DOKTORANDSKÉ DNY 2017

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

10. a 24. listopadu 2017

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2017
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 201, Vydání 1.

Seznam příspěvků

Předmluva

Workshop Doktorandské dny je každoročním setkáním doktorandů oboru Matematické inženýrství zajišťovaného na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze v rámci doktorského studijního programu Aplikace přírodních věd. Na výchově doktorandů se kromě kateder matematiky, fyziky a softwarového inženýrství podílejí i spřátelené ústavy Akademie věd ČR, zejména ÚTIA, ÚI, MÚ, FZÚ a ÚJF.

Témata prezentovaná našimi doktorandy se týkají především matematických modelů fyzikálních či socioekonomických procesů, ale také základního výzkumu v matematice a teoretické informatice.

Věříme, že i letošní ročník přinese prezentujícím důležitou zpětnou vazbu a podpoří jejich vědecký růst. Za podporu konání workshopu děkujeme Studentské grantové soutěži, projektu SVK 33/17/F4.

Editoři

Bayesian Source Term Determination with Unknown Covariance of Measurements*

Alkomiet Belal

2nd year of PGS, email: belalalk@fjfi.cvut.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, CAS

Abstract. Determination of a source term of release of a hazardous material into the atmosphere is a very important task for emergency response. We are concerned with the problem of estimation of the source term in the conventional linear inverse problem $\mathbf{y} = M\mathbf{x}$ is described using the source-receptor-sensitivity (SRS) matrix $\mathbf{M}$ and the unknown source term $\mathbf{x}$. Since the system is typically ill-conditioned, the problem is recast as an optimization problem

$$\min_{R,B} (y - Mx)^T R^{-1} (y - Mx) + x^T B^{-1} x. \tag{1}$$

The first term minimizes the error of the measurements with covariance matrix $\mathbf{R}$, and the second term is a regularization of the source term [2]. There are different types of regularization arising for different choices of matrices $\mathbf{R}$ and $\mathbf{B}$, for example, Tikhonov regularization assumes covariance matrix B as the identity matrix multiplied by scalar parameter.In this contribution, we adopt a Bayesian approach to make inference on the unknown source term $\mathbf{x}$ as well as unknown $\mathbf{R}$ and $\mathbf{B}$.We assume prior on $\mathbf{x}$ to be a Gaussian with zero mean and unknown diagonal covariance matrix $\mathbf{B}$.The covariance matrix of the likelihood $\mathbf{R}$ is also unknown. We consider two potential choices of the structure of the matrix $\mathbf{R}$. First is the diagonal matrix and the second is a locally correlated structure using information on topology of the measuring network. Since the inference of the model is intractable, iterative variational Bayes algorithm is used for simultaneous estimation of all model parameters. The practical usefulness of our contribution is demonstrated on an application of the resulting algorithm to real data from the European Tracer Experiment (ETEX).

Keywords: Bayesian inference, atmospheric transport model, inverse modeling

Abstrakt. Určení zdrojového členu úniku nebezpečného materialu do atmosféry je velmi důležitým úkolem pro krizové řízení vzniklé situace. Zabýváme se problémem odhadu zdrojového členu v běžném lineárním inverzním problému $\mathbf{y} = M\mathbf{x}$, který je definován pomocí matice citlivosti (source-receptor-sensitivity, SRS) $\mathbf{M}$ a neznámého vektrou zdrojového členu $\mathbf{x}$. Protože soustava lineárních rovnic je obvykle špatně podmíněna, problém je řešen jako optimalizační úloha s regularizací

$$\min_{R,B} (y - Mx)^T R^{-1} (y - Mx) + x^T B^{-1} x. \tag{2}$$

Prvni člen minimalizuje chybu měření pomocí kovarianční matice $\mathbf{R}$, a druhý je regularizace zdrojového členu. Existují různé typy regularizace pro různé možnosti matic $\mathbf{R}$ a $\mathbf{B}$, například

*This research is supported by EEA/Norwegian Financial Mechanism under project MSMT-28477/2014 SourceTerm Determination of Radionuclide Releases by Inverse Atmospheric Dispersion Modelling (STRADI).

Tichonovova regularizace, která předpokládá kovarianční matici B jako jednotkovou matici vynásobenou skalárním parametrem. V tomto příspěvku, používáme Bayesovský přístup k odvození jak zdrojového členu $\mathbf{x}$ tak neznámých matic $\mathbf{R}$ a $\mathbf{B}$. Předpokládáme, že apriorní rozložení $\mathbf{x}$ je Gaussovske s nulovou střední hodnotou a neznámou diagonální kovarianční maticí $\mathbf{B}$. Kovarianční matice $\mathbf{R}$ je také neznáma. Uvažujeme dvě možnosti výběru struktury matice $\mathbf{R}$. První je diagonální matice a druhá je lokálně korelovaná struktura využívající informaci o topologii měřicí na sítě. Vzhledem k tomu, že analytické řešení modelu neexistuje, používáme metodu variační Bayes pro simultánní odhad všech parametrů modelu. Praktická užitečnost našeho přístupu je demonstrována na datech z experimentu ETEX (European Tracer Experiment).

Klíčová slova: Bayesovská statistika, atmosférický transportní model, inverzní modelování

1 Introduction

The task of determination of a source term of an atmospheric pollutant is important in many situations such as radioactive release from nuclear power plants or emission of greenhouse gases.The source term is the vector of amounts of the pollutant released in regularly sample time.The location of the release is assumed to be known.Uncertainty in the source term is one of the largest source of errors in modeling and prediction of the pollutant dispersion in the atmosphere, hence, any improvement of the reliability of the source term estimation has significant impact The common approach for determination of the source term is to combine the data measured in the environment (e.g., radionuclide concentrations) with an atmospheric transport model.The quality of the estimated source term to a given measurements can be modeled and optimized using various approaches including the Bayesian approach[2]. Typically, the problem is formulated as a linear regression.The vector of measurements is assumed to be a product of a computed source-receptor-sensitivity (SRS) matrix determined using an atmospheric dispersion model and an unknown source term vector.

2 Bayesian inference

The process of inferring data from observations can be described by using Bayesian inference, Here we formalize a Bayesian inference framework to make use of the observations to infer the parameter values by updating our prior knowledge. This inferring process can be formalized using the Bayes' theorem:

$$p(\mathbf{x}, R, B|\mathbf{y}, M) = \frac{p(\mathbf{y}|\mathbf{x}, M)p(\mathbf{x}, B)p(R)p(B)}{\int p(\mathbf{y}|\mathbf{x}, M)p(\mathbf{x})d\mathbf{x}} \tag{3}$$

where $p(\mathbf{x})$ is the prior distribution, $p(\mathbf{y}|x, M)$ is the likelihood of the measurements. For the choice of Gaussian models [1]

$$p(\mathbf{y}|\mathbf{x}, M) = \mathcal{N}(M\mathbf{x}, R^{-1}), \quad p(\mathbf{x}|B) = \mathcal{N}(0, B^{-1}) \tag{4}$$

The result of the Bayes' theorem (2) is a Gaussian distribution $\mathcal{N}(\hat{x}, B^{-1})$, where $\hat{x}$ corresponds to the solution of the optimization problem (1).

3 Prior Models of Covariance Matrix of Source term

We use modified Cholesky factorization of a source term $\mathbf{x}$ unknown covariance matrix $\mathbf{B} = (W\Upsilon W^T)^{-1}$, where W is a lower diagonal matrix.We assume correlation only between the time adjacent parameters, i.e

$$W = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \mathbf{w}_1 & 1 & 0 & 0 \\ 0 & \ddots & 1 & 0 \\ 0 & 0 & \mathbf{w}_{m-1} & 1 \end{bmatrix}, \Upsilon = diag(\nu_i).$$

We define prior distribution of the model as follows:

$$\nu \sim \prod_{i=1}^{m} G(\nu_0, \rho_0), w \sim \prod_{i=1}^{m-1} N(w_0, \tau_0), \tau \sim G \prod_{i=1}^{m-1} (\omega_0, \kappa_0),$$

with selected prior constants $\nu_0, \rho_0, \tau_0, \omega_0, \kappa_0$. The system is that ill-conditioned is usually related to rapidly oscillation solutions, and using this structure for modeling the covariance matrix of source term favors in fact the smooth solutions.

4 Prior Models of Covariance Matrix of residue model

The main problem is the fact that small errors in the (SRS) lead to large errors in the source determination. The errors in this matrix are caused by inaccurate priori knowledge of meteorological conditions such as the wind field [1]. This can cause either spacial or temporal displacement of the model. We model spatially- and temporally-correlated matrix of the Gaussian distribution of the error. We consider the following structure of matrix R:

$$R = L^{\top} D L, \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & 1 & 0 & 0 \\ \mathbf{l}_1 & \vdots & 1 & 0 \\ \vdots & \mathbf{l}_k & \mathbf{l}_{n-1} & 1 \end{pmatrix}, \quad D = \begin{pmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & d_n \end{pmatrix}.$$

where the vectors of unknowns are $\mathbf{l}_1, \ldots, \mathbf{l}_{n-1}, \boldsymbol{d} = [d_1, \ldots, d_n]$. The Bayesian formalism requires to define prior distribution on all unknowns. We define prior distribution on all unknowns vectors $p(d_i) = G(a_0, b_0)$ and $p(\mathbf{l_j}|\psi_j){=}\mathcal{N}(\mathbf{l_0}, {\psi_j}^{-1})$. The spatial correlation matrix is designed by vectors $\mathbf{l_0}$. For example,we assume all elements in the vector $\mathbf{l_0}$ to have value (-1), if the distance between the measuring stations is less than 100 km, and zero otherwise.

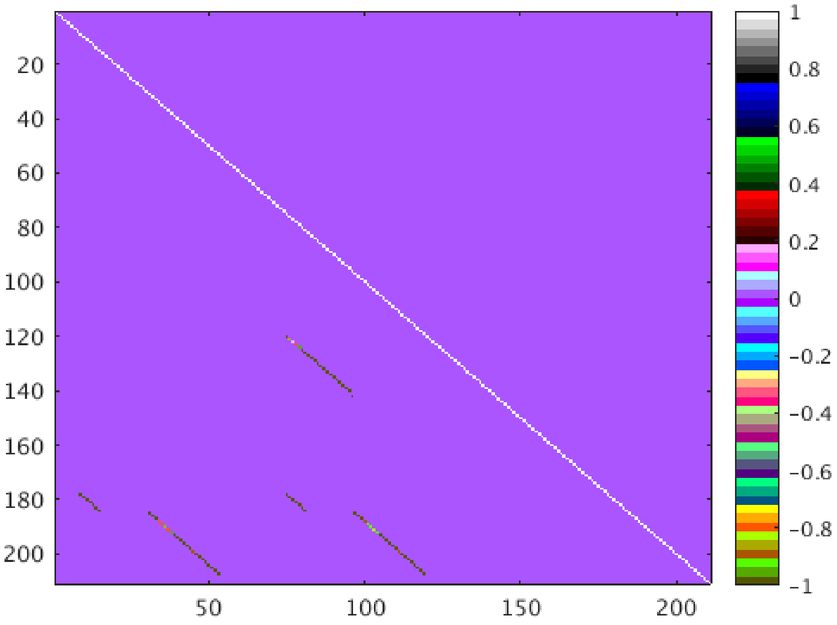


Figure 1: Estimated correlation matrix of residue model.

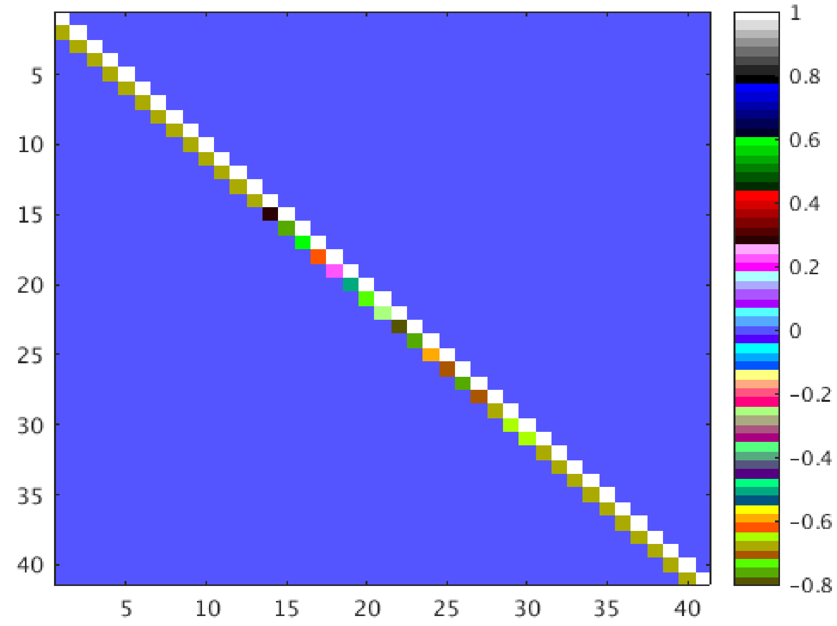


Figure 2: Estimated correlation matrix of Source term model.

5 Approximations of posterior distribution

The task is to calculate the posterior distribution of parameters and hyperparameters based on the Bayes' theorem (3) which gives the posterior probability of the parameters given the data and the model $p(\mathbf{x}|\mathbf{y}, M)$ where $p(\mathbf{x})$ is the prior distribution, $p(\mathbf{y}|x, M)$ is the likelihood of the measurements. The associated Byaes rule is

$$p(\mathbf{x}|\mathbf{y}, M) \propto p(\mathbf{y}|\mathbf{x}, M)p(\mathbf{x}), \tag{5}$$

where symbol $\propto$ denotes equality up to a normalizing constant.

It may not be possible to evaluate the posterior probability distribution analytically. Minimising the Kullback-Liebler divergence (KL distance), also known as the Relative Entropy, between the solution and the hypothetical true posterior, leads to a set of implicit equations which have to be solved iteratively and convergence to local minima is guaranteed [3]. To avoid negative results, truncated normal of prior $p(x)$ to positive domain are considered, to enforce the positivity of the retrieved source term:

$$p(x_j) = t\mathcal{N}(0, \sigma_{x_j}^{-1}, \langle 0, \infty \rangle),$$

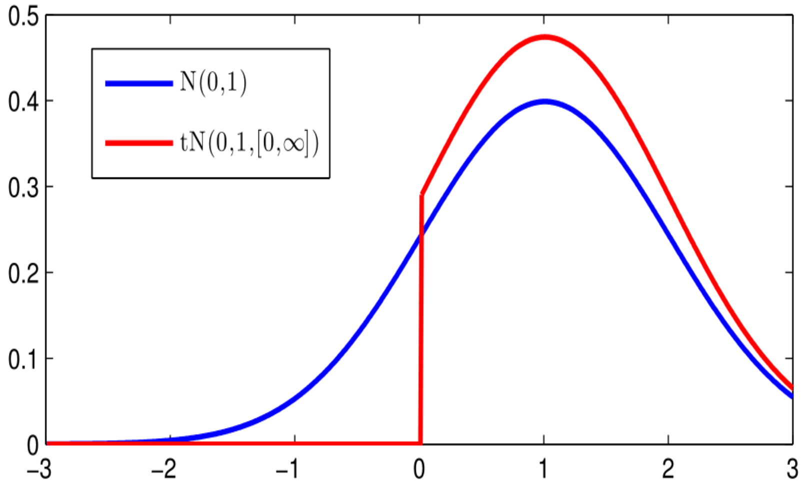


Figure 3: Example of the normal distribution $\mathcal{N}(1, 1)$, blue line, and the truncated normal distribution $t\mathcal{N}(1, 1, < 0, 3 >)$, red line.

6 Experiment

The European tracer experiment (ETEX) were two releases of perfluorocarbon that took place in autumn of 1994 in north-western part of France. These releases were tracked across Europe using a network of 168 ground stations with limited airborne support. The aim of the experiment was to simulate an emergency response situation for meteorological modellers whose task was to create long-range dispersion prediction models in real time. In the first one, 340kg of perfluorocarbon was released in range of 12 hours.

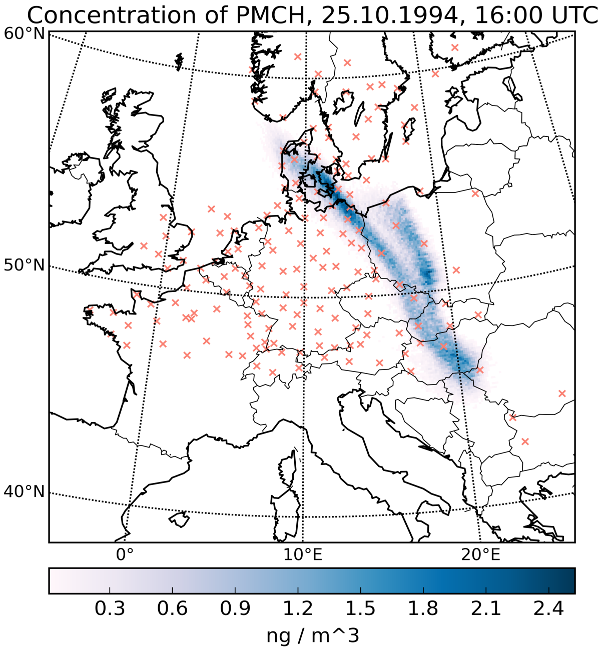


Figure 4: Domain of the ETEX experiment with source (red triangle) and receptors (blue crosses).

7 Example results

We study three models:

- independent source term and residue models. $\mathbf{R} = \mathbf{w}^{-1}\mathbf{l_p}$, $\mathbf{B} = \Upsilon^{-1}$
- correlated source term model with independent residue model. $\mathbf{R} = \mathbf{w}^{-1}\mathbf{l_p}$, $\mathbf{B} = (W\Upsilon W^T)^{-1}$
- correlated source term model with correlated residue model $\mathbf{R} = (LDL^T)^{-1}$, $\mathbf{B} = (W\Upsilon W^T)^{-1}$

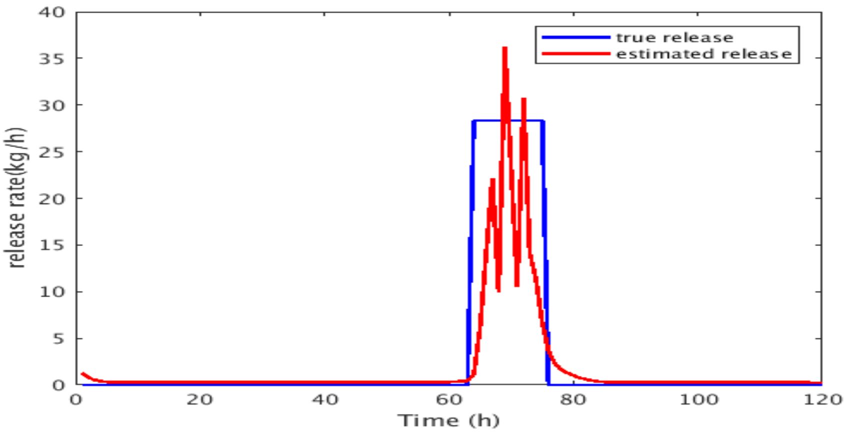


Figure 5: Shows estiamated source term with correlated source term and residue models.

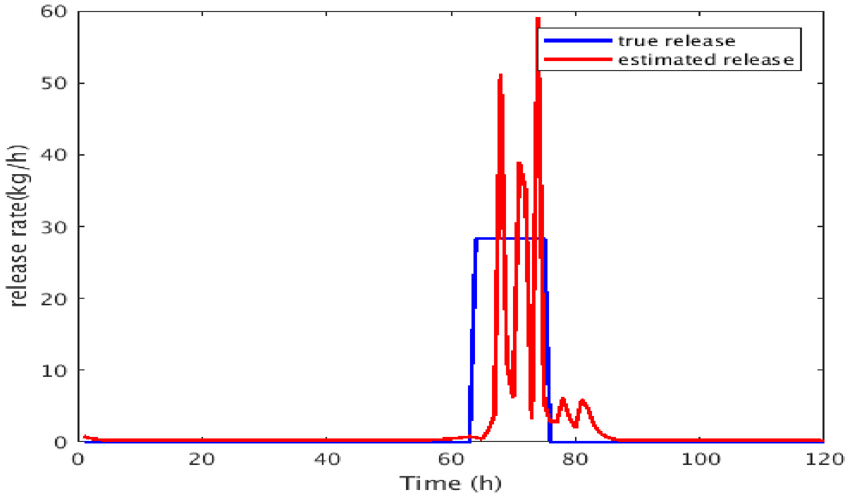


Figure 6: Shows estiamated source term with correlated source term model and independent residue model.

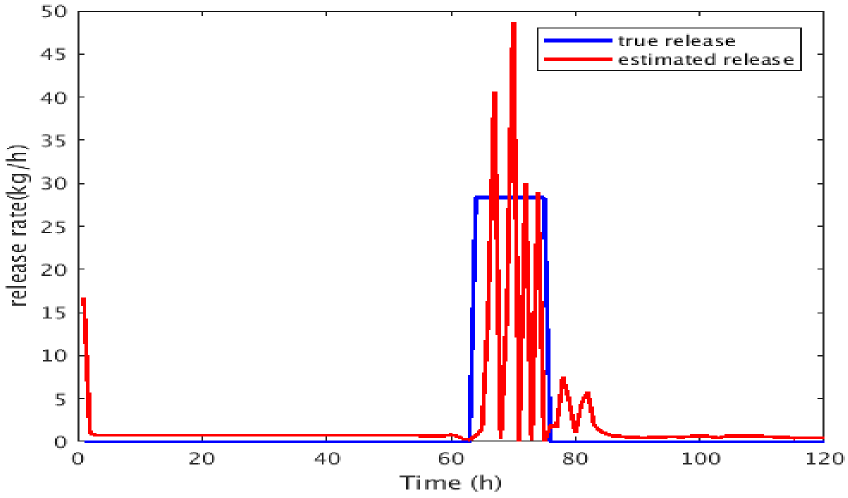


Figure 7: Shows estiamated source term with correlated source term model and correlatedt residue model.

8 Conclusions

- The models of linear regression with two prior models of covariance matrix of residue model and source term are studied.

- If smooth solutions are preferred, a model of correlated source term could be appropriate.

- The models of covariance matrix of residue source term model is estimated from the observations.

References

[1] Bishop, Christopher M *Pattern recognition and machine learnin*, springer, (2006).

[2] Tichý, Ondrej and Smídl, Václav and Hofman, Radek and Stohl, Andreas *LS-APC v1.0: a tuning-free method for the linear inverse problem and its application to source-term determination*, springer, (2016).

[3] Šmídl, V and Quinn, A *The Variational Bayes Method in Signal Processing*, Springer-Verlag, Berlin/Heidelberg, (2006).

Modified Homogeneity Testing for Weighted Data*

Petr Bouř

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

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

Abstract. We introduce several modifications of classical statistical tests applicable to weighted data sets in order to test homogeneity of weighted and unweighted samples, e.g. Monte Carlo simulations compared to the real data measurements. The asymptotic approximation of p-value and power of our weighted variants of homogeneity tests are investigated by means of simulation experiments. The simulation is performed for various probability distributions of samples. Finally, our methods of homogeneity testing are applied to Monte Carlo samples and real data sets measured at the particle accelerator Tevatron in Fermilab at DZero experiment originating from top-antitop quark pair production in two decay channels (electron, muon) with 2, 3 or 4+ jets detected. Consequently, the final variable selection is carried out and the resulting subsets chosen from 46 dimensional physical parameters are recommended for further top quark cross section analysis.

Keywords: statistical homogeneity testing, data weighting, top quark

Abstrakt. Je představeno několik modifikací klasických statistických testů pro vážená pozorování za účelem testování homogenity rozdělení váženého a neváženého vzorku, tj. Monte Carlo simulace v porovnání se skutečně naměřenými daty. Řadou simulačních experimentů je prověřena asymptotická aproximace p-hodnoty i síla vážených variant testů homogenity. Výslednými metodami jsou porovnány vzorky Monte Carlo simulace a skutečná data naměřená na částicovém urychlovači Tevatron ve Fermilabu při experimentu DZero pocházející z produkce páru top-antitop kvarku ve dvou rozpadových kanálech (elektron a mion) se 2, 3 nebo 4 a více jety. Následně je provedena finální selekce vhodných fyzikálních proměnných. Tato podmnožina ze 46 kompletních parametrů je doporučena pro další analýzu účinného průřezu top kvarku.

Klíčová slova: statistické testování homogenity, vážení dat, top kvark

1 Introduction

Homogeneity testing is an important step in many analysis techniques, particularly in machine learning (ML) applications in physics research. It is often the case that physicists apply a field-specific data preprocessing procedure called data weighting. Via assigning weights $w_1, \ldots, w_n > 0$ to simulated observations $x_1, \ldots, x_n$, they are able to fine-tune

[*] This work has been supported by the grants LG15047 (MYES), LM2015068 (MYES), SGS15/214/OHK4/3T/14 (CTU) and GA16-09848S (GACR).

their Monte Carlo (MC) simulation dataset so that it meets their requirements. A typical example is shifting data distribution so that the resulting distribution is positively skewed. However, theory concerning statistical homogeneity tests does not handle any weighting procedures, nor associates weights with observations. Therefore, the classical homogeneity tests must be adjusted for weighted datasets. Despite relatively straightforward incorporation of weights into the classical homogeneity tests and their modification, asymptotic properties of these tests can be no longer guaranteed. Thus, our goal is to investigate the validity of asymptotic properties of homogeneity testing for weighted observations.

The underlying problem the physicist might require us to solve may be a simple signal/backgrounds binary classification task. In this typical ML application, we often use MC simulation for both training and testing our ML classifier. We may then apply the trained classifier to a real measured dataset (DATA). Naturally, we expect both MC $\sim F$ and DATA $\sim G$ to be identically distributed: $F \equiv G$. Otherwise, the classification model will not perform well. Thus, we indeed need to test homogeneity of MC and DATA prior to the modelling step.

2 Weighted Tests of Homogeneity

Prior to subsequent utilization of ML methods, it is vital to guarantee homogeneity of DATA and MC distributions. For this purpose, we first define an analogy with empirical distribution function (EDF) for weighted data set.

Definition 1. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be iid random variables distributed by cumulative distribution function (CDF) $F(x)$ and let $(w_1, \ldots, w_n)$ be respective weights, where $W = \sum_{i=1}^{n} w_i$. We define the weighted empirical distribution function (WEDF) to be*

$$F_n^W(x) = \frac{1}{W} \sum_{i=1}^{n} w_i I_{(-\infty,x]}(X_i), \tag{1}$$

where $I_A(X)$ is the indicator of the set A.

Remark 1. In the case of $w_i = 1$ for all $i \in \widehat{n}$, that is the unweighted DATA, the definition of WEDF goes over to usual EDF.

In order to avoid an investigation of an unknown parametric family, we shall pursue our homogeneity testing only with nonparametric approaches. Thus, proceeding further in this section, we present the Kolmogorov-Smirnov test based upon EDFs of two data sets $\boldsymbol{X}_1 = \left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}\right)$, $\boldsymbol{X}_2 = \left(X_1^{(2)}, \ldots, X_{n_2}^{(2)}\right)$, with respective distribution functions F, G. Also, we provide another class of nonparametric tests based upon ϕ-divergences, with the purpose of verifying precedent homogeneity results. By the homogeneity hypothesis, as our null hypothesis is H_0, we understand

$$H_0 : F = G \qquad \text{vs} \qquad H_1 : F \neq G \qquad \text{at significance level} \quad \alpha \in (0, 1). \tag{2}$$

We require our homogeneity tests to meet the condition

$$P(W_C | H_0) \leq \alpha, \tag{3}$$

where W_C is a critical region for the specific test statistic T, i.e. we reject hypothesis H_0 if $T \in W_C$.

The nature of homogeneity testing prompted us to look for the p-value, i.e., the lowest significance level α for which we reject hypothesis H_0. Thus, for every $\alpha > p$-value we may automatically reject hypothesis H_0.

2.1 Two Sample Kolmogorov-Smirnov Test

Let F_{n_1}, G_{n_2} denote the EDFs of the two data samples $\boldsymbol{X}_1, \boldsymbol{X}_2$ with respective sample sizes n_1, n_2. We consider the test statistic

$$D_{n_1,n_2} = \sup_{x\in\mathbb{R}} |F_{n_1}(x) - G_{n_2}(x)|. \tag{4}$$

It is clear from the Glivenko-Cantelli lemma that under the true H_0 it holds $D_{n_1,n_2} \xrightarrow{a.s.} 0$ for $n_1, n_2 \to \infty$. Furthermore, due to [6] it holds for the true H_0 and $\lambda > 0$ that

$$\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2} \le \lambda\right) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2\lambda^2}. \tag{5}$$

Therefore, we obtain the approximate p-value as $2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda_0^2}$, where $\lambda_0 = \sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2}$.

However, for weighted data sample we are forced to replace EDFs F_{n_1}, G_{n_2}, and the numbers of entries n_1, n_2, with their respective WEDFs $F_{n_1}^{W_1}, G_{n_2}^{W_2}$, and the sums of weights W_1, W_2 in (4) and (5). Instead of (4), we thus obtain the test statistic

$$D_{n_1,n_2}^{W_1,W_2} = \sup_{x\in\mathbb{R}} \left|F_{n_1}^{W_1}(x) - G_{n_2}^{W_2}(x)\right|. \tag{6}$$

Remark 2. The Definition 1 of WEDF makes it clear that the statistic $D_{n_1,n_2}^{W_1,W_2} \xrightarrow{a.s.} 0$ for $n_1, n_2 \to \infty$ and $W_1, W_2 \to \infty$. Nevertheless, it is important to notice some of the weaknesses inherent in the above approach. This modified test for the weighted data sample does not have to obey the asymptotic property (5). Let us stress that the p-value obtained using the statistic $D_{n_1,n_2}^{W_1,W_2}$ can not be considered a regular approximate p-value without subsequent detailed research. This is why we propose numerical verification of our approach in Section 3.

2.2 Divergence Tests of Homogeneity

This particular class of tests converts the problem (2) to testing homogeneity in multinomial populations. It does not utilize the EDF and therefore serves as an independent verification. We recall our notation of two samples $\boldsymbol{X}_1 = \left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}\right)$, $\boldsymbol{X}_2 = \left(X_1^{(2)}, \ldots, X_{n_2}^{(2)}\right)$, and the pooled sample $\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ with $N = n_1 + n_2$ observations. Let $\{t_0, \ldots, t_k\}$ denote a partition of $\mathbb{R}$ such that for all $x \in \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ it holds $x \in [t_0, t_k]$. Hereby we make binning over the populations $\boldsymbol{X}_1, \boldsymbol{X}_2$ consisting of k bins. For $i \in \{1, 2\}$

and $j \in \widehat{k}$ we denote by p_{ij} the probability that a randomly chosen observation from $\boldsymbol{X}_i$ lies in the j-th bin $[t_{j-1}, t_j]$. Instead of (2) we now test equivalently the hypothesis

$$H_0 : p_{1j} = p_{2j} \qquad \text{for all} \quad j \in \widehat{k} \qquad \text{vs} \qquad H_1 : H_0 \text{ is not true.} \tag{7}$$

For $i \in \{1, 2\}$ it holds $\sum_{j=1}^{k} p_{ij} = 1$. This will provide us with the $k - 1$ free parameters for each sample $\boldsymbol{X}_1, \boldsymbol{X}_2$. Thus let us denote the free parameters by $\boldsymbol{\theta}_i = (p_{i1}, \dots, p_{i(k-1)})$. The parametric space of the dimension $2(k - 1)$ for the task (7) is therefore generated by

$$\Theta = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (p_{11}, \dots, p_{1(k-1)}, p_{21}, \dots, p_{2(k-1)})\}. \tag{8}$$

Under the true H_0 we carry out the maximum likelihood estimate (MLE)

$$\widehat{\boldsymbol{\theta}} = \left(\frac{N_1}{N}, \dots, \frac{N_{k-1}}{N}, \frac{N_1}{N}, \dots, \frac{N_{k-1}}{N} \right), \tag{9}$$

where N_j stands for the number of observations $x \in \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ lying in j-th bin. In what follows, for $i \in \{1, 2\}$ we write $\boldsymbol{p}(\boldsymbol{\theta}_i) = (p_{i1}, \dots, p_{ik})$ for the vector of probabilities assigned to the bins. Hence for $\boldsymbol{p}(\boldsymbol{\theta}_i)$ we have MLE

$$\boldsymbol{p}(\widehat{\boldsymbol{\theta}}_i) = \left(\frac{N_{i1}}{n_i}, \dots, \frac{N_{ik}}{n_i} \right), \tag{10}$$

where N_{ij} denotes the number of observations $x \in \boldsymbol{X}_i$ belonging to the j-th bin. First, we construct the vector of joint probabilities

$$\widehat{\boldsymbol{p}} = \left(\frac{n_1}{N} \boldsymbol{p}(\widehat{\boldsymbol{\theta}}_1), \frac{n_2}{N} \boldsymbol{p}(\widehat{\boldsymbol{\theta}}_2) \right) = \left(\frac{N_{11}}{N}, \dots, \frac{N_{1k}}{N}, \frac{N_{21}}{N}, \dots, \frac{N_{2k}}{N} \right). \tag{11}$$

Secondly, we consider the vector

$$\boldsymbol{p}^*(\boldsymbol{\theta}) = \left(\frac{n_1}{N} \boldsymbol{p}(\boldsymbol{\theta}_1), \frac{n_2}{N} \boldsymbol{p}(\boldsymbol{\theta}_2) \right) = \left(\frac{n_1}{N} p_{11}, \dots, \frac{n_1}{N} p_{1k}, \frac{n_2}{N} p_{21}, \dots, \frac{n_2}{N} p_{2k} \right). \tag{12}$$

Furthermore, adopting the definition of ϕ-divergence from [4], we arrive at

$$D_\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}^*(\boldsymbol{\theta})) = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{n_i}{N} p_{ij} \phi \left(\frac{N_{ij}}{n_i p_{ij}} \right), \tag{13}$$

where ϕ is a certain function selected from the convex family of real non-negative valued functions on $(0, \infty)$. We now apply the previous MLE $\widehat{\boldsymbol{\theta}}$ of (9) to $\boldsymbol{p}^*(\boldsymbol{\theta})$. Thereafter we can define the statistic of the divergence test of homogeneity

$$H_\phi(\widehat{\boldsymbol{\theta}}) = \frac{2N}{\phi''(1)} D_\phi \left(\widehat{\boldsymbol{p}}, \boldsymbol{p}^*(\widehat{\boldsymbol{\theta}}) \right) = \frac{2N}{\phi''(1)} \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{n_i}{N} \frac{N_j}{N} \phi \left(\frac{N_{ij} N}{n_i N_j} \right). \tag{14}$$

The distribution of (14) is $\chi^2(k - 1)$. Accordingly, the approximate p-value can be computed as

$$p\text{-value} = 1 - \chi^2_{(k-1)}(H_\phi(\widehat{\boldsymbol{\theta}})). \tag{15}$$

Remark 3. An important special case of (14) is the χ^2 test of homogeneity for $\phi(x) = \frac{1}{2}(x-1)^2$. Moreover, the test (14) coincides with the likelihood ratio test for $\phi(x) = x\log x - x + 1$. Notwithstanding, the $\chi^2(k-1)$ distribution of (14) holds independently on the underlying convex function ϕ (numerically verified in [1]). Throughout what follows, we shall use only the case of the χ^2 test of homogeneity though.

Remark 4. The statistic (14) makes it evident that the divergence test of homogeneity is dependent of the choice of the number of bins k as well as the subsequent binning $\{t_0, \ldots, t_k\}$. That is why we consider histograms with robust equiprobable binning from [1]. However, in order to carry out the construction of the test, we must make a judicious choice of k. Because of the large number of observations in DATA (or sum of weights W in MC, as $W \approx \#\text{DATA}$ for each ensemble under consideration), the test would loose its power with increasing bin number k, see [3], (numerically validated in [1]). Thus, we choose the following wise number of bins $k = \lceil 1 + \log_2 W \rceil$, due to [2]. Finally, let us state once more, we might want to supersede the members N_{ij}, N_j, n_i, N, in (14) with the corresponding sums of weights at the sacrifice of losing some control on the asymptotic property (15).

Figure 1 provides comparison of the divergence test of homogeneity (represented by χ^2 test) with the Kolmogorov-Smirnov and Anderson-Darling [5] test. Notice that the divergence tests produce generally higher p-values compared with the tests based on the EDF.

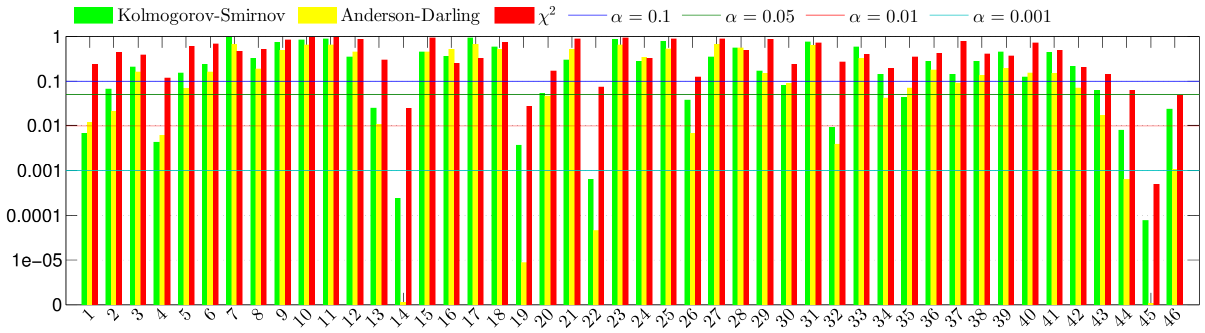


Figure 1: Homogeneity tests of MC and DATA distributions: p-value for all $m = 46$ variables in MC channel Electron 4+ Jets.

3 Simulation

3.1 Ensemble Modification and p-value Validation

In Remarks 2 and 4, we have already mentioned the problem of insecure asymptotic properties when applying weighted modifications of the standard tests. We now turn our attention over the numerical simulation. For our purposes here, the best way would be to validate the asymptotic properties using standard, unweighted tests. This requires us to plug into the testing an unweighted data set (only instead of weighted MC; DATA is unweighted already). We shall do this by appropriate transformation of the weighted ensemble MC into the unweighted ensemble $\mathrm{MC}_\dagger$.

We make two requirements for the transformation. Firstly, we desire to preserve or exploit information contained in weighting in MC. Since the weight of an observation states to what extent the distribution should be present in the neighbourhood of the observation. Secondly, we require that the sum of weights in MC corresponds to the number of observations in the unweighted $\mathrm{MC}_\dagger$. Continuing in this manner, we now proceed as follows.

Denote by $\boldsymbol{X} = \big(X_{(1)}, \ldots, X_{(n)}\big)$ the ordered sample in MC with weights $(w_1, \ldots, w_n)$ and let $W = \sum_{i=1}^{n} w_i$. Let $N = \lfloor W \rfloor$ denote the desired number of observations in the new transformed ensemble $\mathrm{MC}_\dagger$. Given both our requirements regarding $\mathrm{MC}_\dagger$, we are constructing special weighted averages from $\boldsymbol{X}$. For simplicity, we presume $0 \leq w_i \leq 1$ for all $i \in \widehat{n}$. Into the set intended for the first weighted average we include the smallest possible number of observations $\big(X_{(1)}, \ldots, X_{(k_1)}\big)$ such that

$$1 \leq \sum_{i=1}^{k_1} w_i < 2. \tag{16}$$

Thereby, for all $l < k_1$

$$\sum_{i=1}^{l} w_i < 1. \tag{17}$$

The portion of weight w_{k_1} of the observation $X_{(k_1)}$ which contributes above 1 to the sum (16) will not be included into the first weighted average. Hence, we denote this residual portion as

$$r_{k_1} = \sum_{i=1}^{k_1} w_i - 1. \tag{18}$$

Thereafter the first observation $Y_{(1)}$ in $\mathrm{MC}_\dagger$ can be defined as the following weighted average

$$Y_{(1)} = \frac{\sum_{i=1}^{k_1} X_{(i)} w_i - X_{(k_1)} r_{k_1}}{\sum_{i=1}^{k_1} w_i - r_{k_1}}. \tag{19}$$

From (18) we arrive at

$$Y_{(1)} = \sum_{i=1}^{k_1} X_{(i)} w_i - X_{(k_1)} r_{k_1} = \sum_{i=1}^{k_1 - 1} X_{(i)} w_i + X_{(k_1)} (w_{k_1} - r_{k_1}). \tag{20}$$

The residual portion r_{k_1} will be added to the next weighted average for $Y_{(2)}$. In general, for $Y_{(j)}$ we write

$$r_{k_j} = \sum_{i=k_{j-1}+1}^{k_j} w_i - r_{k_j - 1} - 1 \tag{21}$$

$$Y_{(j)} = X_{(k_{j-1})} r_{k_{j-1}} + \sum_{i=k_{j-1}+1}^{k_j - 1} X_{(i)} w_i + X_{(k_j)} (w_{k_j} - r_{k_j}). \tag{22}$$

Repeating the same steps we transform the original weighted ensemble MC with $\boldsymbol{X} = \left(X_{(1)}, \ldots, X_{(n)}\right)$ into the new unweighted ensemble MC$_\dagger$ with $\boldsymbol{Y} = \left(Y_{(1)}, \ldots, Y_{(\widetilde{n})}\right)$. We have distributed the weights from the MC so that there is the unit weight for each observation $Y_{(j)}$. Therefore, we are authorized to apply standard homogeneity tests, which guarantees the asymptotic properties.

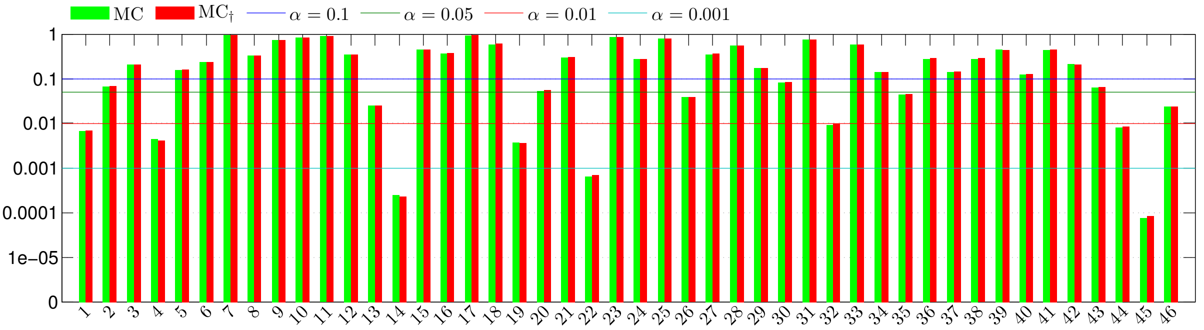


Figure 2: Kolmogorov-Smirnov test: p-value for all $m = 46$ variables in MC/MC$_\dagger$ channel Electron 4+ Jets.

Now, we can finally verify the correctness of the modified tests, used to weighted data. Indeed, the resulting p-values from the standard tests, performed over MC$_\dagger$, remarkably matches with accuracy p-values from the modified tests performed over MC. This is true even for small orders of magnitudes of p-values, as evidenced by comparison in Figure 2.

3.2 Generic Validation

As we verified eligible usage of modified weighted tests in previous section with datasets originating from high energy physics, we aim to provide more general verification now. Thus, we consider several different distributions for $\boldsymbol{X} = (X_1, \ldots, X_n)$: Beta, Cauchy, Exponential, Laplace, Logistic, Lognormal, Normal, Uniform and Weibull. On the contrary, weights $\boldsymbol{W} = (W_1, \ldots, W_n)$ are taken from Beta distribution as we may easily tune the expected value:

$$W \sim Beta(\alpha, \beta) \implies E\left[W\right] = \frac{\alpha}{\alpha + \beta}. \tag{23}$$

The appropriate number of simulation data points was determined by preliminary convergence studies. Otherwise, the simulation steps proceed as follows:

1. Generate n random weighted data points $(\boldsymbol{X}, \boldsymbol{W})$, e.g. $n = 3,500,000$.
2. Estimate weighted distribution from all the observations $(\boldsymbol{X}, \boldsymbol{W})$ (using kernel density estimation). Repeat all the following k times, e.g. $k = 1,000$:
   (a) Draw $m_w = \frac{n}{k}$ weighted observations from $(\boldsymbol{X}, \boldsymbol{W})$ as your current MC sample, e.g. $m_w = 3,500$.
   (b) Generate $m_u \approx \sum_{i=1}^{m_w} w_i$ unweighted observations from estimated weighted distribution as your current DATA sample, e.g. $m_u = 1,000$.
   (c) Apply weighted homogeneity test MC vs DATA.

(d) Rearrange MC into unweighted sample $\mathrm{MC}_\dagger$ and apply standard unweighted test.

Thus, we obtain k p-values from the weighted tests and also another k corresponding p-values from the unweighted tests. We may now check asymptotic properties of both weighted and unweighted tests.

For all the distributions under consideration we arrived at two main results. First, the significance level condition (3) is uniformly satisfied as shown in Figure 3, i.e. both EDFs are located under the diagonal in graph. Second, both weighted modifications and unweighted tests have the same resulting p-value distribution. This can be tested via ordinary classical homogeneity tests for unweighted data. Nevertheless, the extraordinary correspondence is obvious from the graph already.

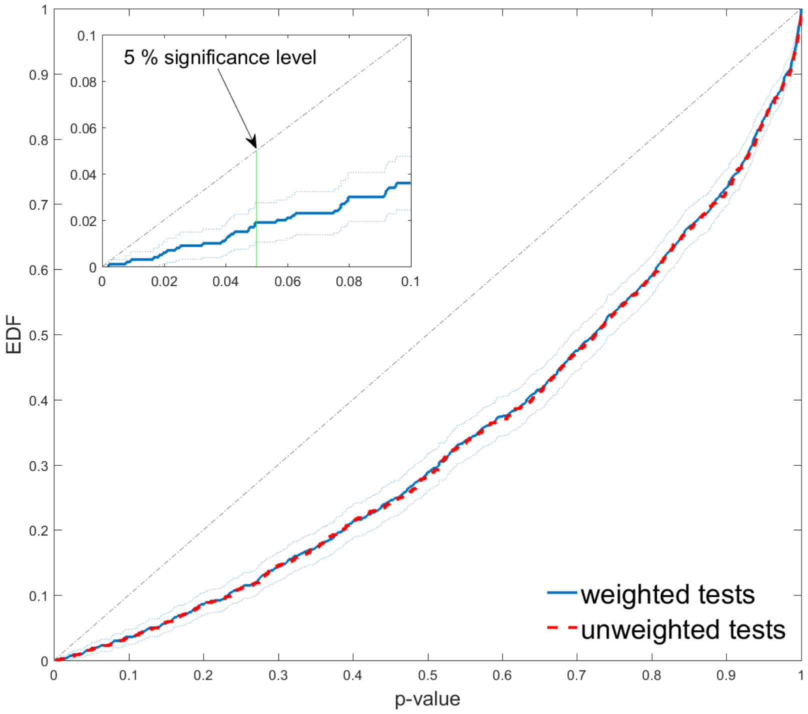


Figure 3: EDF of p-value for weighted and unweighted tests of homogeneity. Underlying data are taken from the lognormal distribution.

4 Conclusion

We performed numerical validation of modified statistical homogeneity tests for weighted data. Our simulation verifies that the approximate asymptotic properties remain the same for both weighted and unweighted tests. In consequence, in practice, we may either utilize modified weighted tests or we may apply the rearranging technique from Section 3.1 directly with the unweighted standard tests (where the asymptotics are proven). In future research, we aim to investigate the effect of various homogeneity tests and different weights distribution on the overall significance and power. We also plan to explore the possibility of proving the validity of weighted tests for arbitrary data distribution as well as potentially perform multivariate testing. The former may be reached by limiting the possibilities for the weights distribution as there exist only limited number of physical motivations for weighting procedures in practice.

References

[1] P. Bouř. *Testing of separation approaches with attributes selections for physical data (FNAL/CERN).* Research report. FNSPE CTU in Prague (2015).

[2] D. B. Kececioglu. *Reliability and Life Testing Handbook, 1st ed.* Prentice Hall (1993).

[3] H. B. Mann, A. Wald. *On the Choice of the Number of Class Intervals in the Application of the Chi Square Test.* Ann. Math. Stat. **13** (1942), 306–317.

[4] L. Pardo. *Statistical Inference Based on Divergence Measures.* Chapman & Hall/CRC (2006).

[5] F. W. Scholz, M. A. Stephens. *K-sample Anderson-Darling tests.* J. Am. Stat. Assoc. **82** (1987), 920.

[6] N. J. Smirnov. *Approximate laws of distribution of random variables from empirical data.* Usp. Mat. Nauk. **10** (1944), 179–206.

Multivariate Chebyshev-like Polynomials of the Fourth Kind*

Adam Brus

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

advisor: Jiří Hrivnák, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. With use of the multivariate trigonometric functions, the Chebyshev polynomials of the fourth kind are generalized to orthogonal polynomials of several variables. The general form of recurrence relations is obtained. These polynomials are further investigated in dimension three, exact form of recurrence relations is obtained and the first four polynomials are calculated using trigonometric identities. Then the first ten multivariate Chebyshev-like polynomials of fourth kind are generated.

Keywords: Chebyshev Polynomials, Multivariate Trigonometric Functions, Orthogonal Polynomials

Abstrakt. Za užití trigonometrických funkcí více proměnných jsou Čebyševovy polynomy čtvrtého druhu zobecněny na ortogonální polynomy více proměnných. Je získán obecný tvar rekurentních relací. Pro dimenzi tři jsou tyto polynomy dále zkoumány, je získán přesný tvar rekurentních relací a první čtyři polynomy jsou spočteny za užití trigonometrických identit. Následně je vygenerováno prvních deset více dimenzionálních Čebyševových polynomů čtvrtého druhu.

Klíčová slova: Čebyševovy polynomy, Trigonometrické funkce více proměnných, Ortogonální Polynomy

1 Introduction

In mathematics and physics we often encounter special functions on n-dimensional Euclidean space which are symmetric or antisymmetric with respect to permutation of variables. Example of such functions are multivariate trigonometric functions defined by Klimyk and Patera [10] as determinants and permanents of matrices, which entries are one dimensional trigonometric transforms. These functions inherit many important properties from the classical trigonometric functions and properties of determinants and permanents and due that are extensively studied [7, 8, 9].

One of application of multivariate trigonometric functions is to use them for generalization of discrete trigonometric transforms [1]. For multivariate discrete sine transforms

*This work was supported by the Grant Agency of Czech Technical University in Prague, grant No. SGS16 /239/OHK4/3T/14

this was done in [7] and for multivariate discrete cosine transforms in [2]. Another approach is to use the multivariate trigonometric functions as a starting point to define multivariate orthogonal polynomials.

Orthogonal polynomials [4, 5] are appearing in many parts of mathematics and physics and are intensively studied. Orthogonal polynomials which are connected to trigonometric functions are the Chebyshev polynomials [6, 11]. These polynomials are connected to effective methods of numerical interpolation and approximation and thus their multivariate generalizations is interesting topic to study. Using the multivariate trigonometric functions one can generalize the classical Chebyshev polynomials and obtain multivariate Chebyshev-like polynomials. In total there exist four kinds of Chebyshev polynomials, each of them can be generalized using symmetric or antisymmetric multivariate trigonometric functions. The generalization of the Chebyshev polynomials of first and third kind was done in [7] the generalization of Chebyshev polynomials of second kind in [3]. The generalization of Chebyshev polynomials of fourth kind is part of this paper.

2 Multivariate trigonometric functions

The multivariate generalizations of trigonometric functions are defined as determinants and permanents of matrices with entries $\cos(\pi\lambda_i x_j)$ resp. $\sin(\pi\lambda_i x_j)$ in [10]. The antisymmetric trigonometric functions $\cos^-_\lambda(x)$, $\sin^-_\lambda(x)$ and symmetric trigonometric functions $\cos^+_\lambda(x)$, $\sin^+_\lambda(x)$ of variable $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ with parameter $\lambda = (\lambda_1, \dots, \lambda_n)$ in the form:

$$\begin{aligned}\cos^-_\lambda(x) &= \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma)\cos(\pi\lambda_{\sigma_1}x_1)\cos(\pi\lambda_{\sigma_2}x_2)\cdots\cos(\pi\lambda_{\sigma_n}x_n),\\ \sin^-_\lambda(x) &= \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma)\sin(\pi\lambda_{\sigma_1}x_1)\sin(\pi\lambda_{\sigma_2}x_2)\cdots\sin(\pi\lambda_{\sigma_n}x_n),\end{aligned} \tag{1}$$

for the antisymmetric trigonometric functions and

$$\begin{aligned}\cos^+_\lambda(x) &= \sum_{\sigma\in S_n} \cos(\pi\lambda_{\sigma_1}x_1)\cos(\pi\lambda_{\sigma_2}x_2)\cdots\cos(\pi\lambda_{\sigma_n}x_n),\\ \sin^+_\lambda(x) &= \sum_{\sigma\in S_n} \sin(\pi\lambda_{\sigma_1}x_1)\sin(\pi\lambda_{\sigma_2}x_2)\cdots\sin(\pi\lambda_{\sigma_n}x_n),\end{aligned} \tag{2}$$

for the symmetric trigonometric functions.

For our applications we will only consider parameters λ in form $\lambda = k$ or $\lambda = k + \rho$ where $k \in \mathbb{Z}^n$ and $\rho = \left(\frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2}\right)$. Further, due (anti)symmetries, we will consider parameters k only lexicographically ordered, i.e.,

$$k_1 \geq k_2 \geq \ldots \geq k_n. \tag{3}$$

Due to properties of determinants and permanents the functions can be considered only on closure of the fundamental domain $F(\widetilde{S}^{\text{aff}}_n)$, of the form:

$$F(\widetilde{S}^{\text{aff}}_n) = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n \mid 1 \geq x_1 \geq x_2 \geq \ldots \geq x_n \geq 0\}, \tag{4}$$

which can be further restricted by omitting boundaries in specific cases due to additional properties discussed in [2], i.e,

- $x_i = x_{i+1}, i \in \{1, \dots, n-1\}$ for $\cos_k^-(x)$
- $x_i = x_{i+1}, i \in \{1, \dots, n\}$, $x_1 = 1$ or $x_n = 0$ for $\sin_k^-(x)$
- $x_i = x_{i+1}, i \in \{1, \dots, n-1\}$ or $x_1 = 1$ for $\cos_{k+\rho}^-(x)$
- $x_i = x_{i+1}, i \in \{1, \dots, n-1\}$ or $x_n = 0$ for $\sin_{k+\rho}^-(x)$
- $x_1 = 1$ or $x_n = 0$ for $\sin_k^+(x)$
- $x_1 = 1$ for $\cos_{k+\rho}^+(x)$
- $x_n = 0$ for $\sin_{k+\rho}^+(x)$

3 Chebyshev polynomials

The classical Chebyshev polynomials of one variable are connected to effective methods of interpolation and numerical integration and due that they are well known and extensively used class of orthogonal polynomials [6, 11]. There exist four kinds of the Chebyshev polynomials defined as

$$\begin{aligned} \mathcal{P}_n^{I}(x) &= T_n(x) = \cos(n\theta), & \mathcal{P}_n^{III}(x) &= V_n(x) = \frac{\cos\left(\left(n+\frac{1}{2}\right)\theta\right)}{\cos\left(\frac{1}{2}\theta\right)}, \\ \mathcal{P}_n^{II}(x) &= U_n(x) = \frac{\sin((n+1)\theta)}{\sin(\theta)}, & \mathcal{P}_n^{IV}(x) &= W_n(x) = \frac{\sin\left(\left(n+\frac{1}{2}\right)\theta\right)}{\sin\left(\frac{1}{2}\theta\right)}, \end{aligned} \tag{5}$$

with variable $x = \cos(\theta)$, $x \in [-1, 1]$.

For further uses we will focus mainly on the Chebyshev polynomials of the fourth kind. These polynomials are orthogonal on interval $(-1, 1)$. i.e.,

$$\int_{-1}^{1} W_n(x) W_m(x) (1-x)^{\frac{1}{2}} (1+x)^{-\frac{1}{2}} \, dx = 0, \quad n \neq m. \tag{6}$$

The first two polynomials can be obtained using of trigonometric formulas as:

$$W_1(x) = 1, \qquad W_2(x) = 2\cos(\theta) + 1 = 2x + 1. \tag{7}$$

The recurrence relations for following polynomials can be obtained using theory of orthogonal polynomials. However it is easier to obtain it using the following trigonometric identity:

$$\sin\left(\left(\left(n+\frac{1}{2}\right)+1\right)\theta\right) + \sin\left(\left(\left(n+\frac{1}{2}\right)-1\right)\theta\right) = 2\cos(\theta)\sin\left(\left(n+\frac{1}{2}\right)\theta\right) \tag{8}$$

which in coordinates $x = \cos(\theta)$ gives recurrence relations:

$$W_n(x) = 2x W_{n-1}(x) - W_{n-2}(x), \qquad n = 2, 3, \dots . \tag{9}$$

This together with knowledge of the first two polynomials generates all polynomials.

4 Multivariate Chebyshev-like polynomials of the fourth kind

The multivariate generalizations of trigonometric functions can be used to define multivariate Chebyshev-like polynomials. In total for any dimension there exist eight multivariate Chebyshev-like polynomials. Every classical Chebyshev polynomials can be generalized by use of symmetric or antisymmetric multivariate trigonometric functions. The Chebyshev polynomials of the first and the third kind were generalized in [7]. In this paper we focus on symmetric multivariate Chebyshev-like polynomials of fourth kind. Lets introduce variables $X_1, X_2, \ldots, X_n$:

$$X_1 = \cos^+_{(1,0,\ldots,0)}, \quad X_2 = \cos^+_{(1,1,\ldots,0)}, \quad \ldots, \quad X_n = \cos^+_{(1,1,\ldots,1)}. \tag{10}$$

Now the multivariate symmetric generalization of the Chebyshev polynomials of the fourth kind can be introduced in a form:

$$\mathcal{P}^{IV,+}_k(X_1, X_2, \ldots, X_n) = \frac{\sin^+_{k+\rho}(x)}{\sin^+_\rho(x)}, \tag{11}$$

where $\rho = \left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)$. These functions are well defined for all points of interior of fundamental domain $F(\tilde{S}^{aff}_n)$.

We use ordering of the polynomials from [7], we say that a polynomial $\mathcal{P}^{IV,+}_k$ is greater than polynomial $\mathcal{P}^{IV,+}_{k'}$, $k \neq k'$ if for all i, $k_i \geq k'_i$ and smaller if for all i, $k_i \leq k'_i$.

4.1 Recurrence relations

To obtain recurrence relations for the generalized Chebyshev-like polynomials $\mathcal{P}^{IV,+}_{k'}$ one has to consider generalized trigonometric identity which can be obtained using the classical identity (8):

$$\sin^+_k(x)\cos^+_l(x) = \frac{1}{2^n}\sum_{\sigma\in S_n}\sum_{\substack{a_i=\pm 1\\ i=1,\ldots,n}} \sin^+_{\left(k_1+a_1 l_{\sigma(1)},\ldots,k_n+a_n l_{\sigma(n)}\right)}(x). \tag{12}$$

Specially case where $l = \rho_1 = (1, 1, \ldots, 1)$, i.e,

$$\sin^+_k(x)\cos^+_{\rho_1}(x) = \frac{n!}{2^n}\sum_{\substack{a_i=\pm 1\\ i=1,\ldots,n}} \sin^+_{(k_1+a_1,\ldots,k_n+a_n)}(x), \tag{13}$$

the recurrence relations then obtain form:

$$\sin^+_k = \frac{2^n}{n!}\sin^+_{k-l_1-l_2-\ldots-l_n} X_n - \sum_i^n \sin^+_{k-2l_i} - \sum_{\substack{i,j=1\\ i<j}}^n \sin^+_{k-2l_i-2l_j} - \ldots - \sin^+_{k-2l_1-2l_2-\ldots-2l_n}. \tag{14}$$

where l_i is vector with 1 on i-th coordinate and 0 for the rest.

Using identity (14) each polynomial can be expressed as linear combination of lower polynomials and combination of products of lower polynomial with variables X_i. Therefore each polynomial can be defined recursively.

4.2 Three-dimensional polynomials

Properties of generalized sine functions together with generalized trigonometric identity (14) leads to the following set of recurrence relations for $\mathcal{P}^{IV,+}_{(k_1,k_2,k_3)}$. The first four polynomials are obtained using trigonometric identities in form:

$$\begin{aligned} \mathcal{P}^{IV,+}_{(0,0,0)} &= 1, & \mathcal{P}^{IV,+}_{(1,0,0)} &= \frac{1}{3}X_1 + 1, \\ \mathcal{P}^{IV,+}_{(1,1,0)} &= \frac{2}{3}X_2 + \frac{2}{3}X_1 + 1, & \mathcal{P}^{IV,+}_{(1,1,1)} &= \frac{4}{3}X_3 + 2X_2 + X_1 + 1. \end{aligned} \tag{15}$$

Following polynomials are then obtained using recurrence relations:

$$\begin{aligned}
k_1 \geq 2, k_2 = k_3 = 0: \quad & \mathcal{P}^{IV,+}_{(k_1,0,0)} = \mathcal{P}^{IV,+}_{(k_1-1,0,0)}X_1 - \mathcal{P}^{IV,+}_{(k_1-2,0,0)} - 2\mathcal{P}^{IV,+}_{(k_1-1,1,0)} + 2\mathcal{P}^{IV,+}_{(k_1-1,0,0)} \\
k_1 - 1 > k_2 > k_3 = 0: \quad & \mathcal{P}^{IV,+}_{(k_1,k_2,0)} = \mathcal{P}^{IV,+}_{(k_1-1,k_2,0)}X_1 - \mathcal{P}^{IV,+}_{(k_1-2,k_2,0)} + \mathcal{P}^{IV,+}_{(k_1-1,k_2,0)} \\
& \quad - \mathcal{P}^{IV,+}_{(k_1-1,k_2+1,0)} - \mathcal{P}^{IV,+}_{(k_1-1,k_2-1,0)} - \mathcal{P}^{IV,+}_{(k_1-1,k_2,1)} \\
k_1 - 1 >, k_2 = k_3 > 0: \quad & \mathcal{P}^{IV,+}_{(k_1,k_2,k_2)} = \mathcal{P}^{IV,+}_{(k_1-1,k_2,k_2)}X_1 - \mathcal{P}^{IV,+}_{(k_1-2,k_2,k_2)} \\
& \quad - 2\mathcal{P}^{IV,+}_{(k_1-1,k_2+1,k_2)} - 2\mathcal{P}^{IV,+}_{(k_1-1,k_2,k_2-1)} \\
k_1 - 1 >, k_2 > k_3 > 0: \quad & \mathcal{P}^{IV,+}_{(k_1,k_2,k_3)} = \mathcal{P}^{IV,+}_{(k_1-1,k_2,k_3)}X_1 - \mathcal{P}^{IV,+}_{(k_1-2,k_2,k_3)} - \mathcal{P}^{IV,+}_{(k_1-1,k_2+1,k_3)} \\
& \quad - \mathcal{P}^{IV,+}_{(k_1-1,k_2-1,k_3)} - \mathcal{P}^{IV,+}_{(k_1-1,k_2,k_3+1)} - \mathcal{P}^{IV,+}_{(k_1-1,k_2,k_3-1)} \\
k_1 - 1 = k_2 > k_3 = 0: \quad & \mathcal{P}^{IV,+}_{(k_1,k_1-1,0)} = \frac{1}{2}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,0)}X_1 - \mathcal{P}^{IV,+}_{(k_1-1,k_1-2,0)} \\
& \quad - \frac{1}{2}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,1)} + \frac{1}{2}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,0)} \\
k_1 - 1 = k_2 > k_3 > 0: \quad & \mathcal{P}^{IV,+}_{(k_1,k_1-1,k_3)} = \frac{1}{2}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3)}X_1 - \mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_3)} \\
& \quad - \frac{1}{2}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3+1)} - \frac{1}{2}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3-1)} \\
k_1 - 1 = k_2 = k_3 > 0: \quad & \mathcal{P}^{IV,+}_{(k_1,k_1-1,k_1-1)} = \frac{1}{3}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-1)}X_1 - \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-2)}
\end{aligned} \tag{16}$$

$$k_1 = k_2 = 2, k_3 = 0: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(2,2,0)} &= 2\mathcal{P}^{IV,+}_{(1,1,0)}X_2 - 2\mathcal{P}^{IV,+}_{(1,0,0)}X_1 + \mathcal{P}^{IV,+}_{(1,1,0)}X_1 \\ &\quad - \mathcal{P}^{IV,+}_{(1,1,1)}X_1 + \mathcal{P}^{IV,+}_{(0,0,0)} + 6\mathcal{P}^{IV,+}_{(1,1,0)} - 4\mathcal{P}^{IV,+}_{(1,0,0)} \\ &\quad + \mathcal{P}^{IV,+}_{(2,1,1)} - \mathcal{P}^{IV,+}_{(1,1,1)} \end{aligned}$$

$$k_1 = k_2 > 2, k_3 = 0: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(k_1,k_1,0)} &= 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,0)}X_2 - 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,0)}X_1 - \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,1)}X_1 \\ &\quad + \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,0)}X_1 + \mathcal{P}^{IV,+}_{(k_1-2,k_1-2,0)} + 4\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,0)} + 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,1)} \\ &\quad + 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,0)} + 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-3,0)} + \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,2)} - \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,1)} \end{aligned}$$

$$k_1 = k_2 > k_3 + 2 > 2: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(k_1,k_1,k3)} &= 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k3)}X_2 - 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k3)}X_1 \\ &\quad - \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3+1)}X_1 - \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3-1)}X_1 \\ &\quad + \mathcal{P}^{IV,+}_{(k_1-2,k_1-2,k_3)} + 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_3+1)} \\ &\quad + 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_3-1)} + 4\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3)} \\ &\quad + 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-3,k_3)} + \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3+2)} \\ &\quad + \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_3-2)} \end{aligned}$$

$$k_1 = k_2 = k_3 + 2 = 3: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(3,3,1)} &= 2\mathcal{P}^{IV,+}_{(2,2,2)}X_2 - 2\mathcal{P}^{IV,+}_{(2,1,1)}X_1 - \frac{2}{3}\mathcal{P}^{IV,+}_{(2,2,2)}X_1 \\ &\quad - \mathcal{P}^{IV,+}_{(2,2,0)}X_1 + \mathcal{P}^{IV,+}_{(1,1,1)} + 5\mathcal{P}^{IV,+}_{(2,2,1)} \\ &\quad + 4\mathcal{P}^{IV,+}_{(2,1,0)} + \mathcal{P}^{IV,+}_{(2,2,0)} \end{aligned}$$

$$k_1 = k_2 = k_3 + 2 > 3: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(k_1,k_1,k_1-2)} &= 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-2)}X_2 - 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_1-2)}X_1 \\ &\quad - \frac{2}{3}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-1)}X_1 - \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-3)}X_1 \\ &\quad + \mathcal{P}^{IV,+}_{(k_1-2,k_1-2,k_1-2)} + 5\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-2)} \\ &\quad + 4\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_1-3)} + \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-4)} \end{aligned}$$

$$k_1 = k_2 = k_3 + 1 = 2: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(2,2,1)} &= \frac{2}{3}\mathcal{P}^{IV,+}_{(1,1,1)}X_2 - \mathcal{P}^{IV,+}_{(1,1,0)}X_1 \\ &\quad + \mathcal{P}^{IV,+}_{(1,0,0)} + \mathcal{P}^{IV,+}_{(1,1,1)} - \mathcal{P}^{IV,+}_{(1,1,0)} \end{aligned}$$

$$k_1 = k_2 = k_3 + 1 > 2: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(k_1,k_1,k_1-1)} &= \frac{2}{3}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-1)}X_2 - \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-2)}X_1 \\ &\quad + \mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_1-2)} + \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-1)} \\ &\quad + \mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-3)} \end{aligned}$$

$$k_1 = k_2 = k_3 = 2: \quad \begin{aligned} \mathcal{P}^{IV,+}_{(2,2,2)} &= \frac{4}{3}\mathcal{P}^{IV,+}_{(1,1,1)}X_3 - 6\mathcal{P}^{IV,+}_{(1,1,0)}X_2 + 3\mathcal{P}^{IV,+}_{(1,0,0)}X_1 - 3\mathcal{P}^{IV,+}_{(1,1,0)}X_1 \\ &\quad + 2\mathcal{P}^{IV,+}_{(1,1,1)}X_1 - \mathcal{P}^{IV,+}_{(0,0,0)} - 9\mathcal{P}^{IV,+}_{(1,1,0)} + 6\mathcal{P}^{IV,+}_{(1,0,0)} + 3\mathcal{P}^{IV,+}_{(1,1,1)} \end{aligned} \tag{17}$$

$$\begin{aligned} k_1 = k_2 = k_3 = 3: \quad \mathcal{P}^{IV,+}_{(3,3,3)} &= \frac{4}{3}\mathcal{P}^{IV,+}_{(2,2,2)}X_3 - 6\mathcal{P}^{IV,+}_{(2,2,1)}X_2 + 3\mathcal{P}^{IV,+}_{(2,1,1)}X_1 \\ &+ 2\mathcal{P}^{IV,+}_{(2,2,2)}X_1 + 3\mathcal{P}^{IV,+}_{(2,2,0)}X_1 - \mathcal{P}^{IV,+}_{(1,1,1)} \\ &- 9\mathcal{P}^{IV,+}_{(2,2,1)} - 6\mathcal{P}^{IV,+}_{(2,1,0)} + \mathcal{P}^{IV,+}_{(2,2,0)} \end{aligned}$$

$$\begin{aligned} k_1 = k_2 = k_3 > 3: \quad \mathcal{P}^{IV,+}_{(k_1,k_1,k_1)} &= \frac{4}{3}\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-1)}X_3 - 6\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-2)}X_2 \\ &+ 3\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_1-2)}X_1 + 2\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-1)}X_1 \\ &+ 3\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-3)}X_1 - \mathcal{P}^{IV,+}_{(k_1-2,k_1-2,k_1-2)} \\ &- 9\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-2)} - 6\mathcal{P}^{IV,+}_{(k_1-1,k_1-2,k_1-3)} \\ &- 3\mathcal{P}^{IV,+}_{(k_1-1,k_1-1,k_1-4)}, \end{aligned} \tag{18}$$

which are obtained from the generalized trigonometric identity (14).

With the use of recurrence relations one can obtain the exact form of first ten polynomials ($k \leq (2,2,2)$) as follows:

$$\begin{aligned} \mathcal{P}^{IV,+}_{(0,0,0)} &= 1, \\ \mathcal{P}^{IV,+}_{(1,0,0)} &= \frac{1}{3}X_1 + 1, \\ \mathcal{P}^{IV,+}_{(1,1,0)} &= \frac{2}{3}X_2 + \frac{2}{3}X_1 + 1, \\ \mathcal{P}^{IV,+}_{(1,1,1)} &= \frac{4}{3}X_3 + 2X_2 + X_1 + 1, \\ \mathcal{P}^{IV,+}_{(2,0,0)} &= \frac{1}{3}X_1^2 - \frac{4}{3}X_2 + \frac{1}{3}X_1 - 1, \\ \mathcal{P}^{IV,+}_{(2,1,0)} &= \frac{1}{3}X_1^2 + \frac{1}{3}X_2X_1 - \frac{2}{3}X_3 - \frac{2}{3}X_2 - \frac{2}{3}X_1 - 1, \\ \mathcal{P}^{IV,+}_{(2,1,1)} &= \frac{1}{3}X_1^2 + \frac{4}{9}X_3X_1 - \frac{2}{3}X_2X_1 - \frac{2}{3}X_2 - \frac{1}{3}X_1 - 1, \\ \mathcal{P}^{IV,+}_{(2,2,0)} &= -\frac{4}{3}X_1^2 + \frac{4}{3}X_2^2 - \frac{8}{9}X_3X_1 - \frac{4}{3}X_3 + \frac{4}{3}X_2 - \frac{1}{3}X_1 + 1, \\ \mathcal{P}^{IV,+}_{(2,2,1)} &= -\frac{2}{3}X_1^2 + \frac{4}{3}X_2^2 + \frac{8}{3}X_3X_2 + \frac{4}{3}X_3 + 2X_2 + \frac{5}{3}X_1 + 1, \\ \mathcal{P}^{IV,+}_{(2,2,2)} &= \frac{16}{9}X_3^2 - 4X_2^2 + X_1^2 + \frac{8}{3}X_3X_2 - 4X_3X_1 - 2X_2X_1 + \frac{16}{3}X_3 - 12X_2 + X_1 - 1. \end{aligned} \tag{19}$$

From the first ten polynomials one can see that the polynomial $\mathcal{P}^{IV,+}_{(k_1,k_2,k_3)}$ is of order k_1 for $k_1 \leq 2$, which can be proven generally for any k_1 using the generalized trigonometric identity (14).

5 Conclusion

The generalization of the Chebyshev polynomials of fourth kind was done using the symmetric multivariate trigonometric function. The generalization by antisymmetric function follow similar procedure but is slightly more complicated due to the antisymmetry condition. This generalizations completes the set of eight multivariate Chebyshev-like polynomials. The multivariate Chebyshev-like polynomials inherited many usable properties from the one dimensional cases, and thus are interesting topic for further study.

One of possible applications of the multivariate Chebyshev-like polynomials is to obtain cubature formulas. Cubature formulas allow replacing weighted integral of polynomial function with a linear combination of polynomial values at some points. This allows faster computation and therefore can lead to effective numerical methods. For the multivariate Chebyshev polynomials of first and third kind this was already done in [7]. The cubature formulas obtained from multivariate Chebyshev polynomials of second and fourth kind are point of current study.

References

[1] V. Britanak, K. Rao, and P. Yip, *Discrete cosine and sine transforms: general properties, fast algorithms and integer approximation*, Elsevier/Academic Press, Amsterdam, (2007).

[2] A. Brus, *Discrete Multivariate (Anti)Symmetric Trigonometric functions*, Doktorandské dny 2015, pp. 13-22, (2015).

[3] A. Brus, *Multivariate Generalizations of Chebyshev Polynomials of Second Kind*, Doktorandské dny 2016, pp. 7-14, (2016).

[4] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, Science Publishers, Inc., 1978.

[5] C. F. Dunkel and Y. Xu, *Orthogonal Polynomials of Several Variables*, Encyclopedia Math. Appl. 81, Cambridge University Press, Cambridge, UK, 2001.

[6] D. C. Handscomb and J. C. Mason, *Chebyshev Polynomials*, Champman & Hall/CRC, Boca Raton, FL, 2003.

[7] J. Hrivnák and L. Motlochová, *Discrete Transforms and orthogonal polynomials of (anti)symmetric multivariate cosine functions*, SIAM J. Numer. Anal., Vol. 52, No 6, pp. 3021-3055 (2014).

[8] J. Hrivnák, L. Motlochová and J. Patera, *Two dimensional symmetric and antisymmetric generalization of sine functions*, J. Math. Phys., 51 (2010), 073509.

[9] J. Hrivnák and J. Patera, *Two dimensional symmetric and antisymmetric generalization of exponential and cosine functions*, J. Math. Phys., 51 (2010), 023515.

[10] A. Klimyk and J. Patera, *(Anti)symmetric multivariate trigonometric functions and corresponding Fourier transforms*, J. Math. Phys., 48 (2007), 093504.

[11] T. J. Rivlin, *The Chebyshev Polynomials*, John Wiley & Sons, New York, 1990.

Evaluation of Pedestrian Density Distribution with Respect to the Velocity Response[*]

Marek Bukáček

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

advisors:
Milan Krbálek, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. There are many approaches to evaluate density within pedestrian scenarios, including point approximation, Voronoi cells or more sophisticated methods. In this project we focus on the individual density, where each pedestrian is considered as a source of density distribution. A cone can be considered as a reasonable shape, with its diameter as a blur parameter. Naturally, pedestrians adapt their velocity and path selection with respect to the conditions around them in given range. The correlation of density and velocity, respective density and exit angle was evaluated on laboratory experiment data for all acceptable blur – range combination. Because negative correlation corresponds to more significant response of velocity (exit angle) to the density, the correlations seem to be a perfect tool to estimate density parameters.

Keywords: crowd dynamics, individual density, velocity response

Abstrakt. Existuje mnoho přístupů k vyhodnocování hustoty v rámci systémů chodců, jako je bodová aproximace, Voronoiské buňky nebo další, sofistikovanější metody. V tomto projektu se zaměřujeme na individuální hustotu, kde je každý chodec považován za zdroj distribuce hustoty. Za vhodný tvar může být považován kužel jednotkového objemu, jehož průměr vyjadřuje parametr rozostření. Chodci zřejmě přizpůsobují svou rychlost a výběr cesty okolním podmínkám v daném okolí. Korelace hustoty a rychlosti, popřípadě hustoty a úhlu k výstupu byla vyhodnocována na základě údajů z laboratorních experimentů pro všechny myslitelné kombinace parametrů rozostření a rozsahu okolí. Vzhledem k tomu, že více negativní korelace odpovídá výraznější odezvě rychlosti (úhlu výstupu) na hustotu, zdá se, že tyto korelace jsou vhodným nástrojem pro odhad parametrů hustoty.

Klíčová slova: dynamika davu, individuální hustota, reakce rychlosti

1 Introduction

The pedestrian movement, including egress situations, walking in corridors or in cross-section areas has been widely studied in last twenty years [4]. This period seems long enough to bring the answer to such fundamental question as "how pedestrians react to

[*] GAČR 15-15049S & SGS15/214/OHK4/3T/14

their surrounding", but so far, there are only qualitative studies or macroscopic approximations. Moreover, the definitions of fundamental quantities are not unified [5] and the only criteria to use some method is to bring the prettiest data.

In this paper, the study of pedestrian reaction starts with quantification of state of his neighborhood and quantification of his reaction. The main idea has been presented and described at the conference PED 2017 [3]. This paper partially discusses some part of density evaluation and concludes preliminary results.

The reaction consisting of velocity and direction changes is considered to be induced by the trend of density. There are many ways to evaluate density and even the reaction range should be parametrized, thus the pedestrian behavior in front of the exit is analyzed on parametric grid with respect to multiple defined densities (defined bellow). This parametric grid is generally based on two features:

- blur, e.g. the size of area affected by one pedestrian,
- range, e.g. the size of area affecting one pedestrian.

At the end, Pearson correlation coefficient

$$\mathcal{R}_t\left(\rho_{\omega_\alpha}, v_\alpha\right) = \frac{\mathrm{Cov}\left(\rho_{\omega_\alpha}, v_\alpha\right)}{\sqrt{\mathrm{Var}\left(\rho_{\omega_\alpha}\right)\,\mathrm{Var}\left(v_\alpha\right)}} \tag{1}$$

is used as a metric to select the density with the best fit to pedestrian reactions.

Numerical study is based on the egress experiment organized in the study hall of FNSPE CTU in Prague in 2014, see [1], [2].

2 Definitions

As mentioned above, the analysis is provided on pedestrian trajectory data. The velocity $v_\alpha(t)$ of pedestrian α is defined as usual using central differences of space coordinates. The exit angle $\vartheta_\alpha(t) \in [0, \pi]$ is defined as angular deflection from the ideal direction of the pedestrian α to the exit

The density is the only flexible variable in this study. Its value is integrated over the distribution generated by each pedestrian α individually

$$\rho = \frac{N}{|A|} = \frac{\int_A p(\vec{x})\,\mathrm{d}\vec{x}}{|A|} = \frac{\int_A \sum_{\alpha=1}^{N} p_\alpha(\vec{x})\,\mathrm{d}\vec{x}}{|A|} = \sum_{\alpha=1}^{N} \frac{\int_A p_\alpha(\vec{x})\,\mathrm{d}\vec{x}}{|A|}. \tag{2}$$

There are several methods to define the individual density distribution function (kernel) $p_\alpha(\vec{x})$:

- point approximation

$$p_\alpha(\vec{x}) = \delta_{\vec{x},\vec{x}_\alpha},$$

where $\int_A \delta_{\vec{x},\vec{x}_\alpha}\,\mathrm{d}\vec{x} = \begin{cases} 1 & \text{if } \vec{x}_\alpha \in A, \\ 0 & \text{otherwise,} \end{cases}$

- stepwise function

$$p_\alpha(\vec{x}) = \begin{cases} \frac{1}{|A_\alpha|} & \text{if} \quad \vec{x} \in A_\alpha, \\ 0 & \text{otherwise,} \end{cases}$$

  where special cases are

  1. cylindrical distribution

  $$p_\alpha(\vec{x}, R) = \begin{cases} \frac{1}{R^2\pi} & \text{if} \quad \|\vec{x} - \vec{x}_\alpha\| < R, \\ 0 & \text{otherwise,} \end{cases}$$

  2. Voronoi distribution, where A_α is a voronoi cell – the whole space is segregated into pedestrian cells A_α according to a simple rule: each point $\vec{x}$ is assigned to the nearest pedestrian $\vec{x}_\alpha$,

- linear (conic) distribution

$$p_\alpha(\vec{x}, R) = \begin{cases} \frac{3}{R^3\pi}(R - \|\vec{x} - \vec{x}_\alpha\|) & \text{if} \quad \|\vec{x} - \vec{x}_\alpha\| < R, \\ 0 & \text{otherwise,} \end{cases}$$

- Gaussian distribution

$$p_\alpha(\vec{x}, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}}\, \mathrm{e}^{-\frac{1}{2}(\vec{x}-\vec{x}_\alpha)^T\Sigma^{-1}(\vec{x}-\vec{x}_\alpha)}$$

  with covariance matrix $\Sigma = \sigma^2\, \mathrm{I}_{2\times2}$, where $\mathrm{I}_{2\times2}$ represents identity matrix.

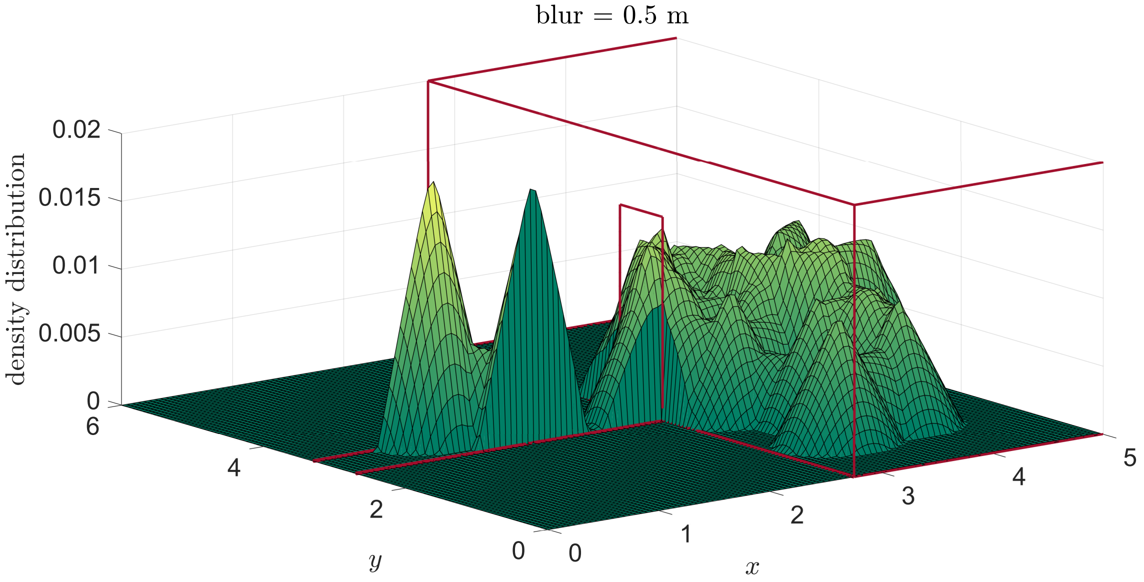


Figure 1: Example of density distribution

In this paper, linear (conic) distribution was used due to its decreasing trend with increasing distance, limited support and independence of one pedestrian to others. An example of density distribution generated by the method mentioned above is visualized in Figure 1.

3 Analysis

Basic overview is provided by of velocity – density, resp. direction – density relation of all trajectories. For each blur and range parametric set, Pearson correlation coefficient was evaluated over the whole trajectory, and then averaged over all trajectories of the experiment, see Figure 2.

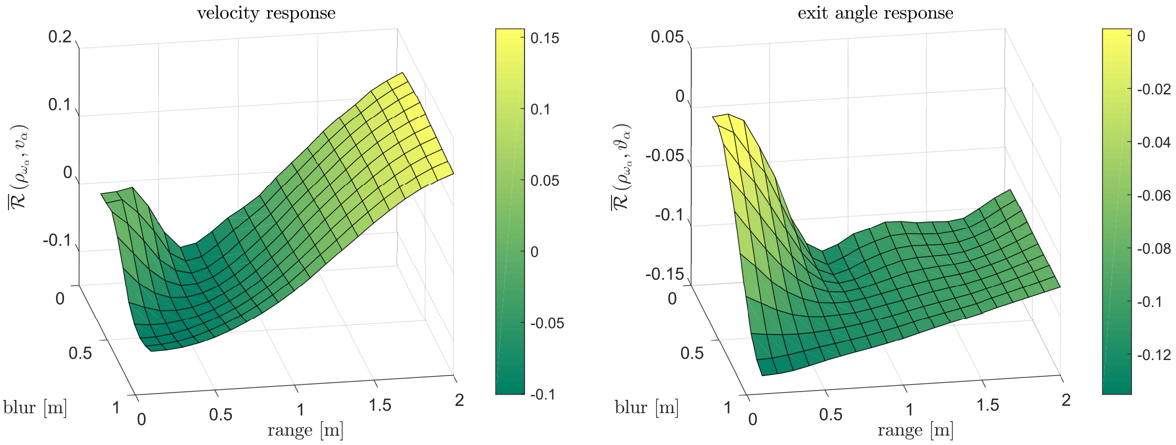


Figure 2: Correlation coefficient over the whole trajectory, mean over all trajectories

We can see expected zero correlation for zero range point approximation in case of both, velocity and the exit angle as well as natural negative velocity correlation for short range narrow approximation. On the other hand, positive velocity correlation for any long range approximation and negative exit angle correlation for all reasonable sets of parameters weren't expected at all. Moreover, the absolute value of correlation is rather small, indicating week dependency of density and pedestrian reaction.

To see the source of positive or negative correlation, we have to go to individual level and check rolling correlation (window width 1.56 s) for segments of one trajectory, see Figure 3.

There is strong positive correlation of velocity and long-range density in free flow area that can be explained by competitiveness between pedestrians. Strong negative correlation of velocity and all densities in avoiding/joining the cluster area corresponds to adjusting velocity to higher density. And at the end, positive correlation of velocity and all densities in the cluster area is caused by the flow conservation law – closer the exit, lower number of participants carry the flow, the velocity at the exit is much higher than inside the crowd, even the density is higher as well.

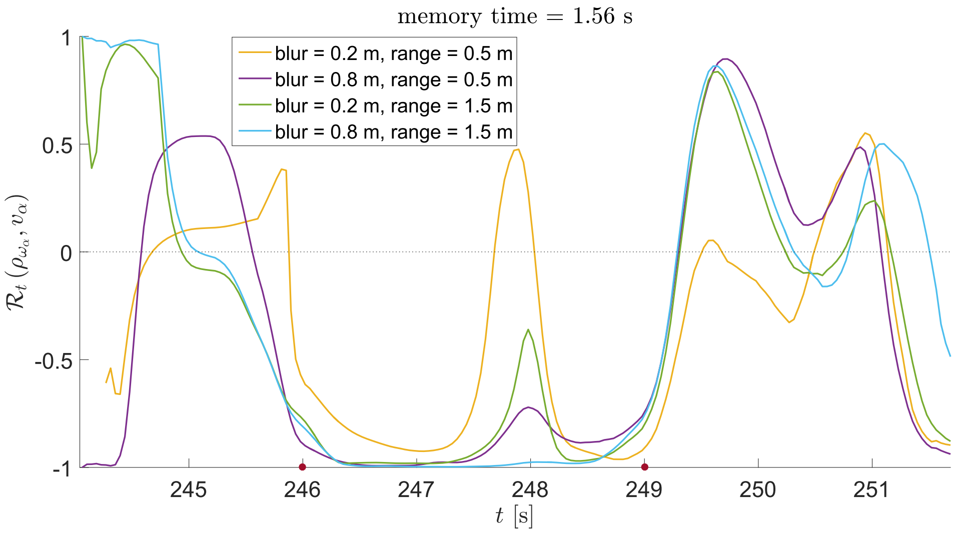


Figure 3: Correlation coefficient of four density combinations and velocity

4 Conclusions

Correlation between velocity and density isn't obviously as clear as expected on the first sight. Expected decrease of velocity implied by increasing density is observed only in transition phase between free flow and congested areas. Others situations produce different behavior due to the complex dynamics.

In general, individual pedestrian density reflects phase transition changes very well, as can be seen in Figure 4. The value of correlation of velocity and one specific density is not stable, but differs with the traffic mode around, personal preferences and individually selected strategy. The analysis of such complexity is a subject of further research.

Yet these preliminary results described and explained unexpected positive correlation in the exit area by the flow conservation law. We hope that deep decomposition and clustering of trajectories reveal more fundamental facts that increase our ability to predict the pedestrian reactions.

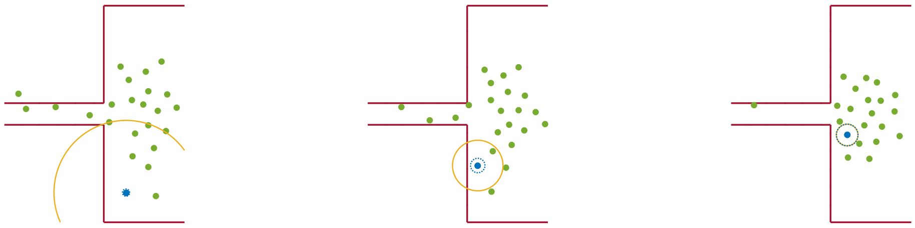


Figure 4: Changes of blur (blue dotted neighborhood) and range (yellow neighborhood) parameters according to the phase transitions.

References

[1] M. Bukáček, P. Hrabák and M. Krbálek. *Experimental Study of Phase Transition in Pedestrian Flow.* In PED 2014, Transportation Research Procedia **2** (2014), 105 – 113.

[2] P. Hrabák, M. Bukáček and M. Krbálek. *Individual Microscopic Results Of Bottleneck Experiments.* In Traffic and Granular Flow '15, Springer (2016), 105 – 112.

[3] M. Bukáček and J. Vacková. *Evaluation of pedestrian density distribution with respect to the velocity response.* In Traffic and Granular Flow '17, submitted 2017.

[4] A. Schadschneider, D. Chowdhury and K. Nishinari, *Stochastic Transport in Complex Systems*, Elsevier (2010).

[5] B. Steffen and A. Seyfried, *Methods for Measuring Pedestrian Density, Flow, Speed and Direction with Minimal Scatter*, Physica A **389(9)** (2010), 1902–1910.

