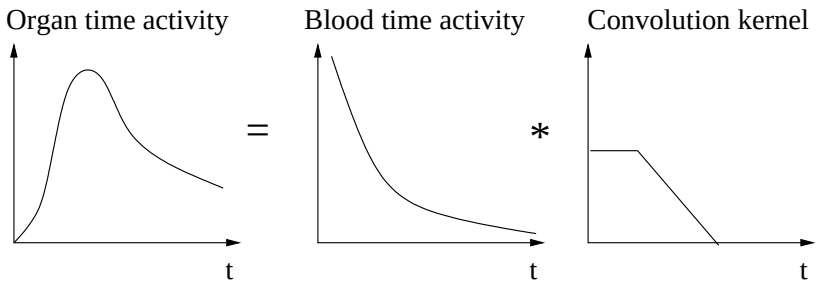


Figure 1: Illustration of assumed shapes of curves in convolution.

Parameters of the model $\mathbf{u}_{j\in 1,\ldots,r}$ and input curve $\mathbf{b}$ are unknown and must be estimated.

Traditional methods of deconvolution are well established method in analysis of dynamic medical image sequences analysis [6]. However, these methods require to know the input curve $\mathbf{b}$ which must be done manually.

2.3 Noise Model

Properties of the noise e_t in (1) determine the quality of separation of the signal. Estimation of the noise properties and its elimination is a crucial step in medical imaging, [3].

The noise may vary across pixels, as well as in time. The noise e_t is assumed to be generated from a Gaussian distribution with zero mean and variance $\sigma_{i,t}$ which may be different for each pixel i and time t. The typical assumption of isotropic noise is $\sigma_{i,t} = \omega^{-1}$, where ω is known as precision. However, it is unrealistic in many modalities. In general, the noise variance is also unknown and should be estimated from the observed data.

Classical methods estimate the noise properties using asymptotic analysis. An example is the correspondence analysis approach [1], where

$$\sigma_{i,t} = \omega^{-1} \sqrt{\sum_{\tau=1}^{n} d_{i,\tau} \sum_{j=1}^{p} d_{j,t}} \tag{3}$$

with unknown precision ω. Correspondence analysis can be interpreted as preprocessing of the data before the factor analysis algorithm.

3 Variational Source Separation

Estimation of parameters of the models described above can be achieved using Bayesian approach. The main advantage of this approach is its ability to determine also the number of relevant factors, r. In such a case, probabilistic formulation of the measurement model (1) must be complemented by prior probabilities of all model parameters. The estimates are obtained by application of the Bayes rule. Exact evaluation of the posterior distribution is however intractable. Therefore, we use an approximate technique known as the Variational Bayes method [11].

We will illustrate the method on the basic model of the factor analysis (1). This model can be written in matrix form $D = AX' + E$, where $D = [\mathbf{d}_1, \ldots, \mathbf{d}_n]$, $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$, and $X = [\mathbf{x}_1, \ldots, \mathbf{x}_r]$. The unknown parameters are matrices A, X and scalar ω. The intractable posterior distribution is

$$f(A, X, \omega|D) = \frac{f(D|A, X, \omega)f(A, X, \omega)}{f(D)}. \tag{4}$$

where $f(A, X, \omega)$ is the prior distribution.

The Variational Bayes approximation is based on restriction of the posterior density to the class of conditionally independent distributions:

$$f(A, X, \omega|D) \equiv f(A|D)f(X|D)f(\omega|D). \tag{5}$$

Under this assumption, necessary conditions for approximate posterior distributions $f(A|D)$, $f(X|D)$, and $f(\omega|D)$ minimizing Kullback-Leibler divergence to the true posterior can be found analytically [11]. The posterior distributions are solutions of a set of implicit equations, typically obtained by an iterative algorithm.

The Variational Bayes method has been applied to the FA model with positivity restrictions in [11], and also extended for unknown noise properties. Extension of the method using the convolution kernels is published in [12]. The Variational solution for the FA model with unknown ROI is presented in [10]. These methods will be now compared on real data and their results will be discussed from diagnostic point of view.

4 Results

The methods will be tested on representative clinical data sets from renal scintigraphy. At first, we briefly describe scintigraphy and biological aspects of dynamics of kidneys. Then, we will discuss the results of the proposed models.

4.1 Renal Scintigraphy

Scintigraphy is a well established and important diagnostic method in nuclear medicine. We are concerned with planar dynamic scintigraphy where the measurements are in the form of a sequence of images of the same scanned region of a body. Each pixel in the sequence is a summation of radioactive particles coming from a whole part of the body under the detector. Therefore, each pixel accumulates activity from potentially many factors. The factors has to be separated using a source separation method such as factor analysis.

A healthy kidney is composed of two main structures, parenchyma and pelvis. There are two important specific properties of a structure and dynamic of these structures: (i) the parenchyma is typically surrounding the whole kidney including the pelvis, and (ii) only the parenchyma is active at the first $100 - 180$ seconds (depending on the patient's state) [5]; this time is called uptake. After the uptake time, the activity passes from parenchyma through pelvis to urinary bladder. Diagnostic parameters related to the uptake time are:

PTT Parenchymal Transit Time (PTT) is the time from the beginning of the sequence to that when pelves are activated.

RRF Relative Renal Function (RRF) can be estimated from an activity in the left (L) and in the right (R) parenchyma as $rel_L = \frac{L}{R+L} \times 100$. Historically, the activity is taken only from the uptake time.

If the assumptions (i) and (ii) are not satisfied, the factor separation is incomplete and could cause significant error in diagnostics. There could be some exceptions in case of abnormal or harmed kidney, this case must be carefully considered by physicians.

4.2 Factor Analysis

The basic model of factor analysis from section 2 was applied to a selected clinical data set from dynamic renal scintigraphy. The sequence is composed of 180 images taken after each 10 seconds. The size of each image is 128×128 pixels.

Four factors were found to be relevant using ARD; however, we shown six factors for following comparison. The results are shown in Fig. 2, on the left side.

The estimates of blood and tissue background, the first and the third factors, are reasonable. The main issue of these results is in a bad separation of parenchyma and pelves, the second factor. There are pelves, dark structures in the inner bound of parenchyma, mixed with the whole parenchyma covering the whole kidneys. Consequently, factor curves of parenchyma and pelves are superposed in this factor too.

Due to the bad separation of the most important structures in our task, we are not able to estimate the PTT.

4.3 Factor Analysis with Regions of Interest

The factor analysis with integrated estimation of regions of interest (FAROI), section 2.1, is applied to the same sequence as in the previous section. The results are shown in Fig. 2, right. The factors are displayed in the same order as in case of the FA.

The main difference between the FA and FAROI algorithms is in separation of parenchyma and pelves. In contrast to the FA algorithm, the FAROI algorithm separated pelves as an independent factor. The assumption of the zero plateau in the beginning of the curve is well satisfied; hence, the diagnostic coefficient PTT could be easily estimated from this result. In this case, $PTT = 130$ seconds.

The second factor, parenchyma, is well separated from pelves; however, the resulting factor image suffer from bad separation from the tissue background. This fact is due to the similar shape of activities of the structures. The sixth factor seems to be an artifact, a residual activity of the urinal process.

We stress that FAROI algorithm, in general, provides comparable or better result then the basic FA algorithm without additional assumptions.

4.4 Factor Analysis with Convolution

The assumption of the convolution model from section 2.2 is not valid for the whole sequence but well satisfied for the uptake part of a sequence, where only blood, parenchyma,

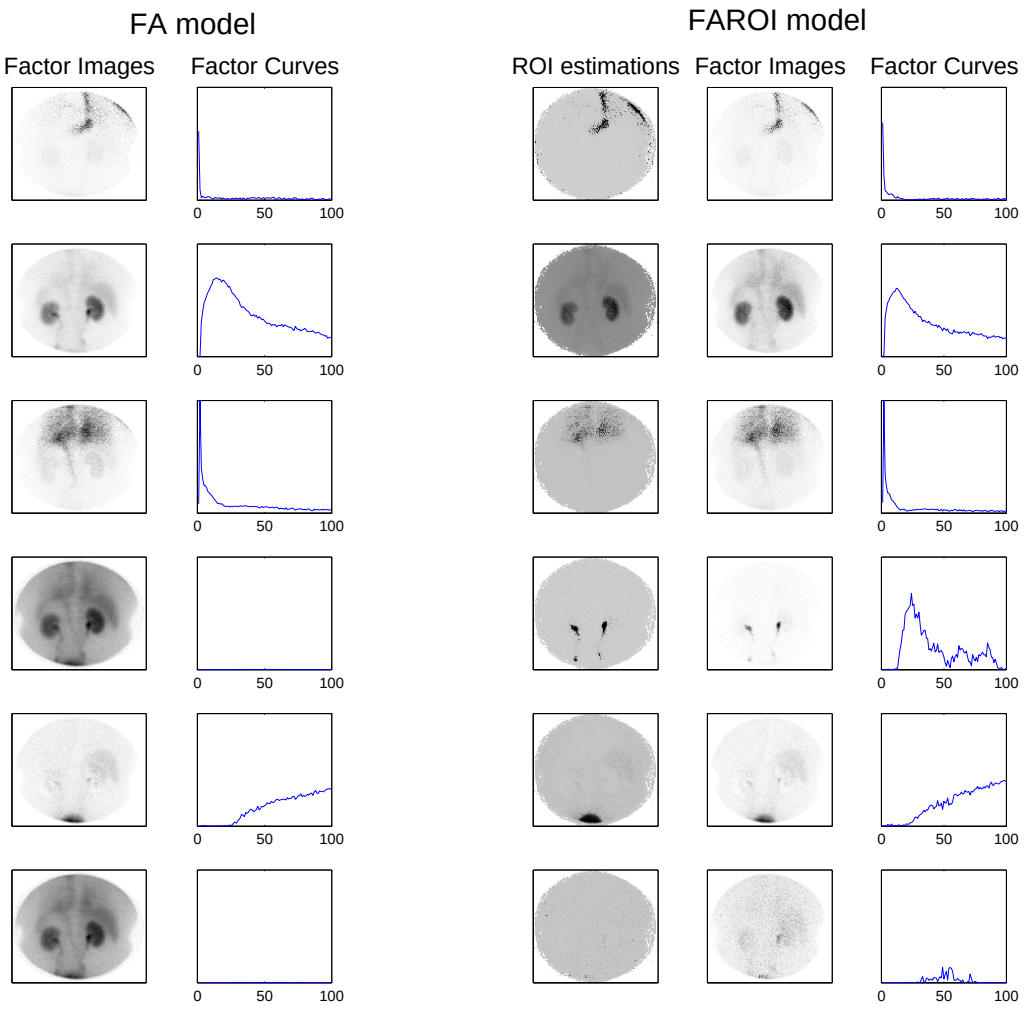


Figure 2: Results from the FA (left) and FAROI (right) models. In the case of FA model, there are (from the top): heart, parenchyma mixed with pelves, lungs and tissue background, dummy factor, urinary bladder, and dummy factor. Estimated factor images are in the first column and estimated factor curves are in the second column. Results from the FAROI algorithm, section II.A., are in the right. There are (from the top): heart, parenchyma, lungs and tissue background, pelves, urinary bladder, and tissue artifact. Estimated parameters are: ROI in the left column, factor images in the middle column, and factor curves in the right column.

and tissue background are activated. This limitation is due to the assumed shape of the convolution kernel of biological structures. The shape in Fig. 1, right, is valid only for structures activated from the beginning of the sequence, e.g. not for the pelves and urinary bladder. Hence, we applied the FA combined with convolution model of factor curves (CFA) only on uptake part of the sequence. The number of images in the uptake part can be estimated using FA or FAROI algorithms automatically. This task is very important part of diagnosis. Here, the parenchyma should be separated from the blood and the tissue backgrounds. After that, the Relative Renal Function (RRF) can be estimated, see section 4.1.

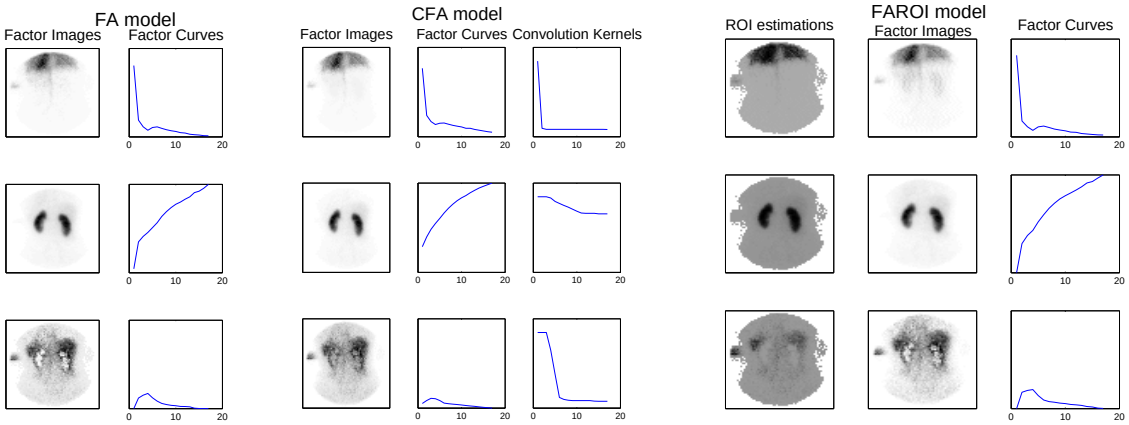


Figure 3: Results from the FA (left), CFA (middle), and FAROI (left) models are shown on the uptake part of the sequence (data set IM3). Estimative procedures estimated in each case three factors (from the top): blood background, parenchyma, and tissue background. In columns are shown (from the left to the right): FA: factor images and factor curves; CFA: factor images, factor curves, and estimated convolution kernels; FAROI: estimated ROI, factor images, and factor curves.

Table 1: Comparison of estimates of RRF coefficient of the left kidney obtained by expert, FA, CFA and FAROI algorithms.

data	expert	FA	CFA	FAROI
IM1	28%-31%	34%	29%	30%
IM2	69%-76%	93%	75%	81%
IM3	48%-51%	48%	49%	49%

The RRF determination is typically performed by an expert using various sets of tools including manually ROI selection, deconvolution, or FA. For our experiment, we roughly selected rectangular ROI around the kidneys and then ran the FA, CFA, and FAROI algorithms on this narrow sequences.

We applied the CFA model on three selected clinical data sets from renal scintigraphy: one set with healthy kidneys (IM3) and two data sets with pathological kidneys (IM1 and IM2). The sequences are composed of images taken after every 10 seconds. Here, the size of each image is 64×64 pixels.