Model of Planar and Spherical Phase Interface Geometries for Multi-Component Mixtures*

David Celný

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

advisors:

Jan Hrubý, Department of Thermodynamics
Institute of Thermomechanics, CAS

Václav Vinš, Department of Thermodynamics
Institute of Thermomechanics, CAS

Abstract. This study presents a mathematical model of density profile computation for multicomponent mixtures of two commonly used phase geometries. The model unifies the description of multicomponent systems of planar and spherical interface geometry. The mathematical model is supplied with PC-SAFT equation of state for thermodynamic property evaluation. The fundamentals of the presented model lie in the gradient theory approximation used to formulate the governing differential equation. An innovative approach to the problem formulation divides the solution into two simple parts. The solution method applicable for arbitrary geometry was developed and a special case for planar and spherical interfaces was solved. In addition to the density profile and the surface tension are computed for modelled system. Binary system CO_2, C_4H_{10} was investigated and compared with available experimental data. Surface tension estimate was found to be in good agreement with experiment.

Keywords: phase interface, gradient theory, multicomponent system, surface tension

Abstrakt. Předmětem studie je zkoumání fázových rozhraní dvou základních typů geometrií. Jedná se o rovinné a sférické geometrie, které jsou ve studii zkoumány jednotným modelem. Tento model využívá poznatky gradientní teorie a je doplněn o stavovou rovnici PC-SAFT, která vyčísluje termodynamické vlastnosti zkoumaného systému. Pomocí originálního přístupu je model rozdělen na dva výpočetní kroky. Ve studii jsou zkoumány obě geometrie na vybrané reálné směsi obsahující CO_2, C_4H_{10}. Vypočtené výsledky jsou následně srovnány s dostupnými experimentálními daty. Výsledky srovnání pro povrchová napětí jsou v dobré shodě s vytvořeným modelem.

Klíčová slova: fázové rozhraní, gradientní teorie, vícesložkové systémy, povrchové napětí

1 Theoretical background

The methods for accurate modelling of phase interfaces are important for the understanding of natural processes and applications in technology. One such application is

[*] The research has received funding from the Norwegian Financial Mechanism 2009-2014 under Project Contract no. 7F14466

carbon capture and storage (CCS). In particular, the prediction of non-equilibrium phase transitions requires a detailed knowledge of the phase interfaces.

The gradient theory (GT) framework presented here was initially used for pure systems only. The initial aim was to predict interface properties of said pure system. Through the recent years the originally simple description of pure systems was extended into multiple component systems for example [15, 16]. The authors derived the formulas and governing equation system for the multi-component problems and provided comparison with available experimental data. But during the derivation authors restricted themselves to the planar phase interface geometry. There also exist group of authors who extend the theory to the more complicated spherical interface geometry [6, 17, 20]. While these authors derived the terms for the special geometry e.g droplets, they also restricted themselves and constructed the models for the pure systems only. Based on our observation there is no unified framework which describes how to approach spherical phase geometry in multi-component systems.

The presented study continues in line with mixture systems research by Vinš *et. al.* [19] and combines the spherical interface geometry research by Planková *et. al.* [18]. The aim of this study is the prediction of multi-component systems with spherical phase interface geometry initially outlined in [4]. The method is extended into derivation of the generalized computational approach for two interface geometries in multiple-component system. This study also present the comparison of investigate two-component system with experimental data in the last section.

2 Theoretical background

2.1 Cahn-hilliard gradient theory

The main advantage of gradient theory approach is the computational speed and the overall simplicity compared to the full density functional theory (DFT) or molecular simulation models. But the simplicity of the approach comes at the cost of lowered accuracy in regions with large gradients of Helmholtz energy.

Gradient theory formulate the work of formation $\Delta\Omega$ and uses it to describe the optimal density profile. The work of formation is defined as the difference between the homogeneous system and the non-homogeneous system where the phase interface effects are accounted for. Same formulation can be expressed in multiple thermodynamic potentials, but for the case of multi-component mixtures the grand potential is the most suitable:

$$\Delta\Omega(\rho) = \Omega_{\text{inhom}}(\rho) - \Omega_{\text{hom}}(\rho). \tag{1}$$

Grand potentials here depend on the molar density $\boldsymbol{\rho}$ which can be understood as an universal descriptor of a system. It is also usual to search for the molar density system description in form of density profile. In an arbitrary system such density profile is a function of the systems coordinates for example $\boldsymbol{\rho} = \rho(s_1, s_2, s_3)$. With this formulation the solution becomes substantially complex. Therefore, it is usual to assume that the system is non-uniform in only single coordinate s_1 denoted further simply as s. Helmholtz energy of inhomogeneous system is then expressed as Taylor expansion around homogeneous Helmholtz energy with higher order terms omitted as follows:

$$f_{\text{inhom}} = f_{\text{hom}}(\boldsymbol{\rho}) + C_1 \cdot \nabla^2 \boldsymbol{\rho} + \frac{1}{2} C_2 \cdot (\nabla \boldsymbol{\rho})^2 \ldots \tag{2}$$

According to the approach used by Cahn and Hilliard [3] the Taylor expansion is utilised in the formulation of the grand potential. With a simplified notation the following equation for grand potential difference is obtained.

$$\Delta\Omega = \int_s \left(\Delta\omega\left(\boldsymbol{\rho}\left(s\right)\right) + \frac{1}{2} C_3 \left(\frac{\partial \boldsymbol{\rho}}{\partial s} \right)^2 \right) S ds, \tag{3}$$

Here C_3 parameters contain the Taylor expansion coefficients and $\Delta\omega$ is the grand potential density which can be also expressed in following form:

$$\Delta\omega\left(\boldsymbol{\rho}\right) = f_{\text{hom}}\left(\boldsymbol{\rho}\right) - \sum_{i=1}^{n} \mu_i^{\text{G}} \rho_i + p^{\text{G}}. \tag{4}$$

It can be noted that formation work in eq. (3) was derived for generalized type of interface geometry parametrized with s and S. This geometry can be specified later with the choice of coordinate system best describing the intended interface geometry. Selecting Cartesian coordinates the planar geometry can be described and similarly spherical coordinates can be used for droplets.

2.1.1 Core problem derivation

When the task is transferred into grand potential formulation it can be noticed that it is also a functional formulation for an unknown density profile function ρ. With the problem then understood as functional problem of finding the saddle point the variational calculus can be used with advantage. The required criterium for optimal density profile can be formulated accordingly as:

$$\delta\Delta\Omega\left[\rho\right]_{\rho=\rho^0} = 0 \tag{5}$$

The extremal point of previous formulation is found by Euler-Lagrange equations.

$$S \frac{\partial \Delta\omega\left(\boldsymbol{\rho}\left(s\right)\right)}{\partial \rho_k} + \frac{S}{2} \sum_{i,j=1}^{n} \frac{\partial c_{i,j}}{\partial \rho_k} \left(\frac{\partial \boldsymbol{\rho_i}}{\partial s} \right) \left(\frac{\partial \boldsymbol{\rho_j}}{\partial s} \right) - \frac{d}{ds} S \sum_{i=1}^{n} c_{i,k} \left(\frac{\partial \boldsymbol{\rho_i}}{\partial s} \right) = 0,\ k \in 1 \ldots n. \tag{6}$$

In such form the set of equations is overly complex. The following three simplifications are proposed for the iterative solution approach taken in the model.

$$\frac{\partial \Delta\omega\left(\boldsymbol{\rho}\left(s\right)\right)}{\partial \rho_k} = \Delta\mu_k \tag{7}$$

Secondly the equation (6) also contains the non-diagonal influence parameter $c_{i,j} i \neq j$. This type of influence parameter is rarely tabulated and has to be inferred from experimental data. This approach is available only for narrow substance range therefore the approximation of parameter is usually used instead.

$$c_{i,k} \doteq \sqrt{c_{i,i} \cdot c_{k,k}} \tag{8}$$

Lastly the influence parameter $c_{i,j}$ is also assumed to be independent on molar density. This assumption is valid for most systems exhibiting a very weak density dependence.

$$\frac{\partial c_{i,j}}{\partial \rho_k} = 0 \tag{9}$$

By combining the (7,8,9) together with a special derivative notation $\rho_i' = \partial\rho_i/\partial s$ the set of equations (6) is substantially reduced into:

$$\sum_{i=1}^{n} \sqrt{c_{i,i} \cdot c_{k,k}} \left(\frac{dS}{ds} \rho_i' + \rho_i'' \right) = \Delta\mu_k, \ k \in 1 \ldots n. \tag{10}$$

Core problem is now formulated as the set of second order differential equations with non-zero right hand side (RHS).

3 Model description

As stated in the previous section 2 the core problem lies within the solution of the second order differential equation set. Moreover the RHS of equations (10) is generally analytically non-integrable due to the fact that $\Delta\mu_k$ is computed from the EoS. Complex equation of state without analytically integrable chemical potential μ_k (such as PC-SAFT) prohibit the analytical solution. Additionally the left hand side contains dS/ds factor dependent on the interface geometry. To answer both problem simultaneously an unified numerical method for the two investigated geometries is proposed here.

$$\sum_{i=1}^{n} \sqrt{c_{i,i}} \left(\frac{dS}{ds} \rho_i' + \rho_i'' \right) = \frac{\Delta\mu_k}{c_{k,k}}, \ k \in 1 \ldots n \tag{11}$$

While solution of aforementioned core problem is possible in this form. It would require a substantial computational effort coupled with the increased error of solution and fundamentally problematic situation for system with more than two components. It is therefore quite favourable to modify the form of a problem. A similar approach as [5, 13, 9, 12] was used to transform the original set into algebraic problem and simplified differential problem. An idea similar is to restructure the set (10) in such a way that all elements with k index are transferred to the right hand side of the set and subtract the first equation form the rest. This creates the system of nonlinear equations and single differential equation to solve.

According to this schema the differential equation has to be modified to preserve the connection between sections. The connection can be expressed as a single variable X also referred as an artificial variable.

$$X = \frac{\Delta\mu_1}{\sqrt{c_{1,1}}} \tag{12}$$

In addition to the variable X the partial densities are also treated. Introduced modification is inspired by the problem of monotonous density. It is known that multiple component systems in gradient theory require at least one density to have a monotonous character along the coordinate axis. The same requirement was formulated by Cahn and

Hilliard [2] and later further investigated by Liang *et. al.* [10]. This requirement implies the remaining partial densities are expressed as functions of the one selected monotonous density.

Proposed approach inspired by [10], introduces a new modified density $\tilde{\rho}$. With this modified density the problem with selection can be softened and all partial densities are processed in same manner as a functions of $\tilde{\rho}$.

$$\tilde{\rho} = \frac{\sum_{i=1}^{n} \sqrt{c_{i,i}} \rho_i}{\sum_{i=1}^{n} \sqrt{c_{i,i}}} \tag{13}$$

Here n is the number of components in mixture and $c_{i,i}$ is influence parameter of pure i-th component. Monotonous character is justified by the existence of monotonous component with high influence parameter as in case of investigated system.

With modified density (13) and artificial variable (12) the differential section of problem can be written as:

$$\frac{dS}{ds}\tilde{\rho}' + \tilde{\rho}'' = \frac{X}{\left(\sum_{i=1}^{n} \sqrt{c_{i,i}}\right)} \tag{14}$$

This shape of equation is expressed for arbitrary geometry and specialized solver can be used for individual geometries. For example, when the factor $dS/ds = 1$ the problem can be numerically integrated. In other cases a numerical solution of differential equation is searched for.

The algebraic section is also treated with notation (12,13). Consequentially one equation has to be added into a system for modified density. The linear system is then composed of n nonlinear equations:

$$\begin{aligned} \frac{\Delta\mu_2}{\sqrt{c_{2,2}}} &= -X \\ \vdots \quad &\vdots \quad \vdots \\ \frac{\Delta\mu_n}{\sqrt{c_{n,n}}} &= -X \\ \sum_{i=1}^{n} \sqrt{c_{i,i}} \rho_i &= \tilde{\rho} \sum_{i=1}^{n} \sqrt{c_{i,i}}. \end{aligned} \tag{15}$$

Algebraic system here does not depend on the type of interface geometry as a trivial result of the previous derivation. This feature of system permits the independent solution regardless of the geometry type.

3.1 Algebraic system solution

This subsection offer a solution method for the nonlinear algebraic set of equation obtained from core problem derivation. Because of the nonlinear character of problem the Newton-Rhapson solver was selected. The numerical properties of a solver were further improved with rearrangement of the set so that Jacobean matrix is symmetric. The fact is straightforward consequence of the partial derivatives interchangeability also previously shown by [10].

$$
\begin{aligned}
\Delta\mu_2 + X\sqrt{c_{2,2}} &= 0 \\
\vdots \quad &\vdots \quad \vdots \\
\Delta\mu_n + X\sqrt{c_{n,n}} &= 0 \\
\sum_{i=1}^{n} \sqrt{c_{i,i}}\rho_i - \tilde{\rho}\sum_{i=1}^{n} \sqrt{c_{i,i}} &= 0
\end{aligned} \tag{16}
$$

The computational procedure of the algebraic solver can be therefore developed around the Newton-Rhapson iterator with the Jacobean inversion method. The whole procedure is in steps applied across the modified density discretization and individual solution are found. These values are coupled into the following data structure evaluated for discrete modified densities $\tilde{\rho}^1, \tilde{\rho}^2, \ldots \tilde{\rho}^{\text{disc}}$.

This data structure is fundament for the piecewise cubic interpolation used afterwards. The interpolation enables to use fewer discretization points and alleviate some computational strain without suffering much greater error. It is also useful for following solution to hold the algebraic solver results as functions $\rho_1(\tilde{\rho}), \rho_2(\tilde{\rho}), \ldots, \rho_n(\tilde{\rho}), X(\tilde{\rho})$.

3.2 Differential equation solution

The initial algebraic solution is followed by the differential solver. In developed solver an artificial variable interpolation $X(\tilde{\rho})$ is used and a general approach is undertaken to produce the density profile dependence $\tilde{\rho}(s)$

Utilising the previous knowledge of selected interface geometry permits the specialized differential solver to be developed. This is especially useful for planar geometry case where solution can be found analytically. The analytical solution is presented in next section 3.3 . In this study we develop the general solution method primarily used for the spherical geometry. Therefore, the following equation is written with dS/ds factor substituted for spherical geometry case.

$$
\frac{2}{r}\tilde{\rho}' + \tilde{\rho}'' = \frac{X(\tilde{\rho})}{\left(\sum_{i=1}^{n}\sqrt{c_{i,i}}\right)} \tag{17}
$$

From the performed analysis of the problem and through the trial and error it has been determined that the shooting method coupled with the predictor corrector type solver can be used. Wide range of methods were tested and deemed to be not useful because of the widespread convergence issues.

The solution method in theory translates the originally boundary value problem into the initial value problem. Therefore, the investigation of droplets in this case can access an information about gas density of surroundings. Also the initial bulk liquid density derivative is known and understood as being zero, because of the requirement of homogeneously distributed density in volume in the centre of droplet. The task for the shooting method is to find initial density that yields the density profile finishing at the a priory known gas density. For droplets the shooting parameter is the initial liquid density which correspond with experimentally measured systems.

The shooting method is also supplied with a decision criteria. The criteria is responsible for the selection of the next shooting parameter α_{next}. It was found out that a bisection method construct a reliable criteria and is able to cope with steep nature of investigate searching task for α_{optimal} shooting parameter.

3.3 Density profile computation

As mentioned in the previous section the core problem can be solved analytically in special case of planar phase interface geometry. This was well investigated [13, 8, 11] and found out that the shape of planar phase interface density profile can be computed as following integral:

$$z(\rho) = z_0 + \int_{\rho_0}^{\rho} \sqrt{\frac{\sum_{i,j=1}^{n} c_{i,j} \left(\frac{\partial \boldsymbol{\rho_i}}{\partial z}\right) \left(\frac{\partial \boldsymbol{\rho_j}}{\partial z}\right)}{2\Delta\omega}}\, d\rho \tag{18}$$

Here the ρ_0 and z_0 stand for initial selected values for initial density of integration as the centre of profile respectively. These two parameters determine how is the profile oriented and where it begins. This approach also replace differential for numerical integral computation and only the partial densities are left to be determined.

In spherical case geometry the differential solver produces result in a form of modified density function of radius $\tilde{\rho}(r)$ further modified into $\rho_i r$. The process includes transformation of modified density and partial density computation base on algebraic set solution. With interpolated functions $\rho_i(\tilde{\rho})$ the transformation of $\tilde{\rho}(r)$ into $\rho_i(r)$ becomes trivial. This operation depends on the monotonousness of modified density which implies injectivity required for transformation.

The main property of interface is surface tension. This property states the force exerted onto the dividing surface that holds the phases separate. For the systems with planar interface geometry the generally known [21, 12, 14] expression for surface tension is used as:

$$\sigma = \int_{\rho^{\mathrm{G}}}^{\rho^{\mathrm{L}}} \sqrt{2\Delta\omega \sum_{i,j=1}^{n} c_{i,j} \left(\frac{\partial \boldsymbol{\rho_i}}{\partial z}\right) \left(\frac{\partial \boldsymbol{\rho_j}}{\partial z}\right)}\, d\rho \tag{19}$$

Following the argument by Liang *et. al.* [10] the integration can be also performed in modified density which gives a negligible boost to the accuracy, because this way the computation does not rely on modified density backward transformation. The second case of spherical geometry offers no such direct approach and the Young-Laplace equation have to be used for computation. It should be noted that saturation of system plays important role as input parameter in droplet density profile computation. This state can be identified with the Laplace pressure Δp. After a simple treatment the equation for spherical surface tension is obtained.

$$\sigma = \sqrt[3]{\frac{3\Delta\Omega\Delta p^2}{16\pi}} \tag{20}$$

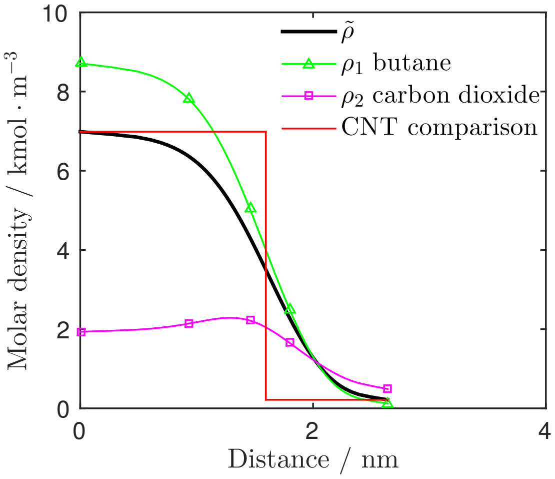


Figure 1: Planar density profiles of $C_4H_{10} - CO_2$ mixture for $T = 300\,\mathrm{K}$, $p = 1.36\,\mathrm{MPa}$ and $\Delta p = 0\,\mathrm{MPa}$

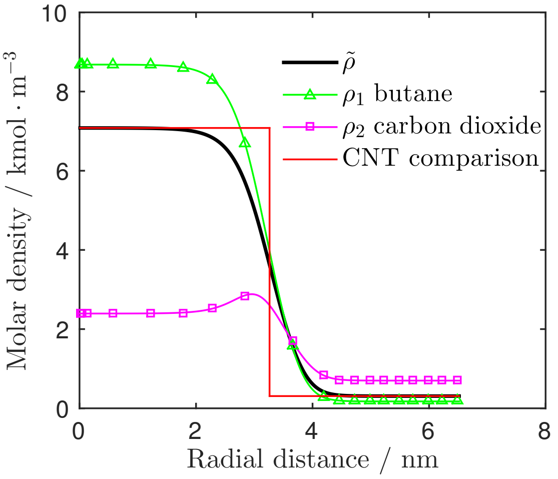


Figure 2: Spherical density profiles of $C_4H_{10} - CO_2$ mixture for $T = 300\,\mathrm{K}$, $p = 1.85\,\mathrm{MPa}$ and $\Delta p = 5.53\,\mathrm{MPa}$

4 Results

Density profiles give the information about phase interface and they are computed with either (18) equation for planar geometry or according to the method described in section about differential equation solution. These solutions are presented for the $C_4H_{10} - CO_2$ mixture depicting both investigated geometries. Distinct feature of both figures is the substantial adsorption of carbon dioxide. The adsorption is more pronounced with increased Δp illustrated with figures for $\Delta p = 0$ and $\Delta p = 5.53$MPa. It can be also noted that profiles are computed until the stop criteria evaluation which in spherical case result in longer gaseous part of profile. Because of a direct computation of planar case geometry the Fig. 1. have no such feature. Additionally the planar geometry has an arbitrary selected initial distance of computation here set to $z = 0$. This means it should be used only as reference for interface width in contrary to the spherical geometry where radial distance is directly related to the size of droplet.

For the more complete comparison we also calculated the surface tension of $C_4H_{10} - CO_2$ mixture and compared it with the measurements of surface tension performed by Brauer and Haugh [1] at Fig. 3. and Hsu, Nagarajan and Robinson [7] at Fig. 4. Both figures depict the planar case because the experimental data for surface tension of droplets are presently non-existent.

Figure 3. show good agreement of the experimental data with model across measured temperatures. The model is qualitatively very well aligned to experiment with constant over-prediction under 10% of modelled value. More troubling is problem with aborted computation visible for lower temperatures $T <= 327.59\,\mathrm{K}$. These points were omitted because the computation was terminated prematurely due to the improper equilibrium conditions. Such problem is caused by non-compatible prediction of equilibrium state from equation of state as compared with experimentally measured values. This issue remain a task for future development with the aim for more robust equilibrium evaluation.

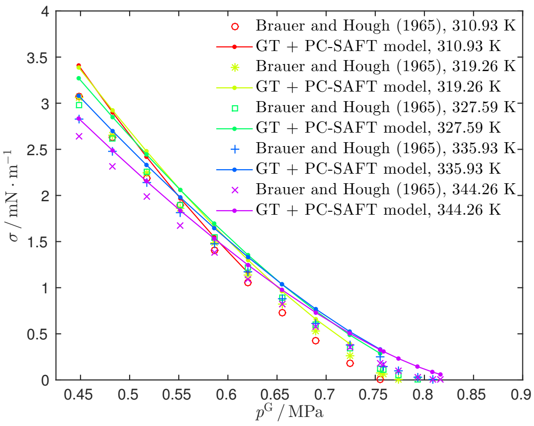


Figure 3: Comparison of model and experimental data surface tension for $C_4H_{10} - CO_2$ mixture for $T = 310.93\,\mathrm{K} - 344.26\,\mathrm{K}$.

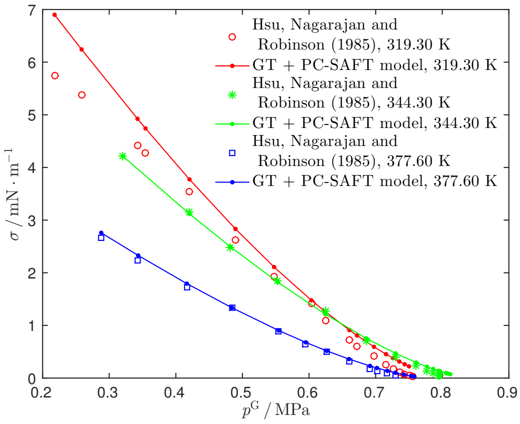


Figure 4: Comparison of model and experimental data surface tension for $C_4H_{10} - CO_2$ mixture for $T = 319.30\,\mathrm{K}$, $344.30\,\mathrm{K}$ and $377.30\,\mathrm{K}$.

In second comparison for $C_4H_{10} - CO_2$ mixture at Fig. 4 three datasets are compared. System conditions were well reproduced by model with a stunning precision for higher temperatures 344.30K and 377.30K. The prediction for temperature $T = 319.30$K provides appropriate estimation for higher pressures and deviates slightly more in region of lower pressures around $0.2 - 0.35$MPa. Aforementioned precision of prediction can be attributed to selected EoS and system combination with medium carbohydrate and carbon dioxide. Similar behaviour is expected for larger carbohydrates where prediction of thermodynamic properties is better.

5 Conclusions

This study presents the unified mathematical model for two types of phase interface geometry targeted at multi-component mixtures. The model is based on gradient theory description of interface and utilise an advanced PC-SAFT EoS for equilibrium and system properties calculation. The study also present an overview of proposed model together with derivation of model key points. At the end of derivation the used formulas for density profile and surface tension results are presented.

The proposed solution utilize the special shape of the simplified problem and enables the innovative two step solution. The presented solution also unifies the two types of investigated geometry previously not mentioned in literature. The model was tested on binary system of carbon dioxide and methane which falls into the category of CCS relevant mixtures. Modelled results were compared with experimental values of surface tension and a close correspondence of prediction and data was observed.

References

[1] E.B. Brauer and E.W. Hough. Interfacial tension of the normal butane-carbon dioxide system. *Producers Monthly*, 29(8):13–..., 1965.

[2] J.W. Cahn. Free energy of a nonuniform system .2. thermodynamic basis. *J. Chem. Phys.*, 30(5):1121–1124, 1959.

[3] J.W. Cahn and J.E. Hilliard. Free energy of a nonuniform system .1. interfacial free energy. *J. Chem. Phys.*, 28(2):258–267, 1958.

[4] D. Celný, V. Vinš, B. Planková, and J. Hrubý. Mathematical modeling of planar and spherical vapor–liquid phase interfaces for multicomponent fluids. In *EPJ Web of Conferences*, volume 114, page 02011. EDP Sciences, 2016.

[5] P.M.W. Cornelisse, C.J. Peters, and J. de Swaan Arons. Application of the peng-robinson equation of state to calculate interfacial tensions and profiles at vapour-liquid interfaces. *Fluid Phase Equilib*, 82:119 – 129, 1993.

[6] J. Hrubý, D. G. Labetski, and M. E. H. van Dongen. Gradient theory computation of the radius-dependent surface tension and nucleation rate for n-nonane clusters. *The Journal of Chemical Physics*, 127(16):164720, 2007.

[7] J.J.C. Hsu, N. Nagarajan, and R.L. Robinson. Equilibrium phase compositions, phase densities, and interfacial-tensions for co2 + hydrocarbon systems .1. co2 + normal-butane. *J. Chem. Eng. Data*, 30(4):485–491, 1985.

[8] Heike K. and Sabine E. Calculation of surface properties of pure fluids using density gradient theory and saft-eos. *Fluid Phase Equilibria*, 172(1):27 – 42, 2000.

[9] T. Lafitte, B. Mendiboure, M. Pineiro, D. Bessieres, and Ch. Miqueu. Interfacial properties of water/co2: A comprehensive description through a gradient theory-saft-vr mie approach. *J. Phys. Chem. B*, 114(34):11110–11116, SEP 2 2010.

[10] Xiaodong Liang, Michael Locht Michelsen, and Georgios M. Kontogeorgis. Pitfalls of using the geometric-mean combining rule in the density gradient theory. *Fluid Phase Equilib.*, 415:75–83, MAY 15 2016.

[11] Hong Lin, Yuan-Yuan Duan, and Qi Min. Gradient theory modeling of surface tension for pure fluids and binary mixtures. *Fluid Phase Equilibria*, 254(1–2):75 – 90, 2007.

[12] J. M. Miguez, Matias G. J., F. J. Blas, H. Segura, A. Mejia, and M. M. Pineiro. Comprehensive characterization of interfacial behavior for the mixture co2 + h2o + ch4: Comparison between atomistic and coarse grained molecular simulation models and density gradient theory. *J. Phys. Chem. C*, 118(42):24504–24519, OCT 23 2014.

[13] C. Miqueu, B. Mendiboure, A. Graciaa, and J. Lachaise. Modelling of the surface tension of pure components with the gradient theory of fluid interfaces: a simple

and accurate expression for the influence parameters. *Fluid Phase Equilib.*, 207(1-2):225–246, MAY 30 2003.

[14] C. Miqueu, B. Mendiboure, A. Graciaa, and J. Lachaise. Modeling of the surface tension of multicomponent mixtures with the gradient theory of fluid interfaces. *Industrial & engineering chemistry research*, 44(9):3321–3329, APR 27 2005.

[15] C. Miqueu, B. Mendiboure, C. Graciaa, and J. Lachaise. Modelling of the surface tension of binary and ternary mixtures with the gradient theory of fluid interfaces. *Fluid Phase Equilibria*, 218(2):189 – 203, 2004.

[16] Erich A. Müller and Andrés Mejía. Interfacial properties of selected binary mixtures containing n-alkanes. *Fluid Phase Equilibria*, 282(2):68 – 81, 2009.

[17] A. Obeidat, M. Gharaibeh, H. Ghanem, F. Hrahsheh, N. Al-Zoubi, and G. Wilemski. Nucleation rates of methanol using the saft-0 equation of state. *ChemPhysChem*, 11(18, SI):3987–3995, DEC 17 2010.

[18] Barbora Planková, Jan Hrubý, and Václav Vinš. Prediction of the homogeneous droplet nucleation by the density gradient theory and pc-saft equation of state. In *Nucl. and Atmos. Aerosols: 19th Int. Conf.*, 2013.

[19] V. Vinš, B. Planková, J. Hrubý, and D. Celný. Density gradient theory combined with the pc-saft equation of state used for modeling the surface tension of associating systems. *EPJ Web Conferences*, 67(02129), 2014.

[20] O. Wilhelmsen, D. Bedeaux, and D. Reguera. Communication: Tolman length and rigidity constants of water and their role in nucleation. *J. Chem. Phys.*, 142(17), MAY 7 2015.

[21] Y.X. Zuo and E.H. Stenby. Calculation of interfacial tensions with gradient theory. *Fluid Phase Equilib.*, 132(1-2):139–158, MAY 31 1997.

Bayesian Approach to Hurst Exponent Estimation*

Martin Dlask

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

školitelé:

Jaromír Kukal, Katedra softwarového inženýrství
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Pavel Sovka, Katedra teorie obvodů
Fakulta elektrotechnická, ČVUT v Praze

Abstract. Fractal investigation of a signal often involves estimating its fractal dimension or Hurst exponent H when considered as a sample of a fractional process. Fractional Gaussian noise (fGn) belongs to the family of self-similar fractional processes and it is dependent on parameter H. There are variety of traditional methods for Hurst exponent estimation. Our novel approach is based on zero-crossing principle and signal segmentation. Thanks to the Bayesian analysis, we present a new axiomatically based procedure of determining the expected value of Hurst exponent together with its standard deviation and credible intervals. The statistical characteristics are calculated at the interval level at first and then they are used for the deduction of the aggregate estimate. The methodology is subsequently used for the EEG signal analysis of patients suffering from Alzheimer disease.

Keywords: fractal dimension, Hurst exponent, Bayesian approach, EEG, Alzheimer disease

Abstrakt. Hurstův exponent H je užitečnou charakteristikou pro fraktální analýzu signálu, který je zkoumán jako realizace náhodného zlomkového procesu. Zlomkový Gaussův šum (fGn) patří do třídy soběpodobných zlomkových procesů a je závislý na stejném parametru H. V současné době existuje řada tradičních metod, které slouží pro odhad Hurstova exponentu. Nový přístup k odhadu je založen na charakteristice průchodů signálu nulou a využívá jeho segmentaci. S využitím Bayesovské analýzy je představena nová axiomaticky založená procedura odhadu H, která poskytuje jeho standardní odchylku a konfidenční interval. Statistické charakteristiky jsou nejprve odhadovány na úrovni jednoho segmentu a následně jsou použity pro stanovení celkového odhadu. Metoda je použita na analýzu signálu EEG pro identifikaci pacientů, kteří trpí Alzheimerovou chorobou.

Klíčová slova: fraktální dimenze, Hurstův exponent, Bayesovský přístup, EEG, Alzheimerova choroba

Plná verze: M. Dlask, J. Kukal, O. Vysata. *Bayesian Approach to Hurst Exponent Estimation.* Methodology and Computing in Applied Probability **19** (2017), 973–983.

*Tato práce byla podpořena grantem SGS ČVUT č. SGS14/208/OHK4/3T/14.

Area-Level Gamma Mixed Model[*]

Ondřej Faltys

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

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

Abstract. In practise we can encounter many problems where is useful (and sometimes necessary) to employ small area estimation (SAE) methods to obtain reliable estimates of characteristics of interest (means, totals, quantiles, etc.). The contribution deals with an area-level gamma mixed model that can be useful in some applications involving only positive responses (e.g. in a financial sector). To obtain estimates of regression parameters and predictors of random effects the PQL algorithm and the ML Laplace approximation algorithm are introduced. In order to check the behaviour of the fitting algorithms we perform simulation experiments and compare acquired results of both of them.

Keywords: Area-level model, Generalized linear mixed model, PQL algorithm, ML Laplace approximation algorithm

Abstrakt. V praxi lze narazit na řadu problémů, kde je užitečné (a často nezbytné), použít metody odhadování v malých oblastech, abychom získali odhady charakteristik, které nás zajímají (středních hodnot, kvantilů, atd.). Tento článek pojednává o statistickém modelu na úrovni oblastí, kde předpokládáme, že odezvy mají gamma rozdělení. Domníváme se, že by tento model mohl být užitečný v praktických aplikacích vyžadujících pouze kladné odezvy (např. ve finančním sektoru). K odhadu regresních parametrů a predikci náhodných efektů použijeme PQL algoritmus a ML Laplaceův aproximační algoritmus. Následně provedeme simulační experiment, abychom ověřili kvalitu výstupů obou algoritmů.

Klíčová slova: Model na úrovni oblastí, Zobecněný lineární smíšený model, PQL algoritmus, ML Laplaceův aproximační algoritmus.

1 Introduction

Small area estimation models can be divided into two parts: area-level models and unit-level models. Considering area-level models, data are available (unlike unit-level models) only at the area level. Data collected for each domain are usually used to compute the direct estimate of investigated characteristic (e.g. mean). In unit-level models there are some auxiliary data even at the individual level. One of the most basic area-level models is the Fay-Herriot model that can be expressed as (see [1])

$$y_d = \mathbf{x}_d^T \boldsymbol{\beta} + v_d + e_d, \quad d = 1, \dots, D,$$

[*] This work was supported by the grant SGS15/214/OHK4/3T/14. This work has arisen in cooperation with Domingo Morales: the author used some parts from a still not published article dealing with this topic.

where $\boldsymbol{\beta}$ is a vector of regression parameters, $e_d \sim N(0, \sigma_d^2)$ are independent sampling errors and $v_d \sim N(0, \sigma_v^2)$ are independent random effects. It is also assumed that the random effects are independent on the samplings errors and the variances $\sigma_1^2, \dots, \sigma_D^2$ are known. The model has $p+1$ unknown parameters: $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ and σ_v^2. The task is then to estimate the quantity $\mu_d = \mathbf{x}_d^T \boldsymbol{\beta} + v_d$. In this work we suppose that the responses have the gamma distribution and we try to estimate unknown parameters.

2 Model

We consider a set of random effects $\{v_d : d = 1, \dots, D\}$ such that $v_d \overset{\text{iid}}{\sim} N(0,1)$. In matrix notation we have $\mathbf{v} = (v_1, \dots, v_D)^T \sim N_D(\mathbf{0}, \mathbf{I}_D)$, i.e.

$$f_{\mathbf{v}}(\mathbf{v}) = \frac{1}{(2\pi)^{D/2}} \exp\left\{ -\frac{1}{2} \mathbf{v}^T \mathbf{v} \right\}.$$

The conditional distribution of the target variable y_d given v_d is

$$y_d | v_d \sim Gamma\left(\nu_d, a_d = \frac{\nu_d}{\mu_d} \right), \quad d = 1, \dots, D$$

and the density follows

$$f(y_d|v_d) = \frac{a_d^{\nu_d}}{\Gamma(\nu_d)} y_d^{\nu_d - 1} \exp\{-a_d y_d\} I_{(0,\infty)}(y_d) = \left(\frac{\nu_d}{\mu_d} \right)^{\nu_d} \frac{y_d^{\nu_d - 1}}{\Gamma(\nu_d)} \exp\left\{ -\frac{\nu_d}{\mu_d} y_d \right\} I_{(0,\infty)}(y_d).$$

The expectation and variance of the conditional random variable y_d given v_d are

$$E[y_d|v_d] = \frac{\nu_d}{a_d} = \mu_d, \quad \operatorname{var}[y_d|v_d] = \frac{v_d}{a_d^2} = \frac{\mu_d^2}{\nu_d}.$$

The canonical link for the gamma distribution (see [2]) is the inverse link, $g(x) = \frac{1}{x}$, then we model the conditional expectation μ_d as

$$g(\mu_d) = \frac{1}{\mu_d} = \mathbf{x}_d^T \boldsymbol{\beta} + \phi v_d, \quad d = 1, \dots, D,$$

where $\beta = (\beta_1, \dots, \beta_p)^T$ and $\mathbf{x}_d^T = (x_{d1}, \dots, x_{dp})$. Considering the data $\mathbf{y} = (y_1, \dots, y_D)^T$ satisfy the assumptions of GLMM the random variables $y_d|v_d$, $i = 1, \dots, D$, are independent, i.e. $f(\mathbf{y}|\mathbf{v}) = \prod_{i=1}^D f(y_d|v_d)$. Finally, we get

$$f(\mathbf{y}) = \int_{\mathbb{R}^D} f(\mathbf{y}|\mathbf{v}) f_{\mathbf{v}}(\mathbf{v}) \mathrm{d}\mathbf{v} = \int_{\mathbb{R}^D} \psi(\mathbf{y}, \mathbf{v}) \mathrm{d}\mathbf{v}, \tag{1}$$

where

$$\begin{aligned} \psi(\mathbf{y}, \mathbf{v}) &= (2\pi)^{-D/2} \exp\left\{ -\frac{\mathbf{v}^T \mathbf{v}}{2} \right\} \prod_{d=1}^D \left(\frac{\nu_d}{\mu_d} \right)^{\nu_d} \frac{y_d^{\nu_d - 1}}{\Gamma(\nu_d)} \exp\left\{ -\frac{\nu_d}{\mu_d} y_d \right\} \\ &= (2\pi)^{-D/2} \exp\left\{ -\frac{\mathbf{v}^T \mathbf{v}}{2} \right\} \left(\prod_{d=1}^D \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} \right) \exp\left\{ \sum_{d=1}^D \nu_d \log(\mathbf{x}_d^T \boldsymbol{\beta} + \phi v_d) \right\} \times \\ &\times \exp\left\{ -\sum_{k=1}^p \left(\sum_{d=1}^D \nu_d y_d x_{dk} \right) \beta_k - \phi \sum_{d=1}^D \nu_d y_d v_d \right\}. \end{aligned}$$

The partial derivatives of $\mu_d = \frac{1}{\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d}$ are

$$\frac{\partial \mu_d}{\partial \beta_r} = -\frac{x_{dr}}{(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^2} = -x_{dr}\mu_d^2, \quad \frac{\partial \mu_d}{\partial \phi} = -\frac{v_d}{(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^2} = -v_d\mu_d^2.$$

There are $p+1$ unknown parameters in this model: $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)^T$ and ϕ. Due to the fact that the integral in (1) cannot be calculated explicitly we employ two different methods to obtain estimates of these parameters: PQL algorithm and ML Laplace approximation algorithm.

Remark 1 In practise, y_d is a direct estimate of a domain total or mean with estimated design-based variance $\sigma_d^2 = \mathrm{var}_\pi(y_d)$. By equating $\mathrm{var}(y_d|v_d)$ to σ_d^2 and substituting μ_d by y_d, we get $\sigma_d^2 = \frac{y_d^2}{\nu_d}$.

3 PQL algorithm

The ML-PQL estimator of $\boldsymbol{\beta}$ and predictor of $\mathbf{v}$ (see [3]) maximizes the joint log-likelihood

$$l = \log\psi(\mathbf{y},\mathbf{v}) = -\frac{D}{2}\log 2\pi - \frac{1}{2}\sum_{d=1}^D v_d^2 + \sum_{d=1}^D(\nu_d\log\nu_d + (\nu_d-1)\log y_d - \log\Gamma(\nu_d))$$
$$+\sum_{d=1}^D \nu_d\log(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d) - \sum_{k=1}^p\left(\sum_{d=1}^D y_d\nu_d x_{dk}\right)\beta_k - \phi\sum_{d=1}^D y_d\nu_d v_d.$$

We use the Newton-Raphson algorithm to maximize $l = l(\boldsymbol{\beta},\mathbf{v})$. The first derivatives of l with respect to $\boldsymbol{\beta}$ and $\mathbf{v}$ are

$$U_r = \frac{\partial l}{\partial \beta_r} = \sum_{d=1}^D \frac{\nu_d x_{dr}}{\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d} - \sum_{d=1}^D y_d\nu_d x_{dr}, \quad r=1,\dots,p,$$
$$U_{p+d} = \frac{\partial l}{\partial v_d} = -v_d + \frac{\nu_d\phi}{\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d} - \phi y_d\nu_d, \quad d=1,\dots,D.$$

The second derivatives of l with respect to $\boldsymbol{\beta}$ and $\mathbf{v}$ are

$$H_{r_1r_2} = \frac{\partial^2 l}{\partial\beta_{r_1}\partial\beta_{r_2}} = -\sum_{d=1}^D \frac{\nu_d x_{dr_1}x_{dr_2}}{(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^2}, \quad r_1,r_2=1,\dots,p,$$
$$H_{r,p+d} = \frac{\partial^2 l}{\partial\beta_r\partial v_d} = -\frac{\nu_d x_{dr}\phi}{(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^2}, \quad r=1,\dots,p,\ d=1,\dots,D,$$
$$H_{p+d,p+d} = \frac{\partial^2 l}{\partial v_d^2} = -1 - \frac{\nu_d\phi^2}{(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^2}, \quad d=1,\dots,D,$$
$$H_{p+d_1,p+d_2} = \frac{\partial^2 l}{\partial v_{d_1}\partial v_{d_2}} = 0, \quad d_1,d_2=1,\dots,D,\ d_1\neq d_2.$$

The updating equation for the Newton-Raphson algorithm with fixed ϕ is

$$\boldsymbol{\xi}^{(k+1)} = \boldsymbol{\xi}^{(k)} - \mathbf{H}^{-1}(\boldsymbol{\xi}^{(k)})\mathbf{U}(\boldsymbol{\xi}^{(k)}), \tag{2}$$

where $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \mathbf{v}^T)^T$, $\mathbf{U} = \mathbf{U}(\boldsymbol{\xi}) = (U_1, \ldots, U_{p+D})^T$ and $\mathbf{H} = \mathbf{H}(\boldsymbol{\xi}) = (H_{rs})_{r,s=1,\ldots,p+D}$. At the step k of the algorithm, the penalized maximum likelihood estimation of ϕ maximizes the joint likelihood of linear predictors $\eta_1^{(k)}, \ldots, \eta_D^{(k)}$ where $\eta_d^{(k)} = \mathbf{x}_d^T \boldsymbol{\beta}^{(k)} + \phi^{(k)} v_d^{(k)}$ and

$$\eta_d^{(k)} \sim N(\mathbf{x}_d^T \boldsymbol{\beta}^{(k)}, \phi^2), \quad d = 1, \ldots, D.$$

The joint log-likelihood of $\eta_1^{(k)}, \ldots, \eta_D^{(k)}$ is

$$l^{(k)} = -\frac{D}{2} \log 2\pi - D \log \phi - \frac{1}{2\phi^2} \sum_{d=1}^{D} (\eta_d^{(k)} - \mathbf{x}_d^T \boldsymbol{\beta}^{(k)})^2.$$

By taking the first derivative of $l^{(k)}$ with respect to ϕ and equating to zero, we get

$$0 = U^{(k)} = \frac{\partial l^{(k)}}{\partial \phi} = -\frac{D}{\phi} + \frac{1}{\phi^3} \sum_{d=1}^{D} (\eta_d^{(k)} - \mathbf{x}_d^T \boldsymbol{\beta}^{(k)})^2,$$

$$\phi^2 = \frac{1}{D} \sum_{d=1}^{D} (\eta_d^{(k)} - \mathbf{x}_d^T \boldsymbol{\beta}^{(k)})^2 = \phi^{(k)^2} \frac{1}{D} \sum_{d=1}^{D} v_d^{(k)^2}.$$

Finally, the ML-PQL updating equation for ϕ is

$$\phi^{(k+1)^2} = \phi^{(k)^2} \frac{1}{D} \sum_{d=1}^{D} v_d^{(k)^2}. \tag{3}$$

3.1 Algorithm

The PQL algorithm calculates predictors of $\mathbf{v}$ and estimators of $\boldsymbol{\beta}$ and ϕ. Steps of the algorithm:

1. $k := 1$ (k denotes iterations), set the values $\boldsymbol{\beta}^{(0)}$, $\mathbf{v}^{(0)}$ and $\phi^{(0)}$.
2. Run (2). Use $\phi^{(k-1)}$ as known value and $\boldsymbol{\beta}^{(k-1)}$, $\mathbf{v}^{(k-1)}$ as algorithm seeds. Let $\boldsymbol{\beta}^{(k)}$ and $\mathbf{v}^{(k)}$ be the output.
3. Update ϕ by using the updating equation (3), i.e.
$$\phi^{(k)^2} = \phi^{(k-1)^2} \frac{1}{D} \sum_{d=1}^{D} v_d^{(k)^2}.$$
4. Repeat the steps 2-3 until the convergence of $\boldsymbol{\beta}^{(k)}$, $v_d^{(k)}$ and $\phi^{(k)}$.

4 ML Laplace approximation algorithm

4.1 Laplace approximation to the likelihood

Let $h : \mathbb{R} \mapsto \mathbb{R}$ be a twice continuously differentiable function with a global maximum at x_0, i.e. $\dot{h}(x_0) = 0$ and $\ddot{h}(x_0) < 0$. Taylor's series expansion of $h(x)$ around x_0 yields to

$$h(x) = h(x_0) + \frac{1}{2}\ddot{h}(x_0)(x - x_0)^2 + o(|x - x_0|^2) \approx h(x_0) + \frac{1}{2}\ddot{h}(x_0)(x - x_0)^2.$$

The univariate Laplace approximation is

$$\begin{aligned}\int_{-\infty}^{\infty} \mathrm{e}^{h(x)} &\approx \int_{-\infty}^{\infty} \mathrm{e}^{h(x_0)} \exp\left\{-\frac{1}{2}(-\ddot{h}(x_0))(x-x_0)^2\right\} \mathrm{d}x \\ &= (2\pi)^{1/2}(-\ddot{h}(x_0))^{-1/2}\mathrm{e}^{h(x_0)} \int_{-\infty}^{\infty} \frac{\exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{(-\ddot{h}(x_0))^{-1/2}}\right)^2\right\}}{(2\pi)^{1/2}(-\ddot{h}(x_0))^{-1/2}} \mathrm{d}x \\ &= (2\pi)^{1/2}(-\ddot{h}(x_0))^{-1/2}\mathrm{e}^{h(x_0)}. \end{aligned} \tag{4}$$

Recalling assumptions, $v_1, \dots, v_d \sim N(0,1)$ are independent and

$$y_d|v_d \overset{\text{ind}}{\sim} \text{Gamma}\left(\nu_d, \frac{\nu_d}{\mu_d}\right), \quad \mu_d = \mu_d(v_d) = (\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)^{-1}, \quad d = 1, \dots, D.$$

The marginal density of y_d can be expressed as

$$\begin{aligned} f(y_d) &= \int_{-\infty}^{\infty} f(y_d|v_d)f(v_d)\mathrm{d}v_d \\ &= \int_{-\infty}^{\infty} \frac{\nu_d^{\nu_d} y_d^{\nu_d-1}}{(2\pi)^{1/2}\Gamma(\nu_d)} \exp\{\nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\} \exp\left\{-\frac{1}{2}v_d^2\right\} \mathrm{d}v_d \\ &= \frac{\nu_d^{\nu_d} y_d^{\nu_d-1}}{(2\pi)^{1/2}\Gamma(\nu_d)} \int_{-\infty}^{\infty} \exp\left\{-\frac{v_d^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\} \mathrm{d}v_d \\ &= \frac{\nu_d^{\nu_d} y_d^{\nu_d-1}}{(2\pi)^{1/2}\Gamma(\nu_d)} \int_{-\infty}^{\infty} \exp\{h(v_d)\}\mathrm{d}v_d, \end{aligned}$$

where

$$\begin{aligned} h(v_d) &= -\frac{v_d^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d), \qquad (5) \\ \dot{h}(v_d) &= -v_d + \frac{\nu_d\phi}{\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d} - \phi\nu_d y_d = -v_d + \phi\nu_d\mu_d(v_d) - \phi\nu_d y_d, \\ \ddot{h}(v_d) &= -\left(1 + \frac{\phi^2\nu_d}{(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)^2}\right) = -(1 + \phi^2\nu_d\mu_d^2(v_d)). \end{aligned}$$

Let v_{0d} denote the global maximum of h then $\dot{h}(v_{0d}) = 0$ and $\ddot{h}(v_{0d}) < 0$. By applying (4) in $v_d = v_{0d}$, we get

$$\begin{aligned} f(y_d) \approx \frac{\nu_d^{\nu_d} y_d^{\nu_d-1}}{\Gamma(\nu_d)}&(1 + \phi^2\nu_d\mu_d^2(v_{0d}))^{-1/2} \times \\ &\times \exp\left\{-\frac{v_{0d}^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d}) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d})\right\}. \end{aligned}$$

It holds that $y_1, \dots, y_D$ are unconditionally independent and then the likelihood has the form $L(\boldsymbol{\beta}, \phi) = \prod_{i=1}^D f(y_i)$. The log-likelihood is $l(\boldsymbol{\beta}, \phi) = \sum_{d=1}^D l_d$, where

$$\begin{aligned} l_d = \log f(y_d) \approx l_{0d} = \log \frac{\nu_d^{\nu_d} y_d^{\nu_d-1}}{\Gamma(\nu_d)} - \frac{1}{2}\log\xi_{0d} - \frac{v_{0d}^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d}) \\ - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d}), \end{aligned}$$

where $\xi_{0d} = 1 + \phi^2 \nu_d \mu_{0d}^2$ and $\mu_{0d} = \mu_d(v_{0d})$. The first derivatives of μ_{0d} and ξ_{0d} are

$$\begin{aligned}\frac{\partial \mu_{0d}}{\partial \beta_r} &= -x_{dr}\mu_{0d}^2, \quad \eta_{0dr} = \frac{\partial \xi_{0d}}{\partial \beta_r} = -2\phi^2 \nu_d x_{dr} \mu_{0d}^3, \\ \frac{\partial \mu_{0d}}{\partial \phi} &= -v_{0d}\mu_{0d}^2, \quad \eta_{0d} = \frac{\partial \xi_{0d}}{\partial \phi} = 2\phi \nu_d \mu_{0d}^2 - 2\phi^2 \nu_d v_{0d} \mu_{0d}^3.\end{aligned}$$

The first derivatives of l_{0d} with respect to β_r and ϕ are

$$\frac{\partial l_{0d}}{\partial \beta_r} = -\frac{1}{2}\frac{\eta_{0dr}}{\xi_{0d}} + \nu_d x_{dr} \mu_{0d} - \nu_d x_{dr} y_d, \quad \frac{\partial l_{0d}}{\partial \phi} = -\frac{1}{2}\frac{\eta_{0d}}{\xi_{0d}} + \nu_d v_{0d} \mu_{0d} - \nu_d v_{0d} y_d.$$

It holds that

$$\begin{aligned}\frac{\partial \eta_{0dr}}{\partial \beta_s} &= 6\phi^2 \nu_d x_{dr} x_{ds} \mu_{0d}^4, \quad \frac{\partial \eta_{0dr}}{\partial \phi} = -4\phi \nu_d x_{dr} \mu_{0d}^3 + 6\phi^2 \nu_d x_{dr} v_{0d} \mu_{0d}^4, \\ \frac{\partial \eta_{0d}}{\partial \beta_r} &= -4\phi \nu_d x_{dr} \mu_{0d}^3 + 6\phi^2 \nu_d v_{0d} x_{dr} \mu_{0d}^4, \quad \frac{\partial \eta_{0d}}{\partial \phi} = 2\nu_d \mu_{0d}^2 - 8\phi \nu_d v_{0d} \mu_{0d}^3 + 6\phi^2 \nu_d v_{0d}^2 \mu_{0d}^4.\end{aligned}$$

The second partial derivatives of l_d are

$$\begin{aligned}\frac{\partial^2 l_{0d}}{\partial \beta_s \partial \beta_r} &= -\frac{1}{2}\frac{\frac{\partial \eta_{0dr}}{\partial \beta_s}\xi_{0d} - \eta_{0dr}\eta_{0ds}}{\xi_{0d}^2} - \nu_d x_{dr} x_{ds} \mu_{0d}^2, \\ \frac{\partial^2 l_{0d}}{\partial \phi \partial \beta_r} &= -\frac{1}{2}\frac{\frac{\partial \eta_{0dr}}{\partial \phi}\xi_{0d} - \eta_{0dr}\eta_{0d}}{\xi_{0d}^2} - \nu_d v_{0d} x_{dr} \mu_{0d}^2, \\ \frac{\partial^2 l_{0d}}{\partial \phi^2} &= -\frac{1}{2}\frac{\frac{\partial \eta_{0d}}{\partial \phi}\xi_{0d} - \eta_{0d}^2}{\xi_{0d}^2} - \nu_d v_{0d}^2 \mu_{0d}^2.\end{aligned}$$

For $r, s = 1, \dots, p+1$, the components of the score vector and the Hessian matrix are

$$U_{0r} = \sum_{d=1}^{D} \frac{\partial l_{0d}}{\partial \beta_r}, \quad U_{0p+1} = \sum_{d=1}^{D} \frac{\partial l_{0d}}{\partial \phi},$$

$$H_{0rs} = H_{0sr} = \sum_{d=1}^{D} \frac{\partial^2 l_{0d}}{\partial \beta_s \partial \beta_r}, H_{0rp+1} = H_{0p+1r} = \sum_{d=1}^{D} \frac{\partial^2 l_{0d}}{\partial \phi \partial \beta_r}, H_{0p+1p+1} = \sum_{d=1}^{D} \frac{\partial^2 l_{0d}}{\partial \phi^2}.$$

In matrix form we have $\mathbf{U}_0 = \mathbf{U}_0(\boldsymbol{\theta}) = (U_{01}, \dots, U_{0p+1})^T$ and $\mathbf{H}_0 = \mathbf{H}_0(\boldsymbol{\theta}) = (H_{0rs})_{r,s=1,\dots,p+1}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$. The Newton-Raphson algorithm maximizes $l_0(\boldsymbol{\theta})$, with fixed $v_d = v_{0d}$, $d = 1, \dots, D$. The updating equation is

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}_0^{-1}(\boldsymbol{\theta}^{(k)})\mathbf{U}_0(\boldsymbol{\theta}^{(k)}). \tag{6}$$

For $d = 1, \dots, D$, the Newton-Raphson algorithm maximizes $h(v_d) = h(v_d, \boldsymbol{\theta})$, defined in (5), with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ fixed. The updating equation is

$$v_d^{(k+1)} = v_d^{(k)} - \frac{\dot{h}(v_d^{(k)}, \boldsymbol{\theta}_0)}{\ddot{h}(v_d^{(k)}, \boldsymbol{\theta}_0)}. \tag{7}$$

4.2 Algorithm

The ML Laplace approximation algorithm is

1. Set the initial values $k = 0$, $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(-1)} = \boldsymbol{\theta}^{(0)} + \mathbf{1}_{p+1}$, $v_d^{(0)} = 0$, $v_d^{(-1)} = 1$, $d = 1, \dots, D$.

2. Until $||\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}|| < \varepsilon_1$, $|v_d^{(k)} - v_d^{(k-1)}| < \varepsilon_2$, $d = 1, \dots, D$, do

   (a) Apply algorithm (7) with seeds $v_d^{(k)}$, $d = 1, \dots, D$, convergence tolerance ε_2 and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ fixed. Output: $v_d^{(k+1)}$, $d = 1, \dots, D$.

   (b) Apply algorithm (6) with seed $\boldsymbol{\theta}^{(k)}$, convergence tolerance ε_1 and $v_{0d} = v_d^{(k+1)}$ fixed, $d = 1, \dots, D$. Output: $\boldsymbol{\theta}^{(k+1)}$.

   (c) $k \leftarrow k + 1$

3. Output: $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k)}$, $\hat{v}_d = v_d^{(k)}$, $d = 1, \dots, D$.

5 Simulation experiments

The target of simulations is to check the behaviour of the fitting algorithms: PQL and Laplace approximation algorithm. We set the true values of parameters as $\beta_0 = 0.05$, $\beta_1 = 0.1$ and $\phi = 0.01$, i.e. $p = 2$. Let $D = 50, 100, 150, 200$ be the number of domains to be considered. For $d = 1, \dots, D$, we generate $\nu_d = 100$, $x_d = \frac{d}{D}$, $v_d \sim N(0, 1)$ and

$$y_d \sim \text{Gamma}\left(\nu_d, \frac{\nu_d}{\mu_d}\right), \text{where} \quad \mu_d = (\beta_0 + \beta_1 x_d + \phi v_d)^{-1}.$$

Steps of the algorithm

1. Repeat $K = 1000$ times ($k = 1, \dots, D$)

   (a) Generate a sample $\{y_d | d = 1, \dots, D\}$.

   (b) Calculate $\hat{\beta}_0^{(k)}$, $\hat{\beta}_1^{(k)}$ and $\hat{\phi}^{(k)}$.

2. For $\theta \in \{\beta_0, \beta_1, \phi\}$, calculate

$$BIAS = \frac{\sum_{k=1}^{K} (\hat{\theta}^{(k)} - \theta)}{K}, \quad MSE = \frac{\sum_{k=1}^{K} (\hat{\theta}^{(k)} - \theta)^2}{K}.$$

As can be seen from tables 1 and 2, ML Laplace approximation algorithm seems to work well. Despite of the very small values of both BIAS and MSE for the PQL algorithm, there is a problem with estimation of the parameter ϕ. We suppose that the true value of ϕ is 0.01 but the output of the PQL algorithm for ϕ is smaller by several orders. The estimations of the regression parameters β_0 and β_1 by PQL are, however, very well.

	$D = 50$		$D = 100$		$D = 150$		$D = 200$	
	PQL	Lap	PQL	Lap	PQL	Lap	PQL	Lap
$\hat{\beta}_0$	-0.0019	0.0042	-0.002	0.0042	-0.0018	0.0041	-0.0018	0.0041
$\hat{\beta}_1$	0.0016	0.0013	0.0017	0.0014	0.0015	0.0016	0.0014	0.0016
$\hat{\phi}$	-0.01	0.0051	-0.01	0.0052	-0.01	0.0052	-0.01	0.0052

Table 1: BIAS depending on the number of the domains D.

	$D = 50$		$D = 100$		$D = 150$		$D = 200$	
	PQL	Lap	PQL	Lap	PQL	Lap	PQL	Lap
$\hat{\beta}_0$	1.96e-05	0.00006	1.17e-05	3.92e-05	8.39e-06	3.27e-05	7.07e-06	3.09e-05
$\hat{\beta}_1$	6.32e-05	0.00012	3.37e-05	6.98e-05	2.16e-05	4.94e-05	1.71e-05	4.24e-05
$\hat{\phi}$	9.99e-05	0.00003	1e-04	2.88e-05	9.99e-05	2.80e-05	9.99e-05	2.79e-05

Table 2: MSE depending on the number of the domains D.

References

[1] T. Hobza. *Model-based methods for small area estimation.* Habilitation Thesis (2017), 23–26.

[2] Ch. E. McCulloch, S. R. Searle. *Generalized, Linear, and Mixed Models.* Wiley Series in Probability and Statistics (2001), 135–142.

[3] N. E. Breslow, D. G. Clayton. *Approximate Inference in Generalized Linear Mixed Models.* Journal of the American Statistical Association, **Vol. 88**, **No. 421** (1993), 9–25.

Towards Reliable Anomaly Detection on Difficult Data

Martin Flusser

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

advisors:

Petr Somol, Cognitive Research at Cisco Systems

Vladimír Jarý, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The anomaly detection is sub field of artificial intelligence the aim of which is identifying data that are somehow different from an expected pattern. Anomaly detection is also known as one-class classification because it is a similar task to the classification with the only difference: The training set contains the only class. This makes the task difficult because the character of the anomalous data is unknown when the model is trained. We give a survey of neural network based models for anomaly detection and their noise robust modifications. The performance is evaluated on the most advanced benchmark data for the anomaly detection

Keywords: Anomaly detection, autoencoder, replicator neural network

Abstrakt. Detekce anomálií je podoborem umělé inteligence a zabývá se nalezením anomálních prvků. Jako anomální se dají považovat data (pozorování), která jsou rozdílná buď od vzorových dat, nebo od očekávaného vzoru. Tato úloha se někdy nazývá jako jednotřídní klasifikace a to proto, že pro trénování modelu jsou k dispozici pouze data z jedné konkrétní třídy. Avšak detekce anomálií je mnohem složitější a obtížnější úkol než klasifikace, protože při detekci anomálii není předem znám charakter anomálních dat a je nutné rozhodovat, jak velké výchylky musí data dosáhnout, aby byla detekována jako anomální. V textu jsou popsány již známé modely neuronových sítí pro detekci anomálií včetně těch robustních vůči šumu. V závěru je testována přesnost těchto metod na zatím nejpokročilejších testovacích datech pro anomální detekci.

Klíčová slova: Detekce anomálií, autoencoder, neuronové sítě

1 Introduction

Representation Learning is enabler of many types of models - classifiers, anomaly detectors, etc. We focus on anomaly detection as the field that is relatively least researched, while constantly gaining on importance. The anomaly detection is identifying data, items and observations that are different from the other data or does not conform the expected pattern. It is widely applied in many fields such as medicine, banking and credit card fraud detection, system health monitoring, intrusion detection and network security.

Our ultimate aim is to define models well usable in large scale data modeling in the area of network security. This, however, will be the next step. First, we aim at verifying our models on smaller scale benchmark data. The choice of benchmark data itself is a problem (see Sec. 3.4) - currently there is not available many good data sets allowing reliable evaluation of methods [9] [12]. The existing anomaly detection models very often fail to generalize well - some models work on some data but not on others, with other existing models, it is the other way round. Hence our initial work focused on 1) reviewing existing models (see Sec. 2.1), 2) finding best methodology for performance evaluation 3) researching options to utilize representation learning models to improve anomaly detection (autoencoders have been used before but to limited extent only, while in other fields - other than anomaly detection - they are known to provide significant results), see Sec. 2.1.

2 Anomaly detection

Anomaly detection is a subfield of machine learning and is also known as one-class classification and is similar to outlier detection. The goal is to detect a sample that is somehow different from expected pattern or other observations. Contrary to the other machine learning tasks such as classification, the anomaly detection is more difficult because the character of the anomalous data is unknown when the model is trained. In addition to that, the decision how much the sample must be different from others, to be detected as anomalous, is a problem. To solve the anomaly detection problem, we need to address the following concerns: 1) Choice of the model/ method with properties suitable for the problem. 2) Address conceptual problems including thresholding and evaluation (see Sec.3)

There is a number of methods for anomaly detection the survey of which is given in [8], [21] and [25]. An example of a simple and popular method is one-class KNN [17] that is beneficial for small scale data with an adequate structure. Next, there are methods such as kernel PCA [23], kernel density estimation (KDE), robust KDE and one-class support vector machine (SVM) that all have been dominated by neural network based method proposed in [32] because deep architectures can learn and represent behaviour and structure of the data more efficiently than shallow architectures like SVMs. Hence the following text will be focused mainly on the neural networks. A paretical focus of the work is on evaluation on real based data where the prior art is mostly lacking.

2.1 Neural networks in anomaly detection

Neural networks are utilized for anomaly detection, intrusion detection etc. in two different ways. The first is that the neural network detector is learned with the only regular data as usual in anomaly detection. The result should be an anomaly score or an another similar metric which can be thresholded. Such networks are autoencoders (see Sec.2.1.1). The second way is a usage of knowledge about the possible outliers thus the problem is more related to the classification. Despite that, it is applied as an anomaly detector (see Sec.2.1.2). The following text expects a basic knowledge of neural networks which could be found in 1992 Neural networks and fuzzy systems [18], 2014 Neural network design [10] and 2016 Deep learning book [15]

2.1.1 Autoencoders

The autoencoders are applied under various conditions with more or less sufficient results. First the autoencoder was applied on several problems in a simple way and the parameters of the neural network was the main issue. Then the autoencoder was extended to deionising autoencoder which is powerful for noisy data. Finally, a few other types of autoencoder have been introduced in last several years.

One of the earlier application was the autoencoder for credit card fraud detection [3] introduced by Aleskerov in 1997. The paper also highlights the difficulty of discovering the optimal setup of the autoencoder and demonstrates the developed user friendly GUI tool box for tuning the parameters. In 2005, Han proposed a paper about the methodology of constructing an optimal structure of the autoencoder using evolutionary algorithm [16]. Thompson demonstrated utilizing autoencoder in novelty assessment in [29]. They recognized simulated anomalous behavior of computer with the CPU's load metrics.

In 2008, the deionising autoencoder was introduced in [30] and extended in [31] by Vincent. The main point of the deionising autoencoder is that the training data are noised and as a result, the network becomes noise robust. Salt and pepper noise is frequently used in the literature for that purpose. Sakurada utilized autoencoder and extended denoising autoencoder for the problem of processing the spacecrafts' telemetry data in [27]. The paper shows an effectiveness of dimensionality reduction with autoencoder on a noised and correlated data from spacecrafts' sensors. In 2014 the potential of autoencoder's utilizing in general on a real data is demonstrated in [9] by Dau. The paper points out the problem of comparison among methods and tests the autoencoder on six data sets based on a real data.

Two different types of autoencoder were developed in last years. The main difference is the substitution of reconstruction error which forms the loss function that is minimized while training and in addition it represents the anomaly score for each sample. The reconstruction error (see Sec. 2.2) is used standardly in all the presented papers above. Variational Autoencoder based Anomaly Detection using Reconstruction Probability [4], introduced in 2015, utilizes the reconstruction probability instead of reconstruction error. Moreover the autoencoder is learned such that the training data must have a Gaussian distribution in the hidden layer. The second method Deep Structured Energy Based Models for Anomaly Detection[32], published in 2016, defines energy model that minimizes the energy for the training set while learning. The energy has an inverse relation to the reconstruction probability from [4]. Both methods are demonstrated as a noise robust.

2.1.2 Other neural networks

In 1998 Cannady designed a neural network for misuse detection [7]. The neural network has two output neurons that represent anomalous and legitim sample. It has nine fixed input neurons and the number of hidden layers was determined empirically. The disadvantage compared to the autoencoder is that the training needs samples of outliers. Meanwhile Ryan introduced neural network for intrusion detection [26] that is trained on computer's logs and commands to recognize individual users. Then a log is detected as anomalous if it is assigned to another user instead of the author. This is an example of a good utilizing of classification and the author obtained results with testing on random commands , however, the network was not tested for commands and logs not seen before. In 2005 Sarasamma proposed Hierarchical Kohonenen net for anomaly detection in network security [28]. Single and multi-layer network is performed with KDD-99 based data set. The method is designed with an expert knowledge of the data thus feature selection is performed in advice and network is predefined according to types of anomalous data. Each neuron of the layer except one represents a class of anomaly and the one is active of the anomaly is represented in following layer. In other words, in the first layer the only neuron is activated during detection and then either the neuron represents type of anomaly or it is the only one without label that suggests to go to the next layer. In addition to these methods there are many others which are similar such as [14], [33], and [24].

2.2 Autoencoder principle

The autoencoder which is also known as replicator neural network or autoassociative neural network is feed forward neural network that encodes the input to a compressed form and then decode back to replicate the input.

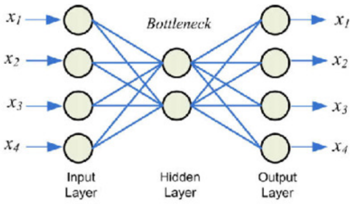


Figure 1: Structure of the autoencoder as an feed forward neural network that encodes the four-dimensional vector into two-dimensional (the hidden layer) and consequently decodes to the original space. (Credit: https://www.researchgate.net/figure/222834127_fig1_-Fig-1-The-structure-of-a-four-input-four-output-auto-encoder)

The autoencoder is composed of the encoder and the decoder such that the encoder observes and performs nonlinear dimensionality reduction with minimal loss of information and similarly the decoder performs a projection from the reduced space back to the original one. In other words, the input vector $\mathbf{x} \in \mathbb{R}^d$ is encoded to $\mathbf{y} \in \mathbb{R}^{d'}$ which is projected consequently to $\mathbf{x}' \in \mathbb{R}^d$.

The encoding is performed as:

$$\mathbf{y} = f_\theta(\mathbf{x}) = a(\mathbf{W}\mathbf{x} + \mathbf{b})$$

where f is parameterized by $\theta = \{\mathbf{W}, \mathbf{b}\}$, a is an activation function, $\mathbf{W}$ is a $d' \times d$ weight matrix and $\mathbf{b}$ is a bias vector. Similarly the decoding (reconstruction) is performed as:

$$\mathbf{x}' = g_{\theta'}(\mathbf{y}) = a(\mathbf{W}'\mathbf{x} + \mathbf{b}')$$

The parameters of the model are optimized with a training set $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}\}$ thus each vector $\mathbf{x}^{(i)} \in \mathbf{X}$ can be projected to $\mathbf{y}^{(i)}$ and $\mathbf{x}'^{(i)}$ such that the average reconstruction error is minimized:

$$\theta^*, \theta'^* = \arg\min_{\theta',\theta} \frac{1}{n} \sum_{i=1}^{n} L\Big(\mathbf{x}^{(i)}, \mathbf{x}'^{(i)}\Big) = \arg\min_{\theta',\theta} \frac{1}{n} \sum_{i=1}^{n} L\Big(\mathbf{x}^{(i)}, g_{\theta'}\big(f_\theta(\mathbf{x}^{(i)})\big)\Big)$$

where L represents a loss function which may be defined in many ways, however, the squared error $L(x, x') = ||x - x'||^2$ is the most common. [30]

Since the autoencoder is trained to minimize the reconstruction error for the training data, tested observations that do conform to the pattern of the training data will have smaller reconstruction error than observations that do not. As a consequence, the reconstruction error could represent the anomaly score and its analyses can be applied for determining outliers (see Sec.3.2).

2.2.1 Denoising autoencoder

The denoising autoencoder is a modification of the basic method which should be noise robust. The only difference is that the training data are noised for each training iteration. The already proposed methods (see Sec.2.1.1) utilize salt and pepper noise such that the only pepper corruption is performed. However the gaussian noise was not utilized in the searched papers.

3 Thresholding and evaluation

3.1 Sensitivity

The sensitivity is an essential issue of all anomaly detection problems. In practice, different setups are required according to the problem. For example, the medical tests need to be performed high sensitively not to neglect an ill patient. On contrary, the system health monitoring must not be too sensitive because the operator would ignore the alarm after many false alarms. Such a widely used setup of sensitivity gives an opportunity for a failure of the detection thus a health patient could be redundantly treated and detained in the hospital and a system could not run optimally without an alarm. However, this is still a better case, than a dead patient or a crashed system due to alarm ignorance.

3.2 Threshold

The threshold is a numerical representation of the sensitivity and it decides whether the tested sample is anomalous or not according to the anomaly score. The threshold is tuned to the optimal value for the certain application. Theoretically, if the tested subject is simple or the test is preformed perfectly, it is possible to find a perfect threshold with a total true rate. In other words, the informative value of the test's result is in the separability of the distribution of regular and anomalous samples (see Fig.2). In addition to the method's quality, the training set has a significant influence on the result of the

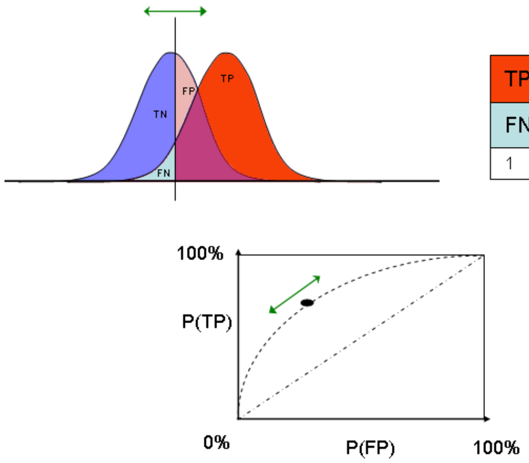


Figure 2: Thresholding - The graph in the upper left corner shows the distribution of anomaly score for the regular samples (left peak) and anomalous samples (right peak). Possible threshold is demonstrated with the vertical line and the consequential classification is indicated with colors and labels (True negative, false negative, false positive, true positive). The ROC curve, which is plotted in lower part, demonstrates all possible thresholds and their probability of true positive and false negative.(Credit: https://en.wikipedia.org/wiki/Receiver_operating_characteristic)

thresholding. Therefore it is tuned as one of the last parameters depending on the known and current data. Anyway, since the thresholds may be different, it is more complicated to define a metric for anomaly detection performance. If there was the only threshold, the percentage of success could be used. [5]

3.3 Receiver operator characteristics and AUC

The performance measuring of the anomaly detection method must take into account all possible thresholds. Receiver operator characteristics (ROC) is utilized to analyze the performance over all thresholds. The graphical representation, which is shown in Fig. 2, is a parametric plot that shows proportion of true positive and false positive rate for all possible thresholds. Note that these proportions are based on the data-set as described in previous paragraph. The curve always starts and finishes in the corners because the lowest threshold classify all samples as positive thus the false and true positive rate is 1. Similarly the highest threshold hits the opposite corner. In an optimal case, the curve is plotted near the third corner that represents high true positive and low false positive rate. On contrary, thresholding an random variable will form the curve as an diagonal. Which means that none method should have the curve under the diagonal. To conclude, it has been shown that the better the method is the higher the curve is plotted which allows us to represent the quality of the method as a scalar that is independent on a specific threshold. This metric is called area under the ROC curve (AUC) and it is often used in anomaly detection. [6] [13] [22] [20]

3.4 Benchmarks

Several different benchmark sets and metrics have been used for the anomaly detection performance evaluation thus the comparison among the anomaly detectors is difficult. However, several benchmark sets are more frequent in the literature than others because they are built for a specific purpose such as intrusion detection or image recognition and are widely used by their community.

KDD-99 [1] is a data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between "bad" connections, called intrusions or attacks, and "good" normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

MNIST [19] is a database of handwritten digits. It has been created as a sample of NIST database and the data have been preprocessed and formatted for easier usage. These data are real world based and widely used for image recognition and many other machine learning branches due to the simplification of MNIST set.

99 DARPA IDEVAL [2] is a data set for intrusion detection. It contains network traffic and audit logs collected on a simulation network in three weeks. The first and third week does not contain any attack contrary to the second week when the network faced various types of attack.

The great advantage of using one of these sets is the comparability of the results among methods. On the other hand, the data sets presented above could be declared as obsolete for the issues in present. In addition to that, the sets are narrowly focused on a specific problem thus they are inappropriate to create a general benchmark for anomaly detection. As a consequence, many authors in the filed of anomaly detection rather constructed their own artificial data because the existing data sets were too different from their problem.

In 2014 Sakurada [27] constructed artificial data from Lorenz system for the purpose of processing the spacecrafts' telemetry data. In 2014 Dau [9] created the data sets by their own from the multi-class problem in the UCI machine learning repository. In 2005 Sarasamma [28] used an expert knowledge of KDD-99 (internet security) to present his method to operate optimal. He selected only the most representative features in advice, predefined several classes of outliers to the model and moreover, modified the data set. However, this could have significantly affect the performance. Such an approach prefers the best results under given conditions (typically used in practise) rather than measure the performance of the proposed method in general.

In 2013 Emmott probably reacted on the situation of missing general comparison data set for anomaly detection and introduced his methodology of creating such sets with using multi-class data set from the UCI repository in [12]. Besides creating a number of carefully selected sets, they also measured performance of 6 popular methods for anomaly detection and demonstrated their score. There is a large number of various multi-class data sets usually based on a real data in the UCI repository hence the constructed sets for anomaly detection are real-based. The performance evaluation could be more efficient and general due to utilizing a number of different sets. This might be a breakthrough in anomaly detection performance measurement if other researchers start to utilize it. In 2014 Dau considered these methodology as the most advanced [9].

4 Proposed experiment

The aim of the experiment is to evaluate the selected state-of-the-art approach with the most advanced benchmark data for anomaly detection because their evaluation is not covered properly with a uniform and well defined data set in the literature (see Sec. 3.4). A similar idea was implemented in [9] but the author did not manage the original set and did not replicate the methodology from the Emmott's work [12].

A feed forward replicator neural network is utilized with several different setups. The number of input and output neurons is equivalent to the dimension of the data set. We use the following approach to find out the near-optimal size of the "bottle neck (see Fig. 1)" :

1. The required variations are predefined. Exactly: 0.7, 0.8, 0.9, 0.95, 0.97 and 0.98.
2. Number of dimensions (neurons) is computed to preserve the variations in the following way:
   (a) PCA is performed and the variation of each component is the matter.
   (b) The components are sorted with respect to the variation.
   (c) The components are excluded consequently from the smallest one until the variance of the rest forms the required proportion.
   (d) The number of the included components is the result.
3. The experiment runs for each number of neurons in the "bottleneck" many times and the results are averaged.
4. The best number of neurones is selected according to the results.

The algorithm above is an heuristic algorithm applicable generally. The best results are expected for the chosen variance. However, the optimal number of neurons can only be found with trial and error method for all possible values. Such an approach is mentioned in the literature and is well applicable if the number of sets is low.

The utilized activation functions is ReLU ($f(x) = \max(0, x)$) and linear ($f(x) = x$). The experiment is performed with autoencoder consisting of 4 layers: Input(ReLU), bottleneck(ReLU), output-hidden(ReLU), output (linear). The anomaly score is computed as the reconstruction error in and the AUC of ROC evaluates the results (See Sec. 3.3).

The evaluation is performed with 29 data sets that were created in accordance to the Emmott methodology proposed in [12]. The utilized datasets represent various problems from the real world and have different properties such as dimension and number of elements. Each data set is composed of the target class (regular data) and anomalous data at four levels of difficulty to detect: easy, medium, hard, very hard which are tested separately and are assumed as separate data sets in the following text. Random sampling is performed such that 75% of the regular data are included to the training and the rest to the validation. The number of sampling iterations is eight and input data are normalized to [0,1].

Evaluation over multiple data sets offers many sophisticated methods that are not described in detail. However the survey is given in Statistical Comparisons of Classifiers. over Multiple Data Sets [11].

The first experiment compares the performance of the basic autoencoder and the PCA with kernel density estimation. Pairwise comparison over multiple data set is carried out with scoring a point for each data set as shown in Tab.1. In other words the comparison counts the number of sets where the method outperforms the other.

Table 1: Performance comparison of basic autoencoder and PCA with kernel density

Winning method	easy	medium	hard	very hard	Sum
Basic autoencoder	14	14	13	7	48
Tie or missing data	1	1	4	8	14
PCA and kernel density estimation	14	14	12	14	54

The second experiment compares the performance among the four selected methods (see Tab. 2). The noise "intesity" was selected from values 0.2, 0.1, 0.05 and 0.01 in order to optimize the performence. The "intensity" represents proportion of corupted features for the pepper noise and variance for the gaussian noise. Friedman ranking is utilized for comparison such that lower rank means better performance. The Table 2 shows that denoising autoencoders outperofrm the PCA and that the gaussian noise is more suitable for the real-based data.

Table 2: Performance comparison among all methods

Method	Friedman rank
Basic autoencoder	2.94
Denoising autoencoder with pepper noise	2.53
Denoising autoencoder with Gaussian noise	2.02
PCA and kernel density estimation	2.51

4.1 Discussion

The results indicate that the Gauss denoising autoencoder have better performance than PCA and other methods on real data in general. An unexpected observation is that the Gaussian noise has a better performance despite that the salt and pepper noise is mainly used in the literature. Possible explanation is that the "salt and pepper" deionising autoencoder is robust to the missing values and that could be the case of their testing data.

The performance comparison of autoencoder over such many sets that are constructed on more difficulty levels has never been done. The statistical significance should be proved to declare that any method is significantly better than other. The performance of the methods in the first proposed experiment is not significantly different according to the Wilcoxon signed-rank test. The same for the second experiment where the Friedman test was performed. The obtained critical value is $Q = 6.1$ but the required value for $\alpha = 0.1$ is $Q = 7.78$.

5 Conclusion

The anomaly detection topic was introduced with a focus on the neural networks and especially the autoencoders, the principle of which is explained in Sec.2.2. The difficulties of evaluation with respect to sensitivity and the state of the benchmark sets in present were discussed in Sec.3.

The performance of four methods for anomaly detecion (PCA based and three types of AE) was compared with using 116 different problems (data sets). The experiment showed that the noise robust autoencoder could outperform PCA. However, the comparison of these methods over multiple data sets, does not proof that any method is better for all sets but only for more sets than any other method. In other words, there might be a number of data sets for which the worst ranked method is the most suitable. Moreover, the tests (Wilcoxon and Friedman) did not prove the significance of the results.

It was discovered that there an universal method has not been Discovered yet (At least among the autoencoders) and the existing have many imperfections such as abilities to detect difficult data, no general key to find out the optimal structure and properties of the neural network etc... Solving that is a future challenge. Especially with respect to the increasing importance of applications on big data with difficult properties, both robust and sensitive methods will be required.

References

[1] KDD Cup 1999 Data. `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`. Accessed: 2017-08-30.

[2] MIT Lincoln Laboratory: DARPA Intrusion Detection Evaluation. `https://ll.mit.edu/ideval/data/1999data.html`. Accessed: 2017-08-30.

[3] E. Aleskerov, B. Freisleben, and B. Rao. *Cardwatch: a neural network based database mining system for credit card fraud detection.* In 'Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)', 220–226, (1997).

[4] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. Technical report, (2015).

[5] J. Beck and E. Shultz. *The use of relative operating characteristic (ROC) curves in test performance evaluation.* Archives of Pathology and Laboratory Medicine **110** (January 1986), 13—20.

[6] A. P. Bradley. *The use of the area under the roc curve in the evaluation of machine learning algorithms.* Pattern Recognition **30** (1997), 1145 – 1159.

[7] J. Cannady. *Artificial neural networks for misuse detection.* In 'National Information Systems Security Conference', 368–81, (1998).

[8] V. Chandola, A. Banerjee, and V. Kumar. *Anomaly detection: A survey.* ACM Computing Surveys (CSUR) **41** (2009), 15.

[9] H. A. Dau, V. Ciesielski, and A. Song. *Anomaly Detection Using Replicator Neural Networks Trained on Examples of One Class*, 311–322. Springer International Publishing, Cham, (2014).

[10] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan. *Neural Network Design.* Martin Hagan, (2014).

[11] J. Demšar. *Statistical comparisons of classifiers over multiple data sets.* J. Mach. Learn. Res. **7** (December 2006), 1–30.

[12] A. F. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. *Systematic construction of anomaly detection benchmarks from real data.* In 'Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description', ODD '13, 16–21, New York, NY, USA, (2013). ACM.

[13] T. Fawcett. *An introduction to ROC analysis.* Pattern Recognition Letters **27** (2006), 861 – 874. ROC Analysis in Pattern Recognition.

[14] A. K. Ghosh, A. Schwartzbard, and M. Schatz. *Learning program behavior profiles for intrusion detection.* In 'Workshop on Intrusion Detection and Network Monitoring', volume 51462, 1–13, (1999).

[15] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, (2016). `http://www.deeplearningbook.org`.

[16] S.-J. Han and S.-B. Cho. *Evolutionary neural networks for anomaly detection based on the behavior of a program.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **36** (2005), 559–570.

[17] E. M. Knorr, R. T. Ng, and V. Tucakov. *Distance-based outliers: algorithms and applications.* The VLDB Journal **8** (Feb 2000), 237–253.

[18] B. Kosko. *Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence/book and disk.* Vol. 1Prentice hall (1992).

[19] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. `http://yann.lecun.com/exdb/mnist/`. Accessed: 2017-08-30.

[20] C. Marrocco, R. Duin, and F. Tortorella. *Maximizing the area under the ROC curve by pairwise feature combination.* Pattern Recognition **41** (2008), 1961 – 1974.

[21] D. Martinus and J. Tax. *One-class classification: Concept-learning in the absence of counterexamples.* PhD thesis, Delft University of Technology, (2001).

[22] C. E. Metz. *Basic principles of ROC analysis.* Seminars in Nuclear Medicine **8** (1978), 283 – 298.

[23] S. Mika, S. Schölkopf, et al. *Kernel PCA and de-noising in feature spaces.* In 'Advances in neural information processing systems', 536–542, (1999).

[24] S. Mukkamala, G. Janoski, and A. Sung. *Intrusion detection using neural networks and support vector machines.* In 'Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on', volume 2, 1702–1707. IEEE, (2002).

[25] T. Pevný. *Loda: Lightweight on-line detector of anomalies.* Machine Learning **102** (2016), 275–304.

[26] J. Ryan, M.-J. Lin, and R. Miikkulainen. *Intrusion detection with neural networks.* In 'Advances in Neural Information Processing Systems', 943–949, (1998).

[27] M. Sakurada and T. Yairi. *Anomaly detection using autoencoders with nonlinear dimensionality reduction.* In 'Proceedings of the MLSDA 2014 2Nd Workshop on Machine Learning for Sensory Data Analysis', MLSDA'14, 4:4–4:11, New York, NY, USA, (2014). ACM.

[28] S. T. Sarasamma, Q. A. Zhu, and J. Huff. *Hierarchical kohonenen net for anomaly detection in network security.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **35** (2005), 302–312.

[29] B. B. Thompson, R. J. Marks, et al. *Implicit learning in autoencoder novelty assessment.* In 'Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on', volume 3, 2878–2883. IEEE, (2002).

[30] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. *Extracting and composing robust features with denoising autoencoders.* In 'Proceedings of the 25th international conference on Machine learning', 1096–1103. ACM, (2008).

[31] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.* Journal of Machine Learning Research **11** (2010), 3371–3408.

[32] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang. *Deep structured energy based models for anomaly detection.* In 'Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48', ICML'16, 1100–1109. JMLR.org, (2016).

[33] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson, and J. Ucles. *Hide: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification.* In 'Proc. IEEE Workshop on Information Assurance and Security', 85–90, (2001).

Podpora distribuovaných výpočetních systémů v knihovně TNL pomocí MPI[*]

Vít Hanousek

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

školitel: Tomáš Oberhuber, Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. This paper presents first part of support of distributed computing systems in the Template Numerical Library (TNL). This library is developed at Department of mathematics at FNSPE. The TNL library uses the Message Passing Interface (MPI) for communication between compute nodes, since it is the most often communication standard on high performance computing clusters. This paper shortly presents a domain decomposition of a regular rectangular mesh and some implementation details which is used in the TNL library. A performance measurement is presented at final section.

Keywords: Cluster, Domain decomposition, MPI, TNL

Abstrakt. V této práci prezentujeme první kroky v přidání podpory distribuovaných výpočetních systémů do knihovny Template Numerical Library (*TNL*), která je aktivně vyvíjena na katedře matematiky na FJFI. Pro komunikaci mezi výpočetními uzly využívá knihovna TNL standarad Message Passing Interface (*MPI*), protože je jedním z nejrozšířenějších způsobů komunikace mezi výpočetními servery na clusterech pro vysoce výkonné počítání. V tomto článku nejdříve představíme použití pravidelných pravoúhlých sítí v TNL a dále se zaměříme na implementaci distribuovaných sítí v knihovně TNL. Závěrem této práce představíme výsledky měření rychlosti synchronizací distribuované sítě.

Klíčová slova: Cluster, Doménová dekompozice, MPI, TNL

1 Úvod

Meassage Passing Interface (MPI) je standard pro komunikaci na clusterech pro vysoce výkonné počítání. Je primárně navržen pro komunikaci mezi servery, ale dá se využít i pro meziprocesovou komunikaci bez jakéhokoli zásahu do aplikace. Tento standard má více implementací, mezi nejznámější patří OpenMPI [2], MPICH [3], Intel MPI [1] a další. Pro testování jsme zvolili knihovnu OpenMPI, ovšem díky standardizaci je možné přeložit knihovnu TNL i s jinou implementací MPI. Mezi základní funkce MPI patří blokující a neblokující zasílání zpráv, dále rozesílání hromadných zpráv a redukce.

Template Numerical Library (TNL)[4] je numerická knihovna vyvíjená na katedře matematiky FJFI a je zaměřená na výpočty na vícejádrových procesorech (CPU) a na

[*] Tato práce vznikla za podpory projektů CERIT Scientific Cloud (LM2015085) a CESNET (LM2015042) financovaných z programu MŠMT Projekty velkých infrastruktur pro VaVaI.

grafických kartách firmy nVidia podporujících technologii CUDA (GPU). Pomocí šablon jsou implementovány základní i pokročilé objekty pro různý hardware, což umožňuje pouhou změnou šablonového parametru změnit hardware, na kterém úloha bude počítána, bez dalších zásahů do kódu. Knihovna TNL podporuje výpočty na strukturovaných pravoúhlých sítích, tak i nestrukturovaných sítích.

Prvním částí knihovny TNL s podporou distribuovaných systémů je podpora dekompozice větších strukturovaných pravoúhlých sítí mezi více výpočetních uzlů. V tomto článku představíme doménovou dekompozici 1D, 2D a 3D sítí. Knihovna TNL s touto podporou bude schopná provádět například výpočty explicitních řešičů na distribuovaných systémech. Jako příklady budeme uvádět 2D síť, implementována byla i 1D a 3D síť.

2 Dekompozice 2D a 3D sítě

Dekompozice sítě mezi více uzlů probíhá následujícím způsobem. Síť rozdělíme na podsítě, které jsou pokud možno stejně velké, a dále tyto lokální sítě zvětšíme o překryv se sousedním výpočetním uzlem. Velikost překryvu volíme dle úlohy. Například řešíme-li Laplaceovu rovnici pomocí explicitního schématu konečných diferencí, pak nám stačí překryv jednoho prvku. Na obrázku 1 je dekompozice 1D sítě a na obrázku 2 je dekompozice 2D sítě. Na obrázku 3 je šipkami naznačena komunikace pro 8-mi okolí. Volba okolí také závisí úloze. Například výše zmíněný diskretizovaný Laplaceův operátor závisí pouze na 4 okolních bodech. Pak je zbytečné v rámci dekompozice sítě uvažovat 8-mi okolí, které bere v úvahu i rohové sousedy. Stejný způsobem lze provést i dekompozici ve 3D. Zde se můžeme bavit o 6-ti okolí, pro sousedství přes stěny, o 18-ti okolí pro sousedství přes hrany a stěny a plné 26 okolí.

Volba okolí také určuje počet navázaných spojení mezi výpočetními uzly, což může mít vliv na rychlost komunikace. Druhý parametr, který má zásadní vliv na rychlost komunikace je množství přenášených dat. Zde mají největší příspěvek hrany pro 2D a stěny pro 3D. Množství přenášených dat závisí na velikosti sítě a na počtu výpočetních uzlů a jejich distribuci. V následujícím příkladu uvažujme 2D síť a 4 okolí. Nechť sít má $n \times m$ prvků a máme N výpočetních uzlů, dále nechť N lze rozložit na součin $i * j$. Pak počet přenášených prvků sítě S je

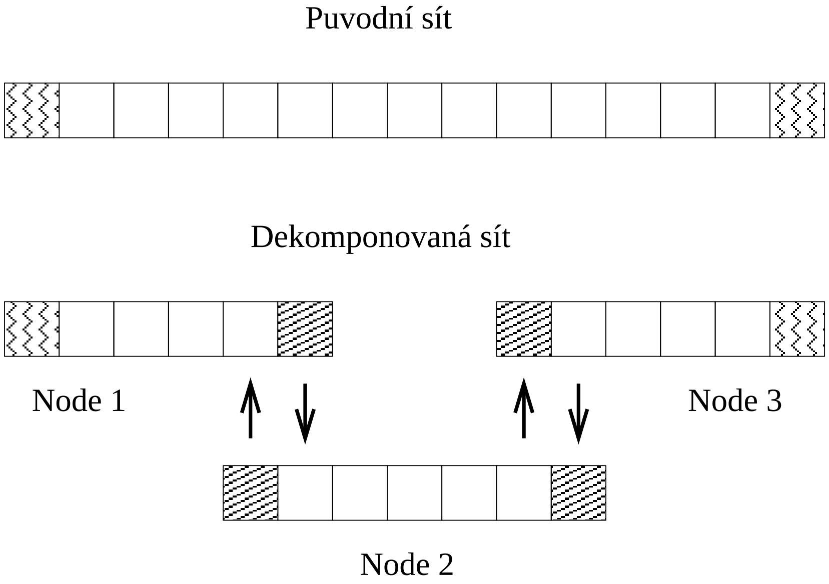
$$S = m(j - 1) + n(i - 1)$$

a počet navázaných spojení je

$$P = (i - 1)j + (j - 1)i$$

Pro lepší představu vlivu distribuce uzlů na tyto parametry uveďme tabulku pro různé distribuce pro síť 100×100 a pro 24 uzlů. Teoretické minimum přenesených dat nastává pro $i = j = \sqrt{N}$, vychází-li celočíselně.