Results of the methods are shown in Tab. 1. For the healthy kidneys (data set IM3), all methods provide comparable estimates corresponding to expert values. Results are different in the case of pathological kidneys (data sets IM1 and IM2). Here, the CFA algorithm provides more reasonable results then the FA and FAROI algorithms due to better background separation from parenchyma, especially for very harmed kidneys (e.g. data set IM2).

An example of results of the algorithms is shown in Fig. 3. For illustration, there are shown results from the whole images, not only for rectangular parts. The ARD procedures estimated in each case three factors. Factor curves are slightly different and

as we can see on comparison of the second factor, the activity of parenchyma by the CFA algorithm suffer from the non-zero start. It is caused by inaccurate parametrization of the convolution kernels, Fig. 1. Factor images are comparable; however, a difference is in separation of parenchyma from tissue backgrounds. The background activity is well estimated by the CFA algorithm in contrast to the FA or FAROI algorithms where the activity is slightly oversubstracted.

A comparison of the FA and CFA algorithms was given in [12]. Generaly, the CFA algorithm provides more relevant estimations of the RRF coefficient then the FA algorithm due to the better separation of parenchyma and blood background. The FAROI algorithm gives promising results, the estimates of the RRF is close to that from an expert; however, the issue with background separation is still not corrected. Note that the difference between the algorithms is more significant especially by harmed kidneys.

4.5 Notes on Noise Estimation

Correspondence analysis from section 2.3 is used in presented algorithms as a preprocessing step. Without this step, there are incorrectness of the background separation.

Various method for online noise-parameters estimation were studied [11]; however, the results are not so different from the used correspondence analysis on typical data sets. Hence, we recommend it for its reasonable results and computational low cost.

5 Conclusion

In this contribution, we summarize various extensions of the model of the factor analysis (FA) for medical image sequences analysis. The extensions of noise, the convolution assumption, and the regions of interest estimation were studied. It is shown that factor analysis provides more physiologically reasonable results with additional, biologically-motivated, extensions.

We discussed the estimation of two diagnostic parameters: parenchymal transit time (PTT) and relative renal function (RRF). For the purpose of PTT estimation, we compared the basic model of FA and the model of FA with regions of interest estimation (FAROI). The FAROI algorithm provides more biologically reasonable results then the FA algorithm. The main difference can be seen on separation of parenchyma and pelves where the FAROI outperforms the FA algorithm. In the case of RRF estimation, we compared FA, FA with convolution (CFA), and FAROI algorithms with estimates provided by an expert. It is shown that the results are similar for healthy kidneys; however, the CFA algorithm provides better results then the other methods on harmed kidneys. Note that all proposed algorithms exploit correspondence analysis as a preprocessing step and automatic relevance determination for significant factors selection. Moreover, we stress that all proposed procedures provide results automatically, without excessive intervention of an expert.

The models were tested on the data from renal scintigraphy; however, the resulting algorithms can be applied in other imaging modalities.

References

[1] H. Benali, I. Buvat, F. Frouin, J. Bazin, and R. Paola. *A statistical model for the determination of the optimal metric in factor analysis of medical image sequences (famis).* Physics in medicine and biology **38** (1993), 1065.

[2] C. Bishop and M. Tipping. *Variational relevance vector machines.* In 'Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence', 46–53. San Francisco: Morgan Kaufmann Publishers, (2000).

[3] R. Boellaard, N. Krak, O. Hoekstra, and A. Lammertsma. *Effects of noise, image resolution, and roi definition on the accuracy of standard uptake values: a simulation study.* Journal of Nuclear Medicine **45** (2004), 1519–1527.

[4] M. Caglar, G. Gedik, and E. Karabulut. *Differential renal function estimation by dynamic renal scintigraphy: influence of background definition and radiopharmaceutical.* Nuclear medicine communications **29** (2008), 1002.

[5] 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).

[6] A. Kuruc, J. Caldicott, and S. Treves. *Improved Deconvolution Technique for the Calculation of Renal Retention Functions.* COMP. AND BIOMED. RES. **15** (1982), 46–56.

[7] G. Liney, P. Gibbs, C. Hayes, M. Leach, and L. Turnbull. *Dynamic contrast-enhanced mri in the differentiation of breast tumors: User-defined versus semi-automated region-of-interest analysis.* Journal of Magnetic Resonance Imaging **10** (1999), 945–949.

[8] R. Reyment. *Applied factor analysis in the natural sciences.* Cambridge University Press, (1997).

[9] M. Šámal, M. Kárný, H. Surová, E. Maříková, and Z. Dienstbier. *Rotation to simple structure in factor analysis of dynamic radionuclide studies.* Physics in medicine and biology **32** (1987), 371.

[10] V. Šmídl and O. Tichý. *Automatic Regions of Interest in Factor Analysis for Dynamic Medical Imaging.* In '2012 IEEE International Symposium on Biomedical Imaging (ISBI)'. IEEE, (2012).

[11] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing.* Springer, (2006).

[12] V. Šmídl, O. Tichý, and M. Šámal. *Factor analysis of scintigraphic image sequences with integrated convolution model of factor curves.* In 'Proceedings of the second international conference on Computational Bioscience'. IASTED, (2011).

Autoregressive Models in Alzheimer's Disease Classification from EEG*

Lucie Tylová

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

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

Abstract. Fluctuation of EEG signal is an useful symptom of EEG quasi-stationarity. Linear predictive models of three types and their prediction error are studied via traditional and robust measures. Resulting EEG characteristics are applied to diagnosis of Alzehimer's disease. The aim is to decide between: foward, backward, and predictive models, EEG channels, and also robust and non-robust variability measures, and then to find statistically significant measures, which should be useful in Alzheimer's disease classification from EEG.

Keywords: Alzheimer's disease, EEG, linear predictive model, quasi-stationarity, robust statistics, multiplestesting, FDR.

Abstrakt. Fluktuace signálu EEG je užitečným příznakem EEG kvazistacionarity. Pomocí tradičních a robustních měr jsou studovány lineární prediktivní modely tří typů a jejich chyba predikce. Výsledné charakteristiky EEG jsou aplikovány v diagnostice Alzheimerovy choroby. Cílem je rozhodnout se mezi: dopřednou predikcí, zpětnou predikcí a vyhlazováním spolu s výběrem EEG kanálů a mezi robustními a nerobustními mírami. Pak je třeba najít statisticky signifikantní míry, které by mohly být užitečné při klasifikaci Alzheimerovy choroby na základě EEG.

Klíčová slova: Alzheimerova choroba, EEG, lineární prediktivní model, kvazistacionarita, robustní statistiky, mnohonásobné testování, FDR.

1 Introduction

Biological rest is an endogenously dynamic process. Transient EEG events identify and quantify brain electric microstates as time epochs with quasi-stable field topography. We can hypothesised better predictability inside microstates, lower predictability during changes between microstates. Higher fluctuations of the EEG predicability may be connected with higher frequency of microstates changes.

2 Models

The main hypotesis of this work is that predictability of brain activity differs between groups of patients with Alzheimer's disease (AD) and normal controls (CN). The ac-

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

tivity of human brain is measured via multichannel EEG which produces time series. Respecting the quasi-stationarity of EEG signal, the time series were decomposed into nonoverlaping segments of constant length. Every segment of given EEG channel and individual patient produced a short time series whose properties were studied via linear autoregressive models of three types.

2.1 Predictive model

Let m and n be length of segment and model size as number of parameters respectively. Let x_1, ..., x_m be EEG [1] data segment. The linear predictive model has the form

$$x_k = \sum_{i=1}^{n} a_i x_{k-i} + e_k, \tag{1}$$

for $k = n + 1$, ..., m where e_k is model error in k-th measurement and a_i is model parameter for $i = 1$, ..., n. Formula (1) represents traditional AR (autoregressive) model [2].

2.2 Back-predictive model

The predictive AR model (1) can be also used in opposite time direction. The resulting model is

$$x_k = \sum_{i=1}^{n} a_i x_{k+i} + e_k, \tag{2}$$

where e_k is again the model error but for $k = 1$, ..., $m - n$.

2.3 Symmetric model

The third AR model is symmetric and thus with lower prediction error for smooth signals. Supposing n is even, the adequate model is

$$x_k = \sum_{i=1}^{n/2} a_i x_{k-i} + \sum_{i=1}^{n/2} a_{n/2+i} x_{k+i} + e_k, \tag{3}$$

where e_k is model error for $k = n/2 + 1$, ..., $m - n/2$.

2.4 Model error

The three AR models above are easily comparable because they produce an overdetermined system of $M = m - n$ linear equations for n unknown variables a_1, ..., a_n. The unknown parameters a_1, ..., a_n were estimated by the method of least squares (LSQ) [3] and the residues r_1, ..., r_M are determined. The estimate of prediction error inside given segment is

$$s_{\mathrm{e}} = \sqrt{\frac{\sum_{i=1}^{M} {r_i}^2}{M - n}}. \tag{4}$$

3 Fluctuation of model error

Three basic characteristics were used to characterize EEG fluctuations: standard deviation (STD), mean of absolute differences from mean value (MAD_1), and mean of absolute differences from median value (MAD_2), which are too sensitive to outlier values. We preferred robust measures of EEG fluctuations: median of absolute differences from median (MAD_3), interquartile range (IQR), and first quartile of absolute mutual differences (MED).

Let N be the number of EEG signal segments. Let $\mathsf{s} = (s_1, s_2, ..., s_N)$ be vector of errors [4] in all segments. Let Q_1, Q_2, Q_3, E be the first, second, and third quartile and mean value functions. The fluctuation criteria are defined as

$$STD = (\mathrm{E}(\mathsf{s} - \mathrm{E}(\mathsf{s}))^2)^{1/2} \tag{5}$$

$$MAD_1 = \mathrm{E}(|\mathsf{s} - \mathrm{E}(\mathsf{s})|) \tag{6}$$

$$MAD_2 = \mathrm{E}(|\mathsf{s} - \mathrm{Q}_2(\mathsf{s})|) \tag{7}$$

$$MAD_3 = \mathrm{Q}_2(|\mathsf{s} - \mathrm{Q}_2(\mathsf{s})|) \tag{8}$$

$$IQR = \mathrm{Q}_3(\mathsf{s}) - \mathrm{Q}_1(\mathsf{s}) \tag{9}$$

$$MED = \mathrm{Q}_1(|s_i - s_j|). \tag{10}$$

We obtained STD, MAD_1, MAD_2, MAD_3, IQR, and MED values of model fluctuations of every channel for all AD and CN patients. Null hypothesis H_0: $\mu_{\mathrm{AD}} = \mu_{\mathrm{CN}}$ was tested via two-sample t-test [4] against alternative H_A: $\mu_{\mathrm{AD}} \neq \mu_{\mathrm{CN}}$. Here, $\mu_{\mathrm{AD}} = \mathrm{E}\ \ln$ *fluctuation* (5-10) for AD group and $\mu_{\mathrm{CN}} = \mathrm{E}\ \ln$ *fluctuation* (5-10) for CN group.

4 Experimental part

Groups of 26 AD and 139 CN patients were used for testing. Every patient was measured on 19 channels with sampling frequency 200 Hz. Predictive model (1), back-predictive model (2), and symmetric model (3) were identified and model errors (4) and their fluctuations were studied for $m = 150$, $n = 50$. The number of EEG segments varies patient by patient and satistisfies the inequality $352 \leq N \leq 762$.

The testing was performed on significance level $\alpha = 0.001$. The hypoteses of mean equity were tested on 19 EEG channels, three predictive models, and six fluctuation characteristics. It is a kind of multiple testing with 342 potentially dependent tests. The standard methodology of False Discovery Rate (FDR) [5] was used to eliminate the false hypothesis acceptance.

The corrected critical value was determined as $\alpha_{\mathrm{FDR}} = 4.8347{\times}10^{-6}$. The numerical results are collected in Tabs. 1, 2, 3. The results show p-values of all three models which

describe ability to separete AD and CN patients. The hypothesis was rejected only on channels 2, 3, 4 which correspond to the frontal domain of human brain. Only three fluctuation characteristics are significant: ln MAD_3, ln IQR, and ln MED. The best $p_{\text{value}} = 1.8885 \times 10^{-7}$ was obtained on the third channel for symmetric model and ln MAD_3 criterion.

The second channel is significant only for ln MED or symmetrical prediction. The third channel is significant only for ln MED, ln MAD_3 or symmetric prediction. The fourth channel is significant only for ln MAD_3 together with symmetrical prediction. These results are collected in Tab. 4.

5 Discussion

While autoregressive model is linear and require stationary signal, higher fluctuation of model error in Alzheimer's subject may reflect different structure of brain microstates comparing healthy subjects. It may reflect alterations in brain anatomical cortical connectivity in resting-state networks.

6 Conclusion

Using the symmetric predictive model of EEG signal and robust measures MAD_3, IQR, and MED of predictive error fluctuations, I recognize significant differences between AD and CN groups in the case of frontal electrodes, which are represented by second, third, and fourth channel of EEG. This result is directly aplicable to the diagnosis of Alzheimer's disease.

References

[1] E. Niedermeyer, F. Lopes da Silva. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, Lippincott Williams & Wilkins, 2005.

[2] M. B. Priestley. *Non-linear and Non-stationary Time Series Analysis*, Academic Press, 1988.

[3] A. Björck. *Numerical Methods for Least Squares Problems*, SIAM, 1996.

[4] M. Meloun, J. Militky. *The statistical analysis of experimental data*, Academia, 2004.

[5] Y. Benjamini, Y. Hochberg. *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* In 'Journal of the Royal Statistical Society', Vol.57, No.1, 1995, 289–300.

[6] T. Fawcett. *An Introduction to ROC Analysis.* In 'Pattern Recognition Letters', Vol.27, No.8, 2006, 861–874.

[7] D. R. Anderson, D. J. Sweeney, T. A. Williams. *Introduction to Statistics: Concepts and Applications*, West Group, 1994.

[8] H. Laufs, K. Krakow, P. Sterzer, E. Eger, A. Beyerle, A. Salek-Haddadi, A. Kleinschmidt. *Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest.* In 'PNAS', Vol.100, No.19, 2003, 11053-11058.

[9] H. Laufs. *Endogenous brain oscillations and related networks detected by surface EEG-combined fMRI.* In 'Hum Brain Mapp', Vol.29, No.7, 2008, 762–769.

[10] F. Musso, J. Brinkmeyer, A. Mobascher, T. Warbrick, G. Winterer. *Spontaneous brain activity and EEG microstates. A novel EEG/fMRI analysis approach to explore resting-state networks.* In 'Neuroimage', Vol.52, No.4, 2010, 1149–1161.