Nakonec této části uveďme, že velké výpočetní clustery mívají kruhové síťové topologie, které je pro tento typ distribuce sítě velmi vhodný. Kruhová síťová technologie má přímé propojení sousedních uzlů. Při správném namapování naší úlohy na cluster mají sousední uzly, ve smyslu dekomponované sítě, přímé propojení a neblokují síťový provoz



Obrázek 1: Dekompozice 1D sítě s 15 prvky mezi 3 výpočetní uzly. Úhlopříčným šrafováním jsou vyznačeny překryvy mezi výpočetními uzly, šipkami je naznačena komunikace mezi nimi a svislým vlnitým šrafováním jsou vyznačeny hraniční prvky sítě.

Distribuce	Počet prvků	Počet spojení
1×24	2300	23
2×12	1200	34
3×8	900	37
4×6	800	38

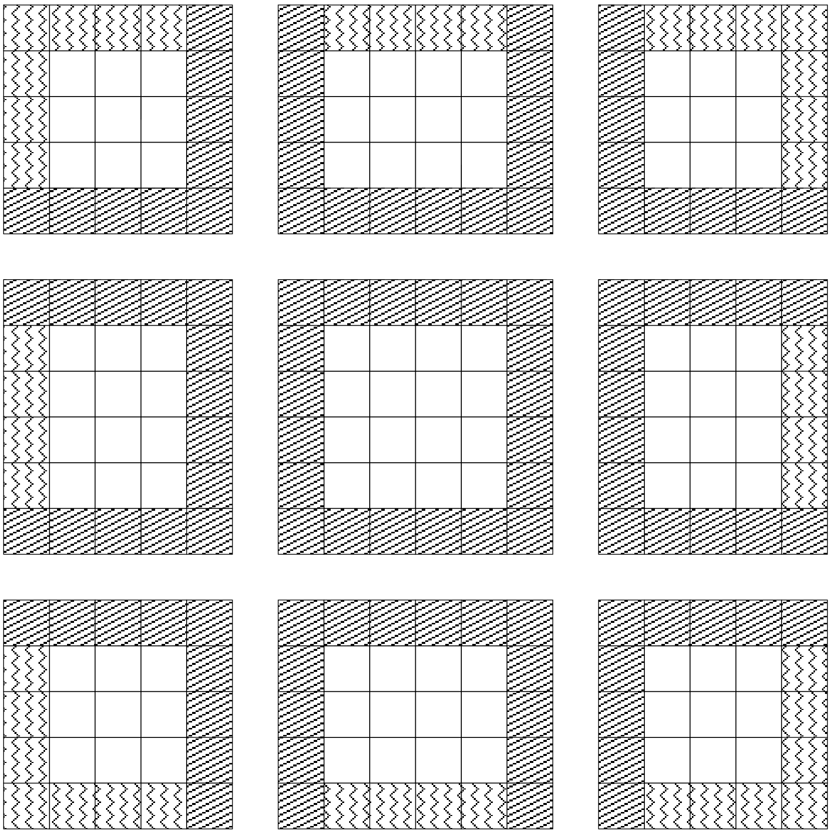
Tabulka 1: Počet přenesených prvků sítě a počet navázaných spojení mezi výpočetními uzly v závislosti na zvolené rozložení 24 výpočetních uzlů do dvojrozměrné mříže. Dekomponovaná síť má 100×100 elementů.

jiné komunikaci, přenosy pak probíhají plně paralelně. Má-li cluster kruhovou síť o nižší dimenzi, než naše síť, pak je výhodné dekomponovat síť právě v dimenzi kruhové sítě.

3 Distribuovaný Grid v TNL

Nejdříve se podíváme jak je strukturovaná pravoúhlá síť v TNL implementována. Tato síť je reprezentována šablonovou třídou *Grid.* Třída gridu sama nenese data síťové funkce vyhodnocované na této síti. Pouze popisuje prostorové uspořádání uzlů, buněk či hran, obsahuje souřadnice počátku a prostorový krok. Nad tímto gridem se vytváří síťová funkce, reprezentovaná třídou *MeshFunction.* Tato třída spojuje informace o gridu s pamětí alokovanou pro jednotlivé hodnoty funkce. Ty se ukládají většinou do třídy *Vector.*

Grid poskytuje pro práci s síťovou funkcí tři základní *Traversary.* První z nich vyhodnocuje pouze vnitřní prvky sítě, druhý vyhodnocuje všechny prvky sítě a poslední



Obrázek 2: Dekompozice 2D sítě s 12 × 12 prvky mezi 9 výpočetních uzlů. Úhlopříčným šrafováním jsou vyznačeny překryvy mezi výpočetními uzly a svislým vlnitým šrafováním jsou vyznačeny hraniční prvky sítě.

vyhodnocuje pouze okrajové prvky sítě, kde prvky mohou být buňky, hrany nebo uzly sítě. Tyto základní traversary jsou využívány třídami operátorů, či jinými třídami pracujícími se síťovou funkcí. Použití pravidelné pravoúhlé v TNL je pak následující:

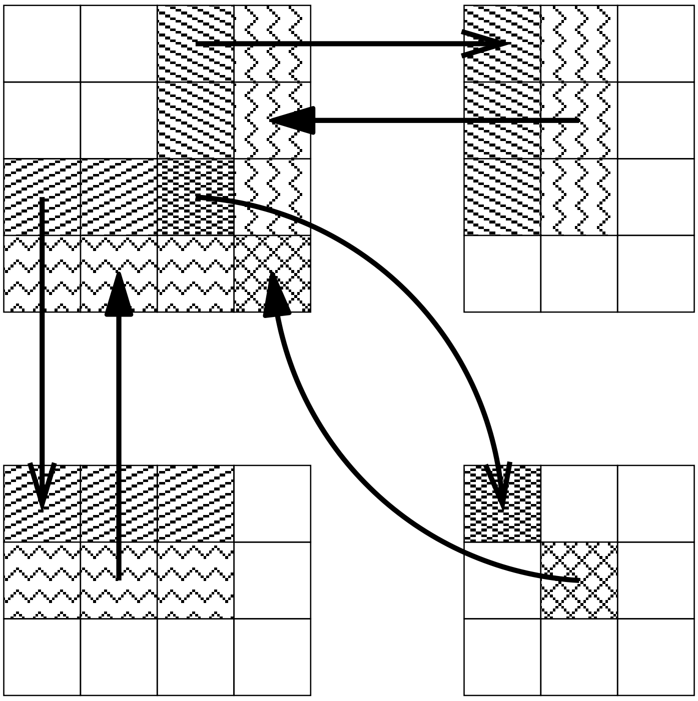
```
typedef MeshType Grid<2,double,Host,int>;

MeshType grid(size);
int dofsize=grid.getEntitiesCount()
Vector<double, Host, int> dof(dofsize);

MeshFunction<MeshType,2,double> meshFunction;
meshFunction.bind(grid,dof)

functionevaluator.evaluateAllEntities(meshFunction,
                                      somefunction);
```

Pro implementaci dekomponované sítě jsme zavedli třídu *DistributedGrid*. Tento objekt



Obrázek 3: Detail komunikace 2D dekomponované sítě. Vyznačené jsou kopírované entity pro výpočetní uzel v levé horní části obrázku. úhlopříčně jsou vyšrafovány odesílané entity tohoto uzlu, vlnitě jsou vyšrafovány prvky přijímané tímto výpočetním uzlem.

nenahrazuje původní grid, pouze uchovává informace o distribuci sítě mezi výpočetními uzly, velikosti lokální sítě, velikosti přesahů a podobně. Distribuovaný grid na každém výpočetním uzlu také předpočítá čísla sousedních výpočetních uzlů všemi směry, pokud existují. Pokud je výpočetní uzel na kraji původní sítě, pak nemá tímto směrem souseda a distribuovaný grid si pro tento směr uloží číslo -1. Díky tomu je snadné a rychlé ve výpočtu určit, zda výpočetní uzel obsahuje daným směrem okrajové entity, či zda má daným směrem přesah. Nakonec distribuovaný grid obsahuje metodu, která nastaví parametry lokální sítě představované původním gridem tak, aby jednotlivé lokální části na sebe navazovali. Lokální grid pak obsahuje pouze přesahy ve směrech kde má daný výpočetní uzel souseda.

Pro správnou funkčnost traversarů přibyla gridu reference na distribuovaný grid. Pokud není nastavena, pak se použijí původní traversary. Pokud je nastavena, vyhodnocují se pouze entity mimo přesahy a hranice se vyhodnocují jen na výpočetních uzlech zpracovávající okraj sítě. Tyto informace získávají traversary právě z objektu distribuovaného gridu.

Po vyčíslení síťové funkce je potřeba doplnit hodnoty síťové funkce v přesazích. K tomuto účelu byl sestaven nástroj *DistributedGridSynchronizer*, který uživateli zakrývá veškerou práci s MPI. Tato třída v konstruktoru podle distribuovaného gridu, který přebírá jako parametr, předpočítá velikosti posílaných dat jednotlivými směry a vytvoří zasílací a přijímací buffery. Po zavolání funkce synchronize, která bere jako parametr třídu síťové funkce, která má být synchronizována, naplní přijímací a odesílací buffery daty z lokální síťové funkce a zajistí komunikaci, pomocí asynchronního zasílání zpráv MPI. Funkce provede zahájení posílání všech zpráv pomocí funkce *MPI_Isend* a zahájení příjmu všech zpráv pomocí funkce *MPI_Irecv* a poté počká na dokončení všech operací pomocí funkce *MPI_Waitall*. Pro funkce *MPI_Isend* a *MPI_Irecv* byly vytvořeny šablonové , které automaticky doplňují parametr *MPI_Type*, dle typu zasílaných dat. Dříve uvedený příklad použití gridu v TNL se při rozšíření na distribuovaný systém změní následujícím způsobem:

```
typedef MeshType Grid<2,double,Host,int>;

MeshType globalGrid(size);
DistributedGrid<MeshType,2> distributedGrid(globalGrid);
MeshType localGrid;
distributedGrid.SetupGrid(localGrid);

int dofsize=localgrid.getEntitiesCount()
Vector<double, Host, int> dof(dofsize);

MeshFunction<MeshType,2,double> meshFunction;
meshFunction.bind(localgrid,dof);

functionevaluator.evaluateAllEntities(meshFunction,
                                       somefunction);
distributeGridSynchronizer.Synchronize(distributedGrid,
                                        meshFunction);
```

Nakonec uveďme, že distribuovaný grid v TNL pro rozmístění výpočetních uzlů do 2D či 3D mříže využívá funkci *MPI_Dims_create*. Tato funkce umožňuje uživateli vynutit distribuci uzlů v nějakém směru ručně. Distribuovaný grid tento způsob ovlivnění rozmístění výpočetních uzlů umožňuje pomocí volitelného parametru, který pracuje stejným způsobem. Díky tomu můžeme dosáhnout jednodimenzionální dekompozice 2D sítě. Distribovaný grid i synchronizer podporují plnohodnotná okolí, tedy ve 2D 8-mi okolí, a ve 3D 26-ti okolí. Podpora volby okolí bude přidána později.

4 Měření

Pro testování naší implementace distribuovaného gridu jsme sestavili následující aplikaci. Aplikace vytvoří 2D distribuovaný grid na kterém několikrát vyhodnotí lineární funkci. Po každém vyhodnocení funkce provede synchronizaci síťové funkce. Měříme průměrnou dobu synchronizace, průměrnou dobu vyhodnocení lineární funkce a celkovou dobu běhu programu. Velikost sítě a počet opakování zápisů jsou programu předány jako parametr. Počet výpočetních uzlů je dán parametrem předávaným spouštěcímu programu *mpirun*.

Distribuce	500	1000	2000	4000	8000	16000	32000
(2,2)	0,60	0,42	0,64	0,58	0,76	1,44	2,41
(4,1)	0,40	0,45	0,52	0,83	1,62	2,32	4,48
(1,4)	0,64	0,69	0,50	0,73	0,63	1,02	1,30

Tabulka 2: průměrná doba synchronizace v milisekundách pro různá rozdělení 4 výpočetních serverů pro různě velké sítě. Rozdělení je uvedeno v prvním sloupci ve tvaru uspořádané dvojce počtu uzlů v ose X a v ose Y. Sítě byly čtvercové o hraně uvedené v prvním řádku.

	500	1000	2000	4000	8000	16000	32000
(3,2)	0,47	0,73	0,57	0,84	1,13	1,66	2,95
(6,1)	0,38	0,18	0,28	0,52	1,10	1,85	3,70
(1,6)	0,13	0,14	0,21	0,26	0,59	0,49	0,86

Tabulka 3: průměrná doba synchronizace v milisekundách pro různá rozdělení 6 výpočetních serverů pro různě velké sítě. Rozdělení je uvedeno v prvním sloupci ve tvaru uspořádané dvojce počtu uzlů v ose X a v ose Y. Sítě byly čtvercové o hraně uvedené v prvním řádku.

Celkově byly sestaveny 3 aplikace, první volí rozložení výpočetních uzlů pomocí zmiňované funkce *MPI_Dims_create*, druhá vynucuje rozložení výpočetních uzlů pouze v ose X, a třetí pouze v ose Y.

Důvodem pro porovnání lineárních rozložení výpočetních uzlů v osách X a Y je skutečnost, že data síťové funkce jsou v paměti uloženy v jednorozměrném poli po řádcích. Při rozdělení výpočetních uzlů v ose Y se do posílacích bufferů kopíruje první a poslední řádek, tedy data v paměti uložená za sebou, zatímco při rozdělení výpočetních uzlů v ose X se do posílacích bufferů kopíruje vždy první a poslední prvek každého řádku, tudíž se s pamětí nepracuje efektivně. Jak ukázalo měření má tato skutečnost zásadní vliv na dobu synchronizace při komunikaci po rychlém rozhraní InfiniBand.

Měření byla provedena s 20 zápisovými cykly na sítích o rozměrech 500 × 500, 1000 × 1000, 2000 × 2000, 4000 × 4000, 8000 × 8000, 16000 × 16000 a 32000 × 32000 elementů. Postupně byly všechny tři aplikace spouštěny na 1 až 9 výpočetních uzlech. Výpočetní uzly byly exkluzivně vyhrazeny pouze pro toto měření, ovšem síťové prvky Infinibandu exkluzivně vyhrazeny nebyly, což mohlo ovlivnit měření. Měření na 2 výpočetních uzlech bylo ukončeno chybou, pravděpodobně způsobenou infrastrukturou výpočetního clusteru na kterém byl výpočet spouštěn, proto je ve výsledcích neuvádíme.

Z naměřených dat jsme vybrali následující výsledky. V prvních třech tabulkách jsou uvedeny průměrné časy synchronizace dat pro různá rozdělení výpočetních serverů a různě velké sítě. V tabulce 2 jsou rozdělení čtyř uzlů, v tabulce 3 jsou rozdělení šesti uzlů a v tabulce 4 rozdělení osmi uzlů. Z prezentovaných výsledků je vidět, že rozdělení serverů v ose Y je v synchronizaci nejrychlejší i za cenu více přenášených dat. Z ostatních výsledků, zde neprezentovaných je patrné že lineární rozdělení výpočetních uzlů v ose Y je vždy výhodnější než rozdělení uzlů v ose X.

V tabulce 5 uvádíme porovnání dob synchronizace pro lineární rozdělení výpočetních

	500	1000	2000	4000	8000	16000	32000
(4,2)	0,44	0,48	0,68	0,77	0,94	1,49	2,86
(8,1)	0,66	0,50	0,83	1,07	1,54	2,35	3,78
(1,8)	0,52	0,70	0,71	1,00	0,85	0,87	1,37

Tabulka 4: průměrná doba synchronizace v milisekundách pro různá rozdělení 8 výpočetních serverů pro různě velké sítě. Rozdělení je uvedeno v prvním sloupci ve tvaru uspořádané dvojce počtu uzlů v ose X a v ose Y. Sítě byly čtvercové o hraně uvedené v prvním řádku.

	3	4	5	6	7	8	9
500x500	0,06	0,64	0,09	0,13	0,40	0,52	0,11
1000x1000	0,08	0,69	0,11	0,14	0,40	0,70	0,37
2000x2000	0,15	0,50	0,40	0,21	0,21	0,71	0,43
4000x4000	0,19	0,73	0,21	0,26	0,50	1,00	0,22
8000x8000	0,25	0,63	0,28	0,59	0,57	0,85	0,36
16000x16000	0,35	1,02	0,56	0,49	0,77	0,87	0,51
32000x32000	0,62	1,30	0,68	0,86	0,83	1,37	1,06

Tabulka 5: průměrná doba synchronizace v milisekundách pro různé počty výpočetních uzlů v lineární distribuci v ose Y a různé velikosti sítě. Velikost sítě je uvedena v prvním sloupci, a počty výpočetních uzlů v prvním řádku.

uzlů v ose Y pro různé počty výpočetních uzlů a různě velké sítě. Z výsledků je patrné, že měření bylo ovlivněno vnějšími vlivy, protože průměrná doba synchronizace pro 8 výpočetních uzlů vychází znatelně delší než doba synchronizace pro 9 výpočetních uzlů. Pro porovnání uvádíme také tabulku 6 s průměrnými dobami vyhodnocení lineární funkce na synchronizované síti. Pro největší dvě testované sítě synchronizace představuje méně než 5% celkového času.

Nakonec uveďme standardní porovnání celkové doby běhu aplikace pro různý počtech výpočetních uzlů a různé sítě. Pro porovnání byly zvoleny časy pro lineární rozložení uzlů v ose Y protože většinou byly nejrychlejší. Tabulka 7 uvádí dobu běhu aplikace v závislosti na velikosti sítě a počtu výpočetních uzlů, tabulka 8 uvádí vypočtené urychlení a tabulka 9 uvádí vypočtenou efektivitu. Z naměřených dat je vidět, že i velmi rychlá synchronizace má negativní velký vliv na celkovou efektivitu. Proto bude dále do knihovny TNL přidána podpora pro překrytí výpočtů a synchronizace. Uveďme také, že v celkové době je zahrnuta také úvodní část programu, která má také na celkovou efektivitu vliv.

5 Závěr

V tomto článku byla prezentována implementace dekompozice pravidelné pravoúhlé sítě v knihovně TNL. Implementována byla dekompozice 1D, 2D i 3D sítí, princip synchronizace dekomponované sítě byl vysvětlen na 1D a 2D síti. Pro 2D síť byla sestavena a spuštěna testovací aplikace, která odhalila, že na rychlém rozhraní Infiniband má velký vliv na

	1	3	4	5	6	7	8	9
500x500	0,34	0,75	1,06	1,09	1,09	1,57	1,66	1,80
1000x1000	1,06	1,12	1,24	1,52	1,42	1,72	1,61	1,65
2000x2000	4,62	2,34	2,25	2,22	2,16	2,28	2,14	1,96
4000x4000	16,74	6,78	5,48	4,64	4,31	4,02	3,62	3,95
8000x8000	59,20	23,31	17,89	14,68	13,48	11,87	10,43	10,30
16000x16000	222,94	79,94	60,78	102,50	45,66	41,78	35,94	33,74
32000x32000	899,60	305,72	225,66	182,22	165,01	139,85	121,45	249,77

Tabulka 6: průměrná doba vyčíslení lineární funkce v milisekundách pro různé počty výpočetních uzlů v lineární distribuci v ose Y a různé velikosti sítě. Velikost sítě je uvedena v prvním sloupci, a počty výpočetních uzlů v prvním řádku.

	1	3	4	5	6	7	8	9
500x500	0,008	0,017	0,035	0,024	0,025	0,040	0,044	0,039
1000x1000	0,030	0,026	0,040	0,034	0,032	0,043	0,047	0,042
2000x2000	0,113	0,058	0,061	0,057	0,051	0,053	0,060	0,052
4000x4000	0,413	0,169	0,148	0,115	0,106	0,103	0,104	0,095
8000x8000	1,471	0,570	0,456	0,365	0,340	0,302	0,272	0,260
16000x16000	5,412	1,957	1,500	1,261	1,109	1,015	0,889	0,845
32000x32000	21,237	7,332	5,476	4,453	3,983	3,391	2,977	2,847

Tabulka 7: doba běhu aplikace v sekundách pro různé počty výpočetních uzlů a různě velké sítě.

	3	4	5	6	7	8	9
500x500	0,5	0,2	0,3	0,3	0,2	0,2	0,2
1000x1000	1,1	0,7	0,9	0,9	0,7	0,6	0,7
2000x2000	2,0	1,8	2,0	2,2	2,1	1,9	2,2
4000x4000	2,4	2,8	3,6	3,9	4,0	4,0	4,3
8000x8000	2,6	3,2	4,0	4,3	4,9	5,4	5,7
16000x16000	2,8	3,6	4,3	4,9	5,3	6,1	6,4
32000x32000	2,9	3,9	4,8	5,3	6,3	7,1	7,5

Tabulka 8: urychlení aplikace v sekundách pro různé počty výpočetních uzlů a různě velké sítě.

	3	4	5	6	7	8	9
500x500	16	6	7	6	3	2	2
1000x1000	37	18	17	15	10	8	8
2000x2000	65	46	39	37	30	23	24
4000x4000	81	70	72	65	57	50	48
8000x8000	86	81	81	72	70	68	63
16000x16000	92	90	86	81	76	76	71
32000x32000	97	97	95	89	89	89	83

Tabulka 9: efektivita paralelizace aplikace v procentech pro různé počty výpočetních uzlů a různě velké sítě.

rychlost synchronizace uspořádání kopírovaných prvků sítě v paměti. Z naměřených dat se nejvýhodnější jeví lineární rozdělení uzlů v ose Y. Měření bylo prozatím provedeno na malém počtu výpočetních uzlů, do budoucna bude rozšířeno alespoň na 20 uzlů. Pro větší testy nám prozatím není dostupná infrastruktura.

Mezi další kroky pro dokončení této části patří implementace ukládání dekomponované síťové funkce do souboru, podpora překrytí výpočtu se synchronizací a podpora synchronizace menších okolí. Následovat by měla podpora dekompozice sítě mezi více GPU.

Literatura

[1] Intel co: *Intel MPI Library*
`https://software.intel.com/en-us/intel-mpi-library`,
Online, [29.9.2016]

[2] Open MPI: *OpenMPI Homepage*
`https://www.open-mpi.org/`,
Online, [29.9.2016]

[3] MPICH: *MPICH Homepage*
`https://www.mpich.org/`,
Online, [29.9.2016]

[4] TNL: *Template Numerical Library Homepage*
`http:tnl-project.org/`,
Online, [29.9.2016]

Using CMA-ES for Black-box Tuning of Coupled PID Controllers in Simulations*

Kateřina Henclová

2nd year of PGS, email: katerina.henclova@fjfi.cvut.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, CAS

Abstract. Proportional integral derivative (PID) controllers are important and widely used tools of system control. However, tuning their gains is a laborious task, especially for complex systems with multiple coupled controllers. To minimize the time and effort spent tuning the gains in a simulation software, we propose to formulate the problem as a black-box optimization problem and solve it with an appropriate method.

We introduce two applications of tuning PID controllers in simulations: combustion engines and an AC filter. For each, a befitting objective function is derived and the resulting problem is successfully solved by a variant of CMA-ES. For the first application, the performance of CMA-ES, PSO and SHADE is compared and the winning method's practical applicability is verified on models of real production engines.

Keywords: CMA-ES, black-box optimization, PID controller

Abstrakt. PID (proporční, integrační, derivační) regulátory jsou důležitým a široce používaným nástrojem pro řízení systémů. Ovšem naladit jejich jednotlivé složky může být složité, obzvlášť v případě komplexních systémů s více navzájem se ovlivňujícími regulátory. Cílem této práce je minimalizovat čas a úsilí nutné k nalezení správného naladění regulátorů v simulačním softwaru. Problém formulujeme jako black-box optimalizační úlohu, kterou následně řešíme pomocí vhodné metody.

Zabýváme se dvěmi konkrétními aplikacemi ladění PID regulátorů pomocí simulací: vznětové motory a AC filtr. V obou případech odvodíme vhodné účelové funkce a výslednou úlohu řešíme pokročilou verzí metody CMA-ES. V úloze s motory srovnáváme CMA-ES s PSO a SHADE a užitečnost vítězné metody je ověřena na ladění regulátorů v modelech skutečně používaných motorů.

Klíčová slova: CMA-ES, black-box optimalizace, PID regulátor

1 Introduction

In a controlled system, PID controllers ensure that given quantities remain constant or within given range. For example, in a room with air-conditioning and/or heating and a temperature sensor, a PID controller keeps the temperature at the pre-set 21°C. The principle remains the same for more complex systems such as a running combustion engine

*This work has been supported by grant 260007 Stradi.

or an AC filter, where multiple controllers may be present and affect each other (i.e. be coupled).

An engineer's task is to tune the gains of all the controllers, so that the system's behavior is satisfactory, i.e. all controlled quantities get to and remain at desired levels. For financial and time reasons, this is often done first with the help of simulations before dealing with physical equipment. This work focuses on the use of such simulations and suggests a method that is to aid engineers in their task without the need to analyze the given system. For combustion engine simulations, 1D dynamics simulation software WAVE is used [18]. This part is largely based on the author's preprint paper [10]. For AC filter simulation, Matlab Simulink [16] and PLECS [1] software combination is used.

The need to solve both these problems arose from industrial applications. Presently, manual work makes up a major part of the controller tuning process. This lengthy procedure is based on trial and error and requires a knowledgeable and experienced control engineer. For systems with a single controller (or multiple but decoupled controllers), simple rules of thumb (e.g. Ziegler-Nichols) can be employed. Similar, already-solved problems can also provide a guideline. However, when having a complicated or unique system of coupled controllers, the complexity of the task makes it very difficult to solve even for an experienced control engineer. Moreover, in our application of PID controllers in combustion engine models, other professionals need to tune the controllers as well, creating the need for a simple-to-use, robust tool. We aim to deliver a method that would eliminate or significantly lower the need for manual tuning. It should find a solution within acceptable time and with as little user interaction as possible. When combined with simple tuning rules or educated guess, our method is to use the provided solution approximation as a starting point and quickly find a more refined solution.

The PID tuning problem with either one controller or multiple but decoupled or symmetric controllers can be and has been reformulated as a black-box optimization problem and solved with an appropriate method. Evolutionary algorithms have been used, e.g. genetic algorithm [15], differential evolution (DE) [3, 11], particle swarm optimization (PSO) [4, 5] and many hybrids [11, 14].

The tuning problem with multiple coupled controllers can too be formulated as an optimization problem. However, compared to other research on controller tuning [3, 4, 5, 11, 14, 15], dealing with coupled controllers requires an extra level of complexity. Its multiple objectives can be efficiently combined into one, enabling us to solve the problem with usual, and faster, algorithms.

The time budget poses the greatest limitation. With simulations taking up to several minutes each, we aim for few thousand simulation runs at most. This imposes high expectations upon efficiency of the method used.

Considering properties of the problem and with the support of experimental evidence, we choose to use a variant of the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [8, 6, 9, 13], an evolutionary algorithm founded deep in probability theory. It has proven to be very effective and robust method in the extensive testing of Black-Box Optimization Benchmarking (e.g. [7, 2]), surpassing the above mentioned algorithms and many others (on the relevant sort of problems). Despite its fame in the optimization community and large number of practical applications, it has so far been little used for tuning PID controllers [11, 12, 20] or similar problems.

In this work, we derive fitting objective functions for both problems and show the applicability of CMA-ES. For the combustion engine problem, we compare performance of CMA-ES, PSO and SHADE (Success-History based Adaptative Differential Evolution [19]).

2 Formulation of the problem

2.1 PID controllers in simulations

PID controllers are well known and powerful tools in system control [17]. Their input is the error

$$e(t) = actual(t) - target(t),$$

i.e. the time-dependent difference between the desired target value and the actual value of a quantity (as measured by a sensor or computed by a model). The output control signal that defines the system's subsequent reaction is given as

$$C(t) = Pe(t) + I \int_0^t e(\tau)d\tau + D\,\frac{d}{dt}e(t),$$

where P, I and D are the proportional, integral and derivative gains, respectively.

In both our applications, the controllers' implementation is provided within the simulation software. Having k controllers within a system, each determined by three constant gains P, I and D, there are $3k$ gains to be tuned: $x = (P_1, I_1, D_1,\ \dots\ ,\ P_k, I_k,\ D_k)$. When the controllers' gains are set and the whole simulation is run, it outputs the above-mentioned error functions' $e_i(t) = e_i(x,t),\ i = 1, \dots, n.$ development over time.

It remains to process $e_i(x,t)$ so that the final function value contains all information about the input's quality. We do so in the next sections by defining an objective function $F(x,t)$ that will be minimized (without loss of generality, we always assume that that higher quality inputs have lower function values).

Our goal is to find such vector x that the corresponding controlled quantities converge to the target values (for constant targets) or start mirroring the target value functions (for targets changing in time) and do so as quickly as possible. For practical purposes, the minimizer found need not be unique.

2.2 Objective function for combustion engine simulations

In the case of combustion engine simulations, construction of the objective function is rather straightforward. Figures 1 and 2 show how the simulation output looks like (on a simple testing model with 3 controllers and 3 controlled quantities). The objective function must then reflect that: 1) all controlled quantities must converge to the target values, 2) the convergence should be as fast as possible, 3) larger error in the beginning of the simulation is OK, 4) each controlled quantity uses different units.

Placing more emphasis on errors with greater time, we weight the error function by time and integrate over time interval $[t_0, t]$. Finally, we scale each objective by the inverse of the (constant) target value, so that their numerical values are comparable and do not

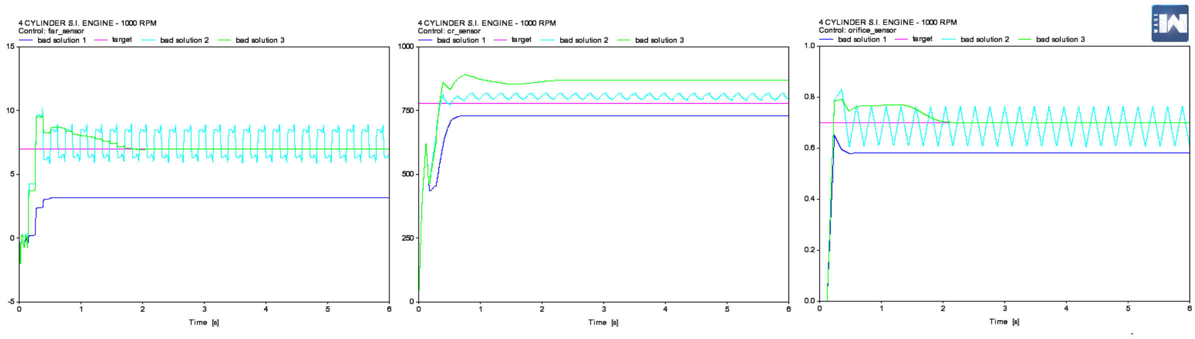


Figure 1: Unsatisfactory solutions: at least one of the controlled quantities does not converge to the target value.

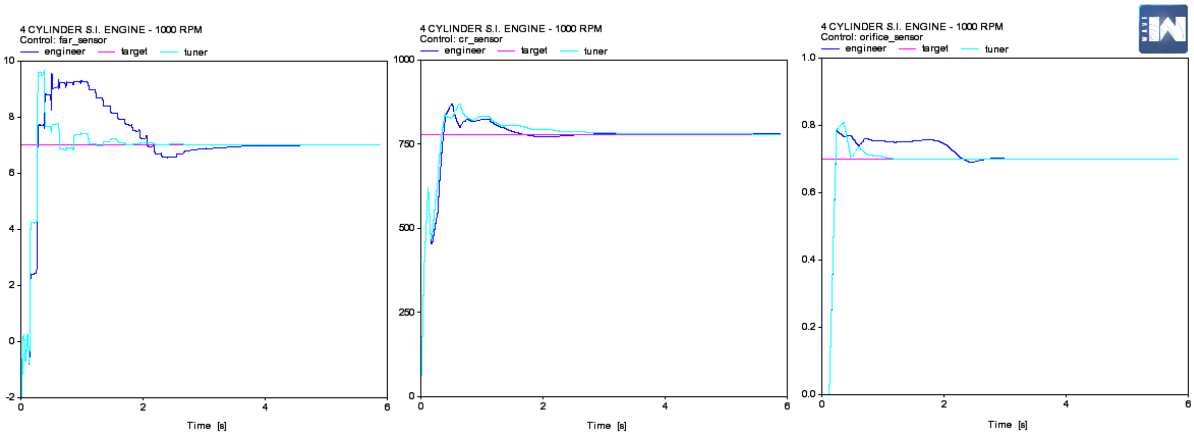


Figure 2: Good solutions: all controlled quantities converge to the target values.

depend on the units of the corresponding quantity. Note that $|target_i|$ is the remainder of the integral over time of the target's (constant) function.

$$F(x,t) \;=\; \sum_{controlled\ quantities} \frac{1}{|target_i|}\int_{t_0}^{t} (\tau+1)|e_i(x,\tau)|\; d\tau$$

Time t corresponds with the end of the simulation. Time $t_0 \geq 0$ is, however, subject to choice. It must be selected manually as a time point just before the output starts to follow a trend. The meaning of $t_0 > 0$ is that it cuts out from the objective function the information that – in this particular application – is essentially noise. Setting $t_0 > 0$ is not neccessary but it can significantly shorten the optimization computation time.

2.3 Objective function for AC filter simulations

For the AC filter, we must take a more general approach. Figure 3 depicts the typical outputs and the prescribed smooth sinusoidal target value functions. The actual value functions tend not to be smooth and overshoot significantly (large overshoot is forbidden due to practical restrictions on not burning the equipment). Moreover, there are 6 sections for each phase (= controlled quantity), the beginning of each being the most troublesome.

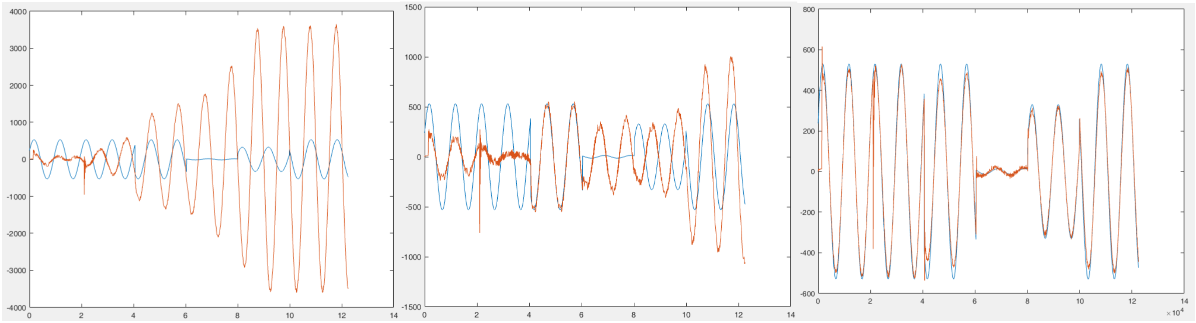


Figure 3: Various outputs of AC filter simulation.

To derive an appropriate objective function $F(x,t)$, the weighted sum over the objectives is used again, this time over all phases and all sections. Each objective is then composed of two parts: scaled L1-norm of the error function (this time we do not use time to weight the error within the integral because the error is most important in the beginning of each section) and L1-norm of the error function's derivative (i.e. its bounded variation), which penalizes non-smooth outputs.

$$F(x,t) = \sum_{phases} \sum_{sections} D_0 \frac{1}{\|target(\tau)\|_{L_1}} \|e(x,\tau)\|_{L_1} + D_1 \left\| \frac{de(x,\tau)}{d\tau} \right\|_{L_1}.$$

Resetting the time counter to 0 at the begining of each phase, the L1-norm is defined as $\|f(\tau)\|_{L_1} = \int_0^t |f(\tau)| d\tau$. Based on typical values of the corresponding L1-norms, the constants were set to $D_0 = 10,\ D_1 = 1e-09$.

3 The optimization method

Clearly, the objective functions will be non-convex, non-differentiable, possibly ill-condititoned, multimodal and must be taken as a black box (since the simulations are such). Metaheuristic and evolutionary methods have been extremely successful when tackling this sort of problems. Based on the results of the extensive Black-Box Optimization Benchmarking [7, 2], the Covariance Matrix Adaptation Evolution Strategy with bi-population restart scheme (BIPOP-CMA-ES) was the method of first choice for our application.

The Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [8] is an evolutionary algorithm that uses stochastic and algebraic tools to define optimally diverse population of candidate solutions in an area that seems to be most promising. The size of the area and its location are determined based on the algorithm's previous experience with the objective function. New candidate solutions are sampled from a multivariate normal distribution, whose mean and covariance matrix are adapted in each generation along with the general step size. For details see Algorithm 1.

There are many upgrades available for basic algorithm. In our application, supported by numerical experiments, we use the elitist BIPOP-aCMA-ES version, i.e. Covariance Matrix Adaptation Evolution Strategy [8] with active covariance matrix updates (including information about detrimental directions [13]), elitist scheme of parent selection (best candidate solutions are parents of new generations until they are superseded [6]) and bi-population restart strategy (method alternanates between 2 regimes with small and large population sizes [9]).

Several important properties of CMA-ES make it so effective in our application. First and foremost, CMA-ES does not use gradients and it does not even presume their existence. Moreover, it does not even use the actual values of the objective function once relative ranking has been assigned to the candidate solutions (except for some stopping/restart criteria). As a result, transformations of the objective function that have no effect upon the relative ranking of individuals do not effect the method's performance, making it more robust. Further, the method exhibits invariance to invertible linear transformations of the search space. In particular, CMA-ES is invariant to scaling of variables (coordinate axes), which is the key property that makes it well-suited for tuning multiple controllers: parameters of one controller are usually of roughly the same scale, but with multiple controllers, the scaling may differ by many orders. A reference point (a vector of typical or expected magnitudes of the controllers' gains) provided by a user then determines how the coordinates are rescaled.

Algorithm 1: Elitist BIPOP-aCMA-ES

```
set $\lambda, \mu$
initialize $m, \sigma, C = I, p_\sigma = 0, p_c = 0$
initialize $restart\_regime = 1, count_1 = 0, count_2 = 0$

while termination criteria not met do
    while restart criteria not met do
        if not first generation in a restart then
            for $i = 1, \dots, \mu$ do
                $x_{i+\mu} = x_i$                      // relabel parents of previous generation
                $f_{i+\mu} = f_i$                      // relabel parents' objective function values

        for $i = 1, \dots, \lambda$ do
            $x_i \sim \mathcal{N}(m, \sigma^2 C)$      // sample new population from normal distribution
            $f_i = evaluate(x_i)$                      // evaluate $x_i$ with objective function

        sort $x_i, i = 1, \dots, \lambda + \mu$ acc. to $f_i$      // assign relative (descending) ranking
        $m^* = m$
        $m = update\_m(x_i, \dots, x_\mu)$                         // move the mean utilizing the parents
        // the evolution paths contain information about past progress
        $p_\sigma = update\_p_\sigma(p_\sigma, \sigma^{-1} C^{-1/2}(m - m^*))$      // isotropic evolution path update
        $p_c = update\_p_c(p_c, \sigma^{-1}(m - m^*), \|p_\sigma\|)$             // anisotropic evolution path update
        $C = update\_C(C, p_c, (x_1 - m^*)/\sigma, \dots, (x_{\lambda+\mu} - m^*)/\sigma)$   // covariance matrix update
        $\sigma = update\_\sigma(\sigma, \|p_\sigma\|)$                          // step size update

        if $restart\_regime = 1$ then
            $count_1 = count_1 + \lambda$
        else
            $count_2 = count_2 + \lambda$

    if $count_1 < count_2$ then
        restart_regime = 1
    else
        restart_regime = 2

    reinitialize parameters and variables acc. to selected restart regime
```

Table 1: Results of 5 CMA-ES runs on real-world models with one (M.1.1, M1.2), two (M2.1, M2.2, M2.3) and three (M3.1) controllers with reference points of various quality. Minimum, maximum and average number of simulation runs is provided.

model	reference p.	min	max	aver.
M1.1	PI baseline	2	68	28
	10^1 PI b.	35	153	79
	10^2 PI b.	95	519	225
	10^{-1} PI b.	20	120	66
	10^{-2} PI b.	49	296	123
M1.2	PI baseline	1	22	9
	10^1 PI b.	4	28	11
	10^2 PI b.	80	225	187
	10^{-1} PI b.	34	100	51
	10^{-2} PI b.	57	181	94

model	reference p.	min	max	aver.
M2.1	PI baseline	11	66	35
	10^1 PI b.	244	280	255
	10^{-1} PI b.	4	32	21
M2.2	PI baseline	8	98	29
	10^1 PI b.	60	770	364
	10^{-1} PI b.	44	107	64
M2.3	PI baseline	9	78	32
	10^1 PI b.	250	757	629
	10^{-1} PI b.	49	1188	347
M2.3	PID baseline	10	91	57
	10^1 PID b.	274	857	522
	10^{-1} PID b.	82	1576	749
M3.1	PID baseline	41	331	152
	10^1 PID b.	827	1763	1268
	10^{-1} PID b.	179	3867	2476

4 Experiments

Extensive experiments were performed using a simplified model of a combustion engine to determine the best setting of CMA-ES for our application and to verify its robustnes. The method's practical usability was tested on models of real engines (see table 1). CMA-ES was also compared to Particle Swarm Optimization (PSO) and Success-History based Adaptative Differential Evolution (SHADE), clearly defeating both (see table 2), especially regarding reliability. For details see [10].

In all cases, we aimed for computation times that are acceptable for engineers using an ordinary PC, i.e. cca 3000 objective function evaluations (= simulation runs) at most. The methods were provided with starting reference points of various quality with "PID baseline" and "PI baseline" (i.e. with D gains set to 0) being the easiest and the other reference points adding orders of magnitude to each of the baseline vector's elements.

The AC filter testing shows very promising preliminary results as well, yet more tests must still be performed.

5 Conclusion

This paper has shown how to construct fitting objective functions for two problems of tuning PID controllers in simulations. It was also shown that CMA-ES can solve the corresponding numerical optimization problem. CMA-ES reaches satisfactory run times and outperforms PSO and SHADE, especially in terms of reliability and robustness.

Table 2: Results of algorithm testing on the basic model. 5 PSO runs, 5 SHADE runs and 10 CMA-ES runs were performed for each of 13 reference points. The value of "-" means that a satisfactory solution (i.e. solution with function value less than 0.5) was not found within the provided budget of 10000 function evaluations. Average run time was not computed if one or more runs did not finish within the given budget.

	PSO runs					SHADE runs					CMA-ES runs		average run time		
reference p.	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5	min	max	PSO	SHADE	CMA-ES
PI baseline	113	61	-	-	-	130	396	341	2100	155	12	168	-	624	76
10^1 PI b.	-	-	-	-	-	-	-	-	-	-	540	2064	-	-	1049
10^2 PI b.	-	-	-	-	-	-	-	-	-	-	821	5700	-	-	2061
10^{-1} PI b.	56	76	48	39	32	342	156	1560	260	538	61	816	50	571	317
10^{-2} PI b.	893	-	483	226	403	5155	1514	2783	1437	2123	222	1334	-	2602	592
10^{-3} PI b.	-	-	-	-	1600	4118	1090	2102	1243	1422	259	1617	-	1995	812
PID baseline	562	28	105	71	28	428	307	497	441	658	32	256	159	466	67
10^1 PID b.	-	-	-	-	-	-	-	-	-	-	62	1462	-	-	1102
10^2 PID b.	-	-	-	-	-	-	-	-	-	-	953	4130	-	-	2022
10^{-1} PID b.	33	164	104	280	43	239	418	209	437	390	141	782	125	339	343
10^{-2} PID b.	387	624	301	1186	-	6524	1018	1150	3297	1556	202	941	-	2709	580
10^{-3} PID b.	-	-	1362	1734	1680	1673	2125	1945	1477	2163	416	2324	-	1877	1138
calibration ref. p.	-	-	-	-	-	-	-	-	-	-	268	2267	-	-	1098

References

[1] J. Allmeling and W. Hammer. *PLECS - piece-wise linear electrical circuit simulation for Simulink.* In 'Proceedings of the IEEE 1999 International Conference on Power Electronics and Drive Systems', volume 1, 355–360, (July 27 1999).

[2] A. Auger, S. Finck, N. Hansen, and R. Ros. BBOB 2010: Comparison Tables of All Algorithms on All Noiseless Functions. Technical Report RT-388, INRIA, (September 2010).

[3] Z. Bingul. *A new pid tuning technique using differential evolution for unstable and integrating processes with time delay.* In 'Neural Information Processing: 11th International Conference, ICONIP 2004, Calcutta, India, November 22-25', 254–260, Berlin, Heidelberg, (2004). Springer Berlin Heidelberg.

[4] Gaing. Z.-L. *A particle swarm optimization approach for optimum design of PID controller in AVR system.* IEEE Transactions on Energy Conversion **19** (June 2004), 384–391.

[5] S. P. Ghoshal. *Optimizations of PID gains by particle swarm optimizations in fuzzy based automatic generation control.* Electric Power Systems Research **72** (2004), 203–212.

[6] N. Hansen, D. Arnold, and A. Auger. *Evolution Strategies.* In 'Springer Handbook of Computational Intelligence', J. Kacprzyk and W. Pedrycz, (eds.), Springer Berlin Heidelberg (2015), chapter 44, 871–898.

[7] N. Hansen, A. Auger, R. Ros, S. Finck, and P. Posik. *Comparing Results of 31 Algorithms from the Black-Box Optimization Benchmarking BBOB-2009.* Workshop Proceedings of the GECCO Genetic and Evolutionary Computation Conference 2010 (2010), 1689–1696.

[8] N. Hansen and A. Ostermeier. *Completely derandomized self-adaptation in evolution strategies.* Evolutionary Computation **9** (2001), 159–195.

[9] Hansen, N. *Benchmarking a BI-population CMA-ES on the BBOB-2009 Function Testbed.* In 'Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers', GECCO '09, 2389–2396, (2009).

[10] K. Henclová. *Using cma-es for tuning coupled pid controllers within models of combustion engines.* Submitted to Engineering Optimization. Preprint version available at arxive.org, (2017).

[11] M. W. Iruthayarajan and S. Baskar. *Evolutionary Algorithms Based Design of Multivariable PID Controller.* Expert Syst. Appl. **36** (July 2009), 9159–9167.

[12] W. M. Iruthayarajan and S. Baskar. *Covariance Matrix Adaptation Evolution Strategy Based Design of Centralized PID Controller.* Expert Syst. Appl. **37** (August 2010), 5775–5781.

[13] G. A. Jastrebski and D. V. Arnold. *Improving evolution strategies through active covariance matrix adaptation.* In 'IEEE Congress on Evolutionary Computation – CEC 2006', 2814–2821, (2006).

[14] W. M. Korani, H. T. Dorrah, and H. M. Emara. *Bacterial foraging oriented by Particle Swarm Optimization strategy for PID tuning.* In 'Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on', 445–450, (Dec 2009).

[15] Kwok, D.P. and Sheng, F. *Genetic algorithm and simulated annealing for optimal robot arm PID control.* In 'Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on', 707–713. IEEE, (1994).

[16] Matlab simulink, (Version 2016b). The MathWorks Inc., Natick, MA, USA.

[17] K. Ogata. *Modern Control Engineering.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, (1990).

[18] Ricardo Software. WAVE Manual, (2016).

[19] R. Tanabe and A. Fukunaga. *Success-history based parameter adaptation for differential evolution.* In 'Evolutionary Computation (CEC), 2013 IEEE Congress on', 71–78. IEEE, (2013).

[20] Y. Wakasa, S. Kanagawa, K. Tanaka, and Y. Nishimura. *Pid controller tuning based on the covariance matrix adaptation evolution strategy.* IEEJ Transactions on Electronics, Information and Systems **130** (2010), 737–742.

Kohonen SOM Learning Strategy and Country Classification*

Radek Hřebík

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

advisors:
Josef Jablonský, Department of Econometrics
Faculty of Informatics and Statistics, University of Economics, Prague
Jaromír Kukal, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The Self-Organized Mapping (SOM) represents a traditional tool for multidimensional data analysis overperforming analytical power of cluster analysis. But there are possible difficulties when the SOM is applied to data patterns of large size. We present testing example using iris dataset. Our approach is mainly used for macro-economical data analysis which is based on logarithmic differences, pattern dimensionality reduction and finalization of data analysis using Kohonen SOM learning. General methodology was applied to main economic indicators describing the situation of thirty five countries during more than twenty years. The used dataset comes from regularly published statistics of European Commission. The main aim is to identify the similarities of countries. The role of SOM topology, learning strategy and reduced pattern size can be also used to predict behaviour during crisis based on the identified similarity and known.

Keywords: SOM, Kohonen learning, iris dataset, artificial neural network, macroeconomic indicators, crisis prediction

Abstrakt. Samoorganizující se mapy (SOM) představují tradiční nástroj pro multidimenzionální analýzu dat, který přesahuje analytickou sílu shlukové analýzy. Pokud se SOM aplikuje na datové vzory velkých rozměrů, vyskytují se problémy. V příspěvku nechybí detailní testovací příklad. Náš přístup se používá hlavně pro makroekonomickou analýzu dat, která je založena na logaritmických diferencích, snížení dimenze a učení pomocí Kohonenových map (SOM). Obecná metodika byla aplikována na hlavní ekonomické ukazatele, které popisují situaci třiceti pěti zemí během více než dvaceti let. Použitá datová sada pochází z pravidelně publikované statistiky Evropské komise. Hlavním cílem je určit podobnosti zemí. Úloha topologie SOM, strategie učení a redukci dimenze lze také použít k predikci chování v průběhu krize, a to na základě zjištěné podobnosti.

Klíčová slova: SOM, Kohonenovo učení, úloha identifikace kosatců, neuronová síť, makroekonomické ukazatele, predikce krize

*This work has been supported by the grants SGS 17/196/OHK4/3T/14 and SGS 17/197/OHK4/3T/14.

1 Introduction

In our research we deal with basic economical indicators which are published on regular basis. The Self-Organized Mapping (SOM) represents a traditional tool for multidimensional data analysis which overperforms analytical power of cluster analysis. We face possible difficulties applying the SOM to data patterns of large size. So we have to make data preprocessing. Our approach of macroeconomic data analysis is based on logarithmic differences, pattern dimensionality reduction and finalization of data analysis using Kohonen SOM learning.

This general methodology was applied to the statistic data describing the economic situation of more than thirty countries during more than twenty years. The regularly published data come from statistics of European Commission. The aim is to identify similar groups of countries and characterize the similarity. The role of SOM topology, learning strategy and reduced pattern size can be also used to crisis prediction based on similarities with countries already suffering with crisis.

2 Kohonen Learning

Kohonen Self Organized Map (SOM) is organized as follows. Let $m, n, H \in \mathbb{N}$ be number of patterns, pattern dimensionality and number of SOM neurons [4]. The individual patterns are $\mathbf{x}_j \in \mathbb{R}^n$ where $j = 1, ..., m$ and form the pattern set $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$. The topology of SOM [8] is described by undirected graph $\mathcal{G}$ of H vertices which are connected with unit length edges. The SOM topology matrix $\mathbb{G} \in \{0,1\}^{H \times H}$ generates mutual vertex distances $\Delta_{i,j}$ for $1 \leq i, j \leq H$. The result of SOM learning is the system of weights [10] $\mathbf{w}_i \in \mathbb{R}^n$ where $, i = 1, ..., H$. We begins with random weights setting $\mathbf{w}_i(0)$. The weights evolve during learning process and their values are denoted as $\mathbf{w}_i(q)$ where $q \in \mathbb{N}_0$.

Kohenen learning rules [7] are very simple. The weight of i-th neuron is changed in q-th step by rule

$$\mathbf{w}_i(q) = \mathbf{w}_i(q-1) + \alpha(q) \cdot c_{i,q} \cdot (\mathbf{x}_q - \mathbf{w}_i(q-1)) \tag{1}$$

for $i = 1, ..., H$, $\mathbf{x}_q \sim \mathrm{U}(\mathcal{S})$ is uniformly selected pattern from $\mathcal{S}$, $c_{i,q}$ is space factor and $\alpha(q) > 0$ is ageing function which is supposed to be non-increasing. The winner is also selected according to Kohonen rule [7] as

$$\varphi_q \in \underset{k=1,\ldots,H}{\arg\min} \|\mathbf{x}_q - \mathbf{w}_k\|_2. \tag{2}$$

We recommend generate the initial weights from the multi-varietal Gaussian distribution as

$$\mathbf{w}_i(0) \sim \mathrm{N}(\mathrm{E}\mathbf{X}, \mathrm{var}\mathbf{X}/100) \tag{3}$$

for $i = 1, ..., H$. The space factor $c_{i,q}$ is calculated using mutual vertex distances as follows. Using learning radius $R_q > 0$ and index of winner vertex ϕ_q, we directly evaluate

$$c_{i,q} = \exp\left(-\frac{\Delta^2_{i,\phi_q}}{2R^2_q}\right)$$

according to Gaussian decay. The final learning strategy consists of $E \in \mathbb{N}$ learning epoch which we characterized by triplets (α_k, R_k, N_k) for $k = 1, ..., E$. Here, α_k is ageing factor, R_k is learning radius, and N_k is number of learning steps in k-th epoch.

3 Quality Measures

The basic way of quality measurement design is based on measuring distances. The Euclidean distance of points $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$ is denoted $\mathrm{d}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$.
Using the pattern $\mathbf{x}_j$ we can investigate the distances to weights $\mathbf{w}_k$ and define winner as

$$\mathrm{win}(j) \in \arg\min_{k=1,...,H} \mathrm{d}(\mathbf{x}_j - \mathbf{w}_k) \tag{4}$$

but the function win(j) is of stochastic nature due to possible distance equities. In some cases we found the winner but one i. e. the second winner which is defined as

$$\mathrm{win2}(j) \in \arg\min_{k \in \mathcal{M}_j} \mathrm{d}(\mathbf{x}_j - \mathbf{w}_k) \tag{5}$$

where $\mathcal{M}_j = \{1, ..., H\} \setminus \{\mathrm{win}(j)\}$.
Using distances and winners we can design traditional measures of various nature.

3.1 Distance penalization

The Quantization Error (QE) is traditionally related to all forms of vector quantization and clustering algorithms [9]. Using linear penalisation we directly penalise the distances between patterns and corresponding winner weights as

$$QE_1 = \sum_{j=1}^{m} \mathrm{d}(\mathbf{x}_j, \mathbf{w}_{\mathrm{win}(j)}). \tag{6}$$

The quadratic penalisation

$$QE_2 = \sum_{j=1}^{m} \mathrm{d}^2(\mathbf{x}_j, \mathbf{w}_{\mathrm{win}(j)}) \tag{7}$$

is also frequently used but has higher sensitivity to outliers.

3.2 Topographic error

General topographic rule is: if two objects are close in reality they must be closed also in the map. Using this principle the Topographic error (TE) [5] is defined as

$$TE = 1 - \frac{1}{m} \sum_{j=1}^{m} g_{\mathrm{win}(j),\mathrm{win2}(j)} \tag{8}$$

where $\mathbb{G} \in \{0, 1\}^{H \times H}$ is SOM topology matrix with $g_{u,v} = \mathrm{I}(\|\mathbf{p}_v - \mathbf{p}_v\|_2 \leq 1)$. The main advantage of TE is in its robustness to outliers. Therefore we use this criterion as main quality measure in this study.

3.3 Correlation based measures

The correlations between mutual distances of patterns and mutual distances of winner weights can be directly used as quality measures.
Let i, j be pattern indexes. The mutual pattern distances can be defined as
$d_{i,j} = \mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)$. The mutual distances of corresponding weights are
$\delta_{i,j} = \mathrm{d}(\mathbf{w}_{win(i)}, \mathbf{w}_{win(j)})$.
Finally, we obtain $m(m-1)/2$ pairs of corresponding distances and directly calculate Pearson correlation coefficient r, Spearmann ρ or Kendall τ coefficient as quality measure.

3.4 Time Complexity of Measures

The evaluations of QE_1, QE_2 and TE are very fast with time complexity $\mathrm{O}(mnH)$. The evaluation of correlation measures is more complex. The Pearson r has time complexity $\mathrm{O}(mnH + m^2)$ due to simple statistics over $m(m-1)/2$ distance pairs. The Spearmann ρ is complicated with pair sorting and its time complexity is $\mathrm{O}(mnH + m^2 \log(m))$. The Kendall τ is not recommended for large pattern sets due to time complexity $\mathrm{O}(mnH+m^4)$.

4 Testing Example

The SOM and its learning as testing example was studied for nineteen neurons placed in 2D space in hexagonal topology with unit neighborhood distances, i.e. $H = 19$, $N = 2$. Artificial two dimensional data were generated in the first case as follows. Total number of 5 000 patterns were generated randomly from seven classes with uniform probability. The center of the first class was placed in the origin. The centres of remaining six classes were placed around in unique distance in the vertices of hexagon. Individual patterns were generated from this Gaussian mixture with standard deviation $\sigma = 0.2$.
Basic quality measures are included in table 1. Resulting weights are depicted in figure 1, meanwhile the density map figure (pattern number in given neuron) and traditional U-map [1]are depicted in figure 2. All neurons and SOM properties were interpolated on convex hull of SOM neurons using cubic interpolation. This convention is useful for weight and density interpretation. As seen the algorithm is able to map the weights proportionally to data coordinates and corresponding contours are approximately uniformly placed parallel lines in figure 1. The density map shows higher central density and six density regions in the network corners meanwhile U-map is approximately constant due to data homogeneity.
Traditional iris flower classification task [2] was originally designed for classifier testing but we apply them for SOM learning with final class density evaluation. Total number of 150 patterns of three classes (Iris setosa, Iris virginica, Iris versicolor) are described by four properties (sepal length, sepal width, petal lenght, petal width). The initial weights, ageing factor and number of learning steps were the same as in previous case. Resulting weight maps are depicted in figure 3 together with class densities and U-map of iris flower problem in figure 4. The SOM learning results can be interpreted using class membership knowledge. As seen in figure 4 the class of Iris setosa is well separated in right corner but remaining two classes are not separable but placed in opposite part of SOM in the left

Measure	Hexagonal Test	Iris Dataset
QE_1	0.2389	0.3121
QE_2	0.2339	0.3450
TE	0.0000	0.0000
p-value of r	0.0953	0.0120
p-value of ρ	0.1054	0.0165
p-value of τ	0.2682	0.1030

Table 1: Quality of SOM learning for hexagonal test

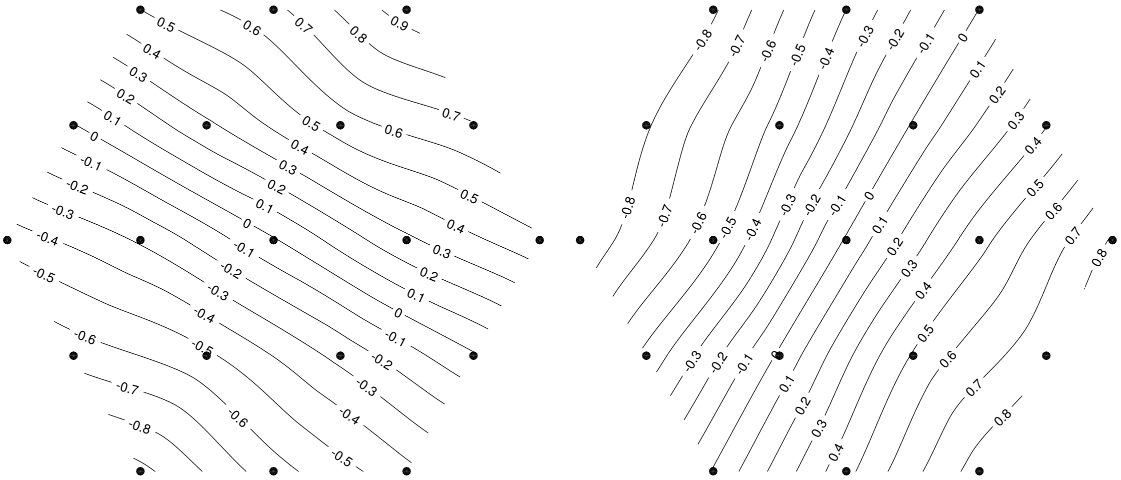


Figure 1: Resulting weights w_1(left) and w_2(right) for hexagonal test

top corner. The remaining part of SOM is not occupied by patterns as also demonstrated as maximal values in U-map.

The subjective evaluation was followed by quality measures evaluation. The results of traditional Graph SOM [4] with Kohonen learning, Gaussian characteristic and $H = 19$ was learned for $E = 9$ with
$\alpha = (0.1, 0.08, 0.07, 0.06, 0.05, 0.04, 0.03, 0.02, 0.01)$,
$R = (5, 3, 3, 1.5, 1, 0.7, 0.5, 0.3, 0.2)$ and $N_k = 1000$. The results are collected in table 1.

5 Case Study: Economical Indicators

As input data we used the main economic indicators. Data has been selected from Statistical Annex of European Economy presented by European Commission in autumn 2016 [3]. As analysis input serve the thirty five countries from the whole world, majority are the European countries. The indicators are observed in years 1993 to 2016. Selected indicators are the total population, unemployment rate, gross domestic product at current market prices, private final consumption expenditure at current prices, gross fixed capital formation at current prices, domestic demand including stocks, exports of goods and services, imports of goods and services and gross national saving. Nine indicators are monitored in total. The main aim of our research is based on data for each country.

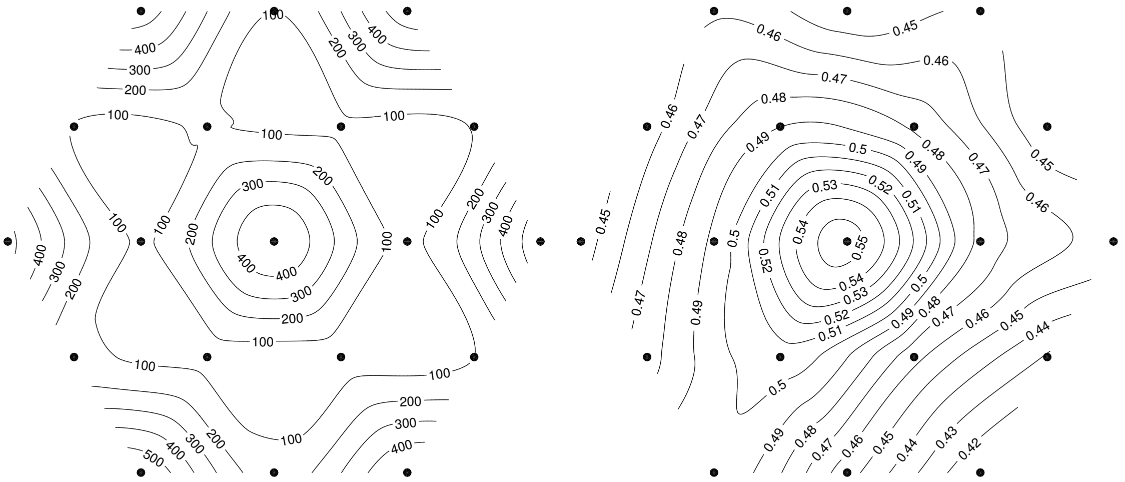


Figure 2: Density map (left) and U-map (right) for hexagonal test

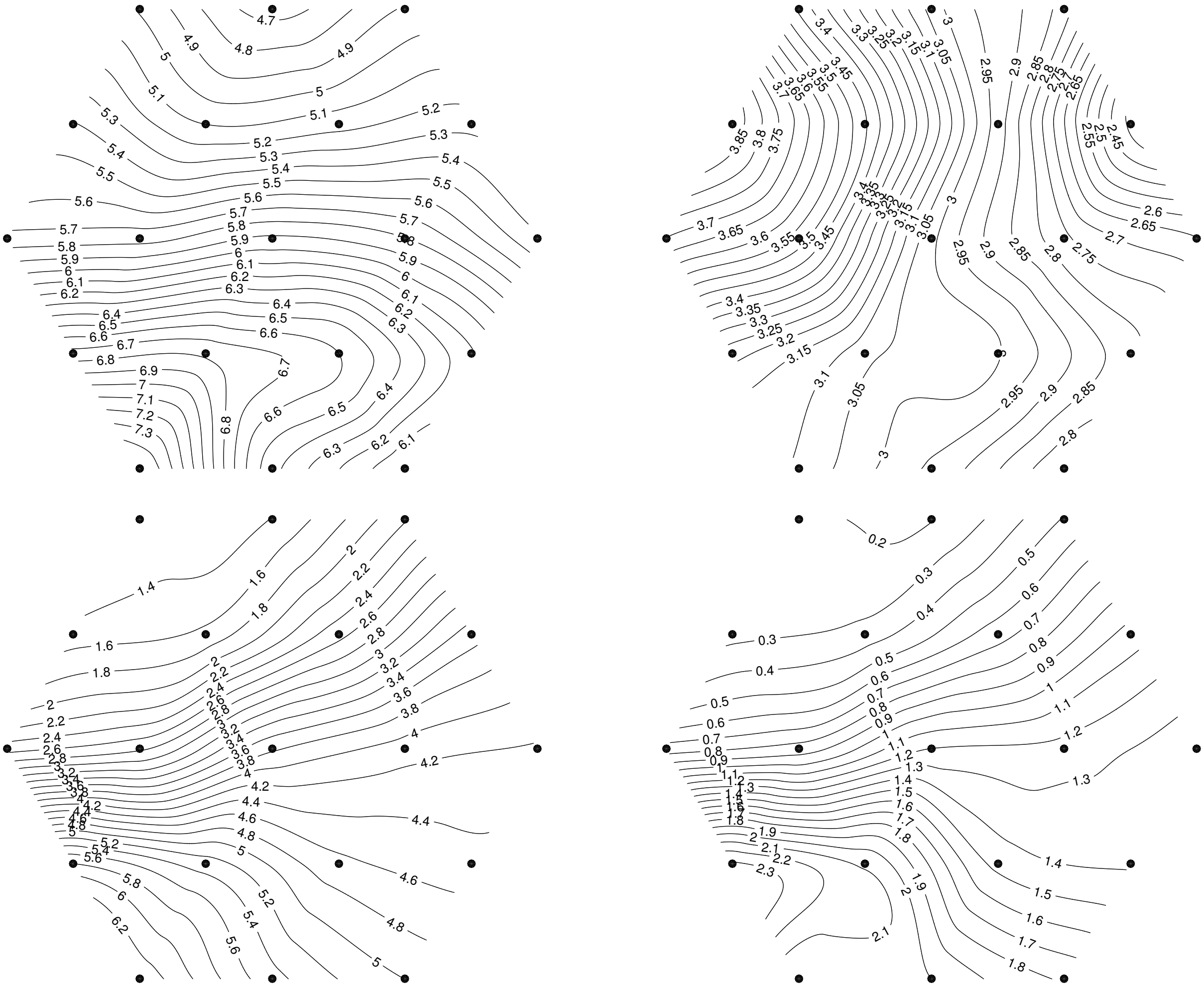


Figure 3: Resulting weights w_1(left top), w_2(right top), w_3(left bottom), w_4(right bottom) for iris flowers

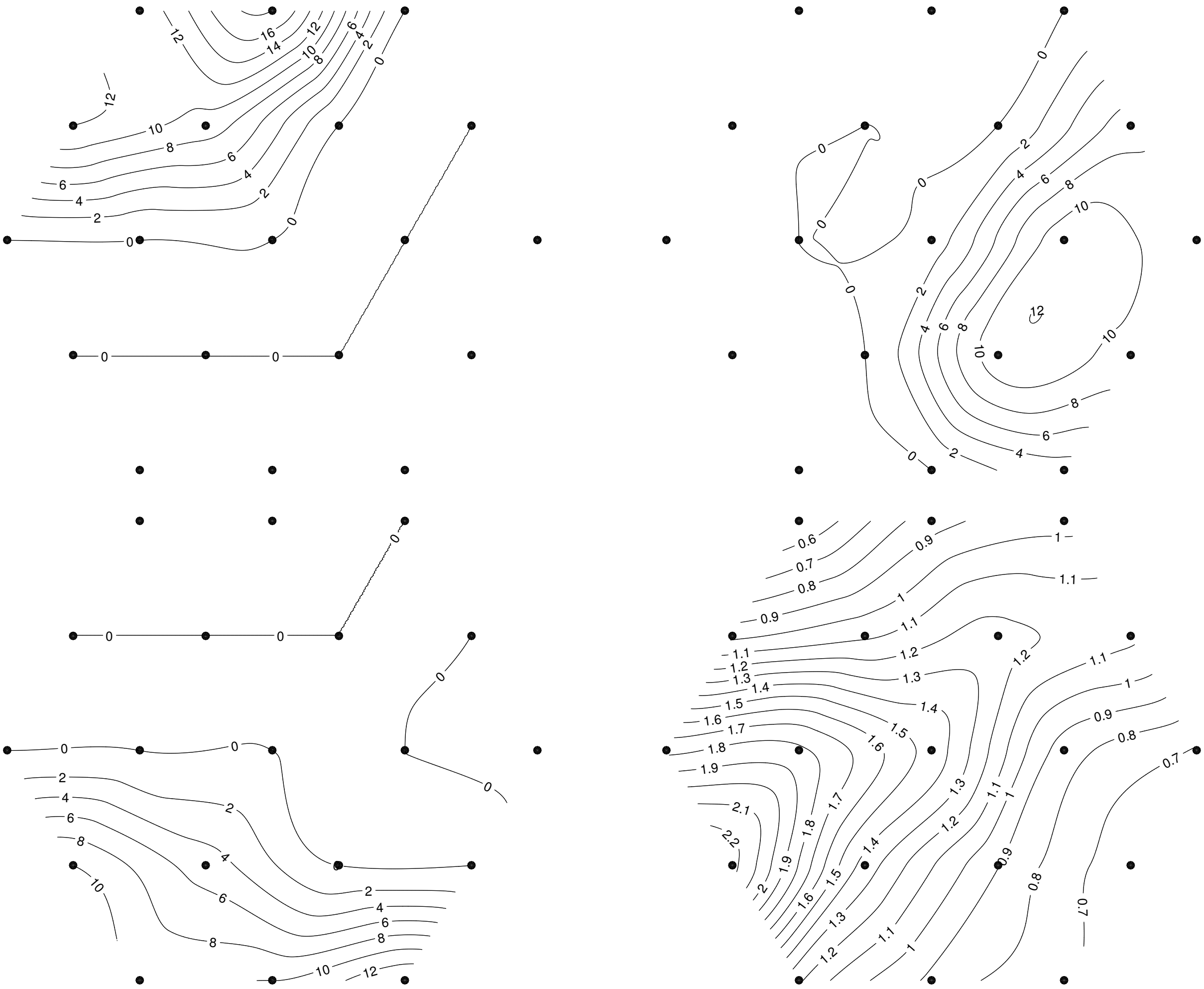


Figure 4: Resulting class densities – setosa (left top), versicolor (right top), virginica (left bottom) and U-map (right bottom) for iris flowers

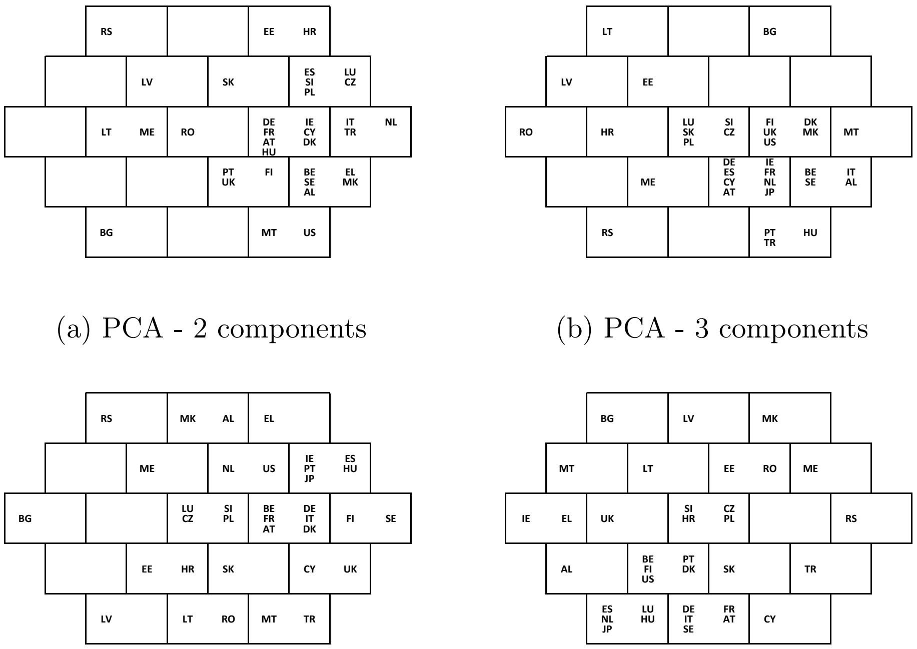
As the dimensionality of input data is quite high, represented by main nine indicators in each year, we use principal component analysis for data dimension reduction. We prefer the standardize variant of PCA which divides the components into square roots of adequate eigenvalues. This approach is frequently called data whitening. The main advantage of the standardization is in identity covariance matrix which generates the components in unified form. We studied data whitening for $D = 2, 3, 4, 5$. Then we applied Kohonen SOM with hexagonal topology with node number $H = 7, 19$. The SOM learning with Gaussian decay was driven by two strategies. For $H = 7$ we used only $E = 2$ with $\alpha = (0.1, 0.05)$, $R = (2, 1)$, $N_k = 1000$. The larger SOM with $H = 19$ was learned for $E = 9$ with $\alpha = (0.1, 0.08, 0.07, 0.06, 0.05, 0.04, 0.03, 0.02, 0.01)$, $R = (5, 3, 3, 1.5, 1, 0.7, 0.5, 0.3, 0.2)$ and $N_k = 1000$. Our aim was to obtain the SOM with zero topographical error (TE) and minimum possible quadratic penalisation (QE_1). The results of QE_1 are captured in table 2.

Table 2: Optimal QE_1 measures

D	SOM_7	SOM_{19}
2	0.002	0.001
3	0.003	0.002
4	0.010	0.007
5	0.020	0.010

6 Results

In all cases we obtained zero values of TE which means that learning was executed well. It is evident from table 2 that SOM_{19} generates results with lower value of QE_1 which is rising with growing dimension. The distribution of countries is captured in figure 5. We see the PCA with 2 components as the best solution and resulting SOM. The different groups of countries were identified. They tell us about the similarities of the concrete countries. The main thing what we can see is the position of Germany, which is usually in the same group as France. In the case of Czech Republic its position depends on number of components but we are in the same group with Poland and Slovenia in all cases. In all cases there are relative compact group of traditional countries which slightly differs each other. The positions of countries with extreme macro-economical behaviours differ with whitening dimensionality. The results are also in accordance to our previous research based on PCA and data whitening [6]. We see some countries which are complicated to be predicted and forms separate groups in each case. This group is represented by Bulgaria and Latvia. The country classification serves also as indicator of upcoming crisis to the closest countries.



(a) PCA - 2 components

(b) PCA - 3 components

(c) PCA - 4 components

(d) PCA - 5 components

Figure 5: Results for $H = 19$ and different number of components

7 Conclusion

Kohonen SOM learning was used to country self-organization in hexagonal SOM topology with whitened log differentiated macroeconomic data. The best result were obtained for $H = 19$ and 2 dimensional whitening with topological error 0% and minimum possible quadratic penalisation. The resulting SOM maps are in agreement with general expectations. From the crisis prediction point of view there is a group of leading European countries (DE, FR, AT, DK, CY, IE), the other European countries with standard economies (UK, ES, IT, IR, BE, NL, LU, CZ, SK, PL, HU) are in the neighbourhood with slightly different response during crisis. The countries with extreme behaviour during crisis (RS, BG, LV, LT, ME, RO) are placed far from the previous groups. The Kohonen SOM is not too sensitive to dimension of data whitening and therefore, the resulting maps only differ in details but save the country similarity property.

References

[1] Soledad Delgado, Consuelo Gonzalo, Estibaliz Martinez, and Agueda Arquero. Visualizing high-dimensional input data with growing self-organizing maps. *Computational and Ambient Intelligence*, pages 580–587, 2007.

[2] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[3] Directorate General for Economic and Financial Affairs (ECFIN). Statistical annex to european economy. autumn 2016. Technical report, European Commission, 2016.

[4] D. Graupe. *Deep Learning Neural Networks: Design and Case Studies*. 2016.

[5] L. Hamel. Som quality measures: An efficient statistical approach. In *Proceedings of the 11th International Workshop WSOM 2016*, pages 49–59, Houston, 2016. Springer.

[6] R. Hrebik and J. Kukal. Multivarietal data whitening of main trends in economic development. In *Mathematical Methods in Economics*, pages 279–284, Plzeň, 2015. University of West Bohemia.

[7] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2012.

[8] E. Oja and S. Kaski. *Kohonen Maps*. Elsevier Science, 1999.

[9] Georg Pölzlbauer. Survey and comparison of quality measures for self-organizing maps.

[10] A. Rettberg, M.C. Zanella, M. Amann, M. Keckeisen, and F.J. Rammig. *Analysis, Architectures and Modelling of Embedded Systems: IESS 2009, Langenargen.* Springer Berlin Heidelberg, 2009.

Discrete Fourier Calculus of Hartley Orbit Functions

Michal Juránek

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

advisors:
Jiří Hrivnák, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Lenka Motlochová, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Ten types of discrete Hartley transforms of Weyl orbit functions are developed. These functions form a generalization of the one-dimensional cas transform. Fundamental domains of even affine and dual even affine Weyl groups, governing the argument and label symmetries of orbit functions, are determined. The discrete orthogonality relations are formulated on finite sets of points from the refinements of the dual weight lattices.

Keywords: Weyl-orbit functions, discrete orthogonality, discrete Fourier transform, Hartley-orbit transform

Abstrakt. Cílem je vývoj deseti typů diskrétních Hartleyovsých transformací Weylových orbitních funkcí. Tyto funkce tvoří zobecnění jednorozměrné transformace *cas*. Určili jsme fundamentální domény sudých afinních a duálních sudých afinních Weylových grup, pomocí kterých se řídí symetrie argumentů a symetrie indexovaní orbitních funkcí. Diskrétní ortogonalita je formulována na konečných souborech bodů na zhuštěné duální váhové mříže.

Klíčová slova: Weylové orbitní funkce, diskrétní ortogonalita, diskrétní Fourierova transformace, Hartleyovská orbitní transformace

1 Introduction

The aim of recent research is to complete and extend the discrete Fourier analysis of Weyl-orbit functions from [10, 8, 11]. The discrete Fourier calculus of all ten types of orbit functions with symmetries inherited from all four types of even Weyl groups is unified in full generality. The real-valued versions of the functions and transforms are also developed by modifying the exponential kernels of orbit functions to their Hartley alternatives [1].

Since introduction of the discrete version of the Hartley transform in [1], both continuous and discrete Hartley transforms form fully equivalent real-valued variants of the standard Fourier transforms. As alternatives to complex Fourier transforms, these transforms together with their 2D and 3D versions found applications in many fields including signal processing [18], pattern recognition, geophysics [17], measurement and optics.

In the context of Weyl-orbit functions and their corresponding transforms, the Hartley transforms have not yet been studied. Replacing exponential kernel as in original 1D Hartley transform yields novel families of real-valued special functions of Weyl groups, which inherit (anti)symmetry properties as well as discrete orthogonality relations from the original Weyl-orbit functions. The resulting generalized Hartley transforms together with the original ten types of Weyl-orbit functions offer, especially in 2D and 3D, richer options and application potential due to greater variability of domain shapes and boundary behaviour.

2 Weyl groups and Crystallographic root systems

Consider the root system Π with its associated Lie algebra of rank n. The notation from [10, 11] is taken. The simple system $\Delta = \{\alpha_1, \cdots, \alpha_n\}$ of the root system Π forms a basis of the Euclidean space $\mathbb{R}^n$, with the symbol $\langle\, ,\, \rangle$ denoting its scalar product. Note that the notions of the root system Π and its inherent set of simple roots Δ are also developed independently on Lie theory. There exist two types simple systems — the first type with roots of only one length, denoted by A_n, $n \geq 1$, D_n, $n \geq 4$, E_6, E_7, E_8, and the second type with two different lengths of roots, denoted by B_n, $n \geq 3$, C_n, $n \geq 2$, G_2 and F_4. The following notation of the standard objects [13], which are induced by the set Δ are:

- the highest root $\xi \in \Pi$
- the marks $m_1, \ldots, m_n \in \mathbb{N}$ of the highest root $\xi = m_1\alpha_1 + \cdots + m_n\alpha_n$ together with $m_0 = 1$,
- the Coxeter number $m = m_0 + m_1 + \cdots + m_n$,
- the root lattice $Q = \mathbb{Z}\alpha_1 + \cdots + \mathbb{Z}\alpha_n$,
- the $\mathbb{Z}$-dual lattice to Q,

  $$P^\vee = \{\omega^\vee \in \mathbb{R}^n \mid \langle \omega^\vee, \alpha\rangle \in \mathbb{Z}, \forall \alpha \in \Delta\} = \mathbb{Z}\omega_1^\vee + \cdots + \mathbb{Z}\omega_n^\vee,$$

  with

  $$\langle \alpha_i, \omega_j^\vee \rangle = \delta_{ij}, \tag{1}$$

- the dual root lattice $Q^\vee = \mathbb{Z}\alpha_1^\vee + \cdots + \mathbb{Z}\alpha_n^\vee$, where

  $$\alpha_i^\vee = \frac{2\alpha_i}{\langle \alpha_i, \alpha_i \rangle}, \quad i \in \{1, \ldots, n\}, \tag{2}$$

- the dual marks $m_1^\vee, \ldots, m_n^\vee$ of the highest dual root $\eta = m_1^\vee\alpha_1^\vee + \cdots + m_n^\vee\alpha_n^\vee$ together with $m_0^\vee = 1$
- the $\mathbb{Z}$-dual lattice to $Q^\vee$

  $$P = \{\omega \in \mathbb{R}^n \mid \langle \omega, \alpha^\vee \rangle \in \mathbb{Z}, \forall \alpha^\vee \in Q^\vee\} = \mathbb{Z}\omega_1 + \cdots + \mathbb{Z}\omega_n,$$

- the Cartan matrix C with elements

$$C_{ij} = \langle \alpha_i,\, \alpha_j^\vee \rangle,$$

- the index of connection c of Π equal to the determinant of the Cartan matrix C,

$$c = \det C. \tag{3}$$

The properties of Weyl groups and affine Weyl groups can be found for example in [13]. The finite Weyl group W is generated by n reflections r_α, $\alpha \in \Delta$, over the hyperplane defined by the normal vector α.

$$r_{\alpha_i} a \equiv r_i a = a - \frac{2\langle a,\, \alpha_i\rangle}{\langle \alpha_i,\, \alpha_i\rangle}\alpha_i\,, \qquad a \in \mathbb{R}^n\,.$$

The infinite affine Weyl group W^{aff} is the semidirect product of the Abelian group of translations $Q^\vee$ and of the Weyl group W

$$W^{\mathrm{aff}} = Q^\vee \rtimes W. \tag{4}$$

Let ψ denote the retraction homomorphism $\psi : W^{\mathrm{aff}} \to W$ of the semidirect product. The fundamental region $F \subset \mathbb{R}^n$ of W^{aff} can be chosen as the convex hull of the points $\left\{0, \frac{\omega_1^\vee}{m_1}, \dots, \frac{\omega_n^\vee}{m_n}\right\}$.

Alternatively, W^{aff} is a Coxeter group generated by n reflections r_i and an affine reflection r_0 given as

$$r_0 a = r_\xi a + \frac{2\xi}{\langle \xi,\, \xi\rangle}\,, \qquad r_\xi a = a - \frac{2\langle a,\, \xi\rangle}{\langle \xi,\, \xi\rangle}\xi\,, \qquad a \in \mathbb{R}^n\,.$$

The set of n reflections r_i together with the affine reflection r_0 is denoted by

$$R = \{r_0, r_1, \dots, r_n\}.$$

2.1 Sign homomorphisms

Any homomorphism σ from W to the multiplicative group $\{1, -1\}$ is called a sign homomorphism [8]. Two standard choices of sign homomorphisms are the trivial homomorphism and the determinant denoted as

$$\begin{aligned} \sigma^e(w) &= \det(w), \\ \mathbb{1}(w) &= 1. \end{aligned}$$

The sign homomorphisms σ^l and σ^s are defined on the set of generators $\{r_\alpha \mid \alpha \in \Delta\}$ of W as

$$\sigma^s(r_\alpha) = \begin{cases} -1 & \text{if } \alpha \in \Delta_s, \\ 1 & \text{otherwise,} \end{cases}$$

$$\sigma^l(r_\alpha) = \begin{cases} -1 & \text{if } \alpha \in \Delta_l, \\ 1 & \text{otherwise.} \end{cases}$$

The set of sign homomorphisms Σ of a root system Δ with two different lengths of roots contains only four elements [8], i.e.

$$\Sigma = \{\mathbf{1},\, \sigma^e,\, \sigma^s,\, \sigma^l\}.$$

The set of 'negative' generators from R with respect to the sign homomorphism σ is denoted by R^σ,

$$R^\sigma = \{r \in R \mid \sigma \circ \psi\,(r) = -1\}\,. \tag{5}$$

The set $F^\sigma \subset F$ is given by

$$F^\sigma = \left\{a \in F \,\middle|\, \sigma \circ \psi\,(\mathrm{Stab}_{W^{\mathrm{aff}}}(a)) = \{1\}\right\}. \tag{6}$$

3 Affine even Weyl groups

3.1 Fundamental domains

Kernels of the non-trivial sign homomorphisms of a given Weyl group W form normal subgroups $W^\sigma \subset W$ known as even Weyl groups [16],

$$W^\sigma \equiv \{w \in W \mid \sigma(w) = 1\}\,.$$

The corresponding affine even Weyl groups are the kernels of the expanded sign homomorphisms $\sigma \circ \psi$

$$W^{\mathrm{aff}}_\sigma \equiv \left\{w^{\mathrm{aff}} \in W^{\mathrm{aff}} \mid \sigma \circ \psi(w^{\mathrm{aff}}) = 1\right\}.$$

For any $r_\sigma \in R^\sigma$, the set $F \cup r_\sigma F^\sigma$ is a fundamental domain of W^{aff}_σ [9].

Generalizing relation (6), the set $F^{\widetilde{\sigma},\sigma}$ is given as

$$F^{\widetilde{\sigma},\sigma} = \left\{a \in F \cup r_\sigma F^\sigma \mid \widetilde{\sigma} \circ \psi(\mathrm{Stab}_{W^{\mathrm{aff}}_\sigma}(a)) = \{1\}\right\}. \tag{7}$$

Note also that for the fundamental domain $F \cup r_\sigma F^\sigma$ it holds that

$$F \cup r_\sigma F^\sigma = F^{\mathbf{1},\sigma}. \tag{8}$$

3.2 Dual affine Weyl group and its even subgroups

The dual affine Weyl group $\widehat{W}^{\mathrm{aff}}$ is a semidirect product of the group of shifts from the root lattice Q and the Weyl group W,

$$\widehat{W}^{\mathrm{aff}} = Q \rtimes W. \tag{9}$$

Let $\widehat{\psi}$ denote the dual retraction homomorphism $\widehat{\psi} : \widehat{W}^{\mathrm{aff}} \to W$ of the semidirect product.

Equivalently, the dual affine Weyl group $\widehat{W}^{\mathrm{aff}}$ is generated by reflections r_i and the reflection $r_0^\vee$ given by

$$r_0^\vee a = r_\eta a + \frac{2\eta}{\langle \eta,\, \eta\rangle}, \quad r_\eta a = a - \frac{2\langle a,\, \eta\rangle}{\langle \eta,\, \eta\rangle}\eta, \quad a \in \mathbb{R}^n.$$

The set of generators of $\widehat{W}^{\mathrm{aff}}$ is denoted by $R^\vee$,

$$R^\vee = \{r_0^\vee, r_1, \ldots, r_n\}.$$

Similarly to the fundamental domain F, the fundamental region $F^\vee$ of $\widehat{W}^{\mathrm{aff}}$ is the convex hull of the vertices $\left\{0, \frac{\omega_1}{m_1^\vee}, \ldots, \frac{\omega_n}{m_n^\vee}\right\}$.

The corresponding dual affine even Weyl groups are the kernels of the expanded sign homomorphisms $\sigma \circ \widehat{\psi}$

$$\widehat{W}_\sigma^{\mathrm{aff}} = \left\{\widehat{w}^{\mathrm{aff}} \in \widehat{W}^{\mathrm{aff}} \mid \sigma \circ \widehat{\psi}(\widehat{w}^{\mathrm{aff}}) = 1\right\}.$$

The set of generators of the affine Weyl group $\widehat{W}^{\mathrm{aff}}$ with negative values of the sign homomorphisms $\sigma \circ \widehat{\psi}$ is denoted by $R^{\vee\sigma}$,

$$R^{\vee\sigma} = \left\{r \in R^\vee \mid \sigma \circ \widehat{\psi}(r) = -1\right\}.$$

Similarly to (6) the domain $F^{\vee\sigma}$ is given by,

$$F^{\vee\sigma} = \left\{b \in F^\vee \mid \sigma \circ \widehat{\psi}(\mathrm{Stab}_{\widehat{W}^{\mathrm{aff}}}(b)) = \{1\}\right\}.$$

The fundamental domains of the dual even affine Weyl groups $\widehat{W}_\sigma^{\mathrm{aff}}$ are determined analogously. The set $F^\vee \cup r_\sigma F^{\vee\sigma}$ is for any $r_\sigma \in R^{\vee\sigma}$ a fundamental domain of $\widehat{W}_\sigma^{\mathrm{aff}}$. The dual analogue of $F^{\widetilde{\sigma},\sigma}$ is given as

$$F^{\vee\widetilde{\sigma},\sigma} = \left\{b \in F^\vee \cup r_\sigma^\vee F^{\vee\sigma} \mid \widetilde{\sigma} \circ \widehat{\psi}(\mathrm{Stab}_{\widehat{W}_\sigma^{\mathrm{aff}}}(b)) = \{1\}\right\}. \tag{10}$$

3.3 Orthogonality coefficients and weights

This section defines the coefficients ($h_M^{\vee\sigma}$) and weights (ϵ^σ) necessary in the discrete orthogonality of Weyl-orbit functions and Hartley kernel orbit functions.

The isotropy subgroups of W_σ^{aff} and their orders are for any $a \in \mathbb{R}^n$ denoted by

$$\mathrm{Stab}_{W_\sigma^{\mathrm{aff}}}(a) = \left\{w_\sigma^{\mathrm{aff}} \in W_\sigma^{\mathrm{aff}} \mid w_\sigma^{\mathrm{aff}} a = a\right\}, \quad h^\sigma(a) = |\mathrm{Stab}_{W_\sigma^{\mathrm{aff}}}(a)|.$$

Related functions $\epsilon^\sigma : \mathbb{R}^n \to \mathbb{N}$ are defined by the relation

$$\epsilon^\sigma(a) = \frac{|W^\sigma|}{h^\sigma(a)}. \tag{11}$$

Since for any $w_\sigma^{\mathrm{aff}} \in W_\sigma^{\mathrm{aff}}$ are the stabilizers $\mathrm{Stab}_{W_\sigma^{\mathrm{aff}}}(a)$ and $\mathrm{Stab}_{W_\sigma^{\mathrm{aff}}}(w^{\mathrm{aff}} a)$ conjugated, it holds that

$$\epsilon^\sigma(a) = \epsilon^\sigma(w_\sigma^{\mathrm{aff}} a), \quad w_\sigma^{\mathrm{aff}} \in W_\sigma^{\mathrm{aff}}. \tag{12}$$

The calculation procedure of the coefficients $h^1(a)$ is detailed in §3.7 in [10]. Having calculated the values of $h^1(a)$ from this procedure the remaining values $h^\sigma(a)$ for any $a \in F$ are calculated thusly

$$h^\sigma(a) = \begin{cases} h^1(a) & \text{if } a \in F^\sigma, \\ \frac{1}{2} h^1(a) & \text{otherwise.} \end{cases} \tag{13}$$

The last step is to extend the values of $h^{\sigma}(a)$, $a \in F$ to the entire fundamental domain $F^{1,\sigma}$ of $W^{\mathrm{aff}}_{\sigma}$ via the following relation

$$h^{\sigma}(r_{\sigma}a) = h^{\sigma}(a).$$

Finally, the coefficients $\epsilon^{\sigma}(a)$, $a \in F^{1,\sigma}$ are determined from $h^{\sigma}(a)$ by equation (11).

The dual versions are developed analogously. The isotropy subgroups of $\widehat{W}^{\mathrm{aff}}_{\sigma}$ are for any $b \in \mathbb{R}^n$ denoted by

$$\mathrm{Stab}_{\widehat{W}^{\mathrm{aff}}_{\sigma}}(b) = \left\{ w^{\mathrm{aff}}_{\sigma} \in \widehat{W}^{\mathrm{aff}}_{\sigma} \mid w^{\mathrm{aff}}_{\sigma} b = b \right\}.$$

Consider the discretization factor $M \in \mathbb{N}$, defining the density of the discretization procedure. The orders of the stabilizers $\mathrm{Stab}_{\widehat{W}^{\mathrm{aff}}_{\sigma}}(b/M)$, are denoted by

$$h^{\vee\sigma}_{M}(b) = \left| \mathrm{Stab}_{\widehat{W}^{\mathrm{aff}}_{\sigma}} \left(\frac{b}{M} \right) \right|. \tag{14}$$

The calculation procedure of the coefficients $h^{\vee 1}_{M}(a)$ is detailed in §3.7 in [10]. Having calculated from this procedure the values of $h^{\vee 1}_{M}(b)$ the following relation allows to determine $h^{\vee\sigma}_{M}(b)$ for any $b \in MF^{\vee}$ as

$$h^{\vee\sigma}_{M}(b) = \begin{cases} h^{\vee 1}_{M}(b) & \text{if } b/M \in F^{\vee\sigma}, \\ \frac{1}{2} h^{\vee 1}_{M}(b) & \text{otherwise.} \end{cases}$$

The last step is to extend the values of $h^{\vee\sigma}_{M}(b)$, $b \in MF^{\vee}$ to the entire magnified fundamental domain $MF^{\vee 1,\sigma}$ of $\widehat{W}^{\mathrm{aff}}_{\sigma}$ via the following relation

$$h^{\vee\sigma}_{M}(r_{\sigma}b) = h^{\vee\sigma}_{M}(b).$$

4 Orbit functions

Consider a sign homomorphism $\sigma \in \Sigma$ and the corresponding even subgroup $W^{\sigma} \subset W$. Taking another sign homomorphism $\widetilde{\sigma} \in \Sigma$ and a parameter $b \in \mathbb{R}^n$, the most general form of Weyl-orbit functions $\Psi^{\widetilde{\sigma},\sigma}_{b} : \mathbb{R}^n \to \mathbb{C}$ is introduced as

$$\Psi^{\widetilde{\sigma},\sigma}_{b}(a) = \sum_{w \in W^{\sigma}} \widetilde{\sigma}(w) \mathrm{e}^{2\pi \mathrm{i} \langle wb, a \rangle}. \tag{15}$$

This general definition leads to three types of orbit functions for root systems with one root-length and to ten types of orbit functions for root systems with two root-lengths [9].

The real-valued modification of orbit functions which for $a \in \mathbb{R}$ uses the Hartley kernel

$$\mathrm{cas}(a) = \cos(a) + \sin(a)$$

instead of exponential kernel. Fixing an even subgroup $W^{\sigma} \subset W$, an additional sign homomorphism $\widetilde{\sigma} \in \Sigma$ and a parameter $b \in \mathbb{R}^n$, the Hartley orbit functions $\zeta^{\widetilde{\sigma},\sigma}_{b} : \mathbb{R}^n \to \mathbb{R}$ are defined via relation

$$\zeta^{\widetilde{\sigma},\sigma}_{b}(a) = \sum_{w \in W^{\sigma}} \widetilde{\sigma}(w) \,\mathrm{cas}(2\pi \langle wb, a \rangle). \tag{16}$$

Similarly to (15), such definition leads to three types of real-valued orbit functions for root systems with one root-length and to ten types of orbit functions for root systems with two root-lengths. Note that the relation of exponential function to the cas function implies

$$\zeta_b^{\widetilde{\sigma},\sigma} = \operatorname{Re}\Psi_b^{\widetilde{\sigma},\sigma} + \operatorname{Im}\Psi_b^{\widetilde{\sigma},\sigma}. \tag{17}$$

This property immediately allows to replicate the argument-label symmetries formulated in [8].

Let $b \in P$, then for any $w^{\mathrm{aff}} \in W_\sigma^{\mathrm{aff}}$ and any $a \in \mathbb{R}^n$ it holds that

$$\zeta_b^{\widetilde{\sigma},\sigma}(w^{\mathrm{aff}}a) = \widetilde{\sigma}\circ\psi(w^{\mathrm{aff}})\cdot\zeta_b^{\widetilde{\sigma},\sigma}(a). \tag{18}$$

Let $a \in \frac{1}{M}P^\vee$, then for any $\widehat{w}^{\mathrm{aff}} \in \widehat{W}_\sigma^{\mathrm{aff}}$ and any $b \in \mathbb{R}^n$ it holds that

$$\zeta^{\widetilde{\sigma},\sigma}_{M\widehat{w}^{\mathrm{aff}}\left(\frac{b}{M}\right)}(a) = \widetilde{\sigma}\circ\widehat{\psi}(\widehat{w}^{\mathrm{aff}})\cdot\zeta_b^{\widetilde{\sigma},\sigma}(a).$$

5 Discretization of orbit functions

Following the standard choice in Fourier analysis, only discrete values of labels of orbit functions $b \in P$ are considered. For any resolution factor $M \in \mathbb{N}$, the discrete Fourier calculus of orbit functions is developed on the set of points $F_M^{\widetilde{\sigma},\sigma}$ defined as

$$F_M^{\widetilde{\sigma},\sigma} = \frac{1}{M}P^\vee \cap F^{\widetilde{\sigma},\sigma}.$$

The sets of labels $\Lambda_M^{\widetilde{\sigma},\sigma}$ are defined as

$$\Lambda_M^{\widetilde{\sigma},\sigma} = P \cap MF^{\vee\widetilde{\sigma},\sigma}. \tag{19}$$

For any $\widetilde{\sigma},\sigma \in \Sigma$ and $M \in \mathbb{N}$ it holds, for the numbers of elements of the sets $F_M^{\widetilde{\sigma},\sigma}$ and $\Lambda_M^{\widetilde{\sigma},\sigma}$, that

$$\left|F_M^{\widetilde{\sigma},\sigma}\right| = \left|\Lambda_M^{\widetilde{\sigma},\sigma}\right|. \tag{20}$$

5.1 Discrete orthogonality of orbit functions

The discrete orthogonality relations of the discretized functions $\Psi_b^{\widetilde{\sigma},\sigma}$, $b \in \Lambda_M^{\widetilde{\sigma},\sigma}$ on the finite point sets $F_M^{\widetilde{\sigma},\sigma}$ have the following formulation [9]. For any $\sigma,\widetilde{\sigma} \in \Sigma$ and any $b,b' \in \Lambda_M^{\widetilde{\sigma},\sigma}$ it holds that

$$\left\langle \Psi_b^{\widetilde{\sigma},\sigma}, \Psi_{b'}^{\widetilde{\sigma},\sigma}\right\rangle_{F_M^{\widetilde{\sigma},\sigma}} = c\left|W^\sigma\right| M^n h_M^{\vee\sigma}(b)\delta_{b,b'}, \tag{21}$$

where c, $h_M^{\vee\sigma}$ are defined by (3) and (14), respectively, and $|W^\sigma|$ is the order of the subgroup W^σ.

The discrete orthogonality relations of all types of functions $\Psi^{\widetilde{\sigma},\sigma}$ are also inherited by the related orbit functions with Hartley kernel $\zeta^{\widetilde{\sigma},\sigma}$ i.e. For any $\sigma,\widetilde{\sigma} \in \Sigma$ and any $b,b' \in \Lambda_M^{\widetilde{\sigma},\sigma}$ it holds that

$$\left\langle \zeta_b^{\widetilde{\sigma},\sigma}, \zeta_{b'}^{\widetilde{\sigma},\sigma}\right\rangle_{F_M^{\widetilde{\sigma},\sigma}} = c\left|W^\sigma\right| M^n h_M^{\vee\sigma}(b)\delta_{b,b'}. \tag{22}$$

5.2 Discrete orbit function transforms

An arbitrary function $f : \mathbb{R}^n \to \mathbb{C}$, sampled on the point set $F_M^{\widetilde{\sigma},\sigma}$, can be interpolated by the interpolating function $\mathrm{I}[f]_M^{\widetilde{\sigma},\sigma}$. The interpolating function $\mathrm{I}[f]_M^{\widetilde{\sigma},\sigma}$ is required to coincide with f at all the gridpoints of $F_M^{\widetilde{\sigma},\sigma}$,

$$\mathrm{I}[f]_M^{\widetilde{\sigma},\sigma}(a) = f(a), \quad a \in F_M^{\widetilde{\sigma},\sigma}. \tag{23}$$

The interpolating function $\mathrm{I}[f]_M^{\widetilde{\sigma},\sigma}$ is given in terms of expansion functions $\Psi_b^{\widetilde{\sigma},\sigma}$,

$$\mathrm{I}[f]_M^{\widetilde{\sigma},\sigma}(a) = \sum_{b \in \Lambda_M^{\widetilde{\sigma},\sigma}} k_b^{\widetilde{\sigma},\sigma} \Psi_b^{\widetilde{\sigma},\sigma}(a), \quad a \in \mathbb{R}^n. \tag{24}$$

The frequency spectrum coefficients $k_b^{\widetilde{\sigma},\sigma}$ are uniquely determined by the standard method of calculation of Fourier coefficients

$$k_b^{\widetilde{\sigma},\sigma} = \frac{\left\langle f, \Psi_b^{\widetilde{\sigma},\sigma} \right\rangle_{F_M^{\widetilde{\sigma},\sigma}}}{\left\langle \Psi_b^{\widetilde{\sigma},\sigma}, \Psi_b^{\widetilde{\sigma},\sigma} \right\rangle_{F_M^{\widetilde{\sigma},\sigma}}} = \frac{1}{c\,|W^\sigma|\, M^n h_M^{\vee\sigma}(b)} \sum_{a \in F_M^{\widetilde{\sigma},\sigma}} \epsilon^\sigma(a) f(a) \overline{\Psi_b^{\widetilde{\sigma},\sigma}(a)}. \tag{25}$$

Taking into account equality (23), relations (25) and (24) constitute the forward and backward discrete Fourier-Weyl transforms, respectively, of the discretized function f. Furthermore, using the Parseval equality of the orthogonal basis $\Psi_b^{\widetilde{\sigma},\sigma}, b \in \Lambda_M^{\widetilde{\sigma},\sigma}$ results in the following relation

$$\sum_{a \in F_M^{\widetilde{\sigma},\sigma}} \epsilon^\sigma(a) \left| f(a) \right|^2 = c\,|W^\sigma|\, M^n \sum_{b \in \Lambda_M^{\widetilde{\sigma},\sigma}} h_M^{\vee\sigma}(b) \left| k_b^{\widetilde{\sigma},\sigma} \right|^2.$$

Similarly to the interpolation formulas and discrete transforms of the standard orbit functions, their related real-valued versions are formulated in terms of Hartley orbit functions. An arbitrary real-valued function $g : \mathbb{R}^n \to \mathbb{R}$ sampled on the point set $F_M^{\widetilde{\sigma},\sigma}$ can be interpolated by the real-valued interpolating functions $\mathrm{Ih}[g]_M^{\widetilde{\sigma},\sigma}$. Again, the interpolating function $\mathrm{Ih}[g]_M^{\widetilde{\sigma},\sigma}$ coincides with g at all the gridpoints $F_M^{\widetilde{\sigma},\sigma}$,

$$\mathrm{Ih}[g]_M^{\widetilde{\sigma},\sigma}(a) = g(a), \quad a \in F_M^{\widetilde{\sigma},\sigma},$$

and is given in terms of expansion functions $\zeta_b^{\widetilde{\sigma},\sigma}$,

$$\mathrm{Ih}[g]_M^{\widetilde{\sigma},\sigma}(a) = \sum_{b \in \Lambda_M^{\widetilde{\sigma},\sigma}} l_b^{\widetilde{\sigma},\sigma} \zeta_b^{\widetilde{\sigma},\sigma}(a), \quad a \in \mathbb{R}^n.$$

The frequency spectrum coefficients $l_b^{\widetilde{\sigma},\sigma}$ of the Hartley-Weyl transform are determined by

$$l_b^{\widetilde{\sigma},\sigma} = \frac{\left\langle g, \zeta_b^{\widetilde{\sigma},\sigma} \right\rangle_{F_M^{\widetilde{\sigma},\sigma}}}{\left\langle \zeta_b^{\widetilde{\sigma},\sigma}, \zeta_b^{\widetilde{\sigma},\sigma} \right\rangle_{F_M^{\widetilde{\sigma},\sigma}}} = \frac{1}{c\,|W^\sigma|\, M^n h_M^{\vee\sigma}(b)} \sum_{a \in F_M^{\widetilde{\sigma},\sigma}} \epsilon^\sigma(a) g(a) \zeta_b^{\widetilde{\sigma},\sigma}(a) \tag{26}$$

and the relation between the sum of squared values of g and the sum of squared values of its frequency spectrum is

$$\sum_{a \in F_M^{\widetilde{\sigma},\sigma}} \epsilon^\sigma(a) g^2(a) = c\,|W^\sigma|\, M^n \sum_{b \in \Lambda_M^{\widetilde{\sigma},\sigma}} h_M^{\vee\sigma}(b) (l_b^{\widetilde{\sigma},\sigma})^2.$$

6 Concluding Remarks

Discrete orthogonality relations (21) and decomposition formulas are in [4] exemplified for six types of two-variable $E-$functions of algebras C_2 and G_2. Effectiveness of interpolation formulas (24) of these two-variable $E-$functions is demonstrated on complex-valued model functions in [5]. Comparable interpolating ability of real-valued functions is expected for Hartley orbit functions. Good performance of orbit functions in interpolation tasks indicates great potential in other fields related to digital data processing. The interpolation properties of all types of orbit functions as well as existence of general convergence criteria of the operator sequence $\mathrm{I}_M^{\widetilde{\sigma},\sigma} : f \mapsto \mathrm{I}[f]_M^{\widetilde{\sigma},\sigma}$ deserve further study.

Link between the Weyl-orbit functions and the inherited discrete and continuous orthogonality relations of the generalized multidimensional Chebyshev polynomials is being recently investigated in connection with the corresponding polynomial methods such as polynomial interpolation, approximation and cubature formulas.

The discrete transforms (25) and (26) of orbit functions specialize for the case A_1 to one-variable discrete Fourier, discrete Hartley, discrete cosine and sine transforms [1].

Discrete orthogonality relations (21) and (22) are formulated on the points of the refined dual weight lattice. This choice of the points induces in turn the dual affine Weyl group (anti)symmetry of the orbit function labels. The labels of this discretization share the same (anti)symmetry with the points generated by the given affine Weyl group. The Fourier transforms constructed on the points of the refined (dual) root lattice represent the remaining unresolved discrete transforms related to the four classical Weyl group invariant lattices. The merit of having all four classical lattice transforms available is the possibility of generating novel and relevant transforms on generalized lattices, including the 2D honeycomb lattice. The open problem of detailing the root lattice transforms is however, specifically challenging, since the symmetry groups of the labels are not in general Coxeter groups.

References

[1] R. N. Bracewell, *Discrete Hartley transform*, J. Opt. Soc. Am. **73** (1983) 1832-1835.

[2] J. Cserti, G. Tichy, *A simple model for the vibrational modes in honeycomb lattices*, Eur. J. Phys. **25** (2004) 723–736.

[3] T. Czyżycki, J. Hrivnák, *Generalized discrete orbit function transforms of affine Weyl groups*, J. Math. Phys. **55** (2014) 113508.

[4] L. Háková, J. Hrivnák, J. Patera, *Six types of $E-$functions of Lie group $O(5)$ and* $G(2)$ J. Phys. A: Math. Theor. **43** (2010) 165206.

[5] L. Háková, J. Hrivnák, *Fourier Transforms of $E-$functions of $O(5)$ and $G(2)$*, in: Geometric Methods in Physics, XXXII workshop, Białowieża, Poland, June 30–July 6, 2013, BirkhĊuser/Springer (2014) 243–252.

[6] J. Hrivnák, I. Kashuba J. Patera, *On $E-$functions of semisimple Lie groups*, J. Phys. A: Math. Theor. **44** (2011) 325205.

[7] J. Hrivnák, L. Motlochová, *Discrete cosine and sine transforms generalized to honeycomb lattice*, arxiv:1706.05672.

[8] J. Hrivnák, L. Motlochová, J. Patera, *On discretization of tori of compact simple Lie groups II,* J. Phys. A: Math. Theor. **45** (2012) 255201.

[9] J. Hrivnák, M. Juránek, *On E−discretization of tori of compact simple Lie groups: II,* J. Math. Phys. (in press).

[10] J. Hrivnák, J. Patera, *On discretization of tori of compact simple Lie groups,* J. Phys. A: Math. Theor. **42** (2009) 385208.

[11] J. Hrivnák, J. Patera, *On E−discretization of tori of compact simple Lie groups* J. Phys. A: Math. Theor. **43** (2010) 165206.

[12] J. Hrivnák, M. A. Walton, *Weight-Lattice Discretization of Weyl-Orbit Functions,* J. Math. Phys. **57**, (2016) 083512.

[13] J. E. Humphreys, *Reflection groups and Coxeter groups*, Cambridge Studies in Advanced Mathematics, **29** (1990) Cambridge University Press, Cambridge.

[14] A. U. Klimyk, J. Patera, *Orbit functions,* SIGMA **2** (2006) 006.

[15] A. U. Klimyk, J. Patera, *Antisymmetric orbit functions,* SIGMA **3** (2007) 023.

[16] A. U. Klimyk, J. Patera, *E−orbit functions,* SIGMA **4** (2008) 002.

[17] H. Kühl, M. D. Sacchi, J. Fertig, *The Hartley transform in seismic imaging*, Geophysics, **66** (2001) 1251–1257.

[18] I. Paraskevas, M. Barbarosou, E. Chilton, *Hartley transform and the use of the Whitened Hartley spectrum as a tool for phase spectral processing*, J. Eng. (2015).

Cramér-Rao Induced Bound for Interference-to-Signal Ratio Achievable through Non-Gaussian Independent Component Extraction[*]

Václav Kautský

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

advisor: Zbyněk Koldovský, Institute of Information Technology and Electronics
Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, TUL

Abstract. This paper deals with the Cramér-Rao Lower Bound (CRLB) for a novel blind source separation method called Independent Component Extraction (ICE). The Cramér-Rao Lower Bound is used to determine the best achievable accuracy of blind source separation (BSS) methods. Only efficient methods are able to reach the CRLB. Blind source separation focuses on estimation of unknown source signals from observed mixtures.

The most popular method for BSS in last years is well known Independent Component Analysis (ICA). We have recently performed a novel ICA based method: ICE. Compared to ICA, ICE aims to extract only one independent signal from a linear mixture. The target signal is assumed to be non-Gaussian, while the other signals, which are not separated, are modeled as a Gaussian mixture.

The most frequently used criterion for measurement of the accuracy of a method is Interference-to-Signal Ratio (ISR). Hence, CRLB-induced Bound (CRIB) for ISR is derived. Numerical simulations, performed in MATLAB, compare the CRIB with the performance of an ICA and an ICE algorithm. The results show good agreement between the theory and the empirical results.

Keywords: Blind Source Separation, Cramér-Rao Lower Bound, Independent Component Analysis, Independent Vector Analysis

Abstrakt. V této práci se zabýváme odvozením Crámerovy-Raovy dolní meze pro nově představenou metodu pro slepou separaci signálu zvanou Independent Component Extraction (ICE). Crámerova-Raova mez se využívá pro stanovení maximální dosažitelné přesnosti separace signálů pomocí dané separační metody. Metody dosahující CRLB nazýváme eficientní. Úkolem slepé separace je odhadnout neznámé zdrojové signály z jejich směsi.

V posledních letech je nejrozšířenější metodou pro slepou separaci analýza nezávislých komponent (ICA). Na základě modelu ICA jsme pro slepou separaci signálu vyvinuli novou metodu: ICE. Narozdíl od ICA, se ICE zabývá separací pouze jednoho nezávislého signálu z lineární směsi. Předpokládáme, že cílový signál není Gaussovský. Ostatní signály, které nejsou předmětem separace, pak modelujeme jako Gaussovskou směs.

Nejběžněji používaným kritériem pro měření přesnosti separačních metod je Interference-to-Signal Ratio (ISR). Z tohoto důvodu dále odvodíme mez pro toto kritérium, tzv. CRLB-induced Bound (CRIB). Pro porovnání výsledků metod ICA a ICE s odvozenou mezí CRIB jsme využili

[*]This work was supported by The Czech Science Foundation through Project No. 17-00902S.

numerické simulace v programu MATLAB. Závěry z těchto simulací ukazují na dobrou shodu mezi teoretickými předpoklady a empirickými výsledky.

Klíčová slova: Analýza nezávislých komponent, Analýza nezávislých vektorů, Cramérova-Raova dolní mez, Slepá separace signálu

Full paper: This paper has been accepted for presentation at the 2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2017), which will be held in Curaçao, Dutch Antilles, December 10-13, 2017. V. Kautský, Z. Koldovský, P. Tichavský, *Cramér-Rao Induced Bound for Interference-to-Signal Ratio Achievable Through Non-Gaussian Independent Component Extraction.*

Numerical Solution of Two-Phase Flow in Porous Media Using Unstructured Meshes on GPU[*]

Jakub Klinkovský

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

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

Abstract. In this paper, we present an efficient GPU accelerated solver for the numerical solution of two-phase compositional flow in porous media and potentially other interesting problems. The underlying system of partial differential equations is formulated in general coefficient form to allow us to easily test different models and problem formulations without substantial modifications of the numerical solver. The numerical scheme is based on the mixed-hybrid finite element and discontinuous Galerkin methods, semi-implicit time discretization, and various stabilization techniques. The used numerical methods allow us to consider any spatial dimension and use both structured and unstructured meshes. The solver is implemented in the C++ language with the help of the TNL library, the CUDA framework and OpenMP. We also present multiple key optimizations necessary for high-performance computations such as ordering of the mesh entities and an improved GMRES method. We use a benchmark problem with known semi-analytical solution to verify the convergence of the numerical scheme and present the GPU speed-up compared to single- and multi-thread computations on CPU.

Keywords: two-phase compositional flow, mixed-hybrid finite element method, upwind, GMRES method, parallel implementation on GPU, unstructured meshes

Abstrakt. V této práci předkládáme efektivní řešič akcelerovaný pomocí GPU pro numerické řešení kompozičního dvoufázového proudění v porézním prostředí a potenciálně i dalších zajímavých úloh. Soustava parciálních diferenciálních rovnic je formulovaná pomocí obecných koeficientů, díky čemuž lze jednoduše testovat různé modely a formulace úloh bez zásadních změn v numerickém řešiči. Numerické schéma je založeno na kombinaci hybridní metody smíšených konečných prvků a nespojité Galerkinovy metody, semi-implicitní časové diskretizaci a několika stabilizačních technikách. Použité numerické metody umožňují použití strukturovaných i nestrukturovaných sítí v prostoru libovolné dimenze. Řešič je implementován v jazyce C++ s využitím knihovny TNL, platformy CUDA a OpenMP. Práce také popisuje několik klíčových optimalizací pro zlepšení efektivity výpočtu, jako např. přečíslování entit sítě a modifikace metody GMRES. Konvergence numerického schématu je ověřena pomocí analýzy experimentálního řádu

[*] This work has been supported by the Student Grant Agency of the Czech Technical University in Prague, project No. SGS17/194/OHK4/3T/14.

konvergence pro testovací úlohu se známým semi-analytickým řešením. Pro všechny výpočty je provedena analýza efektivity paralelního výpočtu na GPU a vícejádrovém CPU.

Klíčová slova: dvoufázové kompoziční proudění, hybridní metoda smíšených konečných prvků, upwind, metoda GMRES, paralelní implementace na GPU, nestrukturované sítě

1 Introduction

Numerical simulations of complex practical problems in the field of computational flow dynamics require immense computational power. In recent years, using GPUs for general-purpose computations has become very popular because of their massive computational power and better power efficiency compared to traditional CPUs. However, efficient utilization of the GPU typically requires data structures and algorithms to be designed specifically for this architecture.

In this work, we present a numerical solver for a general system of partial differential equations, which can be used to describe many practical problems. We describe the key aspects of the efficient implementation of the solver for the CPU and GPU architectures. The GPU speed-up compared to single-thread and multi-thread computations on CPU is measured on a benchmark problem of two-phase flow in porous media.

2 General formulation

The numerical scheme is derived for the following system of n partial differential equations in a general coefficient form

$$\begin{aligned}\sum_{j=1}^{n} N_{i,j}\frac{\partial Z_j}{\partial t} + \sum_{j=1}^{n} \boldsymbol{u}_{i,j}\cdot\nabla Z_j +& \\ \nabla\cdot\left[m_i\left(-\sum_{j=1}^{n}\mathbf{D}_{i,j}\nabla Z_j + \boldsymbol{w}_i\right) + \sum_{j=1}^{n} Z_j\boldsymbol{a}_{i,j}\right] + \sum_{j=1}^{n} r_{i,j}Z_j = f_i&\end{aligned} \tag{1}$$

for $i = 1, ..., n$, where the unknown vector function $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ depends on position vector $\boldsymbol{x} \in \Omega \subset \mathbb{R}^d$ and time $t \in [0,T]$, $d = 1,2,3$. The system of equations (1) is supplemented by the initial condition

$$Z_j(\boldsymbol{x},0) = Z_j^{ini}(\boldsymbol{x}), \quad \forall\boldsymbol{x}\in\Omega,\ j = 1,\dots,n, \tag{2}$$

and boundary conditions for all $t \in (0,T)$,

$$Z_j(\boldsymbol{x},t) = Z_j^{\mathcal{D}}(\boldsymbol{x},t), \quad \forall\boldsymbol{x}\in\Gamma_j^{\mathcal{D}}\subset\partial\Omega,\ j = 1,...,n, \tag{3a}$$

$$\boldsymbol{v}_i(\boldsymbol{x},t)\cdot\boldsymbol{n}_{\partial\Omega}(\boldsymbol{x}) = v_i^{\mathcal{N}}(\boldsymbol{x},t), \quad \forall\boldsymbol{x}\in\Gamma_i^{\mathcal{N}}\subset\partial\Omega,\ i = 1,...,n, \tag{3b}$$

where by $\boldsymbol{v}_i$ we denote the velocity

$$\boldsymbol{v}_i = -\sum_{j=1}^{n}\mathbf{D}_{i,j}\nabla Z_j + \boldsymbol{w}_i. \tag{4}$$

Based on the nonlinear coefficients in (1) we refer to the computational method as `NumDwarf`. The choice of coefficients in (1) depends on the problem and its formulation. The details of the choice of coefficients for the immiscible two-phase flow and two-phase compositional flow in porous media can be found in [5, 6].

3 Numerical scheme

The numerical scheme for the solution of the system (1) is based on the combination of the mixed-hybrid finite element and discontinuous Galerkin methods for the spatial discretization, the Euler method for temporal discretization and the semi-implicit approach of the frozen coefficients method for the linearization in time. The scheme is stabilized with upwind and mass-lumping techniques.

The scheme has the following features: it is locally conservative, leads to a linear system with a positive-definite matrix, allows to use unstructured meshes, and it can be efficiently parallelized. Last but not least, a modification of the MHFE method described in [5] can be employed to solve problems with vanishing diffusion.

The detailed derivation of the numerical scheme can be found in [5, 6].

4 Implementation

The solver is implemented in the C++ language with the help of the TNL library, the CUDA platform [8] for the GPU parallelization, and OpenMP [3] for the CPU parallelization. The TNL library is being developed by the team around Tomáš Oberhuber at the Department of Mathematics, FNSPE CTU in Prague, and the key novel algorithms and data structures implemented in TNL for the `NumDwarf` solver are described in the following subsections.

4.1 Data layout

In high-performance computing, data structures and algorithms have to be designed collectively. The `NumDwarf` solver stores many coefficients which are naturally stored in multidimensional arrays. An interesting problem is how to choose the orientation of these arrays, i.e., the order of indices for accessing the elements. In [6], it is explained that the optimal orientation depends on the computational architecture, for example in the case of two-dimensional arrays, the optimal orientation for CPU is row-wise, but for GPU it is column-wise. To avoid code duplication, we need to have a unified interface independent of the architecture and a possible technical solution using multiple C++14 meta-programming techniques has been proposed in [6].

4.2 Parallelization of the numerical scheme

The computation of the numerical solution to (1) consists of initialization and a time loop, which for each $k = 0, \dots, N_k - 1$ computes the approximation of the solution $\boldsymbol{Z}^{k+1}$ at time t_{k+1} from the state $\boldsymbol{Z}^k$ at current time t_k. The computations in each time step involve many local computations on the mesh entities such as coefficient updates which are independent of each other and therefore can be computed in parallel. We also have to assemble many small matrices $\mathbf{Q}_K \in \mathbb{R}^{n,n}$ which are local to each cell $K \in \mathcal{K}_h$ and compute the inverses $\mathbf{Q}_K^{-1}\mathbf{R}_{K,F}$ and $\mathbf{Q}_K^{-1}\boldsymbol{G}_K$ for each $K \in \mathcal{K}_h$ and $F \in \mathcal{E}_K$. The computation on local inverses on GPU can be implemented efficiently using the LU decomposition of matrices stored in the shared memory [6].

Then we have to assemble the sparse matrix for the linear system $\mathbf{A}\boldsymbol{Z}^{k+1} = \boldsymbol{b}$ which has to be solved to obtain the approximation $\boldsymbol{Z}^{k+1}$ for the next time level. In sequential codes, matrices arising from various PDE discretizations are traditionally constructed by initializing all matrix entries to zero, traversing the mesh cells $K \in \mathcal{K}_h$, and adding the coefficients local to K to the corresponding matrix elements. However, when performed in parallel, this simple approach leads to conflicts between multiple cells that contribute to common matrix elements. The conflicts can be avoided by mesh coloring [2] but it still impairs the efficiency of the solver for medium size problems. In the `NumDwarf` scheme, the rows of $\mathbf{A}$ correspond to faces $E \in \mathcal{E}_h$ and the contributing terms originate from faces $F \in \mathcal{E}_{K_1} \cup \mathcal{E}_{K_2}$ of cells K_1 and K_2 adjacent to the face E. Therefore, the matrix can be assembled row–by–row even in parallel without any conflicts. In addition, the number of non-zero elements per row is fixed and depends only on the geometry of mesh cells. This is advantageous for GPUs because it avoids insertion of padding zeros to the sparse matrix storage format as well as divergent threads during the SpMV operation. For the meshes consisting of a single type of cells, i.e., with constant number of faces per cell e, the number of non-zero entries of $\mathbf{A}$ is $(2e-1)n$, en, and 1 for rows corresponding to inner, Neumann boundary, and Dirichlet boundary faces, respectively.

4.3 Linear system solver

The resolution of the linear system $\mathbf{A}\boldsymbol{Z}^{k+1} = \boldsymbol{b}$ at each time level is the computationally most expensive step. The matrix $\mathbf{A} \in \mathbb{R}^{N,N}$ is large, sparse, nonsymmetric, and its structure can be very complex depending on the mesh ordering. Direct methods for the solution of such systems suffer from huge memory requirements due to fill-in, therefore iterative methods such as GMRES, BiCGstab or TFQMR are usually more efficient.

Due to highly non-linear coefficients in (1), the matrix $\mathbf{A}$ is extremely ill-conditioned and methods such as TFQMR need a strong preconditioner in order to converge. The TNL library currently provides only the Jacobi preconditioner for all architectures, therefore we rely on the restarted GMRES(s) method which is robust enough to converge. The GMRES method is based on the Arnoldi's algorithm for the construction of the orthonormal basis of the Krylov subspace $\mathcal{K}_s = \operatorname{span}\{\bar{\boldsymbol{v}}_1, \mathbf{A}\bar{\boldsymbol{v}}_1, \dots, \mathbf{A}^{s-1}\bar{\boldsymbol{v}}_1\}$ which traditionally uses the MGS orthogonalization. It is commonly written as in [9]:

Algorithm 1

```
1.     Set $\bar{\boldsymbol{v}}_1 \in \mathbb{R}^N$ such, that $\|\bar{\boldsymbol{v}}_1\| = 1$.
2.     For $i = 1, \dots, s$:
2.1.       $\bar{\boldsymbol{w}}_i := \mathbf{A}\bar{\boldsymbol{v}}_i$
2.2.       For $k = 1, \dots, i$:
2.2.1.         $h_{ki} = \bar{\boldsymbol{w}}_i^T \bar{\boldsymbol{v}}_k$
```

```
2.2.2.        $\bar{\boldsymbol{w}}_i := \bar{\boldsymbol{w}}_i - h_{ki}\bar{\boldsymbol{v}}_k$
2.3.      $h_{i+1,i} = \|\bar{\boldsymbol{w}}_i\|$. If $h_{i+1,i} = 0$, stop.
2.4.      $\bar{\boldsymbol{v}}_{i+1} = \frac{1}{h_{i+1,i}}\bar{\boldsymbol{w}}_i$
```

Algorithm 1 is inherently a sequential algorithm because the order of steps in the inner loop ensures numerical stability. In practice, we even have to repeat the orthogonalization step 2.2 twice to ensure convergence (MGSR). Overall, only SpMV and Level 1 BLAS operations can be parallelized.

To improve the scalability of the Arnoldi's algorithm, we replace the MGS part by Householder transformations [9]:

Algorithm 2

```
1.   Choose non-zero vector $\boldsymbol{z}_1 \in \mathbb{R}^N$.
2.   For $i = 1, \dots, s+1$:
2.1.      Find Householder vector $\boldsymbol{y}_i \in \mathbb{R}^N$ such, that $(\boldsymbol{y}_i)_j = 0$ for $j = 1, \dots, i-1$ and
          $(\mathbf{P}_i\boldsymbol{z}_i)_j = 0$ for $j = i+1, \dots, N$, where $\mathbf{P}_i = \mathbf{I} - \bar{t}_i\boldsymbol{y}_i\boldsymbol{y}_i^T$, $\bar{t}_i = \dfrac{2}{\|\boldsymbol{y}_i\|^2}$.
2.2.      $\boldsymbol{h}_{i-1} = \left[(\mathbf{P}_i\boldsymbol{z}_i)_j\right]_{j=1}^{s+1}$
2.3.      $\bar{\boldsymbol{v}}_i = \mathbf{P}_1 \dots \mathbf{P}_i\boldsymbol{e}_i$
2.4.      If $i \le s$, compute $\boldsymbol{z}_{i+1} = \mathbf{P}_i \dots \mathbf{P}_1\mathbf{A}\bar{\boldsymbol{v}}_i$.
```

Algorithm 2 is numerically more stable than the original version using MGS (Algorithm 1) and its cost is comparable to MGSR. To expose more parallelism, we replace the sequential application of the Householder transformations with the *compact WY representation* (CWY) for the products $\mathbf{P}_1 \dots \mathbf{P}_i$ and $\mathbf{P}_i \dots \mathbf{P}_1$, which was introduced in [10]. We denote $\mathbf{Y}_i = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_i] \in \mathbb{R}^{N,i}$, $\mathbf{T}_1 = \bar{t}_1 \in \mathbb{R}^{1,1}$ and recursively define an upper triangular matrix $\mathbf{T}_i \in \mathbb{R}^{i,i}$,

$$\mathbf{T}_i = \begin{pmatrix} \mathbf{T}_{i-1} & -\bar{t}_i\mathbf{T}_{i-1}\mathbf{Y}_{i-1}\boldsymbol{y}_i \\ 0 & \bar{t}_i \end{pmatrix}. \tag{5}$$

Then the following relations hold:

$$\mathbf{P}_1 \dots \mathbf{P}_i = \mathbf{I} - \mathbf{Y}_i\mathbf{T}_i\mathbf{Y}_i^T, \tag{6a}$$

$$\mathbf{P}_i \dots \mathbf{P}_1 = \mathbf{I} - \mathbf{Y}_i\mathbf{T}_i^T\mathbf{Y}_i^T. \tag{6b}$$

Application of the compact WY representation leads to the following modification of the Arnoldi's algorithm:

Algorithm 3

```
1.   Choose non-zero vector $\boldsymbol{z}_1 \in \mathbb{R}^N$.
2.   For $i = 1, \dots, s+1$:
2.1.      Compute $\boldsymbol{y}_i$ and $\bar{t}_i$ for the current $\boldsymbol{z}_i$ same as before.
2.2.      Update $\mathbf{Y}_i$ and $\mathbf{T}_i$ using $\bar{t}_i$, $\boldsymbol{y}_i$, $\mathbf{T}_{i-1}$ and $\mathbf{Y}_{i-1}$.
```

```
2.3.    $\boldsymbol{h}_{i-1} = \left[(\mathbf{P}_i \boldsymbol{z}_i)_j\right]_{j=1}^{s+1}$
2.4.    $\bar{\boldsymbol{v}}_i = \left(\mathbf{I} - \mathbf{Y}_i \mathbf{T}_i \mathbf{Y}_i^T\right) \boldsymbol{e}_i$
2.5.    If $i \le s$, compute $\boldsymbol{z}_{i+1} = \left(\mathbf{I} - \mathbf{Y}_i \mathbf{T}_i^T \mathbf{Y}_i^T\right) \mathbf{A} \bar{\boldsymbol{v}}_i$.
```

Algorithm 3 has better numerical stability than Algorithm 1, it has less global synchronizations because there is no explicit inner loop and the orthogonalization can be implemented with Level 2 BLAS operations.

In both variants of the GMRES method, the restarting parameter s can be chosen adaptively, which reduces the computational cost on both CPU and GPU architectures. We use the strategy introduced in [1] and slightly improved in [6].

4.4 Unstructured meshes

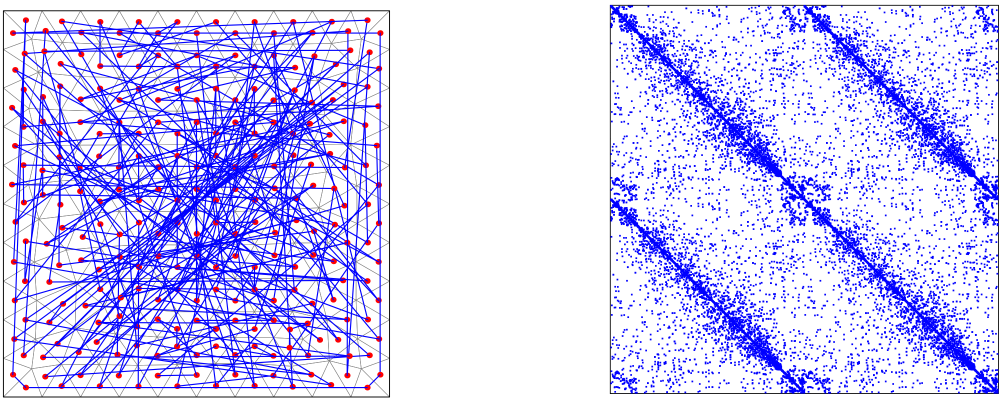
As part of the TNL library, we implemented the `Mesh` template, which is a data structure for working with homogeneous unstructured meshes, i.e. meshes where all cells have the same shape (e.g. triangle, rectangle, tetrahedron or cuboid) and the number of neighbouring cells of a vertex is not constant. Its purpose is to provide storage for numerical meshes and algorithms for accessing topological properties, such as enumerating neighbouring cells of a given vertex. It was designed with efficiency and flexibility in mind, which makes it suitable for integration into complex algorithms for high-performance computations. To achieve these goals, the implementation relies heavily on C++11 features and template meta-programming techniques.

The static compile-time configuration allows to change many parameters, such as the mesh topology determined by the cell shape, dimension of the space in which the mesh is included, coordinate data type (e.g. `float`, `double`), global and local index types (e.g. `int` and `short int`), dimensions of the entities stored in the mesh, and the data representing connectivity information between neighbouring entities.

4.4.1 Optimizations for CPU and GPU

The static configuration affects the size of the mesh itself (unnecessary entities can be omitted), as well as the size of mesh entity structures (unnecessary connections can be omitted). Consequently, the size of a mesh entity depends on its shape, but not on the number of its neighbours. To work with the mesh on GPU, it has to be initialized sequentially on the CPU and then it can be transferred to the GPU. Similarly to the handling of multidimensional arrays, the internal data layout of `Mesh` allows coalesced memory accesses during parallel traversal on GPU.

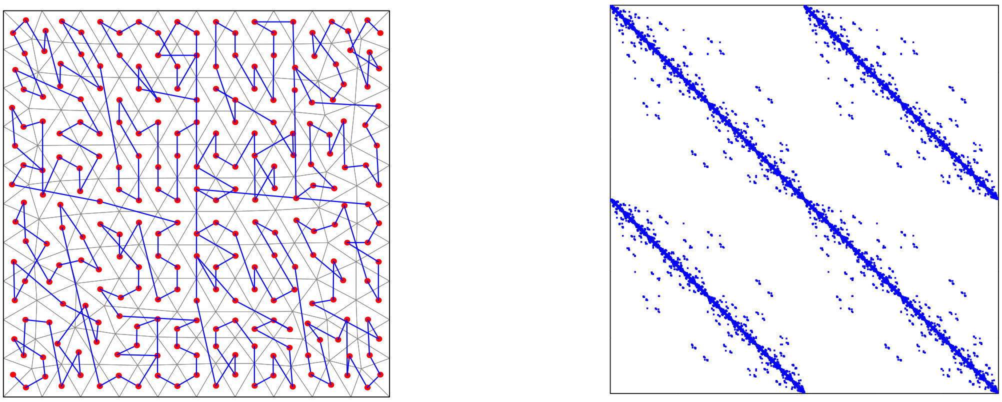
On both CPU and GPU architectures, the efficiency of the solver for (1) is strongly affected by the ordering of mesh entities. Not only direct manipulation with the mesh data structure is affected, most important consequence is the structure of the sparse matrix resulting from the MHFE discretization. We demonstrate this effect on a 2D benchmark problem using two ordering strategies: the original ordering generated by the frontal algorithm of `Gmsh` (see Fig. 1), and a custom ordering based on an in-order traversal of a d-dimensional tree of the entity centres (see Fig. 2). The original ordering does not consider the spatial position of the entities and, consequently, the corresponding sparse



(a) Mesh ordering (polygonal chain connecting the cell centres)

(b) Matrix structure

Figure 1: Ordering of the coarsest triangular mesh with Id. $2\mathrm{D}_1^{\triangle}$ generated by the frontal algorithm of `Gmsh` and the corresponding structure of the global sparse matrix.



(a) Mesh ordering (polygonal chain connecting the cell centres)

(b) Matrix structure

Figure 2: Ordering of the coarsest triangular mesh with Id. $2\mathrm{D}_1^{\triangle}$ generated by the 2-d tree traversal and the corresponding structure of the global sparse matrix.

matrix “does not look sparse”, although every row contains at most 10 non-zero elements. The alternative ordering preserves the spatial locality of neighbouring entities and the corresponding sparsity pattern constitutes of several diagonals and small blocks. The results in Table 1 show that computations using the alternative ordering are significantly faster, which can be attributed to better cache efficiency in the SpMV operation.

5 Results

We use a benchmark problem with known semi-analytical solution to verify the convergence of the numerical scheme by means of the experimental order of convergence. It is a multidimensional extension of the one-dimensional McWhorter and Sunada problem [7] for the special case of incompressible two-phase flow in homogeneous porous medium with neglected gravity and specific initial and boundary conditions. The general semi-

	Intel Core i7-5820K						Nvidia Tesla K40		
	1 core			6 cores					
Id.	Gmsh	tree	t/G	Gmsh	tree	t/G	Gmsh	tree	t/G
$2\mathrm{D}_1^{\triangle}$	0,4	0,4	0,91	0,6	0,6	0,92	1,3	1,3	1,00
$2\mathrm{D}_2^{\triangle}$	5,1	5,0	0,99	3,6	3,6	1,00	7,9	7,9	1,00
$2\mathrm{D}_3^{\triangle}$	99,9	98,5	0,99	36,2	35,7	0,99	47,0	46,2	0,98
$2\mathrm{D}_4^{\triangle}$	2 662	2 383	0,90	632,5	573,6	0,91	374,4	351,5	0,94
$2\mathrm{D}_5^{\triangle}$	64 145	45 953	0,72	15 687	11 976	0,76	3 913,2	3 387,0	0,87

Table 1: Comparison of the computational times for triangular meshes ordered by different strategies: the frontal algorithm of the Gmsh program and using 2-d tree.

analytical solution described in [4] exhibits radial symmetry due to a point injection of one of the phases. The details of the setup for this benchmark problem as well as the choice of parameters for the numerical solution can be found in [5, 6].

The numerical solution of the benchmark problem has been computed in 2D on structured rectangular grids and unstructured triangular meshes and in 3D on structured cuboidal grids and unstructured tetrahedral meshes. A series of refined meshes of each type has been used for the EOC analysis in the L_1 and L_2 norms. The results presented in Table 2 for the capillarity models by Brooks and Corey and by van Genuchten indicate that the scheme converges with the first order of accuracy in all cases.

	Brooks & Corey				van Genuchten			
Id.	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$
$2\mathrm{D}_1^{\square}$	$1{,}52 \cdot 10^{-2}$		$3{,}26 \cdot 10^{-2}$		$1{,}41 \cdot 10^{-2}$		$2{,}17 \cdot 10^{-2}$	
$2\mathrm{D}_2^{\square}$	$8{,}75 \cdot 10^{-3}$	**0,80**	$2{,}08 \cdot 10^{-2}$	**0,65**	$7{,}88 \cdot 10^{-3}$	**0,84**	$1{,}24 \cdot 10^{-2}$	**0,81**
$2\mathrm{D}_3^{\square}$	$4{,}97 \cdot 10^{-3}$	**0,82**	$1{,}35 \cdot 10^{-2}$	**0,62**	$4{,}31 \cdot 10^{-3}$	**0,87**	$6{,}83 \cdot 10^{-3}$	**0,86**
$2\mathrm{D}_4^{\square}$	$2{,}76 \cdot 10^{-3}$	**0,85**	$8{,}93 \cdot 10^{-3}$	**0,60**	$2{,}34 \cdot 10^{-3}$	**0,88**	$3{,}72 \cdot 10^{-3}$	**0,88**
$2\mathrm{D}_5^{\square}$	$1{,}51 \cdot 10^{-3}$	**0,87**	$5{,}79 \cdot 10^{-3}$	**0,63**	$1{,}29 \cdot 10^{-3}$	**0,86**	$2{,}06 \cdot 10^{-3}$	**0,85**
$2\mathrm{D}_1^{\triangle}$	$1{,}54 \cdot 10^{-2}$		$3{,}25 \cdot 10^{-2}$		$1{,}43 \cdot 10^{-2}$		$2{,}13 \cdot 10^{-2}$	
$2\mathrm{D}_2^{\triangle}$	$8{,}14 \cdot 10^{-3}$	**0,97**	$1{,}89 \cdot 10^{-2}$	**0,84**	$7{,}58 \cdot 10^{-3}$	**0,97**	$1{,}16 \cdot 10^{-2}$	**0,93**
$2\mathrm{D}_3^{\triangle}$	$4{,}44 \cdot 10^{-3}$	**0,80**	$1{,}19 \cdot 10^{-2}$	**0,61**	$4{,}01 \cdot 10^{-3}$	**0,84**	$6{,}22 \cdot 10^{-3}$	**0,83**
$2\mathrm{D}_4^{\triangle}$	$2{,}41 \cdot 10^{-3}$	**0,96**	$7{,}79 \cdot 10^{-3}$	**0,67**	$2{,}12 \cdot 10^{-3}$	**1,01**	$3{,}30 \cdot 10^{-3}$	**1,00**
$2\mathrm{D}_5^{\triangle}$	$1{,}29 \cdot 10^{-3}$	**0,86**	$4{,}90 \cdot 10^{-3}$	**0,64**	$1{,}15 \cdot 10^{-3}$	**0,85**	$1{,}79 \cdot 10^{-3}$	**0,84**
$3\mathrm{D}_1^{\square}$	$8{,}28 \cdot 10^{-3}$		$2{,}59 \cdot 10^{-2}$		$8{,}15 \cdot 10^{-3}$		$1{,}64 \cdot 10^{-2}$	
$3\mathrm{D}_2^{\square}$	$4{,}67 \cdot 10^{-3}$	**0,83**	$1{,}59 \cdot 10^{-2}$	**0,70**	$4{,}42 \cdot 10^{-3}$	**0,88**	$9{,}06 \cdot 10^{-3}$	**0,86**
$3\mathrm{D}_3^{\square}$	$2{,}60 \cdot 10^{-3}$	**0,84**	$9{,}87 \cdot 10^{-3}$	**0,69**	$2{,}36 \cdot 10^{-3}$	**0,90**	$4{,}90 \cdot 10^{-3}$	**0,89**
$3\mathrm{D}_4^{\square}$	$1{,}44 \cdot 10^{-3}$	**0,86**	$6{,}12 \cdot 10^{-3}$	**0,69**	$1{,}24 \cdot 10^{-3}$	**0,93**	$2{,}58 \cdot 10^{-3}$	**0,92**
$3\mathrm{D}_1^{\triangle}$	$1{,}15 \cdot 10^{-2}$		$3{,}48 \cdot 10^{-2}$		$1{,}21 \cdot 10^{-2}$		$2{,}43 \cdot 10^{-2}$	
$3\mathrm{D}_2^{\triangle}$	$8{,}02 \cdot 10^{-3}$	**0,69**	$2{,}52 \cdot 10^{-2}$	**0,62**	$8{,}13 \cdot 10^{-3}$	**0,77**	$1{,}66 \cdot 10^{-2}$	**0,73**
$3\mathrm{D}_3^{\triangle}$	$4{,}41 \cdot 10^{-3}$	**0,86**	$1{,}49 \cdot 10^{-2}$	**0,75**	$4{,}26 \cdot 10^{-3}$	**0,93**	$8{,}83 \cdot 10^{-3}$	**0,90**
$3\mathrm{D}_4^{\triangle}$	$2{,}40 \cdot 10^{-3}$	**1,02**	$8{,}62 \cdot 10^{-3}$	**0,93**	$2{,}16 \cdot 10^{-3}$	**1,14**	$4{,}53 \cdot 10^{-3}$	**1,13**
$3\mathrm{D}_5^{\triangle}$	$1{,}26 \cdot 10^{-3}$	**1,01**	$5{,}48 \cdot 10^{-3}$	**0,71**	$1{,}09 \cdot 10^{-3}$	**1,08**	$2{,}28 \cdot 10^{-3}$	**1,08**

Table 2: Errors of numerical solutions and experimental orders of convergence for rectangular, triangular, cuboidal and tetrahedral meshes.

For the same benchmark problem, we also present the GPU speed-up compared to single- and multi-thread computations on CPU. The values in Tables 3 and 4 demonstrate the advantages of massive parallelization for sufficiently large problems. Additionally, we compare the MGSR and CWY variants of the GMRES method which were introduced in Algorithms 1 and 3, respectively. On GPU the CWY variant is significantly faster than the MGSR variant, but on CPU the computational times are more or less the same.

		GPU	CPU											
			1 core		2 cores			4 cores			6 cores			
	Id.	*CT*	*CT*	*GSp*	*CT*	*Eff*	*GSp*	*CT*	*Eff*	*GSp*	*CT*	*Eff*	*GSp*
MGSR	$2\text{D}_1^{\square}$	5,1	0,6	**0,12**	0,7	0,45	0,13	0,8	0,19	0,15	0,9	0,11	**0,17**
	$2\text{D}_2^{\square}$	28,1	11,5	**0,41**	7,9	0,72	0,28	6,4	0,45	0,23	6,8	0,28	**0,24**
	$2\text{D}_3^{\square}$	117,1	173,6	**1,48**	95,9	0,91	0,82	61,2	0,71	0,52	52,8	0,55	**0,45**
	$2\text{D}_4^{\square}$	740,4	4 023,5	**5,43**	2 154,1	0,93	2,91	1 192,1	0,84	1,61	941,6	0,71	**1,27**
	$2\text{D}_5^{\square}$	8 237,3	82 323,5	**9,99**	47 982,0	0,86	5,82	26 919,0	0,76	3,27	19 915,5	0,69	**2,42**
CWY	$2\text{D}_1^{\square}$	1,5	0,7	**0,45**	0,4	0,79	0,28	0,3	0,52	0,22	0,3	0,41	**0,18**
	$2\text{D}_2^{\square}$	11,0	13,2	**1,20**	7,6	0,87	0,69	4,8	0,68	0,44	4,0	0,55	**0,37**
	$2\text{D}_3^{\square}$	46,3	197,0	**4,25**	107,5	0,92	2,32	65,7	0,75	1,42	52,6	0,62	**1,14**
	$2\text{D}_4^{\square}$	380,0	4 325,7	**11,38**	2 360,6	0,92	6,21	1 448,1	0,75	3,81	1 195,8	0,60	**3,15**
	$2\text{D}_5^{\square}$	4 449,9	91 166,3	**20,49**	49 004,3	0,93	11,01	29 182,1	0,78	6,56	24 684,0	0,62	**5,55**
MGSR	$2\text{D}_1^{\triangle}$	4,7	0,3	**0,07**	0,5	0,33	0,11	0,5	0,18	0,10	0,6	0,09	**0,13**
	$2\text{D}_2^{\triangle}$	22,4	5,0	**0,22**	3,9	0,65	0,17	3,1	0,40	0,14	3,6	0,23	**0,16**
	$2\text{D}_3^{\triangle}$	120,0	98,5	**0,82**	59,5	0,83	0,50	38,3	0,64	0,32	35,7	0,46	**0,30**
	$2\text{D}_4^{\triangle}$	778,3	2 382,8	**3,06**	1 298,8	0,92	1,67	701,0	0,85	0,90	573,5	0,69	**0,74**
	$2\text{D}_5^{\triangle}$	7 387,9	45 953,4	**6,22**	25 512,4	0,90	3,45	14 602,7	0,79	1,98	11 976,4	0,64	**1,62**
CWY	$2\text{D}_1^{\triangle}$	1,5	0,4	**0,27**	0,3	0,60	0,22	0,2	0,45	0,15	0,2	0,32	**0,14**
	$2\text{D}_2^{\triangle}$	8,9	6,2	**0,70**	3,7	0,84	0,42	2,3	0,66	0,26	2,0	0,52	**0,23**
	$2\text{D}_3^{\triangle}$	51,1	122,0	**2,39**	67,7	0,90	1,32	40,3	0,76	0,79	32,5	0,63	**0,64**
	$2\text{D}_4^{\triangle}$	396,1	2 695,6	**6,80**	1 480,7	0,91	3,74	855,2	0,79	2,16	671,7	0,67	**1,70**
	$2\text{D}_5^{\triangle}$	4 008,3	57 404,2	**14,32**	32 100,5	0,89	8,01	18 814,1	0,76	4,69	16 414,0	0,58	**4,09**

Table 3: Comparison of computational times *CT*, parallel CPU efficiency *Eff* and GPU/CPU speed-up *GSp* for the 2D benchmark problem.

6 Conclusion

We presented a parallel solver for a general system of PDEs based on the semi-implicit MHFEM/DG numerical scheme. Multiple optimizations were performed to improve the efficiency of the solver, namely a modified GMRES method using the CWY orthogonalization instead of MGSR was employed and the unstructured meshes were suitably reordered. The results of numerical simulations for a benchmark problem with known semi-analytical solution indicate that the numerical scheme converges with the first order of accuracy in all cases. Computations on GPU were about 20 times faster compared to 1-threaded computations on CPU and about 6 times faster compared to 6-threaded computations on CPU, hence, GPU acceleration can be very beneficial for large problems.

		GPU	CPU										
			1 core		2 cores			4 cores			6 cores		
	Id.	*CT*	*CT*	*GSp*	*CT*	*Eff*	*GSp*	*CT*	*Eff*	*GSp*	*CT*	*Eff*	*GSp*
MGSR	$3\mathrm{D}_1^{\square}$	5,9	13,8	**2,34**	7,2	0,96	1,22	4,3	0,80	0,73	3,4	0,67	**0,58**
	$3\mathrm{D}_2^{\square}$	55,7	524,6	**9,42**	304,7	0,86	5,47	173,7	0,76	3,12	128,2	0,68	**2,30**
	$3\mathrm{D}_3^{\square}$	1 234,3	21 128,7	**17,12**	12 770,7	0,83	10,35	7 317,4	0,72	5,93	6 241,6	0,56	**5,06**
	$3\mathrm{D}_4^{\square}$	44 798,3	(not computed on 1, 2 and 4 cores)								272 104,0		**6,07**
CWY	$3\mathrm{D}_1^{\square}$	2,1	15,2	**7,30**	8,0	0,96	3,82	4,4	0,86	2,13	3,4	0,75	**1,62**
	$3\mathrm{D}_2^{\square}$	30,8	564,3	**18,33**	319,5	0,88	10,38	186,7	0,76	6,07	150,3	0,63	**4,88**
	$3\mathrm{D}_3^{\square}$	828,0	20 569,5	**24,84**	12 406,1	0,83	14,98	7 092,6	0,73	8,57	5 533,7	0,62	**6,68**
	$3\mathrm{D}_4^{\square}$	31 805,6	(not computed on 1, 2 and 4 cores)								234 066,0		**7,36**
MGSR	$3\mathrm{D}_1^{\triangle}$	3,8	1,7	**0,44**	1,2	0,71	0,31	0,8	0,53	0,21	0,8	0,33	**0,22**
	$3\mathrm{D}_2^{\triangle}$	6,1	7,2	**1,19**	4,3	0,84	0,70	2,6	0,70	0,43	2,3	0,53	**0,37**
	$3\mathrm{D}_3^{\triangle}$	45,3	274,5	**6,06**	152,6	0,90	3,37	87,5	0,78	1,93	72,4	0,63	**1,60**
	$3\mathrm{D}_4^{\triangle}$	873,1	11 270,0	**12,91**	6 228,3	0,90	7,13	3 414,9	0,83	3,91	3 187,9	0,59	**3,65**
	$3\mathrm{D}_5^{\triangle}$	55 880,2	(not computed on CPU)										
CWY	$3\mathrm{D}_1^{\triangle}$	1,4	2,0	**1,48**	1,2	0,85	0,88	0,7	0,68	0,54	0,6	0,54	**0,46**
	$3\mathrm{D}_2^{\triangle}$	2,6	8,7	**3,30**	4,9	0,89	1,85	2,9	0,75	1,10	2,3	0,64	**0,86**
	$3\mathrm{D}_3^{\triangle}$	23,9	330,9	**13,87**	184,8	0,90	7,75	107,9	0,77	4,53	93,4	0,59	**3,92**
	$3\mathrm{D}_4^{\triangle}$	566,2	12 069,5	**21,32**	6 506,3	0,93	11,49	3 771,0	0,80	6,66	3 306,2	0,61	**5,84**
	$3\mathrm{D}_5^{\triangle}$	37 695,3	(not computed on CPU)										

Table 4: Comparison of computational times *CT*, parallel CPU efficiency *Eff* and GPU/CPU speed-up *GSp* for the 3D benchmark problem.

References

[1] A. H. Baker, E. R. Jessup, and T. V. Kolev. *A simple strategy for varying the restart parameter in GMRES(m).* Journal of computational and applied mathematics 230.2 (2009), pages 751–761.

[2] P. Bauer, V. Klement, T. Oberhuber, and V. Žabka. *Implementation of the Vanka-type multigrid solver for the finite element approximation of the Navier–Stokes equations on GPU.* Computer Physics Communications 200 (2016), pages 50–56.

[3] L. Dagum and R. Menon. *OpenMP: an industry standard API for shared-memory programming.* Computational Science & Engineering, IEEE 5.1 (1998), pages 46–55.

[4] R. Fučík, T. H. Illangasekare, and M. Beneš. *Multidimensional self-similar analytical solutions of two-phase flow in porous media.* Advances in Water Resources, 2016.

[5] R. Fučík, J. Klinkovský, J. Solovský, T. Oberhuber, and J. Mikyška. *Multidimensional mixed-hybrid finite element method for compositional two-phase flow in heterogeneous porous media and its parallel implementation on GPU.* Computer Physics Communications (under review, 2017).

[6] J. Klinkovský. *Mathematical modelling of two-phase compositional flow in porous media.* Master's thesis, FNSPE CTU in Prague (2017).

[7] D. B. McWhorter and D. K. Sunada. *Exact integral solutions for two-phase flow.* Water Resources Research 26 (1990), pages 399–413.

[8] NVIDIA. *CUDA Toolkit Documentation, version 8.0*, 2017, `URL: http://docs.nvidia.com/cuda/index.html`.

[9] Y. Saad. *Iterative methods for sparse linear systems.* SIAM, 2003, ISBN: 0-89871-534-2.

[10] R. Schreiber and C. Van Loan. *A storage-efficient WY representation for products of Householder transformations.* SIAM Journal on Scientific and Statistical Computing 10.1 (1989), pages 53–57.

Unfolding of Spectra in Damped Unitary Ensemble of Hyperbolic Kind*

Ondřej Kollert

2nd year of PGS, email: ondra.kollert@gmail.com
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 work aims to provide the comprehensive interaction analysis of the spectra of random matrices from hyperbolic damped ensembles. For that purpose, the transformation of spectra through unfolding procedure needs to be performed. Here, it is introduced and thoroughly investigated for general counting processes. After its application on the random matrix spectra, the nearest-neighbor spacing of the eigenvalues is studied in the dependence of its position in the spectra.

Keywords: Unfolding, Level Spacing, Counting Process, Random Matrices

Abstrakt. Tato práce si dává za cíl důkladně analyzovat spektra hyperbolických utlumených náhodných matic. Pro tento účel je nezbytné použít transformaci spektra skrze zobrazení unfolding. Zde je tento koncept představen a do detailu probírán z hlediska teorie čítacích procesů. Po aplikaci procedury unfolding je zkoumán odstup mezi sousedními vlastními čísly daných náhodných matic v závislosti na pozici těchto vlastních čísel ve spektru.

Klíčová slova: unfolding, hladinový odstup, čítací proces, nahodné matice

Introductory Talk

The counting process theory has been thoroughly dealt with in [1] where the classical theory is extended by new results of which the most important ones are presented in the paper [2]. The usefulness of those results lies in their ability to describe certain agent systems from the interaction point of view. These agents are assumed to be characterized through the sequence of one-dimensional random variables which often represent arrival times of events or the locations of some objects in space.

Here, we particularly focus on the system of eigenvalues of the so called hyperbolic damped unitary ensembles firstly mentioned in the paper [3] as the numerical implementation of the so called Calogero-Moser hyperbolic random matrices. However, the notion damped unitary ensembles (DUE) was introduced in [4]. In that paper, a very close correspondence between the matrices' eigenvalues and the system of vehicle locations was found. Particularly, it was shown that the nearest-neighbor spacing of the eigenvalues very well describes that of cars located in one lane of the road.

*This work was supported by the grant Detection of stochastic universalities in non-equilibrium states of socio-physical systems by means of Random Matrix Theory within agency GACR, number 15-15049S
.

The first section of the work is devoted to the necessary transformation procedure called unfolding. The concept will be established in its most general way for counting processes. The important transition to finite version of counting processes will be dealt with as well. In the second section, the damped unitary ensemble of random matrices will be defined. Moreover, the origin of these matrices and mainly the guess of the theoretical formula for the level spacing between the matrices' eigenvalues will be provided. The last section belongs to the confirmation and application of the gained knowledge on the numerically generated data. Primarily, the level spacing between the eigenvalues of matrices from hyperbolic DUE will be thoroughly examined.

1 Unfolding of Counting Process

Before we investigate nearest-neighbor spacing distribution, it is necessary to transform the initial system via unfolding. To properly understand the procedure, we will first introduce the basic terms and notation of the counting process theory. First of all, let us define the general counting process itself.

Definition 1. *Let $(R_i)_{i\in\mathbb{N}}$ be a tight sequence of independent a.s. positive random variables where the sequence of the inversions $(1/R_i)_{i\in\mathbb{N}}$ is also tight. Define the variables $T_k = \sum_{i=0}^{k-1} R_i$ for $k \in \mathbb{N}$ and $N_t = \#\{k \in \mathbb{N} \mid T_k \le t\}$ for $t \in \mathbb{R}$. Then $(N_t)_{t\in\mathbb{R}}$ is said to be a counting process.*

The random variable $T_1 = R_0$ in the definition above expresses kind of an initial point of a counting process while the elements of the sequence $(R_i)_{i\in\mathbb{N}}$ represent the nearest-neighbor spacings between the points $(T_k)_{k\in\mathbb{N}}$. Concerning the assumptions of tightness, they ensure that the counting process has expected properties. First, thanks to the tightness of the inversions $(1/R_i)_{i\in\mathbb{N}}$, the sequence of the partial sums $(T_k)_{k\in\mathbb{N}}$ converges to infinity a.s. Using this fact, some other fundamental results can be derived most of which are summarized in the following proposition.

Proposition 1. *Let $(N_t)_{t\in\mathbb{R}}$ be a counting process. Then for $t \in \mathbb{R}_0^+$, it holds*

1) $N_t < \infty$ a.s.,

2) $\mathrm{E}(N_t^r) < \infty$ for $r \in \mathbb{R}_0^+$,

3) $\exists p_0 \in \mathbb{R}$ so that $\mathrm{E}\big(\mathrm{e}^{pN_t}\big) < \infty$ for $p < p_0$,

4) $\lim\limits_{t\to-\infty} N_t = 0$ a.s. and $\lim\limits_{t\to\infty} N_t = \infty$ a.s.

The definition 1 is quite general and there is not actually much more to claim about the corresponding counting process. To derive some more advanced characteristics describing the process, additional assumptions have to be imposed on the sequence $(R_i)_{i\in\mathbb{N}_0}$. The most natural and also simple approach is to assume an identical distribution of the respective random variables and also $R_0 := 0$ a.s. The resulting process then represents a very famous renewal process. As a matter of fact, this is not an ideal type of a counting process to use since its properties are difficult to handle not only from the theoretical point of view. However, performing just a slight change by considering the density for the distribution of the random variable R_0 in the form

$$f_{R_0} = \lambda(1 - F_{R_1}) \tag{1}$$

with positive parameter $\lambda = 1/\mathrm{E}(R_1)$, the major downsides of the corresponding counting process disappear. The formal definition of such a process is given below.

Definition 2. *Let $(\nu_t)_{t\in\mathbb{R}}$ be a counting process with the i.i.d. spacing sequence $(R_i)_{i\in\mathbb{N}}$ and the initial point R_0 distributed according to* (1). *Then $(\nu_t)_{t\in\mathbb{R}}$ is said to be a level counting process and $L_k = R_0 + S_{k-1}$ the kth level where $S_k = \sum_{i=1}^{k} R_i$ is the so called k-fold level spacing for $k \in \mathbb{N}$.*

The process just defined was introduced in the paper [2]. It provides an easier and more slick way of dealing with the properties of counting processes and interacting agent systems in general. Its crucial property is the linearity of the corresponding expected value. Specifically, it holds $\mathrm{E}(\nu_t) = \lambda t$ for $t \in \mathbb{R}$ so the derivative of the expected value, i.e. the density of the counting process, is constant everywhere.

In this work, we aim to present a transformation which maps a general counting process from the definition 1 to the level counting one. The transformation is called unfolding and its definition is provided below.

Definition 3. *Let $(N_t)_{t\in\mathbb{R}}$ be a counting process and $\mathcal{U}\colon \mathbb{R} \to \mathbb{R}_0^+$ a function satisfying*

- $\lim_{t\to-\infty} \mathcal{U}(t) = 0$ *and* $\lim_{t\to\infty} \mathcal{U}(t) = \infty$,
- $\mathcal{U}$ *is continuous and increasing,*
- *there is a level counting process $(\nu_t)_{t\in\mathbb{R}}$ such that $N_t = \nu_{\mathcal{U}(t)}$*

The mapping $\mathcal{U}$ is said to be an unfolding of the counting process $(N_t)_{t\in\mathbb{R}}$.

In real applications, it is usually very difficult, mostly rather impossible, to verify all the assumptions so that it is eligible to apply unfolding. In fact, the most problematic part is the last point of the definition. Particularly, there is usually not enough data to verify that the transformed process satisfies all the properties of a level counting process. However, here we deal with the eigenvalues of random matrices so we can afford to generate enough of them to analyze even this condition. Some numerical tests will be thus given on this topic in the next section. Now let us introduce the specific form of a mapping which satisfies at least part of the assumptions required in the definition 3.

Theorem 1. *Let $(N_t)_{t\in\mathbb{R}}$ be a counting process such that the first member of the partial sum sequence is a countinous random variable. The mapping defined as*

$$\mathcal{U}(t) := \mu \mathrm{E}(N_t) \tag{2}$$

where $\mu = 1/\lambda$ and $\lambda > 0$ then satisfies the first two conditions in the definition 3 and for $N_t = \nu_{\mathcal{U}(t)}$, it also holds $\mathrm{E}(\nu_t) = \lambda t$ for all $t \in \mathbb{R}$. What is more, the mapping with such properties is unique.

Proof: The easier part of the proof is to verify the first two conditions. The very first one is immediately implied from the fourth claim of the proposition 1. The increasing trend of $\mathcal{U}$ follows from the the fact that a counting process is a.s. increasing function.

To proceed further, we need to express the expected value of a counting process through the distribution functions of the corresponding partial sum sequence $(T_k)_{k\in\mathbb{N}}$. In fact, it holds

$$\mathrm{E}(N_t) = \mathrm{E}\big(\#\{k \in \mathbb{N} \,|\, T_k \le t\}\big) = \mathrm{E}\left(\sum_{k=1}^{\infty} \mathrm{I}(T_k \le t)\right) = \sum_{k=1}^{\infty} F_{T_k}(t) \tag{3}$$

whereby the convergence is even uniform which follows from the monotone convergence theorem. Now because we can write $T_{k+1} = T_1 + \sum_{i=1}^{k} R_i$ for all $k \in \mathbb{N}$ and T_1 is continuous, all the members of $(T_k)_{k\in\mathbb{N}}$ must be continuous as well. All the reasoning put together, the expected value of a counting process is a continuous function.

The derivation of the last condition in the claim is the most strenuous part of the proof. First of all, let us denote the random variables transformed via $\mathcal{U}$ and the corresponding counting process as

$$L_k := \mathcal{U}(T_k)\,, \qquad \nu_t := \#\{k \in \mathbb{N} \,|\, L_k \le t\} \tag{4}$$

where $t \in \mathbb{R}$ and $k \in \mathbb{N}$. Due to the increasing trend of the function $\mathcal{U}$, the process $(\nu_t)_{t\in\mathbb{R}}$ is related to the original one $(N_t)_{t\in\mathbb{R}}$ through the relation

$$N_t = \#\{k \in \mathbb{N} \,|\, T_k \le t\} = \#\{k \in \mathbb{N} \,|\, \mathcal{U}(T_k) \le \mathcal{U}(t)\} = \nu_{\mathcal{U}(t)}$$

which is in fact the first part of the condition being proved. To show the linearity of the expected value $\mathrm{E}(\nu_t)$, it would be useful to apply the inversion of $\mathcal{U}$ now. However, that does not have to necessarily exist because $\mathcal{U}$ might not be strictly increasing. We thus need to deal with the areas where the function is constant. According to (3), the function $\mathcal{U}$ is constant on some set if and only if F_{T_k} for all $k \in \mathbb{N}$ are on that set constant. This is equivalent to the statement that $\mathcal{U}$ is strictly increasing on some set if and only if there is $k \in \mathbb{N}$ such that F_{T_k} is strictly increasing on that set. Based on these observations, define the set

$$A := \bigcup_{k=1}^{\infty} \mathrm{supp}(T_k)$$

where the symbol $\mathrm{supp}(T_k)$ denotes the support of the corresponding random variable. From here, it holds that $T_k \in A$ a.s. for all $k \in \mathbb{N}$ and as a consequence, the restriction $\mathcal{V} := \mathcal{U}|_A$ then satisfies

$$\mathcal{U}(T_k) = \mathcal{V}(T_k) \text{ a.s.}$$

Additionally, the equation $\mathcal{V}(A) = \mathbb{R}_0^+$ applies which follows from the properties of $\mathcal{U}$ as the expected value of a counting process. The restriction $\mathcal{V}$ is already strictly increasing which allows one to write

$$\mathrm{E}(\nu_t) = \sum_{k=1}^{\infty} \mathrm{E}\big(\mathcal{U}(T_k) \le t\big) = \sum_{k=1}^{\infty} \mathrm{E}\big(\mathcal{V}(T_k) \le t\big) = \sum_{k=1}^{\infty} \mathrm{E}\big(T_k \le \mathcal{V}^{-1}(t)\big) = \mathrm{E}\big(N_{\mathcal{V}^{-1}(t)}\big)$$

where the relation (4) was used. Using the definition (2) and also the fact that $\mathcal{V}^{-1}(t) \in A$ for all $t \in \mathbb{R}_0^+$, we finally get the equality

$$\mathrm{E}(\nu_t) = \lambda\mathcal{U}\big(\mathcal{V}^{-1}(t)\big) = \lambda\mathcal{V}\big(\mathcal{V}^{-1}(t)\big) = \lambda t\,.$$

The last part of the proof is devoted to the uniqueness of the function chosen as (2) having all the properties claimed. Let $\widetilde{\mathcal{U}}$ be an arbitrary function satisfying the first two conditions in the definition 3 and also the relation $N_t = \nu_{\widetilde{\mathcal{U}}(t)}$ where $\mathrm{E}(\nu_t) = \lambda t$ for all $t \in \mathbb{R}$. The mapping $\widetilde{\mathcal{U}}$ then complies with the definition (2) as

$$\mathrm{E}(N_t) = \mathrm{E}(\nu_{\mathcal{U}(t)}) = \lambda \widetilde{\mathcal{U}}(t)\,.$$

□

According to the relation (4), the function $\mathcal{U}$ maps the points $(T_k)_{k\in\mathbb{N}}$ to the new ones $(L_k)_{k\in\mathbb{N}}$ so that they become uniformly distributed in their state space. If the state space is for instance time, this action results in the loss of the information about the time evolution in the system. On the other hand, if the mapping $\mathcal{U}$ is in addition unfolding, the simplicity of the resulting level process allows one to unfold many useful properties about the corresponding system as will be seen in the third section. Various counting processes can be in this way transformed so that they are examined and consequently compared to each other.

As a result of the theorem 1, the mapping of the form (2) is actually the only candidate which could satisfy even the third assumption for the unfolding of a counting process $(N_t)_{t\in\mathbb{R}}$. That brings up a question what are the requirements to be imposed on the process $(N_t)_{t\in\mathbb{R}}$ so that the function (2) is its unfolding. The intuitive idea is that unfolding only kind of rescales all the nearest-neighbor spacings so that their distributions become the same as it is required in the definition for a level process. Therefore, it is believed that the counting process, on which we intend to apply unfolding, should be formed by the sequence of spacings $(R_i)_{i\in\mathbb{N}}$ all having the same or very similar distribution up to some scale constant. Ideally, it should thus hold

$$\forall i, j \in \mathbb{N}\ \exists s \in \mathbb{R}\ :\ R_i \overset{\mathrm{D}}{=} sR_j\,. \tag{5}$$

Unfolding is shrouded by mysteries and it is still far from being completely understood. Some of the insights will be given in the next sections using real data, but before that, let us present one particular case in which the mapping (2) actually satisfies all the requirements to be unfolding.

Theorem 2. *Let $(N_t)_{t\in\mathbb{R}}$ be a counting process defined through the partial sum sequence $(T_k)_{k\in\mathbb{N}}$. Suppose that the sequence $(T_{k,n})_{k=1}^n$ is the increasingly ordered version of the i.i.d. random variables $(Y_{k,n})_{k=1}^n$ for all $n \in \mathbb{N}$ such that the limit $\lim_{n\to\infty} T_{k,n} \longrightarrow T_k$ holds for all $k \in \mathbb{N}$. Then unfolding of the process $(N_t)_{t\in\mathbb{R}}$ exists and its image results in the homogeneous Poisson process.*

Proof: The theorem will be shown by the transition from the process $(N_t)_{t\in\mathbb{R}}$ to its finite version defined as $N_{t,n} := \#\{k \in \widehat{n}\,|\, T_{k,n} \le t\}$ for $t \in \mathbb{R}$ where n is the number of elements present in the system. The corresponding finite version of unfolding for $\mu < \infty$ then satisfies

$$\mathcal{U}_n(t) := \mu\mathrm{E}\big(N_{t,n}\big) = \mu\mathrm{E}\big(\#\{k \in \widehat{n}\,|\, T_k \le t\}\big) = \mu \sum_{k=1}^{n} F_{T_{k,n}}(t) = \mu n F_{Y_{1,n}}(t) \tag{6}$$

where the distribution function of $Y_{1,n}$ is the mixture of all the distribution functions $(F_{T_{k,n}})_{k=1}^{n}$. Based on this kind of notation, it is clear that $(T_{k,n})_{k=1}^{n}$ represents the ordered statistics of the i.i.d. sequence $(Y_{k,n})_{k=1}^{n}$ established in the claim of the theorem. Analogically, the sequence of levels $L_{k,n} := \mathcal{U}_n(T_k)$ is then the ordered version of the variables $X_{k,n} := \mathcal{U}_n(Y_{k,n})$. According to the relation (6), the sequence $(X_{k,n})_{k=1}^{n}$ is also i.i.d. and moreover, its elements has uniform distribution $\mathrm{U}(0, \mu n)$.

Suppose now that $(\nu_{t,n})_{t\in\mathbb{R}}$ is the finite counting process defined through the partial sum sequence $(L_{k,n})_{k=1}^{n}$. As the consequence of the results obtained in the previous paragraph, the process $(\nu_{t,n})_{t\in\mathbb{R}}$ has the binomial distribution $\mathrm{Bi}(n, t/(\mu n))$. Denoting the limits of $(N_{t,n})_{t\in\mathbb{R}}$ and $\mathcal{U}_n$ as $(N_t)_{t\in\mathbb{R}}$ and $\mathcal{U}$ respectively, it additionally holds

$$\begin{aligned} N_t &= \lim_{n\to\infty} N_{t,n} = \lim_{n\to\infty} \nu_{\mathcal{U}_n(t),n} = \nu_{\mathcal{U}(t)} \qquad (7) \\ \mathrm{E}(\nu_t) &= \lim_{n\to\infty} \mathrm{E}(\nu_{t,n}) = \lim_{n\to\infty} nF_{X_{1,n}}(t) = \lambda t \end{aligned}$$

for all $t \in \mathbb{R}$ where $\lambda := 1/\mu$. Now thanks to the convergence $\lim_{n\to\infty} \mathrm{Bi}(n, \lambda t/n) = \mathrm{Po}(\lambda)$, the resulting counting process $(\nu_t)_{t\in\mathbb{R}}$ is actually the Poisson process. Moreover, since the Poisson process satisfies the properties of a level counting one, the mapping $\mathcal{U}$ is truly unfolding.

□

As a matter of fact, the counting process $(N_t)_{t\geq 0}$ described in the claim of the theorem above is the inhomogeneous Poisson process. Indeed, the relation (7) directly implies that

$$\mathrm{P}(N_t = k) = \mathrm{P}(\nu_{\mathcal{U}(t)} = k) = \frac{\mathcal{U}^k(t)}{k!}\,\mathrm{e}^{-\mathcal{U}(t)}$$

for all $k \in \mathbb{N}$ and $t \in \mathbb{R}$. This relation actually provide the way of generating a general counting process with an arbitrary expected value $\mathrm{E}(N_t) = \lambda\mathcal{U}(t)$ for $t \in \mathbb{R}$. The function $\mathcal{U}$ might then represent a time or a space evolution reflecting some real application system and the parameter λ the intensity which the elements occur in that system with.

The concept of the finite counting process discussed above is crucial while dealing with the real application systems since they naturally always contain a finite number of elements. That is why the results presented for the counting processes based on the definition 1 are just asymptotically approximative ones. As was mentioned, that is also why the tools introduced here and in the paper [2] must be used for the systems with number of elements high enough. Their actual application will be performed on the real data in the following sections.

2 Introduction to DUE of Hyperbolic Kind

The major system studied in this work is the set of eigenvalues of the random matrices called damped unitary ensembles (DUE). Before restricting only to their hyperbolic version, let us introduce the concept generally.

Definition 4. *The random matrix* $\mathbf{D} = \left(D_{ij,n}\right)_{i,j=1}^{n}$ *for natural* n *is said to belong to the damped unitary ensemble if* $D_{ij} \sim \mathrm{N}(0, \sigma^2)$ *for* $i = j$ *and* $D_{ij,n} = \mathrm{i}g/f_n(i-j)$ *a.s. for*

$i \neq j$ where g and σ are positive. The function f_n is required to be continuous, odd and to satisfy $|\lim_{t\to\infty} f_n(t)| = \infty$.

Since the function f_n is odd, the matrix defined above must be always hermitian. That means the corresponding eigenvalues are real random variables so the concepts discussed in the previous section apply to them as well. The word dumped was used because of the increasing character of the function f_n which makes the elements of the matrix $\mathbf{D}$ smaller as they get farther from the diagonal. The simplest type the function f_n is a linear one for which the respective ensemble is called the rational. In this work, we will deal with the hyperbolic ensemble which is obtained by setting

$$\frac{1}{f_n(t)} = \frac{2\pi}{n\sinh(2\pi t/n)}$$

for $t \in \mathbb{R}$. This particular choice of the function f_n comes from the paper [3] where the corresponding matrix ensembles were introduced as the numerical implementation of the so called Calogero-Moser hyperbolic matrices. These represent the Lax matrices of the integrable models characterized by the Hamiltonian

$$H(p_1,\ldots,p_n,q_1,\ldots,q_n) = \frac{1}{2}\sum_{i=1}^{n} p_i^2 + g^2 \sum_{i<j} \frac{1}{f_{4\pi/\gamma}(q_i - q_j)} \tag{8}$$

for n particles with momentum $(p_i)_{i=1}^n$ and one-dimensional positions $(q_i)_{i=1}^n$ where g and γ are positive parameters. The Lax matrix $\mathbf{L}$ is determined by the existence of the pair matrix $\mathbf{M}$ such that the Hamilton equations can be then rewritten in the form

$$\frac{\partial \mathbf{L}}{\partial t} = \mathbf{LM} - \mathbf{ML}\,. \tag{9}$$

Considering the Hamiltonian (8), the condition (9) determines the elements of the matrix $\mathbf{L}$ as $L_{ij} = p_i$ for $i = j$ and $L_{ij} = \mathrm{i}g/f_{4\pi/\gamma}(q_i - q_j)$ for $i \neq j$ where $i,j \in \widehat{n}$. Using the methods of statistical physics, one can then derive the probability density for the momentum and positions in the form

$$f(p_1,\ldots,p_n,q_1,\ldots,q_n) = c\exp\left(-a\left[\sum_{i=1}^{n} p_i^2 + g^2\sum_{i\neq j}\frac{1}{f^2_{4\pi/\gamma}(q_i - q_j)}\right] - b\sum_{i=1}^{n}\cosh(\gamma q_i)\right). \tag{10}$$

where the second term in the exponent represents the confinement potential holding the repulsing particles together. The density above seems to be very cumbersome to work with. That is also why its approximation for the implementation purposes was proposed. Particularly, instead of considering positions to be random, they were chosen to take the values $q_i = i$ a.s. Using this adjusted distribution and setting $\gamma := 4\pi/n$, the matrix $\mathbf{L}$ then matches the one from hyperbolic DUE established in the definition 4.

In the paper [3], the authors also derive the joint distribution of the eigenvalues of the random matrix $\mathbf{L}$. Thanks to the integrability of the underlying system, it is possible to perform the canonical transformation of the momentum and the positions of the particles to the corresponding action-angle variables. As a matter of fact, the action ones turns out to be the eigenvalues of the Lax matrix $\mathbf{L}$. After the transformation of the density (10) and integration over the angle variables, we get the joint density for the eigenvalues in the form

$$f(\lambda_1, \ldots, \lambda_n) = c \exp\left(-a \sum_{i=1}^{n} \lambda_i^2\right) \prod_{i=1}^{n} K_0\left(b \prod_{i \neq j} \left|1 + \frac{\mathrm{i}g\gamma}{\lambda_i - \lambda_j}\right|\right) \tag{11}$$

where K_0 stands for the Macdonald's function of the zeroth order.

From the density above, it is possible to deduce the estimate of the level spacing distribution for the respective eigenvalues. First, let us investigate the character of the distribution for the high values of spacings. In that case the function (11) is mainly determined by its exponential part which corresponds to the distribution of the independent Gaussian random variables. Based on the theorem 2, the spacing after unfolding is of an exponential character. To deduce the behavior of the level spacing density around zero, it is necessary to use the approximation $K_0(t) \sim \sqrt{\pi/(2t)} \exp(-t)$ as $t \to \infty$. Plugging this into the expression (11) and combining it with the estimate for the high values of level spacing, we obtain

$$f_{S_1}(t) = ct^{\alpha} \mathrm{e}^{-\beta/t + dt} \tag{12}$$

for $t > 0$. The constant c normalizes the density and d is often determined by the condition $\mathrm{E}(S_1) = \mu$ applied for the purpose of comparing differently scaled spacings. As a matter of fact, this density determines the generalized inverse Gaussian (GIG) distribution. How well it fits to the data generated from the matrix in the definition (4) will be tested in the following section.

3 Level Spacing for Eigenvalues of DUE_h

In this section, we will thoroughly look at the repulsive interaction kind of dependencies governing in the spectra of the matrices established in the definition (4). Specifically, we are aiming to study the distribution of the spacings between two nearest eigenvalues. We will do so after the application of the rescaling transformation introduced in the theorem (1) and compare individual spacings throughout the whole spectra. For that purpose, the expected value

$$\mathrm{E}(N_{t,n}) = \sum_{k=1}^{n} F_{\Lambda_{k,n}}(t) \tag{13}$$

for $t \in \mathbb{R}$ is required. The sequence $(\Lambda_{k,n})_{k=1}^{n}$ represents the ordered version of the spectra of the matrix from $\mathrm{DUE}_\mathrm{h}(n, g)$.

The intensity function as the derivative of the expected value (13) becomes the density mixture of all the ordered eigenvalues. This mixture is sometimes also denoted as the eigenvalue density. It is well known that the eigenvalue density of the Wigner matrices is a semiellipse according to the famous Wigner semicircular law. In the case of damped matrices, the distribution certainly does not follow this behavior as can be seen in the figure 1. Instead, it gradually changes from the Gaussian distribution to almost uniform one as the parameter g increases. This trend was attempted to be captured in the paper

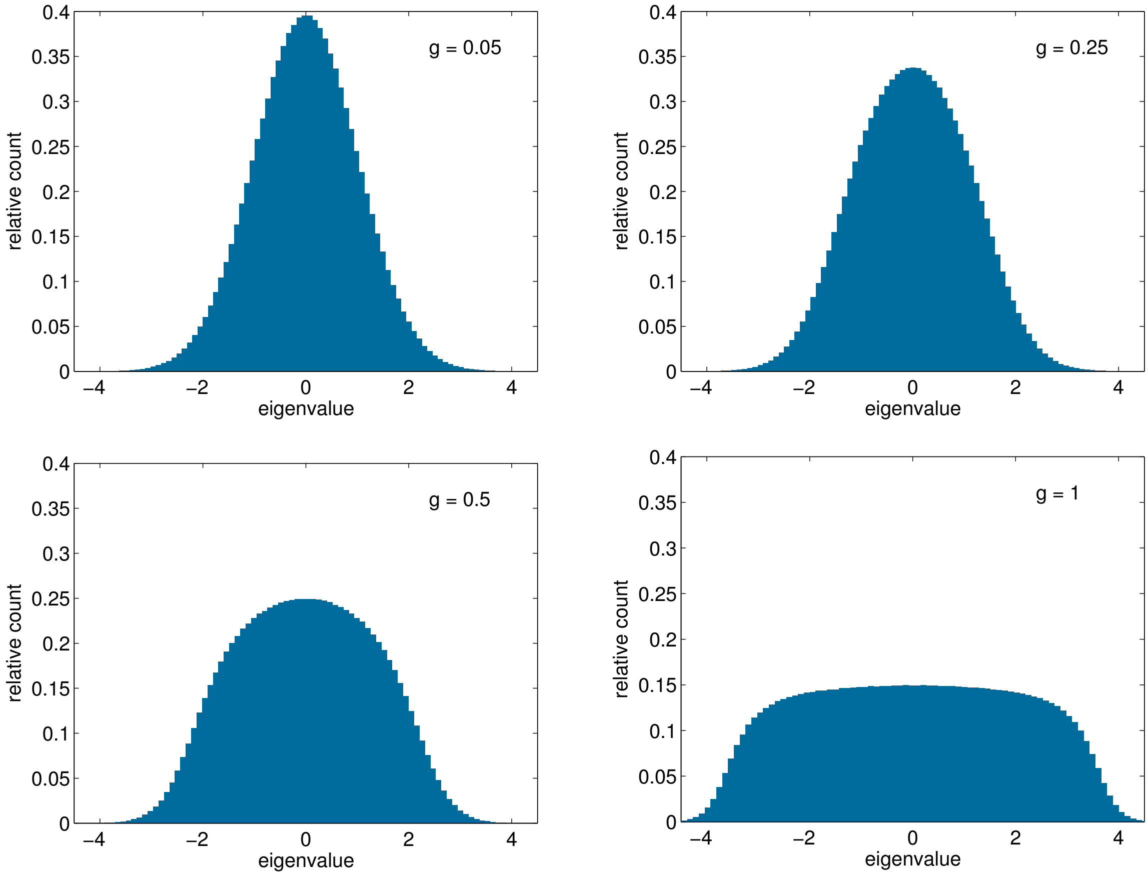


Figure 1: The eigenvalue densities of $\mathrm{DUE_h}$ matrices for various values of the parameter g.