Table 1: Separation ability (p-value) of predictive model

	Traditional			**Robust**		
Ch	**STD**	**MAD$_1$**	**MAD$_2$**	**MAD$_3$**	**IQR**	**MED**
1	0.1027	0.0402	0.0314	0.0029	0.0265	0.0018
2	0.0065	0.0016	6.52×10^{-4}	6.9×10^{-6}	6.2×10^{-5}	4.8×10^{-6}
3	0.0121	0.0038	0.0014	3.5×10^{-6}	6.5×10^{-5}	3.5×10^{-6}
4	0.1408	0.0612	0.0337	2.4×10^{-4}	0.0019	2.2×10^{-4}
5	0.2551	0.1906	0.1277	0.0017	0.0124	0.0022
6	0.0643	0.0476	0.0275	2.9×10^{-4}	0.0047	4.1×10^{-4}
7	0.0279	0.0192	0.0103	2.4×10^{-4}	0.0025	1.9×10^{-4}
8	0.0917	0.1619	0.1290	0.0478	0.0787	0.0390
9	0.1780	0.2093	0.1512	0.0159	0.0490	0.0127
10	0.6823	0.8572	0.8429	0.8614	0.6785	0.7914
11	0.2358	0.1763	0.1203	0.0038	0.0227	0.0034
12	0.0910	0.0598	0.0467	0.0054	0.0182	0.0082
13	0.1183	0.2376	0.1806	0.0177	0.0722	0.0212
14	0.1027	0.1964	0.1744	0.0713	0.1341	0.0873
15	0.2297	0.2925	0.2539	0.0877	0.1351	0.0791
16	0.4478	0.5942	0.5282	0.2547	0.2882	0.2583
17	0.0680	0.1197	0.1094	0.0307	0.0740	0.0359
18	0.0418	0.0634	0.0595	0.0511	0.1400	0.0317
19	0.2875	0.3889	0.3288	0.0491	0.1666	0.0492

Table 2: Separation ability (p-value) of back-predictive model

	Traditional			**Robust**		
Ch	**STD**	**MAD$_1$**	**MAD$_2$**	**MAD$_3$**	**IQR**	**MED**
1	0.0711	0.0338	0.0270	0.0035	0.0236	0.0021
2	0.0031	0.0012	5.06×10^{-4}	1.0×10^{-5}	5.1×10^{-5}	4.8×10^{-6}
3	0.0141	0.0035	0.0013	1.7×10^{-6}	4.2×10^{-5}	3.7×10^{-6}
4	0.1470	0.0540	0.0308	6.6×10^{-4}	0.0026	3.4×10^{-4}
5	0.2573	0.1690	0.1141	0.0038	0.0170	0.0025
6	0.0647	0.0391	0.0223	2.9×10^{-4}	0.0039	3.2×10^{-4}
7	0.0232	0.0166	0.0086	1.7×10^{-4}	0.0019	1.7×10^{-4}
8	0.0947	0.1474	0.1152	0.0495	0.0679	0.0406
9	0.1815	0.1797	0.1308	0.0130	0.0387	0.0123
10	0.7739	0.8309	0.8136	0.8522	0.6462	0.7958
11	0.2218	0.1540	0.1046	0.0021	0.0151	0.0031
12	0.0924	0.0545	0.0446	0.0066	0.0201	0.0121
13	0.0953	0.2120	0.1607	0.0201	0.0730	0.0219
14	0.1114	0.1779	0.1595	0.0676	0.1056	0.0885
15	0.2363	0.2521	0.2174	0.0581	0.0994	0.0631
16	0.4009	0.5395	0.4806	0.2338	0.2464	0.2512
17	0.0437	0.1070	0.0965	0.0277	0.0676	0.0338
18	0.0545	0.0694	0.0654	0.0451	0.1313	0.0375
19	0.2483	0.3431	0.2868	0.0448	0.1625	0.0439

Table 3: Separation ability (p-value) of symmetric model

	Traditional			**Robust**		
Ch	**STD**	**MAD$_1$**	**MAD$_2$**	**MAD$_3$**	**IQR**	**MED**
1	0.0503	0.0131	0.0074	2.6×10^{-4}	0.0025	2.3×10^{-4}
2	0.0015	2.16×10^{-4}	6.00×10^{-5}	3.9×10^{-7}	3.0×10^{-6}	5.1×10^{-7}
3	0.0172	0.0010	2.24×10^{-4}	1.8×10^{-7}	1.6×10^{-6}	3.0×10^{-7}
4	0.0635	0.0081	0.0029	4.8×10^{-6}	4.6×10^{-5}	9.1×10^{-6}
5	0.2063	0.0867	0.0441	1.4×10^{-4}	0.0015	2.1×10^{-4}
6	0.0417	0.0165	0.0064	2.4×10^{-5}	2.4×10^{-4}	3.2×10^{-5}
7	0.0288	0.0214	0.0091	2.9×10^{-4}	0.0014	2.6×10^{-4}
8	0.0572	0.0908	0.0664	0.0174	0.0384	0.0204
9	0.1862	0.0832	0.0504	0.0013	0.0066	0.0023
10	0.6093	0.6226	0.5553	0.3281	0.2948	0.3613
11	0.1234	0.0527	0.0255	1.8×10^{-4}	0.0015	2.4×10^{-4}
12	0.0359	0.0285	0.0216	0.0051	0.0100	0.0083
13	0.0997	0.1113	0.0602	7.0×10^{-4}	0.0085	6.9×10^{-4}
14	0.0706	0.0827	0.0558	0.0040	0.0180	0.0053
15	0.1673	0.1517	0.0985	0.0028	0.0139	0.0050
16	0.3636	0.3136	0.2170	0.0131	0.0368	0.0195
17	0.0288	0.0304	0.0175	4.3×10^{-4}	0.0038	4.7×10^{-4}
18	0.0296	0.0299	0.0219	0.0032	0.0257	0.0033
19	0.1506	0.1568	0.1025	0.0017	0.0253	0.0021

Table 4: Significant channels

	MAD$_3$	**IQR**	**MED**
Predictive	3		2, 3
Back-predictive	3		2, 3
Symmetric	2, 3, 4	2, 3	2, 3

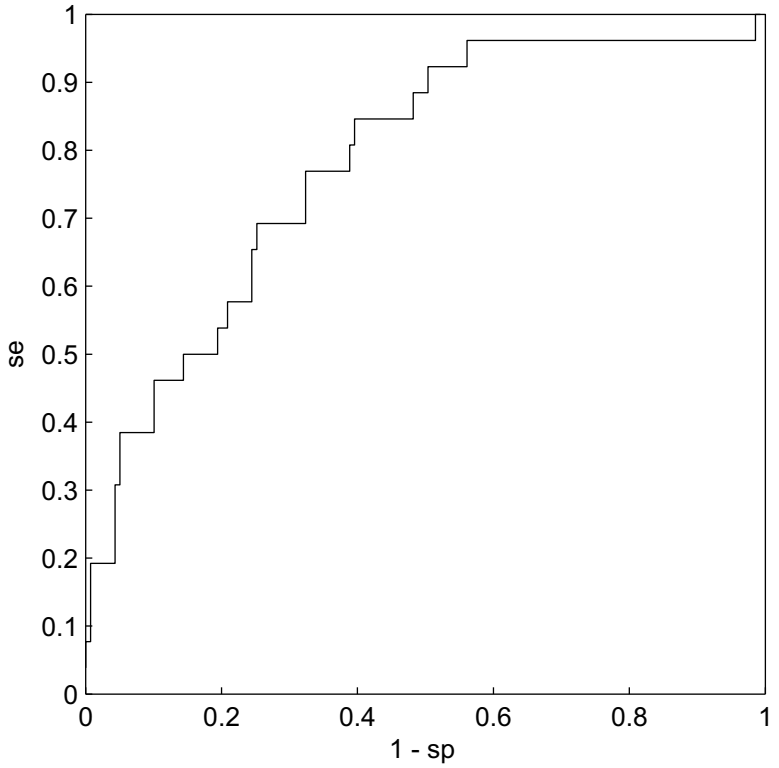


Figure 1: ROC for ln MAD_3

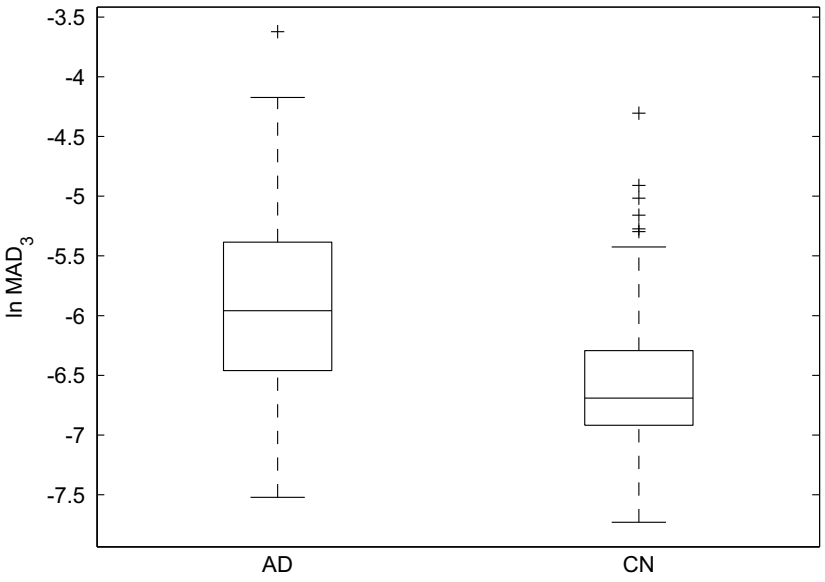


Figure 2: Wishart diagram for ln MAD_3

On Necessary and Sufficient Conditions for Near-Optimal Singular Stochastic Controls*

Petr Veverka (joint work with M. Hafayed and S. Abbas)†

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, Faculty of Mathematics and Physics, CU in Prague

Abstract. This document is an extended abstract to the paper *On Necessary and Sufficient Conditions for Near-Optimal Singular Stochastic Controls* by M. Hafayed, S. Abbas and P. Veverka published in Optimization Letters, Springer; ISSN: 1862-4480, April 2012. In this document this original paper will be refered to just as 'the paper'.
In the paper we discuss the necessary and sufficient conditions for near-optimal singular stochastic controls for the systems driven by a nonlinear stochastic differential equations (SDEs in short). It is well known that optimal singular controls may fail to exist even in simple cases. This justifies the use of near-optimal singular controls, which exist under minimal conditions and are sufficient in most practical cases. Moreover, since there are many near-optimal singular controls, it is possible to choose suitable ones, that are convenient for implementation. This result is a generalization of Zhou's stochastic maximum principle for near-optimality to singular control problem.

Keywords: Near-optimal singular stochastic control, Maximum principle, Necessary and sufficient conditions, Ekeland's variational principle.

Abstrakt. Tento článek je pouze rozšířeným abstraktem ke článku s názvem *On Necessary and Sufficient Conditions for Near-Optimal Singular Stochastic Controls* autorů M. Hafayed, S. Abbas a P. Veverky vydaného v časopise Optimization Letters, Springer; ISSN: 1862-4480, v dubnu 2012. V tomto původním článku jsou zkoumány nutné a postačující podmínky pro přibližnou optimalitu řešení stochastické úlohy singulárního řízení. Jsou dobře známé příklady kdy optimální řízení nemusí existovat dokonce ani v jednoduchých případech. Naproti tomu tzv. přibližně-optimální řízení existuje vždy (je jich vlastně nekonečně mnoho) a tyto kandidáty lze dokonce volit z nějaké vhodné třídy řízení, což může být výhoda pro numerické implementace. Z pohledu praxe je navíc přibližně-optimální řízení vždy postačující. Uvedený výsledek je zobecnění klasického výsledku pro spojité difuzní procesy od X.Y.Zhou-a.

Klíčová slova: Přibližně-optimalní metoda pro singulární stochastickou úlohu řízení, Princip maxima, Nutná a postačující podmínka pro přibližnou-optimalitu, Ekelandův variační princip.

*This work has been supported by Algerian PNR project grant 08/u07/857, Czech CTU grant SGS 2012-2014 and MSMT grant INGO II INFRA LG12020.

†Lab. of Applied Mathematics, Mohamed Khider University, Biskra, Algeria; School of Basic Sciences, Mandi, India.

1 Assumptions and statement of the problem

Singular stochastic control problem is an important and challenging class of problems in control theory. It appears in various fields like mathematical finance (where, for example, it allows to formulate in an elegant way the problem of optimal consumption and portfolio selection with proportional transaction costs), physical models etc. Stochastic maximum principle for singular controls was considered by many authors (for the survey of results see the paper).

The main objective of the paper is to establish necessary as well as sufficient conditions for near-optimal singular control for SDEs where the control domain is not necessarily convex. These conditions are given in terms of second-order adjoint processes corresponding to the controlled SDEs and nearly maximum conditions on the Hamiltonian function. Moreover in a second step, we prove that under additional concavity condition on the Hamiltonian function, these necessary conditions of near-optimality are also sufficient.

In the paper, the singular stochastic control problem for the systems governed by nonlinear controlled diffusion of the following type is considered

$$\begin{cases} dx_t = f(t, x_t, u_t)\, dt + \sigma(t, x_t, u_t)\, dW_t + G_t d\eta_t, \ t \in [s, T] \\ x_s = y, \end{cases} \tag{1}$$

where $T > 0$ is a fixed time horizon, $y \in \mathbb{R}^n$, $(W_t)_{t\in[s,T]}$ is a standard $\mathbb{R}^l-$valued Brownian motion starting at some fixed time $s \in [0, T]$ defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[s,T]}, \mathbb{P})$ satisfying the usual conditions.

The set of admissible controls is defined as follows. Let $\mathbb{A}_1$ be a closed convex subset of $\mathbb{R}^m$ and $\mathbb{A}_2 := ([0, \infty))^m = (\mathbb{R}_+)^m$ for $m \in \mathbb{N}$. We denote the set of stochastic processes

$$\mathbb{U}_1([s, T]) = \{u : [s, T] \times \Omega \to \mathbb{A}_1 | u \text{ is jointly measurable and } \mathcal{F}_t - \text{adapted}\},$$
$$\mathbb{U}_2([s, T]) = \{\eta : [s, T] \times \Omega \to \mathbb{A}_2 | \eta \text{ is jointly measurable and } \mathcal{F}_t - \text{adapted}\}.$$

Definition 1. *An admissible control is a pair* $(u_t, \eta_t)_{t\in[s,T]} \in \mathbb{U}_1([s, T]) \times \mathbb{U}_2([s, T])$ *such that*

1. $\eta(\cdot)$ *is of bounded variation, nondecreasing, continuous on the left with right limits and* $\eta_s = 0$.