[3] by the formula

$$f(t) = \frac{c(\vartheta)}{\varepsilon} \exp\left(\frac{\vartheta^2\varepsilon^2}{\epsilon^2 - t^2}\right) \tag{14}$$

for $|t| < \varepsilon$ and $f(t) = 0$ otherwise. The function $c(\vartheta)$ plays the role of a normalization factor while $\varepsilon, \vartheta > 0$ are the parameters of the distribution. The distribution function of (14) can be used as the decent approximation of the rescaling transformation $\mu\mathrm{E}(N_t, n)$. If no such a theoretical formula is available, a polynomial regression is usually performed to estimate the distribution function of eigenvalue density. Nevertheless, non of these strenuous approximative approaches will be needed in our case. Since we deal with random matrices, we can generate enough of them to precisely normalize the scale of all the spacings manually. By normalization, it is meant here to convert all the respective means to one. Note that this method might not be able to be used in the real systems as only one realization of the finite counting process is usually available. It is also sufficient only when dealing with nearest-neigbor spacing distributions. If one wants to study more advanced characteristics like multi-fold spacing or rigidity, the transformation (2) is necessary to be applied.

Let us now try to fit the spacing distributions by the guessed formula (12) in which the parameter d is determined the scaling condition $\mathrm{E}(S_{1,n}) = 1$. The fits for various positions of the spacings in the spectra are presented in the figure 2. Apparently, the distribution

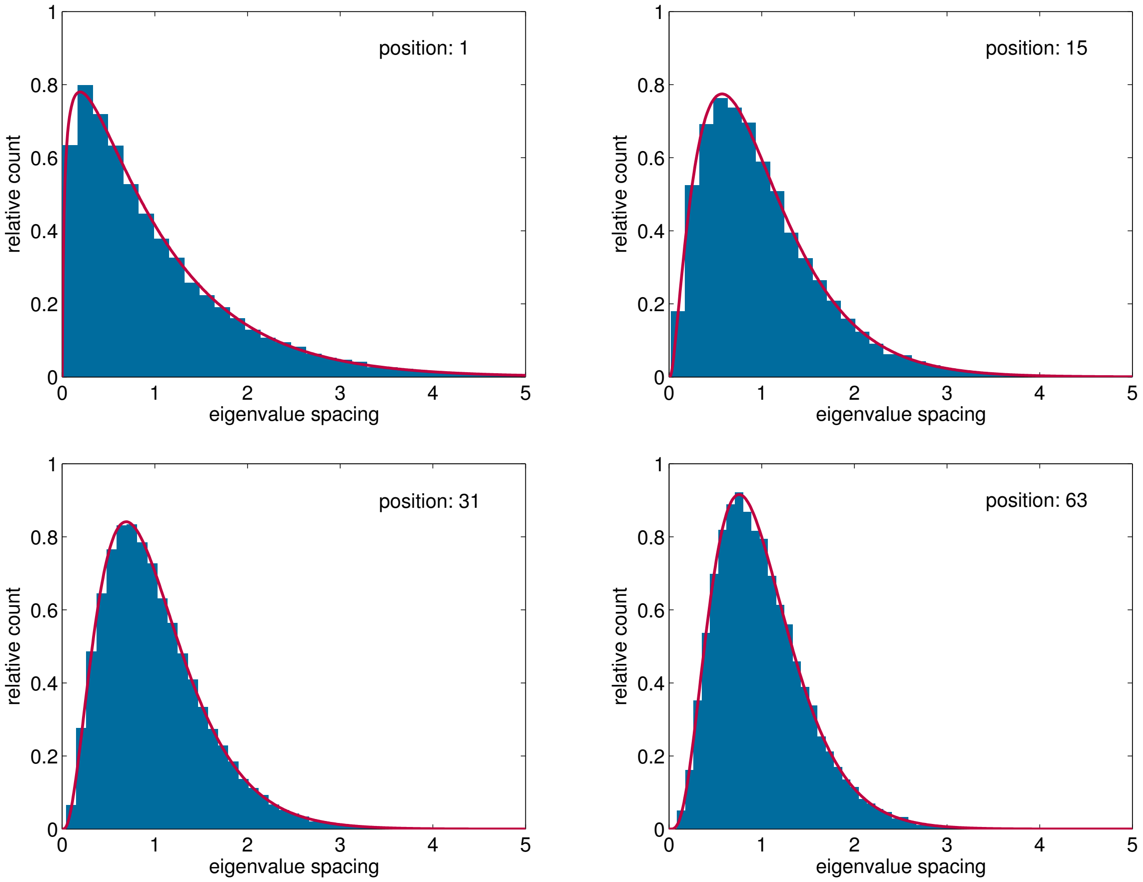


Figure 2: Histograms and fits of spacing densities between two nearest eigenvalues from $\mathrm{DUE_h}$ located in various parts of spectra.

changes significantly determined by the weak repulsion character on the edge of the spectra to quite strong one in the bulk. It is surprisingly different behavior than in the case of the well-known Gaussian random matrices whose eigenvalue spacing distributions appear to be identical no matter the position of the spacing in the spectra. As a result, the finite counting process formed by the eigenvalues of the matrices from $\mathrm{DUE_h}$ does not have an unfolding. Indeed, the application of the mapping (2) to the process would not result in a level counting one since the corresponding spacing distributions would not be the same.

Let us now have a look at the dependence of the spacing distribution on the position in the spectra more in detail. The figure 3 shows the estimates of the parameters α and β of the distribution (12) for all the nearest-neighbor spacings between the eigenvalues of the matrices from $\mathrm{DUE_h}(128, g)$ for various values of g. As expected from the symmetry of the graphs in figure 1, the change in the parameters α and β going from the edges of spectra to its bulk is symmetric as well. The trend of the change seems to be parabolic in the case of the parameter α. Despite the high variability in the estimates of the parameter β, their trend indicates to have a semicircular behavior.

The spectra of the matrices from $\mathrm{DUE_h}$ thus truly cannot be unfolded as a whole.

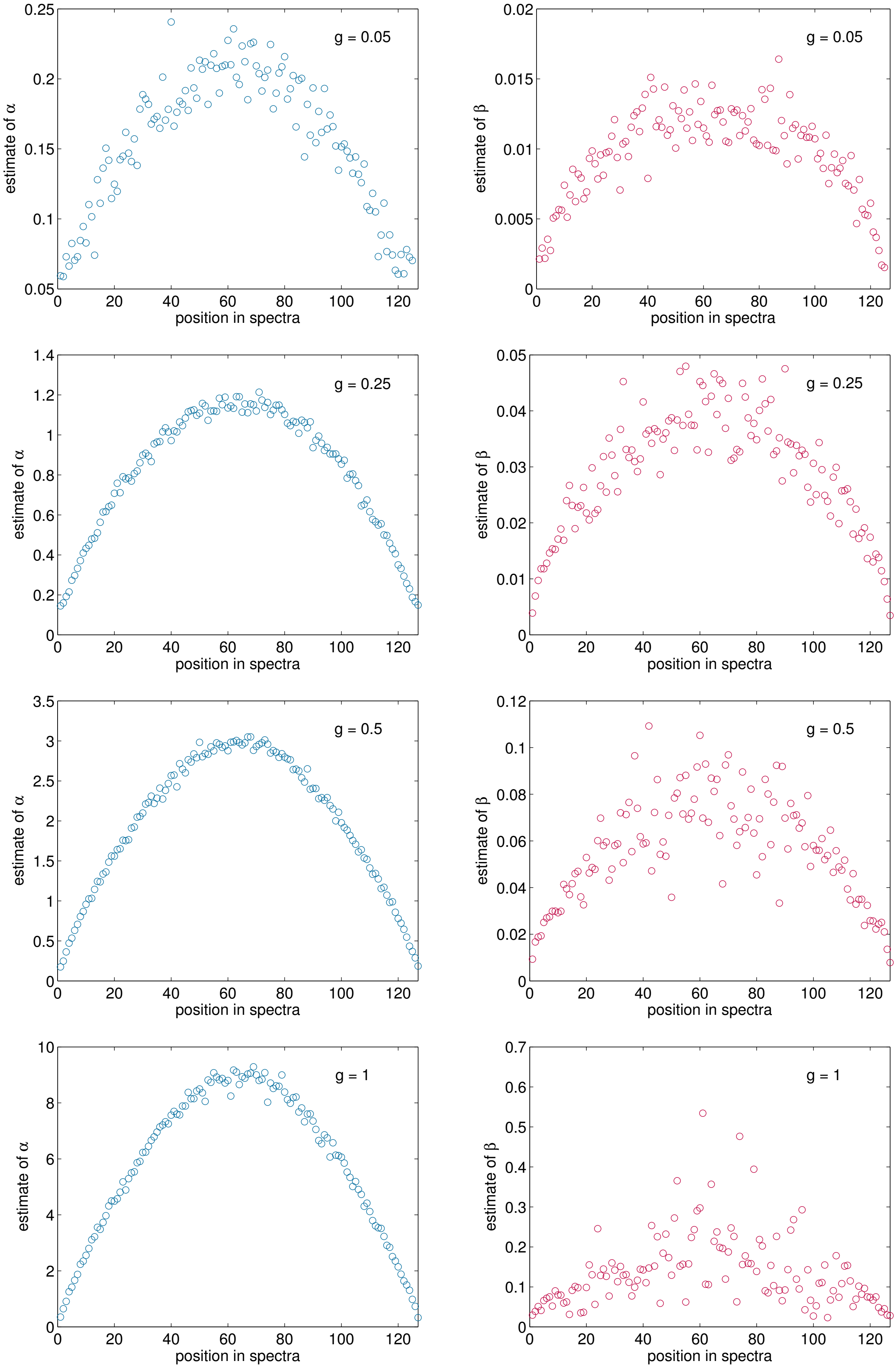


Figure 3: The estimates of the parameters α and β of the density (12) for the spacings of the nearest eigenvalues from $\mathrm{DUE}_{\mathrm{h}}$ with various values of the parameter g.

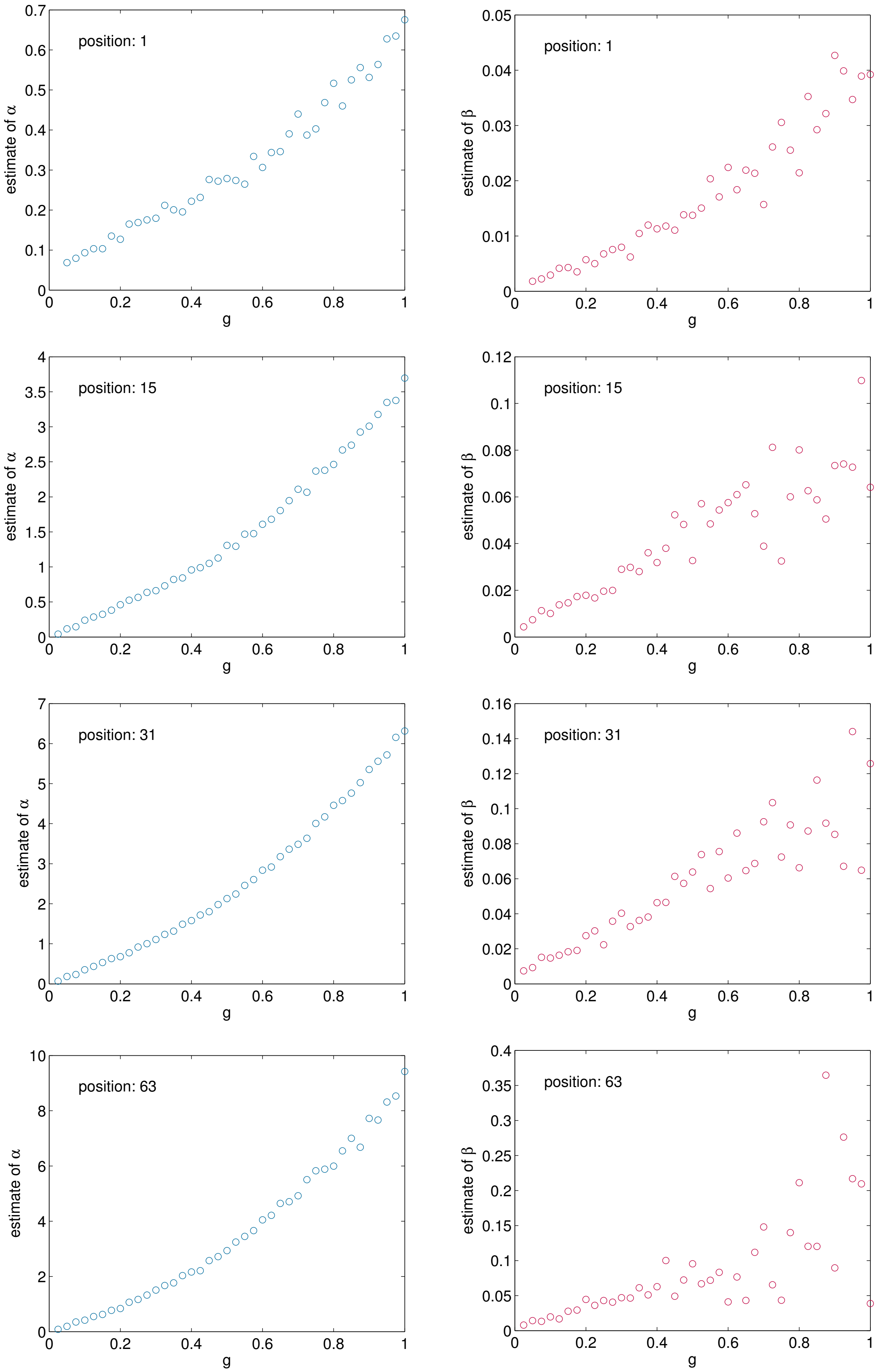


Figure 4: The estimates of the parameters α and β of the density (12) for the spacings of the nearest eigenvalues from $\mathrm{DUE_h}$ located in various parts of spectra.

Nevertheless, the estimates of the parameter α in roughly the middle quarter of the spectra appear to have a steady trend. Considering only those eigenvalues, the unfolding could be performed on corresponding counting process. However, the theoretical prediction (14) does not fit very well this time and the mentioned polynomial regression method has to be used instead.

As a matter of fact, it is again possible to bypass the method using polynomial regression. Setting a threshold h and taking only those eigenvalues $(\Lambda_i)_{i=1}^{128}$ from the middle quarter of the spectra ($i \in \{48, \dots, 80\}$) satisfying $|\Lambda_i| < h$ performs the approximate unfolding[1] as well. The threshold h is chosen with respect to the estimates of the expected values $\mathrm{E}(\Lambda_{48})$ and $\mathrm{E}(\Lambda_{80})$.

So far, we have investigated the spacing distribution in the dependence of the location in the spectra of the matrices from $\mathrm{DUE_h}(128, g)$. Let us now have a look at the dependence of the distribution on the parameter g more thoroughly.

In the figure 4, the respective estimates $\widehat{\alpha}(g)$ and $\widehat{\beta}(g)$ are plotted for various locations of the spacings in the spectra. In the case of parameter α, its estimates seem to have a quadratically increasing trend while those of the parameter β indicate possible linear trend.

References

[1] O. Kollert, *Analysis of Random Matrix Spectra Using Counting Process Theory (diploma thesis)*, ČVUT v Praze, Fakulta jaderná a fyzikálně inženýrská, Praha (2015)

[2] O. Kollert, M. Krbalek, T. Hobza *Level Counting Process Theory*, not published yet

[3] E. Bogomolny, O. Giraud, C. Schmit, *Integrable random matrix ensembles*, Nonlinearity 24 (2011)

[4] T. Hobza, M. Krbalek *Inner Structure of Vehicular Ensembles and Random Matrix Theory*, Physics Letters A 380 (2016)

[1] This procedure is actually used quite frequently in the random matrix theory and it is called the simple unfolding.

Gaussian–Hermite Moments in Object Recognition[*]

Jitka Kostková

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

advisor: Jan Flusser, Department of Image Processing
Institute of Information Theory and Automation, CAS

Abstract. Object recognition is a process for identifying objects in images or video sequences. One of the powerful tools in object recognition is an invariant description of objects. The descriptors ought to be computationally stable and have high discriminative power. Hence, invariants constructed from orthogonal Gaussian–Hermite moments can be used advantageously. Gaussian–Hermite (GH) moments play a special role among various orthogonal moments [1, 2, 3, 5, 6, 8, 10]. They were proved to be very robust w.r.t. additive noise comparing to other common moments [4, 7]. The GH moments are the only moments orthogonal on a rectangle which offer a possibility of an easy and efficient design of rotation invariants. This is guaranteed by the Yang's Theorem [9]. However, the construction of invariants w.r.t. scaling cannot be accomplished easily and a novel approach is needed.

The first paper is concerned with invariants with respect to scaling constructed from Gaussian–Hermite moments. The invariance is achieved owing to modulation of Gaussian–Hermite polynomials using variable parameter σ that depends on the input image. The scale invariance can be easily coupled with the rotation invariance. This approach can be effortlessly applied in 2D and 3D with high numerical stability as demonstrated in experiments on real data.

The second paper is dealing with rotation invariants of vector fields. Vector field images are a new type of data appearing in many engineering areas in the last few years. A 2D vector field $\mathbf{f}(\mathbf{x})$ can be mathematically described as a pair of scalar fields (images) $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$. At each point $\mathbf{x} = (x, y)$, the value of $\mathbf{f}(\mathbf{x})$ show the orientation and the magnitude of a certain vector. Hence, it is necessary to develop new methods and algorithms for dealing with this type of data. In this paper, we propose a method for the description and matching of 2D vector field patterns under an unknown rotation of the field. The considered rotation of a vector field is so-called total rotation, where the action is applied not only on the spatial coordinates but also on the field values. Invariants of vector fields with respect to total rotation constructed from Gaussian–Hermite moments orthogonal on a square and Zernike moments orthogonal on a disk are introduced. Their numerical stability is shown to be better than that of the geometric/complex moment invariants. We demonstrate their usefulness in a real world template matching application of rotated vector fields – a vortex detection in a fluid flow.

Keywords: Scale invariants, Variable modulation, Normalization, Vector field, Total rotation, Invariants, Gaussian-Hermite moments, Zernike moments, Numerical stability.

Abstrakt. Rozpoznávání objektů je proces identifikace objektů v obraze či videu. Jedním z přístupů je použití deskriptorů objektů, které jsou invariantní vůči jistým typům transformací

[*] This work has been supported by grants No. GA15–16928S and SGS15/214/OHK4/3T/14.

v obraze. Výpočet těchto deskriptorů by měl být numericky stabilní a měly by mít vysokou diskriminabilitu. S výhodou lze proto pro jejich konstrukci využít ortogonálních Gaussových–Hermitových (GH) momentů. Tyto momenty hrají důležitou roli mezi ortogonálními momenty [1, 2, 3, 5, 6, 8, 10]. Bylo dokázáno, že GH momenty jsou velmi robustní vůči aditivnímu šumu ve srovnání s jinými běžně používanými momenty [4, 7]. GH momenty jsou jediné momenty ortogonální na obdélníku, ze kterých lze snadno zkonstruovat rotační invarianty. Což je možné díky *Yangově větě* [9]. Bohužel rozšíření na invarianty vůči škálování je netriviální a je třeba zvolit nový přístup.

První z uvedených článků pojednává o invariantech vůči škálování konstruovaných pomocí Gaussových–Hermitových momentů. Invariance je dosaženo díky modulaci Gaussových–Hermitových polynomů proměnným parametrem σ, který závisí na vstupním obrázku. Invariance vůči škálování může být snadno kombinována s invariancí vůči rotaci. Tento přístup lze jednoduše použít jak pro dvourozměrná tak i pro třírozměrná data. Numerická stabilita výpočtů je demonstrována na experimentech s reálnými daty.

Druhý článek se zabývá rotačními invarianty pro vektorová pole. V posledních letech se díky novým způsobům měření a novým typům měřících zařízení setkáváme stále častěji s multidimenzionálními vektorovými poli. 2D vektorové pole $\mathbf{f}(\mathbf{x})$ lze matematicky popsat jako uspořádanou dvojici skalárníích obrázků $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$. V každém bodě $\mathbf{x} = (x, y)$, popisuje hodnota $\mathbf{f}(\mathbf{x})$ velikost a směr daného vektoru. Je proto potřeba k jejich analýze vyvíjet speciální metody a algoritmy či významně modifikovat stávající postupy z tradiční oblasti zpracování obrazu. V tomto článku navrhujeme metodu pro popis a vyhledávání vzorů ve 2D vektorových polích při neznámé rotaci pole. Uvažovaná rotace je tzv. totální rotace, kdy transformace nepůsobí pouze na prostorové souřadnice, ale také na hodnoty pole. Dále představujeme invarianty vektorových polí vzhledem k totální rotaci zkonstruované pomocí Gaussových–Hermitových momentů ortogonálních na čtverci a Zernikeových momentů ortogonálních na kruhu. Ukážeme, že numerická stabilita těchto invariantů je vyšší než stabilita invariantů založených na geometrických/komplexních momentech. Užitečnost těchto invariantů demonstrujeme na reálné problému – detekci vírů v proudění kapalin.

Klíčová slova: Invarianty vůči škálování, proměnná modulace, normalizace, vektorové pole, totální rotace, invarianty, Gaussovy–Hermitovy momenty, Zernikeovy momenty, numerická stabilita.

Full papers:

- B. Yang, J. Kostková, J. Flusser and T. Suk, *Scale-invariants from Gaussian-Hermite Moments*, Signal Processing **132** (2017), 77–84.
- B. Yang, J. Kostková, J. Flusser, T. Suk and R. Bujack, *Rotation Invariants of Vector Fields from Orthogonal moments*, Pattern Recoginition **74** (2018), 110–121.

References

[1] K. M. Hosny. *New set of rotationally Legendre moment invariants.* International Journal of Electrical, Computer, and Systems Engineering **4** (2010), 176–180.

[2] A. Khotanzad and Y. H. Hong. *Invariant image recognition by Zernike moments.* IEEE Transactions on Pattern Analysis and Machine Intelligence **12** (1990), 489–497.

[3] Y. Li. *Reforming the theory of invariant moments for pattern recognition.* Pattern Recognition **25** (1992), 723–730.

[4] J. Shen, W. Shen, and D. Shen. *On geometric and orthogonal moments.* In 'Multispectral Image Processing and Pattern Recognition', J. Shen, P. S. P. Wang, and T. Zhang, (eds.), volume 44 of *Machine Perception Artificial Intelligence*, 17–36. World Scientific Publishing, (2001).

[5] M. R. Teague. *Image analysis via the general theory of moments.* Journal of the Optical Society of America **70** (1980), 920–930.

[6] Å. Wallin and O. Kübler. *Complete sets of complex Zernike moment invariants and the role of the pseudoinvariants.* IEEE Transactions on Pattern Analysis and Machine Intelligence **17** (1995), 1106–1110.

[7] L. Wang, Y. Wu, and M. Dai. *Some aspects of Gaussian-Hermite moments in image analysis.* In 'Third International Conference on Natural Computation ICNC'07', L. P. Suresh, S. S. Dash, and B. K. Panigrahi, (eds.), volume 2 of *Advances in Intelligent Systems and Computing*, 450–454. IEEE, (2007).

[8] B. Yang and M. Dai. *Image analysis by Gaussian–Hermite moments.* Signal Processing **91** (2011), 2290–2303.

[9] B. Yang, G. Li, H. Zhang, and M. Dai. *Rotation and translation invariants of Gaussian–Hermite moments.* Pattern Recognition Letters **32** (2011), 1283–1298.

[10] P.-T. Yap, R. Paramesran, and S.-H. Ong. *Image analysis by Krawtchouk moments.* IEEE Transactions on Image Processing **12** (2003), 1367–1377.

Random Walks with Varying Transition Probabilities

Tomáš Kouřim

4th year of PGS, email: kourim@outlook.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Volf, Department of Stochastic Informatics
Institute of Information Theory and Automation, CAS

Abstract. Random walk is a very well studied object. Since its first introduction by Pearson in 1905 a number of alternative models have been developed. This paper presents a novel approach to a random walk with memory. This memory is introduced by a varying transition probability. Asymptotic properties of such a random walk are described and possible real life applications of such model are introduced.

Keywords: Random walk, memory, varying transition probability

Abstrakt. Náhodná procházka je objekt studovaný více než sto let. Od roku 1905, kdy Pearson poprvé koncept náhodné procházky představil, byla vyvinuta celá řada alternativ k původnímu modelu. Tento článek se věnuje náhodné procházce s pamětí, jež se projevuje proměnlivou přechodovou pravděpodobností. Jsou zkoumány asymptotické vlastnosti takovéto náhodné procházky a naznačena možná použití modelu v praxi.

Klíčová slova: Náhodná procházka, paměť, proměnlivá pravděpodobnost

1 Introduction

Random walks has been subject to extensive study for over a hundred years since they were first introduced in by Pearson in 1905 [1]. Since then, many different variations of a random walk have been introduced. Those variations usually involve different supports (i.e. a random walk on a lattice, graph, finite set) and time properties (discrete or continuous) [3]. Many variations also involve a memory factor added into the random walk, such as self-avoiding walk or reinforced random walk [2]. Introducing a long term memory factor into a random walk leads to a very different asymptotic behavior.

In this paper one-dimensional random walk is considered, in which the position of the walker is controlled by a varying transition probabilities. After each step, the probability that the next step will be in the same direction as the previous one is lowered and the probability that the walker will move in opposite direction is increased accordingly. The transition probabilities evolve in time in a random way and the actual values of the transition probability depend on the entire history, making the walk a non-Markovian stochastic process.

The model is described in the next section and section 3 indicates possible evolution of this theory and concludes this paper.

2 Model

2.1 Previous work

Similar problem has been studied by Turban in [4]. In the paper, the discrete time one-dimensional random walk with the following properties is studied. The step size l_t^+ of the $t-th$ step to the right and step size l_t^- of the $t-th$ step to the left satisfy the condition

$$l_t^+ + l_t^- = 2 \ \forall t$$

and the size of the step is evolving according to the following rules for $t > 1$

$$\sigma_{t-1} = +1 \rightarrow \begin{cases} l_t^+ = \lambda l_{t-1}^+ \\ l_t^- = 2 - \lambda l_{t-1}^+ \end{cases}$$

$$\sigma_{t-1} = -1 \rightarrow \begin{cases} l_t^+ = 2 - \lambda l_{t-1}^- \\ l_t^- = \lambda l_{t-1}^- \end{cases}$$

where the Ising variable $\sigma_i = \pm 1$ with equal probability $p = \frac{1}{2}$, $l_1^+ = l_1^- = 1$ and $0 \leq \lambda \leq 1$. The limit $\lambda = 1$ corresponds with the Bernoulli random walk, the limit $\lambda = 0$ corresponds to a situation when the walker does not move for some time. Turban shows that such a random walk is well controlled and that it is non-diffusive (with Hurst exponent of the mean square displacement $\alpha = 0$) even for λ close to 1.

2.2 The Model

In this paper slightly different approach is considered. Let's take a random walk on integers, with step size $l_t \in \{-1, 1\}$. The probability that in time t the step will be positive is

$$P(l_t = 1) = p_t^+$$

and the probability that the step will be negative is

$$P(l_t = -1) = p_t^- = 1 - p_t^+.$$

The transition probabilities vary in time such that the probability of moving in the same direction as in previous step is lowered by a coefficient $\lambda \in (0, 1)$

$$p_t^+ = \begin{cases} \lambda p_{t-1}^+ & l_{t-1} = 1 \\ 1 - \lambda p_{t-1}^- & l_{t-1} = -1 \end{cases} \tag{1}$$

$$p_t^- = \begin{cases} 1 - \lambda p_{t-1}^+ & l_{t-1} = 1 \\ \lambda p_{t-1}^- & l_{t-1} = -1 \end{cases} \tag{2}$$

As there always holds that $p_t^- = 1 - p_t^+$, it is sufficient to further only consider p_t^+. Let $p_t = p_t^+$. From equations 1 and 2 follows for $t > 1$ that

$$p_t = \begin{cases} \lambda p_{t-1} & l_{t-1} = 1 \\ 1 - \lambda + \lambda p_{t-1} & l_{t-1} = -1 \end{cases} \tag{3}$$

$$p_t = \lambda p_{t-1} + \frac{1}{2}(1-\lambda)(1-l_{t-1}) \tag{4}$$

Let's calculate the expression for p_t using induction. First let's assume $p_1 = \frac{1}{2}$. For $t = 2$ expression 4 gives

$$p_2 = \frac{1}{2}\lambda + \frac{1}{2}(1-\lambda)(1-l_1). \tag{5}$$

For any t let's assume that

$$p_t = \frac{1}{2}\lambda^{t-1} + \frac{1}{2}(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-1-i}(1-l_i) \tag{6}$$

This holds for $t = 2$ (5). For $t = t+1$ we get

$$p_{t+1} = \lambda(\frac{1}{2}\lambda^{t-1} + \frac{1}{2}(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-1-i}(1-l_i)) + \frac{1}{2}(1-\lambda)(1-l_t)$$

$$p_{t+1} = \frac{1}{2}\lambda^{t} + \frac{1}{2}(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-i}(1-l_i) + \frac{1}{2}(1-\lambda)(1-l_t)$$

$$p_{t+1} = \frac{1}{2}\lambda^{t} + \frac{1}{2}(1-\lambda)\sum_{i=1}^{t}\lambda^{t-i}(1-l_i)$$

which is in accordance with 6 and thus 6 holds for any $t > 1$. Since

$$(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-1-i} = (1-\lambda)\sum_{i=0}^{t-2}\lambda^{i} = (1-\lambda)\frac{1-\lambda^{t-1}}{1-\lambda} = 1-\lambda^{t-1}$$

expression 6 can be reduced to

$$p_t = \frac{1}{2}\lambda^{t-1} + \frac{1}{2}(1-\lambda^{t-1}) - \frac{1}{2}(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-1-i}l_i \tag{7}$$

$$p_t = \frac{1}{2}(1-(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-1-i}l_i). \tag{8}$$

Proposition 1. *For $p_1 = \tilde{p}$, $\forall t \geq 1$ it holds*

$$p_t = \tilde{p}\lambda^{t-1} + \frac{1}{2}(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-1-i}(1-l_i),$$

which can be expressed as

$$p_t = (\tilde{p}-\frac{1}{2})\lambda^{t-1} + \frac{1}{2}(1-(1-\lambda)\sum_{i=1}^{t-1}\lambda^{t-1-i}l_i).$$

Proof. Follows directly from 4, 5 and 7. □

Examples of the random walk with memory in probability as well as the model introduced in [4] and the standard random walk can be seen in Figures 1 and 2. It can be seen that the memory coeficient on probability does not limit the position of the walker as much as the step length memory coeficient, but it still significantly affects the random walk development.

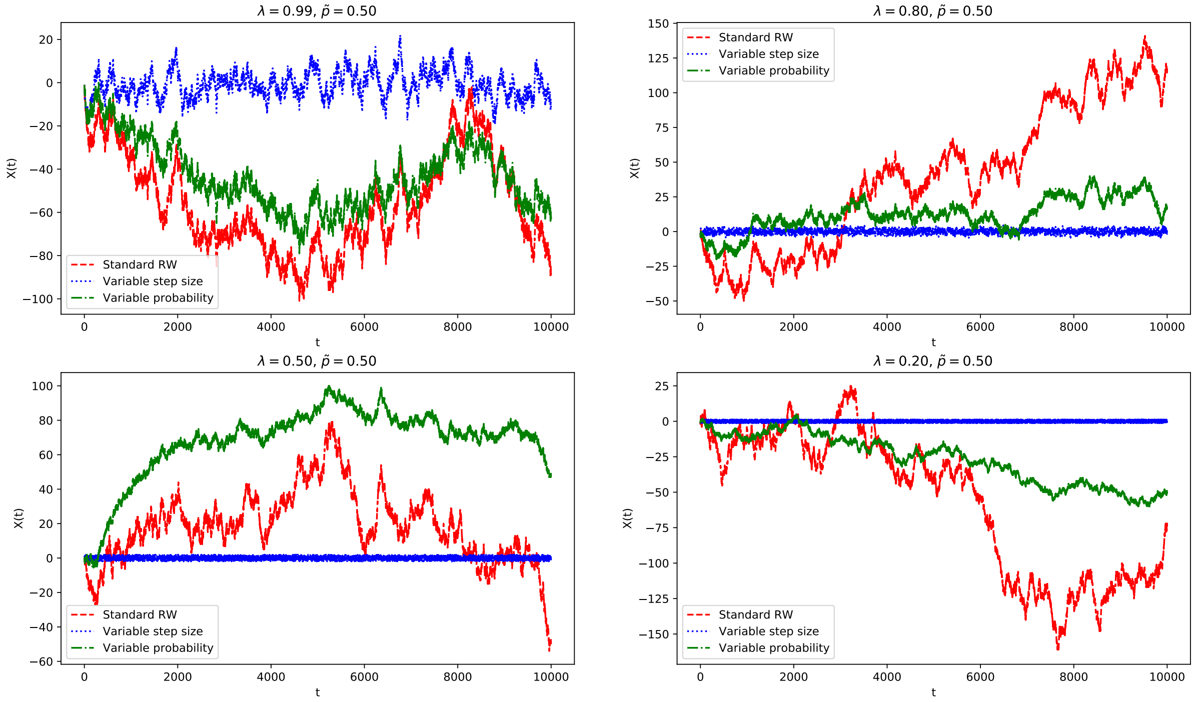


Figure 1: Comparison of random walks generated by the same random numbers for starting probability $\tilde{p} = 0.5$ and varying memory factor λ. Red dashes are the standard random walk, blue dots are random walks with memory introduced by Turban and green dash-dot lines represent random walks generated by the model introduced in this paper.

2.3 Mean values of the process

Let X_t be the position of the walker at time t. It holds that

$$X_t = X_{t-1} + l_t$$

To calculate the expected value EX_t it holds that

$$EX_t = EX_{t-1} + El_t \tag{9}$$

with

$$El_t = 2Ep_t - 1. \tag{10}$$

The expected transition probability Ep_t at time t can be calculated as

$$Ep_t = (Ep_{t-1})^2\lambda + (1 - Ep_{t-1})(1 - (1 - Ep_{t-1})\lambda)$$

$$Ep_t = 2Ep_{t-1}\lambda - \lambda - Ep_{t-1} + 1. \tag{11}$$

Proposition 2. *For $\forall t \geq 1$, it holds that*

$$Ep_t = (2\lambda - 1)^{t-1}\tilde{p} + \frac{1 - (2\lambda - 1)^{t-1}}{2}. \tag{12}$$

Proof. For $t = 1$, equation 12 yields

$$Ep_1 = \tilde{p} \tag{13}$$

and for $t = 2$

$$Ep_2 = (2\lambda - 1)\tilde{p} + \frac{1 - 2\lambda + 1}{2} = 2\tilde{p}\lambda - \lambda - \tilde{p} + 1, \tag{14}$$

which is in accordance with 11. For $t = t + 1$ we get from 11

$$Ep_{t+1} = 2[(2\lambda - 1)^{t-1}\tilde{p} + \frac{1 - (2\lambda - 1)^{t-1}}{2}]\lambda - \lambda - [(2\lambda - 1)^{t-1}\tilde{p} + \frac{1 - (2\lambda - 1)^{t-1}}{2}] + 1$$

$$Ep_{t+1} = 2\lambda(2\lambda - 1)^{t-1}\tilde{p} - \lambda(2\lambda - 1)^{t-1} - (2\lambda - 1)^{t-1}\tilde{p} - \frac{1 - (2\lambda - 1)^{t-1}}{2} + 1$$

$$Ep_{t+1} = (2\lambda - 1)^{t-1}\tilde{p}(2\lambda - 1) + \frac{-2\lambda(2\lambda - 1)^{t-1} - 1 + (2\lambda - 1)^{t-1} + 2}{2}$$

$$Ep_{t+1} = (2\lambda - 1)^{t}\tilde{p} + \frac{1 - (2\lambda - 1)^{t}}{2}$$

and thus 12 holds for all $t \geq 1$. □

Proposition 3. *For $\forall t \geq 1$, it holds that*

$$EX_t = (2\tilde{p} - 1)\frac{1 - (2\lambda - 1)^t}{2(1 - \lambda)}. \tag{15}$$

Proof. For $t = 1$ equations 9, 10 and 13 yield (given $X_0 = 0$, i.e. the walker starts at the beginning)

$$EX_1 = 2\tilde{p} - 1$$

and for $t = 2$ (using 14)

$$EX_2 = 2\tilde{p} - 1 + 2(2\tilde{p}\lambda - \lambda - \tilde{p} + 1) - 1$$

$$EX_2 = 2\lambda(2\tilde{p} - 1),$$

which is the same as the result when using 15. Assuming 15 holds for t we get for $t = t+1$ from 9, 10 and 12

$$EX_{t+1} = EX_t + 2Ep_{t+1} - 1$$

$$EX_{t+1} = (2\tilde{p} - 1)\frac{1 - (2\lambda - 1)^t}{2(1 - \lambda)} + 2[(2\lambda - 1)^t\tilde{p} + \frac{1 - (2\lambda - 1)^t}{2}] - 1$$

$$EX_{t+1} = (2\tilde{p} - 1)(\frac{1 - (2\lambda - 1)^t}{2(1 - \lambda)} + (2\lambda - 1)^t)$$

$$EX_{t+1} = (2\tilde{p} - 1)(\sum_{i=0}^{t-1}(2\lambda - 1)^i + (2\lambda - 1)^t)$$

$$EX_{t+1} = (2\tilde{p} - 1)\frac{1 - (2\lambda - 1)^{t+1}}{2(1 - \lambda)}.$$

□

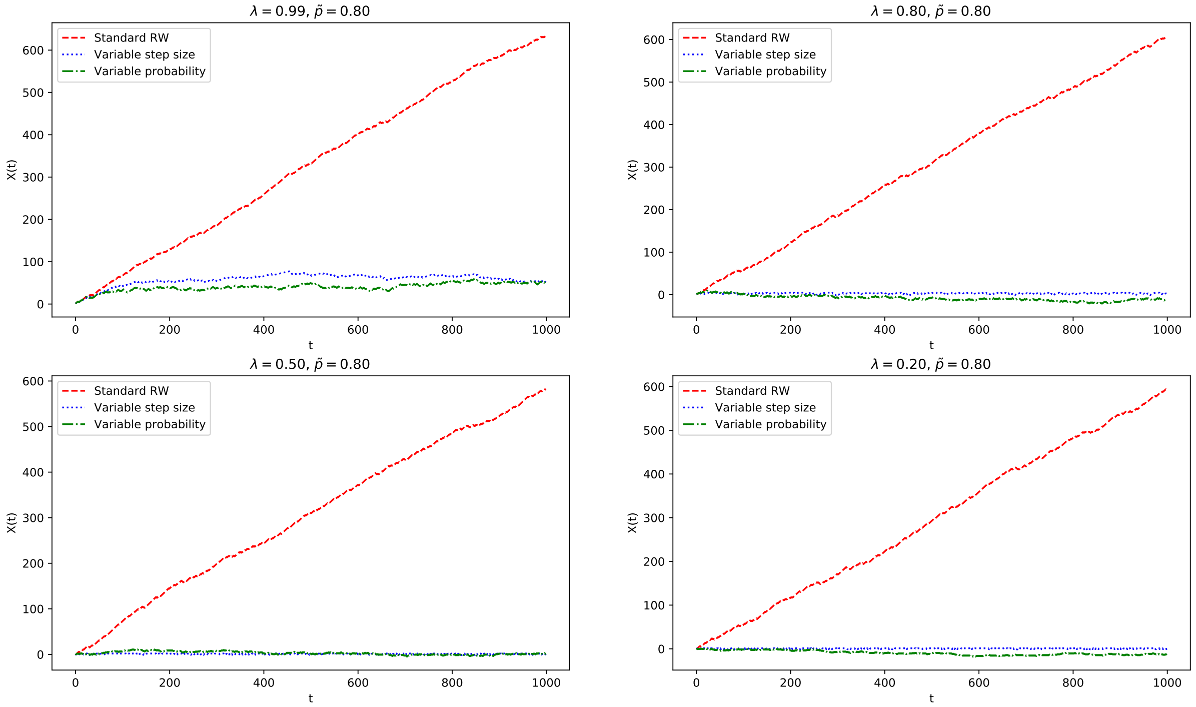


Figure 2: Comparison of random walks generated by the same random numbers for different starting probability $\tilde{p} = 0.8$ and varying memory factor λ.

2.4 Asymptotic behavior

Now let's examine the situation for $t \to \infty$. From Proposition 2 follows that

$$Ep_t \underset{t\to\infty}{=} \frac{1}{2}$$

and from Proposition 3 that

$$EX_t \underset{t\to\infty}{=} \frac{2\tilde{p}-1}{2(1-\lambda)}.$$

In other words the memory introduced by the coefficient λ will in long term eliminate the effect of the starting probability $\tilde{p}$ and drag the transition probability to the value of $\frac{1}{2}$. In a similar manner, the expected position of the walker will remain constant in the long run, at the position given by

$$\frac{2\tilde{p}-1}{2(1-\lambda)}.$$

2.5 Monte Carlo simulations

Monte Carlo simulations have been used to explore the asymptotic properties of the random walk with variable transition probability and to compare it to the standard random walk and to the random walk with memory introduced by Turban [4]. Figure 3 shows the expected position of the walker for the different types of random walk[1]

[1] The case when $\tilde{p} = 0.5$ is trivial, as all three types converge to 0. The standard random walk diverges for $\forall \tilde{p} \neq 0.5$ and is thus not showed.

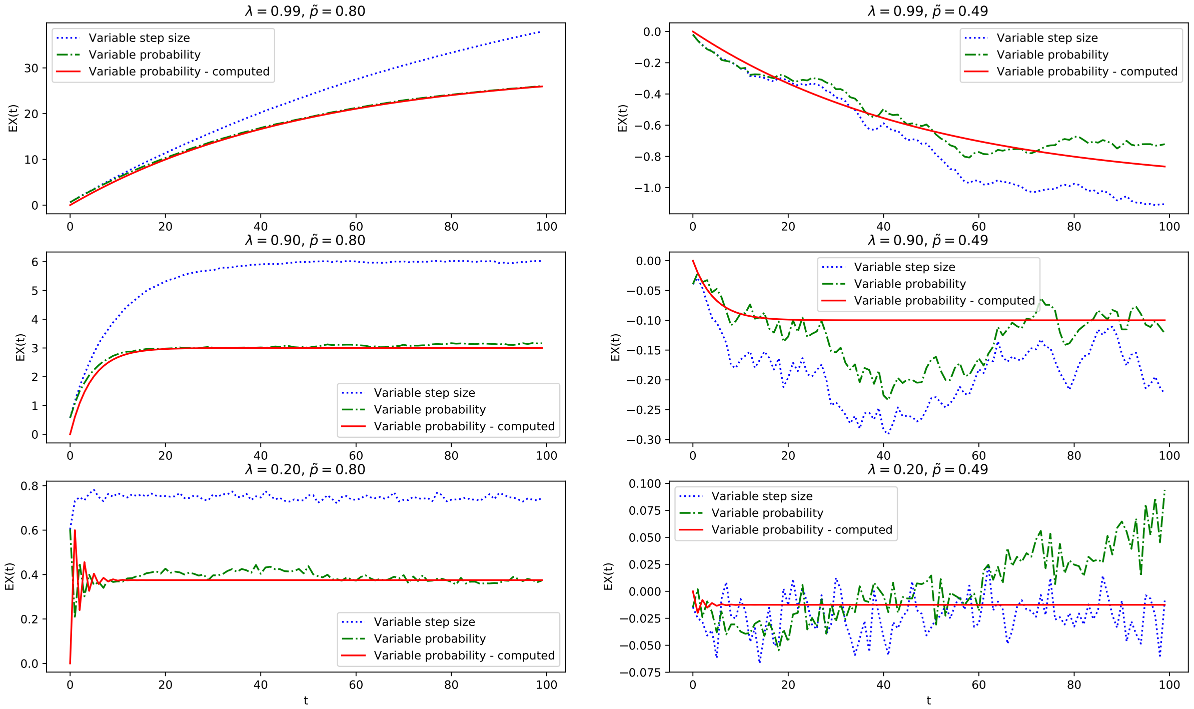


Figure 3: Comparison of the expected position of the walker for different starting probabilities $\tilde{p}$ and memory factor λ. Blue dots are random walks with memory introduced by Turban, green dash-dot lines represent random walks generated by the model introduced in this paper and red line is the computed value of EX_t given by Proposition 3.

and different values of starting probability $\tilde{p}$ and memory factor λ together with the expected position of the walker given by Proposition 3. The expected values of transition probabilities for different $\tilde{p}$ and λ, both theoretical and observed, can be seen in Figure 4.

Finally, Figures 5 and 6 show the observed variance of the walker position and the transition probabilities $Var(X_t)$ and $Var(p_t)$. These observations suggest that the variance of transition probability converges and does not depend on the initial probability $\tilde{p}$ and the variance of the position of the walker diverges linearly with respect to both $\tilde{p}$ and λ.

3 Conclusion

In this paper, a novel approach to a random walk with memory was introduced and the basic properties of such random walk were derived. Asymptotic properties were also demonstrated using Monte Carlo simulations. It seems that there are many real life application of such a model. The evolution of the score in some sports seems to follow the rules introduced in this paper. The (out)performance of a new worker in a company or the reliability of a machine could be another examples of real life applications of the introduced model. However, further research has to be conducted to prove these assumptions.

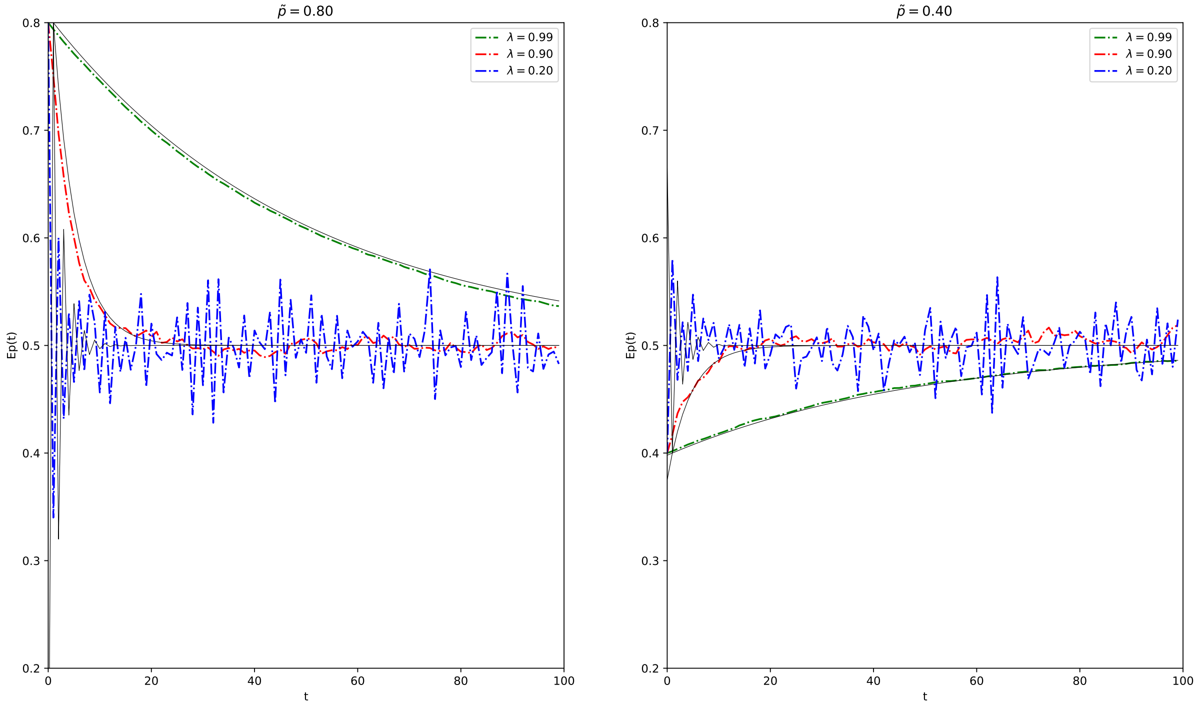


Figure 4: The expected values of transition probabilities. The dot-dashed lines are observed probabilities, the thin black lines their respective theoretical values.

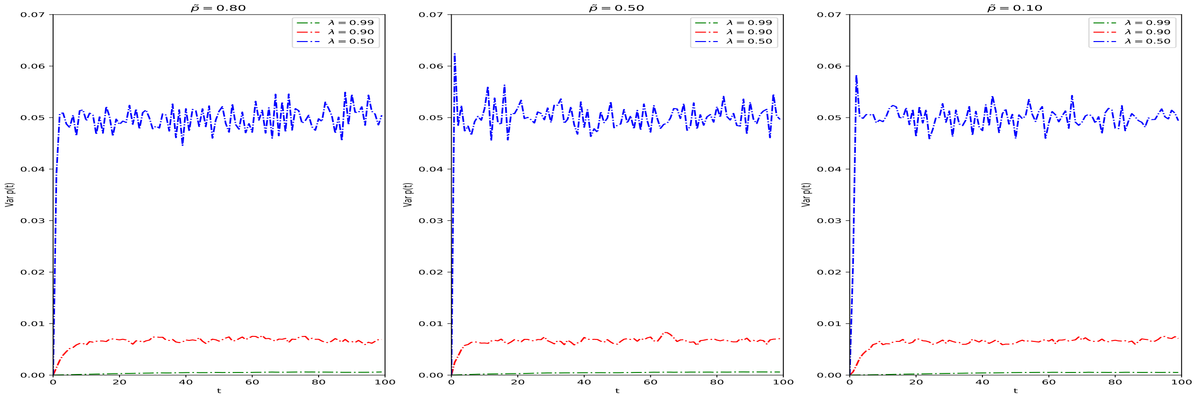


Figure 5: Variance of the transition probabilities.

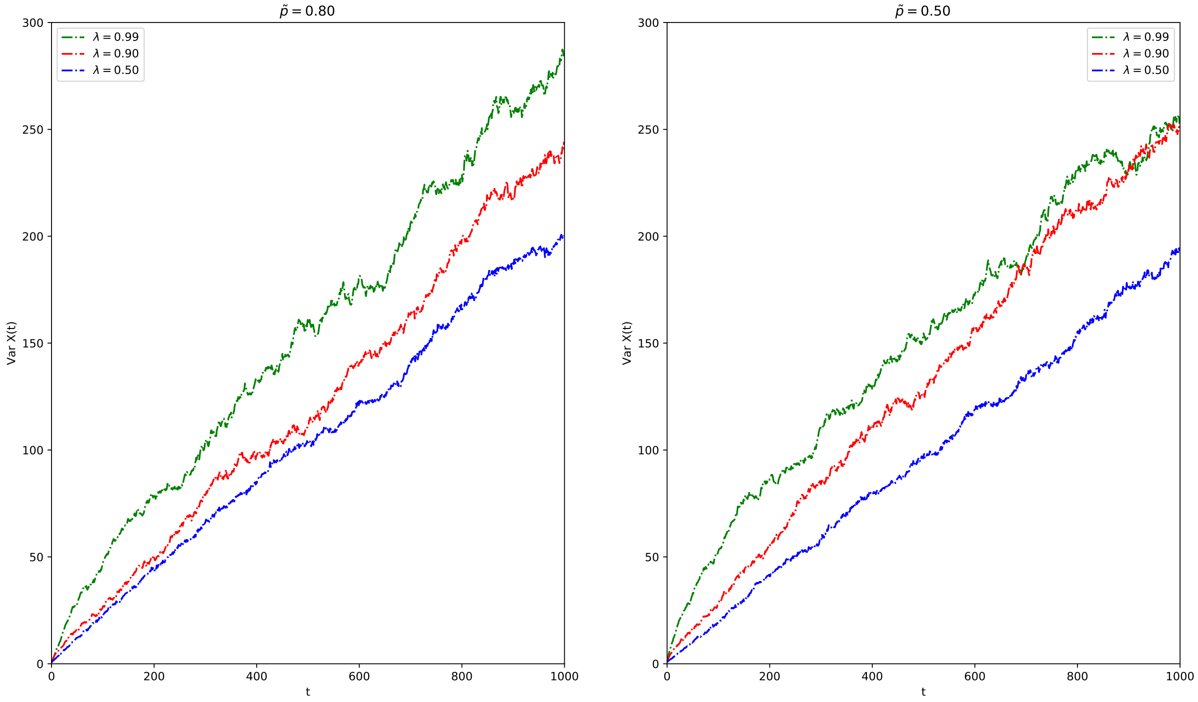


Figure 6: Variance of the position of the walker.

References

[1] Karl Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.

[2] Ryszard Rudnicki and Marek Wolf. Random walk with memory. *Journal of Mathematical Physics*, 40(6):3072–3083, 1999.

[3] Frank Spitzer. *Principles of random walk*, volume 34. Springer Science & Business Media, 2013.

[4] Loïc Turban. On a random walk with memory and its relation with markovian processes. *Journal of Physics A: Mathematical and Theoretical*, 43(28):285006, 2010.

Phase Transition of Chaotic Dynamics in Quantum Purification*

Martin Malachov

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

advisors:
Igor Jex, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Tamás Kiss, Institute for Solid State Physics and Optics
Wigner Research Centre for Physics, Hungarian Academy of Sciences

Abstract. Latest works show that quantum physics allows for new type of chaotic behaviour without analogy in classical physics. This chaos is connected to quantum description of physical state which is subject to nonlinear operation. The chaos has been analytically described in particular set of pure two-qubit states subject to a particular protocol. We aim on investigating chaotic evolution of mixed states which is beyond contemporary knowledge. We work with single-qubit version of protocol originally designed to purify quantum entanglement. We reveal a new phenomenon, half-attractiveness of quantum physical states. Our main result lies in concept of box-counting dimension which is used to characterise structures of chaotic mixed states. We show that the structure undergoes a phase transition where the purity of states plays the role of temperature. These sudden qualitative changes of the structures are very surprising. Finally, we also give quantitative characteristics of basins of attraction which indicate that number of states that can be purified by the protocol explodes exponentially with growing purity.

Keywords: qubit, quantum entanglement, chaos

Abstrakt. V nedávných článcích byla objevena existence nového typu chaotického chování kvantových systémů, který nemá analogii v klasické fyzice. Příčinou chaosu je samotný kvantový popis stavu, na který je aplikován nelineární operátor. Analyticky byl chaos popsán pro speciální třídu dvouqubitových stavů při aplikaci speciálního protokolu. Naším cílem je popsat evoluci smíšených stavů, což jde za hranice současného poznání. Pracujeme s jednoqubitovou verzí purifikačního protokolu. Odhalili jsme nový jev, poloatraktivitu kvantových stavů. Hlavní výsledek naší práce spočívá ve využití tzv. box-counting dimenze k charakterizování struktur stavů s chaotickou evolucí. Tyto struktury podléhají fázovému přechodu, přičemž roli teploty hraje čistota stavů. Tyto náhlé kvalitatvní změny zmíněných struktur jsou velmi překvapující. Rovněž prezentujeme kvantitativní charakteristiku oblastí přitažlivosti atraktorů, která značí, že počet stavů, které purifikační protokol umí 'vyčistit' roste exponenciálně s čistotou stavu.

Klíčová slova: qubit, kvantové provázání, chaos

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/241/OHK4/3T/14 and the Hungarian National Research, Development and Innovation Office under Contract No. K124351.

1 Introduction

Quantum information and computation offer great improvements to classical tasks. Quantum entanglement is one of phenomena that is widely exploited in newly proposed algorithms. However, it suffers form decoherence that cannot be in principle eliminated. Processes aiming on repairing the entanglement are called purification protocols. One of them proposed [2] and generalised [1] lately sacrifices a copy of a piece of information to repair another copy. These exponential costs have to be taken into account for multiple iterations and they are the reason to seek improvements to the protocol. The particular protocol has been shown to induce chaotic behaviour in a special set of pure states.

This type of chaos in the sense of the sensitiveness of the state's evolution to initial conditions has no analogy in classical physics. It is also different from so called quantum chaology (which studies quantum systems corresponding to classically chaotic systems) because the chaos is rooted deeply in the mathematical description of the quantum reality. The reason for this chaotic feature lies in nonlinear maps which can be generally found in physics of open quantum systems but these have not been yet studied. We now aim on showing that the dynamical regimes can be very interesting, rich and surprising.

Because of the complex and intricate nature of the topic we study single-qubit version of the protocol acting on general mixed states. The single-qubit states can be isomorphically mapped onto a particular set of two-qubit states. This allows for reinterpretation of our results to protocol capabilities regarding entanglement purification. We propose a new method to characterise chaotic dynamics inside the Bloch sphere based on study of states that are sensitive to initial conditions. These states form an interesting structure which we characterise in the parameter space of the physical system using concept of box-counting dimension. After explaining the method and we present our main observation. We find that the structure of chaotic states undergoes a phase transition with respect to purity of the initial states.

Additionally, we show that the relative amount of states of given initial purity that converge to the mixed attractor increases with lowering purity. This finding can be interpreted in terms of purification capabilities of the protocol; this purification is meant as increasing the purity of the state here but in two-qubit reinterpretation it manifests in entanglement purification capabilities.

2 Chaos and quantum systems

The nonlinear map acting on mathematical representation of a physical system is the crucial point of our research. General nonlinear maps in quantum physics can be studied only in open systems, because closed system evolve unitarily. In this mode it is impossible to implement expanding or contracting maps. And it is exactly the expanding property that is responsible for the sensitivity to initial conditions, i.e. chaos.

If we would like to examine general nonlinear operator acting on two qubits we would need theory for 15 functions of 15 real variables. Therefore, we choose single-qubit protocol version where three real variables are dealt with. Nevertheless, we remain beyond scope of mathematical books. In this setting we will find many phenomena familiar to classical nonlinear dynamics [6] and theory of complex functions [5]. Amongst these are fractal structures and attractiveness/repulsiveness in certain directions.

The nonlinearity in our protocol is result of interaction of two qubits mediated by measurement-based modification. This is experimentally implemented via CNOT gate which determines computational base of a qubit $(|0\rangle, |1\rangle)$, whole this paper is set into this basis. The CNOT gate is also responsible for the nonlinearity of the protocol. For the detail discussion of the protocol, its construction and physical realisation see [1, 2, 3, 4]; in following text we only present crucial shards of information.

2.1 Protocol iteration

The original purification protocol is constructed to act on two-qubit states but it can be generalised to act an other systems. We choose single-qubit system because of two reasons. The system is simpler but it still goes beyond accessible knowledge as already mentioned. And we can show that the single-qubit states can be mapped to a class of two-qubit states in a way that preserves all physical characteristic and the evolution function. Our examination of single-qubit mixed states than can be easily reinterpreted for that particular set of mixed entangled states.

Let us take the most general single-qubit state and we shall parameterise it in following way with respect to computational basis:

$$\rho = \frac{1}{2}\begin{pmatrix} 1+a & b-ic \\ b+ic & 1-a \end{pmatrix}; a,b,c \in \mathbb{R} : a^2+b^2+c^2 \leq 1 \tag{1}$$

where the conditions ensure that the state is physical. The protocol action on the state given by triplet (a, b, c) yields state with (a', b', c'):

$$(a', b', c') = \mathcal{F}(a, b, c) = \left(\frac{b^2-c^2}{1+a^2}, \frac{2a}{1+a^2}, \frac{-2bc}{1+a^2}\right) \tag{2}$$

The evolution function stirs the states wildly inside the interior of Bloch ball while the surface, the Bloch sphere is invariant. Pure state can be characterised with a complex number $|\psi\rangle = (1+|z|^2)^{-1}(|0\rangle + z|1\rangle)$ and its evolution is expressed via function $f(z) = \frac{1-z^2}{1+z^2}$, for details see [4, 3]. Asymptotic dynamics of a pure state has only two possibilities:

- state belongs to the Fatou set of evolution function f, therefore it is attracted to superattractive cycle $|0\rangle \leftrightarrow 1/\sqrt{2}(|0\rangle + |1\rangle)$;
- state belongs to the Julia set of f, which means it evolves chaotically. The set is a fractal formed by border of the basins of attraction that belong to different parts of the pure cycle. This regime also contains fixed unstable states.

For mixed states we find following new additional possibilities of asymptotic evolution:

- state converges to new attaractor $\rho_0 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, the maximally mixed state;
- state converges to half-attractive mixed cycle $(0.295598, 0, 0) \doteq (a_0, 0, 0) \equiv \rho_a \leftrightarrow \rho_b \equiv (0, b_0, 0) \doteq (0, 0.543689, 0)$ or half-attractive pure state $\rho_2 \equiv (a_2, b_2, 0) =$

$(b_0, \sqrt{1-b_0^2}, 0)$; the numbers a_0, b_0 can be determined analytically by solving equations $\rho'' = \rho$. States in this regime are sensitive to initial conditions, perturbation can deflect them to either of attractors. State ρ_2 is even guaranteed to be chaotic in the pure states dynamics. Now we also (numerically) find it is attractive for certain set of mixed states.

The last possibility is very interesting because it suggests that the state can be resistant to certain types of perturbation (and sensitive to others). Generally, this effect can have relevant impact on experimental usability of some general nonlinear protocols. These finding were obtained from numerical computations, now we give analytical clues that state ρ_0 is indeed an attractor and cycle $\rho_a \leftrightarrow \rho_b$ is half-attractive. To do this we evaluate two protocol iterations.

$$a'' = 4\frac{a^2 - b^2c^2}{(1+a^2)^2 + (b^2-c^2)^2},\ b'' = 2\frac{(1+a^2)(b^2-c^2)}{(1+a^2)^2 + (b^2-c^2)^2},\ c'' = 8\frac{abc}{(1+a^2)^2 + (b^2-c^2)^2} \tag{3}$$

and consider regime of small perturbations to state ρ_0 by setting $|a|, |b|, |c| < 1/8$. Within this regime each state is forced to converge to ρ_0 in sense of converging sequences $a^{(n)}, b^{(n)}, c^{(n)}$. Their convergence is not necessarilly monotonic but it is monotonic when the protocol is applied pairwise. To give clues to half-attractiveness of $\rho_a \leftrightarrow \rho_b$ we consider also two iterations of the protocol and regime of small perturbations $a \doteq a_0, b \doteq 0, c = 0$. Using Taylor series we find $a''|_{a=a_0,b=\varepsilon} = a_0(1 - b^4 + \mathcal{O}(b^8)) \doteq a_0, |b''||_{a=a_0,b=\varepsilon} = \left|\frac{2b^2}{1+a_0^2} - \mathcal{O}(b^6)\right| < |b|$ whenever $|b| < 1/2$. The cycle is therefore resistant to perturbation satisfying particular relations in a, b. This relation basically determines a curve in plane $c = 0$, we see in figure 3 that this curve runs through $c = 0$ plane and separates attractor basins of the mixed and the pure attractors. The relation is very complicated and we have not succeeded in expressing it. Repulsiveness of the cycle can be viewed when considering states $(t, 0, 0)$ or $(0, t, 0); t \in \langle 0, 1\rangle$ subject to two iterations. In these invariant sets of states repulsiveness is proven analytically via derivative of evolution function $t \to t''$.

Particular plane of states $c = 0$ is important for several reasons. It captures all asymptotic features of mixed states because all states (up to negligible set not capturable by numerical calculations) approach this plane; inside this plane they are evolved to the positive-positive quadrant because of the squaring in 2. All critical states are found in this quarterdisc and the attractiveness inside this disc is clearly presented in 3.

2.2 Box-counting method and chaos description

We remind the chaotic behaviour can be described analytically on the Bloch sphere which can be identified with the Riemann sphere which is conformal to complex plane, state and its evolution are then described by single number $z \in \mathbb{C}$ and function $z \to z' = \frac{1-z^2}{1+z^2}$. This function can be examined using theory [5]. The main feature is that the chaotic states are confined to a peculiar fractal structure with deterministically chaotic evolution. Such tool is not available for mixed states. However, we develop a new method of characterising the chaotic evolution in mixed states based on the pure states analysis. We notice that the states 1 with the same purity P are spheres. We identify these states with a plane using stereographical projection

$$P = \frac{1 + a^2 + b^2 + c^2}{2}\,, \quad x = \frac{b}{1 + a}\,, \quad y = \frac{c}{1 + a} \tag{4}$$

Identification with a complex plane does not yield evolution function which can be analysed using theory [5] unless $P = 1$. We stay with the two-dimensional real plane and calculate evolution numerically. After determining asymptotic evolution of the states in the plane, we assign them a colour based on attractor they converge to. In this way we obtain an image we will refer to as *attractor map*. We stress that one such map is created for chosen purity value P and illustrate asymptotic evolution of states that initially have purity P. The evolution typically brings states away from their initial sphere but in this way we can analyse what asymptotic regimes are and are not available depending on the mixedness of the state.

In the attractor maps we find areas of the same colour which are cuts of basins of attraction of attractors. In other words, the islands are states with regular behaviour. On contrary, states forming the borders of these islands are necessarily chaotic because perturbations can deflect them to one or another attractor meaning the states are sensitive to initial conditions. We state that we are going to study the particular structure of borders of attractor islands in attractor maps. This structure in pure state case collapses to fractal structure shown in [4] (the existence and properties are guaranteed by theory). There is a measure capable of characterising the fractalness of the structure, it is the *fractal dimension*, also known as Hausdorff dimension:

Statement 1. *Dimension $\mathcal{D}$ of an object $\mathcal{Y} \subset \mathcal{X}$ in metric space $(\mathcal{X}, \rho)$ is*

$$\mathcal{D} = \lim_{\varepsilon \to 0} \min \frac{\log N_\varepsilon}{\log \frac{1}{\varepsilon}}\,, \tag{5}$$

where N_ε is number of open sets covering the object $\mathcal{Y}$, the minimum is taken over all possible coverings with open sets of diameter $< \varepsilon$.

This quantity captures how finer the structure gets when we study it in finer and finer scales. Nonetheless, it is impossible to determine it for general objects. Therefore, we use following concept of *box-counting dimension* which relieves the definition to estimate the dimension numerically. The method is described in many similar but not same ways, e.g. like in [6] and for its fundamentally simple approach we develop it on our own in MATLAB interface as described later. The crucial idea of the box-counting concept lies in taking boxes instead of challenging all possible coverings. Bypassing the minimum across all possible coverings increases the dimension estimate but allows to easily numerically determine number of covering boxes. We use pictures of fractals which we cover with rigid grid of $m \times m$ squares which is in contrast with [6] where floating boxes are used. Second idea simulates the limit $\varepsilon \to 0$ by taking boxes of smaller and smaller size, in other words m increasing to the resolution of the picture n pixels. Although we can reach only $\varepsilon = n/m \geq 1$ the dimension estimate remains reliable when pictures of high resolution are used. It is because we use another idea: from the 1 we can see that the dimension is a slope of line formed by points $[\log m, \log N_m]$ in limit $m \to \infty$. As this limit is simulated we conclude that the method is implemented in following steps: We choose purity of initial states P. We create attractor map (as described earlier) of the states. This map of resolution $n \times n$ is then cut into $m \times m$ boxes for all possible $m|n$ and number of covering boxes N_m is determined. The dimension is gained as the slope of line fitted by least squares method through points $[\log m, \log N_m]$.

This approach naturally has several pitfalls. If the box structure coincides with the structure of our interest, it cannot capture the structural character properly, especially when the structure is curve, in pathological case like the model of Sierpinski carpet in figure 1 the method fails. The setting of the object in the picture (and in consequence its setting in the box grid) has important influence on the resulting value, see figure 1. Last important caveat lies in the finite resolution of used pictures. These inherently cannot capture infinitely recurring fractal structure but can only approximate it. That is the reason to use images with high resolution. However, when the number of boxes is large $m \sim n$ each box captures only few pixels which do not contain proper structural information. In consequence, we cannot use high values of m to fit the dimension because they underestimate the value. Also, for low values of m a single box contains large pieces of object and does not capture fine details. Aware of these issues we suggest to use various pictures of the object and decide image from image proper values of m to fit the dimension of the structure. We 'calibrated' the method on basic structures to be more reliable but still the method can yield value precisely only to first, maximally second decimal digits. Avoiding pathological objects we conclude the dimension of the structure can give indicative estimate of its fractalness but not precise value. Besides, no other method of characterising the structural features exists.

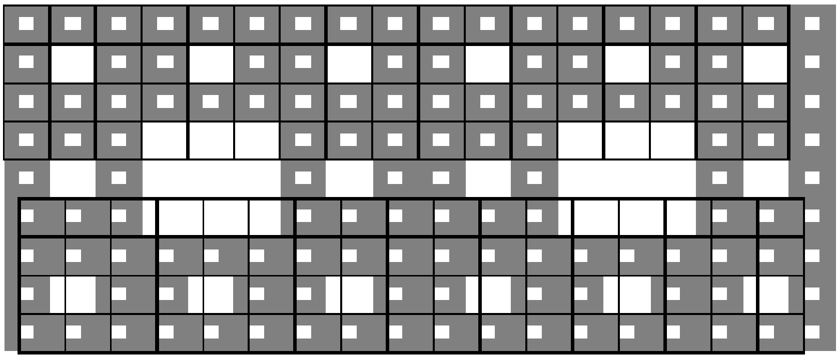


Figure 1: Simplified model of the Sierpinski carpet simulates finite resolution of pictures and also demonstrates position dependence of the box grid. A level finer grid cannot capture border of grey-white.

3 Chaotic dynamics in single-qubit mixed states

3.1 Phase transition in the structure of chaotic states

In figure 2 we illustrate the structure of chaotic states on a sphere within mixed states. From the numerical calculations we immediately make following conclusion. For purity $P = 1 - \varepsilon; \varepsilon > 0$ arbitrary, there are states converging to the mixed attractor. However, visually the structure is very similar to structure of pure states. In order to qualify the structure we use the box-counting method to find the dimension of the borders of the coloured island in these images.

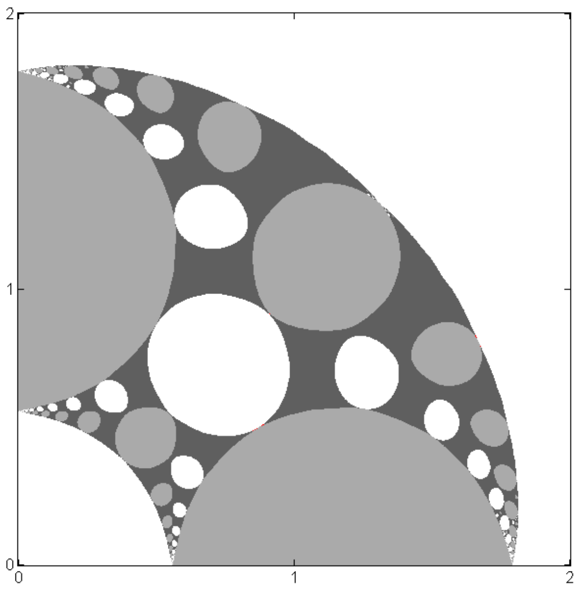


Figure 2: Example of attractor map for initial purity $P = 0.9$. Colour coding of attractors is same in both figures 2,3: white colour marks states converging to $|0\rangle$ after even number of protocol iterations and bright grey states converging to $|0\rangle$ after odd number of iterations; grey colour marks states converging to the mixed attractor; dark colour stand for nonphysical states. Only positive-positive quadrants are shown because of central symmetries.

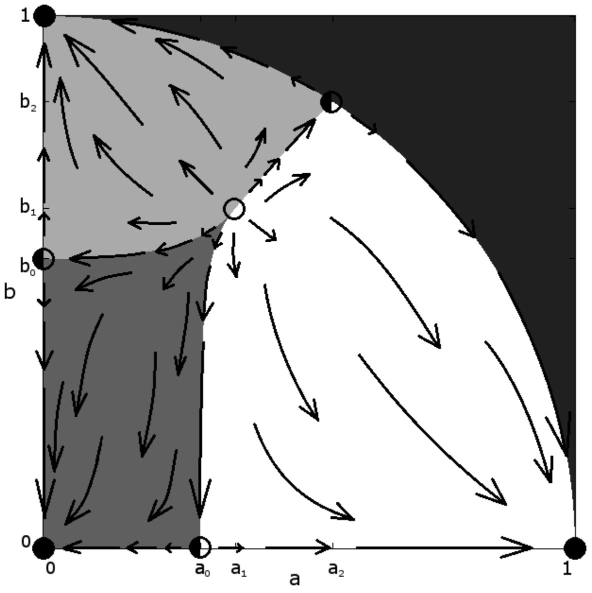


Figure 3: Evolution in quarterdisc of $c = 0$ plane. The arrows symbolise 'the attractive forces' - how fast does a state converge to its attractor when two(!) protocol iterations are used. When the attractive forces of the attractors compensate on the borders of the grey-scaled regions, the states can be attracted to the saddle states marked with half-filled circles. The attractive states are marked with filled circles, the repulsive state with an empty circle.

Results confirm that the fractal structure can be preserved in the mixed states. This is surprising result because mixedness means statistical uncertainty of the physical state. Presence of this uncertainty does not necessarily change the evolution to some trivial regime. Even more surprising is the fact that the dimension remains constant in regime $P = 1 - \varepsilon$ which means that the fractal structure is the same.

The most important result is obvious when we plot the dimension of the structure of chaotic states of chosen initial purity P with respect to this purity. From 4 we can see that the dependence is essentially a phase transition. The structure is the phase and it is in mode *fractal* when the purity of states is in range $P \in (P_1, 1\rangle$. Value P_1 numerically coincides with purity of state

$$\rho_1 = \frac{1}{2}\begin{pmatrix} 1+a_1 & 1-a_1 \\ 1-a_1 & 1-a_1 \end{pmatrix}; a_1 \doteq 0.3611 \quad \rightarrow P_1 \doteq 0.769292 \tag{6}$$

which is a repulsive fixed state also shown in figure 3. The value a_1 can be determined analytically solving $\rho' = \rho$. It seems that this state is the least pure source from which the fractal structure grows. For lower purity, the structure of states that initially have the chosen purity and exhibit sensitivity to perturbations has dimension 1, i.e. is *nonfractal.*

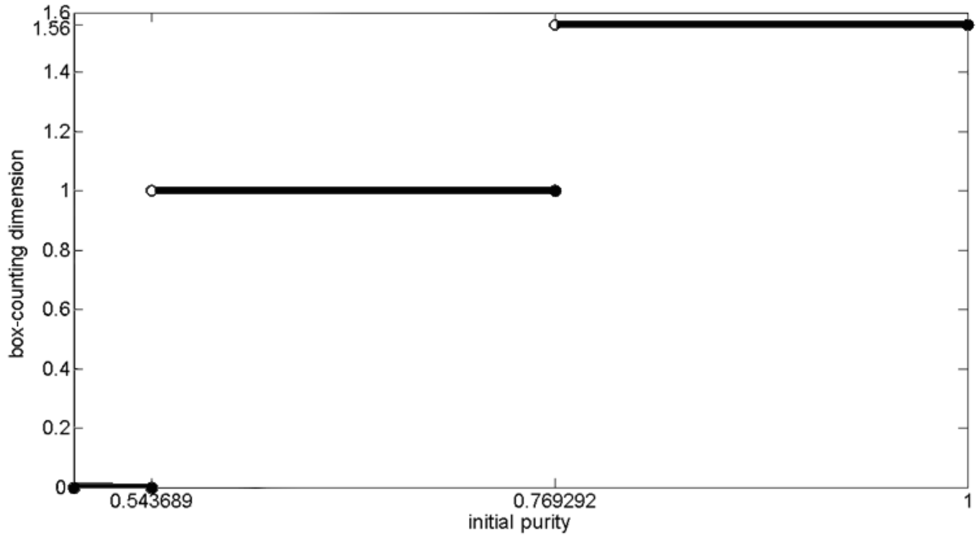


Figure 4: The fractal structure of chaotic states is a phase, its dimension transits sudden changes when the temperature (the purity) changes.

This means that the structure is formed by union of 'common' curves. Another transition in structure happens at the purity P_0 of ρ_a defined earlier as a part of half-attractive cycle. This value is equal to $P_0 = \frac{1+a_0}{2} = b_0 \doteq 0.543689$. The state ρ_a is the least pure state which does not converge to the maximally mixed state ρ_0. This implies that for $P < P_0$ there is *no structure* of chaotic states.

We interpret the sudden change of the fractal dimension when the purity of the initial states is changed as phase transition. The reason is that the structure of chaotic states is not some abstract mathematical construction but truly a phase with its own physical properties, namely exponential sensitiveness to initial conditions, i.e. chaos.

3.2 Quantitative characteristics of attractor basins

The dimension of the structure is its qualitative characteristic and the phase transition expresses that there is single fractal structure changing to nonfractal and than disappearing suddenly. The fact that the fractal structure has its dimension $\mathcal{D} \doteq 1.56$ means that the structure has zero area but infinite length. The dimension expresses the self-similarity and complexness of the structure. In contrast, the nonfractal structure after the transition has finite length. While in preceding subsection we have demonstrated the qualitative properties of the structure of chaotic states, now we have discussed also its quantitative properties.

However, we also present certain quantitative properties of the attractor basins. This structure is formed by points of regular behaviour and in the attractor maps it is formed by coloured islands themselves (not their borders like before). We now want to determine relative amount of states drawn to each attractor. To do this we express the sphere of states as a matrix of elemental areas in spherical coordinates and we assign to an attractor all elemental areas $\sin\vartheta\Delta\vartheta\Delta\varphi$ for each state $\sim \varphi, \vartheta$ that converges to it. By omitting the radius of the sphere we obtain percentage of states of chosen initial purity converging to this and that attractor.

The dependence of relative areas is shown in figure 5. Numerically fitted, it is piecewise composed of exponential functions $A = \exp(\alpha_i P + \beta_i) + \gamma_i$. The parameters undergo

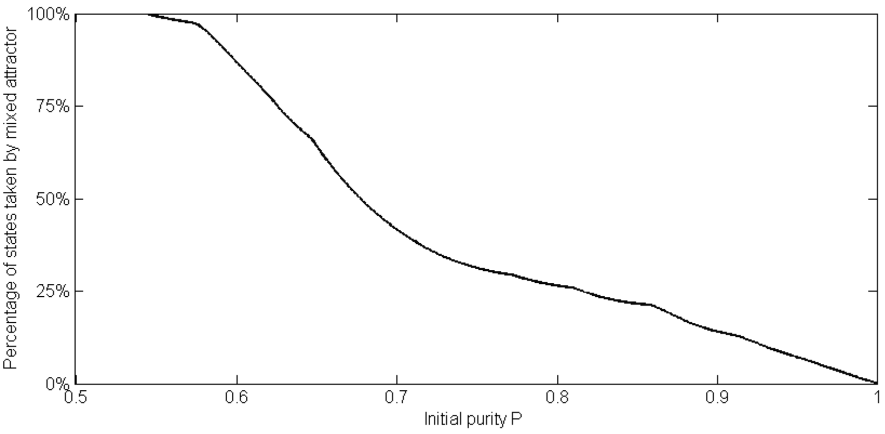


Figure 5: Partition of states with chosen initial purity that converge to the mixed attractors. Curve is formed by piecewise exponential functions $e^{\alpha_i P+\beta_i} + \gamma_i$.

sudden changes, not only in points of phase transition studied before but also other purity values which numerically match with appearance of other sources of fractal structure, i.e. points where the fractal visually emerges from. The quantitative description of attractive basins (more precisely their cuts with hyperplanes of states if constant initial purity) is more complex in purity than the quality of structure of chaotic states. We can interpret the exponential dependence in following sense: Relative amount of states that are not purified by the protocol exponentially explodes as the purity is lowered. This time we use term purification for making a state less mixed, in two-qubit protocol version this leads to analogous statement about purification of entanglement.

4 Conclusion and outlook

When we step outside the unitary dynamics of quantum system we can come across irregular dynamics exhibiting sensitivity to initial conditions. This type of chaos goes far beyond classical physics. As a result of quantum description of physical system it can manifest in interferences or have no analogy at all. The physics of quantum open systems is at its very beginning concerning the chaos in quantum states. Although the theoretical tools demonstrated its presence in pure states subject to particular protocol, it was not clear whether same chaos is present in mixed states which contain uncertainty. Our study shows that this uncertainty is not necessarily amplified during the evolution and even mixed states can be purified and they can be chaotic.

The phase transition presented in our work is not only some abstract mathematical construction but has its physical meaning and properties. The phase is the structure of chaotic states which is understood via its dimension. The temperature is the purity of the initial states which is capable of measuring statistical uncertainty of the physical state. The transition of phase vs. temperature then means sudden dramatic change of the structure of chaotic states of given initial purity. This transition can hardly be experimentally measured because the dimension of the states is still $\mathcal{D} < 2$. Therefore, the experimental chance to prepare such state is also negligible. In contrast to this jump from fractal to nonfractal structure, the jump from nonfractal to no structure means that no state can experimentally exhibit sensitivity to initial conditions. All states with purity $P < P_0$ are doomed to converge to the maximally mixed state under our protocol.

While the dimension gives qualitative description of the structure, we also presented certain quantitative characteristic of the evolution of the mixed states by means of areas of attractor basins captured by attractors within states of chosen purity. We demonstrated that the relative number of purifiable states is reduced exponentially with decreasing purity. Nevertheless, the exponential function changes its parameters with the temperature yielding more complex dependence behaviour than the qualitative characteristics.

The presented results describe dynamics within mixed single-qubit states. There is an isomorphism between the single-qubit mixed states and a particular set of two-qubit states that preserves evolution and all physically relevant properties of the state. In consequence, these results are also valid for these particular two-qubit states when properly interpreted.

The fact that the structure of chaotic state undergoes a transition 'fractal ↔ nonfractal ↔ none' means that the amount of chaotic states is qualitatively and also quantitatively different. The exact nature of the evolution of these states remains unclear because numerical simulations show half-attractive behaviour of certain states (we remark in pure states the theory guarantees deterministic chaos in Julia set of the evolution function). This newly-found property could possibly manifest in experiments. The question we settle now is: What type of chaos can quantum physics allow? What regimes are forbidden by quantum description of the world? The fractal shapes can be possibly change when the protocol is modified. When the Hadamard gate is replaced by another protocol, we can encounter different chaotic patterns and different attractors. We suggest detailed study of the protocol modification. We believe the nonlinear dynamics in quantum physics is unusually rich and exotic and has many to offer.

References

[1] G. Alber, A. Delgado, N. Gisin, I. Jex. J. Phys. A: Math. Gen. **34** (2001), 8821–8833

[2] H. Bechmann-Pasquinucci et. al. Phys. Lett. A **242** (1998), 198–204

[3] T. Kiss, I. Jex, G. Alber, S. Vymětal. Phys. Rev. A **74** (2006), 040301(R).

[4] T. Kiss, S. Vymětal, L.D. Tóth, A. Gábris, I. Jex, G. Alber. Physical Review Letters **107** (2011), 100501.

[5] J.W. Milnor. *Dynamics in One Complex Variable*, 3^{rd} edition. Princeton University Press (2000).

[6] S.H. Strogatz. *Nonlinear Dynamics and Chaos*, 2^{nd} edition (Westview Press)(2015)

Transport Properties of Percolated Coined Quantum Walks on Various Graphs

Jan Mareš

4th year of PGS, email: maresj23@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

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

Jaroslav Novotný, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Quantum walk is a simple abstract model of an excitation spreading in some environment represented by an undirected graph, where the state and the evolution of the system are described by quantum physics. Therefore, quantum walks can be used for simulation of various quantum systems. In this work, we investigate a percolated version of a quantum walk, where the graph undergoes a continuous change during the evolution.

We use our previous general results to determine asymptotic transfer probabilities of an excitation from some given initial vertex to a sink vertex for several examples of 3-regular graphs. First we demonstrate our methods on one of the simplest graph representing a spatial structure - the cube graph. Further we investigate tree graphs and present a closed-form expression for the transfer probability on a class of "snowflake" graphs of arbitrary size.

Keywords: quantum walks, percolation, transfer, asymptotic behaviour

Abstrakt. Kvantová procházka je jednoduchý model šíření excitace v prostředí reprezentovaném neorientovaným grafem, kde je stav a vývoj systému popsán pomocí kvantové mechaniky. Kvantové procházky tedy mohou složit k simulaci kvantových systémů. V této práci se zabýváme kvantovými procházkami s perkolací, kde podkladový graf podléhá nepřetržité změně při časovém vývoji systému.

Používáme zde naše předchozí výsledky ke stanovení asymptotické pravděpodobnosti přenosu excitace z daného počátečního vrcholu do koncového vrcholu pro několik příkladů 3-regulárních grafů. Nejprve demonstrujeme naše metody na jednom z nejjednodušších grafů představujících prostorové těleso - na grafu krychle. Dále zkoumáme stromové grafy a docházíme k výrazu pro pravděpodobnost přenosu na třídě grafů "sněhových vloček" libovolných velikostí.

Klíčová slova: kvantové procházky, perkolace, přenos, asymptotické chování

1 Introduction

Even without quantum computers capable of outperforming the classical ones, there is a need for understanding quantum effects in various systems. Since the number of classical bits needed to simulate a certain number of qubits grows exponentially, it is intrinsically difficult to simulate a quantum system on a classical computer. Fortunately, one does not need a universal quantum computer to deal with this problem. We may just use some other quantum system, which we are able to controll and measure, and use it as a quantum simulator [1] to gain insights about another system of interest. The model of quantum walk can be used just for this purpose. An example of this can be the realised simulation of two-particle dynamics by a 1-walker quantum walk in a 2-dimensional lattice [2].

In our previous work [3], we have presented some general solutions of the asymptotic behaviour of percolated coined quantum walks on general and in particular 3-regular graphs. Now we apply these findings in the study of an asymptotic transfer of an excitation in chosen graphs and classes of graphs. In the whole work we use our modified framework for defining coined quantum walks. We shortly introduce this framework and recapitulate the previous results (without derivations) so that we can use them further.

2 Coined Quantum Walk Definition

The quantum walk is defined on an undirected graph $G(V, E)$, where V is the set of vertices and E is the set of edges. We call G the structure graph of the quantum walk.

The Hilbert Space

The walker is described as standing in some vertex facing towards some other vertex. We associate with the structure graph G a directed graph $G^{(d)}(V, E^{(d)})$ called the state graph. Every undirected edge in the structure graph corresponds to two directed edges of the state graph and these directed edges correspond to base states of the walker. The Hilbert space $\mathcal{H}$ is, therefore, spanned by states $|e^{(d)}\rangle$, where $e^{(d)} \in E^{(d)}$ is some directed edge. Apart from edges going from one vertex to another, the state graph may also contain added loops. (Those may be used to assure regularity of the state graph.)

We will denote subspaces spanned by states corresponding to edges originating in some vertex $v \in V$ as $\mathcal{H}_v$. The Hilbert space $\mathcal{H}$ of a quantum walk can than be written as a direct sum of vertex subspaces: $\mathcal{H} = \bigoplus_{v \in V} \mathcal{H}_v$.

The Time Evolution

The time evolution proceeds in discrete steps and is governed by a unitary evolution operator U:

$$|\psi(t+1)\rangle = U\,|\psi(t)\rangle = U^{t+1}\,|\psi(0)\rangle\,.$$

The operator U can be further decomposed into applications of three unitary operators:

$$U = CPR.$$

Here R is what we call a reflecting shift operators and it moves the walker among vertices - every state is mapped to the other one on the same undirected edge (the initial and the terminal vertex are swapped) or it is left unchanged in the case of loops. Further, the local permutation operator P is applied. It is a permutation operator that only acts locally in vertex subspaces and determines the final direction of the walker in the new vertex. The combined action of R and P represents a so called shift operator. Finally, there is the coin operator, which is an arbitrary vertex-local unitary operation.

Percolated Quantum Walk

By percolation we understand a random disturbance of the underlying structure graph resulting in some broken edges that can not be traversed by the walker. In particular, we will study dynamical percolation, where a new percolated graph (graph obtained from the original structure graph by closing some edges) is generated in every step of the walk. (An edge can, therefore, be closed in one step and open in the following step.)

The Hilbert space is not affected by the percolation, but directed edges corresponding to a closed undirected edge are replaced by loop. Consequently, the reflecting operator R_K (corresponding to some configuration of open edges $K \subset E$) does not move the walker over a broken edge.

The coin operator C and the local permutation P are not affected by percolation.

3 Asymptotic Evolution of Percolated Quantum Walks

The process of percolation brings classical randomness into the system and we now use a density matrix to describe the state of the walk. The time evolution is now governed by a random unitary operation:

$$\rho(t+1) = \sum_{K \subset E} \pi_K U_K \rho(t) U_K^\dagger,$$

where U_K is the evolution operator with the modified reflecting shift operator R_K corresponding to the particular percolated structure graph $G_K(V, K)$ for $K \subset E$ and π_K is the probability of the occurrence of this configuration.

The asymptotic behaviour of a system with such time evolution is studied in [4]. The asymptotic state is determined by so called attractors – solutions of the set of equations:

$$U_K X_\lambda U_K^\dagger = \lambda X_\lambda\,, \quad \text{for all } K \in 2^E, \tag{1}$$

for some given λ fulfilling $|\lambda| = 1$.

The asymptotic state (the limit for infinitely many steps) of a percolated quantum walk is than given as [4]:

$$\rho_{t\to\infty}(t) = \sum_{\lambda,i} \lambda^t \mathrm{Tr}\Big(\rho(0) X_{\lambda,i}^\dagger\Big) X_{\lambda,i},$$

where i distinguishes different attractors for the eigenvalue λ in the orthonormal basis of the solutions of (1) and $\rho(0)$ is the initial state of the quantum walk.

Pure Eigenstates Ansatz

In many cases it is possible to use a simpler approach [6] for finding the set of attractors using common eigenstates of all unitary operators U_K:

$$U_K \left|\phi_{\alpha,i}\right\rangle = \alpha \left|\phi_{\alpha,i}\right\rangle \ , \quad \text{for all } K \subset 2^E, \tag{2}$$

with a corresponding eigenvalue α (i distinguishes different common eigenstates corresponding to α). Then the operator:

$$Y_\lambda = \sum_{\alpha\beta^*=\lambda} A^{\alpha,i}_{\beta,j} \left|\phi_{\alpha,i}\right\rangle \left\langle\phi_{\beta,j}\right|$$

is an attractor corresponding to the superoperator eigenvalue $\lambda = \alpha\beta^*$. It is common that the whole set of attractors can be constructed from these so called p-attractors and a single non-p-attractor resulting from the identity operator. We have shown in the previous work that this is the case for a percolated quantum walk with the grover coin on a 3-regular graph with the reflecting shift operator (the local permutation P is the identity) or cycling shift operators (in every vertex, P can act as a clock-wise or counter-clock-wise permutation).

4 Percolated Grover QWs on 3-regular Graphs

Here we will consider both true 3-regular undirected structure graphs (leading immediately to 3-regular state graphs) and structure graphs with some vertices of lower degree, where we add some loops in the state graph to assure 3-regularity.

We use the 3-dimensional Grover coin in every vertex:

$$G_3 = \frac{1}{3}\begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}.$$

We have dealt with the asymptotic behaviour of such walks in the previous contribution. Here we restrict ourselves to uantum walks with the reflecting shift operator (the local permutation P is the identity), which exhibit an interesting phenomenon of trapping.

Common Eigenstates

Since there are only p-attractors and the identity attractor for this percolated quantum walk, the task of finding the asymptotic state reduces to finding the set of common eigenstates of all evolution operators. There is always one p-attractor corresponding to the eigenvalue 1, which has all the matrix elements the same. The interesting part are the attractors corresponding to -1, where the condition (2) ultimately leads to two rules for the common eigenstates:

1. The sum of vector elements in one vertex must be equal to 0.

2. Vector elements corresponding to directed edges on one undirected edge must be the same.

It can be shown that the equations are independent except from the case of a bipartite structure graph. Then there are $N = 2\#V - \#E$ common eigenstates corresponding to the eigenvalue -1. If the graph is bipartite, one of the equations can be obtained from the others and the number of independent common eigenstates is $N = 2\#V - \#E + 1$.

It is possible to construct a non-orthogonal basis of the subspace of common eigenstates in such a way that all the matrix elements are 1, -1 or 0. (This is no longer true after orthogonalization.) Then the common eigenstates can be represented as paths of non-zero elements in the graph, which are either closed or start and end in loop states. (Due to the zero-sum condition, only two elements in every vertex can be non-zero, so there is no branching.) As a result, the common eigenstates are typically restricted to some subset of vertices and the walker can be trapped in some part of the graph.

Asymptotic Transport

We study a scenario where the walker starts in some given vertex and there is a sink in some other vertex. Whenever the walker enters the sink vertex, he is lost in the sing. This means that the state of the system is projected to a subspace of non-sink states after every step of the walk.

We ask, what is the probability of the walker moving from the initial vertex to the sink (excitation transfer) versus the case of the walker staying trapped in the non-sink vertices of the graph.

If we have the common eigenstates of the percolated walk and we orthogonalize them in such a way that we first use the sates with no sink overlap (preserving this property in the maximal number of states after orthogonalization), we can determine the asymptotic transfer probability easily. We just exclude the common eigenstates with sink overlap and the probability of trapping is given by the overlap of the initial state with the remaining common eigenstates.

5 Example: Percolated Grover QW on a Cube

One of the simplest examples of 3-regular graphs is the cube. Let us position the cube in a coordinate system as shown in figure 1. Every vertex has one edge in the direction of every axis and we use this to denote states of the walk - the computational basis is chosen in the order e_x, e_y, e_z in every vertex.

The graph is bipartite and has 8 vertices and 12 edges. Therefore, we must find $N = 16 - 12 + 1 = 5$ common eigenstates corresponding to the eigenvalue -1. The cube has 6 faces with even number of edges and we simply choose 5 of those and use common eigenstates corresponding to cycles on these faces (denoted as "down", "left", "back", "right", "front"). For example the eigenstate on the left edge will be:$|\psi_l^{(-1)}\rangle = [-1, 0, 1, -1, 0, 1, 0, 0, 0, 0, 0, 0, -1, 0, 1, -1, 0, 1, 0, \ldots, 0]^T$. The only common eigenvector for the eigenvalue 1 has all elements equal: $|\psi^{(+1)}\rangle = \frac{1}{\sqrt{24}}[1, 1, 1, \ldots, 1]^T$.

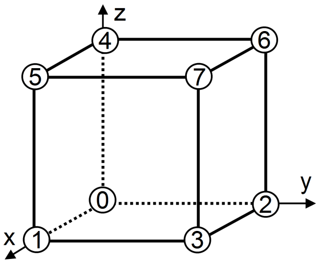


Figure 1: Coordinates on the cube graph. The vertex numbers are chosen so that they correspond to binary numbers given by coordinates zyx. We denote faces by their position so for example the left face has vertices v_0, v_1, v_5 and v_4.

Further calculations are performed using Wolfram Mathematica software, which allows for symbolic solutions. Overall, we obtain a complete set of 37 attractors, 10 corresponding to the eigenvalue -1 and 27 corresponding to $+1$ allowing us to calculate the asymptotic regime when an initial state is given.

The sink is located in the vertex v_7 and the initial state is always localised in the vertex v_0. The common eigenstates with no sink overlap correspond to the "down", "left", and "back" faces. Depending on the initial state, the transfer probability ranges from 70 % to 100 %. The full transfer occurs exclusively for the initial state $|\psi_0\rangle = \frac{1}{\sqrt{3}}[1,1,1,0,\ldots,0]$, because it is orthogonal to all trapped common eigenstates with no overlap with the sink: $|\psi_d^{(-1)}\rangle, |\psi_l^{(-1)}\rangle$ and $|\psi_b^{(-1)}\rangle$. States with the minimum transfer are linear combinations of the states:

$$\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}. \tag{3}$$

Obviously, if the walker begins for example in the state $|\psi_d^{(-1)}\rangle$, he will stay trapped in that state and the transfer probability will be 0, but this state is not localised in the vertex v_0 at the beginning.

This kind of asymptotic trapping has already been shown in [7] for a quantum walk on a line with a coin state corresponding to no movement of the walker ("lazy quantum walk"). Our result demonstrates that the trapping is not associated with the presence o these no-movement states, but rather with the presence of vertices of the degree higher than two.

We also investigate (numerically) the transfer probability in the non-percolated version of the reflecting quantum walk on a cube graph. Obviously, the common eigenvectors present in the percolated version are also eigenvectors for the non-percolated walk, so the trapping is again present for most of the initial states. Nevertheless, more trapped eigenvectors can be identified. There are eigenstates corresponding to the eigenvalue -1 similar to $\left\{|\psi_i^{(-1)}\rangle\right\}_{i\in\{d,l,b,r,f\}}$. The difference is that the values 1 and -1 of the elements oscillate on the level of directed edges. (There is no condition requiring the elements corresponding to the same undirected edge to be the same.) For example the eigenvector

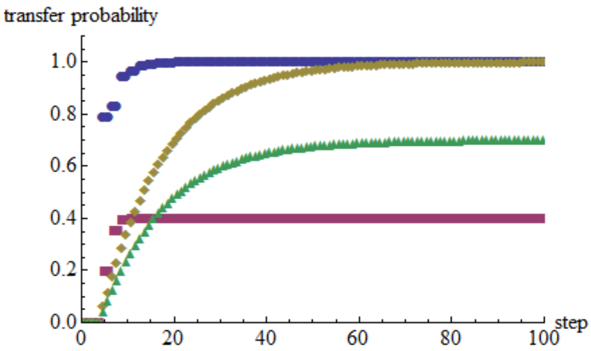


Figure 2: Numerical simulation of Grover quantum walk on a cube graph with a reflecting shift operator: without percolation, initial states $|\psi_0\rangle = \frac{1}{\sqrt{3}}[1,1,1,0,\dots,0]$ (blue circles) and $|\psi_0\rangle = \frac{1}{\sqrt{2}}[1,-1,0,0,\dots,0]$ (purple squares) and with percolation , initial states $|\psi_0\rangle = \frac{1}{\sqrt{3}}[1,1,1,0,\dots,0]$ (yellow diamonds) and $|\psi_0\rangle = \frac{1}{\sqrt{2}}[1,-1,0,0,\dots,0]$ (green triangles). The horizontal axis shows the step of the walk and the vertical axis cumulative transfer probability.

corresponding to the left face is:

$$|\chi_l^{(-1)}\rangle = [-1,0,1,1,0,-1,0,0,0,0,0,0,1,0,-1,-1,0,1,0,\dots,0]^T.$$

The vector $|\psi_0\rangle = \frac{1}{\sqrt{3}}[1,1,1,0,\dots,0]$ is again orthogonal to all the trapped eigenstates and therefore is fully transferred. The minimum transfer probability is again for linear combinations of the states (3). Nevertheless, the transfer probability is only 40 %, so the chance of trapping is doubled compared to the percolated walk. This is associated with the presence of the other set of localised eigenvectors.

Results of a numerical simulation are shown in figure 2. We can see that the percolated walk converges to higher asymptotic transfer probability for the initial state $|\psi_0\rangle = \frac{1}{\sqrt{2}}[1,-1,0,0,\dots,0]$.

6 Example: Percolated Grover QW on Tree graphs

A class o graphs with some interesting properties are tree graphs - graphs with no cycles. Let us now consider 3-regular tree graphs. In fact, an undirected structure graph can not be a 3-regular tree graph, but we add loops in the state graph to achieve the 3-regularity.

The tree structure makes the construction of the set of p-attractors easy. The common eigenstate corresponding to the eigenvalue 1 is trivial (all vector elements are the same). For the eigenvalue -1 we need to find $N = 2\#V - \#E$ common eigenstates. A tree graph with $\#V$ vertices has exactly $\#E = \#V - 1$ undirected edges and therefore $2\#E$ paired directed edges and finally the remainder of $3\#V - \#2E = 2\#V - \#E + 1$ loops.

We can just choose one loop as a starting one and construct independent common eigenstates as paths from this loop to all other loops. Nevertheless, the common eigenstates have to be orthogonalized while keeping in mind that the eigenstates with no sink overlap have to be used first in the Gram-Schmidt process.

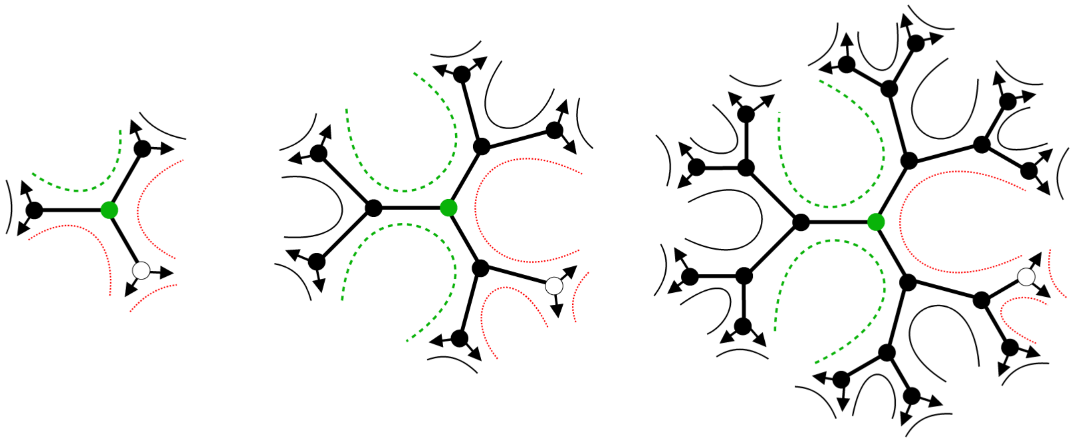


Figure 3: Snowflake graphs of the order 1, 2 and 3. The walker starts in the middle vertex and the sink vertex is not filled. The common eigenstates correspond to paths in the graph depicted by lines. The dotted-line states have an overlap with the sink and therefore will be removed from the asymptotics. The dashed-line states are crucial since those have no overlap with the sink and have an overlap with the initial vertex. In the orthogonalisation, we start with the solid-line states (no overlap with the sink or the origin), then we use the dashed-line states and the dotted-line states must be added last.

"Snowflake" Graphs

Let us consider a class of tree graphs recursively generated in the following way: The graph of the order 0 is just one vertex with three loops. The next order is obtained by replacing every loop by an edge leading to a new vertex with two loops.

Let us now investigate trapping in these graphs with the walker starting in the middle vertex and with a sink in one of the border vertices. The asymptotic transfer is given by the presence of trapped common eigenstates of the eigenvalue -1. A possible choice of those (before orthogonalisation) is shown in figure 3.

After orthogonalisation, we have only two common eigenstates with an overlap with the original vertex and no overlap with the sink (only one for the order 1). Let us denote them as $|T_1\rangle$ (spanning only the two branches without the sink) and $|T_2\rangle$ spanning the whole graph without the sink vertex. Those are the only ones contributing to the trapping. The amount of trapping is given by an overlap of the initial state with these two states.

The state $|T_1\rangle$ is very symmetrical and we will describe it as $\frac{|t_1\rangle}{\sqrt{N_1}}$, where $|t_1\rangle$ is the state scaled to natural numbers. The state $|t_1\rangle$ has elements 2^k and -2^k in the initial vertex and the values are halved in every branching with also gaining the -1 phase.

Let us also denote $|T_2\rangle = \frac{|t_2\rangle}{\sqrt{N_2}}$. The state $|t_2\rangle$ has elements 2^k, 2^k and -2^{k+1} (on the sink branch) in the initial vertex. On the non-sink branches it is similar to $|t_1\rangle$, but the corresponding elements in the two branches have equal signs. The sink branch is more complicated, because the presence of the sink introduces asymmetry. Nevertheless, in the end we only need some information about the normalisation constant, in particular that $N_2 \geq 3N_1$. To prove this, let us first note that the squares of the elements on the non-sink branches contribute N_1 to the sum. Now we can consider every element on the sink branch with two corresponding elements on the non-sink branches. If the non-sink branch was symmetrical, the values on the sink branch would be double the values on the non-sink branches. Then since $(2a)^2 = 4a^2 = 2(a^2 + a^2)$, the normalisation constant would be $N_2 = 3N_1$. The sum of the elements in every vertex must be the same and since uneven splitting will always generate larger sum of squares, it will in fact be $N_2 > 3N_1$ for arbitrary order k.

We note that not only $\langle T_1|T_2\rangle = 0$, but also the restrictions of these states to the initial vertex are orthogonal. Therefore, the trapping will be maximal, if the initial state is only a scaled version of the restriction with the greater magnitude. Since $\frac{\left|[-2^{k+1},2^k,2^k]^T\right|}{\left|[0,2^k,-2^k]^T\right|} = \frac{\left|[-2,1,1]^T\right|}{\left|[0,1,-1]^T\right|} = \sqrt{3}$ and $\sqrt{N_2} \geq \sqrt{3N_1}$, the trapping will be always maximal for the initial state $\frac{1}{\sqrt{2}}[0,1,-1,0,\ldots,0]^T$ having an overlap with $|T_1\rangle$.

Thanks to a simple structure of $|T_1\rangle$, we can explicitly calculate the normalisation constant N_1 as:

$$N_1(k) = 2^{k+1} + \sum_{i=0}^{k-1} 2^{2+i}(2^{k-i})^2 = 2^{k+1}(2^{k+2} - 3).$$

This allows us to express the maximal trapping probability on a snowflake graph for an arbitrary order k as:

$$P_{trap}(k) = \frac{2 \cdot (2^k)^2}{N_1(k)} = \frac{2^k}{2^{k+2} - 3}.$$

The values for the smallest graphs are $P_{trap}(1) = \frac{2}{5} = 0.4$ for the order 1, $P_{trap}(2) = \frac{4}{13} = 0.307692$ for the order 2 and $P_{trap}(3) = \frac{8}{29} = 0.275862$ for the order 3. The maximal trapping probability decreases with k, approaching the value 1/4.

We have also investigated a "disabled" version of the graphs where one of the non-sink branches is missing. Here the asymmetry prevented us from finding nice simple results for a general order of the graph. Nevertheless, our procedure allows for finding trapping rates for some small orders. Using Wolfram Mathematica, the maximal trapping probabilities were found to be $P_{dis}(1) \doteq 0.571$ for the order 1, $P_{dis}(2) \doteq 0.528$ for the order 2 and $P_{dis}(3) \doteq 0.522$. While for the order $k = 1$ the state with maximal trapping is the same as for the non-disabled version, for other orders the states with maximal trapping differ (from the one for $k = 1$ and also among themselves).

We can see, that in the disabled version the trapping is stronger, which is due to a very high weight on the loop in place of the missing branch, which is now a part of the initial vertex. The trapping also decreases slower with increasing order of the graph. Since the first trapped state $|\tilde{T_1}\rangle$ is analogous to the one for non-disabled graph $|T_1\rangle$, where the missing elements are just cut off, we can estimate the the maximal trapping probability by the one for a state $\frac{1}{\sqrt{2}}[0,1,-1,0,\ldots,0]^T$ having a maximal overlap with $|\tilde{T_1}\rangle$. (The true maximal trapping state is different and has a non-zero overlap with $|\tilde{T_2}\rangle$.) The normalisation constant is:

$$\tilde{N_1}(k) = \frac{N_1(k)}{2} + 2^{2k}$$

and the maximal trapping probability

$$P_{dis}(k) \geq \frac{2 \cdot 2^{2k}}{\tilde{N}_1(k)} = \frac{2^{k+1}}{2^{k+2} - 3 + 2^k} = \frac{2}{5 - 3 \cdot 2^{-k}} \geq \frac{2}{5}.$$

Therefore, the maximal trapping probability will not decrease under $\frac{2}{5}$ for an arbitrary order of the graph. Nevertheless, it can stay higher and the limit may be different.

7 Conclusion

In this work we apply general results presented in the previous contribution. Certainly, it is advantageous to make some modifications of the procedure suited for particular graph of interest, but it is demonstrated that our methods are applicable for quantum walks on various graphs.

As seen mainly in the example of the cube graph, percolation can enhance the transfer probability on the studied graph by excluding some trapped states from the asymptotic regime. This result also transfers to other 3-regular graphs, since analogous trapped states will be present. Note also, that the analytical solution of the percolated quantum walk brings a significant insight into transfer properties of the unpercolated walk.

On the example of snowflake graphs we demonstrate that the results may be rather counter-intuitive. By removing a non-sink branch, where the walker could be trapped, we increase the maximal trapping probability. Nevertheless, thanks to the analytical solution, this can be understood mathematically.

References

[1] Johnson, T. H., Clark, S. R., Jacksch, D. *What is a quantum simulator?*. EPJ Quantum Technology **1(10)** (2014).

[2] Schreiber, A. et al. *A 2D Quantum Walk Simulation of Two-Particle Dynamics*. Science **336(6077)** (2012), 55-58.

[3] Mareš, J. *Modified Definition of Coined Quantum Walks for Arbitrary Graphs*. Doktorandské dny 2016. Praha: Česká technika - nakladatelství ČVUT, ČVUT v Praze, (2016)

[4] J. Novotný, G. Alber, I. Jex. *Asymptotic evolution of Random Unitary Operations*. Central European Journal of Physics **8** (2010), 1001–1014.

[5] B. Kollár, T. Kiss, J. Novotný, I. Jex. *Asymptotic dynamics of coined quantum walks on percolation graphs*. Physical Review Letters **108** (2012), 230505.

[6] B. Kollár, J. Novotný, T. Kiss, I. Jex. *Percolation induced effects in 2D coined quantum walks: analytic asymptotic solutions*. New Journal of Physics **16** (2014), 023002.

[7] M. Štefaňák, I. Bezděková, I. Jex, M. Barnett. *Stability of point spectrum for three-state quantum walks on a line*. Quantum Information & Computation **14** (2014), 1213–1226.

Synchronizing Delay of Primitive Sturmian Morphisms

Kateřina Medková

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

advisors:

Vojtěch Rödl, Department of Mathematics and Computer Science
Emory University

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

Abstract. A synchronizing delay is a constant related to circularity of morphism. It is well-known that knowledge of the value of the synchronizing delay is very helpful when analysing the structure of bispecial factors of a given morphism. As shown in this paper, it is also possible to use this connection in the opposite direction: if the structure of bispecial factors is known, a good upper bound on the value of the synchronizing delay can be found. Using this method, a linear upper bound on the minimal value of the synchronizing delay of any primitive Sturmian morphisms is given.

Keywords: circularity, Sturmian morphism, synchronizing delay

Abstrakt. Synchronizační zpoždění je konstanta svázaná s cirkularitou morfismů. Je známo, že znalost hodnoty synchronizačního zpoždění může být s výhodou využita při analýze struktury bispeciálních faktorů daného morfismu. V tomto článku ukazujeme, že tento vztah lze využít také opačným směrem: pokud je struktura bispeciálních faktorů známá, lze toho využít ke stanovení dobrých horních odhadů hodnoty synchronizačního zpoždění. S využitím této metody je nalezen lineární horní odhad hodnoty synchronizačního zpoždění pro všechny primitivní sturmovské morfismy.

Klíčová slova: cirkularita, sturmovský morfismus, synchronizační zpoždění

1 Introduction

The notion of circularity originally comes from theory of codes, where circular codes are well-known. A set $\mathcal{X}$ of finite words is a code if each word in $\mathcal{X}^+$ (the set of all finite concatenations of words from $\mathcal{X}$) has a unique decomposition into words from $\mathcal{X}$. If we slightly modify the requirement of uniqueness, we get the definition of a circular code: $\mathcal{X}$ is a circular code if each word in $\mathcal{X}^+$ written in a circle has a unique decomposition into words from $\mathcal{X}$.

In combinatorics on words, an analogue to codes are morphisms which are injective on their languages. Circularity is defined as slightly relaxed injectivity: a morphism is circular if all long enough factors of its language have a unique preimage in its language except for some prefix and suffix bounded in length by some constant. This constant is called a synchronizing delay and it is studied in this paper.

As explained by Cassaigne in [1], knowledge of the value of the synchronizing delay can be very helpful when analysing the structure of bispecial factors in languages of fixed point of morphisms. This idea was further developed by one of the authors in [3], where an algorithm for generating all bispecial factors is given. This algorithm works for a large family of morphisms and its computational complexity depends on the value of the synchronizing delay. Moreover, as shown in this paper, this link between the synchronizing delay and the structure of bispecial factors can be used also in the opposite direction: if the bispecial factors are known, it is possible to find a good bound on the value of the synchronizing delay.

Mossé in [9] proved that every injective primitive morphism is circular. In fact, circularity is closely related to repetitiveness [8, 6]. Because of this connection, the circularity is decidable by an efficient algorithm [5]. However, if the morphism is circular, the algorithm does not provide any information about the value of the synchronizing delay (except for finiteness). Recently, a theoretical upper bound on this constant for all primitive morphisms was given in [2], but this bound is unreasonably huge and clearly is very far from being optimal. No other general upper bounds are known.

Therefore, we focus on some restricted cases in order to find some more reasonable bounds on the synchronizing delay. We have already studied the case of binary k-uniform morphisms in [4], where we found a polynomial (in k) upper bound. In this paper we focus on primitive Sturmian morphisms, which are well-known and widely studied objects in combinatorics on words [7].

The main result of this work is a linear (in the length of images of letters) upper bound on the synchronizing delay of primitive Sturmian morphisms. In particular, we prove the following result. The details of the proof are given in Section 3.

Theorem 1. *Let ψ be a primitive Strumian morphism. Then its minimal synchronizing delay Z_{min} is bounded as follows:*

$$Z_{min} \leq 3|\psi(0)| + 2|\psi(1)| - 3\,.$$

Moreover, it seems this bound is not far from being optimal. In fact, we suppose that methods similar to those used in [4] will allow us to find the exact value of the synchronizing delay for a given primitive Sturmian morphism.

2 Preliminaries

A finite set of symbols is an alphabet $\mathcal{A}$. A finite word of length n over $\mathcal{A}$ is a string $u = u_0u_1\cdots u_{n-1}$, where $u_i \in \mathcal{A}$ for all $i = 0, 1, \ldots, n-1$. The length of u is denoted by $|u| = n$. The set of all finite words over $\mathcal{A}$ is denoted by $\mathcal{A}^*$, the empty word is ϵ and $A^+ = A^* \setminus \{\epsilon\}$. An infinite word over $\mathcal{A}$ is a sequence $\mathbf{u} = u_0u_1u_2\cdots = (u_i)_{i\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$ with $u_i \in \mathcal{A}$ for all $i \in \mathbb{N} = 0, 1, 2, \ldots$

If a word $u \in \mathcal{A}^*$ is a concatenation of three (possibly empty) words x, y and z from A^*, i.e. $u = xyz$, the word x is a prefix of u, z is a suffix of u and z is a factor of u. A factor is denoted by $y \sqsubset u$. We put $x^{-1}u = yz$ and $uz^{-1} = xy$. Similarly, $w \in \mathcal{A}^*$

is a factor of $\mathbf{u} = u_0u_1u_2\cdots$, denoted by $r \sqsubset \mathbf{u}$, if there are indices $i \leq j$ such that $r = u_iu_{i+1}\dots u_j$. The index i is called the occurrence of r in $\mathbf{u}$.

The language $\mathcal{L}(\mathbf{u})$ of an infinite word $\mathbf{u}$ is the set of all its factors. The mapping $\mathcal{C}_\mathbf{u} : \mathbb{N} \mapsto \mathbb{N}$ defined by $\mathcal{C}_\mathbf{u}(n) = \#\{w \in \mathcal{L}(\mathbf{u}) : |w| = n\}$ is called the factor complexity of the word $\mathbf{u}$. An infinite word $\mathbf{u}$ is eventually periodic if $\mathbf{u} = wvvvvv\dots$ for some $v, w \in \mathcal{A}^*$. Otherwise, $\mathbf{u}$ is aperiodic. It is easy to prove that an infinite word $\mathbf{u}$ is eventually periodic if and only if its factor complexity $\mathcal{C}_\mathbf{u}$ is bounded. Moreover, the factor complexity of any aperiodic word satisfies $\mathcal{C}_\mathbf{u}(n) \geq n + 1$ for every $n \in \mathbb{N}$. An infinite word $\mathbf{u}$ with $\mathcal{C}_\mathbf{u}(n) = n + 1$ for every $n \in \mathbb{N}$ is called Sturmian word.

A word $w \sqsubset \mathbf{u}$ is called right special factor if there are at least two letters $a, b \in \mathcal{A}$ such that both wa and wb belong to the language $\mathcal{L}(\mathbf{u})$. Similarly, a word $w \sqsubset \mathbf{u}$ is called left special factor if there are at least two letters $a, b \in \mathcal{A}$ such that aw, bw belong to $\mathcal{L}(\mathbf{u})$. If a factor w is both left and right special then it is called bispecial factor.

A morphism over $\mathcal{A}^*$ is a mapping $\psi : \mathcal{A}^* \mapsto \mathcal{A}^*$ such that $\psi(vw) = \psi(v)\psi(w)$ for all $v, w \in \mathcal{A}^*$. The domain of the morphism ψ can be naturally extended to $\mathcal{A}^\mathbb{N}$ by

$$\psi(u_0u_1u_2\cdots) = \psi(u_0)\psi(u_1)\psi(u_2)\cdots .$$

A fixed point of the morphism ψ is an infinite word $\mathbf{u}$ such that $\psi(\mathbf{u}) = \mathbf{u}$.

A morphism ψ is non-erasing if $\psi(a) \neq \epsilon$ for all $a \in \mathcal{A}$. A morphism ψ is primitive if there exists a positive integer k such that the letter a occurs in the word $\psi^k(b)$ for each pair of letters $a, b \in \mathcal{A}$. And a morphism ψ is injective if for every $u, v \in \mathcal{A}^*$: $\psi(u) = \psi(v)$ implies that $u = v$.

2.1 Circularity

In [1] circularity is defined using the notion of synchronizing point (see Section 3.2 in [1] for details). We give here an equivalent definition employing the notion of interpretation.

Definition 2. *Let ψ be a non-erasing morphism over $\mathcal{A}^*$ with fixed point $\mathbf{u}$ and $u \sqsubset \mathbf{u}$. A triplet (p, v, s), where $p, s \in \mathcal{A}^*$ and $v \sqsubset \mathbf{u}$, is an* interpretation of the word *u if $\psi(v) = pus$.*

Definition 3. *Let ψ be a morphism over $\mathcal{A}^*$ with fixed point $\mathbf{u}$. We say that two interpretations (p, v, s) and (p', v', s') of a word $u \sqsubset \mathbf{u}$ are* synchronized at position *k, $0 \leq k \leq |u|$, if there exist indices i, j such that*

$$\varphi(v_1 \dots v_i) = pu_1 \dots u_k \quad \textit{and} \quad \varphi(v'_1 \cdots v'_j) = p'u_1 \cdots u_k\,,$$

where $v = v_1 \cdots v_n \in \mathcal{A}^n, v' = v'_1 \cdots v'_m \in \mathcal{A}^m$ and $u = u_1 \cdots u_\ell \in \mathcal{A}^\ell$. (If $k = 0$, we put $u_1 \cdots u_k = \epsilon$.) Two interpretations that are not synchronized at any position are called non-synchronized*. We say that a word $u \sqsubset \mathbf{u}$ has a* synchronizing point at position *k if all its interpretations are pairwise synchronized at position k.*

Definition 4. *Let ψ be a injective morphism over $\mathcal{A}^*$ with fixed point $\mathbf{u}$. We say that ψ is* circular (on $\mathcal{L}(\mathbf{u})$) *if there is a positive integer Z, called a* synchronizing delay*, such that any $u \sqsubset \mathbf{u}$ longer than Z has a synchronizing point. The minimal constant Z with this property is denoted by Z_{min}.*

Example 5. The word $\mathbf{u} = 010010100100101001010\cdots$ is the fixed point of the morphism $\psi : 0 \to 010, 1 \to 01$. For example, the factor 10 is non-synchronized, however, the factor 01 has a synchronizing point at position 0 (before letter 0): $|01$. In fact, it is easy to see that every factor of length 3 has its synchronizing point: $0|01$, $|010$, $10|0$, $1|01$. Since the minimal value of the synchronizing delay represents the length of the longest factor without synchronizing point, the minimal value of the synchronizing delay it this case is 2.

2.2 Sturmian words and morphisms

Sturmian words appear in many various mathematical concepts and so there is a lot of equivalent definitions. For example, any Sturmian word $\mathbf{u}$ can be identified with an upper or lower mechanical word. A mechanical word is described by two parameters: slope and intercept. The slope is an irrational number $\alpha \in (0,1)$ and the intercept is a real number $\rho \in [0,1)$. To define the lower mechanical word $\mathbf{s}_{\alpha,\rho} = (s_n)$ we put $I_0 = [0, 1-\alpha)$. The n^{th} letter of $\mathbf{s}_{\alpha,\rho}$ is as follows:

$$s_n = \begin{cases} 0 & \text{if the number } \alpha n + \rho \mod 1 \text{ belongs to } I_0\,, \\ 1 & \text{otherwise}\,. \end{cases}$$

The definition of the upper mechanical word $\mathbf{s}'_{\alpha,\rho} = (s'_n)$ is analogous, it just uses the interval $I_0 = (0, 1-\alpha]$. Let us stress that $s_n \neq s'_n$ for at most one index $n \in \mathbb{N}$. All upper and lower mechanical words are Sturmian and any Sturmian word equals to a lower or to an upper mechanical word. Language of a mechanical word depends only on α. Many further properties of Sturmian words can be found in [7].

A morphism ψ is called Sturmian if $\psi(\mathbf{u})$ is Sturmian word for any Sturmian word $\mathbf{u}$. It is easy to prove that every Sturmian morphism is injective. As mentioned in Introduction, Mossé [9] proved the following theorem: Every injective and primitive morphism is circular. Since we study only primitive Sturmian morphisms, these morphisms are always circular.

We will work with these four Sturmian morphisms:

$$\varphi_a : \begin{cases} 0 \to 0 \\ 1 \to 10 \end{cases} \qquad \varphi_b : \begin{cases} 0 \to 0 \\ 1 \to 01 \end{cases} \qquad \varphi_\alpha : \begin{cases} 0 \to 01 \\ 1 \to 1 \end{cases} \qquad \varphi_\beta : \begin{cases} 0 \to 10 \\ 1 \to 1 \end{cases}$$

and with the monoid $\mathcal{M}$ generated by them, i.e. $\mathcal{M} = \langle \varphi_a, \varphi_b, \varphi_\alpha, \varphi_\beta \rangle$. For a non-empty word $u = u_0 \cdots u_{n-1}$ over the alphabet $\{a, b, \alpha, \beta\}$ we put

$$\varphi_u = \varphi_{u_0} \circ \varphi_{u_1} \circ \cdots \circ \varphi_{u_{n-1}}.$$

Note that the monoid $\mathcal{M}$ is not free. It is easy to show that for any $k \in \mathbb{N}$ we have

$$\varphi_{\alpha a^k \beta} = \varphi_{\beta b^k \alpha} \quad \text{and} \quad \varphi_{a \alpha^k b} = \varphi_{b \beta^k a}\,.$$

Moreover, Theorem 2.3.14 in [7] says that these two relations are the only relations on the monoid $\mathcal{M}$.

Remark 6. The morphism $E : 0 \to 1, 1 \to 0$ cannot change the factor complexity of an infinite word and so E is clearly Sturmian morphism. But E does not belong to the monoid $\mathcal{M}$, in fact, E is the only missing morphism. More precisely, any Sturmian morphism ψ either belongs to $\mathcal{M}$ or $\psi = \eta \circ E$, where $\eta \in \mathcal{M}$. To generate the whole monoid of Sturmian morphisms, usually denoted by St, one needs only three morphisms, say E, φ_a and φ_b. We have

$$\varphi_\alpha = E\varphi_a E \quad \text{and} \quad \varphi_\beta = E\varphi_b E. \tag{1}$$

Our aim is to study the fixed points of Sturmian morphisms. If $\mathbf{u}$ is a fixed point of ψ, it is also a fixed point of ψ^2. Due to (1), the square ψ^2 always belongs to $\mathcal{M}$. Therefore we may restrict ourselves to fixed points of morphisms from $\mathcal{M}$.

Example 7. The Fibonacci word is the fixed point of the morphism $\tau : 0 \to 01, 1 \to 0$. Morphism τ is Sturmian, but $\tau \notin \mathcal{M}$. We see that $\tau = \varphi_b \circ E$ and by the relations (1) we have $\tau^2 = \varphi_b\varphi_\beta$.

It is easy to prove that a Sturmian morphism φ_w from the monoid $\mathcal{M}$ is primitive if and only if w contains at least one letter from both sets $\{a, b\}$ and $\{\alpha, \beta\}$.

2.3 Conjugate morphisms

We say that a morphism $\varphi : 0 \to w_1, 1 \to w_2$ over $\{0, 1\}^*$ has 1-conjugate, denoted by $\mathrm{conj}_1(\varphi)$, if the last letters of the words w_1 and w_2 are equal. If we denote this letter by x, we put

$$\mathrm{conj}_1(\varphi) : \begin{cases} 0 \to xw_1x^{-1} \\ 1 \to xw_2x^{-1} \end{cases}$$

or equivalently, $\psi = \mathrm{conj}_1(\varphi)$ if there exists a letter $x \in \{0, 1\}$ such that

$$x\varphi(v) = \psi(v)x \quad \text{for each } v \in \{0, 1\}^* .$$

Example 8. In this notation, $\varphi_b = \mathrm{conj}_1(\varphi_a)$ and $\varphi_\beta = \mathrm{conj}_1(\varphi_\alpha)$ as

$$0\varphi_a(v) = \varphi_b(v)0 \quad \text{and} \quad 1\varphi_\alpha(v) = \varphi_\beta(v)1 \quad \text{for each } v \in \{0, 1\}^* .$$

We say that non-erasing morphisms φ and ψ are conjugate if one can be reached from the other one by applying the mapping conj_1 repetitively.

Let ψ be a non-erasing morphism. Denote by $\mathcal{J}_\psi$ the set of all morphism which are conjugate with ψ. Obviously, we get for any $\varphi, \varphi' \in \mathcal{J}_\psi$ that $|\varphi(0)| = |\varphi'(0)|$ and $|\varphi(1)| = |\varphi'(1)|$. Let us put $|\varphi(1)| + |\varphi(0)| = L$. If the morphism ψ is Sturmian, then, by Proposition 2.3.21 in [7], the cardinality of $\mathcal{J}_\psi$ is $L - 1$. Therefore, there are $L - 1$ morphisms in $\mathcal{J}$ and they are all mutually conjugate.

Finally, let us notice that conjugacy could be analogously done also in the opposite direction, in that case the common letter goes from the beginning of images to the end of images. Indeed, the set $\mathcal{J}_\psi$ remains the same.

Example 9. Consider the Sturmian morphism $\psi = \varphi_{b\beta}$: $0 \to 010$, $1 \to 01$. Since $L = 5$, the set $\mathcal{J}_\psi$ contains four distinct morphisms: $\varphi_{b\beta}$: $0 \to 010$, $1 \to 01$, $\varphi_{a\beta}$: $0 \to 100$, $1 \to 10$, $\varphi_{b\alpha}$: $0 \to 001$, $1 \to 01$ and $\varphi_{a\alpha}$: $0 \to 010$, $1 \to 10$. We can also see all these conjugates from the following notation:

$$\frac{\psi(0)u}{\psi(1)u} = \frac{\mathbf{010}\;010}{\mathbf{01}\;010} = \frac{0\;\mathbf{100}\;10}{0\;\mathbf{10}\;10} = \frac{01\;\mathbf{001}\;0}{01\;\mathbf{01}\;0} = \frac{010\;\mathbf{010}}{010\;\mathbf{10}}\,, \tag{2}$$

where u is a sequence of letters which one by one moves from the beginning of images to the end of images. Clearly, it is: $|u| = L - 2$.

3 Upper bound on synchronozing delay

To bound the value of the synchronizing delay, we use knowledge of the structure of bispecial factors in fixed points of Sturmian morphisms. There are several concepts which enable us to describe the structure of bispecial factors in Sturmian words. We use the method similar to the basic ideas from [3].

Let ψ be a primitive Sturmian morphism with a fixed point $\mathbf{u}$. First, we study how bispecial factors change under the application of one of the following morphisms: $\varphi_a, \varphi_b, \varphi_\alpha, \varphi_\beta$. In particular, we show that every bispecial factor longer than 1 has at least one synchronizing point under any of these morphisms. By repeating this process, we can show that every long enough bispecial factor has at least one synchronizing point under the morphism ψ too. Then, we find some suitable bispecial factor r and we bound its length. Finally, we determine how often the bispecial factor r has to appear in $\mathbf{u}$. As a consequence, we are capable to find a length K such that every factor longer than K contains at least one occurrence of a bispecial factor r and so at least one synchronizing point. But it means that we have a upper bound on the value of the synchronizing delay of ψ: $Z_{min} \leq K$.

3.1 Preimages of bispecial factors

Because of (1), the role of φ_a and φ_α and, analogously, the role of φ_b and φ_β are symmetric, so we focus only on images under the morphisms φ_a and φ_b. We use results from [7], more precisely, a small modification of Proposition 2.3.2:

Proposition 10 ([7]). *Let* $\mathbf{x}$ *be an infinite word.*

- *If* $\varphi_b(\mathbf{x})$ *is Sturmian, then* $\mathbf{x}$ *is Sturmian.*
- *If* $\varphi_a(\mathbf{x})$ *is Sturmian and* $\mathbf{x}$ *starts with the letter* 1*, then* $\mathbf{x}$ *is Sturmian.*

Lemma 11. *Let* $\mathbf{u}$ *and* $\mathbf{u}'$ *be Sturmian words such that* $\mathbf{u} = \varphi_b(\mathbf{u}')$. *Let* w *be a bispecial factor of* $\mathbf{u}$ *with* $|w| > 1$. *Then there is a* $w' \sqsubset \mathbf{u}'$ *such that* $w = \varphi_b(w')0$. *Moreover, this factor* w' *is unique, it is a bispecial factor of* $\mathbf{u}'$ *and all the interpretations of* w *are synchronized both at the beginning and at the end of the factor* $\varphi_b(w')$.

Proof. Take a Sturmian word $\mathbf{u}, \mathbf{u}'$ such that $\mathbf{u} = \varphi_b(\mathbf{u}')$. By the form of morphism φ_b, $\mathbf{u}$ can be written as $\mathbf{u} = 0^{k_0}10^{k_1}10^{k_2}1\cdots$, where $k_i > 0$ for all $i \in \mathbb{N}$. Take a bispecial factor w of $\mathbf{u}$ with $|w| > 1$. Then the word w must both begin and end with 0. So we can easily find a word w' such that $w = \varphi_b(w')0$, it suffices to cut w into blocks 0 and 01 (we omit the last letter 0) and desubstitute 0 and 01 for 0 and 1 respectively. It remains to show that this w' is unique. Indeed, it follows from the fact that the morphism φ_b is injective. The factor w' is obviously a bispecial factor, because of the form of φ_b and the fact that w is a bispecial factor. Since there is a synchronizing point before every letter 0, all the interpretations of w are synchronized both at the beginning and at the end of the factor $\varphi_b(w')$. In other words, the occurrences of bispecial factor w in $\mathbf{u}$ are one-to-one to occurrences of bispecial factor w' in $\mathbf{u}'$.

□

Lemma 12. *Let $\mathbf{u}$ and $\mathbf{u}'$ be Sturmian words such that $\mathbf{u} = \varphi_a(\mathbf{u}')$ and $\mathbf{u}$ starts with the letter 1. Let w be a bispecial factor of $\mathbf{u}$ with $|w| > 1$. Then there is a $w' \sqsubset \mathbf{u}'$ such that $w = 0\varphi_a(w')$. Moreover, this factor w' is unique, it is bispecial factor of $\mathbf{u}'$ and all the interpretations of w are synchronized both at the beginning and at the end of factor $\varphi_a(w')$.*

Proof. Take a Sturmian words $\mathbf{u}, \mathbf{u}'$ such that $\mathbf{u} = \varphi_a(\mathbf{u}')$. By the form of morphism φ_a, $\mathbf{u}$ can be written as $\mathbf{u} = 10^{k_0}10^{k_1}10^{k_2}1\cdots$, where $k_i > 0$ for all $i \in \mathbb{N}$. Take a bispecial factor w of $\mathbf{u}$ with $|w| > 1$. Then the word w must both begin and end with 0. So we can easily find a word w' such that $w = 0\varphi_a(w')$, it suffices to cut w into blocks 0 and 10 (we omit the first letter 0) and desubstitute 0 and 10 for 0 and 1 respectively. This w' is unique, since the morphism φ_a is injective. The factor w' is obviously bispecial factor, because of the form of φ_a and the fact that w is a bispecial factor. Since there is a synchronizing point after every letter 0, all the interpretations of w are synchronized both at the beginning and at the end of factor $\varphi_a(w')$. In other words, the occurrences of bispecial factor w in $\mathbf{u}$ are one-to-one to occurrences of bispecial factor w' in $\mathbf{u}'$.

□

The only case which is not covered by Lemmas 11 and 12, namely the case that $\mathbf{u} = \varphi_a(\mathbf{u}')$ and $\mathbf{u}$ begins with 0, can be transformed to one of the previous cases.

Lemma 13. *Let $\mathbf{u}$ be a Sturmian word such that $\mathbf{u}$ starts with the letter 0 and $\mathbf{u} = \varphi_a(\mathbf{u}')$ for some word $\mathbf{u}'$. Then there exists a Sturmian word $\mathbf{v}$ such that $\mathbf{u}' = 0\mathbf{v}$ and $\mathbf{u} = \varphi_b(\mathbf{v})$.*

Proof. By using the following easy observation $\varphi_a(0w) = \varphi_b(w0)$ for every $w \in \{0,1\}^*$, we have $\mathbf{u} = \varphi_a(\mathbf{u}') = \varphi_a(0\mathbf{v}) = \varphi_{\mathbf{b}}(\mathbf{v})$.

Prove the observation by induction on $|w|$. The first step $w = \epsilon$ is trivial since $\varphi_a(0) = \varphi_b(0)$. Suppose that $\varphi_a(0w) = \varphi_b(w0)$ for every $w \in \{0,1\}^*$. Then

$$\varphi_a(0w1) = \varphi_a(0w)\varphi_a(1) = \varphi_a(0w)10 = \varphi_b(w0)10 = \varphi_b(w)010 = \varphi_b(w10)\,,$$
$$\varphi_a(0w0) = \varphi_a(0w)\varphi_a(0) = \varphi_b(w0)0 = \varphi_b(w00)\,.$$

□

In accordance with Lemma 13, in this case we will use the Sturmian word $\mathbf{v}$ instead of the word $\mathbf{u}'$, since we need to maintain the Sturmian property. Indeed, in this case the occurrences of bispecial factor w' in $\mathbf{v}$ are not exactly one-to-one to occurrences of bispecial factor w in $\varphi_a(\mathbf{v})$, the first occurrence of w' in $\mathbf{v}$ (in this case w' is a prefix of $\mathbf{v}$) does not have its complete corresponding occurrence of w in $\varphi_a(\mathbf{v})$ – the first letter 0 is missing. However, this exception is not substantial and we can omit it without lose of correctness: there are infinitely many occurrences of any bispecial factor in every Sturmian word and the new uncomplete bispecial factor is not important at all.

3.2 Suitable bispecial factor

Let $\psi = \varphi_w$ be a primitive Sturmian morphism, $w = w_0 \cdots w_k$ and $\mathbf{u}$ be a fixed point of ψ. Without lose of generality, we can suppose that letter 0 is more frequent in $\mathbf{u}$, since the exchange of letters $0 \leftrightarrow 1$ cannot change the value of the synchronizing delay. It means that 0 is a bispecial factor in $\mathbf{u}$. The aim of this section is to find the shortest bispecial factor r in $\mathbf{u}$ containing $\psi(0)$, prove that r has at least one synchronizing point and bound its length.

First, we apply the morphisms $\varphi_{w_k} \in \{\varphi_a, \varphi_b, \varphi_\alpha, \varphi_\beta\}$ on the infinite word $\mathbf{u}$, the choice of the morphism depends on the last letter of the word w. Because of Lemma 11 or 12 (or their analoques for φ_β, φ_α), the infinite word $\varphi_{w_k}(\mathbf{u})$ is Strumian and we obtain a new bispecial factor $r_1 = s_1\varphi_{w_k}(0)p_1$, where $s_1, p_1 \in \{\epsilon, 0\}$, from the bispecial factor 0. Moreover, the bispecial factor r_1 has a synchronizing point under φ_{w_k} and the occurrences of 0 in $\mathbf{u}$ and r_1 in $\varphi_{w_k}(\mathbf{u})$ are one-to-one. Clearly, we can continue in the same way: application of the morphism $\varphi_{w_{k-1}}$ leads to the new infinite word $\varphi_{w_{k-1}w_k}(\mathbf{u})$ and the bispecial factor $r_2 = s_2\varphi_{w_{k-1}}(r_1)p_2 = s_2\varphi_{w_{k-1}}(s_1)\varphi_{w_{k-1}w_k}(0)\varphi_{w_{k-1}}(p_1)p_2$, which has a synchroninizing point under $\varphi_{w_{k-1}}$ and its occurrences in $\varphi_{w_{k-1}w_k}(\mathbf{u})$ are one-to-one to occurrences of 0 in $\mathbf{u}$.

After repeating this process k-times, we obtain the original infinite word $\mathbf{u}$ again and the bispecial factor $r_k = s_k\varphi_{w_1}p_k = \cdots = s\psi(0)p$. This bispecial factor r_k has some synchronizing point under φ_{w_1} and its occurrences in $\mathbf{u}$ are one-to-one to occurrences of 0 in $\mathbf{u}$. But it means that r_k must have at least one synchronizing point under the morphism ψ too. One can also realize that r_k is the shortest bispecial factor of $\mathbf{u}$ containing $\psi(0)$.

It remains to bound the length of words s and p on the length of $\psi(0)$ and $\psi(1)$. As follows from the notation (2) in Example 9, the number $|s| + |p|$ has to be equal to the length of the word u (from Example 9): $|u| = L-2 = \psi(0)+\psi(1)-2$. Now we summarize all these result in the following observation.

Observation 14. *Let ψ be a primitive Sturmian morphism with a fixed point $\mathbf{u}$ such that 0 is more frequent letter in $\mathbf{u}$. Then the shortest bispecial factor r in $\mathbf{u}$ containing $\psi(0)$ has at least one synchronizing point under ψ and its length is bounded by $|r| \leq 2|\psi(0)| + |\psi(1)| - 2$.*

3.3 Occurrences of suitable bispecial factor

Finally, we have to determine how often the bispecial factor r appears in $\mathbf{u}$, more precisely, we have to determine the length of the longest factor of $\mathbf{u}$ which does not have to contain the whole factor r as its factor.

Let us denote by v the longest factor of $\mathbf{u}$ which does not contain any occurrence of r (the beginning of the factor r). Since the word $\mathbf{u}$ is Sturmian and the letter 0 is more frequent in $\mathbf{u}$, the word 11 is not a factor of $\mathbf{u}$. We also know the occurrences of 0 and r always coincide in $\mathbf{u}$. Based on this two observations one can realize that $|v| \leq |\psi(0)| + |\psi(1)| - 1$. Therefore, we can bound as follows:

$$L \leq |v| + |r| = |\psi(0)| + |\psi(1)| - 1 + 2|\psi(0)| + |\psi(1)| - 2 = 3|\psi(0)| + 2|\psi(1)| - 3\,.$$

In other words, every word longer that L has to contain the word r as its factor and so has to contain at least one synchronizing point under ψ. This concludes the proof since now he have

$$Z_{min} \leq L \leq 3|\psi(0)| + 2|\psi(1)| - 3\,,$$

which is the statement of Theorem 1.

References

[1] J. Cassaigne. *An algorithm to test if a given circular HD0L-language avoids a pattern.* In 'World Computer Congress'94 (North-Holland, 1994)', pp. 459–464.

[2] F. Durand, J. Leroy. *The constant of recognizability is computable for primitive morphisms.* Preprint. *arXiv:1610.05577.*

[3] K. Klouda. *Bispecial factors in circular non-pushy D0L languages.* Theoret. Comput. Sci. **445** (2012), 63–74.

[4] K. Klouda, K. Medková. *Synchronizing delay for binary uniform morphisms.* Theoret. Comput. Sci. **615** (2016), 12–22.

[5] K. Klouda, Š. Starosta. *An algorithm enumerating all infinite repetitions in a D0L-system.* J. Discrete Algorithms **33** (2015), 130–138.

[6] K. Klouda, Š. Starosta. Characterization of circular D0L- systems. Submitted. *arXiv:1401.0038.*

[7] M. Lothaire. *Algebraic Combinatorics on Words, Encyclopedia of Mathematics and its Applications*, volume 90. Cambridge University Press (2002).

[8] F. Mignosi, P. Séébold. *If a D0L language is k-power free then it is circular.* In 'ICALP '93: Proceedings of the 20th International Colloqium on Automata, Languages and Programming (London, 1993)', Spriger-Verlag, pp. 507–518.

[9] B. Mossé. *Notions de reconnaissabilité pour les substitutions et complexité des suites automatiques*, Bull. Soc. Math. France **124** (1996), 101–108. Cambridge University Press (2005).

Reaction-Diffusion Systems with Two Unilateral Sources*

Josef Navrátil

5th year of PGS, email: navrajos@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Milan Kučera, Department of Evolution Differential Equations
Institute of Mathematics, CAS

Martin Väth, Fachbereich Mathematik und Informatik
Free University of Berlin

Abstract. The effect of unilateral sources on the existence of patterns in reaction-diffusion equations has been studied in a vast number of papers. There was proved that this type of sources leads to an emergence of patterns for diffusion rates, for which this cannot happen in systems without sources. In this paper, basic regularity theorems and Hopf lemma are used to prove the existence of bifurcation points in a system with two unilateral condition and the existence of a new class of non-homogeneous solutions (i.e. patterns). The explicit formula for such bifurcation points is derived as well as the form of the solutions.

Keywords: reaction-diffusion equations, bifurcation, unilateral sources

Abstrakt. Vliv jednostranných zdrojů na existenci vzorů v systémech reakce-diffuze byl studová n v mnoha článcích. Ukazuje se, že tento typ zdrojů vede k existenci vzorů i v systémech s hodnotami difúzních parametrů, pro které by bez přítomnosti zdrojů k formování vzorů nedošlo. V tomto článku je pomocí základních vět o regularitě parciálních diferenciálních rovnic a Hopfova lemmatu dokázána existence bifurkačních bodů v množině takových parametrů. Dále je zde odvozen explicitní vzorec pro výpočet těchto bodů a popsána konstrukce příslušných řešení.

Klíčová slova: rovnice reakce-difuze, bifurkace, jednostranné zdroje

1 Introduction

The aim of this paper is to study bifurcation from zero of stationary solutions of the reaction-diffusion system

$$\begin{aligned} d_1\triangle u + b_{11}u + b_{12}v + n_1(u,v) = 0 \;\text{ in } \Omega\backslash\Omega_U, \\ d_2\triangle v + b_{21}u + b_{22}v + n_2(u,v) = 0 \;\text{ in } \Omega\backslash\Omega_U, \end{aligned} \tag{1}$$

*This work has been supported by the grant SGS16/239/OHK4/3T/14

$$
\begin{aligned}
&u \geq 0, \quad d_1 \triangle u + b_{11}u + b_{12}v + n_1(u,v) \leq 0 \ \text{ in } \ \Omega_U,\\
&u \cdot (d_1 \triangle u + b_{11}u + b_{12}v + n_1(u,v)) = 0 \ \text{ in } \ \Omega_U,\\
&v \geq 0, \quad d_2 \triangle v + b_{21}u + b_{22}v + n_2(u,v) \leq 0 \ \text{ in } \ \Omega_U,\\
&v \cdot (d_2 \triangle v + b_{21}u + b_{22}v + n_2(u,v)) = 0 \ \text{ in } \ \Omega_U,\\
&u = v = 0 \ \text{ on } \partial\Omega,
\end{aligned}
\tag{2}
$$

in a bounded domain $\Omega \subset \mathbb{R}$ with Lipschitz boundary and with unilateral obstacles in the set $\Omega_U \subset \mathbb{R}$. This is a system containing a mechanism which prohibits the decrease of concentrations of u an v below zero in the area Ω_U.

Let $d_1 > 0$ be fixed, $d_2 \in \mathbb{R}$ be a bifurcation parameter and $n_1, n_2 \equiv 0$. If the obstacle is not present, i.e. $\Omega_U = \emptyset$, then under the assumption

$$b_{11} > 0 > b_{22}, \ b_{21} > 0 > b_{12}, \ \text{Tr}\, B = b_{11} + b_{22} < 0, \ \det B = b_{11}b_{22} - b_{12}b_{21} > 0, \tag{3}$$

the set of all positive critical points can visualized as a system of hyperbolas in the space $\mathbb{R}^2$ with the asymptotes x_i, see Fig. 1. More precisely, for any fixed positive

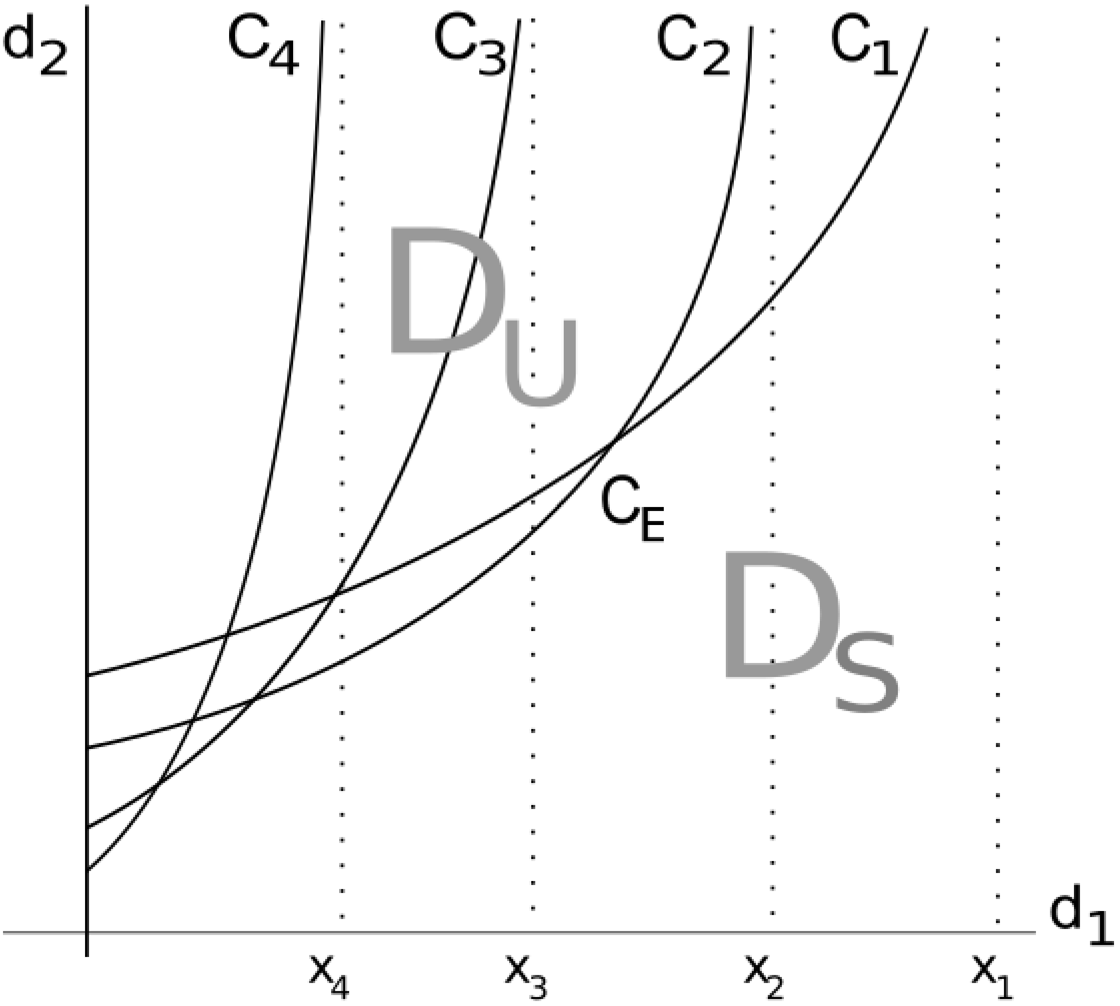


Figure 1: Sketch of hyperbolas.

$d_1 \in (0, x_1) \backslash \{x_2, x_3, \cdots\}$ it is possible to find a value d_2 for which there exists a nontrivial solution of the system without obstacles. The sets D_S, D_U are called the domain of stability and instability respectively. In the domain of stability, the trivial solution of the system (1) with $\Omega_U = \emptyset$ and with the Dirichlet b.c. is stable and hence, there cannot appear any non-homogeneous solutions. On the other hand, in the domain of instability the trivial solution of this system is unstable, and there are nontrivial non-homogeneous stationary solutions, i.e. patterns.

In a general system with $n_1, n_2 \neq 0$, (4) and (3), these critical points can be under additional assumptions also bifurcation points. For particular systems there can exists non-homogeneous solutions (patterns) even in D_S, but there is no guarantee that it will happen for an arbitrary system. Let us note that the assumptions (4) guarantee that the system has a trivial solution.

However, if the unilateral sources are active and (3) is true, there exist a branch of critical points, which interfere into D_S, and therefore there are nontrivial solutions. Under some additional assumptions these critical points can be also bifurcation points of the problem (1), (2); see Theorem 1. This shows that the addition of unilateral sources leads to an occurrence of non-homogeneous stationary solutions, i.e. patterns, for the diffusion parameters, for which it is impossible in the system without these unilateral sources. In addition to the previously published result [1], the new bifurcation branch will be described by an *exact* formula, depending only on parameters b_{ij}, d_1 of the system and eigenvalues of Laplacian on the set $\Omega \backslash \Omega_U$ with Dirichlet boundary conditions. The analytic results will be demonstrated on particular examples.

This paper is a natural generalization of the results proved in [3] for the case of Laplace equation. Although the generalization to the system of two partial differential equations is straightforward, there are several technical problems which have to be treated.

1.1 Abstract formulation

Let $\Omega \subset \mathbb{R}^2$ and $\Omega_U \subset \mathbb{R}^2$ be bounded domains with a Lipschitz and C^2 boundary respectively. Let $\overline{\Omega_U} \subset \Omega$. The nonlinear functions $n_1, n_2 \in C^1(\mathbb{R}^2)$ are supposed to satisfy

$$n_1(0,0) = n_2(0,0) = 0, \quad n_1'(0,0) = n_2'(0,0) = 0, \tag{4}$$

where prime denotes the total derivative, and the growth conditions

$$\begin{aligned} &|n_1(\xi,\chi)| + |n_2(\xi,\chi)| \leq C(1 + |\xi|^{p-1} + |\chi|^{p-1}) \text{ for all } \chi, \xi \in \mathbb{R} \\ &\left|\frac{\partial n_i}{\partial \xi}(\xi,\chi)\right| + \left|\frac{\partial n_i}{\partial \chi}(\xi,\chi)\right| \leq C(1 + |\xi|^{p-2} + |\chi|^{p-2}) \text{ for } i = 1,2 \text{ for all } \chi, \xi \in \mathbb{R}, \end{aligned} \tag{5}$$

with $2 < p < \infty$. The Sobolev space $W_0^{1,2}(\Omega)$ and a convex cone K will be defined in a standard way as

$$W_0^{1,2}(\Omega) := \{u \in W^{1,2}(\Omega) |\ u|_{\partial\Omega} = 0 \text{ in the sense of traces}\}, \tag{6}$$

$$K := \{u \in W_0^{1,2}(\Omega) |\ u \geq 0 \text{ on } \Omega_U\}. \tag{7}$$

The scalar product and norm on this space will be defined by

$$\langle u, v\rangle = \int_\Omega \nabla u \cdot \nabla v \text{ dx}, \quad \|u\| = \left(\int_\Omega |\nabla u|^2 \text{ dx}\right)^{\frac{1}{2}} \quad \text{for all } u, v \in W_0^{1,2}(\Omega).$$

The weak formulation of the system (1) is

Find $u, v \in K$:

$$\begin{aligned} &\int_\Omega d_1 \nabla u \cdot \nabla(\varphi - u) - b_{11} u(\varphi - u) - b_{12} v(\varphi - u) - n_1(u,v)(\varphi - u) \geq 0, \\ &\int_\Omega d_2 \nabla v \cdot \nabla(\psi - v) - b_{21} u(\psi - v) - b_{22} v(\psi - v) - n_2(u,v)(\psi - v) \geq 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for all } \varphi, \psi \in K. \end{aligned} \tag{8}$$

The linearization of this system is a problem

$$\text{Find } u, v \in K : \quad \begin{aligned} &\int_\Omega d_1 \nabla u \cdot \nabla(\varphi - u) - b_{11}(\varphi - u) - b_{12}(\varphi - u) \geq 0 \ \text{ for all } \varphi \in K, \\ &\int_\Omega d_2 \nabla v \cdot \nabla(\psi - v) - b_{21}(\psi - v) - b_{22}(\psi - v) \geq 0 \ \text{ for all } \psi \in K. \end{aligned} \tag{9}$$

Let $d_1 \in (0, y_1)$ be fixed. A significant role will play here two Laplace eigenvalue problems. The first one is

$$\begin{aligned} &\Delta u + \hat{\kappa} u = 0 \ \text{ in } \Omega \backslash \Omega_U, \\ &u = 0 \ \text{ on } \partial\Omega \cup \partial\Omega_U, \end{aligned} \tag{10}$$

and the second one is

$$\begin{aligned} &\triangle u + \kappa u \ = 0 \text{ in } \Omega, \\ &u = 0 \ \text{ on } \partial\Omega. \end{aligned} \tag{11}$$

Remark 1. *If is well known, that the first (smallest) eigenvalue $\hat{\kappa}_1$ of the problem (10) is simple, and the respective eigenfunction does not change its sign in the set $\Omega \backslash \Omega_U$. The second smallest eigenvalue of (10) will be denoted as $\hat{\kappa}_2$.*

The first eigenvalue and the respective eigenfunction of (11) have the same properties. Because the eigenvalues of Laplacian with Dirichlet b.c. are monotone w.r.t. domain, there is $\hat{\kappa}_1 < \kappa_1$.

Remark 2. *Let $\Omega_U = \emptyset$, i.e. the obstacle is not present. It can be proved that the k-th hyperbola from the Fig. 1 is described by the formulas*

$$d_{2,k}(d_1) = \frac{1}{\kappa_k} \left(\frac{b_{12} b_{21}}{d_1 \kappa_k - b_{11}} + b_{22} \right),$$

see e.g. [2]. If $d_1 \in (b_{11}/\kappa_2, b_{11}/\kappa_1)$, then $d_{2,1}$ is positive and $d_{2,k}$ is negative for any $k \geq 2$. And in general, if $d_1 \in (b_{11}/\kappa_{i+1}, b_{11}/\kappa_i)$, then $d_{2,j} > 0$ for all $j \leq i$ and $d_{2,j} < 0$ for all $j > i$. The envelope of these hyperbolas is denoted by D_E. The set which is to the right from the envelope is called the domain of stability, D_S and the set to the left from the envelope is called the domain of instability, D_U.

Definition 1. *The point $d_2 > 0$ is a critical point of the system (9) with fixed $d_1 > 0$ if and only if there exists a solution $u, v \in K$, $(u, v) \neq 0$ of this system.*

Definition 2. *The point $d_2 > 0$ is a bifurcation point of the system (8) with fixed $d_1 > 0$ if and only if in any neighborhood of $(d_2, 0, 0)$ in $\mathbb{R} \times K^2$ there exists $(\tilde{d_2}, u, v) \in \mathbb{R} \times K^2$ with $(u, v) \neq 0$ solving this system.*

2 Main Theorem

Theorem 1. *Let $d_1 \in (b_{11}/\hat{\kappa}_2, b_{11}/\hat{\kappa}_1)$. Under the assumptions (3) – (5) the number*

$$d_2^K := \frac{b_{12}b_{21}}{\hat{\kappa}_1(d_1\hat{\kappa}_1 - b_{11})} + \frac{b_{22}}{\hat{\kappa}_1} \tag{12}$$

is a bifurcation point of (8) with fixed d_1.

Let $\kappa_2 < \hat{\kappa}_1$. There exists $d_{1,m}, d_{1,M} \in (b_{11}/\hat{\kappa}_2, b_{11}/\hat{\kappa}_1)$ such that if $d_1 \in (d_{1,m}, d_{1,M})$, then $d_2^K \in D_S$.

Proof. First step is to prove that the point d_2^K is a critical point of the system (9) with fixed $d_1 \in (0, b_{11}/\hat{\kappa}_1)$. For further purposes we will define the space $W_0^{1,2}(\Omega \backslash \Omega_U)$ in the same way as $W_0^{1,2}(\Omega)$ in (6) and consider an auxiliary problem

$$\begin{aligned} d_1 \triangle u + b_{11}u + b_{12}v = 0 \quad &\text{in } \Omega\backslash\Omega_U, \\ d_2 \triangle v + b_{21}u + b_{22}v = 0 \quad &\text{in } \Omega\backslash\Omega_U, \end{aligned} \tag{13}$$

and with the Dirichlet b.c. on $\partial\Omega \cup \partial\Omega_U$. It can be by a direct computation verified that for (d_1, d_2^K) there exists a nontrivial solution of (13), with the respective eigenfunction

$$(u_0, v_0) = \left(\frac{b_{12}}{d_1\hat{\kappa}_1 - b_{11}} e_1, e_1 \right),$$

where e_1 is the eigenfunction respective to $\hat{\kappa}_1$ with unit norm, and because e_1 does not change its sign in $\Omega\backslash\Omega_U$, see Remark 1, it can be chosen either positive or negative a.e. in Ω. Even though the sign does not play role for such linear system, it will play a crucial role for variational inequality. For further purposes e_0 will be chosen *negative* a.e. Since also $b_{12} < 0$ and $d_1\kappa_1 - b_{11} < 0$, the functions u_0, v_0 have the same constant sign a.e. in Ω. Since $d_1 \in (\hat{\kappa}_2/b_{11}, \hat{\kappa}_1/b_{11})$ is fixed, and because $\hat{\kappa}_1$ is simple, there exists only one couple (u_0, v_0) (up to multiples) solving (13) with the parameters (d_1, d_2^K).

To get the bifurcation, the Dancer Theorem will be employed.

Theorem 2 (Dancer Theorem)**.** *Let $L : \mathbb{H} \to \mathbb{H}$ be a compact linear operator, $N : \mathbb{R} \times \mathbb{H} \to \mathbb{H}$ be a nonlinear compact operator, λ_0 be a simple characteristic value of the operator L, u_0 be the eigenfunction corresponding to the characteristic value λ_0. Moreover let for any bounded set $\mathcal{M} \subset \mathbb{R}$ the operator N satisfy a condition*

$$\lim_{\|u\|\to 0} \frac{N(\lambda, u)}{\|u\|} = 0 \quad \textit{uniformly for all } \lambda \in \mathcal{M}. \tag{14}$$

Denote S the closure of all solutions of the equation

$$\lambda u - Lu + N(\lambda, u) = 0 \tag{15}$$

with $u \neq 0$, i.e.

$$S = \overline{\{(\lambda, u) \mid u \neq 0,\ u \textit{ is a solution of } (15)\}}.$$

Then $(\lambda_0, 0) \in S$*, i.e.* λ_0 *is a bifurcation point of the equation (15). Denote* C *the component of* S *which contains* $(\lambda_0, 0)$*. Then* C *consists of two connected sets* C^+, C^-, $C = C^+ \cup C^-$ *such that*

$$C^+ \cap C^- \cap B((\lambda_0, 0); \rho) = \{(\lambda_0, 0)\} \textit{ and } C^\pm \cap \partial B((\lambda_0, 0); \rho) \neq 0,$$

where $B((\lambda_0, 0); \rho)$ *is a ball with sufficiently small radius* ρ*. The sets* C^+ *and* C^- *are either both unbounded or*

$$C^+ \cap C^- \neq \{(\lambda_0, 0)\}.$$

By use of (5) and Theorem about Nemyckii operator there can be defined the operators $A : W_0^{1,2}(\Omega \backslash \Omega_U) :\to W_0^{1,2}(\Omega \backslash \Omega_U)$, $N_1, N_2 : \left(W_0^{1,2}(\Omega \backslash \Omega_U)\right)^2 \to W_0^{1,2}(\Omega)$ as

$$\begin{aligned} \langle Au, w \rangle &= \int_\Omega uw \ \mathrm{dx}, \quad \text{for all } u, w \in W_0^{1,2}(\Omega \backslash \Omega_U), \\ \langle N_i(u, v), w \rangle &= \int_\Omega n_i(u, v) w \ \mathrm{dx} \quad \text{for all } u, v, w \in W_0^{1,2}(\Omega \backslash \Omega_U), \ i = 1, 2. \end{aligned}$$

Due to the compact embedding $W_0^{1,2}(\Omega \backslash \Omega_U) \hookrightarrow^c L^p(\Omega)$ the operators A, N_1, N_2 are compact. The weak formulation of the system

$$\begin{aligned} d_1 \triangle u + b_{11} u + b_{12} v + n_1(u, v) = 0 \ \text{ in } \Omega \backslash \Omega_U, \\ d_2 \triangle v + b_{21} u + b_{22} v + n_2(u, v) = 0 \ \text{ in } \Omega \backslash \Omega_U, \end{aligned} \tag{16}$$

is equivalent to a system of two operator equations

$$\begin{aligned} d_1 u - b_{11} A u - b_{12} A v - N_1(u, v) = 0, \\ d_2 v - b_{21} A u - b_{22} A v - N_2(u, v) = 0, \end{aligned}$$

and this system can be written in a form

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} d_1^{-1} & 0 \\ 0 & d_2^{-1} \end{bmatrix} \left(\begin{bmatrix} b_{11} A & b_{12} A \\ b_{21} A & b_{22} A \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} N_1(u, v) \\ N_2(u, v) \end{bmatrix} \right) = 0. \tag{17}$$

The linearization of this equation is

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} d_1^{-1} & 0 \\ 0 & d_2^{-1} \end{bmatrix} \left(\begin{bmatrix} b_{11} A & b_{12} A \\ b_{21} A & b_{22} A \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \right) = 0,$$

and as d_1 is fixed, it is a characteristic value problem

$$\mathbf{w} - \lambda(d_2) \mathbb{L} \mathbf{w} = 0,$$

where $\mathbf{w} = (u, v) \in W_0^{1,2}(\Omega \backslash \Omega_U)^2$, $\mathbb{L}$ is a linear compact operator (due to compactness of A), and $\lambda(d_2)$ is an characteristic value, depending on the parameter d_2. This problem is equivalent to a weak formulation of (13). Since $\hat{\kappa}_1$ is simple and $d_1 \in (b_{11} \hat{\kappa}_2, b_{11}/\hat{\kappa}_1)$ the characteristic value d_2^K is simple. The vector formulation of 17 is

$$\mathbf{w} - \lambda(d_2) \mathbb{L} \mathbf{w} + N(\lambda(d_2), \mathbf{w}) = 0,$$

which is suitable for Dancer Theorem. The operator $\mathrm{N}(u,v) := (N_1(u,v), N_2(u,v))$ is compact and due to (4) it satisfies (14). Hence, the assumptions of the Dancer Theorem are fulfilled and the point $\lambda(d_2^K)$ is according to this theorem a (global) bifurcation point of the equation (17) and therefore also of the equation (16) with Dirichlet b.c. and fixed d_1. Moreover, there exists two branches of solutions bifurcating in the directions $\pm(u_0, v_0)$ from $(d_2^K, 0) \in \mathbb{R} \times W_0^{1,2}(\Omega \backslash \Omega_U)$. More precisely, there exists two sequences $\{d_{2,n}, u_n^+, v_n^+\}$ $\{d_{2,n}, u_n^-, v_n^-\}$ of weak solutions of (16) with Dirichlet b.c. such that

$$\lim_{n\to\infty} d_{2,n} = d_2^K, \quad \lim_{n\to\infty} \frac{u_n^-}{\sqrt{\|u_n^-\|^2 + \|v_n^-\|^2}} = u_0, \quad \frac{v_n^-}{\sqrt{\|u_n^-\|^2 + \|v_n^-\|^2}} = v_0 \tag{18}$$

$$\lim_{n\to\infty} \frac{u_n^+}{\sqrt{\|u_n^+\|^2 + \|v_n^+\|^2}} = -u_0, \quad \frac{v_n^+}{\sqrt{\|u_n^+\|^2 + \|v_n^+\|^2}} = -v_0, \tag{19}$$

$$\lim_{n\to\infty} u_n^+ = \lim_{n\to\infty} u_n^- = \lim_{n\to\infty} v_n^+ = \lim_{n\to\infty} v_n^- = 0, \tag{20}$$

the limits of $u_n^\pm, v_n^\pm$ are w.r.t. $W_0^{1,2}(\Omega \backslash \Omega_U)$. Let us remind here that u_0, v_0 were chosen to be negative a.e. in Ω. For the purposes of this proof the branch $\{d_{2,n}, u_n^+, v_n^+\}$ will be discarded, and the sequence $\{d_{2,n}, u_n^-, v_n^-\}$ will be relabeled as $\{d_{2,n}, u_n, v_n\}$. The next step is to prove the regularity of solutions in a neighborhood of the set $\partial\Omega_U$.

Let Ω_V be a domain with C^2 boundary satisfying

$$\overline{\Omega \backslash (\Omega_U \cup \Omega_V)} \subset \Omega \backslash \Omega_U, \quad \partial\Omega_V \cap \partial\Omega_U = \partial\Omega_U. \tag{21}$$

The growth conditions (5) and standard regularity arguments can be used to prove that that $u_n|_{\Omega_V}, v_n|_{\Omega_V} \in W^{3,2}(\Omega_V)$ and moreover

$$\lim_{n\to\infty} \left\| u_0 - \frac{u_n}{\sqrt{\|u_n\|^2 + \|v_n\|^2}} \right\|_{W^{3,2}(\Omega_V)} = \lim_{n\to\infty} \left\| v_0 - \frac{v_n}{\sqrt{\|u_n\|^2 + \|v_n\|^2}} \right\|_{W^{3,2}(\Omega_V)} = 0, \tag{22}$$

the step-by-step procedure for a case of Laplacian is described in [3].

Since $u_0, v_0 \in W^{3,2}(\Omega_V)$, it is possible to use the Hopf Lemma together with negativeness of u_0, v_0 the get a result

$$\frac{\partial u_0}{\partial \vec{n}}(x) > 0 \text{ for a.a. } x \in \partial\Omega_U, \quad \frac{\partial v_0}{\partial \vec{n}}(x) = \frac{b_{12}}{d_1 \kappa_k - b_{11}} \frac{\partial u_0}{\partial \vec{n}} > 0 \text{ for a.a. } x \in \partial\Omega_U. \tag{23}$$

Now we define the function u_0 by

$$\tilde{u}_0(x) = \begin{cases} 0 & \text{if } x \in \Omega_U \\ u_0(x) & \text{if } x \in \Omega \backslash \Omega_U \end{cases}$$

and v_0 similarly. Substituting $\tilde{u}_0, \tilde{v}_0$ in (9) and using that $\tilde{u}_0(x) = 0$ for a.a. $x \in \partial\Omega_U$ leads to

$$\begin{aligned} &\int_\Omega d_1 \nabla\tilde{u}_0 \cdot \nabla(\varphi - \tilde{u}_0) - (b_{11}u_0 + b_{12}v_0)(\varphi - \tilde{u}_0) = \\ &= d_1 \int_{\Omega\backslash\Omega_U} -\Delta\tilde{u}_0 - (b_{11}u_0 - b_{12}v_0)(\varphi - \tilde{u}_0)\ \mathrm{dx} + \int_{\partial\Omega_U} \frac{\partial\tilde{u}_0}{\partial\vec{n}}(\varphi - \tilde{u}_0) = \int_{\partial\Omega_U} \frac{\partial\tilde{u}_0}{\partial\vec{n}}\varphi \geq 0, \end{aligned}$$

because $\varphi \geq 0$ a.e. in $\partial\Omega_U$. Similarly the second equation gives

$$\int_\Omega d_2 \nabla\tilde{v}_0 \cdot \nabla(\psi - \tilde{v}_0) - (b_{21}u_0 + b_{22}v_0)(\psi - \tilde{v}_0) =$$
$$= \int_{\Omega\backslash\Omega_U} -d_2\Delta\tilde{u}_0 - (b_{21}u_0 + b_{22}v_0)(\psi - \tilde{v}_0)\ \mathrm{dx} + \int_{\partial\Omega_U} \frac{\partial\tilde{v}_0}{\partial\vec{n}}(\psi - \tilde{v}_0) = \int_{\partial\Omega_U} \frac{\partial\tilde{v}_0}{\partial\vec{n}}\psi \geq 0.$$

Therefore u_0, v_0 are nontrivial solution of (9) and d_2^K is a critical point of this system. We construct the functions

$$\tilde{u}_n(x) = \begin{cases} 0 & \text{if } x \in \Omega_U \\ u_n(x) & \text{if } x \in \Omega\backslash\Omega_U \end{cases}$$
$$\tilde{v}_n(x) = \begin{cases} 0 & \text{if } x \in \Omega_U \\ v_n(x) & \text{if } x \in \Omega\backslash\Omega_U \end{cases}$$

Due to (22), (23) there exists n_0 such that for any $n > n_0$ the normal derivatives of u_n and v_n on $\partial\Omega_U$ satisfy

$$\frac{\partial u_n}{\partial\vec{n}}(x) > 0 \text{ for a.a. } x \in \partial\Omega_U, \qquad \frac{\partial v_n}{\partial\vec{n}}(x) > 0 \text{ for a.a. } x \in \partial\Omega_U.$$

Similar procedure as for linear case gives

$$d_1\int_\Omega \nabla\tilde{u}_n \cdot \nabla(\varphi - \tilde{u}_n) - (b_{11}u_n + b_{12}v_n - n_1(\tilde{u}_n, \tilde{v}_n))(\varphi - \tilde{u}_n)\ \mathrm{dx} =$$
$$= d_1\int_{\Omega\backslash\Omega_U} -\Delta\tilde{u}_n - (b_{11}u_n - b_{12}v_n - n_1(\tilde{u}_n, \tilde{v}_n)(\varphi - \tilde{u}_n)\ \mathrm{dx} +$$
$$+ \int_{\partial\Omega_U} \frac{\partial\tilde{u}_n}{\partial\vec{n}}(\varphi - \tilde{u}_n)\ \mathrm{dS} = \int_{\partial\Omega_U} \frac{\partial\tilde{u}_n}{\partial\vec{n}}\varphi\ \mathrm{dS} \geq 0,$$
$$d_{2,n}\int_\Omega \nabla\tilde{v}_n \cdot \nabla(\psi - \tilde{v}_n) - (b_{21}u_n + b_{22}v_n - n_2(\tilde{u}_n, \tilde{v}_n))(\psi - \tilde{v}_n)\ \mathrm{dx} =$$
$$= d_{2,n}\int_{\Omega\backslash\Omega_U} -\Delta\tilde{v}_n - (b_{21}u_n - b_{22}v_n - n_2(\tilde{u}_n, \tilde{v}_n)(\psi - \tilde{v}_n)\ \mathrm{dx} +$$
$$+ \int_{\partial\Omega_U} \frac{\partial\tilde{v}_n}{\partial\vec{n}}(\psi - \tilde{v}_n)\ \mathrm{dS} = \int_{\partial\Omega_U} \frac{\partial\tilde{v}_n}{\partial\vec{n}}\psi\ \mathrm{dS} \geq 0,$$

i.e. the functions $\tilde{u}_n, \tilde{v}_n$ are solutions of (8). Therefore d_2^K is a bifurcation point of (8).

The key to the proof of the last statement is in Proposition 3.1 in [2]. Let

$$d_2(\hat{\kappa}, d_1) := \frac{1}{\hat{\kappa}}\left(\frac{b_{12}b_{21}}{d_1\hat{\kappa} - b_{11}} + b_{22}\right).$$

For $d_1 \in (b_{11}/\hat{\kappa}_2, b_{11}/\hat{\kappa}_1)$ there is $d_2^K = d_2(\hat{\kappa}_1, d_1)$, as follows from the definition of d_2^K. If $\hat{\kappa}_i < \hat{\kappa}_j$ are different positive numbers, then there exists exactly one positive $d_1 < b_{11}\hat{\kappa}_i$ such that $d_2(\hat{\kappa}_i, d_1) = d_2(\hat{\kappa}_j, d_1)$. In simple terms, the hyperbolas intersects exactly at one point. The points of intersection satisfy

$$\hat{\kappa}_i\hat{\kappa}_j b_{22}d_1^2 - (\hat{\kappa}_i + \hat{\kappa}_j)d_1\frac{\det B}{b_{11}} + \frac{b_{11}\det B}{b_{22}} = 0.$$

Let $\kappa_2 < \hat{\kappa}_1 < \kappa_1$. The intersection points are

$$\begin{aligned}\hat{\kappa}_1\kappa_1 b_{22} d_{1,M}^2 - (\hat{\kappa}_1 + \kappa_1) d_{1,M} \frac{\det B}{b_{11}} &= -\frac{b_{11}\det B}{b_{22}},\\ \hat{\kappa}_1\kappa_2 b_{22} d_{1,m}^2 - (\hat{\kappa}_1 + \kappa_2) d_{1,m} \frac{\det B}{b_{11}} &= -\frac{b_{11}\det B}{b_{22}}.\end{aligned}$$

Dividing these equations gives

$$\frac{\hat{\kappa}_1\kappa_1 b_{22} d_{1,M}^2 - (\hat{\kappa}_1 + \kappa_1)\frac{\det B}{b_{11}}}{\hat{\kappa}_1\kappa_2 b_{22} d_{1,m}^2 - (\hat{\kappa}_1 + \kappa_2)\frac{\det B}{b_{11}}} = 1.$$

Since $\hat{\kappa}_2 < \kappa_1 < \hat{\kappa}_1$ this can be true only if $d_{1,m} < d_{1,M}$. Since $d_2(\kappa_2, d_1)$ is negative for all $d_1 \in (b_{11}/\kappa_2, b_{11}/\kappa_1)$, cf. Remark 2, and because $d_2(\hat{\kappa}, d_1) < d_2(\kappa_1, d_1)$ for any $d_1 \in (d_{1,m}, d_{1,M})$ it must be $(d_1, d_2^K) \in D_S$. □

3 Applications

The set of all positive critical points $(d_1, d_2) \in \mathbb{R}_+^2$ of the problem (1) with Dirichlet b.c. on $\partial\Omega$, i.e. of the problem *without* unilateral terms, is

$$C := \bigcup_{i=1}^{\infty} \left\{ (d_1, d_2) \in \left(0, \frac{b_{11}}{\kappa_1}\right) \times \mathbb{R}^+ \middle|\ d_2 := \frac{1}{\kappa_i}\left(\frac{b_{12}b_{21}}{d_1\kappa_i - b_{11}} + b_{22}\right)\right\}.$$

It is easy to verify that there are no positive critical points if $d_1 > b_{11}/\kappa_1$. The set of bifurcation points of (1), (2) is described by the exact formula (12), and it only suffices to (numerically) compute the eigenvalues $\hat{\kappa}_k$. To demonstrate the results Thomas model from [4] in the set $\Omega = [-1, 1]^2$ and $\Omega_U = B_{0.05}(0, 0)$ was chosen. In particular,

$$\begin{aligned}u_t &= d_1\Delta u + \gamma(a - u - h(u, v)),\\ v_t &= d_2\Delta u + \gamma(\alpha b - \alpha v - h(u, v)),\end{aligned}$$

with $a = 150$, $b = 100$, $\alpha = 1.5$, $\gamma = 252$, $K = 0.05$, $\rho = 13$ and with Dirichlet boundary condition and small random initial condition. This system has a stationary solution $(\bar{u}, \bar{v}) = (37.738, 25.1588)$. The system has to be shifted by $u \equiv u - \bar{u}$, $v \equiv v - \bar{v}$, in order to this stationary solution be equal to zero and (4) be true. The stationary system with unilateral sources is then as follows:

$$\begin{aligned}d_1\Delta u + 226.7u - 1124.5v + n_1(u, v) = 0 \ \text{ in } [-1, 1]^2,\\ d_2\Delta v + 478.7u - 1502.5v + n_2(u, v) = 0 \ \text{ in } [-1, 1]^2,\end{aligned} \tag{24}$$

$$\begin{aligned}&u \geq 0, \quad d_1\triangle u + 226.7u - 1124.5v + n_1(u, v) \leq 0 \ \text{ in } \ B_{0.05}(0, 0),\\ &u \cdot (d_1\triangle u + 226.7u - 1124.5v + n_1(u, v)) = 0 \ \text{ in } \ B_{0.05}(0, 0),\\ &v \geq 0, \quad d_2\triangle v + 478.7u - 1502.5v + n_2(u, v) \leq 0 \ \text{ in } \ B_{0.05}(0, 0),\\ &v \cdot (d_2\triangle v + 478.7u - 1502.5v + n_2(u, v)) = 0 \ \text{ in } \ B_{0.05}(0, 0),\\ &u = v = 0 \ \text{ on } \partial\Omega,\end{aligned} \tag{25}$$

where n_1, n_2 satisfy (4). The eigenvalues of the Laplace operator on Ω are known to be

$$\kappa_k = \frac{(k\pi)^2}{4}.$$

The first eigenvalue $\kappa_1 = \pi^2/4$, the second one is $\kappa_2 = \pi^2 = 9.9$. The first eigenvalue of the Laplacian (10) was numerically computed as $\hat{\kappa}_1 = 9.1 \pm 0.1$. The situation is sketched in the Fig. 2. The red curve represents the set of bifurcation points of the problem (24), (25) partially interfering into D_S, which is impossible for a system without sources, whose critical points generates the hyperbolas in this figure (cf. the Fig. 1). Although there is infinitely many hyperbolas, there are plotted only five of them in the Fig. 2.

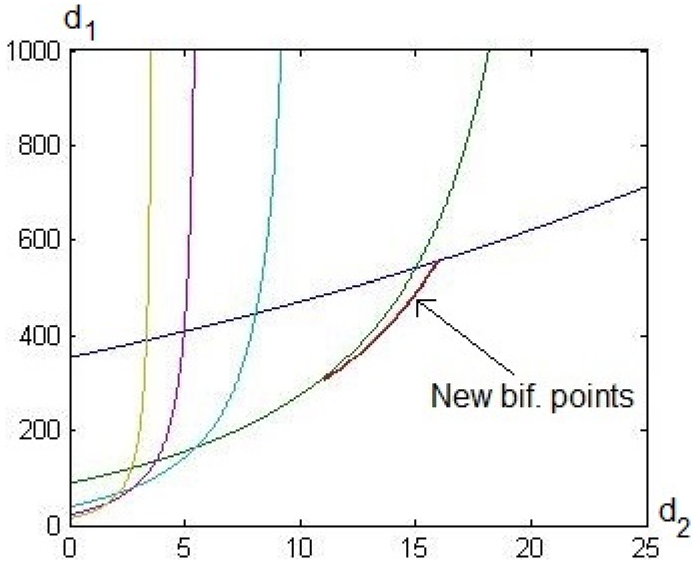


Figure 2: Hyperbolas for Thomas system

The further research will focus on numerical solution of this particular problem and other systems, and on a study of resulting patterns.

References

[1] J. Eisner M. Kučera, M. Väth, *Bifurcation points for a reaction-diffusion system with two inequalities.* J. Math. Anal. Appl. **365** (2010), 176–194.

[2] J. Eisner, M. Väth, *Degree, instability and bifurcation of reaction-diffusion systems with obstacles near certain hyperbolas.* Nonlinear Analysis: Theory, Methods & Applications **135** (2016), 158–193.

[3] J. Navrátil, *Bifurcation of the Laplace Equation with an Interior Unilateral Condition.* DDNY Workshop (2014), available online: `http://kmwww.fjfi.cvut.cz/ddny/historie/14-sbornik.pdf`

[4] V. Rybář, T. Vejchodský: *Irregularity of Turing patterns in the Thomas model with a unilateral term.* in Proceedings of the Programs and Algorithms of Numerical Mathematics 17 (2014), Institute of Mathematics AS CR, Prague, 2015, 188–193

On the Kramers-Fokker-Planck Equation with Decreasing Potentials in Dimension One*

Radek Novák

5th year of PGS, email: novakra9@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

David Krejčiřík, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Xue Ping Wang, Jean Leray Laboratory for Mathematics, University of Nantes

Abstract. In this work we study the Kramers-Fokker-Planck equation with a potential whose gradient tends polynomially fast to zero at the infinity. For this class of short-range potentials in one position variable, we show that complex eigenvalues do not accumulate at low-energies. The first threshold zero is always a resonance and the corresponding resonant state is uniquely determined. This allows us to obtain the low-energy resolvent asymptotics, which, in combination with more general high energy pseudospectral estimates, gives the large-time asymptotics of solutions to the KFP equation in appropriate spaces. These are expressed in terms of the equilibrium state, the Maxwellian.

Keywords: return to equilibrium, threshold spectral analysis, pseudo-spectral estimates, Kramers-Fokker-Planck equation.

Abstrakt. V tomto článku studujeme Kramers-Fokker-Planckovu rovnici s potenciálem, jehož gradient v nekonečnu klesá polynomiálně rychle k nule. Pro tuto třídu krátkodosahových potenciálů v jedné proměnné polohy ukazujeme, že komplexní vlastní hodnoty neakumulují poblíž nízkých energií. První prahová hodnota nula je vždy rezonancí a odpovídající rezonantní stav je jednoznačně určen. To nám umožňuje získat asymptotiky rezolventy pro nízké energie, jež, společně s více obecnými vysokoenergetickými pseudospektrálními odhady, nám dává ve vhodných prostorech aysmptotiky řešení KFP rovnice pro velké časy. Tyto jsou vyjádřeny pomocí rovnovážného stavu, Maxwelliánu.