2. $\mathbb{E}\left[\sup_{t\in[s,T]} |u_t|^2 + |\eta_T|^2\right] < +\infty$.

The set of all admissible controls is denoted as $\mathbb{U}([s, T])$.

Since $d\eta_t$ may be singular with respect to Lebesgue measure dt, we call $\eta(\cdot)$ the singular part of the control and the process $u(\cdot)$ its absolutely continuous part.

Further, we denote by $\mathbb{L}^2_{\mathcal{F}}([s,T];\mathbb{R}^n)$ the Hilbert space of $\mathcal{F}_t-$progresivelly measurable processes $(x_t)_{t\in[s,T]}$ with values in $\mathbb{R}^n$ such that $\mathbb{E}\int_s^T |x_t|^2\,dt < +\infty$.

The criteria to be minimized associated with the state equation (1) is defined by the functional

$$J(s,y,u(\cdot),\eta(\cdot)) = \mathbb{E}\left[h(x_T) + \int_s^T \ell(t,x_t,u_t)\,dt + \int_s^T k_t d\eta_t\right], \tag{2}$$

and the associated value function is defined as

$$V(s,y) = \inf_{(u(\cdot),\eta(\cdot))\in\mathbb{U}([s,T])} J(s,y,u(\cdot),\eta(\cdot)). \tag{3}$$

1.1 Optimality and near-optimality

The usual goal in control theory is to find the optimal control $(u^*(\cdot),\eta^*(\cdot)) \in \mathbb{U}([s,T])$ so that the infimum in (3) is attained, i.e. $V(s,y) = J(s,y,u^*(\cdot),\eta^*(\cdot))$.

In this place it is worth mentioning that optimal (not only singular) controls may not exist in many (even trivial) situations, while the following concept, the near-optimal singular controls, always exist.

Definition 2. *For a given $\varepsilon > 0$ the admissible control $(u^\varepsilon(\cdot),\eta^\varepsilon(\cdot))$ is called near-optimal if*

$$|J(s,y,u^\varepsilon(\cdot),\eta^\varepsilon(\cdot)) - V(s,y)| \le \mathcal{O}(\varepsilon), \tag{4}$$

where $\mathcal{O}(\cdot)$ is a function of ε satisfying $\lim_{\varepsilon\to 0}\mathcal{O}(\varepsilon) = 0$. The estimator $\mathcal{O}(\varepsilon)$ is called an error bound.
If $\mathcal{O}(\varepsilon) = C\varepsilon^\delta$ for some $\delta > 0$ independent of the constant $C > 0$ then $(u^\varepsilon(\cdot),\eta^\varepsilon(\cdot))$ is called near-optimal control of order ε^δ.
If $\mathcal{O}(\varepsilon) = \varepsilon$ the admissible control $(u^\varepsilon(\cdot),\eta^\varepsilon(\cdot))$ called $\varepsilon-$optimal.

1.2 Standing assumptions

Throughout the paper we assume the following:

(H1) $f : [s,T]\times\mathbb{R}^n\times\mathbb{A}_1 \to \mathbb{R}^n$, $\sigma : [s,T]\times\mathbb{R}^n\times\mathbb{A}_1 \to \mathbb{R}^{n\times l}$ and $\ell : [s,T]\times\mathbb{R}^n\times\mathbb{A}_1 \to \mathbb{R}$ are measurable in (t,x,u), twice continuously differentiable in x and there exists a constant $C > 0$ such that for $\varphi = f,\sigma,\ell$:

$$|\varphi(t,x,u) - \varphi(t,x',u)| + |\varphi_x(t,x,u) - \varphi_x(t,x',u)| \le C|x-x'|, \tag{5}$$

$$|\varphi(t,x,u)| \le C(1+|x|). \tag{6}$$

(H2) $h : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable in x and there exists a constant $C > 0$ such that

$$|h(x) - h(x')| + |h_x(x) - h_x(x')| \le C|x-x'|. \tag{7}$$

$$|h(x)| \le C(1+|x|). \tag{8}$$

(H3) $G : [s,T] \to \mathbb{R}^{n\times m}$, $k : [s,T] \to \mathbb{A}_2$. G is continuous, bounded and k is continuous.

It is a classical result that under the assumptions (H1)-(H3) the following SDE

$$\begin{aligned} d\tilde{x}_t &= f\left(t,\tilde{x}_t,u_t\right)dt + \sigma\left(t,\tilde{x}_t,u_t\right)dW_t,\ t\in[s,T] \\ \tilde{x}_s &= y, \end{aligned}$$

has a unique strong solution $(\tilde{x}_t)_{t\in[s,T]}$ for each $u(\cdot)$ coming from an admissible control $(u(\cdot),\eta(\cdot))$. Then the solution to SDE (1) is obtained as $x_t = \tilde{x}_t + \int_s^t G_r d\eta_r,\ \forall t\in[s,T]$.

1.3 Hamiltonian and adjoint equations

For any $(u(\cdot),\eta(\cdot)) \in \mathbb{U}([s,T])$ and the corresponding state trajectory (x_t), we define the first-order adjoint processes (Ψ_t, K_t) and the second-order adjoint processes (Q_t, R_t) as the solution to the following two backward SDEs respectively

$$\begin{cases} d\Psi_t = -\left[f_x^*\left(t,x_t,u_t\right)\Psi_t + \sigma_x^*\left(t,x_t,u_t\right)K_t + \ell_x\left(t,x_t,u_t\right)\right]dt \\ \qquad\quad + K_t dW_t, \\ \Psi_T = h_x\left(x_T\right), \end{cases} \tag{9}$$

and

$$\begin{cases} dQ_t = -\left[f_x^*\left(t,x_t,u_t\right)Q_t + Q_t f_x^*\left(t,x_t,u_t\right) + \sigma_x^*\left(t,x_t,u_t\right)Q_t\sigma_x^*\left(t,x_t,u_t\right)\right. \\ \qquad\quad \left. + \sigma_x^*\left(t,x_t,u_t\right)R_t + R_t\sigma_x\left(t,x_t,u_t\right) + \Gamma_t\right]dt + R_t dW_t, \\ Q_T = h_{xx}\left(x_T\right), \end{cases} \tag{10}$$

where

$$\Gamma_t = \ell_{xx}\left(t,x_t,u_t\right) + \sum_{i=1}^{n}\left(\Psi_t^i f_{xx}^i\left(t,x_t,u_t\right) + K_t^i\sigma_{xx}^i\left(t,x_t,u_t\right)\right).$$

As it is well known that under conditions (H1), (H2) and (H3) the first-order adjoint equation (9) (the second-order adjoint equation (10) respectively) admits a unique solution pair $(\Psi_t,K_t) \in \mathbb{L}^2_{\mathcal{F}}\left([s,T];\mathbb{R}^n\right)\times\mathbb{L}^2_{\mathcal{F}}\left([s,T];\mathbb{R}^{n\times l}\right)$ (a unique solution pair $(Q_t,R_t) \in \mathbb{L}^2_{\mathcal{F}}\left([s,T];\mathbb{R}^{n\times n}\right)\times\mathbb{L}^2_{\mathcal{F}}\left([s,T];\mathbb{R}^{n\times n\times l}\right)$ respectively).
Now we define the usual Hamiltonian function

$$H\left(t,x,u,p,q\right) := -pf\left(t,x,u\right) - q\sigma\left(t,x,u\right) - \ell\left(t,x,u\right), \tag{11}$$

for $(t,x,u,p,q) \in [s,T]\times\mathbb{R}^n\times\mathbb{A}_1\times\mathbb{R}^n\times\mathbb{R}^{n\times l}$. Furthermore, we define the so called $\mathcal{H}$-function corresponding to a given admissible pair (x_t,u_t) as follows

$$\begin{aligned} \mathcal{H}^{(x,u)}(t,x,u) &= H\left(t,x,u,\Psi_t,K_t - Q_t\sigma\left(t,x,u\right)\right) \\ &\quad -\frac{1}{2}\sigma^*\left(t,x,u\right)Q_t\sigma\left(t,x,u\right), \end{aligned}$$

for $(t,x,u,p,q) \in [s,T]\times\mathbb{R}^n\times\mathbb{A}_1\times\mathbb{R}^n\times\mathbb{R}^{n\times l}$, where Ψ_t, K_t and Q_t are determined by adjoint equations (9) and (10) corresponding to (x_t,u_t).

2 Main results

2.1 Necessary conditions for near-optimal singular control

The necessary conditions for near-optimality for singular controls is given by the following theorem. The interpretation of the condition is that every near-optimal control has to 'near-maximize' the $\mathcal{H}$ function in some integral sense.

Theorem 1. *Let (H1)-(H3) hold and let $(u^\varepsilon(\cdot), \eta^\varepsilon(\cdot))$ be an arbitrary near-optimal control to the singular control problem (1),(2) and (3) for some arbitrary but fixed $\varepsilon > 0$. Further, let $(\Psi_t^\varepsilon, K_t^\varepsilon)$ and $(Q_t^\varepsilon, R_t^\varepsilon)$ be the solution of adjoint equations (9) and (10) respectively corresponding to $(x_t^\varepsilon, (u_t^\varepsilon, \eta_t^\varepsilon))$.*

Then for any $\delta \in (0, \frac{1}{3}]$ there exists a positive constant $C = C(\delta)$ such that for each admissible control $(u(\cdot), \eta(\cdot)) \in \mathbb{U}([s,T])$ it holds

$$\begin{cases} -C\varepsilon^\delta \leq \mathbb{E}\int_s^T \left\{ \frac{1}{2} \left(\sigma\left(t, x_t^\varepsilon, u_t\right) - \sigma\left(t, x_t^\varepsilon, u_t^\varepsilon\right)\right)^* Q_t^\varepsilon \left(\sigma\left(t, x_t^\varepsilon, u_t\right) - \sigma\left(t, x_t^\varepsilon, u_t^\varepsilon\right)\right) \right. \\ \quad + \Psi_t^\varepsilon \left(f\left(t, x_t^\varepsilon, u_t\right) - f\left(t, x_t^\varepsilon, u_t^\varepsilon\right)\right) + K_t^\varepsilon \left(\sigma\left(t, x_t^\varepsilon, u_t\right) - \sigma\left(t, x_t^\varepsilon, u_t^\varepsilon\right)\right) \\ \quad \left. + \left(\ell\left(t, x_t^\varepsilon, u_t\right) - \ell\left(t, x_t^\varepsilon, u_t^\varepsilon\right)\right) \right\} dt, \end{cases}$$

and

$$-C\varepsilon^\delta \leq \mathbb{E}\left[\int_s^T (k_t + G_t^* \Psi_t^\varepsilon) d\left(\eta_t - \eta_t^\varepsilon\right)\right].$$

Corollary 1. *Under the assumptions of Theorem 1 we have*

$$\mathbb{E}\int_s^T \mathcal{H}^{(x^\varepsilon, u^\varepsilon)}(t, x_t^\varepsilon, u_t^\varepsilon) dt \geq \sup_{u(\cdot) \in \mathbb{U}_1([s,T])} \mathbb{E}\int_s^T \mathcal{H}^{(x^\varepsilon, u^\varepsilon)}(t, x_t^\varepsilon, u_t) dt - C\varepsilon^\delta,$$

and

$$\mathbb{E}\int_s^T (k_t + G_t^* \Psi_t^\varepsilon) d\eta_t^\varepsilon \leq \inf_{\eta(\cdot) \in \mathbb{U}_2([s,T])} \mathbb{E}\int_s^T (k_t + G_t^* \Psi_t^\varepsilon) d\eta_t + C\varepsilon^\delta.$$

2.2 Sufficient condition for near-optimality

Under (H1)-(H3) and some additional assumptions on differentiability and concavity the condition of near-maximality of the Hamiltonian function is in fact a sufficient one. Let us further assume that

(H4) The functions f, σ and ℓ are differentiable in u and there is a constant $C > 0$ such that for each t, x, u, u'

$$|\varphi(t,x,u) - \varphi(t,x,u')| + |\varphi_u(t,x,u) - \varphi_u(t,x,u')| \leq C\,|u - u'|\,, \tag{12}$$

where $\varphi = f, \sigma, \ell$.

Theorem 2. *Let* $(\tilde{u}(\cdot), \tilde{\eta}(\cdot))$ *be an arbitrary admissible control and let* $\left(\tilde{\Psi}_t, \tilde{K}_t\right)$ *and* $\left(\tilde{Q}_t, \tilde{R}_t\right)$ *be the solutions to adjoint equations (9)-(10) associated with* $(\tilde{u}(\cdot), \tilde{\eta}(\cdot))$. *Further, let us assume that the function* $H\left(t, \cdot, \cdot, \tilde{\Psi}_t, \tilde{K}_t\right)$ *is concave (in* (x, u)*) for a.e.* $t \in [s, T]$, $\mathbb{P} - a.s$ *and that* h *is a convex function.*
If for some $\varepsilon > 0$ *and for any admissible control* $(u(\cdot), \eta(\cdot))$ *the following near-maximality conditions hold*

$$\mathbb{E}\int_s^T \mathcal{H}^{(\tilde{x},\tilde{u})}(t, \tilde{x}_t, \tilde{u}_t)dt \geq \sup_{u(\cdot)\in\mathbb{U}_1([s,T])} \mathbb{E}\int_s^T \mathcal{H}^{(\tilde{x},\tilde{u})}(t, \tilde{x}_t, u_t)dt - \varepsilon^{\frac{1}{2}},$$

and

$$\mathbb{E}\left[\int_s^T k_t d\left(\eta_t - \tilde{\eta}_t\right)\right] \geq -C\varepsilon^{\frac{1}{2}},$$

with C *being a positive constant independent of* ε, *then* $(\tilde{u}(\cdot), \tilde{\eta}(\cdot))$ *is in fact near-optimal control for the control problem (1),(2) and (3), i.e.*

$$J\left(s, y, \tilde{u}(\cdot), \tilde{\eta}(\cdot)\right) \leq V\left(s, y\right) + C\varepsilon^{\frac{1}{2}}.$$

3 Proofs and References

All the proofs as well as a full list of references can be found in the original paper.

Higher Roytenberg Bracket and Applications*

Jan Vysoký

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

advisor: Branislav Jurčo, Mathematical Institute, Charles University in Prague

Abstract. Roytenberg bracket is a Courant algebroid structure obtained from Courant-Dorfman bracket by twisting using the background fields on a manifold. A similar procedure starting from higher Dorfman bracket is provided. The most important properties of higher Roytenberg bracket are presented. Higher Roytenberg bracket is derived here using the worldvolume algebra of p-brane action. This is done using the recent results of Ekstrand and Zabzine. The knowledge of the algebra of charges is used to calculate the conditions of their conservation in the time evolution. The coordinate expressions of higher Roytenberg bracket are presented.

Keywords: Roytenberg bracket, Courant algebroid, worldvolume algebra, charges

Abstrakt. Roytenbergova závorka je Courantovým algebroidem, který lze získat "twistováním" Courant-Dorfmanové závorky užitím polí pozadí na varietě. Ukazujeme podobnou proceduru, kde ale vycházíme z vyšší Dorfmanové závorky. Nejdůležitější vlastnosti vyšší Roytenbergovy závorky jsou předvedeny. Vyšší Roytenbergova závorka je odvozena použitím světoobjemové algebry pro akci p-brány. Toho je dosaženo pomocí výsledků Ekstranda a Zabzina. Znalost algebry nábojů je použita k vypočítání podmínek pro jejich zachování v časovém vývoji. V práci jsou vypsány souřadnicové výrazy pro vyšší Roytenbergovu závorku.

Klíčová slova: Roytenbergova závorka, Courantův algebroid, světoobjemová algebra, náboje

1 Introduction

It is a well known fact that a tangent bundle TM of any smooth manifold M is naturally equipped with the bracket of its smooth sections:

$$[\cdot,\cdot] : \Gamma(TM) \times \Gamma(TM) \to \Gamma(TM).$$

Of course it is a Lie bracket of smooth vector fields commutator. This bracket is crucial for the integrability of tangent distributions in M, via the famous Frobenius theorem. In the study of non-regular Lagrangian mechanics, there emerged the need of similar integrability condition in the more general vector bundle, namely $TM \oplus T^*M$. Ted Courant in [3] introduced a new bracket of the sections of this vector bundle:

$$[V+\xi, W+\eta]_C = [V,W] + \frac{1}{2}\big(\mathcal{L}_V(\eta) - \mathcal{L}_W(\xi) + i_V(d\eta) - i_W(d\xi)\big), \tag{1}$$

for all $V, W \in \Gamma(TM)$ and $\xi, \eta \in \Gamma(T^*M)$. This again is a canonical structure for any manifold M. However; the cost was a violation of the Jacobi identity and Leibniz rule.

*Excerpts from the paper written with Branislav Jurčo.

In fact, this bracket appeared to be an example of so called Courant algebroid, which was first axiomatized in [9]. Although this skew-symmetric version was nicely related to strongly homotopy algebras by Dmitry Roytenberg and Alan Weinstein in [12], the anomalies in Jacobi identities and Leibniz rule are difficult to handle. Fortunately, in his thesis [10] Dmitry Roytenberg proved that to every Courant algebroid there exists its non-skew-symmetric version, for which Jacobi identity and Leibniz rule holds, and vice versa. This notion of Courant algebroid, related closely to Leinbniz algebroids, is now widely accepted as more convenient definition. The non-skew-symmetric version of the original Courant bracket is called Dorfman bracket or Courant-Dorfman bracket.

This bracket can be twisted in various ways, incorporating the background structures of the manifold M, without spoiling the Courant algebroid properties. The most general way was discussed by Dmitry Roytenberg in [11], which explains the origin of the name Roytenberg bracket.

In [1], Anton Alekseev and Thomas Strobl observed that Courant bracket can be derived using so called worldsheet algebra, the algebra of Noetherian charges in theory of sigma models with respect to canonical Poisson bracket. For more general charges (of more general sigma models), the same procedure was done by Nick Halmagyi in [6], to obtain the Roytenberg bracket.

A situation becomes more complicated when one tries to work on a more general vector bundle, $E = TM \oplus \Lambda^p T^*M$, for $p > 1$. There exists straightforward generalization of Courant bracket, using the same formula (1), or its non-skew-symmetric version. This bracket again has very convenient properties, and was studied by Yoshuke Hagiwara in [5] or by Yanhui Bi and Yunhe Sheng in [2]. However; it is still unclear how to axiomatize such brackets, finding some "higher analogues" of Courant algebroids.

In this paper we present the generalization of Roytenberg bracket. We derive it using the worldvolume algebra of lately proposed p-brane action (or Nambu sigma model), which was introduced by Branislav Jurčo and Peter Schupp in [7]. We show how they can be applied to calculate the conditions for the conservation of the Noetherian charges.

2 Higher Roytenberg bracket

Let $E = TM \oplus \Lambda^p T^*M$. We define a non-degenerate and $C^\infty(M)$-bilinear pairing $\langle \cdot, \cdot \rangle : \Gamma(E) \times \Gamma(E) \to \Omega^{p-1}(M)$ as

$$\langle V + \xi, W + \eta \rangle = i_V(\eta) + i_W(\xi), \tag{2}$$

for vector fields $V, W \in \mathfrak{X}(M)$ and p-forms $\xi, \eta \in \Omega^p(M)$. We define the anchor map $\rho : E \to TM$ as the projection onto the first direct summand of E, and denote by the same character also the induced map of sections $\rho(V + \xi) = V$. The higher Dorfman bracket is the $\mathbb{R}$-bilinear bracket on sections $[\cdot, \cdot]_D : \Gamma(E) \times \Gamma(E) \to \Gamma(E)$, defined as

$$[V + \xi, W + \eta]_D = [V, W] + \mathcal{L}_V(\eta) - i_W(d\xi), \tag{3}$$

for all $V, W \in \mathfrak{X}(M)$ and $\xi, \eta \in \Omega^p(M)$. This bracket is a particular example of a Leibniz algebroid bracket, see [2]. If we define $\mathcal{D} : \Omega^{p-1}(M) \to \Gamma(E)$ as $\mathcal{D} = j \circ d$, where $j : \Omega^p(M) \hookrightarrow \Gamma(E)$ is the inclusion, we have the following properties of higher Dorfman bracket:

1.
$$[e_1,[e_2,e_3]_D]_D=[[e_1,e_2]_D,e_3]_D+[e_2,[e_1,e_3]_D]_D, \tag{4}$$
for all $e_1,e_2,e_3\in\Gamma(E)$.
$$[e_1,fe_2]=f[e_1,e_2]+(\rho(e_1).f)e_2, \tag{5}$$
for all $e_1,e_2\in\Gamma(E)$ and $f\in C^\infty(M)$.

2. $\langle\cdot,\cdot\rangle$ is E-invariant in the following sense:
$$\mathcal{L}_{\rho(e_1)}(\langle e_2,e_3\rangle)=\langle[e_1,e_2]_D,e_3\rangle+\langle e_2,[e_1,e_3]_D\rangle, \tag{6}$$
for all $e_1,e_2,e_3\in\Gamma(E)$.

3. Higher Dorfman bracket is skew-symmetric up to "coboundary", that is
$$[e_1,e_1]=\frac{1}{2}\mathcal{D}\langle e_1,e_1\rangle, \tag{7}$$

This bracket can be easily modified in two ways. First, assume that on M we have a closed $(p+2)$-form $H\in\Omega^{p+2}(M)$. Then we can define H-twisted higher Dorfman bracket on E as
$$[V+\xi,W+\eta]_D^{(H)}=[V,W]+\mathcal{L}_V(\eta)-i_W(d\xi)+i_Wi_VH. \tag{8}$$
The form H has to be closed to keep the property (4), all the other properties of higher Dorfman bracket are valid also for the H-twisted case.

Now, assume that we have an arbitrary $C^\infty(M)$-linear map of sections $\Pi^\#:\Omega^p(M)\to\mathfrak{X}(M)$, for example the map induced by a $(p+1)$-vector Π on M. Define new anchor $\rho:E\to TM$ as
$$\rho(V+\xi)=V+(-1)^{p+1}\Pi^\#(\xi), \tag{9}$$
and the "twisted" inclusion of $\Omega^p(M)$ into $\Gamma(E)$ as
$$j(\xi)=\xi+(-1)^p\Pi^\#(\xi). \tag{10}$$
Denote as pr_2 the projection onto the second summand of E. Using this notation, one can define new non-degenerate pairing $\langle\cdot,\cdot\rangle_R$:
$$\langle e_1,e_2\rangle_R=i_{\rho(e_1)}(pr_2(e_2))+i_{\rho(e_2)}(pr_2(e_1)), \tag{11}$$
for all $e_1,e_2\in\Gamma(E)$. Finally, we define the following bracket on $\Gamma(E)$:
$$[e_1,e_2]_R=[\rho(e_1),\rho(e_2)]+ \tag{12}$$
$$+j\big(\mathcal{L}_{\rho(e_1)}(pr_2(e_2))-i_{\rho(e_2)}(d(pr_2(e_1)))+i_{\rho(e_2)}i_{\rho(e_1)}H\big),$$
for all $e_1,e_2\in\Gamma(E)$. We refer to $[\cdot,\cdot]_R$ as higher Roytenberg bracket. This bracket together with the anchor (9) defines again a Leibniz algebroid, that is it satisfies (4) and (5). More interestingly, it also satisfies (6) and (7) with respect to the pairing (11). All of the properties are straightforward to check.

3 A p-brane action, basic properties

Let us consider a $(p+1)$-dimensional worldvolume Σ with set of local coordinates $(\sigma^0, \dots, \sigma^p)$. We assume that σ^μ are Cartesian coordinates for Lorentzian metric h of signature $(-,+,\dots,+)$ on Σ .

Next, we consider an n-dimensional target manifold M, equipped with a $(p+1)$-vector Π and a $(p+1)$-form B. We also choose some local coordinates $(y^1, \dots, y^n)$ on M. Lower case Latin characters will always correspond to these coordinates. We will use upper case Latin characters to denote strictly ordered multi-indices (mostly p-indices), that is $I = (i_1, \dots, i_p)$, where $i_1 < \cdots < i_p$.

For a smooth map $X : \Sigma \to M$ we will use the notation $X^i = y^i(X)$, $dX^I = dX^{i_1} \wedge \ldots \wedge dX^{i_p}$, and $\widetilde{\partial X}^I = (dX^I)_{1\dots p}$ for the $1\dots p$ component of the worldvolume form dX^I.

We will also assume that M is equipped with a metric tensor field G with local components G_{ij}, and a fiberwise metric $\widetilde{G}$ on the vector bundle $\Lambda^p TM$ with components $\widetilde{G}_{IJ}$ in local section basis $\partial_I \equiv \frac{\partial}{\partial y^{i_1}} \wedge \cdots \wedge \frac{\partial}{\partial y^{i_p}}$. Metric matrices with upper indices denote as usually the corresponding inverses.

The action is the following one:

$$S[\eta, \widetilde{\eta}, X] := \int d^{p+1}\sigma \big[-\frac{1}{2}(G^{-1})^{ij}\eta_i\eta_j + \frac{1}{2}(\widetilde{G}^{-1})^{IJ}\widetilde{\eta}_I\widetilde{\eta}_J + \eta_i\partial_0 X^i + \tag{13}$$

$$+\widetilde{\eta}_I\widetilde{\partial X}^I - \Pi^{iJ}\eta_i\widetilde{\eta}_J - B_{iJ}\partial_0 X^i\widetilde{\partial X}^J\big],$$

where $\eta_i, \widetilde{\eta}_J \in C^\infty(\Sigma)$ are the auxiliary fields, which transform under change of local coordinates on M accordingly to their index structure.

Canonical momenta corresponding to the fields X^i are

$$P_i = \eta_i - B_{iJ}\widetilde{\partial X}^J. \tag{14}$$

Going to the canonical Hamiltonian $H_{can}[X, P, \widetilde{\eta}] = \int d^p\sigma P_i\partial_0 X^i - \mathcal{L}(X, P, \widetilde{\eta})$, and substituting the Lagrange-Euler equation for $\widetilde{\eta}_J$, we obtain the Hamiltonian

$$H[X, P] = \frac{1}{2}\int d^p\sigma\big[(G^{-1})^{ij}K_iK_j + \widetilde{G}_{IJ}\widetilde{K}^I\widetilde{K}^J\big], \tag{15}$$

where

$$K_i := \eta_i = P_i + B_{iK}\widetilde{\partial X}^K, \tag{16}$$

$$\widetilde{K}^I = \widetilde{\partial X}^I + (-1)^{p+1}\Pi^{Im}K_m. \tag{17}$$

Here an in the rest of the paper, the integration over $d^p\sigma$ means the integration over the space-like coordinates $(\sigma^1, \dots, \sigma^p)$ of Σ.

4 Charge algebra

The canonical Poisson bracket is

$$\{X^i(\sigma), P_j(\sigma')\} = \delta^i_j\delta(\sigma - \sigma'),$$

where by σ, σ' we mean the space-like p-tuples of coordinates, and all the Poisson brackets are the equal time ones.

We consider the following generalized charges, corresponding to the currents K^i and $\widetilde{K}_J$ appearing explicitly in the Hamiltonian:

$$Q_f(V+\xi) = \int d^p\sigma f(\sigma)[V^i K_i + \xi_J \widetilde{K}^J], \tag{18}$$

where $V + \xi \in \Gamma(E)$, and $f \in C^\infty(\Sigma)$ is a test function.

The appearance of Courant algebroid structures in the current algebra was first observed by Anton Alekseev and Thomas Strobl for $p = 1$ in [1]. Here we follow the idea of Joel Ekstrand and Maxim Zabzine, who integrated the currents to generalized charges, and calculated their algebra for $p \geq 1$. We consider more general charges, involving background fields Π and B. This can be done in a straightforward way; however it is easier to use the results of [4]:

Let $\widetilde{Q}_f(V+\xi)$ be defined as

$$\widetilde{Q}_f(V+\xi) = \int d^p\sigma f(\sigma)\big[V^i P_i + \xi_J \widetilde{\partial X}^J\big]. \tag{19}$$

Then for their Poisson bracket we get

$$\{\widetilde{Q}_f(V+\xi), \widetilde{Q}_g(W+\eta)\} = -\widetilde{Q}_{fg}([V+\xi, W+\eta]_D) - \tag{20}$$

$$- \int d^p\sigma g(\sigma)(df \wedge X^*(\langle V+\xi, W+\eta\rangle))_{1\ldots p},$$

where $[\cdot,\cdot]_D$ is the higher Dorfman bracket (3) and $\langle\cdot,\cdot\rangle$ is the pairing (2).