Klíčová slova: návrat do rovnováhy, prahová spektrální analýza, pseudospektrální odhady, Kramers-Fokker-Planckova rovnice

*This work is based on the submission to an impacted journal

Plane-Parallel Waves as Duals of the Flat Background II: T-Duality with Spectators[*]

Filip Petrásek

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

advisor: Ladislav Hlavatý, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. We give the classification of T-duals of the flat background in four dimensions with respect to one-, two-, and three-dimensional subgroups of the Poincaré group using non-Abelian T-duality with spectators. As duals, we find backgrounds for sigma models in the form of plane-parallel waves or diagonalizable curved metrics often with torsion. Among others, we find exactly solvable time-dependent isotropic pp-wave, singular pp-waves, or generalized plane wave (K-model).

Keywords: sigma model, pp-wave background, string duality, non-Abelian T-duality, isometry group, spectator

Abstrakt. Předkládáme klasifikaci T-duálů plochého pozadí ve čtyřech rozměrech vzhledem k jednorozměrným, dvourozměrným a trojrozměrným podgrupám Poincarého grupy s využitím neabelovské T-duality s přihlížeči. Jako duály nalézáme pozadí pro sigma modely ve tvaru pp-vln nebo diagonalizovatelných křivých metrik často s torzí. Mimo jiné nalézáme exaktně řešitelnou časově závislou izotropní pp-vlnu, singulární pp-vlny nebo zobecněnou rovinnou vlnu (K-model).

Klíčová slova: sigma model, pp-vlna, strunová dualita, neabelovská T-dualita, grupa isometrií, přihlížeč

Full paper: F. Petrásek, L. Hlavatý, and I. Petr. *Plane-parallel waves as duals of the flat background II: T-duality with spectators.* Class. Quantum Grav. **34** (2017) 155003. arXiv:1612.08015 [hep-th].

[*] This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/239/OHK4/3T/14.

Adaptivity of Gaussian Process-Based Versions of the CMA-ES[*]

Zbyněk Pitra[†]

4th year of PGS, email: z.pitra@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Martin Holeňa, Department of Nonlinear Dynamics & Complex Systems
Institute of Computer Science, CAS

Abstract. An area of increasingly frequent applications of evolutionary optimization to real-world problems is continuous black-box optimization. However, evaluating real-world black-box fitness functions is sometimes very time-consuming or expensive, which interferes with the need of evolutionary algorithms for many fitness evaluations. Therefore, surrogate regression models replacing the original expensive fitness in some of the evaluated points have been in use since the early 2000s [3]. The Surrogate Covariance Matrix Adaptation Evolution Strategy (S-CMA-ES) [1] and its successor the Doubly Trained S-CMA-ES (DTS-CMA-ES) [4] represent two surrogate-assisted versions of the state-of-the-art algorithm for continuous black-box optimization CMA-ES [2]. In [5] and [9], we have investigated extensions of S- and DTS-CMA-ES that control the usage of the model according to the model's error. In [6] and [7], we have compared the ordinal and metric Gaussian process regression model using in combination with the DTS-CMA-ES. Moreover, we have presented an overview of several algorithms using surrogate models to speed up the original CMA-ES [8].

Keywords: benchmarking, black-box optimization, surrogate model, Gaussian process

Abstrakt. Oblastí se stále se zvyšujícím množstvím aplikací evoluční optimalizace na problémy z praxe je spojitá black-box optimalizace. Vyhodnocení takovéto skutečné black-box fitness funkce ale bývá velice časově nebo výpočetně náročné, což koliduje s faktem, že evoluční algoritmy vyžadují mnoho vyhodnocení fitness funkce. Proto se již téměř od roku 2000 využívají náhradní regresní modely namísto skutečné fitness funkce pro některé z vyhodnocovaných bodů [3]. Algoritmy Surrogate Covariance Matrix Adaptation Evolution Strategy (S-CMA-ES) [1] a jeho následník Doubly Trained S-CMA-ES (DTS-CMA-ES) [4] představují dvě varianty v současnosti nejlepšího algoritmu na spojitou black-box optimalizaci jménem CMA-ES [2], které používají náhradní modely. V článcích [5] a [9], jsme představili rozšíření S- a DTS-CMA-ESu, která řídí používání modelu v závislosti na jeho chybě. Porovnání ordinálních a metrických modelů založených na gaussovských procesech v kombinaci s DTS-CMA-ESem jsme provedli v [6] a [7]. Dále jsme také vypracovali porovnání několika algoritmů používajících náhradní modely k urychlení původního CMA-ESu [8].

Klíčová slova: benchmarking, black-box optimalizace, náhradní modelování, gaussovské procesy

[*] The reported research was supported by the Czech Science Foundation grant No. 17-01251, by the Grant Agency of the Czech Technical University in Prague with its grant No. SGS17/193/OHK4/3T/14, and by the project Nr. LO1611 with a financial support from the MEYS under the NPU I program.

[†] This study has been provided in cooperation with Lukáš Bajer and Jakub Repický.

References

[1] L. Bajer, Z. Pitra, and M. Holeňa. *Benchmarking Gaussian processes and random forests surrogate models on the BBOB noiseless testbed.* In 'Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation', GECCO Companion '15, 1143–1150, New York, NY, USA, (July 2015). ACM.

[2] N. Hansen. *The CMA Evolution Strategy: A Comparing Review.* In 'Towards a New Evolutionary Computation', J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, (eds.), number 192 in Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg (January 2006), 75–102.

[3] Y. S. Ong, P. B. Nair, and A. J. Keane. *Evolutionary optimization of computationally expensive problems via surrogate modeling.* AIAA Journal **41** (2003), 687–696.

[4] Z. Pitra, L. Bajer, and M. Holeňa. *Doubly trained evolution control for the Surrogate CMA-ES.* In 'Parallel Problem Solving from Nature – PPSN XIV: 14th International Conference, Edinburgh, UK, September 17-21, 2016, Proceedings', J. Handl, E. Hart, P. R. Lewis, M. López-Ibáñez, G. Ochoa, and B. Paechter, (eds.), 59–68, Cham, (2016). Springer International Publishing.

[5] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. *Adaptive doubly trained evolution control for the Covariance Matrix Adaptation Evolution Strategy.* In 'ITAT 2017 Proceedings', J. Hlaváčová, (ed.), volume 1885, 120–128. CEUR Workshop Proceedings, (September 2017).

[6] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. *Comparison of ordinal and metric Gaussian process regression as surrogate models for CMA evolution strategy.* In 'Proceedings of the Genetic and Evolutionary Computation Conference Companion', GECCO '17, 1764–1771, New York, NY, USA, (2017). ACM.

[7] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. *Ordinal versus metric Gaussian process regression in surrogate modelling for CMA evolution strategy.* In 'Proceedings of the Genetic and Evolutionary Computation Conference Companion', GECCO '17, 177–178, New York, NY, USA, (2017). ACM.

[8] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. *Overview of surrogate-model versions of Covariance Matrix Adaptation Evolution Strategy.* In 'Proceedings of the Genetic and Evolutionary Computation Conference Companion', GECCO '17, 1622–1629, New York, NY, USA, (2017). ACM.

[9] J. Repický, L. Bajer, Z. Pitra, and M. Holeňa. *Adaptive generation-based evolution control for Gaussian process surrogate models.* In 'ITAT 2017 Proceedings', J. Hlaváčová, (ed.), volume 1885, 136–143. CEUR Workshop Proceedings, (September 2017).

Monotonicity in Bayesian Networks for Computerized Adaptive Testing[*]

Martin Plajner

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

advisor: Jiří Vomlel, Department of Decision-Making Theory
Institute of Information Theory and Automation, CAS

Abstract. Artificial intelligence is present in many modern computer science applications. The question of effectively learning parameters of such models even with small data samples is still very active. It turns out that restricting conditional probabilities of a probabilistic model by monotonicity conditions might be useful in certain situations. Moreover, in some cases, the modeled reality requires these conditions to hold. In this article we focus on monotonicity conditions in Bayesian Network models. We present an algorithm for learning model parameters, which satisfy monotonicity conditions, based on gradient descent optimization. We test the proposed method on two data sets. One set is synthetic and the other is formed by real data collected for computerized adaptive testing. We compare obtained results with the isotonic regression EM method by Masegosa et al. which also learns BN model parameters satisfying monotonicity. A comparison is performed also with the standard unrestricted EM algorithm for BN learning. Obtained experimental results in our experiments clearly justify monotonicity restrictions. As a consequence of monotonicity requirements, resulting models better predict data.

Keywords: computerized adaptive testing, monotonicity, isotonic regression EM, gradient method, parameters learning

Abstrakt. V dnešní době se umělá inteligece využívá v mnoha oblastech lidské činnosti a to s pomocí rozličných modelů. Otázka možnosti efektivního učení takových modelů je proto stále velmi aktuální. Ukazuje se, že, v případě omezení modelu dodatečnými podmínkami monotonicity, je v určitých podmínkách přínosné. V mnoha aplikacích je dokonce nezbytné, aby byly tyto podmínky splněny, protože vychází z modelované reality. Tento článek se zaměřuje na podmínky monotonicity uplatněné v modelech bayesovských sítí. Představujeme algoritmus založený na gradientním sestupu k učení parametrů modelů splňujících podmínky monotonicity. Tyto algoritmy testujeme na dvou datových sadách. První sada je tvořena syntetickými daty, zatímco druhá se skládá z reálných dat sesbíraných pro tento účel. Získané výsledky porovnáváme s EM isotoní regresí vytvořeným autory Masegosa et al., který také učí model bayesovské sítě splňující podmínky monotonicity. Srovnání je též provedeno s neomezeným EM algoritmem pro učení bayesovských sítí. Získané výsledky z našich experimentů jasně potvrzují užitečnost podmínek monotonicity. Jako důsledek jejich vynucení při učení parametrů, výsledné model lépe předpovídají data.

[*] This work was supported by the Czech Science Foundation (project No. 16-12010S) and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/198/OHK4/3T/14.

Klíčová slova: počítačové adaptivní testování, monotonicita, EM isotoní regrese, gradientní metoda, učení parametrů

References

[1] Plajner, M., Vomlel, J. *Monotonicity in Bayesian Networks for Com- puterized Adaptive Testing*, Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, Cham, 2017. pp. 125-134. ISSN 0302-9743. ISBN 978-3-319-61580-6

[2] Masegosa, A. R., Feelders, A. J., and van der Gaag, *L. Learning from in- complete data in Bayesian networks with qualitative influences.* International Journal of Approximate Reasoning, (2016). 69:18–34.

Logaritmická Schrödingerova rovnice a její $\mathcal{PT}$ symetrické analogie

František Růžička

2. ročník PGS, email: fruzicka@gmail.com
Katedra fyziky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Miloslav Znojil, Oddělení teoretické fyziky
Ústav jaderné fyziky AV ČR, v. v. i.

Abstrakt. Nelineární Schrödingerovou rovnicí se v principu rozumí jakákoli z obecné třídy rovnic $-i\psi_t(x,t) = \Delta\psi(x,t) + F[\psi(x,t), \psi^*(x,t)]\psi(x,t)$, kde F je libovolný nekonstantní funkcionál. Pro různé volby F se následně objevují různé možnosti fenomenologického uplatnění této rovnice. V praxi se setkáváme zejména s případem kvadratické nelinearity $F[\psi(x,t)] = \psi^*(x,t)\psi(x,t)$ v teorii supravodivosti a při studiu Bose-Einsteinova kondenzátu. Z nepolynomiálních funkcionálů se lze (v podobných aplikacích) v literatuře nejčastěji setkat s logaritmickou nelinearitou $F[\psi(x,t)] = \ln[\psi^*(x,t)\psi(x,t)]$.

Nelineární Schrödingerova rovnice je pro libovolnou volbu F rovnicí lokální, což je také nutná podmínka většiny současných fyzikálních teorií. V nedávné době se ale objevilo několik možných aplikací tzv. $\mathcal{PT}$-symetrických Hamiltoniánů (jak v klasické, tak v kvantové mechanice), které mohou v některých případech vést na nelokální (efektivní) teorie. To bylo také popudem k nedávnému studiu modifikované NLSE s (nelokálním) funkcionálem $F = \psi^*(-x,t)\psi(x,t)$ (cit. no. 43, 44).

V tomto článku se zabýváme dalším logických krokem v této úvaze: srovnáním logaritmické NLSE a její nelokální analogie $F[\psi(x,t)] = \ln[\psi^*(-x,t)\psi(x,t)]$. Jelikož nelokální "hustota pravděpodobnosti" $\psi^*(-x,t)\psi(x,t)$ je obecně komplexní funkcí pro $x \in \mathbb{R}$, studujeme tuto rovnici (inspirováni cit. no. 42) na modifikovaném definičním oboru, který tvoří správně zvolený kontur v komplexní rovině. Nakonec diskutujeme několik explicitně zkonstruovaných referenčních řešení lokální i nelokální logaritmické NLSE, a to jak pro případ jednočásticové vlnové funkce, tak pro její vektorovou (vícečásticovou) formu.

Klíčová slova: nelineární Schrödingerova rovnice, logaritmická Schrödingerova rovnice, $\mathcal{PT}$-symetrie

Abstract. In its most general meaning, the nonlinear Schrödinger equation is understood to be any of the family of equations $-i\psi_t(x,t) = \Delta\psi(x,t) + F[\psi(x,t), \psi^*(x,t)]\psi(x,t)$, with F being an arbitrary nonconstant functional. For varying F we may encounter vastly different possibilities of phenomenogical appllications of the equation. The most often discussed case is probably the quadratic nonlinearity $F[\psi(x,t)] = \psi^*(x,t)\psi(x,t)$ relevant e.g. when studying superconductivity and Bose-Einstein condensates. Among non-polynomial functionals, one may encounter also the $F[\psi(x,t)] = \ln[\psi^*(x,t)\psi(x,t)]$.

The NLSE is a local equation for any choice of F, which is also a strict requirement of the vast majority of current physics theories. However, a number of possible applications of $\mathcal{PT}$-symmetric Hamiltonians (in both classical and quantum mechanics) emerged recently, which could sometimes lead to nonlocal (effective) theories. This was also the principal motivation for studying a modified NLSE with a nonlocal functional $F = \psi^*(-x,t)\psi(x,t)$ in cit. no. 43, 44.

In the present paper, we take another step in this direction and provide a comparison of the logarithmic NLSE and its nonlocal analogue $F[\psi(x,t)] = \ln[\psi^*(-x,t)\psi(x,t)]$. Since the nonlocal "probability density" $\psi^*(-x,t)\psi(x,t)$ is in general complex-valued for $x \in \mathbb{R}$, we study the equation (iinspired by cit. no. 42) on a modified domain, consisting of a carefully selected contour in the complex plane. We finally construct several reference solutions to these equations, both for the case of single-particle wavefunction and its many-body matrix counterpart.

Keywords: nonlinear Schrödinger equation, logarithmic Schrödinger equation, $\mathcal{PT}$-symmety

Plná verze: M. Znojil, F. Růžička and K. G. Zloshchastiev, *Schrödinger Equations with Logarithmic Self-Interactions: From Antilinear PT-Symmetry to the Nonlinear Coupling of Channels*, Symmetry **9** (2017), 165.

A Machine Learning Method for Incomplete and Imbalanced Medical Data*

Issam Salman

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

advisor: Jiří Vomlel Vomlel, Department of Decision-Making Theory
Institute of Information Theory and Automation, CAS

Abstract. Our research reported in this paper is twofold. In the first part of the paper we use standard statistical methods to analyze medical records of patients suffering myocardial infarction from the third world Syria and a developed country - the Czech Republic. One of our goals is to find whether there are statistically significant differences between the two countries. In the second part of the paper we present an idea how to deal with incomplete and imbalanced data for tree-augmented naive Bayesian (TAN). All results presented in this paper are based on a real data about 603 patients from a hospital in the Czech Republic and about 184 patients from two hospitals in Syria.

Keywords: Machine Learning, Data analysis, Bayesian networks, Missing data, Imbalanced data, Acute Myocardial Infarction.

Abstrakt. Náš výzkum, kterým se zabýváme v tomto článku, má dvě části. V první části používáme standardní statistické metody k analýze lékařských záznamů pacientů, kteří prodělali infarkt a pocházeli buď ze země třetího světa (Sýrie) nebo z rozvinuté země (Česká republika). Jedním z našich cílů je zjistit, zda mezi oběma zeměmi existují statisticky významné rozdíly. V druhé části článku předkládáme myšlenku zabývat se neúplnými a nevyrovnanými daty pro klasifikátor Tree-Augmented Naive Bayes (TAN). Všechny naše výsledky jsou prezentovány v tomto článku a vycházejí z reálných údajů o 603 pacientech z nemocnice v České republice a přibližně 184 pacientů ze dvou nemocnic v Sýrii.

Klíčová slova: strojové učení, analýza dat, bayesovské sítě, neúplná data, nevyrovnaná data, akutní infarkt myokardu

1 Introduction

Acute myocardial infarction (AMI) is commonly known as a heart attack. A heart attack occurs when an artery leading to the heart becomes completely blocked and the heart doesn't get enough blood or oxygen. Without oxygen, cells in that area of the heart die. AMI is responsible for more than a half of deaths in most countries worldwide. Its treatment has a significant socioeconomic impact.

One of the main objectives of our research is to design, analyze, and verify a predictive model of hospital mortality based on clinical data about patients. A model that predicts

*This work has been supported by the SGS grant CTU SGS16/253/OHK3/3T/14.

well the mortality can be used, for example, for the evaluation of the medical care in different hospitals. The evaluation based on mere mortality would not be fair to hospitals that treat often complicated cases. It seems better to measure the quality of the health care using the difference between predicted and observed mortality.

A related work was published by [1]. The authors analyze the mortality data in U.S. hospitals using the logistic regression model. Other work was published by [2]. The authors compare different machine learning methods using a real medical data from a hospital.

2 Data

Our dataset contains data about 787 patients characterized by 24 variables. 603 patients of them are from the Czech Republic [2] and 184 are from Syria. The attributes are listed in the Table 1. Most of the attributes are real valued, four attributes are nominal. Only a subset of attributes was measured for the Syrian patients. Most records contain missing values, i.e., for most patients only some attribute values are available. The thirty days mortality is recorded for all patients. In the Czech Republic the results of blood tests are reported in millimoles per liter of blood. In Syria some of the measurements are reported in milligrams per liter and some in millimoles per liter. We standartize all measurements to the millimoles per liter scale.

We will note $\mathbf{U} = \{X_1, X_2, \dots, X_m\}$ for a discrete domain, where $X_i, i \in \{1, 2, \dots, m\}$ is a discrete attribute and take on values from a finite set, denoted by $Val(X_i)$. We use capital letters such as X, Y, Z for attribute names, and lower-case letters such as x,y,z to denote specific values taken by those variables. Sets of variables are denoted by boldface capital letters such as $\mathbf{X}$,$\mathbf{Y}$,$\mathbf{Z}$ and assignments of values to the variables in these sets are denoted by boldface lowercase letters $\mathbf{x}$,$\mathbf{y}$,$\mathbf{z}$. A classified discrete domain is a discrete domain where one of the attributes is distinguished as "class". We will use $\mathbf{U}_C = \{A_1, A_2, \dots, A_n, C\}$ for a classified discrete domain. A dataset $D = \{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ of instances of $\mathbf{U}_C$, where each $\mathbf{u}_i, i \in \{1, \dots, N\}$ is a tuple of the form $(a_i^1, \dots, a_i^n, c_i)$ where $a_i^1 \in Val(A_1), \dots, a_i^n \in Val(A_n)$ and $c_i \in Val(C)$. Also we note that the class is always known, and a missing value in the dataset is denoted by NA.

3 Preliminary Statistical Analysis

For a preliminary statistical analysis [3] we selected a subset of attributes that are highly correlated with the class [5] and present in both groups, namely, we considered these variables: age, nationality, gender, STEMI location, and the class mortality. The STEMI location encoded by 1 denotes a STEMI.inf, 2 denotes a STEMI.ant, and 3 denotes a STEMI.lat. The nationality is encoded by a binary variable, where 0 means Czech and 1 means Syrian. The Gender is encoded by a binary variable where 0 denotes a man, while 1 stands for a female. The mortality is also encoded as a binary variable, where 0 means that the patient survived 30 days, while 1 means that he/she did not.

Already from Figure 1, where the histogram of the age values is presented, we can see that from patients that didn't survive a high percentage are young patients from Syria.

Table 1: Attributes

Attribute	**Code**	**type**	**value range in data**	**Country**
Age	AGE	real	[23, 94]	SYR, CZ
Height	HT	real	[145, 205]	CZ
Weight	WT	real	[35, 150]	CZ
Body Mass Index	BMI	real	[16.65, 48.98]	CZ
Gender	SEX	nominal	{male, female}	SYR, CZ
Nationality	NAT	nominal	{Czech, Syrian}	SYR, CZ
STEMI Location	STEMI	nominal	{inferior, anterior, lateral}	SYR, CZ
Hospital	Hospital	nominal	{CZ, SYR1, SYR2}	SYR, CZ
Kalium	K	real	[2.25, 7.07]	CZ
Urea	UR	real	[1.6, 61]	SYR, CZ
Kreatinin	KREA	real	[17, 525]	SYR, CZ
Uric acid	KM	real	[97, 935]	SYR, CZ
Albumin	ALB	real	[16, 60]	SYR, CZ
HDL Cholesterol	HDLC	real	[0.38, 2.92]	SYR, CZ
Cholesterol	CH	real	[1.8, 9.9]	SYR, CZ
Triacylglycerol	TAG	real	[0.31, 11.9]	SYR, CZ
LDL Cholesterol	LDLC	real	[0.261, 7.79]	SYR, CZ
Glucose	GLU	real	[2.77, 25.7]	SYR, CZ
C-reactive protein	CRP	real	[0.3, 359]	SYR, CZ
Cystatin C	CYSC	real	[0.2, 5.22]	SYR, CZ
N-terminal prohormone of brain natriuretic peptide	NTBNP	real	[22.2, 35000]	CZ
Troponin	TRPT	real	[0, 25]	CZ
Glomerular filtration rate (based on MDRD)	GFMD	real	[0.13, 7.31]	CZ
Glomerular filtration rate (based on Cystatin C)	GFCD	real	[0.09, 7.17]	CZ

The standard chi-square test of conditional independence between two variables reveals (see Table 2) that there is a significant dependence (at the level 0.05) between the mortality and nationality, the gender and nationality, also there are a significant dependencies between the gender and age, the mortality and gender – the patients from Syria have the lowest probability to survive, also they are younger and there is higher percentage of woman.

Finally, we learned the logistic regression model, that describes the relationship between the considered independent variables and the mortality as the dependent variable. We have got:

$$\begin{aligned}\text{logit } P(C=1|A=a) &= \beta_0+\beta_1 a_1+\ldots+\beta_4 a_4\\ &= -0.034+0.001\cdot a_1+0.027\cdot a_2-0.007\cdot a_3+0.065\cdot a_4\end{aligned}$$

where a_1: age, a_2: gender, a_3: STEMI loc, and a_4: nationality. Variables age and nationality appeared to be statistically significant for mortality prediction.

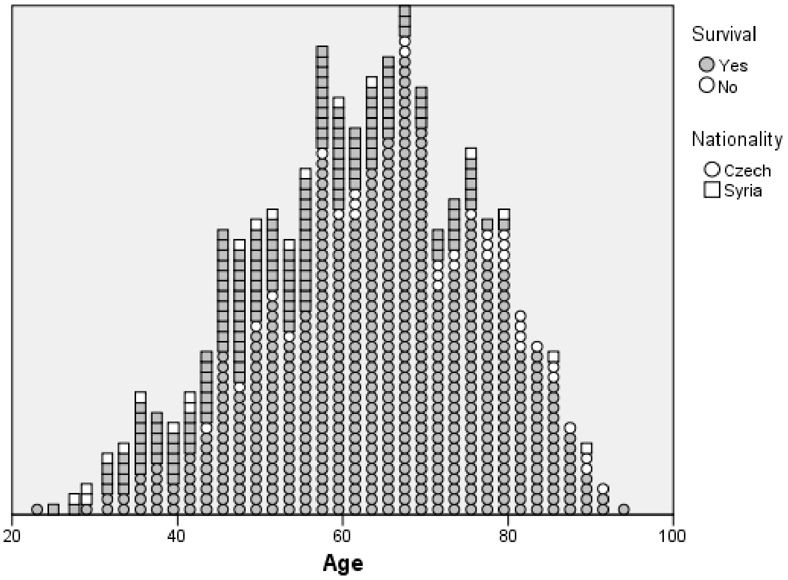


Figure 1: Histogram of the age values

From the preliminary statistical analysis we can conclude that:

- In Syria the mortality from AIM is significantly higher than in the Czech Republic – 87.3% Syrian patients survive, while 94.7% patients from the Czech Republic survive.
- The age of patients in Syria is lower in average (the average difference is 13 years) and there is a higher prevalence of women among the patients with AIM in Syria than in the Czech Republic.
- The STEMI location is related to the mortality.

Table 2: The Chi-Square Test of conditional independence

		gender	STEMI loc.	mortality	nationality
age	value	**.174**	-.010	.048	**-.381**
	sign.	**.0001**	.775	.181	**.0001**
gender	value		.022	.068	**.92**
	sign.		.53	.057	**.01**
STEMI loc.	value			-.026	-.036
	sign.			0.46	.312
mortality	value				**.089**
	sign.				**0.013**

4 Machine Learning Methods

The preliminary statistical analysis studied mostly the pairwise relations only. Since the explanatory variables may combine their influence and the influence of a variable may be mediated by another variable it is worth of studying the relations of variables alltogether. Our data are incomplete and imbalanced. We will present an idea for dealing with that type of data using tree-augmented naive Bayesian (TAN).

4.1 Bayesian networks

A Bayesian network [6] is an annotated directed acyclic graph that encodes a mass probability distribution over a set of random variables $\mathbf{U}$. Formally, a Bayesian network for $\mathbf{U}$ is a pair $B = \langle G, \Theta \rangle$. The first component, G, is a directed acyclic graph whose vertices correspond to the random variables $\mathbf{U} = \{X_1, X_2, \dots, X_m\}$, and whose edges represent direct dependencies between the variables. The graph G encodes independence assumptions: each variable X_i is independent of its non-descendants given its parents in G. The second component of the pair, namely Θ, represents the set of parameters that quantifies the network. It contains the parameter $\theta_{x_i|\Pi_{x_i}} = f(x_i|\Pi_{x_i})$ for each possible value x_i of X_i and Π_{x_i} of Π_{X_i}, where Π_{X_i} denotes the set of parents of X_i in G. Accordingly, a Bayesian network B defines a unique joint probability distribution over $\mathbf{U}$ given by:

$$f(X_1 = x_1, \dots, X_m = x_m) \quad = \quad \prod_{i=1}^{m} f(X_i = x_i | \Pi_{X_i} = \Pi_{x_i}) \quad = \quad \prod_{i=1}^{m} \theta_{x_i|\Pi_{x_i}}$$

for each Π_{X_i} which is a parent of X_i.

4.2 Learning with Trees

A directed acyclic graph on $\{X_1, X_2, \dots, X_n\}$ is a tree if Π_{X_i} contains exactly one parent for all X_i, except for one variable that has no parents (this variable is referred to as the root). A tree network can be described by identifying the parent of each variable [7]. A function $\pi : \{1, \dots, n\} \to \{0, \dots, n\}$ is said to define a tree over $X_1, X_2, \dots, X_n$ if there is exactly one i such that $\pi(i) = 0$ (namely the root of the tree), and there is no sequence $i_1, \dots, i_k$ such that $\pi(i_j) = i_{j+1}$ for $i \leq j < k$ and $\pi(i_k) = i_1$ (i.e., no cycles). Such a function defines a tree network where $\Pi_{X_i} = \{X_{\pi(i)}\}$ if $\pi(i) > 0$ and $\Pi X_i = \emptyset$ if $\pi(i) = 0$.

4.3 Learning Maximum Likelihood TAN

Let $\{A_1, A_2, \dots, A_n\}$ be a set of attribute variables and C be the class variable. We say that B (Bayesian network) is a TAN model if $\Pi_C = \emptyset$ and there is a function π that defines a tree over $\{A_1, A_2, \dots, A_n\}$. The optimization problem consists on finding a tree defining function π over $\{A_1, A_2, \dots, A_n\}$ such that the log likelihood is maximized [8] $LL(B_T|D) = \sum_{\mathbf{u} \in D} \log f(\mathbf{u})$. To learn the maximum likelihood TAN we should use the following equation to compute the parameters [8], $\theta_{a_i, \Pi_{a_i}} = \frac{N_{a_i, \Pi_{a_i}}(a_i, \Pi_{a_i})}{N_{\Pi_{a_i}}(\Pi_{a_i})}$ where $N_{a_i, \Pi_{a_i}}(a_i, \Pi_{a_i})$ stands for the number of times that attribute i has value a_i and its parents have values Π_{a_i} in the dataset. Similarly, $N_{\Pi_{a_i}}(\Pi_{a_i})$ is the number of times that the parents of attribute A_i have values Π_{a_i} in the dataset.

5 Learning TAN from incomplete data

Missing data are a very common problem which is important to consider in a many data mining applications, and machine learning or pattern recognition applications. Some variables may not be observable (i.e. hidden) even for training instances. Now more and more datasets are available, and most of them are incomplete. Therefore, we want to find a way to build a new model from an incomplete dataset. Normally, to learn the maximum likelihood TAN structure [8], we need a complete data, such that all instances $\mathbf{u}_i, i \in \{1, \ldots, N\}$ from $\mathbf{U}_C$ are complete and don't have any missing value. In case the data are incomplete and there is an instance which has a missing value, we will not use the whole instance in TAN structure learning i.e. not use the other known values from that instance in TAN structure learning. Note that the class is always known, and a missing value in the dataset is denoted by NA. Our goal is to learn a tree-augmented naive Bayesian (TAN) from incomplete data. Some previous work by [13] propose maximizing conditional likelihood for BN parameter learning. They apply their method to MCAR (Missing Completely At Random) incomplete data by using available case analysis in order to find the best TAN classifier. In other work by [9] also deals with TAN classifiers and expectation-maximization (EM) principle for partially unlabeled data. In their work, only the variable corresponding to the class can have missing. Also, other work by [10] deals with TAN based on the EM principle, where they have proposed an adaptation of the learning process of Tree Augmented Naive Bayes classifier from incomplete data. In their work, any variable can have missing values in the dataset. The TAN algorithm can be adapted to learn from incomplete datasets, such that most available data will be used in TAN structure learning. The procedure is shown in Algorithm 1, where the Conditional Mutual Information "CMI" is defined as:

$$I(X, Y|Z) = \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} f(\mathbf{x},\mathbf{y},\mathbf{z}) \log \frac{f(\mathbf{z}) f(\mathbf{x},\mathbf{y},\mathbf{z})}{f(\mathbf{x},\mathbf{z}) f(\mathbf{y},\mathbf{z})}$$

where the sum is only over $\mathbf{x},\mathbf{y},\mathbf{z}$ such that $f(\mathbf{x},\mathbf{z}) > 0$ and $f(\mathbf{y},\mathbf{z}) > 0$. In Algorithm 1, on line 25 we build a complete undirected graph in which the vertices are the attributes $A_1, \ldots, A_n$. Annotate the weight of an edge connecting A_i to $A_j, i \neq j$ by $I_{p_ij} = I(A_i, A_j|C)$ One line 26 we build a subgraph from G, without any cycles and with the maximum possible total edge weight. On line 27 we transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it. On line 28 we add the class C to the graph as a node and add edges from C to all other nodes in the graph

The idea behind Algorithm 1 is that we believe if we use more data then the estimates of conditional mutual information are more reliable.

6 Imbalanced Data

In case of imbalanced data the classifiers are more sensitive to detecting the majority class and less sensitive to the minority class. Thus, if we don't take care of the issue, the classification output will be biased, in many cases resulting in always predicting the majority class. Many methods have been proposed in the past few years to deal with imbalanced data. In our research the mortality rate of patients with myocardial infarction refers to the percentage of patients who have not survived more than 30 days, where the results are 89% of patients survive and 11% of patients do not survive, therefore the data are quite imbalanced. One of the most common and simplest strategies to handle imbalanced data is to under-sample the majority class [11, 12]. While different techniques have been proposed in the past, they did not bring any improvement

Algorithm 1 TAN For Incomplete Data

```
procedure CMI($A_i, A_j, C\}$)                ▷ // Conditional Mutual Information
    $\overline{D} = \{\overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_N\}, \overline{\mathbf{u}}_m = (a_i, a_j, c), m \in \{1, \ldots, N\}$, such that $\mathbf{u}_m = (a_1, \ldots, a_n, c) \in D$
    Foreach $\overline{\mathbf{u}}_m \in \overline{D}$
        If($a_i == NA|a_j == NA$)
            Delete $\overline{\mathbf{u}}_m$ from $\overline{D}$
    endfor
    Compute $I_p = I(A_i, A_j|C)$ from $\overline{D}$
    return $I_p$
Endprocedure
Read $D = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}, \mathbf{u}_m = (a_1, \ldots, a_n, c), m \in \{1, \ldots, N\}$
var:
$n$ the number of attribute variables $A$;
$\mathbb{I}_p[n][n]$ the WeightMatrix;
$UG$ the UndirectedGraph;
$UT$ the UndirectedTree;
$T$ the DirectedTree;
TAN the DirectedGraph;
Foreach $A_i, i \in \{1, \ldots, n-1\}$
    Foreach $A_j, j \in \{2, \ldots, n\}$
        $I_{p_ij} = CMI(A_i, A_j, C)$
        $\mathbb{I}_p[i][j] = I_{p_ij}$
        $\mathbb{I}_p[j][i] = I_{p_ij}$
    EndForeach
EndForeach
$G$ = ConstructUndirectedGraph($\mathbb{I}_p[i][j]$)
$UT$ = MaximumWeightedSpanningTree($G$);
$T$ = MakeDirected($UT$);
TAN = AddClass($T$);
```

with respect to simply selecting samples at random. So, for this analysis we propose the following steps:

- Let M be the number of samples for the majority class, and N be the number of samples for the minority class, and M be L times greater than N.

- Divide the instances which have majority class into L distinct clusters.

- Train L predictors, where each predictor is trained on only one of the distinct clusters, but on all of the data from the rare class. To be clear, the data from the minority class are used in the training of all L predictors.

- Use model averaging for the L learned predictors as your final predictor. i.e (in our case we will compute a conditional mutual information between each pair of attributes $(A_i, A_j), i, j \in 1, 2, \ldots, n, i \neq j$ given the class L times for each pair, in each time will use only one of the distinct clusters and all data from the minority class, then we will use the average of conditional mutual information for each pair to compute a weight matrix).

After integrating this step into the Algorithm 1, we will have a TAN algorithm which deals with an incomplete and imbalance data 2:

Algorithm 2 TAN for incomplete and imbalance data

```
var
    $M$ The number of samples for the majority class
    $N$ The number of samples for the minority class
    $D_T$ All instances of the majority class, $D_T \subset D$
    $D_F$ All instances of the minority class, $D_F \subset D$
integer division $L = M/N$
Divide $D_T$ to $L$ parts, $D_{T_k}, k \in \{1, \ldots, L\}$
**Foreach** $D_{T_k}$
    $D_k = D_{T_k} \cup D_F$
**EndForeach**
Compute WeightMatrix $\mathbb{I}_{p_k}[n][n]$ foreach $D_k$
$\hat{\mathbb{I}}_p[n][n]$ = the average of $\mathbb{I}_{p_k}[n][n], k \in 1, \ldots, L$    ▷ // $\hat{\mathbb{I}}_p$ is the WeightMatrix which
    wwill be used in Algorithm 1
Continue from line 26 in Algorithm 1 using $\hat{\mathbb{I}}_p$
```

7 Results

For each data record classified by a classifier there are four possible classification results. Either the classifier got a positive example labeled as positive (in our data the positive example is the patient survived) or it made a mistake and marked it as negative. Conversely, a negative example may have been mislabeled as a positive one, or correctly marked as negative. Our results are summarized in Figure 2 using the ROC curves. We use the 10 fold cross validation as the model evaluation method. The ROC curve shows how the classifier can sacrifice the true positive rate (TP rate: number of positive examples, labeled as such over total positives) for the false positive rate(FP rate: number of negative examples, labeled as positive over total negatives) (1-specificity) by plotting the TP rate to the FP rate. In other words, it shows how many correct positive classifications can be gained as you allow for more and more false positives by changing the threshold.

In Figure 2 we compare our results with normal TAN ([8]) and SMOTE algorithm ([4]) for TAN. Algorithm 2 has achieved the highest area under the ROC curve (AUC) with 0.82. The results of Algorithm 1 (ROC = 0.77) is better than the normal TAN algorithm (ROC = 0.62). But SMOTE algorithm with TAN (ROC = 0.802) is better than Algorithm 1.

8 Conclusions

First, we used medical data on patients with AIM for preliminary statistical analysis. We found a significant difference between Syrian patients and Czech patients. Second, Bayesian networks are a tool of choice for reasoning in uncertainty, with incomplete data. However, often, Bayesian network structural learning only deals with complete data. We have proposed here an adaptation of the learning process of the Tree Augmented Naive Bayes classifier from incomplete and imbalanced datasets. This methods have been successfully tested on our dataset. We have seen that our Algorithm 2 performed better than normal TAN and TAN-SOMTE.

References

[1] H. M. Krumholz, S.-L. T. Normand, D. H. Galusha, J. A. Mattera, A. S. Rich, Y. Wang and Y. Wang, *Risk-Adjustment Models for AMI and HF 30-Day Mortality, Methodology*, Harvard Medical School, Department of Health Care Policy, (2007).

[2] J. Vomlel and H. Kružík and P. Tůma and J. Přeček, and M. Hutyra, *Machine Learning Methods for Mortality Prediction in Patients with ST Elevation Myocardial Infarction*, In the Proceedings of The Nineth Workshop on Uncertainty Processing WUPES'12, Czech Republic, 204–213, (2012).

[3] L. Wasserman. *All of Statistics*, Springer-Verlag New York, (2004).

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, Volume 11, Issue 16, 321–357, (2002).

[5] M. Hall and E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and H. Witten, *The WEKA Data Mining Software: an Update*, In 'ACM SIGKDD Exploration ACM SIGKDD Explorations', Volume 11, Issue 1. (2009), 10–18.

[6] F. V. Jensen, *An Introduction to Bayesian Networks*, Springer, (1996).

[7] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, (1988).

[8] N. Friedman, D. Geiger, and M. Goldszmidt, *Bayesian network classifiers*, Machine Learning Journal, Volume 29, Issue 2. (1997). 131–163.

[9] I. Cohen and F. Cozman and N. Sebe and M. C. Cirelo and T. S. Huang, *Semi-supervised learning of classifiers: theory, algorithms and their application to human-computer interaction*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 26, Issue 12, 1553–1568, (2004).

[10] O. C. H. Francois and P. Leray, *Learning the Tree Augmented Naive Bayes Classifier from incomplete datasets*,

Third European Workshop on Probabilistic Graphical Models, 91–98, (2006).

[11] R. Laza, R. Pavon, M. Reboiro-Jato and F. Fdez-Riverola R. Laza and et al, Evaluating the effect of unbalanced data in biomedical document classification, Journal of Integrative Bioinformatics, Volume 16, Issue 3, pp. 177, (2011).

[12] M. M. Rahman and D. N. Davis, Addressing the Class Imbalance Problem in Medical Datasets, International Journal of Machine Learning and Computing, volume 3, Issue 2, 224-228,(2013)

[13] R. Greiner and W. Zhou, Structural extension to logistic regression, Eighteenth Annual National Conference on Artificial Intelligence (AAI02), 167–173, (2002).

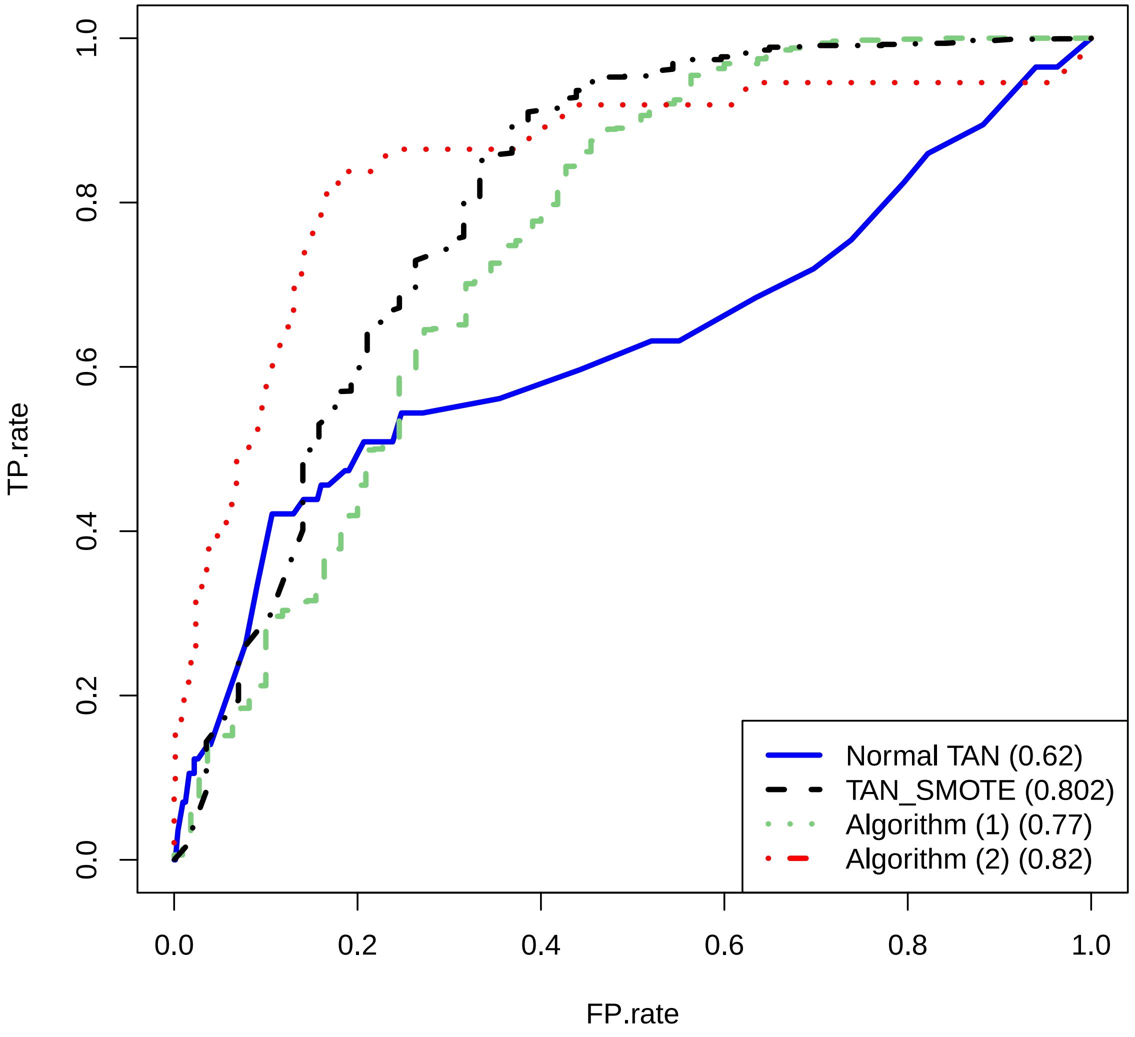


Figure 2: ROCs (TAN , TAN_SMOTI , Algorithm(1) , Algorithm(2))

The Problem of Coexistence of Several Non-Hermitian Observables in $\mathcal{PT}$-Symmetric Quantum Mechanics[*]

Iveta Semorádová

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

advisor: Miloslav Znojil, Department of Theoretical Physics
Nuclear Physics Institute, CAS

Abstract. During the recent developments of quantum theory it has been clarified that the observable quantities (like energy or position) may be represented by operators Λ (with real spectra) which are manifestly non-Hermitian in a preselected "friendly" Hilbert space $\mathcal{H}^{(F)}$. The consistency of these models is known to require an upgrade of the inner product, i.e., mathematically speaking, a transition $\mathcal{H}^{(F)} \to \mathcal{H}^{(S)}$ to another, "standard" Hilbert space. We prove that whenever we are given more than one candidate for an observable (i.e., say, two operators Λ_0 and Λ_1) in advance, such an upgrade *need not* exist in general.

Keywords: non-Hermitian operator, two observables, $\mathcal{PT}$-symmetry, metric operator

Abstrakt. Během nedávného rozvoje kvantové teorie bylo vyjasněno, že pozorovatelné veličiny (jako energie či poloha) mohou být reprezentovány operátory Λ (s reálným spektrem), které jsou zřejmě nehermitovské v předvybraném „přátelském" Hilbertově prostoru $\mathcal{H}^{(F)}$. Konzistence takovýchto modelů vyžaduje změnu skalárního součinu, to jest, matematicky řečeno, přechod $\mathcal{H}^{(F)} \to \mathcal{H}^{(S)}$ do jiného, „standardního" Hilbertova prostoru. Ukazujeme, že kdykoliv máme více než jednoho kandidáta na pozorovatelnou (to jest například dva operátory Λ_0 a Λ_1), takovýto přechod nemusí obecně vůbec existovat.

Klíčová slova: nehermitovský operátor, dvě pozorovatelné, $\mathcal{PT}$-symetrie, metrický operátor

Full paper: M. Znojil, I. Semorádová, F. Růžička, H. Moulla, and I. Leghrib. *The problem of coexistence of several non-hermitian observablesin PT-symmetric quantum mechanics.* Phys. Rev. A **95** (2017), 042122.

[*] This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/239/OHK4/3T/14

Numerical Method for Two Phase Flow Problems in Porous Media*

Jakub Solovský

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

advisor: Radek Fučík, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This work deals with testing of a numerical method for solving two phase flow problems in porous media. We briefly describe the numerical method, it's implementation, and benchmark problems. First, the method is verified using test problem in homogeneous porous media in 2D and 3D. Results show that the method is convergent and the experimental order of convergence is slightly less than one. On the problem in heterogeneous porous media, the method produces oscillations at the interface between different porous media and we demonstrate that these oscillations are not caused by the coarseness of the grid. To overcome the oscillations, we use the mass lumping technique which eliminates the oscillations at the interface. Tests on the problems in homogeneous porous media show that although the mass lumping technique slightly decreases the accuracy of the method, the experimental order of convergence remains the same.

Keywords: two phase flow, heterogeneity, mixed hybrid finite element method, mass lumping, porous media, upwind

Abstrakt. Článek se věnuje testování numerické metody pro řešení úloh dvoufázového proudění v porézním prostředí. Na začátku je stručně popsána numerická metoda, její implementace a testovací úlohy. Metoda je nejprve testována na úloze v homogenním prostředí ve 2D i 3D. Ukazuje se, že numerické schéma je konvergentní s experimentálním řádem konvergence o něco menším než jedna. Při řešení úlohy v heterogenním prostředí se na rozhraní mezi různými prostředími objevují oscilace, u kterých ukážeme, že nejsou způsobeny použitou sítí. Pro odstranění oscilací použijeme techniku mass lumping, která oscilace na rozhraní výrazně omezí. Na testech v homogenním prostředí se pak ukazuje, že ačkoli použití mass lumpingu nepatrně zhorší přesnost numerické metody, experimentální řád konvergence zůstává stejný.

Klíčová slova: dvoufázové proudění, heterogenity, hybridní metoda smíšených konečných prvků, mass lumping, porézní prostředí, upwind

1 Introduction

Mathematical modeling of two phase flow in porous media can be used in many applications. For instance prediction of contaminant transport can be used for protection of

*The work was supported by the Czech Science Foundation project no. 17-06759S: Investigation of shallow subsurface flow with phase transitions and by grant No. SGS17/194/OHK4/3T/14 of the Grant Agency of the Czech Technical University in Prague.

water resources or for sanitation of dangerous substances leakage. Except for special cases, there is no known way how to solve these problems exactly but with numerical methods, we can find at least a good approximation of the solution.

This paper focuses on the verification of the proposed numerical method. The method is implemented in parallel using `MPI` [12, 13]. Firstly, we test the method on two phase flow problems in homogeneous porous media in 2D and 3D. We further proceed with a problem in heterogeneous porous media which shows limitations of the method. Therefore, we propose a modification using mass lumping technique which helps to solve problems in heterogeneous porous media correctly. Finally, we compare both approaches on problems with known exact solution.

2 Numerical method

Here, we briefly describe the numerical method. A detailed description of the method together with a different approach to parallelism, using CUDA, is described in [7]. The method can be used for solving a system of n partial differential equations in the following coefficient form:

$$\sum_{j=1}^{n} N_{i,j} \frac{\partial Z_j}{\partial t} + \nabla \cdot \left[m_i \left(- \sum_{j=1}^{n} \boldsymbol{D}_{i,j} \nabla Z_j + \boldsymbol{w}_i \right) \right] = f_i, \tag{1}$$

where $Z_j = Z_j(\boldsymbol{x}, t)$, $j = 1, \ldots, n$, are unknown functions $(\forall t > 0,\ \forall \boldsymbol{x} \in \Omega)$, $\Omega \subset \mathbb{R}^d$ is the computational domain, and d is the spatial dimension, $d \in \{1, 2, 3\}$. $N_{i,j}$, f_i, and m_i are scalar coefficients, $\boldsymbol{w}_i$ are vector coefficients and $\boldsymbol{D}_{i,j}$ are symmetric, second order tensors. The coefficients can be functions of time t and spatial coordinates $\boldsymbol{x}$, but also of the unknown functions Z_j.

The method was implemented in `C++` and for the parallel implementation, `MPI` was used. Serial implementation of the method is described in detail in [7], parallel implementation in 2D, using MPI, is described in [13]. The parallelism in 3D which is used in this paper is a direct extension of the 2D case.

Triangular and tetrahedral meshes used in this paper were generated by `Gmsh` [8].

2.1 Coefficients in general formulation

All benchmark problems presented here are represented by the following choice of coefficients in the general formulation of the method given by Eq. (1):

$$\boldsymbol{N} = \begin{pmatrix} -\Phi \rho_w \frac{dS_w}{dp_c} & 0 \\ -\Phi \rho_n \frac{dS_w}{dp_c} & \Phi S_n \frac{d\rho_n}{dp_n} \end{pmatrix}, \qquad \boldsymbol{m} = \begin{pmatrix} \rho_w \frac{\lambda_w}{\lambda_t} \\ \rho_n \frac{\lambda_n}{\lambda_t} \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} \lambda_t \boldsymbol{K} & -\lambda_t \boldsymbol{K} \\ 0 & \lambda_t \boldsymbol{K} \end{pmatrix}, \qquad \boldsymbol{w} = \begin{pmatrix} -\lambda_t \rho_w \boldsymbol{K} \boldsymbol{g} \\ \lambda_t \rho_n \boldsymbol{K} \boldsymbol{g} \end{pmatrix}, \qquad \boldsymbol{f} = \begin{pmatrix} -f_w \\ f_n \end{pmatrix},$$

where:

Φ	$[-]$	is the porosity,
S_α	$[-]$	is the α-phase saturation,
ρ_α	$[\mathrm{kg}\cdot\mathrm{m}^{-3}]$	is the α-phase density,
f_α	$[\mathrm{kg}\cdot\mathrm{m}^{-3}\cdot\mathrm{s}^{-1}]$	are the sinks/sources,
$\boldsymbol{g}$	$[\mathrm{m}\cdot\mathrm{s}^{-2}]$	is the gravity vector,
$\boldsymbol{K}$	$[\mathrm{m}^2]$	is the permeability tensor,
$k_{r\alpha}$	$[-]$	is relative permeability (Burdine [2] or Mualem [11] model),
μ_α	$[\mathrm{kg}\cdot\mathrm{m}^{-1}\cdot\mathrm{s}^{-1}]$	is dynamic viscosity of the phase α,
$\lambda_\alpha = \frac{k_{r\alpha}}{\mu_\alpha}$	$[\mathrm{kg}^{-1}\cdot\mathrm{m}\cdot\mathrm{s}]$	is the α-phase mobility ($\lambda_t = \lambda_w + \lambda_n$),
p_α	$[\mathrm{Pa}]$	is the α-phase pressure,
$\alpha \in \{w, n\}$		denotes the wetting or non–wetting phase.

These coefficients represent mass conservation law and Darcy's law for both phases, refer to [6] for details .

3 Homogeneous porous media

In this section, we verify the numerical method on benchmark problems in 2D and 3D in homogeneous porous media. For these problems, the exact solution can be found and, therefore, we can compute the errors of the numerical solution and experimental order of convergence.

3.1 Benchmark problems

The benchmark problem used in this section is the extension of the McWhorter and Sunada problem into an arbitrary dimension. We only briefly describe the configuration of the problem, a more detailed description together with the method to find the exact solution can be found in [5, 10]. We assume a radially symmetric domain with the prescribed initial saturation S_i and the inflow at the origin in the form:

$$Q_0(t) = At^{\frac{d-2}{2}}. \tag{2}$$

The problem configuration in 2D is illustrated in Fig. 1. This setting together with the neglected gravity and the assumption of incompressible phases allow us to find the exact semi-analytical solution of the problem [5, 10].

The problem is defined in the whole $\mathbb{R}^2$ or $\mathbb{R}^3$ but due to the assumed radial symmetry, we restrict ourselves only to one quadrant in 2D or one octant in 3D, respectively. We also have to restrict ourselves to a domain of finite length and compare the results at a certain time when the head of the solution does not reach the boundary representing infinity.

In this paper, the computational domains are a square with 1 m long side and a cube with 1 m long edge in 2D and 3D, respectively. In both cases, we compare the solutions at time $t = 20\,000$ s.

The exact solution requires prescribing a flux at the origin (point-wise). Numerical method used in this paper cannot handle to prescribe a flux in one point, therefore, we

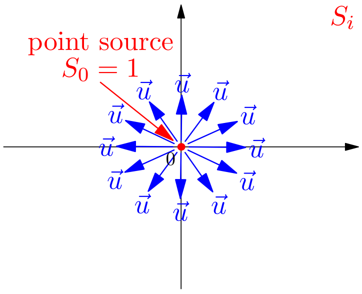
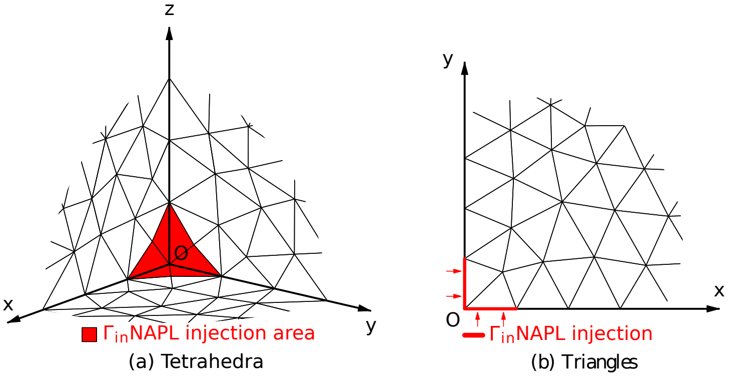


Figure 1: Benchmark problem configuration in 2D.

approximate the point inflow condition via a boundary condition by prescribing the flux through all element boundaries (edges, faces) that are adjacent to the origin as illustrated in Fig. 2. The corresponding value of the Neumann boundary condition is computed so that the total volume injected through the boundary is the same as the volume given by Eq. (2).

We set coefficients $A = 10^{-5}\,\mathrm{m}^2 \cdot \mathrm{s}^{-1}$ for the 2D case and $A = 10^{-7}\,\mathrm{m}^3 \cdot \mathrm{s}^{-\frac{3}{2}}$ for the 3D case. Initial saturation in the domain is $S_i = 0.95$ for both cases.


z
x
y
O
🟥 Γ_{in}NAPL injection area
x
y
O
Γ_{in}NAPL injection

(a) Tetrahedra

(b) Triangles

Figure 2: Approximation of the point injection flux at the origin in 2D and 3D.

3.2 Numerical analysis

In this paper, Brooks–Corey [1] and van Genuchten [14] models for capillary pressure together with Burdine [2] and Mualem [11] models for relative permeability, respectively, are used.

Numerical solutions in 2D (contours) and 3D (isosurfaces) together with the comparison with the exact solution in radial coordinates are shown in Fig. 3.

With the known exact solution, we can compute errors of the numerical solution and the experimental order of convergence. Results for 2D and 3D are shown in Table 2 and 3, respectively. Properties of the used meshes are given in Table 1, the following notation is used:

Mesh ID	h	Elements	Degrees of freedom
$2\mathrm{D}_1^{\triangle}$	$6.71 \cdot 10^{-2}$	242	766
$2\mathrm{D}_2^{\triangle}$	$3.49 \cdot 10^{-2}$	944	2 912
$2\mathrm{D}_3^{\triangle}$	$1.64 \cdot 10^{-2}$	3 714	11 302
$2\mathrm{D}_4^{\triangle}$	$8.73 \cdot 10^{-3}$	14 788	44 684
$2\mathrm{D}_5^{\triangle}$	$4.23 \cdot 10^{-3}$	59 336	178 648
$3\mathrm{D}_1^{\triangle}$	$2.13 \cdot 10^{-1}$	1 312	5 874
$3\mathrm{D}_2^{\triangle}$	$1.27 \cdot 10^{-1}$	3 697	15 546
$3\mathrm{D}_3^{\triangle}$	$6.29 \cdot 10^{-2}$	29 673	121 678
$3\mathrm{D}_4^{\triangle}$	$3.48 \cdot 10^{-2}$	240 372	973 750
$3\mathrm{D}_5^{\triangle}$	$1.84 \cdot 10^{-2}$	1 939 413	7 807 218

Table 1: Properties of the meshes used in the benchmarks described in Section 3.1.

	Brooks & Corey				van Genuchten			
Id.	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$
$2\mathrm{D}_1^{\triangle}$	$1{,}45 \cdot 10^{-2}$	**0,92**	$3{,}17 \cdot 10^{-2}$	**0,78**	$1{,}42 \cdot 10^{-2}$	**0,98**	$2{,}12 \cdot 10^{-2}$	**0,94**
$2\mathrm{D}_2^{\triangle}$	$7{,}94 \cdot 10^{-3}$	**0,78**	$1{,}91 \cdot 10^{-2}$	**0,60**	$7{,}51 \cdot 10^{-3}$	**0,86**	$1{,}15 \cdot 10^{-2}$	**0,84**
$2\mathrm{D}_3^{\triangle}$	$4{,}40 \cdot 10^{-3}$	**0,95**	$1{,}21 \cdot 10^{-2}$	**0,69**	$3{,}93 \cdot 10^{-3}$	**1,05**	$6{,}11 \cdot 10^{-3}$	**1,03**
$2\mathrm{D}_4^{\triangle}$	$2{,}41 \cdot 10^{-3}$	**0,85**	$7{,}84 \cdot 10^{-3}$	**0,66**	$2{,}03 \cdot 10^{-3}$	**0,90**	$3{,}19 \cdot 10^{-3}$	**0,89**
$2\mathrm{D}_5^{\triangle}$	$1{,}30 \cdot 10^{-3}$		$4{,}85 \cdot 10^{-3}$		$1{,}06 \cdot 10^{-3}$		$1{,}68 \cdot 10^{-3}$	

Table 2: Errors of the numerical solution and experimental orders of convergence in 2D for the benchmark problem described in Section 3.1.

h mesh element size. To compute h, we circumscribe a circle (ball) to each triangle (tetrahedron) of the mesh and take h as the radius of the largest such circle (ball).

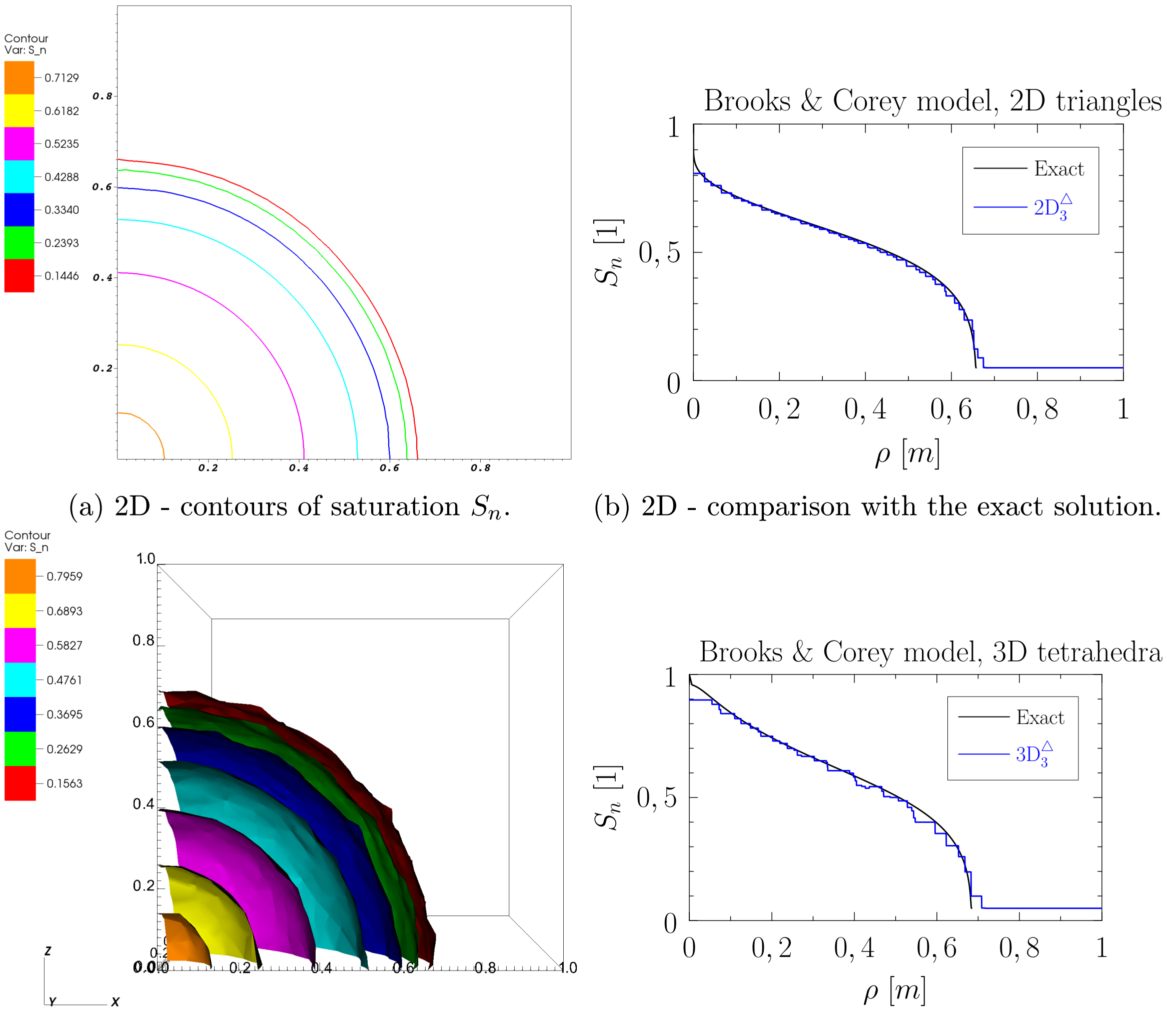
$\|E_{h,S_n}\|_p$ is the L_p norm of the difference between the exact and numerical solution of the saturation S_n on mesh with element size h.

$eoc_{S_n,p}$ is the experimental order of convergence in L_p norm, see [7] for details.

Different results for the Brooks–Corey and van Genuchten models are caused by different capillary pressure - saturation relationships for the near-water-saturated state.

	Brooks & Corey				van Genuchten			
Id.	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$
$3D_1^\triangle$	$1{,}12 \cdot 10^{-2}$	**0,69**	$3{,}38 \cdot 10^{-2}$	**0,60**	$1{,}21 \cdot 10^{-2}$	**0,77**	$2{,}43 \cdot 10^{-2}$	**0,73**
$3D_2^\triangle$	$7{,}82 \cdot 10^{-3}$	**0,84**	$2{,}47 \cdot 10^{-2}$	**0,72**	$8{,}13 \cdot 10^{-3}$	**0,93**	$1{,}66 \cdot 10^{-2}$	**0,90**
$3D_3^\triangle$	$4{,}35 \cdot 10^{-3}$	**1,03**	$1{,}49 \cdot 10^{-2}$	**0,92**	$4{,}25 \cdot 10^{-3}$	**1,14**	$8{,}84 \cdot 10^{-3}$	**1,12**
$3D_4^\triangle$	$2{,}37 \cdot 10^{-3}$	**0,82**	$8{,}63 \cdot 10^{-3}$	**0,79**	$2{,}17 \cdot 10^{-3}$	**1,04**	$4{,}56 \cdot 10^{-3}$	**1,02**
$3D_5^\triangle$	$1{,}41 \cdot 10^{-3}$		$5{,}23 \cdot 10^{-3}$		$1{,}12 \cdot 10^{-3}$		$2{,}39 \cdot 10^{-3}$	

Table 3: Errors of the numerical solution and experimental orders of convergence in 3D for the benchmark problem described in Section 3.1.



(a) 2D - contours of saturation S_n.

(b) 2D - comparison with the exact solution.

(c) 3D - isosurfaces of saturation S_n.

(d) 3D - comparison with the exact solution.

Figure 3: Numerical results and comparison with the exact solution. In radial coordinates, ρ denotes the distance from the origin (injection point).

4 Heterogeneous porous media

In this section, we focus on problems in heterogeneous porous media. As was shown in [12], the numerical method cannot correctly capture the effects at the interface between two different porous media. Oscillations appear in the solution and are more apparent in the case of flow from finer to coarser sand.

To demonstrate the oscillations in this work, we use the same benchmark problem as in [12] which was originally proposed in [9]. The problem setup is shown in Fig. 4. We consider three layers of sand, the middle one finer than the remaining two, initially fully saturated with water. NAPL is injected through the upper boundary with a given flux.

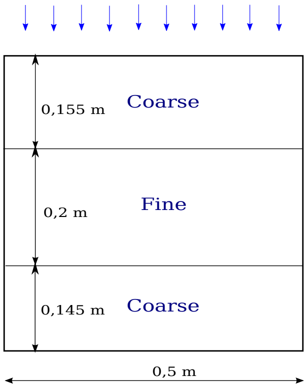


Figure 4: Heterogeneous problem setup based on [9, 12].

We use the numerical solution obtained using the vertex centered finite volume method in 1D on a very fine mesh as a reference solution to which we compare our numerical results. The 1D solution taken from [4] is in a good match with results provided in [9]. We want to compare our 2D results with this 1D solution. We do not use only the values over single crossection through the center of the domain, but we plot superposed values from all the elements of the mesh using their y position of the center.

Numerical results for the original variant of the method are shown in Figs. 5a, 5c, and 5e. We can see the oscillations that are present for several mesh refinements and, therefore, are not caused by the coarseness of the mesh.

4.1 Mass Lumping

To overcome the oscillations at the material interface we use the mass lumping technique. One of the steps of the MHFEM method used in this paper is to discretize numerical fluxes between elements. This is done by computing matrices $\boldsymbol{B}_{i,j,K}$, with elements defined by the following integral [12]:

$$B_{i,j,K,E,F} = \int_K \boldsymbol{\omega}_{K,F}^T \boldsymbol{D}_{i,j}^{-1} \boldsymbol{\omega}_{K,E}, \tag{3}$$

where K is the element, $\boldsymbol{\omega}_{K,F}$ and $\boldsymbol{\omega}_{K,E}$ are the basis functions of the lowest order Raviart-Thomas-Nédelec space. Element K is a simplex (line segment, triangle or tetrahedron depending on the dimension of the problem) and integrated functions are polynomials

	Brooks & Corey				van Genuchten			
Id.	$\Vert E_{h,S_n}\Vert_1$	$eoc_{S_n,1}$	$\Vert E_{h,S_n}\Vert_2$	$eoc_{S_n,2}$	$\Vert E_{h,S_n}\Vert_1$	$eoc_{S_n,1}$	$\Vert E_{h,S_n}\Vert_2$	$eoc_{S_n,2}$
$2D_1^{\triangle}$	$1{,}48 \cdot 10^{-2}$		$3{,}22 \cdot 10^{-2}$		$1{,}44 \cdot 10^{-2}$		$2{,}16 \cdot 10^{-2}$	
		0,91		**0,76**		**0,98**		**0,95**
$2D_2^{\triangle}$	$8{,}17 \cdot 10^{-3}$		$1{,}96 \cdot 10^{-2}$		$7{,}59 \cdot 10^{-3}$		$1{,}17 \cdot 10^{-2}$	
		0,77		**0,59**		**0,86**		**0,85**
$2D_3^{\triangle}$	$4{,}56 \cdot 10^{-3}$		$1{,}25 \cdot 10^{-2}$		$3{,}95 \cdot 10^{-3}$		$6{,}15 \cdot 10^{-3}$	
		0,96		**0,69**		**1,04**		**1,04**
$2D_4^{\triangle}$	$2{,}49 \cdot 10^{-3}$		$8{,}10 \cdot 10^{-3}$		$2{,}04 \cdot 10^{-3}$		$3{,}20 \cdot 10^{-3}$	
		0,86		**0,68**		**0,90**		**0,89**
$2D_5^{\triangle}$	$1{,}33 \cdot 10^{-3}$		$4{,}96 \cdot 10^{-3}$		$1{,}06 \cdot 10^{-3}$		$1{,}68 \cdot 10^{-3}$	

Table 4: Errors of the numerical solution and experimental orders of convergence in 2D for the mass lumping variant of the method.

of the second order and, therefore, the integral in Eq. (3) can be computed exactly (in the following using notation exact integration). The value of this integral can be also approximated using a quadrature rule [3]. We use the following quadrature rule to approximate the integral of arbitrary function over simplex K.

$$\int_K f \approx |K| \frac{1}{k} \sum_{i=1}^{k} f(\boldsymbol{x}_i), \tag{4}$$

where k is the number of vertices of the simplex (line segment $k = 2$, triangle $k = 3$, tetrahedron $k = 4$) and $\boldsymbol{x}_i$ are the positions of the vertices. In our case, the function f is the integrated function on the right hand side of Eq. (3).

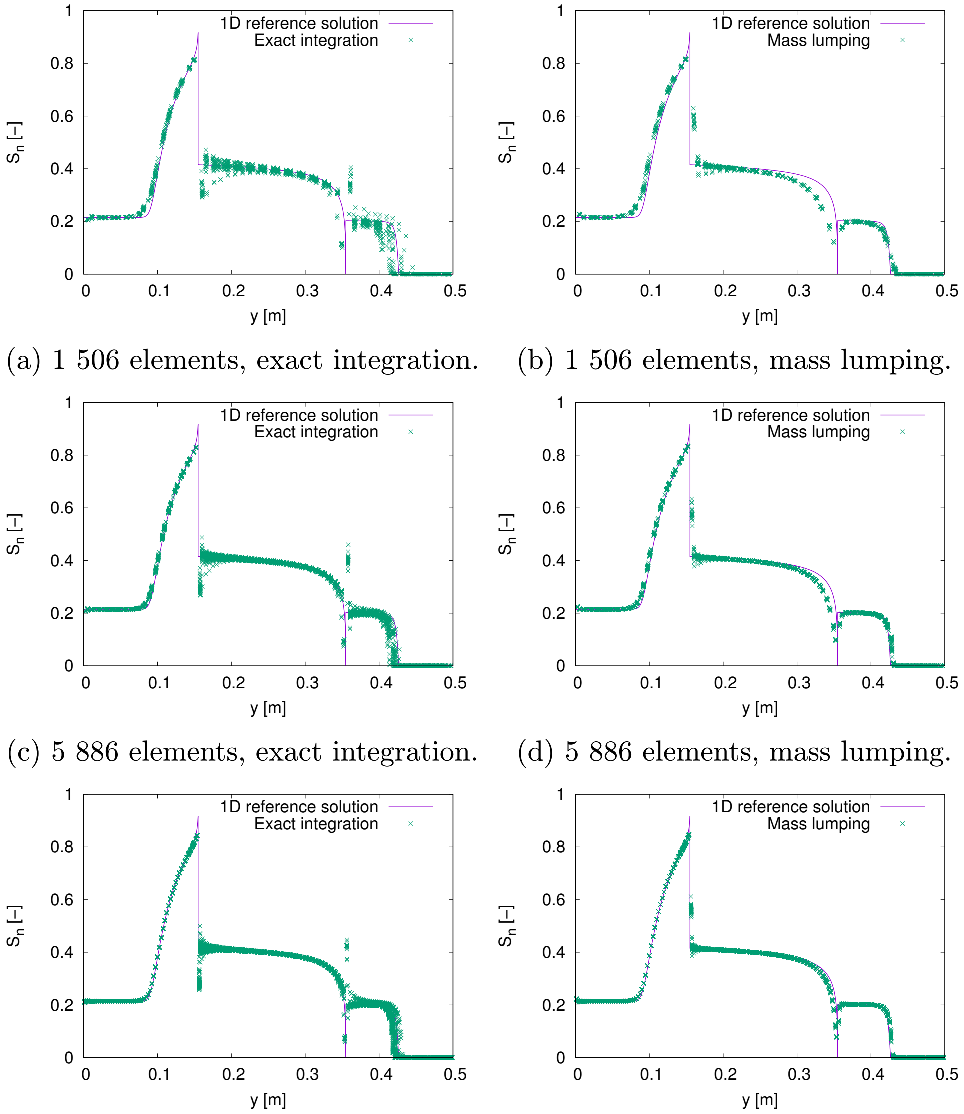
Numerical solutions using mass lumping technique are shown in Figs. 5b, 5d, and 5f. In the comparison with the basic variant of the method using exact integration, it can be seen that the use of the mass lumping technique eliminates the oscillations at the material interface.

5 Mass Lumping in homogeneous porous media

In the previous section, we showed that use of mass lumping eliminates the oscillations at the material interface. In this section, we show how the mass lumping technique affects the accuracy of the method in the case of homogeneous porous media where we can compare the results with exact solutions. We use the benchmark problem described in Section 3.1, solve it with the mass lumping variant of the method, and compare the results with those given in Section 3.2.

Errors of the solution and experimental orders of convergence in the 2D and 3D cases are shown in Table 4 and 5, respectively.

Results show that in both 2D and 3D cases, the errors of the mass lumping variant of the method are slightly worse than without mass lumping but the method is still convergent with the same experimental order of convergence.



(a) 1 506 elements, exact integration. (b) 1 506 elements, mass lumping.

(c) 5 886 elements, exact integration. (d) 5 886 elements, mass lumping.

(e) 23 308 elements, exact integration. (f) 23 308 elements, mass lumping.

Figure 5: Comparison between the exact integration and the mass lumping technique on various meshes for the solution in a heterogeneous porous medium.

	Brooks & Corey				van Genuchten			
Id.	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$	$\|E_{h,S_n}\|_1$	$eoc_{S_n,1}$	$\|E_{h,S_n}\|_2$	$eoc_{S_n,2}$
$3D_1^{\triangle}$	$1{,}13 \cdot 10^{-2}$	**0,67**	$3{,}46 \cdot 10^{-2}$	**0,61**	$1{,}22 \cdot 10^{-2}$	**0,77**	$2{,}49 \cdot 10^{-2}$	**0,74**
$3D_2^{\triangle}$	$7{,}96 \cdot 10^{-3}$	**0,82**	$2{,}52 \cdot 10^{-2}$	**0,72**	$8{,}22 \cdot 10^{-3}$	**0,93**	$1{,}70 \cdot 10^{-2}$	**0,91**
$3D_3^{\triangle}$	$4{,}50 \cdot 10^{-3}$	**1,01**	$1{,}53 \cdot 10^{-2}$	**0,92**	$4{,}30 \cdot 10^{-3}$	**1,13**	$8{,}97 \cdot 10^{-3}$	**1,12**
$3D_4^{\triangle}$	$2{,}47 \cdot 10^{-3}$	**0,83**	$8{,}64 \cdot 10^{-3}$	**0,79**	$2{,}20 \cdot 10^{-3}$	**1,04**	$4{,}63 \cdot 10^{-3}$	**1,02**
$3D_5^{\triangle}$	$1{,}44 \cdot 10^{-3}$		$5{,}26 \cdot 10^{-3}$		$1{,}15 \cdot 10^{-3}$		$2{,}41 \cdot 10^{-3}$	

Table 5: Errors of the numerical solution and experimental orders of convergence in 3D for the mass lumping variant of the method.

6 Conclusion

In this work, we tested the numerical method for solving two phase flow problems in porous media. We showed that for homogeneous porous media, the method is convergent for both 2D and 3D cases with the experimental order of convergence slightly less than one. In the case of heterogeneous porous media, the method produces oscillations at the interface between different porous media when exact evaluation of the integrals in matrix $\boldsymbol{B}$ is used. To overcome the difficulties, we used the mass lumping technique which eliminates the oscillations and only very slightly affects the accuracy of the method as was shown in the comparison of the solutions using the benchmark problems in 2D and 3D with known exact solutions.

References

[1] R. Brooks and A. Corey. *Hydraulic Properties of Porous Media.* Colorado State University, Hydrology Paper **3** (1964), 27.

[2] N. Burdine. *Relative Permeability Calculations From Pore Size Distribution Data.* Journal of Petroleum Technology **5** (1953), 71–78.

[3] G. Chavent and J. Roberts. *A unified physical presentation of mixed, mixed-hybrid finite elements and usual finite differences for the determination of velocities in waterflow problems* . [Research Report] RR-1107, INRIA (1989).

[4] R. Fučík. *Advanced Numerical Methods for Modelling Two-Phase Flow in Heterogeneous Porous Media.* PhD thesis, FNSPE of Czech Technical University Prague, (2010).

[5] R. Fučík, T. H. Illangasekare, and M. Beneš. *Multidimensional self-similar analytical solutions of two-phase flow in porous media.* Advances in Water Resources **90** (2016), 51–56.

[6] R. Fučík and J. Mikyška. *Mixed-hybrid finite element method for modelling two-phase flow in porous media.* Journal of Math for Industry **3** (2011), 9–19.

[7] R. Fučík, J. Mikyška, J. Solovský, J. Klinkovský, and T. Oberhuber. *Multidimensional Mixed–Hybrid Finite Element Method for Compositional Two–Phase Flow in Heterogeneous Porous Media and its Massively Parallel Implementation on GPU.* In review in Computer Physics Communications (2017).

[8] C. Geuzaine and J. F. Remacle. *Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities.* International Journal for Numerical Methods in Engineering **79** (2009), 1309–1331.

[9] R. Helmig. *Multiphase Flow and Transport Processes in the Subsurface, A contribution to the Modelling of Hydrosystems.* Springer, (1997).

[10] D. B. McWhorter and D. K. Sunada. *Exact integral solutions for two-phase flow.* Watter Resources Research **26** (1990), 399–413.

[11] Y. Mualem. *A new model for predicting the hydraulic conductivity of unsaturated porous media.* Water Resources Research **12** (1976), 513–522.

[12] J. Solovský. Matematické modelování dvoufázového vícesložkového proudění v porézním prostředí v problematice ochrany životního prostředí, (2016). Diplomová práce, ČVUT v Praze.

[13] J. Solovský and R. Fučík. *A parallel mixed–hybrid finite element method for two phase flow problems in porous media using MPI.* Accepted to Computer Methods in Materials Science (2017).

[14] M. van Genuchten. *A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils.* Soil Science Society of America Journal **40** (1980), 892–898.

Transmutation of Statistics in Financial Time Series Data*

Václav Svoboda

2nd year of PGS, email: svobova5@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. The concept of continuous time fractional Levy processes and its discrete time counterpart ARFIMA model are introduced. This class contains wide range of processes exhibiting so called fractional behaviour. Methods for computationally simple estimation of key ARFIMA parameters are presented.
The theory of fractional Levy processes is then applied to financial time series data. Key ARFIMA parameters are estimated on rolling window basis for S&P 500 daily data and transmutations of statistics are detected in the original data based on time evolution of these parameters. This transmutation reminds phase transitions in statistical physics.

Keywords: Fractional Levy processes, ARFIMA, Transformation of statistics, Finacial time series

Abstrakt. Představíme frakční Levyho procesy a jejich diskrétní verzi ARFIMA model. Tato třída obsahuje širokou škálu procesů vyznačující se takzvaným frakčním chováním. Efektivní metody pro odhad ARFIMA parametrů jsou představeny.
Tato teorii je poté aplikována na finanční data. ARFIMA parametry jsou odhadnuty na posouvající se podmnožině uvažovaných dat a transmutace statistika je detekována na základě jejich časového vývoje.

1 Introduction

Fractional processes have been successfully applied to number of problems in physics, biology or economy [1,12]. They are closely related to anomalous (non-Brownian) diffusion and they lead to non- standard scaling relations between temporal and positional coordinates, i.e. standard Brownian scaling

$$\langle x^2(t)\rangle \sim \sigma t \tag{1}$$

is no longer valid. For self similar processes this can be caused by two mechanisms - by correlations between increments of the process or by infinite variance of underlying process [2]. We will see that fractional Levy processes can compass both of these mechanism in unified framework.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/239/OHK4/3T/14 and by Czech Science Foundation Grant No. 17-33812L.

There is a number of ways which may give a rise to fractional dynamics. Trapping or long range memory effects may for example lead to this behaviour. However probably the most illustrative way to derive fractional processes is from continuous time random walk. It is well known result that continuous time random walk with finite average waiting time between jumps and finite jump size variance leads to Brownian motion in the limit. However if one of these assumptions fails to be satisfied the resulting process exhibits fractional behaviour. The infinite average waiting time leads to processes with various memory effects, fractional Brownian motion being the most prominent example, while infinite jump variance leads to stable processes. These are the two already mention mechanism leading to fractional behaviour.

Fractional processes can be well described using fractional differential equations [3]. Changing the order of temporal and spatial derivative in Fokker-Planck equation to non-integer order leads to desired distortion of standard Brownian scaling. However there is a number of different non-equivalent definitions of fractional differentiation and unified framework for fractional processes defined as solutions of fractional differential equations is still missing.

That is why we will not pursue the approach based on fractional differential equations in this paper. Instead we will start from definition of so called Levy fractional processes [4]. This class of processes directly combines fractional behaviour observed for fractional Brownian motion with behaviour of heavy tailed stable processes. This means that fractional behaviour of these processes is caused simultaneously by both correlations between increments and infinite variance of underlying process. While these processes cover incredibly wide range of processes and they likely provide general enough framework to model any type of fractional behaviour their analytical tractability is a major issue. Even simulation of such processes is a complicated issue with no satisfying solution am aware of.

However in the limit fractional Levy processes can be written as normalized sums of so called ARFIMA processes [5]. ARFIMA is discrete time stochastic model which directly generalizes well known ARMA linear model. Broadly speaking ARFIMA model describes (in the limit) behaviour of increments of Levy fractional processes.

Fitting ARFIMA model is complicated but tractable process [6]. However for our purposes we will only need to fit two main parameters of ARFIMA process - parameters effecting deformation of scaling of temporal and positional coordinate. Discontinuities and local extremes in time evolution of these parameters may be regarded as points of transmutation of underlying statistics. Furthermore in analogy to truncated Levy flights another so called damping coefficient is introduced which essentially cuts off extreme values produced by stable noise which lead to unrealisticly high fourth moments which are not observed in financial time series.

The class of fractional Levy processes contains essentially all self-similar processes with stationary increments. Financial data have typically fractal nature [11] i.e. are self-similar and the assumption of stationary increments is also in most of the cases reasonable. That is why we believe that fractional Levy processes provide appropriate and sufficient framework for financial time series modelling.

The paper is organized as follows - in the first part theoretical background behind fractional Levy processes is presented. Basic properties of these processes are discussed and

in particular two cases are mentioned - fractional Brownian motion and Levy stable processes. Then discrete time counterpart of Levy fractional processes - ARFIMA model - is introduced. The basic properties of this model are presented and connection with fractional Levy processes is established.
The second part of this paper will focus mostly on numerical methods used to estimate ARFIMA parameters and their application to real data. Computationally tractable method is introduced which allows for effective estimation of these parameters. The method is applied to daily data observed on financial markets during last sixty years. Analysis of evolution of these parameters on rolling window data is then used for detection of transmutation of statistics.

2 Fractional Levy processes and ARFIMA model

General Levy fractional process can be define in analogical way as fractional Brownian motion as an integral [7]

$$L_H^\alpha(t) = \int_{\mathbb{R}} \left((t-x)_+^d - (-x)_+^d\right) dL_\alpha(x) \tag{2}$$

where L_α is α-stable symmetric process, $d = H - 1/\alpha$ with $H \in (0,1)$ and $0 < \alpha \le 2$ and $(x)_+ = \max(x,0)$. In what follows we will always also consider $\alpha > 1$ because Levy processes with $\alpha < 1$ have number of undesirable properties.
Parameter α is called stability index and H is famous Hurst self similarity index. Hurst index is connected with fractal dimension of graph of the process which is equal to $2 - H$ [8].
Fractional Levy processes are H-self similar processes with stationary increments. They can be described via their characteristic function [4]

$$\varphi_t^{H,\alpha}(z) = e^{-(ct^H|z|)^\alpha} \tag{3}$$

The density of fractional Levy processes is not available in closed form in the general case.
The alternative to describe very similar class of processes exhibiting this type of fractional behaviour is through fractional Fokker-Planck equation. One of possible forms of this equation is [1]

$$\frac{\partial W}{\partial t} = {}_0D_t^{1-\gamma} K_\alpha \frac{\partial^2}{\partial x^2} W(x,t) \tag{4}$$

where ${}_0D_t^{1-\gamma}$ is Riemann-Liouville derivative. Riemann-Liouville operator is integral operator and therefore this equation is non-local and resulting process exhibits non-trivial memory effects. Parameter γ effects scaling - it holds $\langle x^2(t) \rangle \sim t^\gamma$ which means that case $\gamma > 1$ corresponds to super-diffusion and $\gamma < 1$ to sub-diffusion.
There are two special of fractional Levy processes we should mention.

Levy stable processes: The case $d = 0$ i.e. $H = 1/\alpha$ leads obviously to Levy stable processes [2]. Levy stable processes are H-self similar processes with stationary and independent increments. The general form of characteristic function of Levy stable process is

$$\ln \varphi_t(k) = it\gamma k - \sigma t|k|^\alpha (1 + i\beta \frac{|k|}{k}\omega(k,\alpha)) \tag{5}$$

where

$$\omega(k,\alpha) = \begin{cases} -tan(\pi\alpha/2) & \text{for } \alpha \neq 1 \\ (2/\pi)ln|k| & \text{for } \alpha = 1 \end{cases}$$

However in this paper we will focus only on symmetric case i.e. $\beta = 0$.
The stability index α determines the tail behaviour of density of stable process (if $\alpha < 2$)

$$p_\alpha(x) \sim \frac{1}{|x|^{\alpha+1}} \quad |x| \to \infty \tag{6}$$

The closed form density for Levy processes is available only in several cases - most important being case $\alpha = 2$ which leads to Brownian motion. All Levy processes other than Brownian motion have infinite variance.
The theoretical importance of stable distributions follows from generalized central limit theorem - stable distributions are attractors for normalized sums of iid variables with infinite variance [10].
Stable processes with $\alpha < 2$ have qualitatively different behaviour than Brownian motion, one of important differences is the fact that fractional dimension of a trail of a stable process is equal to $\max(\alpha, 1)$ [8]. This means that Brownian motion can fill two dimensional space while any other stable process cannot. This behaviour is due to the fact that heavy tailed stable processes move by very small jumps with occasional large jump - this means that they form clusters instead of filling the whole space.

Fractional Brownian motion: The case $\alpha = 2$ leads to integration with respect to Brownian motion which yields fractional Brownian motion [8].
Fractional BM is H-self similar Gaussian process with stationary but not with independent increments. The increments of fractional BM are positively correlated in the case $H > 1/2$ and negatively for $H < 1/2$. This means that the case $H > 1/2$ leads to superdiffusion and long range dependence of increments, if $H < 1/2$ increments are negatively correlated and process is sub-diffusive. The case $H = 1/2$ is just Brownian motion.

2.1 ARFIMA model

Autoregressive fractionally integrated moving average model (ARFIMA) [5,13] generalizes the standard linear ARMA model in two ways, naturally these two generalizations represent the two mechanisms leading to fractional behaviour. The general form of ARFIMA model is

$$\mathcal{A}_p(B)X_t = \mathcal{B}_q(B)(1-B)^{-d}Z_t \tag{7}$$

where B is a lag operator, $\mathcal{A}, \mathcal{B}$ are polynomials of order p respectively q and Z_t are iid α-stable variables representing random noise.
The term $(1-B)^{-d}$ is defined via Taylor expansion as

$$(1-B)^{-d}Z_t = \sum_{i=0}^{\infty} \frac{\Gamma(i+d)}{\Gamma(i)\Gamma(d+1)} Z_{t-i} \tag{8}$$

We denote the above defined model as $ARFIMA(p, d, q, \alpha)$, for it to be correctly specified (converge a.s.) the following must hold [6]

$$H = d + 1/\alpha < 1 \tag{9}$$

Furthermore if roots of polynomial $\mathcal{A}_p$ lie outside of unit circle the ARFIMA process is stationary.

Stationary ARFIMA process is asymptotically H self-similar with $H = d + 1/\alpha$. The most important result for us is the following limiting relation, let X be ARFIMA process then

$$N^{-H} \sum_{i=1}^{\lfloor Nt \rfloor} X_i \overset{\mathcal{D}}{\to} L_H^{\alpha}(t) \qquad N \to \infty \tag{10}$$

So ARFIMA model can be considered as discrete time version of Levy fractional processes. The case $d = 0$ leads to ARMA processes (with α-stable noise) and exponentially decaying autocorrelation functions. The case $d > 0$ is similar to the case of fractional Brownian motion and leads to long range dependence

$$\sum_{k=0}^{\infty} E[X(0)X(k)] = \infty \tag{11}$$

The case $d < 0$ is analogical to the case of fractional Brownian motion with $H < 1/2$ and leads to short and negative correlations.

Even though ARFIMA is discrete time model it is quite complicated and even simulation of ARFIMA is quite tricky. However it is much more tractable than fractional Levy processes and at the same time it exhibits fractional dynamics caused by both non-trivial correlation structure and by infinite variance of its noise process.

3 Parameter estimation and transmutation of statistics

The methods for estimating parameters α and d of ARFIMA model are presented in this section and applied to S&P 500 daily data.

3.1 Numerical estimation of ARFIMA parameters

ARFIMA model is defined by four parameters d, α, p, q and by $p + q$ coefficients of polynomials $\mathcal{A}$ and $\mathcal{B}$. Due to large complexity of ARFIMA model parameters p, q are often assumed to be equal to one at most which still gives the ARFIMA model sufficient generality. The estimation of coefficient of polynomials $\mathcal{A}, \mathcal{B}$ can then be formulated as well defined optimization problem and solved numerically [6].

However the most important parameters of ARFIMA model are the two parameters defining the fractional nature of the model d and α. We will introduce computationally simple methods to estimate these parameters in the following paragraphs.

Estimation of anomalous diffusion parameter

The parameter d effects the memory effects of the underlying process, it is sometimes called memory or long range dependence parameter.

Most common way to estimate this parameter is so called rescaled range (R/S) method [11]. It estimates Hurst exponent as

$$E[\frac{R(t)}{\sigma(t)}] \sim Ct^H \qquad t \to \infty \tag{12}$$

where $R(t)$ is a range of the cumulative sum of the underlying stationary (noise) process and $\sigma(t)$ denotes standard deviation of the noise process.

However this method returns the true parameter H only in Gaussian case, it generally gives value $d + 1/2$ which is equal to the true Hurst exponent only in the case $\alpha = 2$. In other words this method assumes that the fractional behaviour of the underlying self-similar process is caused solely by the correlations between increments (i.e. the underlying process is fractional BM) and therefore fails in the general case of fractional Levy processes.