We can use this result to find the Poisson brackets for charges Q. The key is the following relation between charges Q and $\tilde{Q}$:

$$Q_f(V+\xi) = \widetilde{Q}_f\big(V + (-1)^{p+1}\Pi^\#(\xi) + \xi + i_{V+(-1)^{p+1}\Pi^\#(\xi)}(B)\big). \tag{21}$$

The calculation is tedious but straightforward and we omit it here. For the Poisson bracket of the charges (18) we have:

$$\{Q_f(V+\xi), Q_g(W+\eta)\} = -Q_{fg}([V+\xi, W+\eta]_R) - \tag{22}$$

$$- \int d^p\sigma g(\sigma)(df \wedge X^*(\langle V+\xi, W+\eta\rangle_R))_{1\ldots p},$$

where $[\cdot,\cdot]_R$ is the higher Roytenberg bracket (12) and $\langle\cdot,\cdot\rangle_R$ is the pairing (11).

Let us note that choosing $f = g = 1$, one finds that the charge algebra (22) closes and it is described by higher Roytenberg bracket. This was already observed by Halmagyi [6] for $p = 1$.

5 Applications

Using this result, we can determine conditions for conservation of such charges. To get rid of anomalous term in (22), we consider only the charges

$$Q(V+\xi) := Q_1(V+\xi), \tag{23}$$

that is choosing $f=1$. Hence we would like to obtain the conditions on $V+\xi \in \Gamma(E)$, which would guarantee that

$$\{Q(V+\xi), H\} = 0, \tag{24}$$

where H is the Hamiltonian (15). The left hand side of this condition can be conveniently rewritten using the Leibniz rule for Poisson bracket as

$$\begin{aligned}\{Q(V+\xi), H\} = \\ = \frac{1}{2}\{Q(V+\xi), Q_{K_i}((G^{-1})^{ij}\partial_j)\} + \frac{1}{2}\{Q(V+\xi), Q_{(G^{-1})^{ij}\partial_j}(\partial_i)\} + \\ + \frac{1}{2}\{Q(V+\xi), Q_{\widetilde{K}^I}(\widetilde{G}_{IJ}dy^J)\} + \{Q(V+\xi), Q_{\widetilde{G}_{IJ}\widetilde{K}^J}(dy^I)\}.\end{aligned} \tag{25}$$

Now we can use the (22) to carry out the calculation. After a tedious, but straightforward calculation, one arrives to the following result. Put $W = V + (-1)^{p+1}\Pi^{\#}(\xi)$. The charge $Q(V+\xi)$ conserves, if the following set of (sufficient) conditions is satisfied:

$$\mathcal{L}_W(G)_{ij} = (-1)^{p+1}G_{in}\Pi^{Ln}\big((d\xi)_{jL} - W^m dB_{mjL}\big) + (i \leftrightarrow j). \tag{26}$$

$$\mathcal{L}_W(\widetilde{G})_{IJ} = (-1)^{p+1}\widetilde{G}_{IL}\Pi^{Ln}\big((d\xi)_{nJ} - W^m dB_{mnJ}\big) + (I \leftrightarrow J). \tag{27}$$

$$\mathcal{L}_W(\Pi)^{Ik} = (-1)^p\big(\Pi^{In}\Pi^{Lk} - (\widetilde{G}^{-1})^{IL}(G^{-1})^{kn}\big)\big((d\xi)_{nL} - W^m dB_{mnL}\big). \tag{28}$$

Here $\widetilde{G}$ is viewed as $2p$-times covariant tensor field on M. Let us note that there exists a particular simplification of the these conditions; if one assumes

$$d\xi = i_W(dB), \tag{29}$$

all the right-hand sides vanish, and we get the set of conditions

$$\mathcal{L}_W(G) = \mathcal{L}_W(\widetilde{G}) = \mathcal{L}_W(\Pi) = 0. \tag{30}$$

The assumption (29) can be rewritten as

$$\mathcal{L}_W(B) = d(\xi - i_W(B)). \tag{31}$$

Obviously, the particular solution (30) to the more general conditions (26-28) says that the image of $V+\xi$ under the anchor (9) preserves the background fields $G, \widetilde{G}, \Pi$ and preserves B up to an exact term.

Conditions (26-28) have interesting geometrical meaning. Let $(\cdot,\cdot)$ be a fiberwise metric on $TM \oplus \Lambda^p T^*M$ given by

$$(V+\xi, W+\eta) := \begin{pmatrix} V \\ \xi \end{pmatrix}^T \begin{pmatrix} G & G\Pi^T \\ \Pi^T G & \widetilde{G}^{-1} + \Pi^T G \Pi^T \end{pmatrix} \begin{pmatrix} W \\ \eta \end{pmatrix}. \tag{32}$$

Note that this is in fact the inverse of the matrix in the Hamiltonian, where we put $B = 0$. Let $e = V + (-1)^{p+1}\Pi^{\#}(\xi) + \xi$. The conditions (26 - 28) are equivalent to the equation

$$pr_1(e).(e_1, e_2) = ([e, e_1]_D^{(dB)}, e_2) + (e_1, [e, e_2]_D^{(dB)}), \tag{33}$$

for all $e_1, e_2 \in \Gamma(TM \oplus \Lambda^p T^*M)$, $[\cdot,\cdot]_D^{(dB)}$ is a dB-twisted higher Dorfman bracket (8), and pr_1 is a projection onto TM. In the other words, $Q(V + \xi)$ conserves, if $e = V + (-1)^{p+1}\Pi^{\#}(\xi) + \xi$ is a "Killing section" of the fiberwise metric $(\cdot,\cdot)$ (32) w.r.t. to dB-twisted higher Dorfman bracket.

6 Coordinate expressions for the higher Roytenberg bracket

Here we recall the local form of the higher Roytenberg bracket. Let $(y^1, \ldots, y^n)$ be a set of local coordinates on M. Denote $\partial_k = \frac{\partial}{\partial y^k}$ and $dy^K = dy^{k_1} \wedge \ldots \wedge dy^{k_p}$. Then, one has

$$[\partial_k, \partial_l]_R = F_{kl}{}^m \partial_m + H_{klL} dy^L, \tag{34}$$

$$[\partial_k, dy^J]_R = Q_k^{Jm} \partial_m + D^J_{k\,L} dy^L, \tag{35}$$

$$[dy^I, dy^J]_R = R^{IJm} \partial_m + S^{IJ}{}_L dy^L. \tag{36}$$

Structure functions have the following form:

$$F_{kl}{}^m = (-1)^p dB_{klJ} \Pi^{Jm}, \tag{37}$$

$$H_{klJ} = dB_{klJ}, \tag{38}$$

$$Q_k^{Jm} = (-1)^{p+1} \Pi^{Jm}{}_{,k} - dB_{klL} \Pi^{Lm} \Pi^{Jl}, \tag{39}$$

$$D^J_{k\,L} = (-1)^{p+1} \Pi^{Jl} dB_{klL}, \tag{40}$$

$$R^{IJm} = \Pi^{In} \Pi^{Jm}{}_{,n} - \Pi^{Jn} \Pi^{Im}{}_{,n} - \sum_{r=1}^{p} \Pi^{Ij_r}{}_{,k} \Pi^{j_1 \ldots k \ldots j_p m} + (-1)^p \Pi^{Ik} \Pi^{Jl} \Pi^{Lm} (dB)_{klL}, \tag{41}$$

$$S^{IJ}{}_L = (-1)^{p+1} \sum_{r=1}^{p} \Pi^{Ij_r}{}_{,k} \delta_L^{j_1 \ldots k \ldots j_p} + \Pi^{Ij} \Pi^{Jl} (dB)_{klL}. \tag{42}$$

7 Conclusion

We have shown how can be the higher Roytenberg bracket obtained by twisting the ordinary higher Dorfman bracket. This approach simplifies the proofs of the properties, which will be quite impossible using the coordinate expressions (see section 6). Similar formula was derived by Yvette Kosmann-Schwarzbach in [8] for $p = 1$ case.

We have derived this bracket using the generalized charges of the p-brane action, where we have used the calculation in [4]. Appearance of higher Roytenberg bracket in the Poisson algebra of the charges is very useful for many calculations. We show how

it can be used to find the sufficient conditions for the charge conservation and we have given them a geometrical meaning.

We would like to find a sufficient axiomatization for higher Roytenberg bracket, possibly finding more interesting examples of higher Courant-like structures. Moreover, usual Courant algebroids come as derived brackets of symplectic supermanifolds, which should be more or less generalizable to the $p > 1$ case. The formulation using supermanifolds could possibly help us with the AKSZ formulation of p-brane actions.

Anomalies present in the bracket (22) lead to the discussions of possible secondary constraints. We have recently partially solved this problem.

We would like to thank Noriaki Ikeda, for his helpful comments and discussion.

References

[1] A. Alekseev and T. Strobl. *Current Algebras and Differential Geometry.* JHEP **03** (2005).

[2] Y. Bi and Y. Sheng. *On higher analogues of Courant algebroids.* Sci. China Math. **54** (2011), 437–447.

[3] T. J. Courant. *Dirac Manifolds.* Transactions of the American Mathematical Society **319** (1990), 631–661.

[4] J. Ekstrand and M. Zabzine. *Courant-like brackets and loop spaces.* Journal of High Energy Physics **1103:074** (2011).

[5] Y. Hagiwara. *Nambu-Dirac manifolds.* Journal of Physics A: Math. Gen. **35** (2002), 1263.

[6] N. Halmagyi. *Non-geometric String Backgrounds and Worldsheet Algebras.* JHEP **0807:137** (2008).

[7] B. Jurco and P. Schupp. *Nambu sigma model and effective membrane actions.* Physics Letters B **713** (2012), 313–316.

[8] Y. Kosmann-Schwarzbach. *Quasi, twisted, and all that... in Poisson geometry and Lie algebroid theory.* The Breadth of Symplectic and Poisson Geometry (2005), 363–389.

[9] Z.-J. Liu, A. Weinstein, and P. Xu. *Manin triples for Lie bialgebroids.* Journal of Differential Geometry **45** (1995), 547–574.

[10] D. Roytenberg. *Courant algebroids, derived brackets and even symplectic supermanifolds.* PhD thesis, University of California at Berkeley, (1999).

[11] D. Roytenberg. *Quasi-Lie Bialgebroids And Twisted Poisson Manifolds.* Lett.Math.Phys **61** (2002), 123–137.

[12] D. Roytenberg and A. Weinstein. *Courant Algebroids and Strongly Homotopy Lie Algebras.* ArXiv Mathematics e-prints (1998).

Design of a General-purpose Unstructured Mesh in C++*

Vítězslav Žabka

3rd 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. This article explains the motivation and design decisions behind a new general-purpose mesh library. The library provides a unified interface for traversing unstructured meshes of different dimensions and types, such as triangular, tetrahedral and hexahedral. It is designed in the C++ language using templates. Placing the emphasis on the ability to customize the internal representation of meshes, the library is particularly suitable for porting to systems with small memory, e.g., GPUs.

Keywords: Unstructured mesh, C++

Abstrakt. Tento příspěvek popisuje návrh nové knihovny pro práci se sítěmi a objasňuje motivaci pro její vývoj. Knihovna poskytuje jednotné rozhraní pro procházení nestrukturovaných sítí různých dimenzí a typů, např. trojúhelníkových, čtyřstěnových nebo šestistěnových. Je navržena v jazyku C++ s využitím šablon. Vzhledem k možnosti přizpůsobit interní reprezentaci sítí je knihovna vhodná pro adaptaci na systémy s omezeným množstvím paměti, jako např. grafické procesory.

Klíčová slova: Nestrukturovaná síť, C++

1 Introduction

We are involved in the development of a multigrid solver for the incompressible Navier-Stokes equations [2, 3]. The solver is a C/C++ implementation of a geometric multigrid method. It relies on the mixed finite element discretization of the two-dimensional Navier-Stokes equations on a hierarchy of unstructured triangular grids. A parallel implementation of the solver for shared-memory systems was developed using OpenMP. In addition, a GPU version of the solver was created.

Our plans for a further development of the solver include its extension into 3D so that it could also handle unstructured tetrahedral grids. However, the triangular grid is an essential component of the solver, and there is no simple way to replace the triangular grid with another grid. Adding a new grid would basically mean that a new solver, similar to the original one, would have to be created. Since maintaining two similar solvers for

*This work has been supported by the grant No. SGS11/161/OHK4/3T/14 of the Student Grant Agency of the Czech Technical University in Prague and the project No. TA01020871 of the Technological Agency of the Czech Republic

one problem is unfavorable, we would prefer to use a library providing a unified C++ interface for accessing grids of different types. Such library would allow us to concentrate on the implementation of the algorithms independently of the grid type.

The main functionality we expect from the library is the ability to store and traverse arbitrary triangular and tetrahedral meshes in a unified manner. Additionally, we would like the same functionality to be also available on the GPU. To our best knowledge, there is no such freely available GPU library, with OP2 [7] being the sole exception. Nevertheless, the OP2 framework seems to be rather experimental. It is poorly documented and provides a low-level C-style API only.

Another possibility is to modify an existing CPU library by adding GPU support. GPU devices are substantially different from CPU-based systems. To utilize GPUs effectively, the library to be modified to support GPUs should meet several requirements. First, the library should allocate memory in contiguous blocks so that, on GPUs, coalesced memory accesses can be achieved by reordering data in memory [10], which increases GPU memory bandwidth. Second, the library should provide a mechanism for adjusting the internal representation of meshes in order not to store unnecessary data because the size of memory available might be a limiting factor on GPUs. Furthermore, the library must be released under an open source license. A brief overview of existing open-source libraries is presented in Section 1.1.

Since no library we are aware of is, without further modifications, suitable for adapting to GPUs, we decided to create a new mesh library with GPU friendliness being the main goal of its design.

1.1 Existing mesh libraries

We only consider libraries written in C++. Besides `libMesh`, all the following libraries make full use of C++ templates.

ViennaGrid [8]
: ViennaGrid is a library for the handling of unstructured meshes. It almost satisfies the requirements. However, it does not store all relevant mesh data in contiguous blocks of memory, and it relies on the Standard Template Library containers, which are not supported on GPUs.

GrAL [4]
: GrAL is a similar library to ViennaGrid. Similarly, it uses the STL containers and, furthermore, containers of containers. Such storage scheme is inconvenient considering the need for data reordering.

dune-grid [1]
: The dune-grid library is one of the core modules of DUNE, a template library for the numerical solution of partial differential equations. From our point of view, the most important drawback of the library is its complexity and the lack of documentation. This makes modifying dune-grid challenging. Moreover, dune-grid does not allow the internal representation of meshes to be tailored for a specific purpose.

`libMesh` [9]
: The `libMesh` library provides a whole framework for the numerical solution of PDEs. It allocates memory for each mesh vertex, edge, etc. separately and accesses them using pointers. Thus, data reordering to increase memory throughput on the GPU would be difficult to implement.

OpenMesh [5]
: The OpenMesh library supplies a generic data structure for manipulating polygon meshes. It is based on the halfedge data structure with the focus on surface meshes. Volumetric meshes are not supported.

CGAL [6]
: The CGAL library aims at providing easy access to efficient and reliable geometric algorithms, e.g., triangulations, convex hull algorithms and geometry processing. Similarly to OpenMesh, it uses halfedge data structures. CGAL is not designed for the purpose of the numerical solution of PDEs.

1.2 Terminology

The terminology of this article might not correspond to that commonly used in mathematics. The intended meaning of basic terms used in this article is as follows:

Mesh and grid
: A mesh is a collection of geometrical and topological objects (e.g., vertices, edges and triangles). By a grid is understood a mesh enriched with various quantities related to numerical computations.

Mesh entity
: Mesh entities are all the objects of which a mesh is composed. For example, a triangular mesh is composed of entities of three types: vertices, edges and triangles.

Mesh dimension
: Dimension of a mesh is the highest topological dimension of its entities. For example, triangular and quadrilateral meshes have dimension two, tetrahedral and hexahedral meshes have dimension three. Mesh dimension is always less than or equal to the dimension of the underlying space. If, e.g., a triangular mesh resides in a three-dimensional space, the dimension of the mesh is still two.

Vertex, cell and facet
: A vertex is a mesh entity of topological dimension zero. Considering a mesh of dimension N, a mesh entity of the maximum topological dimension, i.e., N, is referred to as a cell; a mesh entity of topological dimension $N-1$ is referred to as a facet. The surface of a cell is composed of facets.

Structured and unstructured mesh
: The property of being structured or unstructured is determined by the internal storage scheme of the mesh. Unstructured meshes store the connectivity between entities explicitly whereas structured (i.e., regular) meshes store this connectivity implicitly by the arrangement of the data in memory.

Border and coborder entities

Each mesh entity E of topological dimension $n > 1$ defines entities of topological dimension less than n which form the boundary of E. These entities are called the border entities of the entity E. For example, the border entities of an edge are the two vertices joined by the edge; the border entities of a quadrilateral are his four vertices (i.e., corners) and four edges (i.e., sides). We prefer the name border entities to a presumably more descriptive name boundary entities in order not to confuse entity border with mesh boundary. Coborder entities of a mesh entity E are those of which E is one of their border entities.

Conforming and nonconforming mesh

An N-dimensional mesh is conforming if, for all positive $n \leq N$, the intersection of each pair of its n-dimensional entities is either a border entity of both these entities or empty. Otherwise, the mesh is nonconforming.

2 Design decisions

The design of the mesh library is influenced especially by ViennaGrid. ViennaGrid is written in C++, it is open source, it can be configured to disable the explicit storage of mesh entities of certain types, and it stores cells and vertices in contiguous blocks of memory. On the other hand, ViennaGrid is not designed with regard to the GPU computations. First, it employs the STL containers `std::deque` and `std::map` which are not available on GPUs. And second, it stores the links from entities to their border and coborder entities as pointers. This assumes the border and coborder entities are stored as objects to which the pointers point. On the GPU devices, the data are usually organized differently to coalesce memory accesses. As a result, pointers to the objects are not available.

The principal design feature of the library is that the only containers used for data storage are arrays. All mesh cells, vertices and other entities are stored in single contiguous arrays. This allows random access to the entities and, consequently, iterating through the entities in parallel. In addition, the entity index in the corresponding array—unique among all the mesh entites of the same type—can be used for referring to the particular entity.

Another important characteristic of the design is utilizing C++ templates for generic programming. This approach looks better than the conventional techniques based on run-time polymorphism. Virtual methods suffer from run-time overhead and from the fact that they cannot be inlined. Moreover, run-time polymorphism is only available through pointers or references, and there are no arrays of polymorphic objects in C++. Instead of holding polymorphic objects, arrays can hold pointers to the objects. Using templates, the whole objects with all their attributes can be stored in arrays. However, these objects are not run-time polymorphic. They must be fully determined at compile time.

Compile-time polymorphism is used mainly to implement entity objects. Thus, all types of entities can be treated the same way. For that purpose, the following parameters must be available at compile time:

- dimension of the underlying space,
- type of the mesh cells,
- types of the entities to be stored (the cells and vertices are always stored),
- types of the border entities to be stored (the border vertices are always stored),
- types of the coborder entities to be stored.

The cell type determines the type of the other mesh entities. For example, a tetrahedral mesh is composed of tetrahedra, triangles, edges and vertices, and all these entities can also be stored. The library supports meshes with only one cell type and, accordingly, one entity type of each dimension; i.e., prismatic meshes are not supported.

The border and coborder entities of an entity are stored as their indices in the corresponding arrays. Integral indices are more preferable than pointers from the GPU's point of view. The number of the border entities is always known at compile time, so their indices are stored in statically allocated arrays. On the other hand, the number of the coborder entities depends on the particular mesh; therefore, dynamically allocated arrays have to be used. Meshes are treated as nonconforming.

3 Interface

The library provides functions for loading a mesh from a file, traversing the entities of a mesh, traversing the border and coborder entities of a mesh entity and retrieving the coordinates of a vertex.

Meshes are manipulated through `Mesh` objects. `Mesh` objects are created according to mesh configuration classes where the cell type and the dimension of the underlying space are specified. An example of a mesh configuration class for a tetrahedral mesh in a 3D space follows:

```
class MeshConfig : public MeshConfigBase   // tetrahedral mesh in a 3D space
{
public:
    typedef topology::Tetrahedron Cell;    // cell type

    enum { dimension = Cell::dimension }; // mesh dimension = topological dimension of the cell
    enum { dimWorld = dimension };        // dimension of the underlying space
};
```

The `Mesh` object is then declared by:

```
typedef Mesh<MeshConfig> MeshType; // assigns a shorter name to Mesh<MeshConfig>
MeshType mesh;
```

To load a mesh from a file, write:

```
mesh.load("input.vtk"); // loads the mesh from file input.vtk
```

Access to mesh entities is gained by ranges. A range of the n-dimensional entities is obtained calling:

```
typedef typename EntityRange<MeshType, n>::Type EntityRangeType;
EntityRangeType entityRange = entities<n>(mesh);
```

Individual entities from the range are addressed using brackets. All the entities of the range are iterated as follows:

```
typedef typename EntityRangeType::DataType  EntityType;
typedef typename EntityRangeType::IndexType IndexType;
for (IndexType i = 0; i < entityRange.size(); i++)
{
    const EntityType &entity = entityRange[i];
    // do something with entity here
}
```

The range of the k-dimensional border entities of a mesh entity is obtained by:

```
typedef typename BorderRange<EntityType, k>::Type BorderRangeType;
BorderRangeType borderRange = borderEntities<k>(entity); // entity is of type EntityType
```

Similarly, the range of the k-dimensional coborder entities of a mesh entity is retrieved by:

```
typedef typename CoborderRange<EntityType, k>::Type CoborderRangeType;
CoborderRangeType coborderRange = coborderEntities<k>(entity);
```

Unlike other entities, vertices hold the information about their location in the underlying space. This information is accessible using the `getPoint()` method:

```
typedef typename VertexType::PointType PointType;
const PointType &point = vertex.getPoint(); // vertex is of type VertexType
// point[d] is the d^th vertex coordinate
```

The library is able to optimize the storage scheme for the mesh to reduce its memory consumption. By default, all the entities are explicitly stored in memory. If there is no need for entities of dimension n (other than cells and vertices), their creation and storage can be avoided by adding the following to the mesh configuration class:

```
template<> struct EntityStorage<MeshConfig, n> { enum { enabled = false }; };
```

If the storage of some entities is disabled and a corresponding call to the `entities()` function is made, the code fails to compile.

Similarly, the storage of the k-dimensional border entities (other than vertices) of, e.g, tetrahedra can be disabled using:

```
template<> struct BorderStorage<MeshConfig, topology::Tetrahedron, k>
{ enum { enabled = false }; };
```

As opposed to the border entities, the coborder entities are not stored by default. The storage of the k-dimensional coborder entities of, e.g., triangles can be enabled by:

```
template<> struct CoborderStorage<MeshConfig, topology::Triangle, k>
{ enum { enabled = true }; };
```

For example, the complete mesh configuration for a hexahedral mesh in a four-dimensional space, without the edges, without the border facets of the cells and with the coborder cells of the facets is:

```
class ExampleMeshConfig : public MeshConfigBase
{
public:
    typedef topology::Hexahedron Cell;

    enum { dimension = Cell::dimension };
    enum { dimWorld = 4 };
};

template<> struct EntityStorage<ExampleMeshConfig, 1>
{ enum { enabled = false }; };
```

```
template<> struct BorderStorage<ExampleMeshConfig, topology::Hexahedron, 2>
{ enum { enabled = false }; };

template<> struct CoborderStorage<ExampleMeshConfig, topology::Quadrilateral, 3>
{ enum { enabled = true }; };
```

4 Conclusion

The library is implemented in standard C++ with an extensive use of templates and inheritance. It has no dependencies on external libraries. It is now ready to be tested on the numerical solution of some example problems. If the library proves useful for the implementation of the numerical solvers, we will attempt to adapt it to GPUs.

References

[1] P. Bastian et al. *The Distributed and Unified Numerics Environment (DUNE) Grid Interface HOWTO*, version 2.3-svn, (September 2012). Downloaded from `http://www.dune-project.org/doc/`.

[2] P. Bauer, V. Klement, P. Strachota, and V. Žabka. *Numerical study of flow in a 2D boiler*. In 'Algoritmy 2012', A. Handlovičová, Z. Minarechová, and D. Ševčovič, (eds.), 172–178. Slovak University of Technology in Bratislava, Faculty of Civil Engineering, Department of Mathematics and Descriptive Geometry, (2012).

[3] P. Bauer, V. Klement, and V. Žabka. *FEM for flow and pollution transport in urban canopy*. In 'SNA 2012', 9–12. Technical University of Liberec, (2012).

[4] G. Berti. *GrAL—the grid algorithms library*. Future Generation Computer Systems **22** (2006), 110–122.

[5] M. Botsch, S. Steinberg, S. Bischoff, and L. Kobbelt. *OpenMesh – a generic and efficient polygon mesh data structure*. In '1st OpenSG Symposium', (2002).

[6] The CGAL Project. *CGAL user and reference manual*, version 4.0, (September 2012). Available at `http://www.cgal.org/Manual/4.0/doc_html/cgal_manual/packages.html`.

[7] M. B. Giles, G. R. Mudalige, Z. Sharif, G. Markall, and P. H. J. Kelly. *Performance analysis and optimisation of the OP2 framework on many-core architectures*. Computer Journal **55** (2012), 168–180.

[8] Institute for Microelectronics and Institute for Analysis and Scientific Computing, TU Wien. *ViennaGrid 1.0.1 user manual*, (August 2012). Downloaded from `http://viennagrid.sourceforge.net/`.

[9] B. S. Kirk, J. W. Peterson, R. H. Stogner, and G. F. Carey. *libMesh: A C++ library for parallel adaptive mesh refinement/coarsening simulations*. Engineering with Computers **22** (2006), 237–254.

[10] NVIDIA Corporation. *NVIDIA CUDA C programming guide*, version 4.2, (April 2012). Downloaded from `http://developer.nvidia.com/`.

Model of Soil Freezing*

Alexandr Žák

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

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

Abstract. This contribution is an extended abstract of the paper [1]. The paper studies a mathematical model of two-dimensional two-phase system describing a rectangular cross-section of saturated soil sample in time by its temperature. The model setting and its variational formulation are presented and a basic analysis of mathematical properties of the problem solution is provided. The model is used to control the structural conditions in the medium by coupling to Navier's equations. The computational studies of this coupling are presented at the end of the article.

Keywords: analysis, freezing, model, phase-change, soil

Abstrakt. Tento příspěvek je rozšířeným abstraktem článku [1]. Článek studuje matematický model dvoudimenzionálního dvoufázového systému popisujícího svojí teplotou v čase obdélníkový průřez vzorku saturované půdy. Je představeno nastavení modelu spolu s jeho variační formulací a je poskytnuta zakladní analýza matematických vlastností problému. Model je využit pro řízení strukturálních podmínek v médiu pomocí propojení s Navierovými rovnicemi. Závěrem jsou ukázány počítačové studie tohoto modelového propojení.

Klíčová slova: analýza, zamrzání, model, fázová změna, půda

1 Introduction

As a consequence of seasonal alternation, soil freezing and thawing occur in many regions of the globe, which stimulates structural changes in the upper layers of the saturated soils causing upward movements of the ground surface. The changes differ in rate and depth of occurrence according to the soil properties and the local environmental conditions. The natural process causing them is referred to as frost heave.

The principal cause of frost heave was ascribed by Taber in 1929 ([2]) to the formation of ice lenses in the neighbourhood of the frozen and unfrozen soil material interface. The ice lense growth is due to both the capillarity effect and the regelation mechanism. Referring to the dependence on one of the mechanisms, the terms primary heaving and secondary heaving, respectively, are used. The secondary heaving mechanism was described by Miller, [3], and then the first complex models of frost heave considering the

*Partial support of the project of the "Numerical Methods for Multi-phase Flow and Transport in Subsurface Environmental Applications, project of Czech Ministry of Education, Youth and Sports Kontakt ME10009, 2010-2012" and of the project "Advanced Supercomputing Methods for Implementation of Mathematical Models, project of the Student Grant Agency of the Czech Technical University in Prague No. SGS11/161/OHK4/3T/14, 2011-13".

processes at the microscopic level followed (e.g., Gilpin 1980, [4], O'Neil and Miller 1985, [5], Fowler 1989, [6]). An opposite approach to frost heave modelling can be found in the constitutive models using the definition of frost susceptibility as a property of soil (Michalowski 1993, [7], Michalowski and Zhu 2005, [8]).