Similarly there are methods assuming that fractional behaviour of self-similar process is caused solely by infinite variance of the underlying noise process. Mantegna and Stanley for example proposed the following test [9]:

For process with stationary increments self-similarity implies the following relation $p_t(0) = \frac{1}{t^H}p_1(0)$. First we estimate an empirical density at zero $\hat{p}_t(0)$. This can be done from the histogram for example. Then we get the following relation

$$\ln \hat{p}_t(0) \simeq H \ln \frac{\Delta}{t} + \ln \hat{p}_\Delta(0) \tag{13}$$

Mantegna and Stanley applied this to SP 500 and obtained $H \simeq 0,55$. They concluded the α-stable model with $\alpha \simeq 1,8$. However this test implicitly assumes that the other source of fractional behaviour is not present (i.e. that process has independent increments).

We instead propose the following simple method which seems to provide accurate estimated of parameter d. We define mean sample displacement as follows

$$M_N(t) = \frac{1}{N-t-1}\sum_{i=0}^{N-t}(X_{i+t} - X_i)^2 \tag{14}$$

The key result is that if the process X_t is cumulative sum process of stationary ARFIMA process (with $\alpha > 1$) then the following asymptotic relation holds (for large N) [6]

$$M_N(t) \sim t^{2d+1} \tag{15}$$

So clearly the case $d = H - 1/\alpha > 0$ corresponds to super diffusion and $d = H - 1/\alpha < 0$ leads to sub diffusion. Interesting is the case $d < 0$ with non-Gaussian noise, in this case the large jumps produced by stable noise are compensated by large jumps of opposite sign and on average the diffusion of the process is slower that in the standard Brownian case.

The proposed method of estimation of parameter d is the following:

1. Estimate $M_N(t)$ for $t = 1, 2, .., 10$
2. Run regression $\ln M_N(t) \sim \ln t$
3. Take the calculated slope δ and calculate $d = \frac{\delta - 1}{2}$

The proposed estimator is consistent, it has been tested and seems to produce reliable results. However in some case it is required to calculate $M_N(t)$ for more values of t before running the regression.

Estimation of stability index

There is number of ways in which parameters of stable distribution can be fitted. The most common ones are maximum likelihood method (using approximate likelihood function), methods based and tabulated quantile values and methods based on regression of empirical characteristic function. The regression methods seems to be most reliable [14]. However we are interested only in stability parameter α, so we will follow different approach. We present method for calculation of Hurst index H applicable for general class of fractional Levy processes. Combined with previous method to estimate d we then obtain stability index as $\alpha = \frac{1}{H-d}$.

We will apply concept of p-variation for this analysis, we define sample p-variation of process $X_{i \in \{1..N\}}$ of lag m as [16]

$$V_m^p = \sum_{i=0}^{N/m-1} |X_{(i+1)m} - X_{im}|^p \tag{16}$$

Let us assume that X is cumulative sum process of stationary ARFIMA process, then for sufficiently large N/m it holds [6]

1. if $\alpha = 2$ or if $1 < \alpha < 2$ and $d \geq 0$
$$V_m^p \sim m^{Hp-1} \tag{17}$$
2. if $1 < \alpha < 2$ and $d < 0$
$$V_m^p \sim m^{Hp-p/\alpha} \tag{18}$$

It worth noticing that in the first case variation increases with growing p but it decreases in the second case.

In the first case the following estimation technique can be applied.

1. Estimate V_m^p for $p = 1/\{0.01, 0.02..1\}$
2. For fixed m find p that minimizes $(\frac{V_m^p - V_1^p}{V_1^p})^2$
3. estimate $H = 1/p$

The appropriate choice of m has to done based on sample size, generally it is better to choose larger m as long as N/m remains sufficiently large.

The second case can be transformed into the first case by using concept of surrogate data, which means essentially reshuffling the (stationary increment) data. That should break the correlation structure within the data and essentially set $d = 0$, then the same approach can be applied.

3.2 Transmutation of statistics in financial time series

We will apply the methods presented above to detect the transmutation of statistics in S&P 500 daily data observed between years 1950 to 2007. There is approximately 14500 data points in analysed time series.

Proposed approach assumes that logarithms of observed prices follow fractional Levy process. This is quite standard approach, in fact famous Black-Scholes theory assumes that logarithms of asset prices follow particular case of fractional Levy processes - Brownian motion. We checked stationarity of increments of the process using unit root test and self-similarity of log-levels using rescaled range approach, both these assumptions seems to be satisfied.

We applied methods for estimation of α and d parameters of fractional Levy process to rolling window sample of original S&P 500 daily data. The evolution of these parameters determining the fractional nature of process will allow us to detect transmutations of statistics in original data.

Some parameters of the approach were determined purely by numerical analysis, after testing different specifications we chose

1. the length of rolling window sample to be 3000 data points
2. to replace 600 data points of the rolling window sample in every iteration
3. we chose parameter m used in estimation of Hurst index to be equal to 3

We first applied the above approach to the whole dataset, we obtained

$$H \simeq 0.56, d \simeq 0.01 \Rightarrow \alpha \simeq 1.81 \tag{19}$$

Notice that this is exactly the same result that Mantegna and Stanley obtained for similar dataset using different approach based on self similarity, they ignored the fractionality causes by parameter d however in this case we can see that its effect is negligible.

Application of the above described approach on rolling window sub-sample of size 3000 data points and replacing 600 data points in every iteration yielded time evolution of Hurst index $H = d + 1/\alpha$ depicted in Figure 1. The dates in the graph are always the end dates of the corresponding rolling window sample.

The red lines denotes points where derivative of H changes sign, these will be regarded as point of transmutation of statistics. The blue line denoted point where d changes sign from positive to negative, i.e. point of transmutation from super-diffusion to sub-diffusion. The discontinuity of H in this point is caused by this transmutation, due to nature of rolling window approach this discontinuity can be seen twice, however only the first one interests us.

When we plot the original log-prices we can see that the located points of transmutation of statistics are clearly significant For the first two "red lines" the transmutation of statistics can be seen very clearly, the other two are less clear mainly due to smaller sample size in these windows. This can be also seen from the following table summarizing the different windows (separated by red lines in Figure2)

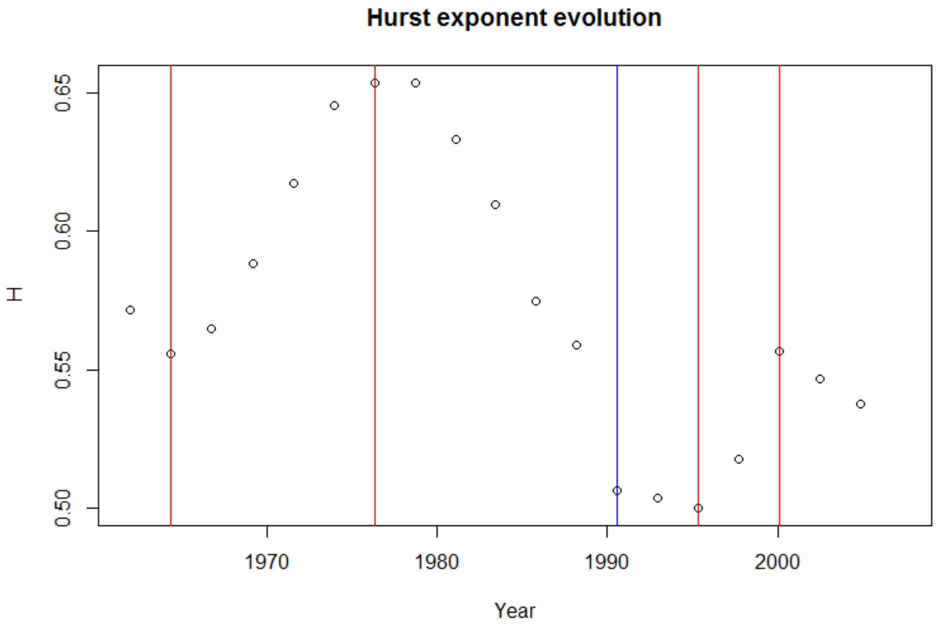


Figure 1: Hurst index evolution

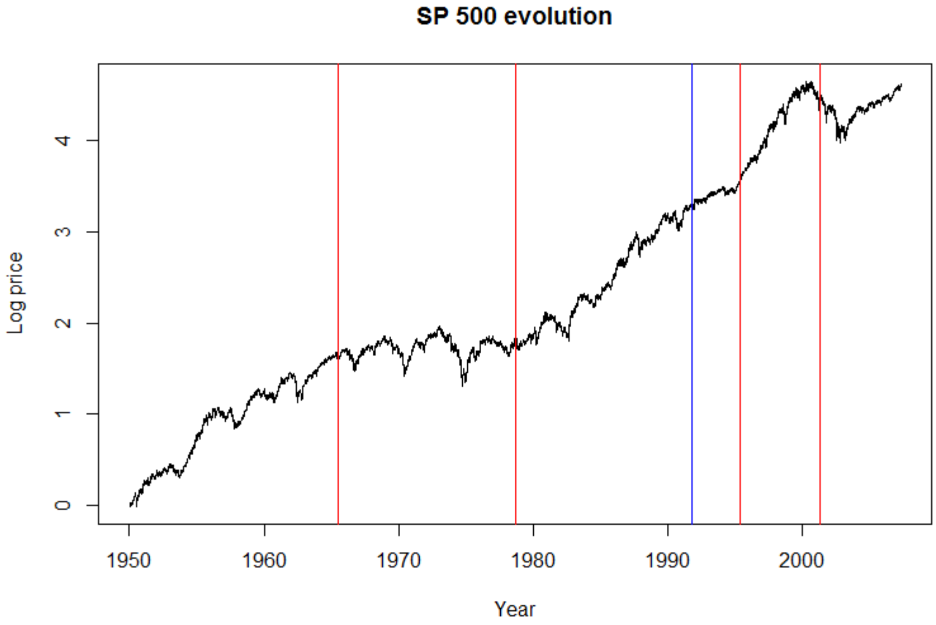


Figure 2: Hurst index evolution

	H	d	α	kurtosis	skewness	damping coefficient
1st	0.55	0.02	1.89	9	-0.7	0.06%
2nd	0.65	0.07	1.72	2.5	0.23	0.16%
3rd	0.53	0	1.89	5.8	-0.42	0.08%
4th	0.52	-0.06	1.72	3.8	-0.31	0.1%
5th	0.48	-0.04	1.9	2.9	0.13	0.14%

where damping coefficient is introduced because kurtosis of simulated values from stable distributions are much higher than observed kurtosis. The damping coefficient determines how many extreme values simulated from given stable distribution we have to exclude for simulated and observed kurtosis to match. We used simulation approach to determine these values. This idea is known from theory of truncated Levy flights [12] and damping is there introduced through cut-off of density function, for fractional processes this is more complicated however.

The transition from super-diffusive regime to sub-diffusive regime can be clearly seen. It also seems that there are two well defined regimes of stability index α, in addition the stability index of whole dataset lies approximately in the middle of these. Interesting is the third transition where H does not change much because the change in α is compensated by change of d .

Based on our analysis of this and few similar samples we can state few empirical rules that seem typically to hold

1. The point of transmutation is typically either point where d or derivative of d changes sign, in this case first and last transition are related to change of sign of derivative of d and middle transitions to sign of d
2. transitions seem often to be followed by change of sign of skewness
3. transitions caused by change of sign of d seem to behave less regularly
4. transition from super diffusive to sub diffusive regime causes discontinuities in rolling window graph of H

Conclusion

We used formalism of fractional Levy processes and ARFIMA model to detect transmutations of statistics in daily S&P 500 data observed on financial markets. The method seems very promising and we can conclude that the underlying dynamics of the observed data clearly changes in located points of transmutation. Proposed parameter estimation technique seems to be quite reliable and on the whole dataset it gave similar results as other methods typically used.

While the initial results are promising the method must be tested for much broader range of datasets. That should also give us better understanding of underlying transitions. We would also like to develop analytical framework for estimation of size of rolling window sample and for number of points that are replaced in every iteration.

The main objective would be to classify transitions observed on financial markets in unified framework, the idea we have in mind at the moment is to build in this framework

in analogy with phase transitions in statistical physics. For example the key diffusion parameter d could play similar role as thermodynamic potentials, because the statistic transmutation is typically related to discontinuity of d or of its derivative.
We also plan to apply this method to higher frequency data in the future, the results there might be quite different due to much higher volatility of diffusion parameter d. The understanding of these transition could also allow us to forecast future volatility of underlying financial time series.

References

[1] R. Metzler, J. Klafter. *The random walks's guide to anomalous diffusion: A fractional dynamics approach.* Physics reports 339 1-77, 2000.

[2] P. Tankov. *Financial modelling with jump processes.* Chapman-Hall/CRC, 2003.

[3] K. S. Miller, B. Ross. *An introduction to the fractional calculus and fractional differential equations.* Wiley, 1993.

[4] M. Teuerle, A. Wylomanska, G. Sykora *Modelling anomalous diffusion by subordinated Levy stable processes.* Journal of statistical mechanics: Theory and experiment. 10.1088/1742-5468/2013/05/P05016.

[5] Granger, C. W. J.; Joyeux, R. . *An introduction to long-memory time series models and fractional differencing.* Journal of Time Series Analysis (1980). 1: 15–30.

[6] K. Burnecki, A. Weron. *Algorithms for testing of fractional dynamics: a practical guide to ARFIMA modelling.* Journal of statistical mechanics: Theory and experiment. 10.1088/1742-5468/2014/10/P10036.

[7] T. Marquardt. *Fractional Levy processes with an application to long memory moving average process.* Bernoulli 12(6), 2006, 1099-1126.

[8] K. Falconer. *Fractal geometry, Mathematical foundations and applications.* Wiley, 1989.

[9] R. N. Mantegna, H. E. Stanley. *An introduction to econophysics.* CUP, Cambridge, 2000.

[10] B. V. Gnedenko, A. N. Kolmogorov. *Limit distributions for sums of independent random variables.* Adison-Wesley, 1968.

[11] B.B. Mandelbrot. *Fractals and scaling in finance.* SELECTA VOLUME E, 1996.

[12] W.Paul, J. Baschnagel. *Stochastic Processes: From physics to finance.* Springer, 2000.

[13] Hosking, J. R. M. *Fractional differencing.* Biometrika. 68 (1): 165–176, 1981.

[14] M. Matsui, A. Takemura. *Goodness of fit tests for symmetric stable distributions.* http://www.e.u-tokyo.ac.jp/cirje/research/03research02dp.html.

[15] B.Oksendal. *Stochastic differential equations.* Springer, 1994.

[16] R. Norvaise, D.Salopek. *Estimating the p-Variation Index of a Sample Function: An Application to Financial Data Set.* Methodology And Computing In Applied Probability, March 2002, Volume 4, Issue 1, pp 27–53.

Bayesian Matrix Factorization in Multiple-Instance Learning Problem

Vít Škvára

1st year of PGS, email: skvarvit@fjfi.cvut.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, CAS

Abstract. Multiple-instance learning (MIL) is a subset of supervised binary classification. In MIL, multiple instances (feature vectors) belonging to a single individual are collected in *bags* which are then labeled as positive or negative. Usually, no indication is given whether the label is given by a number of positive/negative observations in a bag or if the bags differ with their entire structure. In this contribution we research the possibility of representing the internal structure of bags by a set of base vectors and selection matrices which are unique for each bag. This leads to an ill-posed matrix factorization problem which we solve by employing the Bayesian framework. Performance of the resulting algorithm is validated on a testing MIL dataset. Also, motivation is given by describing a real-world MIL problem of detection of malware infected computers.

Keywords: multiple-instance learning, supervised learning, variational Bayes, matrix factorization

Abstrakt. *Multiple-instance learning* (MIL) je druhem binární klasifikace s učitelem. MIL problémy se vyznačují tím, že ke každému jedinci existuje několik vektorů příznaků sdružených do jediné matice - tzv. *bagu.* Každému bagu jako celku je pak přiřazeno označení pozitivní/negativní. Není přitom dáno, zda je např. pozitivní označení způsobeno několika pozitivními vektory mezi zbytkem negativních nebo zda se liší celková struktura bagů. V tomto příspěvku se zabýváme reprezentací vnitřní struktury bagů pomocí množiny základních vektorů a výběrových matic, unikátních pro každý bag. Řešení této špatně podmíněné úlohy je navrženo ve tvaru maticové faktorizace a je hledáno pomocí bayesovského hierarchického modelu. Odvozený algoritmus je otestován na vzorovém MIL datasetu. V textu je také popsána motivace daná problémem vyvstávajícím v detekci malwarem napadených počítačů.

Klíčová slova: multiple-instance learning, učení s učitelem, variační Bayes, maticová faktorizace

1 Introduction

In multiple-instance learning (MIL), the problem of supervised binary classification is made more difficult for the learner due to a number of reasons. Firstly, instead of having a set of instances (feature vectors) labeled as negative or positive, a number of *bags* of instances is received, where the whole bags are labeled as positive or negative. Every bag consists of a (possibly different) number of instances whose individual labels are not known. The common conception is that a bag is labeled negative if all instances in it are negative, but if even a single instance is positive, then the label of the bag is also positive [6]. Secondly, the ratio of negative to positive instances in a bag can be arbitrarily high. In real-world problems however, this presumption can be violated and positive and negative bags may be generated from entirely different sources. In [4], the MIL model was formalized for the first time and a solution using axis-parallel rectangles constructed by the conjunction of the features was proposed.

A number of different MIL problems were solved using different approaches. Decision trees were used [3] for the drug activity prediction problem. Boosting algorithm was proposed to be used for face detection in [2]. The kNN nearest algorithm with Hausdorff distance [11], a variant of the support vector machine algorithm [1] or various set distances [10] were compared on a common MIL dataset.

The problem of malware detection in computers connected to a network whose activity we supervise is of particular interest to us. In such a case, the communication of every computer with the outside world (using a HTTP protocol) goes through a common hub. The observer, for a limited time frame, collects all HTTP requests of the computers in the network. From each request, we substract a number of features (e.g. bytes sent and received, request lenght in ms). A collection of such instances for one computer creates a bag. Additionaly, some computers are known to be infected with malware that communicates with the Internet. Their bags are then labeled as positive and together with bags of some uninfected computers compose a training dataset. Presumably, positive bags should contain a number of positive instances - requests created by the malware. This poses an interesting MIL problem, as the ratio of positively labeled bags to negatively labeled ones is small (∼2%) and it is possible that not all positively labelled bags actually contain a malware-originated request. Decision trees [8] and neural networks [9] were used to tackle this problem.

In the following text, the MIL problem will be formalized. Also, an approach leading to matrix factorization will be outlined. The solution will be sought after using Bayesian formalism. The performance of the resulting algorithm will be presented on a well-known MIL dataset. Finally, some comments will be made about the method and the future outlook

2 MIL and matrix factorization

Let the structure of the training MIL dataset be following: there are N bags - matrices $Y_n \in \mathbb{R}^{L\times M_n}, n \in \widehat{N}$. The columns of each bag are instances $y_{nm} \in \mathbb{R}^L, m \in \widehat{M_n}$. Labels of the bags are stored in the vector $x \in \{0,1\}^N$. Now, let $Y \in \mathbb{R}^{L\times M}$ be a single bag. Consider the following factorization

$$Y = BA^T + E, \tag{1}$$

where $B \in \mathbb{R}^{L\times H}$ is a matrix consisting of a few base, general instances. $A \in \mathbb{R}^{M\times H}$ can be thought of as a selector matrix that chooses the base instances for a given Y. Matrix $E \in \mathbb{R}^{L\times M}$ is the noise. This model can very well represent a true MIL problem as it can be expected that there is a number H of universal instances that repeat across and inside bags. Some of these can be positive and some negative.

Computation of this factorization is an ill-posed problem and has infinitely many solutions. To achieve the properties described above, we must impose some further restrictions on the simple model (1). This will be discussed in the next section.

Now, suppose that we concatenate all the positive and negative bags together in two general matrices

$$T_0 \in \mathbb{R}^{L \times M_0}, M_0 = \sum_{\substack{n=1 \\ x_n=0}}^{N} M_n, \tag{2}$$

$$T_1 \in \mathbb{R}^{L \times M_1}, M_1 = \sum_{\substack{n=1 \\ x_n=1}}^{N} M_n. \tag{3}$$

By computing the factorization (1) for T_0 and T_1, we obtain two base matrices B_0 and B_1. If we computed them in accordance with the properties stated above, then they should differ by a number of base positive instances. When deciding on the label of an unknown bag Y_{N+1}, we compute matrices A_0 and A_1 from (1) with Y_{N+1} on the right side and with a fixed B_0 and B_1, respectively. Then the label is given by the decomposition where the reconstruction has smaller error, i.e.

$$x_{N+1} = argmin\{||Y_{N+1} - B_i A_i^T||_2 : i \in \{0, 1\}\}, \tag{4}$$

where $||.||_2$ is the matrix L^2-norm. The classification is based on the assumption that decomposing with respect to a correct base should be more precise than the using the wrong one. However this might not be true for all MIL datasets.

3 Variational Bayes matrix factorization

In this section, we build a bayesian hierarchical model around the simple model (1) with the proposed factorization properties in mind, i.e. B is a matrix of base instances and A is a selector matrix. To achieve this, we want the matrix A to be sparse. In an ideal case, A would only consist of ones and zeros as it selected the apropriate instances encoded in B. In a Bayesian context, the property of achieving sparsity is called ARD (automatic relevance determination, see [12]).

3.1 The hierarchical model

We will start by choosing the data likelihood and prior for B in accordance with [7], where ARD is implied on the columns of B and A in order to reduce the inner dimension H. However, to achieve the proposed sparsity, we will impose the ARD property on every single element of A by choosing a normal distribution of vectorized matrix A instead of the the original matrix normal distribution. The data likelihood and priors on A, B are chosen as

$$p(Y|B, A, \sigma) = \mathcal{MN}(Y|BA^T, \sigma^{-1} I_L, I_M), \tag{5}$$

$$p(\mathrm{vec}(A^T)|C_A) = \mathcal{N}(\mathrm{vec}(A^T)|0, C_A^{-1}), \tag{6}$$

$$p(B|C_B) = \mathcal{MN}(B|0, I_L, C_B^{-1}). \tag{7}$$

Here, $\mathcal{NM}(.)$ is the matrix normal distribution and $\mathcal{N}(.)$ is the normal distribution, I_d is identity matrix of size d. Prior distributions for covariances are following:

$$p(C_A) = \prod_{h,m=1}^{H,M} \mathcal{G}(C_{Amh}|\alpha_0, \beta_0), \tag{8}$$

$$C_A = \mathrm{diag}(C_{A11}, C_{A12}, \dots, C_{AMH}), \tag{9}$$

$$p(C_B) = \prod_{h=1}^{H} \mathcal{G}(C_{Bh}|\gamma_0, \delta_0) \tag{10}$$

$$C_B = \mathrm{diag}(C_{B1}, \dots, C_{BH}), \tag{11}$$

$$p(\sigma) = \mathcal{G}(\sigma|\eta_0, \zeta_0), \tag{12}$$

where $\mathcal{G}(.)$ denotes the gamma distribution. It is actually through the estimation of the precisions (inverse variances) that the ARD property is achieved.

3.2 The Variational Bayes method

The joint probability distribution of the data and the parameters is now

$$p(Y, \Theta) = p(Y, A, B, C_A, C_B, \sigma) = p(Y|B, A, \sigma)p(A|C_A)p(B|C_B)p(C_A)p(C_B)p(\sigma), \tag{13}$$

where the simplification $\Theta = (A, B, C_A, C_B, \sigma)$ is used. The structure of the model does not permit a direct evaluation of the true posterior $p(A, B, C_A, C_B, \sigma|Y) = p(\Theta|Y)$. Instead of resorting to MCMC methods, we use the computationally less expensive Variational Bayes (VB) framework. Using some approximations, VB will enable us to come to an analytic expression for an equilibrium state that describes the parameters of the posterior.

VB approximates the true posterior distribution with a product of mutually independent posteriors

$$p(\Theta|Y) \approx q(\Theta|Y) = q(A|Y)q(B|Y)q(C_A|Y)q(C_B|Y)q(\sigma|Y). \tag{14}$$

The fixed log marginal probability of Y can be expressed as

$$\ln p(Y) = \int q(\Theta|Y) \ln \left(\frac{p(Y, \Theta)}{q(\Theta|Y)} \right) d\Theta \tag{15}$$

$$+ \int q(\Theta|Y) \ln \left(\frac{q(\Theta|Y)}{p(\Theta|Y)} \right) d\Theta \tag{16}$$

$$= \mathcal{F}(q) + \mathrm{KL}\left(q(\Theta|Y) || p(\Theta|Y) \right). \tag{17}$$

Here, $\mathcal{F}(q)$ is the *free energy* and KL(.) is the Kullback-Leibler divergence between the true and the approximate posterior. It is an integral probability measure and is equal to zero if the two arguments are equal. Because KL divergence is always non-negative, we can minimize it by choosing the right forms of approximate posteriors in $q(\Theta|Y)$ that maximize the negative free energy $\mathcal{F}(q)$, thus bringing the approximate posterior closer to the true one. From the VB theory, the posteriors that maximize the free energy have the form

$$\ln q(\Theta_i|Y) = \mathbb{E}_{q(\Theta_j|Y),j\neq i}\left[\ln p(Y,\Theta)\right], \tag{18}$$

where the expectation is taken over all other approximate posteriors $q(\Theta_j|Y)$ but the i-th. For details see Chapter 3 in [13]. Using conjugate priors, the posterior distributions have known forms and analytical expressions of their parameters.

3.3 The approximate posterior

In this place, we will analytically derive the posterior distribution for variance of the data σ using prescription (18). It is a simple and straightforward computation compared to other posteriors but it ilustrates the principle of the VB method. Recollecting the form of the likelihood, the priors (5) - (12) and using (18) we have

$$\ln q(\sigma|Y) = (\eta_0 - 1)\ln\sigma - \zeta_0\sigma - \sigma\frac{1}{2}\mathrm{tr}\left(\mathbb{E}\left[\left(Y - BA^T\right)^T\left(Y - BA^T\right)\right]\right) + \frac{ML}{2}\ln\sigma + \mathrm{const.} \tag{19}$$

Here, const. stands for terms that are not dependent on σ and that are considered to be part of the integration constant of the posterior distribution. Expectation is computed with respect to the other posteriors. By collecting the terms for σ and $\ln\sigma$, we see that the posterior of σ is again a Gamma distribution of the following form

$$q(\sigma|Y) = \mathcal{G}(\sigma|\eta,\zeta), \tag{20}$$

$$\eta = \eta_0 + \frac{ML}{2}, \tag{21}$$

$$\zeta = \zeta_0 + \frac{1}{2}\mathrm{tr}\left(\mathbb{E}\left[\left(Y - BA^T\right)^T\left(Y - BA^T\right)\right]\right). \tag{22}$$

Clearly, the posterior balances the influence of the prior and the data. Usually, the prior parameters η_0, ζ_0 are set to small values (e.g. 10^{-10}) to keep the estimates unbiased.

Following the procedure outlined above for the rest of the estimated parameters, we arrive at the following posterior distributions

$$q(\mathrm{vec}(A^T)|Y) = \mathcal{N}(\mathrm{vec}(A^T)|\mu_A,\Sigma_A), \tag{23}$$

$$q(B|Y) = \mathcal{MN}(B|M_B, I_L, \Sigma_B), \tag{24}$$

$$q(C_{Amh}|Y) = \mathcal{G}(C_{Amh}|\alpha_{mh},\beta_{mh}), \tag{25}$$

$$q(C_{Bh}|Y) = \mathcal{G}(C_{Bh}|\gamma_h,\delta_h), \tag{26}$$

with their shaping parameters given by this set of equations:

$$\mu_A = \widehat{\sigma}\Sigma_A\mathrm{vec}(\widehat{B}^TY), \qquad \Sigma_A = (\widehat{C_A} + \widehat{\sigma}I_M \otimes \widehat{B^TB})^{-1}, \tag{27}$$

$$M_B = \widehat{\sigma}Y\widehat{A}\Sigma_B, \qquad \Sigma_B = (\widehat{\sigma}\widehat{A^TA} + \widehat{C_B})^{-1}, \tag{28}$$

$$\alpha_{mh} = \alpha_0 + \frac{1}{2}, \qquad \beta_{mh} = \beta_0 + \frac{1}{2}\widehat{A^2_{mh}}, \tag{29}$$

$$\gamma_h = \gamma_0 + \frac{L}{2}, \qquad \delta_h = \delta_0 + \frac{1}{2}\widehat{B_h^TB_h}. \tag{30}$$

Here, the notation $\widehat{.}$ is used for expectation over posterior distribtutions and $\otimes$ is used for Kronecker product. The equations contain a number of lower and higher moments that can be expresed using the shaping parameters and well known properties of used distributions. They have the following form

$$\widehat{A} = \text{devec}(\mu_A)^T, \qquad \widehat{A^TA} = \widehat{A}^T\widehat{A} + \sum_{m=1}^{M} \text{sub}(\Sigma_A, m, H), \tag{31}$$

$$\widehat{B} = M_B \qquad \widehat{B^TB} = \widehat{B}^T\widehat{B} + L\Sigma_B, \tag{32}$$

$$\widehat{C_A} = \text{diag}\left(\frac{\alpha_{11}}{\beta_{11}}, \dots, \frac{\alpha_{MH}}{\beta_{MH}}\right), \qquad \widehat{C_B} = \text{diag}\left(\frac{\gamma_{11}}{\delta_{11}}, \dots, \frac{\gamma_{MH}}{\delta_{MH}}\right), \tag{33}$$

$$\widehat{\sigma} = \frac{\eta}{\zeta}, \tag{34}$$

$$\text{tr}\left(\mathbb{E}\left[\left(Y - BA^T\right)^T\left(Y - BA^T\right)\right]\right) = \text{tr}\left(Y^TY + \widehat{B^TB}\widehat{A^TA} - 2Y\widehat{A}\widehat{B^T}\right). \tag{35}$$

The notation devec(.) is used for the operation of devectorization a vector into a matrix of the original size, $\text{sub}(\Sigma_A, m, H)$ is the m-th diagonal submatrix of Σ_A of size $H \times H$.

To compute the solution of the system of equations, we use an iterative algorithm. It starts with some initial values for the shaping parameters. Then, the shaping parameters of each posterior are updated using the equations (21), (22), (27) - (30) and keeping the shaping parameters of other posteriors fixed. This way, it is guaranteed that a local minimum of KL divergence is found [5]. The algorithm is described in Algorithm 1.

Algorithm 1: VBMF - Variational Bayes Matrix Factorization

```
input  : bag $Y \in \mathbb{R}^{L \times M}$, inner dimension $H$, stopping conditions
         $maxIter \in \mathbb{N}, \varepsilon \in \mathbb{R}$
output: shaping parameters of posterior distrubution $q(\Theta|Y)$
initialization: initialize the values of shaping parameters, set
 $nIter = 0, B_\delta = M_B, \delta = \varepsilon + 1$;
while $niter < maxIter \wedge \delta > \varepsilon$ do
    update shaping parameters of $q(\text{vec}(A^T)|Y)$ using (27) ;
    update shaping parameters of $q(B|Y)$ using (28) ;
    update shaping parameters of $q(C_A|Y)$ using (29) $\forall m \in \widehat{M}, h \in \widehat{H}$ ;
    update shaping parameters of $q(C_B|Y)$ using (30) $\forall h \in \widehat{H}$ ;
    update shaping parameters of $q(\sigma|Y)$ using (21), (22) ;
    set $\delta = \frac{||B_\delta - M_B||}{||M_B||}$;
    set $B_\delta = M_B$;
    set $niter = niter + 1$;
report $B = M_B$, $A = \text{devec}(\mu_A)^T$ and the rest of estimated shaping parameters;
```

3.4 The classification algorithm

This section compiles the whole procedure of training the classification algorithm and then using it to classify a new sample bag. The basic idea of classification was already described in section 2.

```
Algorithm 2: MIL classification using VBMF
  input  : training dataset $\{T_0, T_1\}$, testing dataset $\{Y_1, \dots, Y_D\}$
  output: estimated labels $\{x_1, \dots, x_D\}$
  using Algorithm 1 on the matrix $T_0$, compute negative basis $B_0$ and its
    covariance $\Sigma_{B0}$;
  using Algorithm 1 on the matrix $T_1$, compute positive basis $B_1$ and its covariance
    $\Sigma_{B1}$;
  for $d = 1, \dots, D$ do
      Compute the backward factorization with fixed $B_0$;
        • initalize Algorithm 1 for $Y_d$ with $M_B = B_0, \Sigma_B = \Sigma_{B0}$
        • compute the rest of Algorithm 1, ommiting updates for $q(B|Y)$ ;
        • report estimates and set $A_{d0} = A$
      Compute the backward factorization with fixed $B_1$;
        • initalize Algorithm 1 for $Y_d$ with $M_B = B_1, \Sigma_B = \Sigma_{B1}$
        • compute the rest of Algorithm 1, ommiting updates for $q(B|Y)$ ;
        • report estimates and set $A_{d1} = A$
      set the estimate of label as $x_d = argmin\{||Y_d - B_i A_{di}^T||_2 : i \in \{0, 1\}\}$,
```

4 Validation

The classification algorithm was tested on set of well-known datasets of MIL problems. While on some it did not perform well, on a few particular datasets the classification procedure did achieve some success. This is the case for other MIL algorithms, that are sometimes tuned with a particular dataset in mind. An overview of the size of the datasets is in table 4. In these datasets, the labels of all bags are known.

dataset	number of bags N	instance length L	average bag size M
BrownCreeper	548	38	19
Musk1	166	92	5
WinterWren	548	38	19

On these datasets, the classification algorithm was tested in the following manner: a) a subset of bags was randomly chosen and used as traning data b) for every bag in the remaining (testing) subset, the classification was computed. This was repeated more times. Each time, an error metric called *equal error rate* (EER) was computed. We define it as

$$\mathrm{EER} = \left(\frac{\sum \text{false negatives}}{\sum \text{positive labels}} + \frac{\sum \text{false positives}}{\sum \text{negative labels}} \right) /2 \tag{36}$$

It is used here because of the unbalanced number of negative and positive samples in some datasets.

The matrix Σ_A has a total of M^2H^2 elements. For some datasets, this slows down the computation due to memory allocation and a very difficult inversion of the term in (27). A compromise between precision and speed has been done so that for $MH > 200$ only the diagonal of the matrix is estimated and kept in memory. When compared to the computation of the whole matrix, this does not lead to significantly deteriorated results.

For the 3 datasets, the histograms of EER for different ratio of training and testing data for 100 tries is in Figure 1. Missing entries for smaller percentage of known labels are caused by numerical difficulties when inverting ill-conditioned matrices. Clearly, for larger percentage of known labels the mean error is smaller and is in the range of 10-20%.

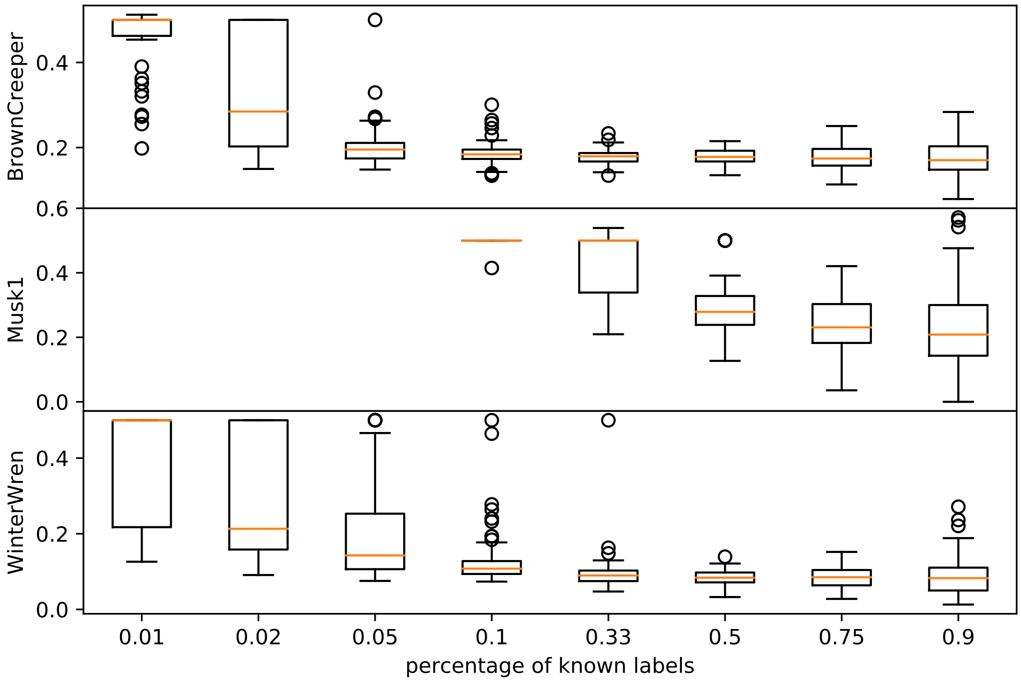


Figure 1: Equal error rate histograms for the classification experiment on available datasets. Internal factorization dimension $H = 5$, 100 retries for each known label percentage.

5 Discussion

In this article, an introduction to multiple-instance learning was given with the motivation for the work given by malware datection in a network of computers. Unfortunately, real-world data from this field are not yet available, so all experiments were only made on a set of well-known MIL problems. In the rest of the paper, basic idea behind the method was described and further elaborated using Bayesian formalism. The proposed hierarchical model was detailed together with the resulting algorithms for matrix factorization and classification of MIL datasets.

In comparison to other MIL algorithms, the classification error of our method is still high, as the cutting-edge approaches achieve EER in the range of 5-10%. Clearly, further work is required to be able to compete. The direction in which to improve is certainly the classification rule, which is now based on a very simple error criterion.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. *Support vector machines for multiple-instance learning.* In 'Advances in neural information processing systems', 577–584, (2003).

[2] B. Babenko, M.-H. Yang, and S. Belongie. *Visual tracking with online multiple instance learning.* IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. (2009), 983–990.

[3] H. Blockeel, D. Page, and A. Srinivasan. *Multi-instance tree learning.* In 'Proceedings of the 22nd international conference on Machine learning', 57–64. ACM, (2005).

[4] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. *Solving the multiple instance problem with axis-parallel rectangles.* Artificial intelligence **89** (1997), 31–71.

[5] Z. Ghahramani and M. J. Beal. *Propagation algorithms for variational bayesian learning.* In 'Advances in neural information processing systems', 507–513, (2001).

[6] O. Maron and T. Lozano-Pérez. *A framework for multiple-instance learning.* In 'Advances in neural information processing systems', 570–576, (1998).

[7] S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. *Global analytic solution of fully-observed variational bayesian matrix factorization.* Journal of Machine Learning Research **14** (2013), 1–37.

[8] T. Pevny and P. Somol. *Discriminative models for multi-instance problems with tree structure.* In 'Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security', 83–91. ACM, (2016).

[9] T. Pevnỳ and P. Somol. *Using neural network formalism to solve multiple-instance problems.* In 'International Symposium on Neural Networks', 135–142. Springer, (2017).

[10] Q. N. Tran, B.-N. Vo, D. Phung, B.-T. Vo, and T. Nguyen. *Multiple instance learning with the optimal sub-pattern assignment metric.* arXiv preprint arXiv:1703.08933 (2017).

[11] J. Wang and J.-D. Zucker. *Solving multiple-instance problem: A lazy learning approach.* (2000).

[12] D. P. Wipf and S. S. Nagarajan. *A new view of automatic relevance determination.* In 'Advances in neural information processing systems', 1625–1632, (2008).

[13] V. Šmídl and A. Quinn. *The variational Bayes method in signal processing*. Springer, (2006).

Lazy Fully Probabilistic Design: Application Potential[*]

Jakub Štěch[†]

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

advisor: Tatiana V. Guy, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS

Abstract. The article addresses an approach to decision making when a decision maker (human or artificial) uses incomplete knowledge of environment and faces high computational limitations. It considers a closed decision-making (DM) loop consisting of *agent-environment* pair described by agent's actions and environment states (possibly partially observable). Agent's DM problem (estimation, filtering, prediction, classification) is to influence the environment behavior in a desired way by choosing and applying a tailored DM policy generating optional actions with respect to environment.
In general LL is an approach that searches and uses relevant information from the past data and use solutions already invented (analogical modelling, memory-based prediction, transfer learning, ...). Particularly, the lazy FPD uses currently observed data to find and employ past closed-loop similar to the actual ideal represents preferences.

Keywords: decision making, lazy learning, fully probabilistic design

Abstrakt. Článek se zabývá přístupem k rozhodování při neúplné znalosti prostředí a vysokým výpočetním omezením. Uvažujeme uzavřenou smyčku složenou z páru agent a prostředí popsaného pomocí akcí agenta a stavů prostředí (částečně pozorovatelných). Cíl agenta je ovlivnit chování prostředí výběrem a uplatněním rozhodovací strategie. Lazy learning je obecný přístup, který vyhledává a používá relevantní informace z pozorovaných dat a používá již vyvinutá řešení.

Klíčová slova: rozhodování, lazy learning, plně pravděpodobnostní návrh

Full paper: This paper has been submitted to the 15th European Conference on Multi-Agent Systems EUAMAS2017, Paris, Evry, December 2017.

[*] **Acknowledement:** The research has been partially supported by the project GA16-09848S.

[†] Co-authors Tatiana V. Guy and Siavash Fakhimi Derakhshan

The Communication Library DIALOG for iFDAQ of the COMPASS Experiment

Ondřej Šubrt

3rd year of PGS, email: subrton2@fjfi.cvut.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Tomáš Liška, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Modern experiments in high energy physics impose great demands on the reliability, the efficiency, and the data rate of Data Acquisition Systems (DAQ). This contribution focuses on the development and deployment of the new communication library DIALOG for the intelligent, FPGA-based Data Acquisition System (iFDAQ) of the COMPASS experiment at CERN. The iFDAQ utilizing a hardware event builder is designed to be able to readout data at the maximum rate of the experiment. The DIALOG library is a communication system both for distributed and mixed environments, it provides a network transparent inter-process communication layer. Using the high-performance and modern C++ framework Qt and its Qt Network API, the DIALOG library presents an alternative to the previously used DIM library. The DIALOG library was fully incorporated to all processes in the iFDAQ during the run 2016. From the software point of view, it might be considered as a significant improvement of iFDAQ in comparison with the previous run. To extend the possibilities of debugging, the online monitoring of communication among processes via DIALOG GUI is a desirable feature. In the paper, we present the DIALOG library from several insights and discuss it in a detailed way. Moreover, the efficiency measurement and comparison with the DIM library with respect to the iFDAQ requirements is provided.

Keywords: Data acquisition system, DIALOG library, DIM library, FPGA, Qt framework, TCP/IP

Abstrakt. Moderní experimenty ve fyzice vysokých energií kladou veliké nároky na spolehlivost, efektivitu a rychlost přenosu dat systémů pro sběr dat (DAQ). Tento článek se zaměřuje na vývoj a nasazení nové komunikační knihovny DIALOG pro inteligentní systém pro sběr dat založeného na FPGA (iFDAQ) experimentu COMPASS v CERNu. iFDAQ čerpá události vytvořené na úrovni hardwaru a je navržen tak, aby umožňoval čtení dat při maximální rychlosti přenosu dat z experimentu. Knihovna DIALOG je komunikační systém jak pro distribuované tak pro smíšené architektury a poskytuje síťovou transparentní meziprocesovou komunikační vrstvu. Pomocí vysoce výkonného a moderního C++ frameworku Qt a jeho Qt Network API představuje knihovna DIALOG alternativu k dříve používané knihovně DIM. Knihovna DIALOG byla plně integrována do všech procesů v iFDAQ během sběru dat v roce 2016. Tato integrace z hlediska softwaru může být považována za významné vylepšení iFDAQ ve srovnání se sběrem dat v předchozím roce. Pro rozšíření možností ladění je DIALOG GUI vítaným nástrojem pro online sledování komunikace mezi procesy. V článku prezentujeme knihovnu DIALOG z několika pohledů a detailně ji diskutujeme. Kromě toho je k dispozici výkonnostní měření a porovnání s knihovnou DIM s ohledem na požadavky iFDAQ.

Klíčová slova: Systém pro sběr dat, knihovna DIALOG, knihovna DIM, FPGA, Qt framework, TCP/IP

Full paper: O. Šubrt, Y. Bai, M. Bodlak, V. Frolov, S. Huber, V. Jary, I. Konorov, D. Levit, J. Novy, D. Steffen, M. Virius. *The Communication Library DIALOG for iFDAQ of the COMPASS experiment.* ICHEP 2017 conference, Paris, France, September 2017. Available at: `http://waset.org/publications/10007840/the-communication-library-dialog-for-ifdaq-of-the-compass-experiment`.

References

[1] P. Abbon et al. *The COMPASS experiment at CERN.* Nucl. Instrum. Methods Phys. Res. (2007), 455 – 518.

[2] V. Y. Alexakhin et al. *COMPASS-II Proposal.* The COMPASS Collaboration, (May 2010). CERN-SPSC-2010-014, SPSC-P-340.

[3] T. Anticic et al. *ALICE DAQ and ECS Users Guide.* CERN, EDMS 616039 (January 2006).

[4] M. Bodlak et al. *Developing Control and Monitoring Software for the Data Acquisition System of the COMPASS Experiment at CERN.* Acta polytechnica: Scientific Journal of the Czech Technical University in Prague (2013).

[5] M. Bodlak et al. *New data acquisition system for the COMPASS experiment.* Journal of Instrumentation **8** (2013), C02009.

[6] M. Bodlak et al. *FPGA based data acquisition system for COMPASS experiment.* Journal of Physics: Conference Series **513** (2014), 012029.

[7] M. Bodlak et al. *Development of new data acquisition system for COMPASS experiment.* Nuclear and Particle Physics Proceedings **273** (2016), 976 – 981. 37th International Conference on High Energy Physics (ICHEP).

[8] CASTOR - CERN Advanced Storage manager [online]. `http://castor.web.cern.ch`. Accessed 2017-09-09.

[9] Electronic developments for COMPASS at Freiburg [online]. `http://hpfr02.physik.uni-freiburg.de/projects/compass/electronics/catch.html`. Accessed 2017-09-09.

[10] The GANDALF Module [online]. `http://wwwhad.physik.unifreiburg.de/gandalf/pages/hardware/the-gandalf-module.php?lang=EN`. Accessed 2017-09-09.

[11] iMUX/HGESICA module [online]. `http://wwwcompass.cern.ch/twiki/pub/Detectors/FrontEndElectronics/imux_manual.pdf`. Accessed 2017-09-09.

[12] Linux at CERN [online]. `http://linux.web.cern.ch/linux/scientific6/`. Accessed 2017-09-09.

[13] S-Link – High Speed Interconnect [online]. `http://hsi.web.cern.ch/HSI/s-link/`. Accessed 2017-09-09.

[14] C. Gaspar and M. Dönszelmann. *DIM – A Distributed Information Management System for the DELPHI Experiment at CERN.* Proceedings of the 8th Conference on Real-Time Computer applications in Nuclear, Particle and Plasma Physics (June 1993). Vancouver, Canada.

[15] C. Gaspar, M. Dönszelmann, and P. Charpentier. *DIM – A Distributed Information Management System for the DELPHI Experiment at CERN.* International Conference on Computing in High Energy and Nuclear Physics (February 2000). Padova, Italy.

[16] C. Gaspar and J. J. Schwarz. *A Highly Distributed Control System for a Large Scale Experiment.* 13th IFAC workshop on Distributed Computer Control Systems – DCCS'95 (September 1995). Toulouse, France.

[17] C. G. Larrea, K. Harder, D. Newbold, D. Sankey, A. Rose, A. Thea, and T. Williams. *IPbus: a flexible Ethernet-based control system for xTCA hardware.* Journal of Instrumentation **10** (2015), C02019.

Diagnosing Alzheimer's Disease Using Granger Causality*

Lucie Tylová

6th year of PGS, email: lucie.tylova@fjfi.cvut.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. The structure and function of human brain is quite complex. Various brain regions communicate with each other. Observing external potentials via EEG electrodes, we can study these communications as dependencies of multichannel EEG signal. The hypothesis presented here is that Alzheimer's diseased and normal control participants can be distinguished due to different distributions of scalp EEG-based causality measurements. The theory of Vector Auto-Regressive model and Granger causality is used to obtain the Causality Index as a novel criterion of brain activity. The general methodology is applied to real 21-channel EEG data obtained from normal control and Alzheimer's diseased groups of patients. The developed method is applicable to the localization of pathophysiological changes of Alzheimer's disease.

Keywords: VAR model, Granger causality, EEG, Alzheimer's disease, multiple-testing

Abstrakt. Struktura a funkčnost lidského mozku jsou velmi složité. Různé oblasti mozku spolu navzájem komunikují. Při sledování externích potenciálů skrz elektrody EEG můžeme studovat tuto komunikaci jako závislosti vícekanálového EEG signálu. Prezentovaná hypotéza předpokládá, že pacienti s Alzheimerovou chorobou a kontrolní účastníci mohou být od sebe odlišeni díky rozdílnému rozložení míry kauzality v naměřeném EEG. Je zde použit vektorový autoregresní model a Grangerova kauzalita k tomu, aby byl určen nový Kauzální index, který popisuje mozkovou aktivitu. Obecná metodologie je aplikována na reálná 21kanálová EEG data od zdravých pacientů a pacientů s Alzheimerovou chorobou. Vyvinutá metoda se dá použít k lokalizaci patofyziologických změn při Alzheimerově chorobě.

Klíčová slova: VAR model, Grangerova kauzalita, EEG, Alzheimerova choroba, mnohonásobné testování

1 Introduction

Alzheimer's disease (AD), the most common form of a neurodegenerative disease, causes brain cells atrophy in parallel with a decline in memory, language and everyday activities. EEG records electrical activity of the neural tissue. Thus, any pathological changes affect the resulting EEG signal [3], [1], [5]. Lower mean levels of channel-to-channel synchronization [11], [17] and greater uniformity in alpha and gamma band activity [14] have been shown in AD patients' EEG data. The dynamic relations between EEG channels, the direction of interactions, and their strength can be studied via Granger causality [8],

*This work has been supported by the grant SGS 17/196/OHK4/3T/14

[10], [4]. An alternative approach to causality investigation was used by McBride et al. [13]. In this research, the Vector Auto-Regressive (VAR) model of optimum length is directly applied to channel pairs followed by Granger causality testing to obtain novel criterion called Causality Index. Channel pairs with maximum significance of differences are localized and interpreted.

2 VAR model of multichannel EEG

The multichannel EEG data are proceeded by VAR model [16] to obtain both an optimum model order [9] and Granger causalities [7]. I suppose time series of length M in k-dimensional space and VAR(p) model of order p in the form [12]

$$\mathbf{x}_n = \mathbf{c} + \sum_{m=1}^{p} \mathbf{A}_m \mathbf{x}_{n-m} + \mathbf{e}_n \tag{1}$$

for $n = p+1, \ldots, M$ where $\mathbf{x}_n, \mathbf{c}, \mathbf{e}_n \in \mathbb{R}^k$ for $n = 1, \ldots, M$, $\mathbf{A}_m \in \mathbf{R}^{k \times k}$ for $m = 1, \ldots, p$, and $\mathbf{e}_n \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ as independent vectors. Unknown matrices $\mathbf{A}_m$ and bias vector $\mathbf{c}$ are estimated by the least squares method. The covariance matrix $\mathbf{\Sigma}$ is estimated from residues $\mathbf{r}_n$ as

$$\mathbf{C} = \frac{1}{T} \sum_{n=p+1}^{M} \mathbf{r}_n \mathbf{r}_n^{\mathrm{T}}, \tag{2}$$

where $T = M - p$ is constrain number.
The quality of VAR(p) model varies with its order p. Schwarz criterion BIC(p) (Bayesian Information Criterion) [6] is frequently used to find the optimum model order as $p_{\mathrm{opt}} \in \operatorname{argmin} \mathrm{BIC}(p)$, where

$$\mathrm{BIC}(p) = \ln|\mathbf{C}| + \frac{pk^2 \ln T}{T}. \tag{3}$$

The optimum value of model order varies segment by segment, but the most frequent value of p_{opt} (over all patients and their segments of length M) is postulated as the best choice for following Granger causality analysis [7].

3 Granger causality in investigation of multichannel EEG interactions

The k-dimensional VAR model of order p is used to investigate EEG signal dependences. Granger causality is focused on EEG channel pairs investigations. We study channels ch_i, ch_j for $i, j \in 1, \ldots, k, i \neq j$. The complete model is studied first as a two-dimensional VAR(p) model with $\mathbf{x_n} = (x_{n,i}, x_{n,j})^{\mathrm{T}} \in \mathbb{R}^2$ and $2p+1$ unknown parameters as producing residual sum SSQ_{c}.
The reduced one-dimensional case produces residual sum SSQ_{r} using also VAR(p) model of $p+1$ unknown parameters but only for channel ch_i, where $\mathbf{x_n} = x_{n,i} \in \mathbb{R}$. Therefore, in this submodel p parameters were constrained to zero values to eliminate the influence of

channel ch_j. The standard F-test of variance equity hypothesis H_0 is based on criterion

$$F = \frac{SSQ_{\mathrm{r}} - SSQ_{\mathrm{c}}}{SSQ_{\mathrm{c}}} \cdot \frac{T - 2p - 1}{p}, \tag{4}$$

which has Fisher-Snedecor distribution F of p and $T - 2p - 1$ degrees of freedom for independent channels. Applying this test to all segments of all patients, we obtain various p-values, but it is a case of multiple-testing. Therefore, False Discovery Rate (FDR) [2] correction has to be performed to obtain decreased critical value α_{FDR} as follows.

Let $H \in \mathbb{N}$ be the number of independently tested hypotheses on critical level $\alpha > 0$. The corresponding p-values are p_k for $k = 1, \ldots, H$. Comparing the sorted $p_{(k)}$ values with diminished critical levels $k\alpha/H$, we find $k^* = \max(k : p_{(k)} \leq \alpha)$ if it exists. The decreased critical value is defined as $\alpha_{FDR} = p_{(k^*)}$ or $\alpha_{FDR} = 0$ in the case of k^* non-existence. Finally, all hypotheses satisfying $p_k \leq \alpha_{FDR}$ are rejected, which is statistically correct as proven in [18].

The FDR technique is employed in this novel approach as a very sensitive tool to localize significant segment dependencies. This approach is used for the design of novel Causality Index of relative channel synchronization.

Let u be patient index, AD, CN be sets of diseased and control patients, and let us denote N_u, $N^*_{i,j,u}$ as the number of all segments of u-th patient and the number of significant segment dependences of ch_i on ch_j of u-th patient.

The *Causality Index* can be defined as the relative frequency of synchronized events

$$S_{i,j,u} = N^*_{i,j,u}/N_u. \tag{5}$$

Variable $S_{i,j,u} \in [0, 1]$ is a measure of synchronization from j-th to i-th channel for a given patient. The final hypothesis is focused on the Causality Index differences between AD and CN groups. For the fixed pair of channels ch_i, ch_j I test the hypothesis H_0 if the median of Causality Index differs, using Wilcoxon-Mann-Whitney (WMW) rank-sum test, again with FDR correction.

4 Data description

General approach was applied to the group of 26 patients with Alzheimer's disease (AD) and 139 control patients (CN). All subjects were recorded under the same resting protocol, i.e. eyes closed, lying on a bed. The standard 10-20 EEG system of electrode placement was used to obtain 21-channel digital EEG via TruScan 32. The sampling frequency was 200 Hz with 22 bit AD converter. Due to quasistationarity, the EEG signal was segmented to two-second segments of 400 samples for separate analysis. The total number of 24 742 segments from all patients were used for statistical investigation.

5 Results

The statistical analysis had three aims. The optimum order of VAR(p) model was investigated first. Then inter-channel causalities in individual segments were tested and segments with statistical significant causalities were localized. In the final step, the main

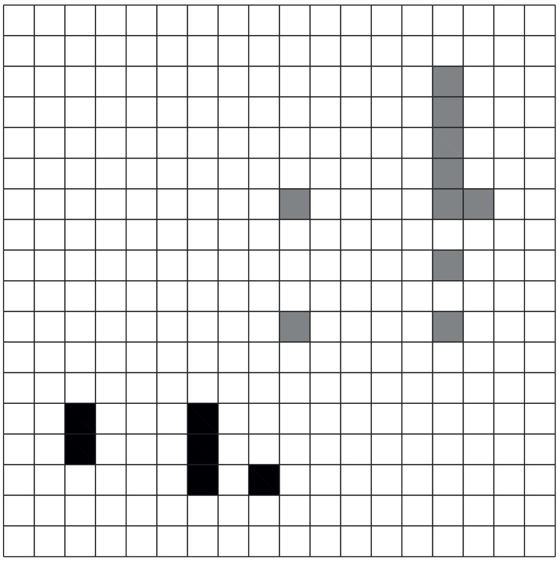


Figure 1: Significant increasing (grey) and decreasing (black) of Causality Index in Alzheimer's disease: consequent channels ch_i as rows, antecedent channels ch_j as columns.

issue, i.e. whether the Causality Index is affected by Alzheimer's disease, was the subject of multiple-testing.

5.1 Optimum length of VAR model

The first aim of separate segment processing was to estimate the optimum model order p_{opt} of VAR(p) model of dimension $k = 19$. Using non-overlapping 24 742 segments of all patients I minimized BIC(p) for $p \leq 100$. The optimum order varied from 11 to 48, and the most frequent value was $p_{\text{opt}} = 26$ as experimental modus. This value was postulated as the recommended model order for the consequent Granger causality investigations.

5.2 Significant channel dependencies

The total number of $19 \times 18 = 342$ channel pairs can potentially significantly interact in the case of 19-channel EEG. The F-test of hypothesis H$_0$: $\sigma^2_{\text{complete}} = \sigma^2_{\text{reduced}}$ was used for all pairs and 24 742 segments with $p = 26$ on critical level $\alpha = 0.05$. The resulting p-values were corrected by FDR to obtain decreased value $\alpha_{\text{FDR}} = 0.0023$. Significant combinations of channels and segments were labelled and counted to obtain Causality Indexes $S_{i,j,u}$.

5.3 Causality Index changes

Being focused on channel pair ch_i, ch_j, The H$_0$: $S_{i,j}(AD) = S_{i,j}(CN)$ hypothesis was tested, where $S_{i,j}(AD)$ is a median of $S_{i,j,u}$ for $u \in$ AD and $S_{i,j}(CN)$ is a median of $S_{i,j,u}$ for $u \in$ CN. The non-parametric WMW test of critical level $\alpha = 0.05$ was applied. The p-values resulting from 342 independent tests were corrected by FDR technique with $\alpha_{\mathrm{FDR}} = 7.3074 \times 10^{-4}$. Significant channel pairs with increasing or decreasing Causality Indexes in Alzheimer's disease are collected in Tab. 1. The dependencies between antecedent (rows) and consequent (columns) channels are depicted in Fig. 1. For better biomedical interpretation, the traditional EEG 10-20 scheme is used to show channel pairs with significant dependencies in Figs 2, 3.

Table 1: Significant changes of Causality Index

i	j	$\tilde{\mathrm{S}}_{AD}$	$\tilde{\mathrm{S}}_{CN}$	p-value
7	10	0.9677	0.7097	3.01×10^{-5}
11	10	0.6882	0.4301	6.35×10^{-5}
3	15	0.9839	0.8495	2.73×10^{-6}
4	15	0.9462	0.7742	2.34×10^{-5}
5	15	0.8763	0.7097	2.73×10^{-4}
6	15	0.9140	0.7312	8.43×10^{-6}
7	15	0.9839	0.7957	7.49×10^{-7}
9	15	0.9194	0.6882	9.14×10^{-5}
11	15	0.8172	0.5806	1.48×10^{-6}
7	16	0.9785	0.8495	2.58×10^{-4}
14	3	0.5000	0.7312	5.48×10^{-4}
15	3	0.5376	0.7849	2.83×10^{-5}
14	7	0.6828	0.8710	4.61×10^{-4}
15	7	0.6344	0.8602	2.08×10^{-4}
16	7	0.6183	0.8065	1.99×10^{-4}
16	9	0.5108	0.6667	8.25×10^{-5}

5.4 Biomedical interpretation

As seen in Tab. 1, there are many significant changes in the Causality Index. The lowest p-value was observed for the pair of 7th and 15th channels. This pair can be used for distinguish between AD and CN. Using rule $S_{7,15,u} > 0.92$ for the diagnosis of AD in the case of uth participant, the sensitivity and specificity were 73 % and 77 % respectively. Similar behaviour was also observed on the other significant channel pairs.

During Alzheimer's disease, the Causality Index exhibits very interesting changes. The significant increase in the Causality Index (Fig. 2) points from parietal to frontal regions of the brain. In AD, it means that neural activity in the frontal lobes is highly activated from the parietal zone. The opposite significant dependencies (Fig. 3) come from the left and right frontal lobes to the parietal zone. This behaviour can be interpreted as a decreasing Causality Index between the inspiring frontal neurons and receiving parietal zone. This interpretation is consistent with the concept of the dynamics of changes in the course of a developing Alzheimer's disease [15].

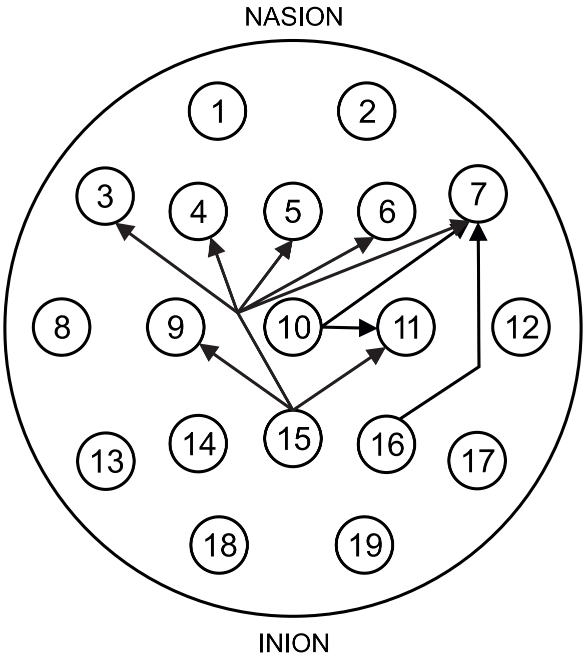


Figure 2: Causality increasing in the case of Alzheimer's disease: arrows from antecedent to consequent channels.

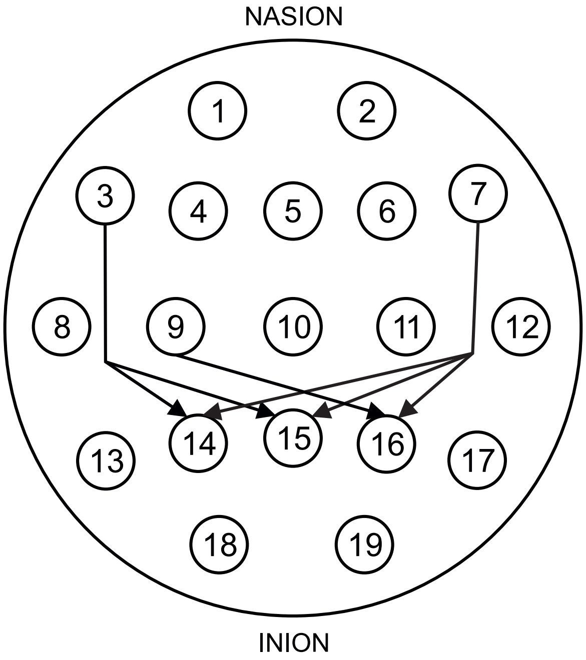


Figure 3: Causality decreasing in the case of Alzheimer's disease: arrows from antecedent to consequent channels

6 Conclusions

The theory of VAR model was applied to multichannel EEG. The optimal order was 26 as an experimental modus value. Significant interchannel causalities were obtained over segments of patients. The False Discovery Rate correction was used as an efficient tool for selecting significant EEG events. The event counting forms a novel Causality Index as a criterion able to distinguish between AD and CN. 10 significant electrode pairs were observed with decreasing Causality Index and 6 electrode pair with increasing Causality Index in AD. The difference in Causality Indexes can help in diagnosing Alzheimer's dementia. Interchannel dependencies observed exhibiting statistically significant changes in the Causality Index have direct biomedical interpretation. In AD, there is a significant increase in the Causality Index between the parietal and frontal domains of the brain. The complementary effect of decreasing Causality Index was also localized, however the direction was opposite.

References

[1] D. Abásolo, R. Hornero, P. Espino, D. Alvarez, and J. Poza. *Entropy analysis of the eeg background activity in alzheimer's disease patients.* Physiological measurement **27** (2006), 241.

[2] Y. Benjamini and Y. Hochberg. *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the royal statistical society. Series B (Methodological) (1995), 289–300.

[3] R. P. Brenner, R. F. Ulrich, D. G. Spiker, R. J. Sclabassi, C. F. Reynolds, R. S. Marin, and F. Boller. *Computerized eeg spectral analysis in elderly normal, demented and depressed subjects.* Electroencephalography and clinical neurophysiology **64** (1986), 483–492.

[4] S. L. Bressler and A. K. Seth. *Wiener–granger causality: A well established methodology.* Neuroimage **58** (2011), 323–329.

[5] P. T. Francis, A. M. Palmer, M. Snape, and G. K. Wilcock. *The cholinergic hypothesis of alzheimer's disease: a review of progress.* Journal of Neurology, Neurosurgery & Psychiatry **66** (1999), 137–147.

[6] A. Gelman, J. Hwang, and A. Vehtari. *Understanding predictive information criteria for bayesian models.* Statistics and Computing **24** (2014), 997–1016.

[7] R. Goebel, A. Roebroeck, D.-S. Kim, and E. Formisano. *Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping.* Magnetic resonance imaging **21** (2003), 1251–1261.

[8] C. W. Granger. *Investigating causal relations by econometric models and cross-spectral methods.* Econometrica: Journal of the Econometric Society (1969), 424–438.

[9] S. Johansen. *Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models.* Econometrica: Journal of the Econometric Society (1991), 1551–1580.

[10] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler. *Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance.* Biological cybernetics **85** (2001), 145–157.

[11] T. König, L. Prichep, T. Dierks, D. Hubl, L. Wahlund, E. John, and V. Jelic. *Decreased eeg synchronization in alzheimer's disease and mild cognitive impairment.* Neurobiology of Aging **26** (2005), 165–171.

[12] H. Lütkepohl. *Vector autoregressive models.* Springer, (2011).

[13] J. C. McBride, X. Zhao, N. B. Munro, G. A. Jicha, F. A. Schmitt, R. J. Kryscio, C. D. Smith, and Y. Jiang. *Sugihara causality analysis of scalp eeg for detection of early alzheimer's disease.* NeuroImage: Clinical **7** (2015), 258–265.

[14] J. C. McBride, X. Zhao, N. B. Munro, C. D. Smith, G. A. Jicha, L. Hively, L. S. Broster, F. A. Schmitt, R. J. Kryscio, and Y. Jiang. *Spectral and complexity analysis of scalp eeg characteristics for mild cognitive impairment and early alzheimer's disease.* Computer methods and programs in biomedicine **114** (2014), 153–163.

[15] S. Sanei and J. A. Chambers. *EEG Signal Processing.* John Wiley & Sons, (2013).

[16] C. A. Sims. *Macroeconomics and reality.* Econometrica: Journal of the Econometric Society (1980), 1–48.

[17] C. Stam, B. Jones, G. Nolte, M. Breakspear, and P. Scheltens. *Small-world networks and functional connectivity in alzheimer's disease.* Cerebral cortex **17** (2007), 92–99.

[18] K. J. Verhoeven, K. L. Simonsen, and L. M. McIntyre. *Implementing false discovery rate control: increasing your power.* Oikos **108** (2005), 643–647.

Iterative Reconstruction of Compressed Sensed MRI Data

Hynek Walner

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

advisors:
Jiří Boldyš, Department of Image Processing
Institute of Information Theory and Automation, CAS
Michal Šorel, Department of Image Processing
Institute of Information Theory and Automation, CAS
Jiří Dvořák, Department of Image Processing
Institute of Information Theory and Automation, CAS

Abstract. Iterative reconstruction techniques find their used in many optimization problems, such as matrix completion in computer vision or reconstruction in image processing. Brief introduction to iterative algorithm based on proximal gradient method will be presented together with connection to the image reconstruction problem. Furthermore, reconstruction of the subsampled (compressed) medical data will be formulated as a variational problem using total variation regularization, ready to be solved using presented method. Finally, we will demonstrate and compare selected methods on real data acquired from MRI scanner at ISI of the CAS in Brno and propose further extension of current model.

Keywords: image reconstruction, TV regularization, proximal algorithms

Abstrakt. Iterativní rekonstrukční metody jsou často využívány v mnoha optimalizačních úlohách jako například doplnění dat ve strojovém učení nebo rekonstrukci obrazu. Provedeme krátké shrnutí iterativního algoritmu založeném na vyhodnocení proximálního operátoru a ukážeme jeho vazbu na úlohu rekonstrukce obrazu. Dále formulujeme rekonstrukci podsamplovaných zdravotnických dat jakožto úlohu variačního počtu s regularizací ve tvaru totální variace v takovém tvaru, aby byla řešitelná uvedenou metodou. Nakonec vybrané algoritmy předvedeme a srovnáme na datech ze skeneru využívající magnetickou rezonanci umístěného na ÚPT AV ČR v Brně a navrhneme další rozšíření modelu.

Klíčová slova: rekonstrukce obrazu, TV regularizace, proximální algorithmy

1 Introduction

Many inverse imaging problems such as image denoising, image deconvolution or image signal reconstrucion can be conveniently formulated as a variational problem

$$\min_{x \in \mathbb{R}^2} \left\{ \lambda \int_{\Omega} |K(x)| + \frac{1}{2} \|y - Ax\|_2^2 \right\}, \tag{1}$$

where $\Omega \subset \mathbb{R}^2$ is image domain, $x \in L^1(\Omega)$ is the desired solution and $y \in L^1(\Omega)$ is the original data which are to be reconstructed. Parameter $\lambda \in \mathbb{R}_0^+$ scales the trade-off between "data" term and regularization term. Data term ensures closeness of the solution and the input, whereas regularization represents effort to improve visual features of an image. Operator A represent transformation of output to comparable domain in which y is acquired. In basic case of medical imaging, A typically denotes Fourier transformation. Modern imaging techniques relies on methods of compressed sensing (CS), where only selected samples of Fourier domain are taken into account, rather than sampling at the full (i.e. Nyquist) rate. Usually, matrix A also models trajectory of given samples and multi-coil sensitivites for more realistic models. If K is assumed to be gradient of input image, proposed model (1) becames so-called Total Variation (TV) regularization model (or ROF model) introduced in [1]. Major advantage of incorporating TV regularization is allowing appearance of sharp discontinuities in the solution. This fact is often sought after in image processing, since edges represent important features such as boundaries of objects. However this formulation of cost functional (1) leads to difficult minimization, given the non-smooth property of the total variation. We will introduce used algorithm based on proximal operators, which can be successfully used to tackle such problems with application to MRI data reconstruction.

2 Iterative Reconstruction Technique

Algorithms based on evaluating proximal operator can be percieved as a generalization of standard gradient descent. We will briefly introduce main idea of this technique and present its use in iterative method to solve optimization problem (1).

2.1 Proximal Operator

Let us suppose, that we want to solve

$$\min_{x \in \mathbb{R}^n} f(x) = \min_{x \in \mathbb{R}^n} g(x) + h(x) \tag{2}$$

where $g : \mathbb{R}^n \mapsto \mathbb{R}^n$ is convex and differentiable while $h : \mathbb{R}^n \mapsto \mathbb{R}^n$ is only convex but not necessarily differentiable. Instead of making quadratic approximation of f around x with step size $t \in \mathbb{R}^+$ to get gradient descent update for the case of g and h both convex and differentiable, it is possible to approximate only g while h stays in its original form. We obtain following

$$\begin{aligned}
x^+ =& \operatorname*{argmin}_z \left\{ g(x) + \nabla g(x)^T (z - x) + \frac{1}{2t} \|z - x\|_2^2 + h(z) \right\} \\
=& \operatorname*{argmin}_z \left\{ \frac{1}{2t} \left(\|z - x\|_2^2 + 2t \nabla g(x)^T (z - x) + t^2 \|\nabla g(x)\|_2^2 \right) + g(x) - \frac{2}{t} \|\nabla g(x)\|_2^2 + h(z) \right\} \\
=& \operatorname*{argmin}_z \left\{ \frac{1}{2t} \|z - (x - t \nabla g(x))\|_2^2 + h(z) \right\} \\
=& \operatorname{prox}_{t,h}((x - t \nabla g(x))),
\end{aligned}$$

where we denoted minimizing term by the symbol prox. Components in prox forces update to be as close to gradient step of g as possible and keeps values of h small. Using this intuitive derivation, we can formally define *proximal operator* $\mathrm{prox}_{t,h} : \mathbb{R}^n \mapsto \mathbb{R}^n$ by

$$\mathrm{prox}_{t,h}(x) = \underset{z}{\mathrm{argmin}} \left\{ \frac{1}{2t} \|z - x\|_2^2 + h(z) \right\}.$$

Combining proximal operator with gradient descent, leads to writing minimizing algorithm of (2) as

Algorithm 1 General proximal operator minimization

```
1. Initialize $x^0 \in \mathbb{R}^n$.
2. Let $x^+ = (x^{k-1} - t\nabla g(x^{k-1}))$.
3. Define $x^k = \mathrm{prox}_{t_k,h}(x^+)$.
```

The last step of **Algorithm 1** can be also written in the gradient descent manner as

$$x^k = x^{k-1} - t_k G_{t_k}(x^{k-1}), \quad G_t(x) = \frac{x - \mathrm{prox}_{t,h}(x - t\nabla g(x))}{t},$$

where $G_t(x)$ is so-called *generalized gradient.* Notice that the evaluation of proximal operator depends only on the gradient of g and h itself, thus it can be conveniently used when proximal operator of h is known. Especially, this is the case of $h = \lambda \| \cdot \|_1$, where respective proximal operator is of form

$$\mathrm{prox}_{\mathrm{t},\lambda\|\cdot\|_1}(x) = \underset{z}{\mathrm{argmin}} \left\{ \frac{1}{2t} \|z - x\|_2^2 + \lambda \|z\|_1 \right\}. \tag{3}$$

The solution to this equation can be written as a *soft thresholding operator* $S_{\lambda t}(x)$ where

$$S_{\lambda t}(x) = \mathrm{sgn}(x)(|x| - \lambda t)_+.$$

It can be easily shown, that $S_{\lambda t}(x)$ minimizes term in (3) and is easily numerically computed.

2.2 Alternating Direction Method of Multipliers

Following algorithm employs alternating minimization of objective functions to tackle variational problems with non-smooth regularization. Such method is called Alternating Direction Method of Mutlipliers (ADMM) and is built on minimizing each function from

$$\min_{x \in \mathbb{R}^n} g(x) + h(x)$$

separately. This technique is known as dual minimization or Douglas-Racheford splitting and its main advantage is when evaluating proximal operator of $f + g$ is more numerically demanding, than computing each proximal operator separately. We will now derive solution to (1) using this method.

Formal steps of ADMM algorithm originates from minimizing augmented Lagrangian [2]. Firstly, rewrite original problem (1) as a constrain optimization

$$\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|z\|_1 \quad \text{s.t.} \quad Kx - z = 0.$$

Furthermore, we write augmented Lagrangian of such problem as

$$L_\varrho(x, z, u) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|z\|_1 + \rho u^T(Kx - z) + \frac{\rho}{2}\|Kx - z\|_2^2, \tag{4}$$

where constant $\varrho > 0$ is called penalty parameter. Notice, that additional terms equal to zero at optimal point by definiton of constraint $Kx - z = 0$. Minimizing of augmented Lagrangian (4) is treated separately over its primal variables x and z, therefore we can write ADMM algorithm in following manner

Algorithm 2 ADMM

```
1. Initialize $x^0, u^0, z^0 \in \mathbb{R}^n$, $\rho \in \mathbb{R}^+$.
2. Let $x^k = \underset{x}{\mathrm{argmin}} L_\varrho(x^k, u^k, z^k) = \underset{x}{\mathrm{argmin}} \left\{ \frac{1}{2}\|y - Ax^k\|^2 + \varrho {u^k}^T K x^k + \frac{\varrho}{2}\|Kx^k - z^k\|^2 \right\}$.
3. Let $z^k = \underset{z}{\mathrm{argmin}} L_\varrho(x^k, u^k, z^k) = \underset{z}{\mathrm{argmin}} \left\{ \lambda\|z^k\|_1 - \varrho {u^k}^T z^k + \frac{\varrho}{2}\|Kx^k - z^k\|^2 \right\}$
4. Update $u^k = u^k + \varrho(x^k - z^k)$.
```

Finding optimal value $x^\star$ in step 2 of **Algorithm 2** can be easily attained using partial derivative of L_ϱ over x in closed-form solution

$$x^\star = (A^T A + \varrho K^T K)^{-1}(A^T y + \varrho K^T(z - u)).$$

To find optimal $z^\star$ one can successfully use evaluation of proximal algorithm, namely soft thresholding operator defined in previous section. We can write

$$z^\star = S_{\lambda/\rho}(u^k + Kx^k).$$

Finally, dual variable u is updated by gained values of constrain to conclude current iteration. Notable feature of ADMM is, that it converges fast at early stage, but requires fair number of iterations for high precision results.

3 MRI Data Reconstruction

Let us now closely describe acquired data that were used in reconstruction and exactly formulate model to simulate measurement and reconstruction.

3.1 Data Description

Data originate from in-vivo experiment with a standard Sprague-Dawley rat at the Bruker 9.4T MRI Small Animal Scanner stationed at the Institute of Scientific Instruments of the CAS. MRI scanner collects signal in k-space (i.e. Fourier domain) and due to the physical constraints of the scanner, data are sampled alongside the radial trajectories. Radials are rotated through the space using *golden angle* technique allowing relatively dense sampling of important regions of k-space [5]. MRI machine compound aquisition from 4 coils and returns 128 complex coefficients of k-space for each coil and radial. During the experiment time of 14 minutes 50 000 projections of 128 coefficients were sampled. In all formulations, coil sensitivities were estimated using ESPIRiT algorithm proposed in [7].

3.2 Formulation of Reconstruction Problem

Firstly, we will omit the element of time for simpler notation and write formulation of reconstruction of *static* data as

$$\min_{x \in \mathbb{C}^2} \left\{ \sum_{i=1}^{4} \frac{1}{2} \|y_i - MFS_i x\|_2^2 + \lambda \|Kx\|_1 \right\}, \tag{5}$$

where S_i maps sensitivity of coil i, F corresponds to the 2D Fourier transform and M interpolates cartesian grid to the sampled radial space. Matrix M together with F can be replaced with *non-uniform* Fourier transformation. Regularization term is in form of Total Variation, therefore K computes gradient of the image.

This formulation can be extended to reconstruct *dynamical* data as

$$\min_{x_t \in \mathbb{C}^2} \left\{ \sum_{t=0}^{T-1} \sum_{i=1}^{4} \frac{1}{2} \|y_{t,i} - M_t F S_i x_t\|_2^2 + \lambda \|K x_t\|_1 \right\}. \tag{6}$$

It is worth noting, that instead of estimating output image for each time-frame separately (as can be achieved iterating static case through all data), this formulation optimizes coefficients of given basis. Let us elaborate more explicitly for the case of modeling dynamics as a polynomial of 2nd degree

$$p(t) = a + bt + ct^2.$$

If plugged into the (6) for x_t we get

$$\min_{a,b,c, \in \mathbb{C}^2} \left\{ \sum_{t=0}^{T-1} \sum_{i=1}^{4} \frac{1}{2} \|y_{t,i} - M_t F S_i (a + bt + ct^2)\|_2^2 + \lambda \|K(a + bt + ct^2)\|_1 \right\}.$$

and reformulated to more compact and tidy matrix notation

$$\min_{x \in \mathbb{C}^2} \left\{ \sum_{t=0}^{T-1} \sum_{i=1}^{4} \frac{1}{2} \|y_{t,i} - M_t F S_i B_t x^T\|_2^2 + \lambda \|K B_t x^T\|_1 \right\},$$

where $B_t = [I\ It\ It^2]$ and $x = [a\ b\ c]$. It is clear, that we can choose various forms of prescribed basis B_t with coefficients x and we will present results of different options in following sections.

Finally, let us briefly note the formulation of Low Rank + Sparse model (L+S) which will be also evaluated in results section. L+S model can be written as

$$\min_{L,S \in \mathbb{C}^2} \left\{ \frac{1}{2} \|y - MFS(L + S)\|_2^2 + \lambda_L \|L\|_* + \lambda_S \|KS\|_1 \right\}, \tag{7}$$

and estimates output image as a sum of two components: low-rank regularized by nuclear norm and sparse regularized by TV norm. Used implementation of L+S model uses non-uniform Fourier transformation together with density compensation technique (coupled in matrix MF) rather than radial interpolation M with uniform Fourier transform F developed in ADMM formulation. For further detail see for example [8].

4 Results Comparison

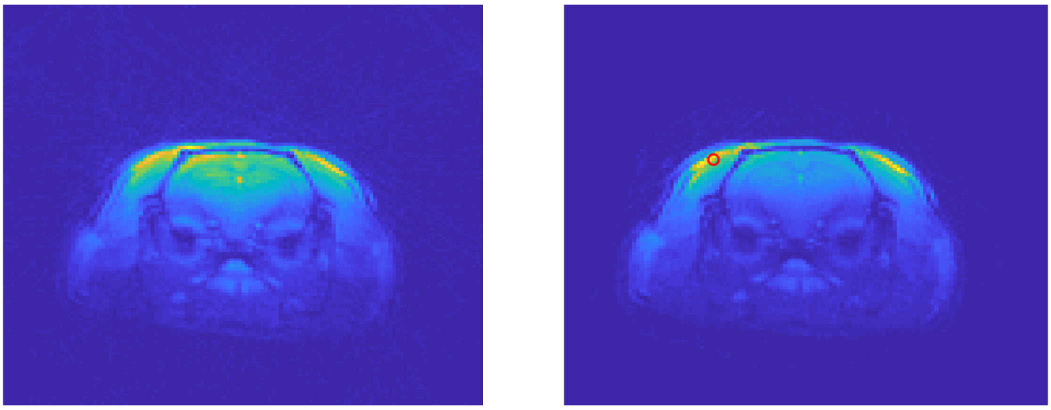
We will now present reconstructed data and several different approaches to attain the most realistic outcome. All ADMM results share the same parameters $\lambda = 1$ and $\varrho = 0.1$, L+S algorithm was used with setting $\lambda_L = 0.025$ and $\lambda_S = 0.5$.

4.1 Regridding and Reconstruction

Simple method how to transform measured signal into image domain is called *regridding* and it is direct, non-iterative approach. Regridding is one-off application of operator A, i.e. matrices M, F and S in our formulation, to the input data. No regularization is employed and it can be easily seen (Figure 1), that this operation suffers from artifacts when compared to the results of the ADMM algorithm on static formulation (5). Namely, notice the streak-like artifacts that originate from radial sampling. Both results were obtained using 200 projections per one frame. Decreasing number of used projections increases temporal-resolution of outcome image (as 60 projections takes roughly 1 second of measurement) but brings significant degradation of image quality (at least in static formulation), as is shown on Figure 2.

4.2 Perfusion Curves

The measured data are not the same during the whole experiment, intensity of signal varies on time and body tissue. One of the main objectives of reconstruction is to get this function of intensity on time (so called *perfusion curve*) as detailed and realistic as possible. Typical perfusion curve has sharp increase at the beginning of the measurement (corresponding to the increased activity of contrast agent) followed by slow decline. Perfusion curves of static reconstruction (i.e. separately reconstructed image through whole data) together with several selected outcomes of dynamic formulation (6) is shown on Figure 3. Selected pixel is marked by red dot in Figure 1 b). Prescribed basis for dynamical formulation was estimated orthogonal basis using singular-value decomposition method on various perfusion curves from static case. In presented case, first 3 singular curves



(a) Regridding.

(b) Static reconstruction.

Figure 1: Comparison of regridding and iterative reconstruction.

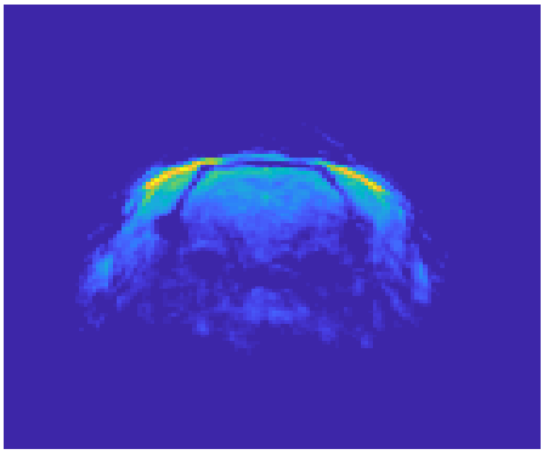


Figure 2: Iterative reconstruction using 28 projections.

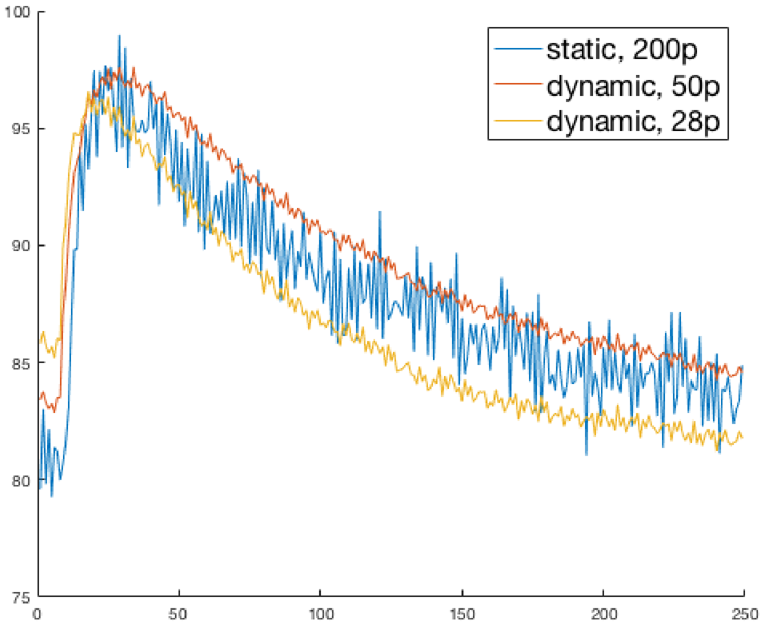


Figure 3: Comparison on static reconstruction using 200 samples and dynamic from 50, or 28 samples.

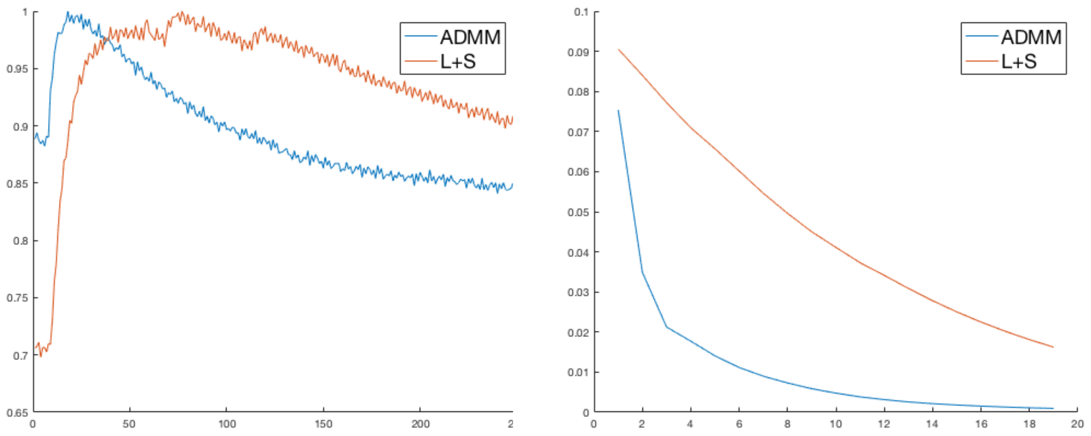
were used. It can be seen, that prescribing basis for dynamic reconstruction can lead to improved stability of perfusion curve and possibility to further reduce used samples, thus increase temporal resolution.

4.3 Comparison with L+S

Measured data were also processed by different formulation of reconstruction problem, the Low Rank +Sparse model (7). L+S model assumes, that perfusion curve consists of one component with low rank and other, that is sparse in Fourier domain. Figure 4 shows comparison of perfusion curve reconstructed from 28 projections per one frame and relative improvement of cost functional in each iteration. Convergence comparison agrees with standard ADMM feature of high convergence speed, namely in the first iterations. Perfusion curves were rescaled by maximum of each curve and prompt to say, that L+S model estimates somewhat more stable decline after the growth phase. It is worth noting, that final rank of L+S model was one, whereas ADMM was the most stable at basis of rank 2 and 3. Nevertheless, these differences are up to more detailed research.

5 Conclusion

We have introduced main idea of proximal operator and demonstrated its application on iterative recontruction of MRI data. Two different formulation of reconstruction problem were shown and results on real data we demonstrated. Further work lies in modeling



(a) Perfusion curves.

(b) Relative improvement of cost.

Figure 4: Comparison of perfusion curve and convergence of ADMM and L+S model

acquisition process in greater detail and developing faster reconstruction techniques to increase both temporal and spatial resolution of outcoming images. This should lead to more reliable perfusion analysis of outcoming data and to improve diagnostics of vascular diseases affecting myocardium, brain and other organs, as well as cancer diseases in the long-term.

References

[1] L. Rudin, S. J. Osher and E. Fatemi. *Nonlinear total variation based noise removal algorithms.* Physica D., 60. (1992), 259–268.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* Foundations and Trends in Machine Learning, Vol. 3, No. 1 (2010) 1–122

[3] N. Parikh, S. Boyd. *Proximal Algorithms.* Foundations and Trends in Optimization, Vol. 1, No. 3 (2013), 123–231.

[4] A. Chambolle, T. Pock. *A first-order primal-dual algorithm for convex problems with applications to imaging.* T. J Math Imaging Vis, 40:120, (2011).

[5] S. Winkelmann, T. Schaeffter, T. Koehler, H. Eggers and O. Doessel. *An Optimal Radial Profile Order Based on the Golden Ratio for Time-Resolved MRI.* IEEE Transactions on medical imaging, Vol. 26, No. 1. (2007).

[6] S. Sykora. *K-space formulation of MRI.* Stan's Library, Vol.I. (2005).

[7] M. Uecker, P. Lai, M. J. M, P. Virtue, M. Elad,J. M. Pauly, S. S. Vasanawala, M. Lustig. *ESPIRiT — An Eigenvalue Approach to Autocalibrating Parallel MRI: Where SENSE meets GRAPPA* Magn Reson Med. Mar; 71(3): 990–1001. (2014).

[8] M. Mangová, P. Rajmic and R. Jiřík. *Dynamic Magnetic Resonance Imaging using Compressed Sensing with Multi-scale Low Rank Penalty* IEEE 40th International Conference on Telecommunications and Signal Processing (TSP). (2017).

[9] M. V. Afonso, J.M. Bioucas-Dias. *Fast Image Recovery Using Variable Splitting and Constrained Optimization.* IEEE Transactions on image processing, Vol. 19., No. 9. (2010).