2 The heat model

Let Ω be the rectangular domain $]0, x_1[\times]z_1, 0[$ and Q denote $]0, T[\times\Omega$ for some $T > 0$. Similarly to [9], the modified heat equation for the soil temperature u (in °C) which covers the phase change in a neighborhood of the freezing point depression $u^\star$, $u^\star < 0$, is considered. The equation has the form

$$C\frac{\partial}{\partial t}u(t,\mathbf{x}) + L\frac{\partial}{\partial t}\theta(u) = \lambda\Delta u(t,\mathbf{x})\,, \quad (t,\mathbf{x}) \in Q\,, \tag{1}$$

where C, L, λ are, for simplicity, constants, which have the meaning of the volumetric heat capacity, the volumetric latent heat of fusion water and thermal conductivity, respectively. The volumetric water content is described by the power function θ,

$$\theta(u) = \eta\phi(u)\,, \quad \phi(u) = \begin{cases} 1 & : \quad u \geq u^\star \\ \frac{|u^\star|^b}{|u|^b} & : \quad u < u^\star \end{cases}\,,$$

where η is the soil porosity of melt-state soil, ϕ represents the liquid pore water fraction and b is a positive constant related to the material characteristic of the soil. The equation is then supplemented by the initial temperature distribution

$$u(0,\mathbf{x}) = u_0(\mathbf{x})\,, \quad \mathbf{x} \in \bar{\Omega}\,, \tag{2}$$

and, for simplicity, the homogeneous Dirichlet boundary conditions can be assumed

$$u = 0\,, \quad \mathbf{x} \in \partial\Omega,\ t \in]0, T[\,. \tag{3}$$

Considering the model settings, it is possible to find an analogy between this problem and the Stefan problem ([10], [11], [12]).

3 Mathematical analysis

3.1 Enthalpy formulation

For the purpose of the mathematical analysis, it is convenient to pass to the enthalpy formulation of equation (1)

$$\frac{\partial}{\partial t}H(u) = \lambda\Delta u\,, \tag{4}$$

which can be obtained by the substitution

$$H(u) = \int_{u_{min}}^{u} C\,\mathrm{d}\xi + L\theta(u)$$

on the left-hand side, where u_{min} is a constant value.

Note that H is continuous and its first derivative is continuous everywhere except for $u^\star$. The value $u^\star$ becomes a singularity of equation (1) or (4).

3.2 Variational formulation

To be possible to define a weak formulation of the problem, equation (4) is multiplied by test functions v from $\mathcal{C}^2(Q) \cap \mathcal{C}^1(\bar{Q})$ which are furthermore zero for all $\mathbf{x} \in \partial\Omega$, $t \in [0,T]$ and for all $\mathbf{x} \in \bar{\Omega}$, $t = T$ and integrate it over Q. Using the Green formula, we gradually get:

$$
\begin{aligned}
0 &= \int_Q \left(\frac{\partial}{\partial t} H(u) v - \lambda \Delta u v \right) \mathrm{d}\mathbf{x}\mathrm{d}t \,, \\
0 &= \int_Q \left(\frac{\partial}{\partial t} H(u) v + \lambda \nabla u \nabla v \right) \mathrm{d}\mathbf{x}\mathrm{d}t - \lambda \int_0^T \int_{\partial\Omega} \nabla u \vec{n} v \, \mathrm{d}s\mathrm{d}t \,, \\
0 &= \left[\int_\Omega H(u) v \, \mathrm{d}\mathbf{x} \right]_0^T - \int_Q \left(H(u) \frac{\partial}{\partial t} v - \lambda \nabla u \nabla v \right) \mathrm{d}\mathbf{x}\mathrm{d}t \,, \\
0 &= \int_Q \left(H(u) \frac{\partial}{\partial t} v - \lambda \nabla u \nabla v \right) \mathrm{d}\mathbf{x}\mathrm{d}t + \int_\Omega H(u_0(\mathbf{x})) v(0,\mathbf{x}) \, \mathrm{d}\mathbf{x} \,.
\end{aligned}
$$

Now, it is possible to define the weak solution.

Definition 3.1. *The weak solution of problem (4) with (2), (3) is the function $u \in H^1(Q)$ which satisfies relation (5) for all test functions $v \in \mathcal{C}^2(Q) \cap \mathcal{C}^1(\bar{Q})$, $v = 0$ for $\forall \boldsymbol{x} \in \partial\Omega$, $t \in [0,T]$ and for $\forall \boldsymbol{x} \in \bar{\Omega}$, $t = T$.*

3.3 Existence of solution

The sequence of auxiliary problems with mollified functions H_k which have everywhere continuous first derivative and their limit is the function H is considered as follows:

$$
\begin{aligned}
\frac{\partial}{\partial t} H_k(u) &= \lambda \Delta u \,, \\
u(0) &= u_0 \,, \\
u|_{\partial\Omega} &= 0 \,,
\end{aligned}
$$

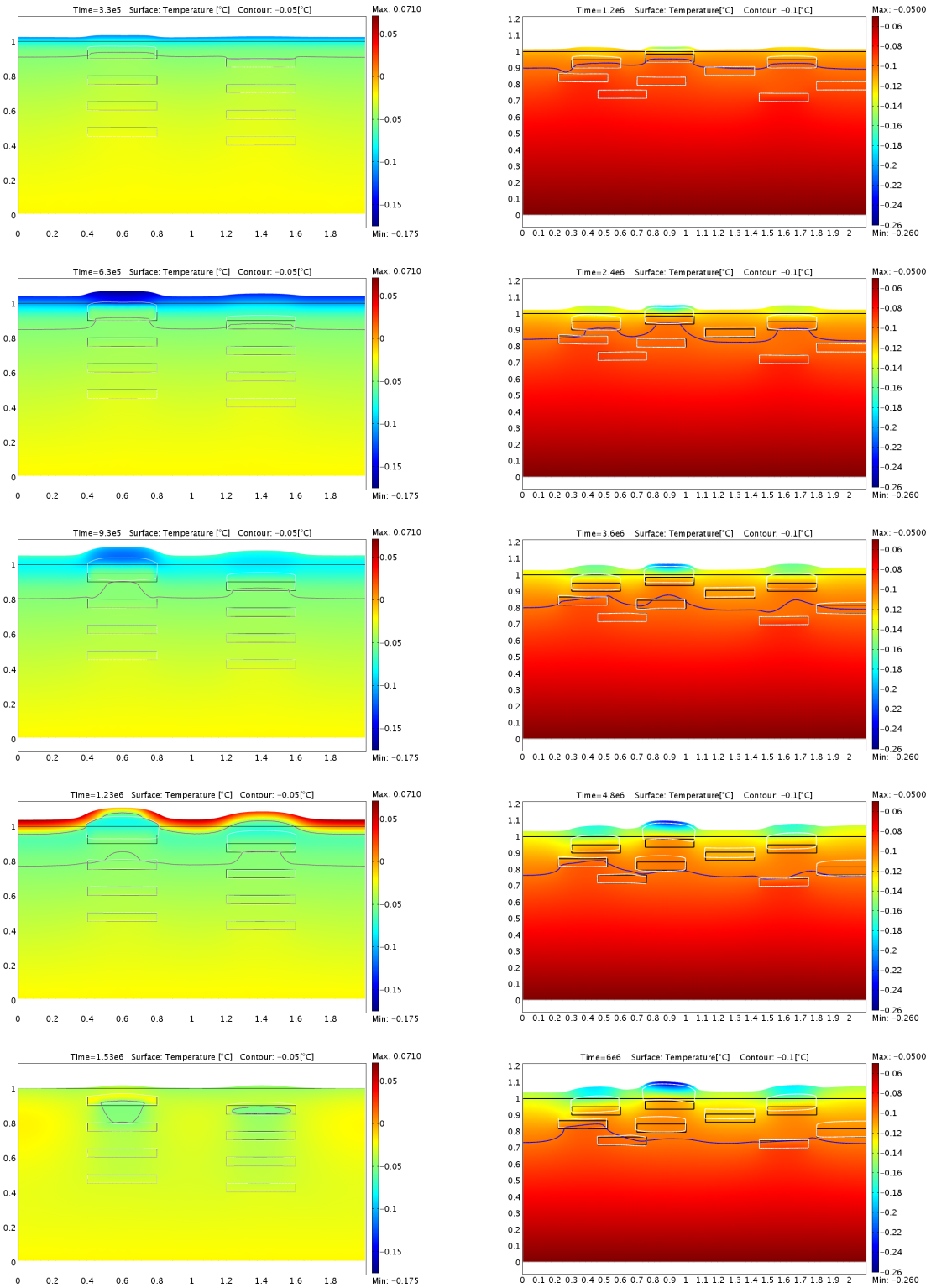
where $k \in \mathbb{N}$, and then the limit of the problems solutions $\{u^k\}_{n\in\mathbb{N}}$ is studied. Using the Galerkin method and appropriate a priori estimates, it is possible to show that the solutions exist and the sequence $\{u^k\}_{n\in\mathbb{N}}$ has a weak limit. Due to this fact and the discussion of the convergence of the nonlinear terms in the equations, it can be seen that it is eligible to pass in equation

$$
0 = \int_Q \left(H_k(u^k) \frac{\partial}{\partial t} v - \lambda \nabla u^k \nabla v \right) \mathrm{d}\mathbf{x}\mathrm{d}t + \int_\Omega H_k(u_0(\mathbf{x})) v(0,\mathbf{x}) \, \mathrm{d}\mathbf{x}
$$

to the weak limit and the limit of the solutions sequence is the weak solution of the original problem. In addition, it can be found that the solution is unique.

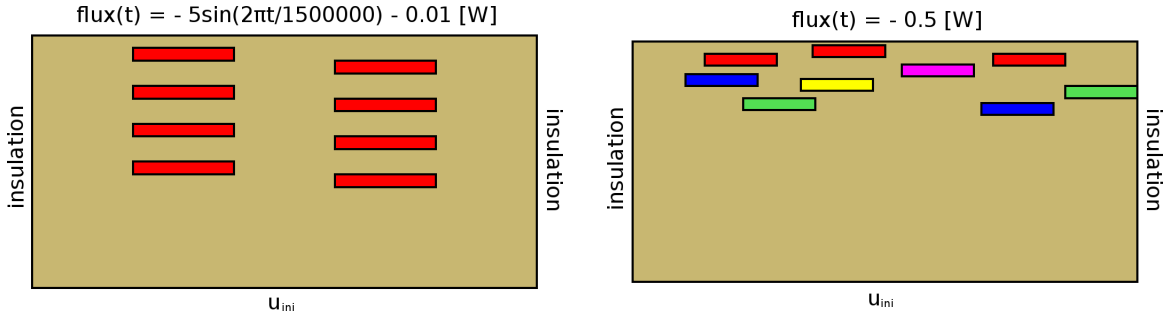
4 Computational Studies

Capturing the empirical knowledge that freezing water in a fixed volume increases abruptly the inner stress, the following switch function can be used to couple the temperature and



(a) (b)

Figure 1: Soil freezing deformation effects.



(a) (b)

Figure 2: The initial conditions and the thermal boundary conditions.

the position:

$$\xi(u) = \chi\vartheta(u^\star - u)\,, \tag{5}$$

where χ is internal stress rate expressing the value of the jump in stress during the cooling the material below $u^\star$ and where ϑ denotes the Heaviside step function. Assuming the soil material as continuum and the stress change induced by (5), Navier's equations for the position vector (v, w) are added as follows

$$\varrho\frac{\partial^2}{\partial t^2}\begin{bmatrix} v \\ w \end{bmatrix} + E\nabla\cdot\Gamma = 0\,, \tag{6}$$

where

$$\Gamma = \begin{bmatrix} \frac{(\nu-1)\frac{\partial}{\partial x}v - \nu\frac{\partial}{\partial y}w}{(1+\nu)(1-2\nu)} + \frac{\xi(T)}{E}\,, & \frac{-1}{2(1+\nu)}\left(\frac{\partial}{\partial y}v + \frac{\partial}{\partial x}w\right) \\ \frac{-1}{2(1+\nu)}\left(\frac{\partial}{\partial y}v + \frac{\partial}{\partial x}w\right)\,, & \frac{-\nu\frac{\partial}{\partial x}v + (\nu-1)\frac{\partial}{\partial y}w}{(1+\nu)(1-2\nu)} + \frac{\xi(T)}{E} \end{bmatrix},$$

E is Young's modulus and ν is Poisson's ratio.

The model governed by (1) and (6) and supplemented by the convenient boundary and initial conditions serves as a simple phase and structure change model. Several computational studies of this model with heterogeneities in the thermal and mechanical properties have been made; see Figure 1a and 1b. The simulation settings are illustrated in Figure 2a and 2b, where inner rectangles denote the distribution of the heterogeneities; the side and bottom boundaries are considered to be fixed.

References

[1] A. Žák and M. Beneš and T. H. Illangasekare, "Analysis of Model of Soil Freezing and Thawing", IAENG International Journal of Applied Mathematics, submitted.

[2] S. Taber, "Frost Heaving", The Journal of Geology, Vol. 37, No. 5, pp. 428–461, Jul. - Aug. 1929.

[3] R. D. Miller, "Frost Heaving in Non-Colloidal Soils", in Proc. 3rd Int. Conference on Permafrost, pp. 707–713, 1978.

[4] R. R. Gilpin, "A Model for the Prediction of Ice Lensing and Frost Heave in Soils", Water Resources Research, Vol. 16, No. 5, pp. 918–930, 1980.

[5] K. O'Neill and R. D. Miller, "Exploration of a Rigid Ice Model of Frost Heave", Water Resources Research, Vol. 21, No. 3, pp. 281–296, 1985.

[6] A. C. Fowler, "Secondary Frost Heave in Freezing Soils", SIAM J. APPL. Math., Vol. 49, No. 4, pp. 991–1008, 1989.

[7] R. L. Michalowski, "A Constitutive Model of Saturated Soils for Frost Heave Simulations", Cold Region Science and Technology, Vol. 22, Is. 1, pp. 47–63, 1993.

[8] R. L. Michalowski and F. ASCE and M. Zhu, "Modeling and Simulation of Frost Heave Using Porosity Rate Function", Geomechanics II: Testing, Modeling, and Simulation, ASCE, pp. 178–187, 2005.

[9] D. Nicolsky, V. Romanovsky, G. Panteleev: *Estimation of soil thermal properties using in-situ temperature measurements in the active layer and permafrost* [online]. Permafrost Laboratory, Geophysical Institute, University of Alaska. Retrieved from: http://permafrost.gi.alaska.edu/content/data-assimilation.

[10] O. A. Olejnik, "One Method for Solving General Stefan Problem", Proc. of USSR Acad. of Sci., pp. 1054–1057, 1960 (in Russian).

[11] B. M. Budak and E. N. Solov'eva and A. B. Uspenskij, "Difference Method with Smoothing Factors for Solution of Stefan Problem", GVM&MF, pp. 828–840, 1965 (in Russian).

[12] A. Visintin, Models of Phase Transitions, Boston, USA: Birkäuser, 1996, ch. 5, pp. 123–152.

[13] S. L. Kamenomostskaja, "On Stefan's Problem", Math. Col. 53, pp. 489–514, 1961 (in Russian).

[14] M. Beneš, " Numerical Solution of Two-Dimensional Stefan Problem by Finite Difference Method", ACTA POLYTECHNICA, pp. 61–87, 1989 (in Czech).